AIGC Image Quality Assessment via Image-Prompt Correspondence

Fei Peng[1] Huiyuan Fu[1*] Anlong Ming[1] Chuanming Wang[1]
Huadong Ma[1] Shuai He[1] Zifei Dou[2] Shu Chen[2]
[1]Beijing University of Posts and Telecommunications, China
[2]Beijing Xiaomi Mobile Software Co., Ltd.
{pf0607, fhy, mal, wcm, mhd, hs19951021}@bupt.edu.cn
{douzifei, chenshu1}@xiaomi.com

Abstract

In the rapidly evolving landscape of deep learning, generative models such as Generative Adversarial Networks (GANs) and diffusion models have significantly advanced the capabilities of Artificial Intelligence Generated Content (AIGC). These technologies have streamlined the creative process, enabling AI to autonomously produce a diverse range of content with minimal human input. Despite the remarkable progress in AI-generated images (AIGIs), evaluating the quality of AIGIs remains a complex challenge. Traditional image quality assessment (IQA), focusing on aspects like distortion and blurriness, are insufficient for capturing the correspondence between AIGIs and their prompts. To address this, we propose a novel AIGC image quality assessment (AIGCIQA) framework that emphasizes the correspondence between images and prompts. Utilizing the CLIP model's pre-trained image and text encoders, our method effectively measures the correspondence between visual and textual inputs. By transforming the assessment into classification probabilities and subsequently into a precise regression task, our method enhances the CLIP model's performance in AIGCIQA. Our method's effectiveness is confirmed by its first place in the image track of the NTIRE 2024 Quality Assessment for AI-Generated Content challenge and its state-of-the-art (SOTA) performance on benchmark datasets AGIQA-1K, AGIQA-3K, and AIGCIQA2023. This research represents a significant advancement in the field, offering an efficient and versatile tool for the evaluation of AIGIs and contributing to the ongoing development of AIGC technologies. Our codes are available at https://github.com/pf0607/IPCE.

1. Introduction

In recent years, the field of generative models within deep learning has seen an extraordinary evolution, particularly with the advent of Generative Adversarial Networks [6, 11, 13, 15, 32] (GANs) and diffusion models [7, 17, 37, 38, 58]. This progress has propelled Artificial Intelligence Generated Content (AIGC) to the forefront of technological innovation, capturing the interest of the computer science community with its significant contributions to both research and practical applications. AIGC has revolutionized the creative process by leveraging the sophisticated capabilities of deep learning to emulate the human approach to content creation. Unlike traditional methods, which require extensive human involvement, contemporary AIGC simplifies the process by merely providing a prompt to an AI generative model. This model then autonomously produces a wide range of content, spanning from texts to intricate images, audio, and dynamic videos. This automated creation process has significantly enhanced efficiency and has democratized creativity, making it more accessible to a broader range of users.

In the realm of image generation, AIGC has made significant strides, with groundbreaking works such as DALLE [34], Imagen [38] and Stable Diffusion [35], showcasing the potential for generating high-quality images. However, the absence of effective supervision often results in a perceptual divide between AI-generated images (AIGIs) and human expectations. Evaluating the quality of AIGIs, including the correspondence with the original prompts and the quality of AIGIs, presents a complex challenge. The accurate assessment of image quality is pivotal for refining these models and remains a critical area of focus. Although existing research [5, 10, 14, 28–30, 41, 42, 46–48, 53, 57, 59, 60] has achieved significant progress in image quality assessment (IQA) by focusing on conventional metrics such as distortion, blurriness, and resolution, these methods often overlook the context provided by prompts when directly inputting images and regressing scores. This oversight limits their effectiveness in accurately assessing the quality of AIGIs, which requires a nuanced evaluation of the correspondence between the generated images and their prompts.

*Corresponding author

To bridge this gap, we introduce a novel method to AIGC image quality assessment (AIGCIQA) that centers on the correspondence between images and prompts. By utilizing the CLIP [33] model, we process the AIGIs and prompts through the model's respective image and text encoders, thereby generating corresponding visual and textual embeddings. This method benefits from the CLIP model's extensive contrastive learning pre-training on vast image-text datasets, which aligns the embeddings within a shared vector space. The cosine similarity calculation between these embeddings quantifies the degree of correspondence between the image and text. Drawing inspiration from previous research[16, 20, 44, 49, 51, 60] employing the language-image model for IQA tasks, we devise text templates that articulate various levels of correspondence between the prompts and images. We further transform the evaluation of image-text correspondence into classification probabilities and then convert the final score calculation, through weighted summation, into a precise regression task, thereby enhancing the CLIP model's efficacy in the domain of AIGCIQA.

Our proposed method, which we name Image-Prompt Correspondence Estimator (IPCE), is validated through its triumph in the image track of the NTIRE 2024 Quality Assessment for AI-Generated Content challenge [27], where it wins the first place. Moreover, our method achieves state-of-the-art (SOTA) performance across existing AIGCIQA datasets, including AGIQA-1K [61], AGIQA-3K [20], and AIGCIQA2023 [45], as demonstrated through rigorous experimentation. These achievements underscore the effectiveness and versatility of our method in evaluating the quality of AIGIs, marking a contribution to the ongoing development of AIGC technologies.

The contributions of this paper can be summarized as follows:

- **AIGCIQA Method based on Image-Prompt Correspondence**: This paper introduces a method that focuses on the correspondence between AIGIs and their prompts, which is achieved by utilizing the CLIP model's image and text encoders to generate embeddings, which are then compared using cosine similarity to quantify correspondence.
- **Integration of Classification and Regression for Assessment**: The research pioneers a method that transforms the assessment of image-text correspondence into classification probabilities and further refines the score calculation into a precise regression task, enhancing the efficiency of the AIGCIQA task.
- **Validation and Benchmark Performance**: The proposed method has been rigorously validated and proven effective through its first place in the NTIRE 2024 challenge and its SOTA performance on benchmark datasets such as AGIQA-1K, AGIQA-3K, and AIGCIQA2023, demonstrating its versatility in AIGC quality assessment tasks.

The rest of this paper is structured as follows. In Sec. 2, we provide a brief overview of existing IQA methods and Language-Image Models. The proposed method is elaborated upon in Sec. 3, followed by experiments in Sec. 4. Lastly, Sec. 5 offers an overall conclusion.

2. Related Work

2.1. IQA

In recent years, the field of IQA has witnessed significant advancements. A multitude of effective methods [5, 10, 14, 28–30, 41, 42, 46–48, 53, 57, 59, 60] has emerged, which are widely applied across various benchmarks [3, 4, 10, 18, 25, 39]. The input for these methods typically consists solely of images, with models directly regressing to output scores, employing Spearman's rank correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) as evaluation metrics. The quality of images is usually correlated with various low-level indicators such as distortions and blurs. The rapid development in the field of AIGC has garnered the attention of researchers. Recently, several AIGCIQA datasets, including AGIQA-1K [61], AGIQA-3K [20], AIGCIQA2023 [45], and PKU-I2IQA[55], have been proposed. However, methods specifically tailored to the characteristics of AIGCIQA task remain relatively scarce. Current research primarily utilizes existing methods [9, 12, 30, 40, 41, 44], without considering the correspondence between the image and the prompt in the AIGCIQA task. Yuan et al. [55] propose an image-to-image (I2I) AIGCIQA method based on references, which is not applicable to I2T task. The PSCR [54] method introduced a contrastive regression approach, which also lacks consideration for the prompt. TIER [56] method concatenates text features and image features and inputs them into a multi-layer perceptron (MLP) to regress scores, considering both the prompt and the image, which aligns with the AIGCIQA task. However, the direct use of different encoders for text and image features may result in features that are distributed across vastly different vector spaces, which may pose a challenge for the network in assessing the similarity between these two types of features. ImageReward [51] In this paper, we employ the CLIP [33] model, which can map text and images to the same vector space, to calculate the correspondence between the image and the prompt for quality assessment. This approach simply and effectively considers the characteristics of the AIGCIQA task.

2.2. Language-Image Model

CLIP [33] has gained popularity due to its powerful image-text alignment capabilities acquired through contrastive learning. It can possess strong capabilities for downstream

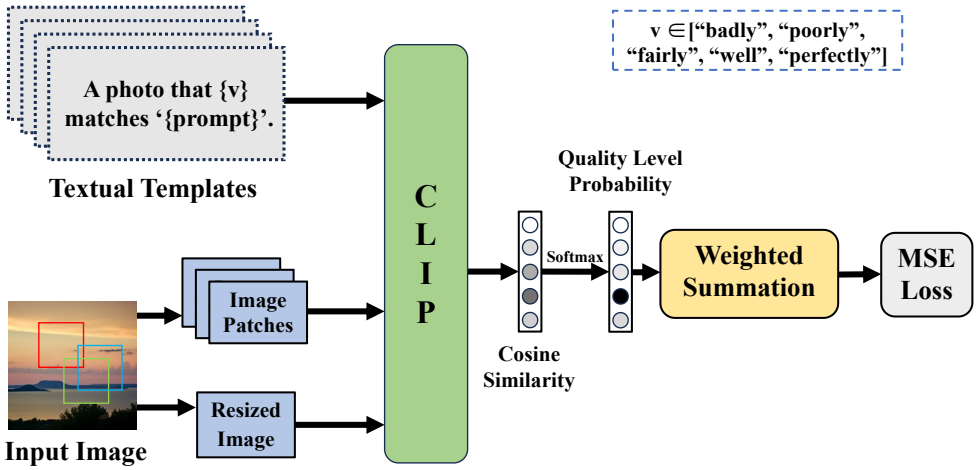

Figure 1. The proposed Image-Prompt Correspondence Estimator network framework.

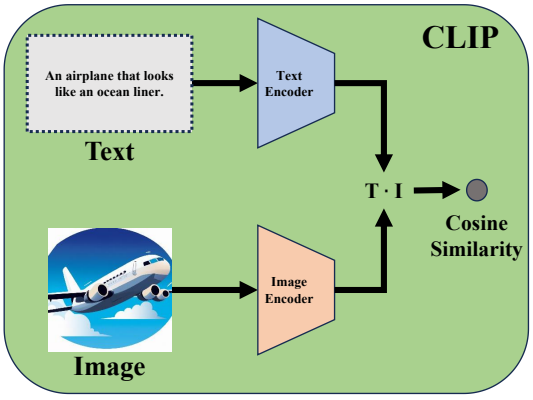

Figure 2. The structure of CLIP [33] model.

tasks [8, 19, 24, 50, 62] such as classification, semantic segmentation and object detection with appropriating textual input. Influenced by CLIP, several stronger language-image models have emerged, such as BLIP [22, 23], LLaVA [26], and MiniGPT-4 [64].

Zhang et al. [60], Wang et al. [44] introduce CLIP into the IQA task by setting different image quality level text templates. Wu et al. [49], Li et al. [20], Kirstain et al. [16] and Xu et al. [51] apply CLIP and BLIP to the AIGCIQA task, simulating human preferences by utilizing image-text similarity. Our method draws inspiration from theirs [16, 20, 44, 49, 51, 60], enhancing the application of CLIP in the AIGCIQA task by embedding prompt content into various designed text templates.

3. Method

3.1. Basic Framework

This study introduces the Image-Prompt Correspondence Estimator (IPCE), an advanced model structure depicted in Fig. 1. The structure of the CLIP [33] model is depicted in Fig. 2, comprising a text encoder and an image encoder. The text encoder adopts a transformer structure, while the image encoder offers two selectable architectures: convolutional neural network (CNN) and visual transformer (VIT). Through extensive contrastive learning of text-image pairs, CLIP is capable of mapping input image and text into the same vector space, thereby computing the cosine similarity between text and image. Let F_i be the extracted image feature and F_t be the extracted text feature, the calculation of their cosine similarity S is as follows:

$$S = \frac{F_t \odot F_i}{||F_t|| \cdot ||F_i||} \tag{1}$$

Our method involves feeding images and text templates describing different levels of correspondence between images and prompts into the CLIP model. This process calculates the cosine similarity between images and prompts at different levels of correspondence and then processes the discrete cosine similarities to obtain weighted continuous quality assessment scores. This method enables consideration of the correspondence between images and prompts in AIGCIQA.

3.2. Text Templates

Our method applies CLIP to the AIGCIQA task, which is not the first to utilize CLIP for IQA tasks. Let's briefly review previous CLIP-IQA [44] and LIQE [60] methods, and compare them with our method.

CLIP-IQA is proposed earlier to assess image quality from the perspective of simulating human perception using powerful language-image models. It employs a prompt-engineering approach by setting text templates describing different levels of image quality to calculate the similarity between images and different quality levels. CLIP-IQA

uses simple text templates as follows:

$$[\text{“}Good\ photo.\text{”}, \text{“}Bad\ photo.\text{”}]$$

Similar to CLIP-IQA, LIQE introduces more text templates of quality levels and fine-tuned the model. Let c be a replaceable adjective describing different quality levels, LIQE uses text templates as follows:

$$\text{“}A\ photo\ of\ \{c\}\ quality.\text{”},$$
$$c \in [\text{“}bad\text{”}, \text{“}poor\text{”}, \text{“}fair\text{”}, \text{“}good\text{”}, \text{“}perfect\text{”}]$$

In the AIGCIQA task, assessing the quality of an image requires considering the correspondence between the image and the prompt used to generate it. These methods aim to solve IQA tasks from the perspective of simulating human perception, although they introduce text information, fundamentally they are general IQA regression methods that only consider the quality of the image itself without incorporating prompt information. To introduce the calculation of correspondence between images and prompts, let v be a replaceable adverb describing different levels, we redesign the text templates as follows:

$$\text{“}A\ photo\ that\ \{v\}\ matches\ \textbf{'prompt'}\text{”},$$
$$v \in [\text{“}badly\text{”}, \text{“}poorly\text{”}, \text{“}fairly\text{”}, \text{“}well\text{”}, \text{“}perfectly\text{”}]$$

By modifying the text templates to describe different levels of correspondence with the prompt, we consider both the prompt and image information, which adapts to the AIGCIQA task.

For a given input image, assuming there are N (N=5 in this paper) different descriptions for quality levels, we can obtain the cosine similarity S_i (1 ≤ i ≤ N) of the image with each text template using Eq. (1). To further normalize the assessment results, we employ $Softmax$ option to convert the cosine similarities into quality probabilities that sum up to 1. Let P_i represent the quality probabilities, which can be calculated using the following formula:

$$P_i = \frac{e^{S_i}}{\sum_{k=1}^{N} e^{S_k}}, \quad (1 \le i \le N) \tag{2}$$

3.3. Image Segmentation

The image sizes in general IQA datasets may vary, and the common use of generalized resizing operations directly affects image quality because the resolution itself is also a component of image quality. The usual practice in existing IQA tasks is to segment images into multiple image patches using sliding windows, and then average the assessment results for each image patch. In the AIGCIQA task, segmenting images may compromise the semantic information contained in the images. Our method adopts the design of two types of image inputs: one part is the segmented image patches, and the other part is an additional resized whole image. The quality probabilities obtained from averaging the results of segmented image patches are further averaged with those of the resized whole image to obtain the final quality probability, balancing original image quality and whole image semantic information.

Suppose the image is segmented into multiple patches using a fixed window size (e.g., 224×224), and M patches are selected from them, along with one resized whole image. Then, by multiplying with N text templates, we obtain $N \times (M+1)$ quality probabilities. The quality probabilities obtained from calculating M image patches are denoted as P_{ij} (1 ≤ j ≤ M+1), and the quality probability for the resized whole image is denoted as P_i^*. Let the final combination be a set of quality weights W_i, which can be calculated as follows:

$$W_i = (\frac{\sum_{j=1}^{M} P_{ij}}{M} + P_i^*)/2 \tag{3}$$

3.4. Regression

After obtaining final weights W_i summing to 1 (1≤ i ≤5). Weight values V_i are set to 1-5 for the five correspondence levels. Finally, the predicted result Q is calculated as the weighted sum of W_i and V_i, with the following calculation formula:

$$Q = \sum_{i=1}^{5} W_i \times V_i, \tag{4}$$
$$V_i = i \quad (1 \le i \le 5)$$

LIQE uses the common IQA loss function fidelity loss [43], which does not require predicting scores as accurately as possible but only needs the order of predicted scores for different inputs to be correct. In the AIGCIQA task, the assessment results between different input prompt-image pairs are influenced by correspondence, making it difficult to directly compare the quality of two sets of inputs. Therefore, we optimize the problem into a precise regression task for score prediction, mapping the weighted scores to the actual score distribution range (0, 5), where the direct weighted score range in Eq. (4) is (1, 5). Then, we use Mean Absolute Error (MAE) as the loss function to allow the model to directly learn the image-prompt correspondence score distribution. Let Q_T be the ground truth score of input prompt-image pair, the fine-tuning loss L for the CLIP model can be calculated as follows:

$$L = |Q_T - (Q-1) \times \frac{5}{4}| \tag{5}$$

4. Experiments

Comparative experiments are conducted to showcase the efficacy of our IPCE model. We utilize three publicly avail-

able AIGCIQA datasets for both training and testing, to thoroughly assess the performance of the model under consideration. Ablation studies are meticulously carried out to evaluate the contribution of each component within the proposed model framework. Through a series of quantitative tests and analyses, we substantiate the effectiveness, robust performance, and distinct advantages that our proposed methodology offers in this section.

4.1. Datasets

We primarily validate our proposed method on datasets AGIQA-1K [61], AGIQA-3K [20], and AIGCIQA2023 [45]. Additionally, we also consider AGIQA-20K [21] dataset from the NTIRE 2024 Quality Assessment for AI-Generated Content challenge [27].

AGIQA-1K. The AGIQA-1K dataset contains 1080 AI-generated images from two Text-to-Image (T2I) models [35]: stable-inpainting-v1 and stable-diffusion-v2. It utilizes Mean Opinion Score (MOS) as the quality assessment annotation, which is obtained through manual scoring by human annotators.

AGIQA-3K. In the AGIQA-3K dataset, there are 2982 AIGIs generated by six T2I models, including GLIDE [31], Stable Diffusion [35, 36], Midjourney, AttnGAN [52], and DALLE2 [34]. This dataset is unique as it combines AIGIs from GAN, auto-regression, and diffusion-based models. It annotates MOS for both image quality and alignment between image and prompt.

AIGCIQA2023. The AIGCIQA2023 dataset includes 2400 images generated by six T2I models like Glide [31], Lafite [63], DALLE [34], and others [1, 35, 58]. Each prompt has four randomly generated images, resulting in a total of 2400 AIGIs across 100 prompts. It annotates MOS from three perspectives: image quality, authenticity, and correspondence.

AGIQA-20K. The AGIQA-20K dataset comprises 20,000 images generated by various state-of-the-art T2I models. It is divided into 70% for training, 10% for validation, and 20% for testing. It annotates MOS by combining image quality and alignment.

4.2. Evaluation Metrics

We primarily utilize two commonly used evaluation metrics for comparing model performance: Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC). These metrics are commonly employed to measure the correlation between two variables and are frequently used to evaluate the performance of models in ranking or regression tasks.

SRCC. SRCC measures the strength and direction of association between two variables by assessing the monotonic relationship between their ranks. It can be calculated as follows:

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

Where d_i represents the difference in ranks between the two variables. N is the number of observations in the sample.

PLCC. PLCC measures the strength and direction of a linear relationship between two continuous variables.It can be calculated as follows:

$$\mathrm{PLCC} = \frac{\sum_{i=1}^{N}\left(s_i-\bar{s}\right)\left(p_i-\bar{p}\right)}{\sqrt{\sum_{i=1}^{N}\left(s_i-\bar{s}\right)^2}\sqrt{\sum_{i=1}^{N}\left(p_i-\bar{p}\right)^2}} \tag{7}$$

Where s_i represents the i-th observed value in the sample, p_i represents the corresponding predicted value, $\bar{s}$ is the mean of the observed values, and $\bar{p}$ is the mean of the predicted values.

4.3. Implementation Details

For the AGIQA-1K, AGIQA-3K, and AIGCIQA2023 datasets, we randomly divide the data into training and testing sets at a ratio of 4:1. In the case of the AGIQA-3K dataset, we ensure that data with the same prompt falls into the same set. Each experiment is repeated 10 times, and the average of the final results is taken. For the AGIQA-20K dataset, we directly follow the dataset partitioning provided by the competition.

Considering the dataset sizes and to prevent overfitting, we use the ViT-B/32 as the image encoder for CLIP on the lightweight datasets AGIQA-1K, AGIQA-3K, and AIGCIQA2023. For the larger dataset AGIQA-20K, we employ the ViT-L/14 as the image encoder for CLIP.

During the training phase, we set the batch size to 16 and utilize the AdamW optimizer with an initial learning rate of 1×10^{-5}. We set the weight decay to 1×10^{-2} and employ a cosine annealing learning rate strategy, with the cosine function completing half a cycle every 5 updates. We use MAE as the loss function. Other implementation details are shown as follows:

- GPU: RTX 4090
- Operating System: Ubuntu 20.04
- CUDA: 11.3
- Language: Python 3.8
- Platform: PyTorch 1.10.0

4.4. Experimental Results

AGIQA-1K. To validate the effectiveness of our method and considering that deep learning methods generally outperform hand-crafted-based methods in current research, we compare our method with SOTA methods ResNet50 [9], StairIQA [42], and MGQA [46] on the AIGC-1K dataset.

Methods	Quality		Authenticity		Correspondence	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
WaDIQaM-NR [2]	0.4447	0.4996	0.3936	0.3906	0.3027	0.2810
CNNIQA [12]	0.7160	0.7937	0.5958	0.5734	0.4758	0.4937
VGG16 [40]	0.7961	0.7973	0.6660	0.6807	0.6580	0.6417
VGG19 [40]	0.7733	0.8402	0.6674	0.6565	0.5799	0.5670
ResNet18 [9]	0.7583	0.7763	0.6701	0.6528	0.5979	0.5564
ResNet34 [9]	0.7229	0.7578	0.5998	0.6285	0.7058	0.7153
IPCE(ours)	**0.8640**	**0.8788**	**0.8097**	**0.7998**	**0.7979**	**0.7887**

Table 1. Quantitative comparison on AIGCIQA2023 [45] dataset. Evaluated on the quality, authenticity and correspondence metrics. The best results are **bolded**.

Methods	MOS	
	SRCC	PLCC
ResNet50 [9]	0.6365	0.7323
StairIQA [42]	0.5504	0.6088
MGQA [46]	0.6011	0.6760
IPCE(ours)	**0.8535**	**0.8792**

Table 2. Quantitative comparison on AGIQA-1K [61] dataset. The best results are **bolded**.

The results are presented in Tab. 2. The results demonstrate that our method outperforms the others in terms of both SRCC and PLCC metrics.

AGIQA-3K. Since the AGIQA-3K dataset provides annotations for both image perception and alignment quality, we compare our method with general SOTA IQA methods DBCNN [30], CLIPIQA [44], CNNIQA [12], and HyperNet [41] on the perception metric and with image-text SOTA IQA methods CLIP [33], ImageReward [51], HPS [49], PickScore [16], and StairReward [20] on the alignment metric. The experimental results are shown in Tab. 3 and Tab. 4, respectively. Our method achieves the best SRCC and PLCC scores on both metrics, demonstrating its effectiveness for the AIGCIQA task.

AIGCIQA2023. On the AIGCIQA2023 dataset, we conduct experiments on quality, authenticity, and correspondence metrics, comparing our method with methods WaDIQaM-NR [2], CNNIQA [12], VGG16 [40], VGG19 [40], ResNet18 [9], and ResNet34 [9]. The experimental results are listed in Tab. 1, where our method achieves the best SRCC and PLCC scores on all three metrics. This indicates that our method performs well in evaluating various aspects of AIGC image quality.

AGIQA-20K. Our method participates in the image track of the NTIRE 2024 Quality Assessment for AI-Generated Content challenge, which aims to discover effective quality assessment algorithms for AI-generated images that align with human perception. The final test results of the challenge are presented in Tab. 5, listing the top 10 ranked participants. Our team achieves first place in the challenge

Methods	MOS_Perception	
	SRCC	PLCC
DBCNN [30]	0.8207	0.8759
CLIPIQA [44]	0.8426	0.8053
CNNIQA [12]	0.7478	0.8469
HyperNet [41]	0.8355	0.8903
IPCE(ours)	**0.8841**	**0.9246**

Table 3. Quantitative comparison on AGIQA-3K [20] dataset. Evaluated on the perception metric. The best results are **bolded**.

Methods	MOS_Alignment	
	SRCC	PLCC
CLIP [33]	0.5972	0.6839
ImageReward [51]	0.7298	0.7862
HPS [49]	0.6349	0.7000
PickScore [16]	0.6977	0.7633
StairReward [20]	0.7472	0.8529
IPCE(ours)	**0.7697**	**0.8725**

Table 4. Quantitative comparison on AGIQA-3K [20] dataset. Evaluated on the alignment metric. The best results are **bolded**.

Methods	Main Score
1st IPCE(Ours)	**0.9175**
2nd	0.9169
3rd	0.9157
4th	0.9138
5th	0.9091
6th	0.9087
7th	0.9065
8th	0.9044
9th	0.9023
10th	0.8835

Table 5. Quantitative comparison on AGIQA-20K [21] dataset from image track of the NTIRE 2024 Quality Assessment for AI-Generated Content challenge. This table only shows part of the participants and the best result is **bolded**.

Methods	MOS	
	SRCC	PLCC
IPCE w/ F	0.8469	0.8375
IPCE w/ M	0.8394	0.8727
IPCE w/ C	0.8261	0.8615
IPCE w/o R	0.8418	0.8724
IPCE	**0.8535**	**0.8792**

Table 6. Ablation studies of IPCE. Evaluated on AGIQA-1K [61] dataset and the best results are **bolded**. "w/ F/M/C" refers to replacing loss function with fidelity loss, MSE loss, and replacing the image encoder with CNN (ResNet50 [9]) architecture. "w/o R" refers to removing resized whole image.

based on the main score, which is calculated as the average of SRCC and PLCC scores.

4.5. Ablation Studies

To validate the effectiveness of our method, we conduct several ablation experiments. The experimental results are shown in Tab. 6. Firstly, we replace the loss function with fidelity loss and MSE loss, denoted as "w/ F" and "w/ M" respectively. As shown in Tab. 6, using fidelity loss leads to a significant decrease in the PLCC score by 0.0417, indicating that the direct regression approach is more suitable for the AIGCIQA task. After using MSE loss, SRCC and PLCC decrease slightly by 0.0141 and 0.0065 respectively, suggesting that MAE is more effective in ensuring the smoothness and robustness of the model in the AIGCIQA task.

We consider two options for the image encoder of the CLIP model: ViT and CNN. Our method exclusively utilizes the ViT architecture, but we also experiment with replacing ViT with the CNN architecture ResNet50 [9], denoted as "w/ C". As observed in Tab. 6, SRCC and PLCC decrease by 0.0274 and 0.0177 respectively, indicating that the ViT architecture is more suitable for extracting image features in the AIGCIQA task.

Finally, to demonstrate the effectiveness of using the complete semantic information provided by the whole image in the AIGCIQA task, we remove the resized whole image input as described in Sec. 3.3, denoted as "w/o R". The results in Tab. 6 show a certain degree of decrease in both SRCC and PLCC.

5. Conclusion

This paper introduces the Image-Prompt Correspondence Estimator (IPCE), a novel method for assessing the quality of AI-generated images (AIGIs). IPCE utilizes the CLIP model's capabilities to measure the correspondence between images and designed textual prompts templates, effectively addressing the unique challenges of AIGCIQA. Our method achieves state-of-the-art results on several benchmark datasets and win first place in the image track of the NTIRE 2024 Quality Assessment for AI-Generated Content challenge, as validated through extensive experiments and ablation studies. In summary, IPCE represents a significant step forward in the assessment of AIGC, providing a robust and efficient solution that aligns well with human perception. Future work will focus on refining this framework to further enhance its performance and applicability in the evolving landscape of AIGC.

ACKNOWLEDGMENTS

This work is supported in part by the NSFC under No.62272059, the National Key R&D Program of China under No.2023YFF0904800, the Beijing Nova Program under No.20230484406, the Innovation Research Group Project of NSFC (61921003), and the 111 Project (B18008).

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. 6

[3] Alexandre Ciancio, Eduardo AB da Silva, Amir Said, Ramin Samadani, Pere Obrador, et al. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on image processing*, 20(1):64–75, 2010. 2

[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. 2

[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. 1, 2

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[7] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 1

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5, 6, 7

[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. 1, 2

[11] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1138–1147, 2021. 1

[12] 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

[13] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 1

[14] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 1, 2

[15] Taehoon Kim, Gwangmo Song, Sihaeng Lee, Sangyun Kim, Yewon Seo, Soonyoung Lee, Seung Hwan Kim, Honglak Lee, and Kyunghoon Bae. L-verse: Bidirectional generation between image and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16526–16536, 2022. 1

[16] 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. 2, 3, 6

[17] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1

[18] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006–011006, 2010. 2

[19] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 3

[20] 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, 3, 5, 6

[21] 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. 5, 6

[22] 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. 3

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[24] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3

[25] 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. 2

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[27] 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 Workshops*, 2024. 2, 5

[28] Kede Ma, Xuelin Liu, Yuming Fang, and Eero P Simoncelli. Blind image quality assessment by learning from multiple annotators. In *2019 IEEE international conference on image processing (ICIP)*, pages 2344–2348. IEEE, 2019. 1, 2

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

[30] A Deep Bilinear Convolutional Neural Network. Blind image quality assessment using a deep bilinear convolutional neural network. 1, 2, 6

[31] 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

[32] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1505–1514, 2019. 1

[33] 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, 3, 6

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 5

[35] 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

[36] 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

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[39] 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. 2

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 6

[41] 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. 1, 2, 6

[42] Wei Sun, Huiyu Duan, Xiongkuo Min, Li Chen, and Guangtao Zhai. Blind quality assessment for in-the-wild images via hierarchical feature fusion strategy. In *2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 01–06. IEEE, 2022. 1, 2, 5, 6

[43] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. Frank: a ranking method with fidelity loss. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 383–390, 2007. 4

[44] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 2, 3, 6

[45] 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. 2, 5, 6

[46] Tao Wang, Wei Sun, Xiongkuo Min, Wei Lu, Zicheng Zhang, and Guangtao Zhai. A multi-dimensional aesthetic quality assessment model for mobile game images. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2021. 1, 2, 5, 6

[47] Zhihua Wang and Kede Ma. Active fine-tuning from gmad examples improves blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4577–4590, 2021.

[48] Zhihua Wang, Haotao Wang, Tianlong Chen, Zhangyang Wang, and Kede Ma. Troubleshooting blind image quality models in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16256–16265, 2021. 1, 2

[49] 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. 2, 3, 6

[50] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 3

[51] 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. 2, 3, 6

[52] 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

[53] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3585, 2020. 1, 2

[54] 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

[55] 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

[56] 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

[57] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 1, 2

[58] 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. 1, 5

[59] 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. 1, 2

[60] 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, 3

[61] 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. 2, 5, 6, 7

[62] 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. 3

[63] 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

[64] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3