Signal, Image and Video Processing (2025) 19:1254
https://doi.org/10.1007/s11760-025-04830-0

ORIGINAL PAPER

Multimodal Quality Assessment of AI-Generated Images using BERT-CLIP Feature Fusion

P Shabari Nath[1] · Rajlaxmi Chouhan[1]

Received: 10 June 2025 / Revised: 11 September 2025 / Accepted: 22 September 2025 / Published online: 3 October 2025

Abstract
AI-generated images continue to pose challenges such as misalignment between the text and generated output image and insufficient naturalness in terms of visual quality. These challenges necessitate the need for image quality assessment of the generated image to quantify the alignment of the images with the input text prompts. In this context, the BERT (Bidirectional Encoder Representation from Transformers) model outperforms OpenAI's GPT on the GLUE (General Language Understanding Evaluation) task. Its expanded tokenization length of upto 512 also helps circumvent CLIP's limitation of reduced text prompt. Therefore, in the present work, we propose a fusion model that applies BERT as a text encoder combined with CLIP as an image encoder. A bidirectional prompt learning approach through BERT is employed to extract the text features of the prompt used for the generation of the images. Further, using cross attention feature fusion, the proposed method obtains better SRCC and PLCC correlation metric results when compared with state-of-the-art methods on both PKUI2IQA and AGIQA-3K. Results of the ablation study and comparative characterization with other quality assessment metrics for AI-generated images demonstrate a noteworthy performance of the proposed method.

Keywords No-reference Image Quality Assessment · Vision-Language Models · AI-generated Images · Deep learning

1 Introduction

Artificial Intelligence (AI)–generated images have witnessed an exponential growth in the past few years in which the users can generate images using the text prompts. However, unlike real-world photographs, AI-generated images (AGIs) often exhibit quality inconsistencies due to limitations in model design and a lack of diverse training data Zhang et al. [27]. Generating photo-realistic scenes requires exposure to rich visual detail and variation, which many models lack. As a result, synthesizing natural images from text remains a complex task. This often forces users to manually evaluate and choose the best output. To streamline this process, there is a growing need for objective image quality assessment to automatically determine the visual quality and suitability of AGIs.

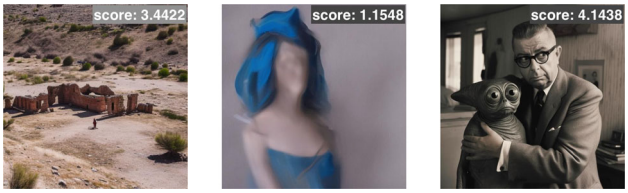
✉ P Shabari Nath
nath.1@iitj.ac.in

Rajlaxmi Chouhan
rajlaxmichouhan@iitj.ac.in

[1] Department of Electrical Engineering, Indian Institute of Technology Jodhpur, Jheepasani, Rajasthan, India

Image quality assessment (IQA) methods are commonly divided into Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR), depending on the availability of a ground truth image. In practice, especially with AGIs, a reference image is not available since the images are synthesized from text. Consequently, NR-IQA (or blind IQA) methods are particularly relevant for assessing the quality of AI-generated content.

Traditional no-reference image quality assessment (NR-IQA) methods typically extract handcrafted features such as visual neuron responses Chang et al. [2], natural scene statistics (NSS), mean-subtracted contrast-normalized (MSCN) coefficients Mittal et al. [13, 14]; Zhang et al. [28] . Later, the field shifted towards data-driven models like convolutional neural networks (CNNs) He et al. [7]; Kang et al. [9], transformers You and Korhonen [24], and other deep learning-based NR-IQA frameworks Zhang et al. [29]; Su et al. [19]; Ke et al. [10]; Golestaneh et al. [6]. These models are mainly designed for natural scene images (NSIs) affected by common distortions such as noise, blur, or compression. However, AI-generated images (AGIs) present unique challenges, requiring evaluation not only of perceptual quality but also of authenticity (realism) and consistency (seman-



(a) small group of nomads staying at ancient ruins, cold color, elevation view

(b) a portrait of girl alchemist in blue dress, blurred detail, sci-fi style

(c) sartre hugging a sad alien

Fig. 1 Example of AI-generated images with different AI models with their prompts and quality scores (human perception scores).

tic alignment with the input prompt). These attributes are independent and must be assessed separately, calling for AGI-specific IQA frameworks. Although AGI quality assessment is an emerging field, existing efforts mostly adapt NSI-oriented methods, which are limited in handling AGI-specific characteristics. Some approaches compare AGIs to NSIs at the distribution level, but this fails to capture individual image quality. Recently, new datasets incorporating prompts and multiple quality labels have enabled fine-tuning of existing models. Yet, these models often overlook text-image correspondence and rely on hand-engineered solutions, underscoring the need for dedicated NR-IQA techniques for AGIs.

For instance, Figure 1 shows AIGIs generated from models such as GAN, auto-regressive model, and diffusion-based approaches from the AGIQA-3k Li et al. [11]. The text prompt for image generation and the corresponding the human perception Mean Opinion Scores (MOS) on perception and T2I alignment are also included in the database. The subjects are asked to score the image in terms of visual quality, alignment score, and consistency. The scores shown on the image are those of visual quality. As visible, the images are substantially different in terms of their style and appearance, yet the image (b) receives the lowest MOS score although the text prompt includes the mention of blurriness.

Recently, Contrastive Language Image Pre-training (CLIP) model has shown better performance through textual prompt tuning Wang et al. [22]. However, there are a few challenges in implementing the CLIP model for quality assessment of AIGI. First, IQA methods must bridge the quality gap between AIGIs and natural scene images (NSIs), as current generative models still struggle to fully replicate NSIs. Second, it is essential to consider the text prompts used during image generation to evaluate how well AGIs align with their corresponding prompts. Lastly, CLIP's transformer-based text encoder, which operates left-to-right, may limit comprehension of prompts, prompting the use of antonym-based prompt pairs. Based on this, we propose that integrating vision-language learning could enhance AIGI quality prediction. Recent advances in AI-based deepfake detection using deep learning techniques like DenseNet and Vision Transformers, providing robust evaluation protocols Siddiqui et al. [17, 18]. Though focused on authenticity, their rigorous evaluation methods offer valuable benchmarks for building reliable AI-generated image quality assessment models.

To handle the above challenges, we propose a multimodal bidirectional prompt learning approach. We use user input text prompts without truncation to guide the optimization of the vision language learning. The main contribution of this paper is as follows:

- Use of the conventional BERT (bidirectional encoder representations from transformers) model as a **text encoder** in the CLIP model
- Three-pronged fusion to combine text and image features

The subsequent sections of this paper are organized as follows. Section 3 presents the proposed approach for the no-reference IQA of AI-generated images. The results, comparative characterization, and the ablation study are discussed in Section 4. Section 5 summarizes important findings from the paper.

2 Related Work

Research on evaluating the quality of AI-generated images (AGIs) is still in its early stages, with limited studies conducted compared to natural scene images (NSIs). Initially, the Inception Score, introduced by Salimans et al. [16], was widely used for assessing AGI quality. However, due to fundamental Fig 2 differences between NSIs and AGIs, more sophisticated metrics such as the Frechet Inception Distance Heusel et al. [8] and Kernel Inception Distance Bińkowski et al. [1] were later developed. These metrics assess AGI quality by measuring distributional differences between AGIs and NSIs. Despite their usefulness, they evaluate AGI quality from a single perspective and are inadequate for assessing individual AGIs, highlighting the need for more advanced evaluation approaches.

Zhang et al. (2023b) were early contributors to the study of multi-dimensional AGI quality assessment, proposing the importance of evaluating factors such as technical intelligence, AI-specific attributes, visual artifacts, deviation from real images, and aesthetic appeal. However, their study lacked a defined evaluation method or a structured way to integrate AGI-specific distortions into the assessment framework.

Recent works have proposed specialized IQA techniques for AGIs, often adapting methods designed for NSIs. Yuan

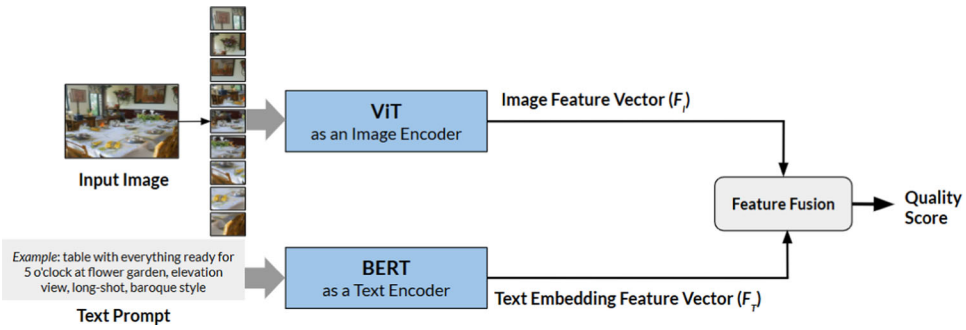


Fig. 2 Framework of proposed model

et al. [25] introduced a contrastive regression network that enhances feature representation across multiple AGIs. Later, Yuan et al. [26] presented a regression framework using separate text and image encoders, merging their features for quality prediction. Other studies leverage LLMs to assess text–image consistency Lu et al. [12], such as LLMScore Lu et al. [12], which generates object- and image-level descriptions and uses LLMs for evaluation. However, these methods demand large labeled datasets and heavy parameters. To improve efficiency, Li et al. [11] proposed StairReward, which segments text prompts into morphemes and images into hierarchical sections, aligning them one-to-one to produce a quality score.

Despite the improvements these AGI quality assessment methods offer over traditional NSI-oriented IQA approaches, they still face challenges that limit their practical applications. Most approaches focus on visual quality alone, overlooking how well the generated content aligns with user intent, where content relevance and consistency are equally important. Additionally, many methods evaluate the AGI in isolation, lacking robust mechanisms for assessing text-image alignment. Some strategies use prompt segmentation and image partitioning to gauge consistency, but these depend on handcrafted rules, limiting scalability. Furthermore, the interplay between multimodal features, especially between text and images, is often underexplored. These issues highlight the need for more versatile and holistic AGI evaluation models.

Due to the limitations of the above methods, we hypothesize that a fusion-based approach has the potential to mitigate the limitation of individual models. We therefore propose a CLIP-BERT-AGIQA framework in this paper (as described in Section 3.)

3 Proposed Method

In this paper, we propose a new approach towards no-reference quality assessment of AI-generated images by integrating the CLIP-based embedding technique with the BERT Devlin et al. [3] text encoder. The recent CLIP models use antonym prompt pairing strategy to estimate the perceptual quality Wang et al. [22]; Fu et al. [5] of an AIGI. However, in the proposed model we utilize the text prompt that was used to generate the image for estimating its perceptual quality. Based on this, we verify the similarity between the AIGI and the user text prompt. In this paper, we leverage multimodal prompt and vision learning by fine-tuning the BERT-based text encoder and CLIP's ViT-based image encoder. The proposed CLIP-BERT based framework, as shown in Fig 2, mainly consists of three blocks: CLIP (image encoder), BERT (text encoder), and quality prediction.

3.1 CLIP Model

Recent CLIP models are mainly focused on the adjective-based prompts for the quality assessment of the images or videos Wang et al. [22]. The CLIP model comprises two encoders: a transformer for text processing and a vision transformer (ViT) or ResNet for image representation. The model uses a dual-tower design in which the text and image encoders perform independently but are aligned in a common embedding space. It connects the gap between language and vision by comprehending visual characteristics and textual semantics, which makes it appropriate for a variety of applications such as assisting visually challenged users, evaluating AI-generated art, and automating content moderation. CLIP is more flexible than standard supervised models since it learns from constraint-free text-image pairings rather than fixed labels. Initially, the vision encoder of the CLIP models was based on ResNet50 pre-trained model He et al. [7], which was later replaced by the transformer-based models Dosovitskiy et al. [4]. Later, a antonym prompt–based pairing strategy of the CLIP model was used for quality assessment, where cosine similarity between the text features (e.g. antonym prompts: "Good Photo", "Bad Photo") and the image features is computed. The prediction score of CLIP is calculated based on the similarity score, s, between the image and text features (for all text prompts) as per equation 1. Then, the final quality prediction score $\bar{s}$ is computed using the soft-

max information as per equation 2 defined as:

$$s_i = \frac{x \odot t_i}{\|x\| \cdot \|t_i\|}, \qquad i \in \{1, 2\} \tag{1}$$

.

Here, x represents the embedded image features, t_i are the prompt-embedded features (for the antonym pair such that $i = 1, 2$), $\|.\|$ denotes l_2 norm, and $\odot$ represents vector dot product.

$$\bar{s} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}} \tag{2}$$

The next stage involves text encoding.

3.2 BERT Model

The model architecture of the BERT is based on multi-layer bidirectional transformer encoder which is inspired by "the annotated transformer" proposed by Vaswani et al. [21]. We utilize the general language understanding evaluation (GLUE) task for fine tuning the BERT model with the quality score annotated text prompts, which is one of the 11 NLP tasks reported by Devlin et al. for the analysis of the BERT model Devlin et al. [3]. Further, the BERT model is fine-tuned using the quality scores of the AIGIs.

In this work, we utilize BERT-based architecture, which has 12 layers of transformer blocks, a hidden size of 768, and 12 self-attention heads. While the BERT transformer is based completely on the bidirectional self-attention method, the GPT uses a constrained self attention–based transformer. The conditional language models can train the text either from left-to-right or right-to-left. Unlike conditional language models, BERT is a bidirectional model that allows it to analyze the context surrounding a target word and predict it within a multi-layered framework. However, for training the deep bidirectional representation in BERT, we predict the masked tokens after randomly masking a certain percentage of the input tokens. This procedure is referred to as 'Masked Language Model (MLM)', and it is often referred to in the literature as the Cloze task.

Based on the findings of Devlin et al. [3], BERT-Base and BERT-Large consistently outperform OpenAI GPT across all GLUE benchmarks, achieving average gains of 4.5% and 7.0% respectively. Owing to its superior ability to generate rich textual embeddings, BERT is selected as the text encoder to extract the feature vector F_T from the input prompt. Unlike CLIP's text encoder, which limits input to 77 tokens, BERT supports sequences up to 512 tokens, allowing for more flexible and detailed prompt encoding. BERT, RoBERTa, and DistilBERT are transformer-based models tailored for NLP, each optimized differently. BERT, by Google, uses masked language modeling and next-sentence prediction to capture deep context in text. RoBERTa, from Facebook AI, refines BERT by omitting the NSP task, leveraging larger datasets, longer training, and dynamic masking for better performance. In contrast, DistilBERT, from Hugging Face, streamlines BERT via knowledge distillation, cutting model size by 40% while preserving most of its accuracy, making it suitable for low-resource or real-time tasks. While BERT sets a strong foundation, RoBERTa typically offers higher accuracy, and DistilBERT prioritizes speed and efficiency, highlighting the trade-offs in choosing a model based on task requirements.

3.3 Quality Prediction

As shown in Fig. 2, the features F_I and F_T are extracted by feeding the AIGIs to the vision and text encoders. For fusing these features, we select three methods which includes: Concatenation, Cosine Similarity (CS), and Cross Attention (CA). For visual quality prediction through CS, we utilize BERT to encode the text prompt (from BERT). We then compute the CS between the image feature F_I (from CLIP) and textual feature $F_T \in \mathbb{R}^{N \times 768}$ (where N is maximum length of the text prompt tokens) as shown below:

$$S_Q = \frac{F_I \odot (F_T)^T}{\|F_I\|_2 \cdot \|F_T\|_2} \tag{3}$$

For the CA mechanism, the image feature vector, F_I, is used as a query, and the prompt feature vector, F_T, is used as both key and value. Later, we predict the score by passing the output feature vector of the CA through a regression network of two Fully Connected (FC) layers.

$$F_v = [F_I, F_T] \tag{4}$$

Moreover, for fusion of image and text features using concatenation, the concatenated vector F_v as represented in Eq. 4, is connected to the FC layer characterized by bias b_1 and weight W_1 (Eq. 5). The final score, Q_s, is represented as the linear transformation of FC1 vector (i.e. the output of first FC layer) with weights W_2 and bias b_2 (Eq. 6).

$$FC1 = ReLU(W_1 F_v + b1) \tag{5}$$

$$Q_s(I, T) = W_2 FC1 + b2 \tag{6}$$

4 Results and Discussions

4.1 Database and Evaluation Criteria

In the present work, we conduct experiments on three publicly available AI-generated image datasets, i.e., AGIQA-3K

Li et al. [11], AIGCIQA2023 Wang et al. [23], and PKU-I2IQA Yuan et al. [26], for evaluating the quality scores of AI-generated images and comparative characterization with state-of-the-art methods. The AGIQA-3K, AIGCIQA2023, and PKU-I2IQA datasets are benchmark resources for assessing AI-generated image quality. AGIQA-3K includes 2,982 images from six text-to-image (T2I) models and is labeled with visual quality, authenticity, and consistency scores based on 300 diverse prompts. AIGCIQA2023 features 2,400 images generated by six T2I models using 100 prompts, with similar quality-related labels. PKU-I2IQA offers 1,600 images produced by two image-to-image models, paired with 200 prompts and annotated for authenticity, consistency, and visual quality.

To quantify the reliability of this quality metric, a correlation must be drawn between the objectively computed metric (from the model) and the mean opinion scores representing human perception (available with the databases) as shown in Fig 3. The correlation metrics used in this study include the Spearman Rank-order Correlation Coefficient (SRCC), and the Pearson Linear Correlation Coefficient (PLCC). These metrics are commonly used in IQA, and have value between 0 and 1, such that larger value denotes greater prediction performance.

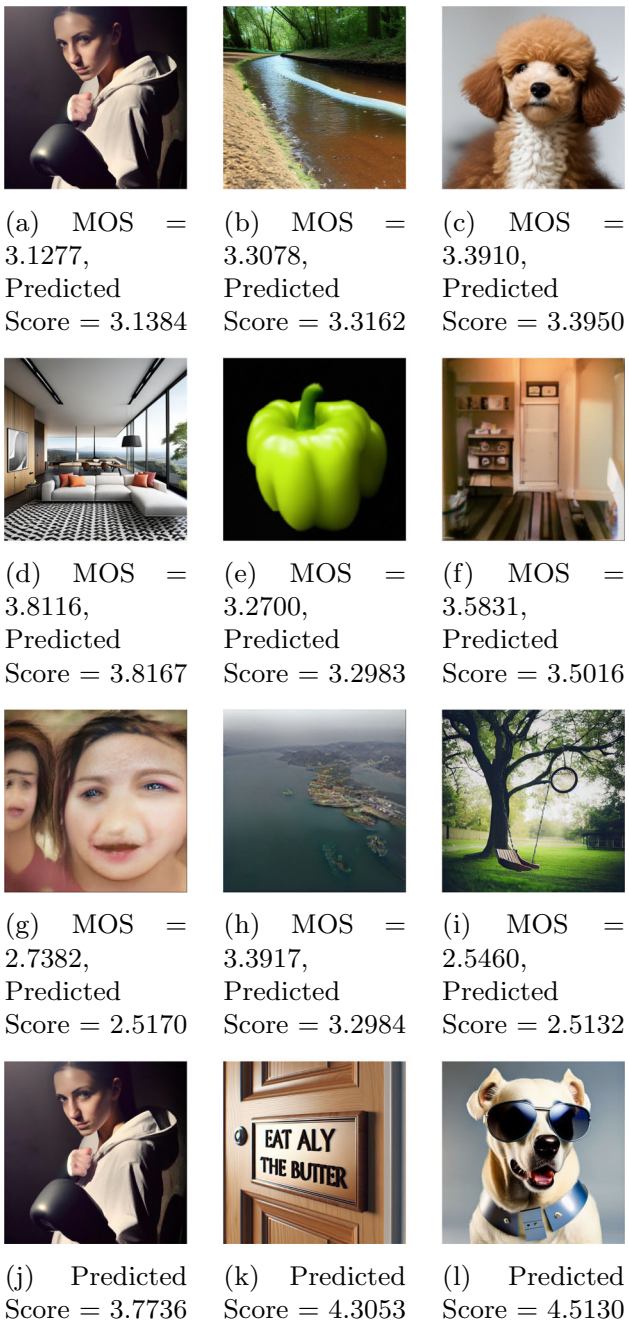
4.2 Implementation Details

Our proposed approach uses the ViT-B/32 as the vision encoder and BERT as the text encoder. As mentioned in Section 3.2, the limit of prompt length for BERT is more than CLIP's text encoder.

Each dataset is partitioned differently: While the AGIQA-3K dataset uses 80:20 ratio between training and testing sets, the AIGCIQA2023 and PKU-I2IQA are partitioned in 75:25 ratio for training and testing. During training, the text prompt is given to the BERT tokenizer for creating the attention mask of the entire text prompt. As the text prompts are of different sizes, the mask provides a padding to the text prompt as per the requirements of the BERT encoder. The Adam optimization technique is utilized to optimize the parameters of the model. We choose $1e^{-4}$ and 100 for learning rate and training epochs, respectively. The proposed method was implemented on Pytorch-based code, and experiment was performed on an Nvidia DGX server available at IIT Jodhpur.

4.3 Performance and Comparative Characterization

The performance of the proposed approach is validated using SRCC and PLCC. The average SRCC values for AGIQA-3K, AIGCIQA2023, and PKU-I2IQA were 0.8813, 0.8397, and 0.7149, respectively. Similarly, the respective average PLCC for for 0.9153, 0.8641, and 0.7268 databases was found to be AGIQA-3k, AIGCIQA2023, and PKUI2IQA. This represents that the proposed metric accurately represents the quality of AIGIs. To evaluate the generalizability of the proposed model, we conducted additional testing on a set of 200 AGI's sourced from unseen T2I models, namely DALL·E 2 and Dreame. The model achieved better performance, with a SROCC of **0.8601** and PLCC of **0.8893**. This result highlights the robustness and potential of the proposed approach for quality assessment in evolving real-world T2I systems. To assess the statistical significance of performance gains



(a) MOS = 3.1277, Predicted Score = 3.1384
(b) MOS = 3.3078, Predicted Score = 3.3162
(c) MOS = 3.3910, Predicted Score = 3.3950
(d) MOS = 3.8116, Predicted Score = 3.8167
(e) MOS = 3.2700, Predicted Score = 3.2983
(f) MOS = 3.5831, Predicted Score = 3.5016
(g) MOS = 2.7382, Predicted Score = 2.5170
(h) MOS = 3.3917, Predicted Score = 3.2984
(i) MOS = 2.5460, Predicted Score = 2.5132
(j) Predicted Score = 3.7736
(k) Predicted Score = 4.3053
(l) Predicted Score = 4.5130

Fig. 3 Example of AI-generated images with different AI models with their quality scores (human perception scores)

Table 1 Ablation study of the Score Fusion on AGIQA-3K database

BERT-version	Concatenation		Cosine Similarity		Cross Attention	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
distillBERT	0.8429	0.8690	0.8318	0.8461	0.8649	0.8906
RoBERT	0.8521	0.8796	0.8591	0.8640	0.8783	0.9072
BERT-Base (proposed)	**0.8579**	**0.8893**	**0.8648**	**0.8755**	**0.8823**	**0.9173**

Table 2 Performance comparison of the proposed method with state-of-the-art methods. Here boldfaced font represents the best performance for each metric. The methods are grouped as (i) traditional (ii) DL-based (iii) CLIP-based in this table

Method	AGIQA-3K		AIGCIQA2023		PKUI2IQA	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE [13]	0.4984	0.5496	0.6289	0.5839	0.4638	0.5281
NIQE Mittal et al. [14]	0.5181	0.5308	0.4971	0.4516	0.4189	0.4319
ILNIQE Zhang et al. [28]	0.6018	0.6194	0.5613	0.5078	0.4585	0.4869
DB-CNN Zhang et al. [29]	0.8391	0.8814	0.8296	0.8520	0.6027	0.6139
HyperIQA Su et al. [19]	0.8427	0.8993	**0.8491**	**0.8736**	0.6813	0.6983
CNNIQA Kang et al. [9]	0.7578	0.8149	0.7618	0.8065	0.5896	0.6037
Re-IQA Saha et al. [15]	0.8207	0.8810	0.8166	0.8347	0.6094	0.6209
AMFF-NET Zhou et al. [31]	0.8537	0.9070	0.8419	0.8537	0.7093	0.7718
StairIQA Sun et al. [20]	0.8215	0.8894	0.8136	0.8501	0.5955	0.6071
LIQE Zhang et al. [30]	0.8671	0.8975	0.8481	0.8530	0.6983	0.7385
ResNet50 He et al. [7]	0.8361	0.8794	0.8273	0.8495	0.6219	0.6382
MUSIQ Ke et al. [10]	0.8356	0.8829	0.8423	0.8602	0.6408	0.6410
TReS Golestaneh et al. [6]	0.8367	0.8973	0.8436	0.8637	0.6374	0.6427
CLIPIQA Wang et al. [22]	0.6846	0.6987	0.4171	0.3970	0.6581	0.6562
CLIP-AGIQA Fu et al. [5]	0.8747	**0.9190**	0.8324	0.8604	0.6515	0.6871
CLIP-BERT-AGIQA (Proposed)	**0.8813**	0.9153	0.8397	0.8641	**0.7349**	**0.7868**

of proposed method, we performed statistical significance testing using Fisher's z-transformation. The SRCC improvement from 0.8747 to 0.8813 on AGIQA-3K corresponds to $z = 1.83 (p \approx 0.06)$, with a 95% confidence interval of [0.874, 0.890]. This indicates that the observed improvement is marginally significant.

We also compare the performance of the proposed method with three traditional feature-based methods, five deep learning–based metrics, and two CLIP-based approaches. Among the methods, the traditional feature–based Mittal et al. [13, 14]; Zhang et al. [28] and deep learning–based Zhang et al. [29]; Su et al. [19]; He et al. [7]; Ke et al. [10]; Golestaneh et al. [6] and CLIP-IQA Wang et al. [22] methods are designed and trained for NSIs. On the other hand, the CLIP-AGIQA method Fu et al. [5] is primarily trained for the AIGI dataset. The obtained quantitative results of the proposed method (CLIP-BERT-AGIQA) and other methods are reported in the Table 2.

The following observations can be derived from the Table 2. The proposed approach delivers superior visual quality predictions on both AGIQA-3K and PKU-I2IQA datasets compared to baseline methods. Traditional feature-based techniques consistently underperform across datasets, likely due to their reliance on natural scene statistics, which do not align well with the distribution of AI-generated images (AGIs) created by complex generative models that lack full photo-realism. On AIGCIQA-2023, the proposed CLIP-BERT model ranks second in PLCC performance, slightly behind the best. This may be due to the visual similarities among AGIs generated by Lafite and ControlNet, limiting the vision encoder's sensitivity to fine-grained variations. Lastly, Hyper-IQA Su et al. [19] surpasses other NSI-based methods, likely due to its content-aware hypernetwork, which effectively captures semantic distortions–an important factor in subjective assessment of AGIs.

Table 3 compares different IQA methods on the AGIQA-3k dataset. With the highest SRCC score of 0.8813, the proposed CLIP-BERT-AGIQA model demonstrates a strong correlation with human visual perception. On comparison with CLIP-AGIQA Fu et al. [5], the proposed approach comes in close second with a PLCC of 0.9153. In terms of both SRCC and PLCC, PSCR Yuan et al. [25] has the lowest performance among the three. Overall, the proposed model is found to have a balanced and effective performance in comparison with other metrics of IQA for AGIs. In our study, we observed some imperfect and noisy text prompts (for example: hr giger in the style of hr giger). Due to the BERT's bidirectional context modeling, and robustness to

Table 3 Comparison of the proposed method with other IQA methods for AGIQA-3k

Method	SRCC	PLCC
PSCR Yuan et al. [25]	0.8394	0.8859
CLIP-AGIQA Fu et al. [5]	0.8747	**0.9190**
CLIP-BERT-AGIQA (Proposed)	**0.8813**	0.9153

incomplete or partially noisy prompts, model produces meaningful embeddings even when portions of the text are missing or imperfect.

4.4 Ablation Study

This section presents an ablation study evaluating the effectiveness of the proposed three-pronged feature fusion strategy. The proposed model's performance is assessed using three different fusion mechanisms: (i) Concatenation, (ii) Cosine Similarity, and (iii) Cross Attention, as detailed in Section 3.3. The comparative results of these fusion techniques are summarized in Table 1.

Evaluation of Fusion Mechanisms:

Concatenation: This approach fuses image and text features by direct concatenation, followed by two fully connected layers for quality prediction. While simple and efficient, it often underperforms due to poor modeling of interactions between modalities, likely caused by misalignment during training.

Cosine Similarity (CS): CS, used in models like CLIP for IQA, measures alignment between image and text embeddings. In our study, it slightly outperforms concatenation but still lags behind cross-attention, as shown in Table 1.

Cross Attention (CA): CA uses image features as queries and text features as keys and values, enabling the model to better capture cross-modal relationships. This results in refined feature representations and more accurate quality predictions. As shown in Table 1, CA consistently achieves the highest SRCC and PLCC scores across all BERT variants.

Performance Comparison Across BERT Variants

The study also analyzes the trade-offs between model accuracy, inference time, and computational cost for different BERT-based text encoders–**BERT-Base**, **DistilBERT**, and **RoBERTa**–in the context of feature fusion performance.

The proposed **BERT-Base** model delivers the highest performance across all fusion techniques, with Cross Attention achieving the best results (SRCC: 0.8823, PLCC: 0.9173). This highlights its ability to capture rich semantic relationships between image and text features. However, this performance comes with increased computational cost with inference time of approximately 217 ms for a single input image; as BERT-Base has a significantly larger number of parameters, it leads to a slower inference time compared to lighter models.

DistilBERT, designed for efficiency, offers faster inference (approximately 70% of model with BERT-Base i.e., 152 ms) and a smaller memory footprint while maintaining reasonably competitive performance. Although it shows slightly lower accuracy than BERT-Base and RoBERTa, it is a practical choice for resource-constrained environments, particularly where real-time processing is required.

RoBERTa achieves performance close to BERT-Base, particularly in the Cross Attention setting (SRCC: 0.8783, PLCC: 0.9072), benefiting from dynamic masking and training on a larger corpus. It provides a balance between performance and computational load (approximately 196 ms), outperforming DistilBERT while being marginally more efficient than BERT-Base.

These findings emphasize the importance of selecting text encoders based on the target application's performance and efficiency requirements.

5 Conclusion

In this paper, we proposed a multimodal bidirectional prompt learning framework based on the BERT model for quality assessment of AI-generated images (AGIs). By replacing CLIP's text encoder with BERT, the proposed approach leverages richer bidirectional context and extended prompt lengths, resulting in more accurate predictions of visual quality. Our results demonstrate that this framework consistently outperforms several existing IQA methods across multiple benchmark datasets. An extended ablation study further shows that cross-attention fusion combined with BERT-based encoders provides the strongest performance. While BERT-Base achieves the highest accuracy, our computational analysis highlights that lighter alternatives such as DistilBERT and RoBERTa deliver competitive results with substantially reduced overhead. This trade-off makes the proposed framework adaptable for both high-accuracy research settings and efficiency-critical real-world applications. Looking ahead, we plan to investigate dynamic masking strategies, alternative lightweight encoders (e.g., T5, LLaMA, Mistral), and scaling experiments on larger and more diverse AGI datasets to further enhance both robustness and deployment feasibility.In our study,all prompts fit within BERT's 512-token limit, so truncation was unnecessary. For longer prompts, methods like hierarchical encoding, segment-wise processing, or larger-context models could be used, which we consider this a promising direction for future work.

Author Contributions Author 1: methodology, manuscript, experimentation, and analysis Author 2: manuscript, analysis, and supervision

Funding The authors declare that this work does not have any funding.

Data Availability There will be no data and code availability until the manuscript is published.
No datasets were generated or analysed during the current study.

Declarations

Conflicts of Interest They have no conflict of interest.

Publication Statement This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.

Original Work All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.

Competing interests The authors declare no competing interests.

References

1. Bińkowski, M., Sutherland, D.J., Arbel, M., et al.: Demystifying mmd gans (2021). arXiv:1801.01401
2. Chang, H.W., Bi, X.D., Kai, C.: Blind image quality assessment by visual neuron matrix. IEEE Signal Process. Lett. **28**, 1803–1807 (2021). https://doi.org/10.1109/LSP.2021.3106579
3. Devlin, J., Chang, M.W., Lee, K., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, Minneapolis, Minnesota, p 2 (2019)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale (2020) . arXiv preprint arXiv:2010.11929
5. Fu, J., Zhou, W., Jiang, Q., et al.: Vision-language consistency guided multi-modal prompt learning for blind ai generated image quality assessment. IEEE Signal Process. Lett. **31**, 1820–1824 (2024). https://doi.org/10.1109/LSP.2024.3420083
6. Golestaneh, S.A., Dadsetan, S., Kitani, K.M.: No-reference image quality assessment via transformers, relative ranking, and self-consistency. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 3989–3999 (2022). https://doi.org/10.1109/WACV51458.2022.00404
7. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
8. Heusel, M., Ramsauer, H., Unterthiner, T., et al.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'17, p 6629–6640 (2017)
9. Kang, L., Ye, P., Li, Y., et al.: Convolutional neural networks for no-reference image quality assessment. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp 1733–1740 (2014). https://doi.org/10.1109/CVPR.2014.224
10. Ke, J., Wang, Q., Wang, Y., et al.: Musiq: Multi-scale image quality transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp 5128–5137 (2021). https://doi.org/10.1109/ICCV48922.2021.00510
11. Li, C., Zhang, Z., Wu, H., et al.: Agiqa-3k: an open database for ai-generated image quality assessment. IEEE Trans. Circuits Syst. Video Technol. **34**(8), 6833–6846 (2024). https://doi.org/10.1109/TCSVT.2023.3319020
12. Lu, Y., Yang, X., Li, X., et al.: Llmscore: unveiling the power of large language models in text-to-image synthesis evaluation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS '23 (2023)
13. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. **21**(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050
14. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Process. Lett. **20**(3), 209–212 (2013). https://doi.org/10.1109/LSP.2012.2227726
15. Saha, A., Mishra, S., Bovik, A.C.: Re-iqa: Unsupervised learning for image quality assessment in the wild. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 5846–5855 (2023). https://doi.org/10.1109/CVPR52729.2023.00566
16. Salimans, T., Goodfellow, I., Zaremba, W., et al.: Improved techniques for training gans. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'16, p 2234–2242 (2016)
17. Siddiqui, F., Yang, J., Xiao, S., et al.: Diffusion model in modern detection: advancing deepfake techniques. Knowl.-Based Syst. **325**, 113922 (2025). https://doi.org/10.1016/j.knosys.2025.113922
18. Siddiqui, F., Yang, J., Xiao, S., et al.: Enhanced deepfake detection with densenet and cross-vit. Expert Syst Appl 267(C) (2025b) . https://doi.org/10.1016/j.eswa.2024.126150
19. Su, S., Yan, Q., Zhu, Y., et al.: 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), pp 3664–3673 (2020). https://doi.org/10.1109/CVPR42600.2020.00372
20. Sun, W., Min, X., Tu, D., et al.: Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. IEEE Journal of Selected Topics in Signal Processing **17**(6), 1178–1192 (2023). https://doi.org/10.1109/JSTSP.2023.3270621
21. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, et al.: (eds) Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc., (2017).https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
22. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. Proceedings of the AAAI Conference on Artificial Intelligence **37**(2), 2555–2563 (2023). https://doi.org/10.1609/aaai.v37i2.25353
23. Wang, J., Duan, H., Liu, J., et al.: Aigciqa 2023: a large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In: Fang, L., Pei, J., Zhai, G., et al. (eds.) Artificial Intelligence, pp. 46–57. Springer Nature Singapore, Singapore (2024)
24. You, J., Korhonen, J.: Transformer for image quality assessment. In: 2021 IEEE International Conference on Image Processing (ICIP), pp 1389–1393 (2021). https://doi.org/10.1109/ICIP42928.2021.9506075
25. Yuan, J., Cao, X., Cao, L., et al.: Pscr: Patches sampling-based contrastive regression for aigc image quality assessment (2023). https://arxiv.org/abs/2312.05897, arXiv:2312.05897
26. Yuan, J., Yang, F., Li, J., et al.: Pku-aigiqa-4k: A perceptual quality assessment database for both text-to-image and image-to-image ai-generated images. arXiv:2404.18409 (2024). https://api.semanticscholar.org/CorpusID:269449873
27. Zhang, C., Zhang, C., Zhang, M., et al.: Text-to-image diffusion models in generative ai: A survey (2024). arXiv preprint arXiv:2303.07909

28. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. IEEE Trans. Image Process. **24**(8), 2579–2591 (2015). https://doi.org/10.1109/TIP.2015.2426416
29. Zhang, W., Ma, K., Yan, J., et al.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans. Circuits Syst. Video Technol. **30**(1), 36–47 (2018)
30. Zhang, W., Zhai, G., Wei, Y., et al.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 14071–14081 (2023). https://doi.org/10.1109/CVPR52729.2023.01352
31. Zhou, T., Tan, S., Zhou, W., et al.: Adaptive mixed-scale feature fusion network for blind ai-generated image quality assessment. IEEE Trans. Broadcast. **70**(3), 833–843 (2024). https://doi.org/10.1109/TBC.2024.3391060