Enhancing prompt-image alignment evaluations via cyclic mutual information maximization

Xingran Liao[1], Duanyu Feng[2], Mingliang Zhou[3], Sam Kwong[4], and Weisi Lin[1⋆]

[1] Nanyang Technological University, 50 Nanyang Avenue, Singapore
[2] Sichuan University, 1st Ring Road South, Chengdu, China
[3] Chongqing University, Shapingba District, Chongqing, China
[4] Lingnan University, Castle Peak Road, Tuen Mun, Hong Kong

Abstract. Assessing AI-generated images remains a significant challenge because both visual quality and prompt-image alignment are vital. Existing metrics prioritize visual quality, but the exploration of semantic alignment with user prompts remains limited. To address this issue, we propose a novel cyclic mutual information maximization framework for AI-generated image quality assessment (CMIM-AIGIQA). Unlike existing methods, our framework focuses on promoting effective multimodal information fusion in the deep feature domain, which has cyclic mutual information maximization phases. In the forward phase, we maximize the mutual information between text and image features to generate refined, cross-aware representations. The refined features are then integrated through a fusion network. In the backward phase, the mutual information between the joint embedding and every single modality is maximized to ensure that the joint embedding retains critical semantic and visual cues. Experiments on three datasets demonstrate that our method effectively bridges the gap between visual quality and prompt-image alignment. The code is available at `https://github.com/Buka-Xing/CMIM-AIGIQA`.

Keywords: AIGC · Image quality assessment · Mutual Information Maximization

1 Introduction

The explosion of generative AI has changed the creation of digital content. By inputting simple prompts, the generative models can synthesize images with various contents [3, 15]. However, two urgent issues hinder the practical application of generative AI [10]: (1) the occurrence of unappealing artifacts, which severely affects the visual quality of AI-generated images (AIGI), and (2) the lack of controllability, where generation models often fail to reflect the semantic constraints specified in user prompts. Manually evaluating AIGIs in terms of both the visual quality and the prompt-image alignment dimension is time-consuming and

⋆ Corresponding author.

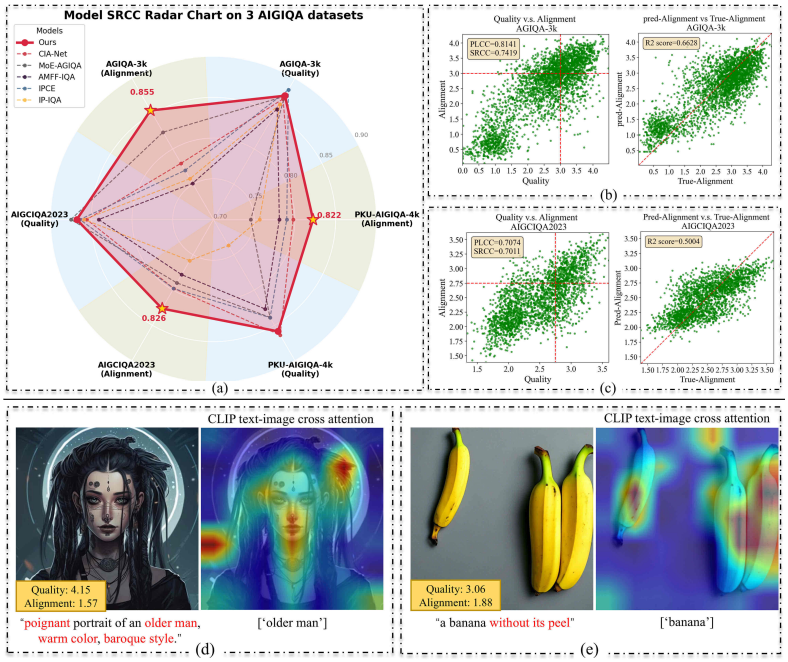

Fig. 1: (a) Spearman-rank order correlation coefficient (SRCC) radar chart on three AIGIQA datasets. (b) and (c), The quality and alignment scatter plots (left) and the fitted linear function prediction results (right). (d) and (e), Two image samples from the AGIQA-3k [13] and the AIGCIQA2023 [34]. The left panels display the AIGIs with their prompts and subjective scores. The unaligned prompts are highlighted in red. The right panels provide text-to-image cross-attention visualizations via Grad-CAM [26] on the pretrained CLIP [25].

labor-intensive. Therefore, developing objective AI-generated image quality assessment (AIGIQA) methods to evaluate both visual quality and prompt-image alignment is urgently needed [31,32].

Current AIGIQA research focuses mainly on evaluating visual quality, yet joint models of visual quality and prompt-image alignment remain limited. These quality-sensitive AIGIQA methods cannot directly adapt to alignment evaluations. We retrain prevailing AIGIQA methods on the alignment dimensions with their open-source settings in Fig. 1(a). Although these metrics demonstrate strong correlations with human-perceived visual quality on the AGIQA-3k [13], AIGCIQA2023 [34], and PKU-AIGIQA-4k datasets [39], their SRCC for alignment consistently falls below 0.80 across the three datasets. We identify two potential reasons for inferior alignment evaluations. First, the relationship between visual quality and prompt-image alignment is complex, where superior visual quality does not guarantee high prompt alignment, and vice versa. This is shown in scatter plots of AGIQA-3k [13] and AIGCIQA2023 [34] in Fig. 1(b) and

(c). Despite the positive SRCC between the quality and alignment dimensions (0.7419 for AGIQA-3k [13] and 0.7011 for AIGCIQA2023 [34]), we find significant local dispersion. As highlighted by the vertical dashed lines, images with high quality scores (e.g., 3.0 for AGIQA-3k [13] and 2.75 for AIGCIQA2023 [34]) exhibit disparate alignment levels, spanning nearly the entire y-axis. A similar issue also occurs along the horizontal dashed lines. We employ a linear regression analysis to model the relationship between the visual quality and the prompt-image alignment in Fig. 1(b) and (c), where the quality score is used to predict the alignment score as "$Alignment = a \times Quality + b$", and a and b are linear function parameters. Nevertheless, low coefficients of determination (R^2<0.7) are obtained across both AIGIQA datasets, which explains why direct retraining of existing quality-sensitive methods yields degraded performance in alignment evaluations. Second, the visual and textual features extracted from pretrained visual-language models (VLMs) often suffer from inherent "modality noise", hindering effective prompt-image alignment. We use Grad-CAM [26] on the pretrained CLIP [25] network to visualize the text-to-image cross-attention, and the results are depicted in Fig. 1(d) and (e). We find that the AIGI visual content typically contains dense details not specified in the sparse textual prompts. During the evaluation of prompt-image alignment, such extra visual information becomes the "modality noise": in Fig. 1(d), the "older man" prompts cause the visual model to focus on image contents that are apart from the human portrait; in Fig. 1(e), the model pays attention to visual contents that are not related to the "banana" prompt. Notably, this "attention drift" widely exists in the AGIQA-3k [13] and AIGCIQA2023 [34] datasets, especially in low-alignment samples, which constitute nearly 50% of the total images in the two datasets. Nevertheless, prevailing methods typically utilize pretrained CLIP [25] but confine their network design to peripheral stages, such as input-level preprocessing or output-level score pooling. The exploration of visual-textual feature refinement and fusion strategies remains underexplored. In conclusion, a more advanced design paradigm is indispensable for developing advanced AIGIQA methods that excel at both visual quality and prompt-image alignment evaluations.

To model complex relationships and promote effective multimodal feature fusion, we propose the cyclic mutual information maximization framework for AI-generated image quality assessment (CMIM-AIGIQA), the first paradigm to treat multimodal fusion in the AIGIQA task as an information-theoretic optimization process. Specifically, a forward-backward mutual information maximization (MIM) framework is developed. In the forward phase, we explicitly refine the visual and textual features by maximizing the MI between them. These refined features are then integrated through a multimodal fusion network to generate a joint embedding. Second, in the backward phase, we implicitly maximize the statistical dependency between the joint embedding and the two refined modality features, ensuring that the joint embedding preserves essential information from both visual and semantic cues. The forward phase aims to extract "cross-aware" information that enjoys the shared semantic core and filter out modality noise. The backward phase aims to impose a semantic preservation

constraint to prevent high-density visual features from overwhelming sparse textual features. By maximizing the dependency between the joint embedding and its constituent modalities, we ensure that the fused representation serves as a sufficient statistic for both modalities. Consequently, CMIM-AIGIQA achieves a balanced, robust evaluation of both perceptual quality and semantic alignment. Our key contributions are summarized as follows.

- We propose CMIM-AIGIQA, the first model that introduces the MIM framework into the AIGIQA task. Moving beyond input-level data preprocessing and output-level score pooling, our framework treats visual-textual feature fusion as an MIM process to balance evaluations of both visual quality and prompt-image alignment.
- We introduce a cyclic mutual information maximization framework. The forward phase processes refined, cross-aware features by filtering modality-specific noise, whereas the backward phase ensures that the joint multimodal embedding preserves the essential semantic and visual information of both inputs, effectively bridging the gap between visual quality and prompt alignment. Such a cyclic pathway, especially the backward constraints, is novel in the AIGIQA task.
- Extensive experiments on the AGIQA-3k [13], AIGCIQA2023 [34], and PKU-AIGIQA-4k [39] validate the superiority of our model. The results show that CMIM-AIGIQA achieves robust performance across both quality and alignment dimensions, demonstrating that the CMIM framework effectively captures the nonlinear relationships between modalities for comprehensive AIGIQA evaluations.

2 Related works

2.1 IQA: From Natural Images to AIGI

Objective IQA models are classified into full-reference, reduced-reference, and non-reference (NR-IQA) models based on reference availability. NR-IQA is most related to the AIGIQA task, as text-to-image generation typically lacks a ground-truth reference. Early NR-IQA methods, such as NIQE [19] and BRISQUE [18], utilized natural scene statistics (NSS) [36] to quantify visual quality as the deviation from the natural image manifold in the spatial or frequency domains [17,40]. With the rise of deep learning, the IQA transitioned to data-driven approaches. These IQA models tend to utilize CNNs [4, 27, 41], transformers [23], or hybrid models [8] to capture both local distortions and long-range dependencies. More recently, driven by vision-language models (VLMs) such as CLIP [25] and BLIP [14], the current NR-IQA paradigm has emphasized multimodal information processing. For instance, CLIP-IQA [33] introduced quality-related textual templates in visual quality evaluations, which is a strategy now central to AIGIQA frameworks such as IPCE [22], AMFF-IQA [44], CIA-Net [43], and CLIP-AGIQA [29]. Additionally, preference-aware models such as ImageReward [37] and MoE-AGIQA [38] utilize reward modelling on web-scale data to

better align with subjective human preferences. Recently, Sun *et al.* [28] also proposed a PICC for prompt-image-caption consistency for the AIGIQA task, which takes an initial exploration in alignment evaluations. Despite these advancements, jointly optimizing visual quality and prompt-image alignment remains a significant challenge. Prevailing strategies, such as handcrafted textual templates [43] or multiscale inputs [21,44], primarily enhance visual quality evaluation but overlook the deep multimodal fusion essential for prompt-image alignment. Therefore, we reframe AIGIQA from an information-theoretic perspective, explicitly modelling cross-modal interactions via a CMIM process in the deep feature domain. This approach enables effective feature integration and achieves superior alignment performance across diverse AIGIQA datasets.

2.2 Mutual information in deep learning

Mutual information is a concept from information theory that estimates the dependency between two random variables and is defined as follows [7]:

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right], \tag{1}$$

where X and Y are two random variables. $p(x,y)$ is their joint probability density function with the marginal probability density functions $p(x)$ and $p(y)$. When X and Y are independent, $p(x,y) = p(x)p(y)$, and $I(X;Y) = 0$. The MIM has emerged as a powerful paradigm for self-supervised representation learning and multimodal information analysis. Early foundational works, such as Deep InfoMax [12], introduce the idea of maximizing the mutual information between local and global feature maps to learn robust visual representations. This concept was further extended by contrastive multiview coding [30], which maximizes the MI across different views of the same data to capture view-invariant features. In multimodal analysis and integration, the MINE [2] provides a foundation for estimating the MI in continuous high-dimensional spaces and has been widely adopted in cross-modal tasks. Moreover, modality-invariant and modality-specific representations (MISAs) [11] are learned by disentangling multimodal features into shared deep feature spaces, achieving superior performance in multimodal sentiment analysis. Furthermore, the MIM has been successfully applied to generative tasks, such as in InfoGAN [5]. Compared with these methods, our CMIM-AIGIQA is the first model to introduce the MIM into the AIGIQA task. Unlike existing works that rely on one-way MIM for representation learning, we propose a cyclic architecture specifically tailored for AIGIQA. The forward phase refines "cross-aware" interactions between dense visual features and sparse prompt features, whereas the backward phase ensures that sparse textual constraints are not overwhelmed by dense visual features in the joint embedding. By explicitly balancing these modalities, our framework effectively addresses the challenge of evaluating both visual quality and prompt-image alignment.

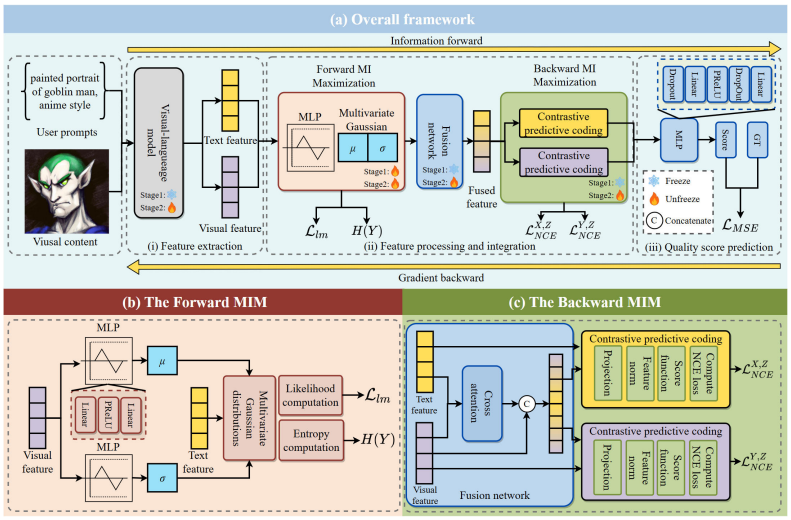

Fig. 2: (a) Overall Framework of our proposed CMIM-AIGIQA. (b) and (c) Details of the forward MIM module and the backward MIM module.

3 Methodology

3.1 Problem formulation

Given the generated image I and the corresponding text prompts P, the AIGI model should include three main components: (1) the multimodal feature encoding network $(E_t(\cdot), E_v(\cdot))$; (2) the multimodal feature fusion network $\phi(\cdot,\cdot)$; and (3) the task-specific score predictor $f(\cdot)$. With these notations, the AIGIQA task can be formulated as follows:

$$\underset{\{(E_t,E_v),\phi,f\}}{\operatorname{argmax}} \quad \operatorname{Corr}(f(\phi(E_t(P),E_v(I))),MOS), \tag{2}$$

where $E_t(P) \in \mathbb{R}^{d_t}$ and $E_v(I) \in \mathbb{R}^{d_v}$ are textual and visual features with dimensions d_t and d_v, respectively. MOS are the mean opinion scores and can be chosen as either visual quality or prompt-image alignment scores. $\operatorname{Corr}(\cdot,\cdot)$ is the correlation coefficient index and is usually chosen as the Pearson linear correlation coefficient (PLCC) or SRCC. Our research focuses on fine-tuning the multimodal encoding network $(E_t(\cdot), E_v(\cdot))$ and designing an effective feature fusion network $\phi(\cdot,\cdot)$ by using the CMIM framework.

3.2 Overall framework

The architecture of our proposed CMIM-AIGIQA is depicted in Fig. 2(a), which consists of three stages: (i) multimodal feature extraction, (ii) feature processing and integration, and (iii) quality score prediction. For feature extraction, given

an AI-generated image I and its corresponding user prompt P, we first employ a pretrained VLM as a backbone to extract the visual features $E_v(I)$ and textual features $E_t(P)$. Unlike existing approaches that incorporate auxiliary visual transformers, we employ only a VLM backbone with partially trainable weights. For feature processing and integration, the CMIM framework collaboratively refines features at the forward and backward phases, where the forward phase computes mainly the MI lower bound between single modality features $E_v(I)$ and $E_t(P)$, and the backward phase maximizes the MI between the fused feature Z and $\{E_v(I),\ E_t(P)\}$. During training, the MI from both phases is backpropagated to jointly fine-tune the pretrained VLM in an end-to-end manner. Z is fused as follows:

$$Z = \text{Concat}(E_v(I), \text{CSA}(E_v(I), E_v(I), E_t(P))), \tag{3}$$

where $\text{Concat}(\cdot,\cdot)$ is the concatenation operation and $\text{CSA}(\cdot,\cdot,\cdot)$ is the cross-attention function defined as $\text{CSA}(Q,K,V) = softmax(\frac{QK^\top}{\sqrt{d_k}})V$. Finally, the joint embedding is fed into a simple regression head to output a final quality score. The entire framework is trained end-to-end in a two-stage process, where the first stage mainly updates weights in the forward MIM module with other weights fixed, and the second stage updates the entire set of weights in the whole network.

3.3 The forward MIM

Denote $X = E_v(I)$ and $Y = E_t(P)$; then, the $I(X;Y)$ between X and Y is defined by Eq. (1). Notably, $I(X;Y) = 0$ denotes the situation in which the image generation is completely independent of the prompt. In the forward phase, the goal of maximizing $I(X;Y)$ is to preserve the modality-invariant semantic content and filter out the modality-specific noise. Directly computing $I(X;Y)$ is intractable because the joint distribution of $p(x,y)$ is unknown. Following Barber and Agakov [1], we compute a tractable lower bound for the MIM. Specifically, we approximate the conditional distribution $p(y|x)$ with a variational counterpart $q_\theta(y|x)$, and the MI lower bound can be derived as follows:

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{q_\theta(y|x)}{p(y)}\right] + \mathbb{E}_{p(x)}[KL(p(y|x)|q_\theta(y|x))], \tag{4}$$

$$\geq \mathbb{E}_{p(x,y)}[\log q_\theta(y|x)] + H(Y), \tag{5}$$

where Eq. (4) to Eq. (5) is because the KL-divergence is nonnegative. $H(Y)$ denotes the differential entropy of the textual features. In our implementation, we treat the image feature as the condition and the textual feature as the target for $q_\theta(y|x)$. This is motivated by our observation that the images enjoy denser information than the textual prompts do. Predicting relatively sparse textual information from information-compact visual features is logical; otherwise, the prediction becomes ill-posed. To compute $E_{p(x,y)}[\log q_\theta(y|x)]$, we model $q_\theta(y|x)$ as a multivariate Gaussian distribution, which is a typical choice in MI research [6].

Specifically, $q_\theta(y|x) = \mathcal{N}(y|\mu_\theta(x), \sigma_\theta^2(x)I)$, where μ_θ and σ_θ are predicted by two independent multilayer perceptrons (MLPs) with feature sample x as the input. Consequently, maximizing $E_{p(x,y)}[\log q_\theta(y|x)]$ is equivalent to maximizing the log-likelihood of the target modality, which is defined as follows:

$$\mathcal{L}_{lm} = -\frac{1}{N}\sum_{i=1}^{N} \log q_\theta(y_i|x_i), \tag{6}$$

where N is the batch size and $\{y_i, x_i\}$ represents the prompt-image feature samples. For $H(Y)$, we also assume that the marginal distribution of the textual features follows a multivariate Gaussian distribution, i.e., $p(y) \sim \mathcal{N}(\mu_y, \Sigma_y)$. The differential entropy can be computed in a closed-form expression:

$$H(Y) = \frac{1}{2}\ln\left((2\pi e)^k \det(\Sigma_y)\right), \tag{7}$$

where k denotes the dimension of the textual feature space and Σ_y is the covariance matrix. During the training process, the mean μ_y and the covariance matrix Σ_y are estimated from each mini-batch of textual features.

We assume that the marginal distribution of Y is Gaussian for two reasons. First, the Gaussian assumption yields a closed-form expression Eq. (7), which ensures a stable and smooth optimization of the mutual information lower bound during end-to-end training. Second, $H(Y)$ acts as a structural regularizer for the textual feature space. The Gaussian prior is a conservative estimate of the topological structure for such a space: the textual representation should be evenly distributed on the surface of a sphere such that the model can preserve an informative semantic space [6, 35]. Without $H(Y)$, the model may converge to a trivial solution where the textual features Y collapse into a narrow region to maximize the likelihood L_{lm} but lose semantic information of the textual prompts. The validity of the Gaussian assumption on the textual features is further checked in the supplementary. Finally, the total loss for the forward phase is $\mathcal{L}_{forward} = -\mathcal{L}_{lm} - H(Y)$.

3.4 The backward MIM

In the backward phase, we expect the fusion result $Z = \phi(X, Y)$ to retain sufficient information to inversely "predict" the essential information of X and Y. We combine contrastive predictive coding (CPC) [20] with the noise-contrastive estimation (NCE) framework [9], resulting in a method that does not need to explicitly assume the distribution shape. Specifically, we employ a score function based on the cosine similarity [20] between the fused embedding and the individual modality features. For each modality $M \in \{X, Y\}$, we use a projection network $G_\psi(\cdot)$, a small MLP with network parameter ψ, to project Z into the same modality space. The correlation between the projected feature and the ground-truth feature is defined as a normalized score function $s(M, Z) = \exp(\overline{M} \bullet \overline{G_\psi(Z)})$, where $\overline{M} = M/||M||_2$, $\overline{G_\psi(Z)} = G_\psi(Z)/||G_\psi(Z)||_2$, and $\exp(\cdot)$ is the exponential function. The score function is further used in the NCE framework [9],

which provides a parameter-free method to maximize the statistical dependency by discriminating the positive pair from a set of negative samples. Specifically, for a given modality feature M_i in a batch $B_M = \{M_1, ..., M_N\}$, we treat M_i as the positive sample and the remaining $N-1$ representations in the same batch as the negative samples. The NCE loss [9] is subsequently calculated as follows:

$$\mathcal{L}_{NCE}(Z, \mathbf{B}_M) = -\mathbb{E}_{\mathbf{H}} \left[\log \frac{s(Z, M_i)}{\sum_{M_j \in \mathbf{B}_M} s(Z, M_j)} \right]. \tag{8}$$

We combine CPC [20] with the NCE [9] framework in the backward phase for two reasons. First, CPC allows Z to inversely predict features across modalities, ensuring that modality-invariant information is effectively passed to the final embedding. Second, the NCE prevents the "modality dominance" issue. By enforcing a discriminative correlation between Z and its corresponding prompt in the batch, the framework forces the model to retain critical textual constraints. We do not use the same MIM framework as the forward phase does because the Gaussian likelihood-based approach leads to feature oversmoothing for Z, which is not good for fine-grained prompt-image alignment evaluations. Moreover, computing $H(X)$ is also nontrivial because X is usually very high-dimensional, and computing the log-determinant of the covariance matrix of X easily leads to numerical instability even with the Gaussian assumption. Finally, the total loss for the backward phase is given by the sum of the NCE losses for both the visual and textual modalities $\mathcal{L}_{backward} = \mathcal{L}_{NCE}^{X,Z} + \mathcal{L}_{NCE}^{Y,Z}$.

4 Experiments

4.1 Training and implementation details

The training for CMIM-AIGIQA has two stages. In the first stage, only the weights of the two MLPs in the forward phase are trained to approximate the conditional distribution $p(y|x)$ with $q_\theta(y|x)$. The loss function is to minimize the negative log-likelihood formulated in Eq. (6). In the second stage, the weights of the whole network are trained, including partially trainable VLM weights, two MLP weights in the forward phase, the fusion network weights, the projection weights in the backward phase, and the quality prediction network weights. Upon the final prediction $\hat{y}$ to the ground truth y, the mean square error (MSE) is computed as the task loss. Consequently, the total loss for this stage is formulated as a weighted summation:

$$\mathcal{L} = \mathcal{L}_{task} + \alpha \mathcal{L}_{forward} + \beta \mathcal{L}_{backward}, \tag{9}$$

where $\alpha = 0.1$ and $\beta = 0.1$ are hyperparameters that control the impact of the MIM and are selected via a grid search where $\alpha, \beta \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$ on the AGIQA-3k [13] dataset. We set $\alpha, \beta < 0.5$ as the upper bound to maintain the regression loss as the primary objective, which balances the MI constraints to provide sufficient guidance for feature alignment without overwhelming the

Table 1: Brief summary of the used AGI quality assessment databases.

Datasets	Published year	Evaluation dimensions	No. of text prompts	No. of images	Score range
AGIQA-3K	2023	Quality Alignment	300	2982	[0,5]
AIGCIQA2023	2023	Quality Alignment	100	2400	[0,5]
PKU-AIGIQA-4k (Text-to-image part)	2025	Quality Alignment	200	2400	[0,5]

main loss. We also set $\alpha, \beta > 0.05$ to ensure that the cyclic MIM is not too weak to provide sufficient constraints. For the VLM, we use the BLIP [14] network architecture and fix the shallow-layer weights of both the image and the text encoders and fine-tune the remaining half of the deep-layer weights, the same settings as in the MoE-AGIQA [38]. The AIGIs are resized to 224×224 as the image input, and the text inputs are the raw prompts without any delicately designed template. The batch size is set to 5 for reproducibility on memory-limited GPUs. We adopt the AdamW optimizer [16] with the default settings to train our model. We set momentum coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to ensure stable adaptive optimization and enough momentum to escape local minima. We also set $\epsilon = 1 \times 10^{-8}$ and a weight decay of 0.001 to maintain numerical stability and mitigate overfitting. The initial learning rate is set to $lr = 1 \times 10^{-5}$, which is decayed by a factor of 0.5 every 4 epochs over a total of 20 training epochs. We adopt a conservative initial learning rate to ensure a stable training process, and decaying every 4 epochs also ensures progressive convergence. A total of 20 epochs is chosen because the scale of existing AIGIQA datasets is relatively small. For all the evaluation benchmarks, we employ a random split of 80% for training and 20% for testing. This process is repeated 10 times, and the average SRCC and PLCC values are reported. All training is conducted on NVIDIA RTX 3090 GPUs.

4.2 Experimental settings

We evaluate our method on three AIGIQA datasets: AGIQA-3k [13], AIGCIQA2023 [34], and the text-to-image part of the PKU-AIGIQA-4k [39]. The details are presented in Table 1. We compare our models with several blind IQA milestones, including the BRISQUE [18], the NIQE [19], the BMPRI [17], the DBCNN [41], the HyperIQA [27], and the WaDIQaM-NR [4] and several SOTA AIGIQA methods, including the IP-IQA [24], IPCE [22], CLIP-AGIQA [29], AMFF-IQA [44], MoE-AGIQA [38], and the CIA-Net [43]. We employ the PLCC and the SRCC to quantify the agreement between the model predictions and ground-truth subjective scores. The definitions of these metrics are detailed in [36].

Table 2: Quantitative results between different IQA methods on three AIGI datasets. Red, green, and blue denote the first, second, and third highest scores in each column.

Method	AGIQA-3K [13]				AIGCIQA2023 [34]				PKU-AIGIQA-4k [39]			
	Quality		Alignment		Quality		Alignment		Quality		Alignment	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BRISQUE	0.6210	0.5657	0.5683	0.4794	0.6389	0.6239	0.4280	0.4219	0.4832	0.4441	0.3431	0.3268
NIQE	0.5171	0.5623	0.5216	0.4369	0.5218	0.5060	0.3485	0.3659	0.5169	0.4765	0.2895	0.2959
BMPRI	0.7912	0.6794	0.6565	0.5794	0.7492	0.6732	0.4346	0.3946	0.5525	0.4961	0.3307	0.3168
DBCNN	0.8759	0.8207	0.7823	0.6329	0.8577	0.8339	0.6787	0.6837	0.5970	0.5975	0.5925	0.6083
HyperIQA	0.8903	0.8355	0.8087	0.6276	0.8689	0.8483	0.7439	0.7541	0.6955	0.6849	0.7062	0.7239
WaDIQaM-NR	0.3930	0.2190	0.3589	0.2568	0.4996	0.4447	0.2810	0.3027	0.6347	0.6231	0.6456	0.6480
IP-IQA	0.9116	0.8634	0.8544	0.7578	0.8794	0.8589	0.7772	0.7583	0.7439	0.7368	0.7743	0.7565
IPCE	0.9246	0.8841	0.8725	0.7697	0.8788	0.8640	0.7887	0.7979	0.8452	0.8391	0.8173	0.7900
CLIP-AGIQA	0.8978	0.8618	0.8554	0.7434	0.8302	0.8140	0.7678	0.7655	0.7388	0.7298	0.7549	0.7213
AMFF-IQA	0.9050	0.8565	0.8476	0.7513	0.8537	0.8409	0.7638	0.7782	0.8312	0.8268	0.8025	0.7809
MoE-AGIQA	0.9282	0.8746	0.8916	0.8238	0.8904	0.8751	0.7832	0.7899	0.8433	0.8387	0.7860	0.7458
CIA-Net	0.9134	0.8709	0.8687	0.7797	0.8756	0.8557	0.7800	0.7977	0.8967	0.8638	0.8466	0.7978
CMIM-AIGIQA	0.9161	0.8755	0.9181	0.8546	0.8898	0.8679	0.8158	0.8263	0.8786	0.8586	0.8930	0.8223

4.3 Experimental results and analysis

The evaluation results for the three datasets are presented in Table 2. There are two key observations. First, our CMIM-AIGIQA consistently achieves the highest alignment performance (marked in red) across all three datasets in both the PLCC and the SRCC. On the AGIQA-3K [13] dataset, CMIM-AIGIQA achieves an SRCC of 0.8546, outperforming the second-best model (MoE-AGIQA [38]) by a substantial margin of 3.74% (0.8546 vs. 0.8238). Similar trends are also observed for AIGCIQA2023 [34] and PKU-AIGIQA-4k [39]. These results strongly demonstrate that our framework captures the core information between visual and textual modalities to achieve superior performance in prompt-image alignment evaluations. Second, our CMIM-AIGIQA also shows competitive performance in terms of the visual quality dimensions, which consistently maintains the top 3 best methods. Specifically, on AIGCIQA2023, we achieve a PLCC of 0.8898 (2nd) and an SRCC of 0.8679 (2nd) in the visual quality dimensions, whereas our alignment SRCC is significantly better than the second-best method (IPCE [22]), with an enhancement of 3.56% (0.8263 vs. 0.7979). These results demonstrate that our framework balances the modality information vital for visual quality assessment. In contrast, the performance of the MoE-AGIQA [38] is imbalanced between the two evaluation dimensions. It delivers superior visual quality SRCCs but significantly degraded alignment SRCCs on AIGCIQA2023 [34] (quality 0.8751 vs. alignment 0.7899) and PKU-AIGIQA-4k [39] (quality 0.8387 vs. alignment 0.7458). Notably, the MoE-AGIQA [38] also adopts the same BLIP architecture as ours but shows imbalanced performance, demonstrating that our mutual information framework is indispensable in bridging the gap between visual quality and semantic alignment.

Table 3: Cross-dataset evaluation results of different AIGIQA methods. "AGIQA-3k ⟶ AIGCIQA2023" denotes training models on the AGIQA-3k dataset and testing on AIGCIQA2023. Red denotes the highest score in each column.

Method	AGIQA-3k ⟶ AIGCIQA2023				AGIQA-3K ⟶ PKU-AIGIQA-4k			
	Quality		Alignment		Quality		Alignment	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
AMFF-IQA	0.6685	0.6776	0.5485	0.8461	0.4557	0.4168	0.4009	0.3431
CIA-Net	0.7141	0.6816	0.7443	0.6506	0.4765	0.4006	0.4861	0.3952
MoE-AGIQA	0.7547	0.7619	0.6474	0.6499	0.3218	0.2676	0.6146	0.5058
CMIM-AIGIQA	0.7743	0.7632	0.6750	0.6795	0.5547	0.4855	0.6484	0.5532
Method	AIGCIQA2023 ⟶ AGIQA-3k				AIGCIQA2023 ⟶ PKU-AIGIQA-4k			
	Quality		Alignment		Quality		Alignment	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
AMFF-IQA	0.6951	0.6542	0.6240	0.5537	0.3516	0.2877	0.3148	0.2229
CIA-Net	0.6528	0.6600	0.5498	0.5287	0.3764	0.2968	0.3539	0.2687
MoE-AGIQA	0.8163	0.7500	0.8217	0.7530	0.2821	0.2459	0.4054	0.2612
CMIM-AIGIQA	0.7841	0.7573	0.8177	0.7441	0.4894	0.4853	0.5767	0.4794
Method	PKU-AIGIQA-4k ⟶ AGIQA-3k				PKU-AIGIQA-4k ⟶ AIGCIQA2023			
	Quality		Alignment		Quality		Alignment	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
AMFF-IQA	0.5216	0.4222	0.6855	0.6158	0.2987	0.2946	0.2734	0.2655
CIA-Net	0.5658	0.4697	0.7278	0.6456	0.3438	0.3336	0.2661	0.2548
MoE-AGIQA	0.6714	0.6502	0.8127	0.7349	0.4962	0.4903	0.5666	0.5646
CMIM-AIGIQA	0.7488	0.7382	0.8047	0.7412	0.5400	0.5436	0.6249	0.6305

4.4 Cross-dataset evaluations

We conduct cross-dataset evaluations to verify the generalization ability of our proposed CMIM-AIGIQA in Table 3. All the models are trained on the source dataset and directly tested on the unseen target dataset. There are two key observations. First, CMIM-AIGIQA demonstrates superior cross-dataset performance, achieving the highest alignment scores in most of the evaluation pairs. When trained on the PKU-AIGIQA-4k [39] and tested on the AIGCIQA2023 [34], the CMIM-AIGIQA outperforms the second-best model (MoE-AGIQA [38]) by a margin of 11.67% in the alignment SRCC (0.6305 vs. 0.5646). Even in the "AIGCIQA2023→AGIQA-3k" pair, where MoE-AGIQA [38] shows strong alignment results, CMIM-AIGIQA remains competitive and achieves a second-best alignment SRCC of 0.7441. These results demonstrate that our model has successfully captured the intrinsic, dataset-invariant features of AI-generated content. Second, our CMIM-AIGIQA delivers robust prediction performance across different training–testing pairs, without a significant decrease in performance. Existing AIGIQA methods usually undergo a sharp performance decline when they encounter a domain shift from the AGIQA-3k [13] and the AIGCIQA2023 [34] to the PKU-AIGIQA-4k [39]. For instance, the MoE-AGIQA [38] performs well in the "AIGCIQA2023 → AGIQA-3k" pairs, but its quality and alignment SRCCs decrease significantly to 0.2459 and 0.2612 in "AIGCIQA2023→ PKU-AIGIQA-4k". Similar trends are also shown for the CIA-Net [43] and the

Table 4: Results of ablation studies on three AIGI datasets. We report mainly the SRCC results. Red denotes the highest score in each column.

Method	AGIQA-3K [13]		AIGCIQA2023 [34]		PKU-AIGIQA-4k [39]	
	Quality	Alignment	Quality	Alignment	Quality	Alignment
Baseline	0.8124	0.7731	0.8015	0.7428	0.7756	0.7204
w/o Forward MIM	0.8432	0.8415	0.8329	0.8110	0.8104	0.8058
w/o Backward MIM	0.8567	0.7984	0.8412	0.7655	0.8321	0.7512
w/o $H(Y)$	0.8654	0.8502	0.8496	0.8187	0.8412	0.8094
$\mathcal{L}_{NCE} \rightarrow \mathcal{L}_{MSE}$	0.8512	0.8156	0.8533	0.7892	0.8445	0.7936
w/o CSA	0.8610	0.8327	0.8422	0.8045	0.8207	0.7984
w/CLIP	0.8931	0.8214	0.8498	0.8103	0.8683	0.8101
CMIM-AIGIQA	0.8755	0.8546	0.8679	0.8263	0.8586	0.8223

AMFF-IQA [44]. In contrast, our CMIM-AIGIQA maintains stable and high performance in the two pairs. We attribute this robustness to our CMIM framework. By explicitly maximizing the mutual information in both forward and backward paths, the model is forced to learn the essential semantic-visual cues rather than overfitting to the specific score.

4.5 Ablation studies

The CMIM framework. To verify the efficacy of CMIM, we evaluated three variants: 1. The baseline model contains only multimodal feature fusion and is trained with MSE loss; 2. w/o forward MIM; and 3. w/o backward MIM. The results are shown in Table 4. We have three observations. First, the full framework significantly outperforms the baseline across all the datasets, with the most notable gains observed in the alignment dimension (e.g., 0.8546 vs. 0.7731 SRCC on AGIQA-3K [13]). Second, removing the forward MIM leads to degradations in both visual quality and prompt-image alignment. On AGIQA-3K [13], the quality of the SRCC decreases from 0.8755 to 0.8432, and the alignment of the SRCC decreases from 0.8546 to 0.8415. This confirms that the forward phase effectively filters out modality-specific noise and refines cross-modal features to achieve superior quality and alignment SRCCs. Third, the absence of the backward MIM leads to mild degradation in quality but significant degradation in alignment. On AIGCIQA2023 [34], the alignment SRCC decreases from 0.8263 to 0.7655, demonstrating that without the backward phase, the dense visual features tend to dominate the fused embedding, causing the sparse textual prompt cues to be marginalized. In conclusion, both the forward and backward phases are indispensable for bridging the gap between visual quality and prompt-image alignment evaluations.

The loss function analysis. We investigate the effectiveness of the textual entropy $H(Y)$ and the NCE loss $\mathcal{L}_{NCE}$. As shown in Table 4, removing $H(Y)$ (variant w/o $H(Y)$) leads to a consistent decrease in performance across all the

benchmarks. For instance, the alignment SRCC decreases from 0.8546 to 0.8502 for the AGIQA-3K [13], 0.8263 to 0.8187 for the AIGCIQA2023 [34], and 0.8223 to 0.8094 for the PKU-AIGIQA-4k [39]. We suspect that the reason is that without the entropy term as the regularization term, the model learns that the textual features collapse to a restricted region to maximize the likelihood $\mathcal{L}_{lm}$ but reduce the overall discriminative capacity of the semantic space. On the other hand, the NCE framework is pivotal in our backward phase. To verify its effectiveness, we replace $\mathcal{L}_{NCE}$ with the MSE loss $\mathcal{L}_{MSE}$ (variant $\mathcal{L}_{NCE} \rightarrow \mathcal{L}_{MSE}$). Such a strategy forces the projected fused feature $G_\psi(Z)$ to match the two single embeddings $M \in \{X, Y\}$, *i.e.*, $\mathcal{L}_{MSE}(Z, M) = ||G_\psi(Z) - M||_2$. Nevertheless, we observe a significant decrease in alignment scores, which decreases from 0.8546 to 0.8156 on AGIQA-3k [13], 0.8263 to 0.7892 on AIGCIQA2023 [34], and 0.8223 to 0.7936 on PKU-AIGIQA-4k [39]. We assume that the reason is that MSE focuses on minimizing pixelwise distances, which often leads to "oversmoothed, averaged" representations that ignore fine-grained semantic cues [42]. In contrast, $\mathcal{L}_{NCE}$ introduces discriminative constraints through negative sampling, forcing the joint embedding Z to retain the unique "identity" of its corresponding prompt.

The fusion network & VLM backbone. We investigate the efficacy of the CSA module and the VLM backbone in Table 4. Removing the CSA module (variant "w/o CSA") and directly concatenating the visual and textual features leads to performance degradation across all benchmarks. On PKU-AIGIQA-4k [39], the alignment SRCC decreases from 0.8223 to 0.7984, indicating that the CSA for visual-textual semantic alignment is necessary. When our VLM backbone is substituted with a standard CLIP [25] (ViT-L/14), we find that the performance on the quality dimensions is enhanced, but the alignment dimensions decrease, as shown in Table 4. Specifically, while it achieves competitive visual quality SRCC on AGIQA-3K [13] and PKU-AIGIQA-4k [39], its alignment performance is inferior to that of our default model (e.g., alignment SRCC 0.8214 vs. 0.8546 on AGIQA-3K). We suspect that this explains why existing CLIP-based methods often show high sensitivity to visual quality but fail to evaluate prompt-image alignment. We also need to note that, even with the CLIP [25], our method still outperforms existing CLIP-based AIGIQA methods in alignment evaluations. As shown in Table 4, the CLIP variant of our model consistently surpasses the 0.80 SRCC threshold in alignment—a performance level unattained by existing CLIP-based methods, demonstrating the effectiveness of our CMIM framework.

5 Conclusion

In this paper, we address a critical challenge in the AIGIQA task: balancing visual quality and prompt-image alignment evaluations. We propose the CMIM-AIGIQA framework, which addresses multimodal assessment as an information-theoretic optimization problem. By introducing a cyclic mutual information maximization process, our model effectively addresses the issue of "modality noise" and prevents the loss of sparse textual constraints during fusion. Extensive exper-

iments on three large-scale AIGIQA benchmarks demonstrate that our method significantly outperforms existing SOTA metrics, particularly in the alignment dimension, while maintaining highly competitive visual quality evaluation. Our findings demonstrate that maximizing the mutual information for both modalities is essential for bridging perceptual quality and semantic alignment. We believe that our work provides a robust foundation for the comprehensive AIGIQA task, ultimately contributing to the development of more controllable and high-fidelity AI-generated content.

Acknowledgements

This work is supported in part by the Ministry of Education, Singapore, under the funding of Tier 1 RG103/24; in part by Chongqing Talent under Grant cstc2024ycjh-bgzxm0082; and in part by Lingnan University under faculty grants F106101 and F105131.

References

1. Barber, D., Agakov, F.: The im algorithm: a variational approach to information maximization. Advances in neural information processing systems **16**(320), 201 (2004)
2. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: International conference on machine learning. pp. 531–540. PMLR (2018)
3. Bie, F., Yang, Y., Zhou, Z., Ghanem, A., Zhang, M., Yao, Z., Wu, X., Holmes, C., Golnari, P., Clifton, D.A., et al.: Renaissance: A survey into ai text-to-image generation in the era of large model. IEEE transactions on pattern analysis and machine intelligence (2024)
4. Bosse, S., Maniry, D., Müller, K.R., Wiegand, T., Samek, W.: Deep neural networks for no-reference and full-reference image quality assessment. IEEE Transactions on image processing **27**(1), 206–219 (2017)
5. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. Advances in neural information processing systems **29** (2016)
6. Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., Carin, L.: Club: A contrastive log-ratio upper bound of mutual information. In: International conference on machine learning. pp. 1779–1788. PMLR (2020)
7. Cover, T.M.: Elements of information theory. John Wiley & Sons (1999)
8. Golestaneh, S.A., Dadsetan, S., Kitani, K.M.: 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. pp. 1220–1230 (2022)
9. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 297–304. JMLR Workshop and Conference Proceedings (2010)

10. Hartwig, S., Engel, D., Sick, L., Kniesel, H., Payer, T., Poonam, P., Glöckler, M., Bäuerle, A., Ropinski, T.: A survey on quality metrics for text-to-image generation. IEEE Transactions on Visualization and Computer Graphics **31**(10), 9464–9483 (2025). https://doi.org/10.1109/TVCG.2025.3585077
11. Hazarika, D., Zimmermann, R., Poria, S.: Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM international conference on multimedia. pp. 1122–1131 (2020)
12. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018)
13. Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai, G., Lin, W.: Agiqa-3k: An open database for ai-generated image quality assessment. IEEE Transactions on Circuits and Systems for Video Technology **34**(8), 6833–6846 (2023)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
15. Li, X., Sheng, X., Wang, M., Ou, F.Z., Chen, B., Wang, S., Kwong, S.: Cofaco: Controllable generative talking face video coding. IEEE Transactions on Image Processing (2026)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
17. Min, X., Zhai, G., Gu, K., Liu, Y., Yang, X.: Blind image quality estimation via distortion aggravation. IEEE Transactions on Broadcasting **64**(2), 508–517 (2018)
18. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050
19. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2013). https://doi.org/10.1109/LSP.2012.2227726
20. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
21. Pan, Z., Yang, Y., Yuan, F., Xie, H., Wang, F.L., Kwong, S.: Assessing ai-generated image quality using a cross-modal hierarchical perception network. IEEE Transactions on Broadcasting (2025)
22. Peng, F., Fu, H., Ming, A., Wang, C., Ma, H., He, S., Dou, Z., Chen, S.: Aigc image quality assessment via image-prompt correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6432–6441 (2024)
23. Qin, G., Hu, R., Liu, Y., Zheng, X., Liu, H., Li, X., Zhang, Y.: Data-efficient image quality assessment with attention-panel decoder. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2091–2100 (2023)
24. Qu, B., Li, H., Gao, W.: Bringing textual prompt to ai-generated image quality assessment. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
26. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In:

Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
27. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: 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. pp. 3667–3676 (2020)
28. Sun, W., Chen, C., Liao, L., Lin, W.: Prompt-image-caption consistency for ai-generated image quality assessment. IEEE Transactions on Multimedia pp. 1–11 (2026). https://doi.org/10.1109/TMM.2026.3668530
29. Tang, Z., Wang, Z., Peng, B., Dong, J.: Clip-agiqa: boosting the performance of ai-generated image quality assessment with clip. In: International Conference on Pattern Recognition. pp. 48–61. Springer (2024)
30. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: European conference on computer vision. pp. 776–794. Springer (2020)
31. Tian, Y., Li, Y., Chen, B., Zhu, H., Wang, S., Kwong, S.: Ai-generated image quality assessment in visual communication. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7392–7400 (2025)
32. Tian, Y., Liu, Y., Wang, S., Kwong, S.: Quality assessment for text-to-image generation: A survey. IEEE MultiMedia (2025)
33. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)
34. Wang, J., Duan, H., Liu, J., Chen, S., Min, X., Zhai, G.: 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. pp. 46–57. Springer (2023)
35. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. pp. 9929–9939. PMLR (2020)
36. Wang, Z., Bovik, A.C.: Modern image quality assessment **Synthesis Lectures on Image, Video, and Multimedia Processing**, 1–156 (2006)
37. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
38. Yang, J., Fu, J., Zhang, W., Cao, W., Liu, L., Peng, H.: Moe-agiqa: Mixture-of-experts boosted visual perception-driven and semantic-aware quality assessment for ai-generated images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6395–6404 (2024)
39. Yuan, J., Li, J., Yang, F., Cao, X., Che, J., Lin, J., Cao, X.: Pku-aigiqa-4k: A perceptual quality assessment database for both text-to-image and image-to-image ai-generated images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3331–3340 (2025)
40. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. IEEE Transactions on Image Processing **24**(8), 2579–2591 (2015)
41. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: 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)
42. Zhou, J., You, C., Li, X., Liu, K., Liu, S., Qu, Q., Zhu, Z.: Are all losses created equal: A neural collapse perspective. Advances in Neural Information Processing Systems **35**, 31697–31710 (2022)

43. Zhou, T., Tan, S., Li, L., Zhao, B., Jiang, Q., Yue, G.: Cross-modality interactive attention network for ai-generated image quality assessment. Pattern Recognition **167**, 111693 (2025)
44. Zhou, T., Tan, S., Zhou, W., Luo, Y., Wang, Y.G., Yue, G.: Adaptive mixed-scale feature fusion network for blind ai-generated image quality assessment. IEEE Transactions on Broadcasting **70**(3), 833–843 (2024)