AIGIQA-20K: A Large Database for AI-Generated Image Quality Assessment

Chunyi Li[1*], Tengchuan Kou[1*], Yixuan Gao[1], Yuqin Cao[1], Wei Sun[1],
Zicheng Zhang[1], Yingjie Zhou[1], Zhichao Zhang[1], Weixia Zhang[1], Haoning Wu[2],
Xiaohong Liu[1,†], Xiongkuo Min[1,†], Guangtao Zhai[1,†]
Shanghai Jiao Tong University[1]
Nanyang Technological University[2]

Abstract

With the rapid advancements in AI-Generated Content (AIGC), AI-Generated Images (AIGIs) have been widely applied in entertainment, education, and social media. However, due to the significant variance in quality among different AIGIs, there is an urgent need for models that consistently match human subjective ratings. To address this issue, we organized a challenge towards AIGC quality assessment on NTIRE 2024 that extensively considers 15 popular generative models, utilizing dynamic hyper-parameters (including classifier-free guidance, iteration epochs, and output image resolution), and gather subjective scores that consider perceptual quality and text-to-image alignment altogether comprehensively involving 21 subjects. This approach culminates in the creation of the largest fine-grained AIGI subjective quality database to date with 20,000 AIGIs and 420,000 subjective ratings, known as AIGIQA-20K. Furthermore, we conduct benchmark experiments on this database to assess the correspondence between 16 mainstream AIGI quality models and human perception. We anticipate that this large-scale quality database will inspire robust quality indicators for AIGIs and propel the evolution of AIGC for vision. The database is released on https://www.modelscope.cn/datasets/lcysyzxdxc/AIGCQA-30K-Image.

1. Introduction

AI Generated Content (**AIGC**) refers to various types of content generated by artificial intelligence, such as images, videos, texts, and music. Among those modalities, AI-Generated Images (AIGIs), especially Text-to-Image (T2I) models, have already revolutionized the paradigm of entertainment, education, and social media. According to huggingface[1], there were 10,000+ T2I models coexisting on the internet that generated results of widely varying quality. As vision is the dominant way for humans to perceive external information, a universal quality indicator for this new visual information is a topic worth investigating in the AIGC era.

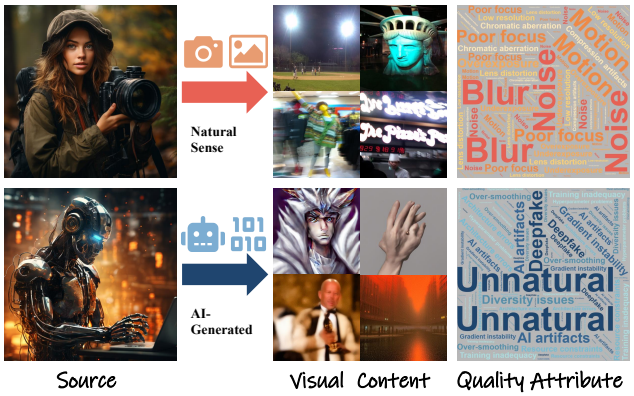


Figure 1. Illustration of the difference between Natural Sense Content and AI-Generated Content, whose perceptual quality are affected by different attributes.

However, existing Image Quality Assessment (IQA) metrics [12, 20, 23] cannot be applied in AIGIs directly. As Figure 1 shows, the quality of Natural Sense Images/Videos (NSIs) is determined by distortion in the imaging process (e.g. blur, noise) [19, 55] while the quality of AIGIs is more closely related to hardware limitations and technical proficiency [22, 40, 41, 43, 53] (e.g. unnatural, deepfake). Besides, T2I alignment is also an important factor for AIGI which is absent in traditional IQA tasks. The quality of AIGC is a mixture of perceptual and alignment quality. Therefore, towards a strong quality indicator specifically for AIGC, an AIGI quality database is highly demanded to illustrate their quality-aware attributes besides NSIs.

In the past year, the emerging demand for AIGI qual-

* Equal contribution.
† Corresponding authors.

[1] https://huggingface.co, data collected in March 2024

Table 1. Existing quality databases for AI-Generated Images/Videos.

Database	Grain	Size	Ratings	Models	CFG	Iteration	Resolution
HPD [44]	Coarse-grained	98,807	98,807	1	Fixed	Fixed	Fixed
ImageReward [45]	Coarse-grained	136,892	136,892	3	Fixed	Fixed	Dynamic
Pick-A-Pic [16]	Coarse-grained	500,000	500,000	6	Fixed	Fixed	Dynamic
AGIQA-1K [50]	Fine-grained	1,080	23,760	2	Fixed	Fixed	Fixed
AGIQA-3K [24]	Fine-grained	2,982	125,244	6	Dynamic	Dynamic	Fixed
AIGCIQA [39]	Fine-grained	2,400	48,000	6	Fixed	Fixed	Fixed
AGIN [4]	Fine-grained	6,049	181,470	18	Dynamic	Fixed	Fixed
AIGIQA-20K	Fine-grained	20,000	420,000	15	Dynamic	Dynamic	Dynamic

ity has spawned several related databases as shown in Table 1 including two main categories: coarse-grained and fine-grained. The former usually has a larger data size, with only one user scoring the images or selecting preferences for image pairs. Thus, such scoring has strong discontinuities and bias; the latter has a smaller scale, but the quality scores are derived from the Mean Opinion Score (MOS) [51, 52, 54] of more than 15 users, which accurately characterizes the image quality. Meanwhile, along with the rapid development of generative models, the AIGI quality database needs to consider an increasing number of models. Besides, the quality of AIGI not only depends on the T2I model itself, where the hyper-parameters also play a decisive role. Thus, to reflect the actual distortion of AIGI, these factors also need to be dynamically adjusted.

Facing the above challenges, this paper lays the foundation of the NTIRE 2024 AIGCQA Grand Challenge [25] to inspire effective quality metrics for AIGC, which contributes (i) a quality database named AIGIQA-20K that extensively covers 15 T2I models. Meanwhile, it dynamically adjusts for both resolution and hyper-parameters for the first time, which comprehensively characterizes the visual distortion of AIGC. (ii) a comprehensive set of subjective quality labels. For the AIGIQA-20K, we organized 21 subjects to produce accurate MOS scores. As a fine-grained AIGI database, it has the largest size to date. (iii) an exhaustive benchmark experiment for AIGC quality assessment. The indicators cover both traditional IQA [17] and T2I alignment methods, which can inspire more accurate quality metrics in the future. The rest of the paper is organized as follows. In Sec 2, details of the proposed AIGIQA-20K are provided. Sec 3 analyze the subjective scoring of AIGIs. Sec 4 validate the several quality indicators on the AIGIs. Finally, a conclusion is provided in Sec 5.

2. Database Construction

2.1. Hyper-parameter Configuration

The quality of model generation is closely related to the hyper-parameters. Due to the limited computational resources and different settings, these hyper-parameter configurations change frequently in the actual generation process. Among them, insufficient iterations will reduce image detail; too high/low Classifier Free Guidance (CFG) will affect the tradeoff between perceptual/alignment quality; non-square resolution will cause a sharp drop in overall quality. Therefore, before the generation process for each T2I model, our AIGIQA-20K database dynamically set these quality-aware configurations with the following criteria:

- Iterations: 50% as default full epochs, 25% as $0.5\times$ epochs, and 25% as $0.25\times$ epochs.
- CFG: 50% as default CFG number, 20% as $0.5\times$ default CFG, 20% as $2\times$ default CFG, and 10% applies $0.5 \sim 2$ default CFG randomly.
- Resolution: 50% as $1:1$ square, four 10% as $3:4$, $4:3$, $9:16$, $16:9$, and final 10% as $9:16 \sim 16:9$ randomly. The longest edge is set as 512 or 1,024 according to the maximum resolution of the model.

where the configuration adjustments for each model are described in the next section.

2.2. Generative Model Collection

Based on the size of previous fine-grained database [8, 18], the AIGIQA-20K includes 2,000 $\times$ 7+1,000 $\times$ 4+500 $\times$ 4=20,000 images. Considering that the overall generation effect of the diffusion-based model is well-developed and widely used, we discard previous Generative Adversarial Network (GAN) and Auto-Regressive (AR) [6] models that have been eliminated for the current T2I generation task. To ensure content diversity, our AIGCQA-20K database considered 15 representative T2I generative models in Figure 2. For each model, with specific configurations in Sec 2.1, we generate the following number of images:

- 2,000 images: Dreamlike, Pixart α, Playground v2, SD1.4, SD1.5, SDXL, SSD1B [3,5,7,9,30,33]. These models have strong generalize ability so we change all three hyper-parameters with default iterations as 40.

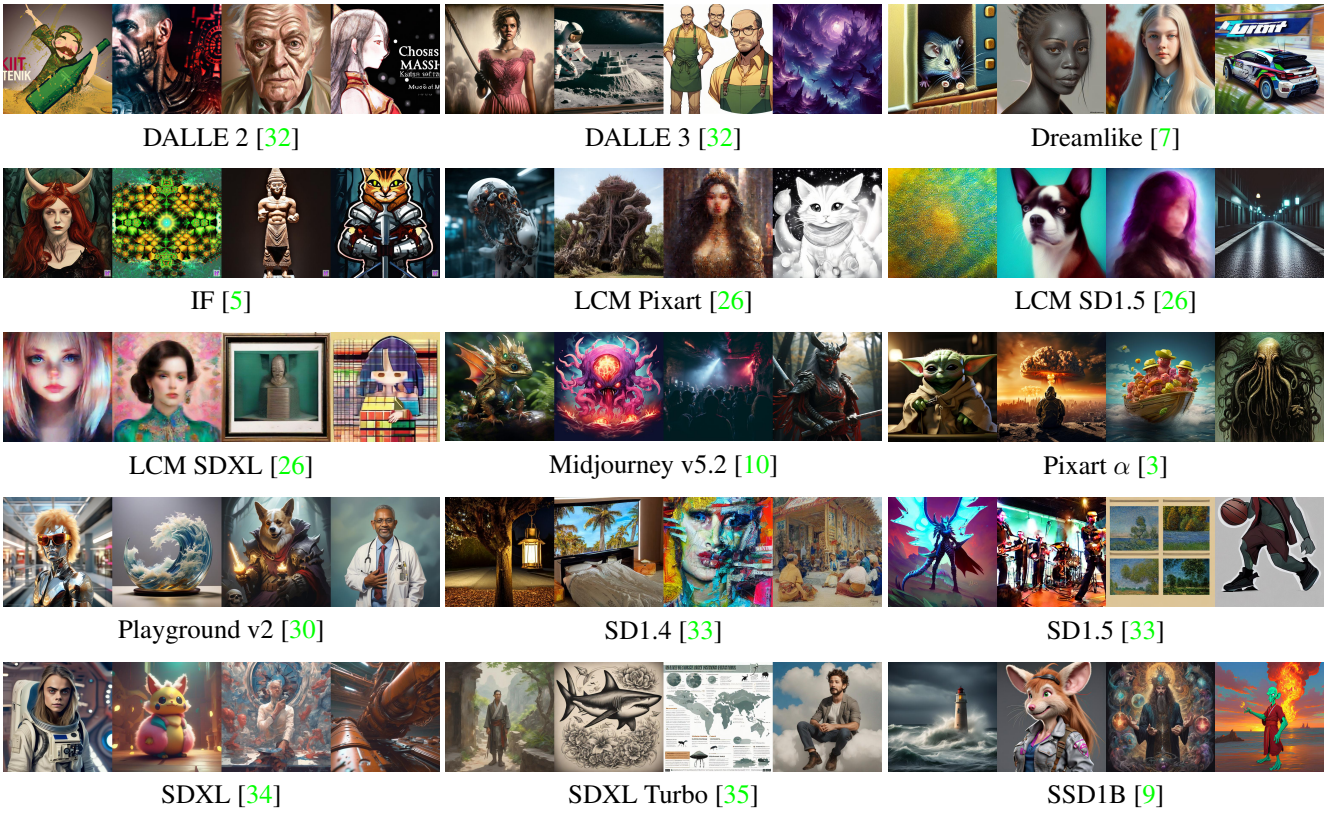


Figure 2. Visulization result of 15 Text-to-Image models in AIGIQA-20K.

- 1,000 images: LCM Pixart, LCM SD1.5, LCM SDXL, SDXL Turbo [26, 35]. At the cost of a fixed CFG, these models use acceleration mechanisms that significantly reduce iteration times. Thus, we only change iterations/resolutions with default iterations as 4.
- 500 images: DALLE2, DALLE3, IF, Midjourney v5.2 [5, 10, 32]. Adjusting the hyper-parameters drastically reduces the quality of their output. Due to overly complex model structures or closed sources, we set all three parameters to their own default values.

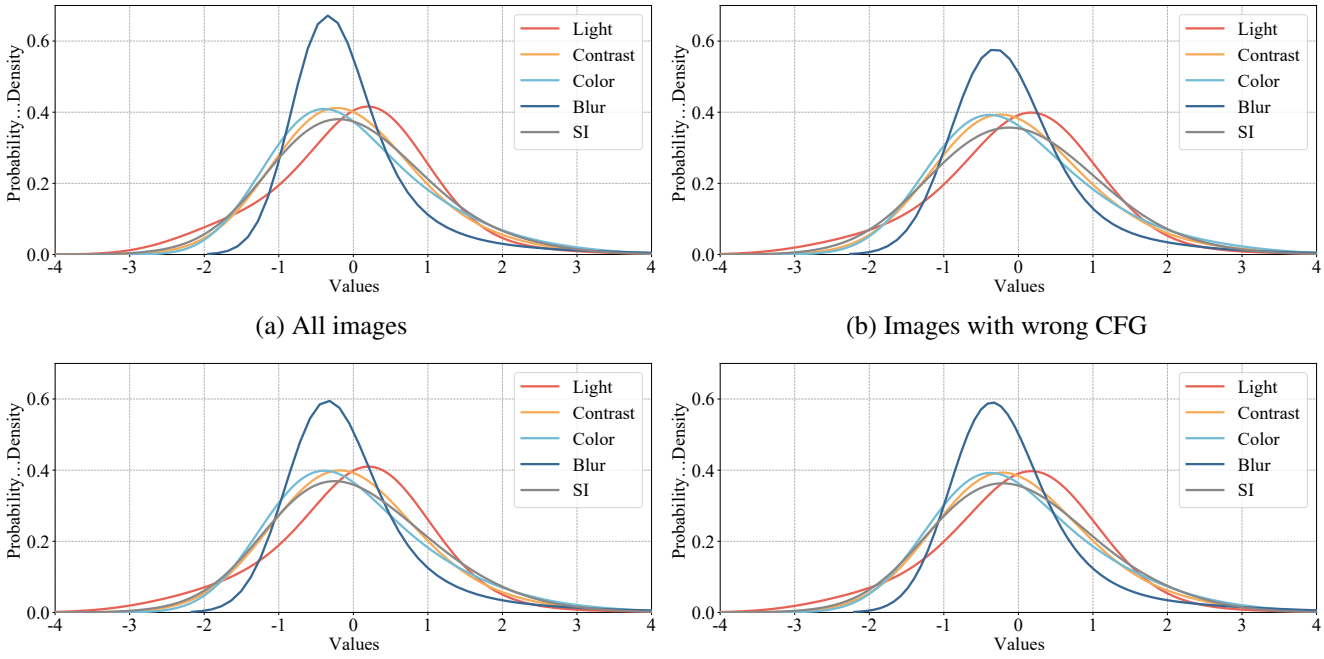
where we generate 20,000 images with different configurations while ensuring 500 results from each model according to the default configuration. Therefore, the database can be used for both IQA tasks and horizontal comparisons of output quality between different models.

2.3. Prompts Selection

For AIGI quality databases, prompts are typically from real input or manually designed. Here, AIGIQA-20K relies on the real input of AIGC community users. Firstly, as a large-scale database, AIGIQA-20K has extensively covered inputs of different lengths/themes/styles, eliminating the need to design manual prompts like previous small databases to ensure the diversity of input content. Secondly, using real user input is more in line with the real usage scenarios of AIGC, and the quality score obtained is also more reasonable. Therefore, we selected 30,000 prompt words from DiffusionDB as the original input. Considering the presence of some junk data in the above prompts, we adopt the following filtering mechanism: (1) Similarity comparison: prompts with 90% consistent content will be merged; (2) Character detection: Remove consecutive spaces, parentheses, punctuation, and non-UTF-8 encoded characters; (3) NSFW avoidance: First, delete prompts containing sensitive words, and then use GPT-4 [29] to delete NSFW prompts in semantic level. From this, we filtered out 20,000 prompts as inputs for T2I models.

2.4. Feature Analysis

After generating images from prompts and hyper-parameters above, to evaluate the impact of different configurations on the images, we calculated the distribution of five quality-related attributes in Figure 3 for the entire database, adjusted subsets of iteration, cfg, and resolution. The quality-related attributes include light, contrast, color, blur, and Spatial Information (SI, representing the content diversity of the image). Detailed explanations of



(a) All images

(b) Images with wrong CFG

(c) Images with limited iterations

(d) Images with non-square resolution

Figure 3. Distribution of quality attribute over the AIGIQA-20K database three sub-datasets with abnormal hyper-parameters.

these attributes can be found in work [11]. As previous works [21, 24] state, AIGIs have more extreme blur distribution than NSIs, while the other four attributes are distributed more evenly. In addition, we also found that for each subset of hyper-parameter anomalies, the maximum probability values never exceed 0.6; however, for the entire AIGIQA-20K, its distribution curve is sharper. This difference indicates a significant gap between the attributes of default and abnormal hyper-parameter subsets, which indirectly demonstrates the strong correlation between hyper-parameters and image quality while demonstrating the necessity of such adjustments.

3. Subjective Experiment

3.1. Experimental Procedures

Compliant with the ITU-R BT.500-13 [37] standard, we invited 21 subjects (12 male, 9 female) in this subjective experiment with normal lighting levels. AIGIs are presented on the iMac display together with the prompt in random order on the screen, with a resolution of up to 4096 $\times$ 2304. Both prompt and image are accessible for subjective, with a translation of their mother tongue like Figure 4. Considering the average between perceptual quality and T2I alignment, subjects were asked to give an overall score within the range of [0, 5], where each one-point interval stands for poor, bad, fair, good, or excellent quality.

Figure 4. User interface for subjective quality assessment. The image is given together with its correlated prompt. The score is an overall consideration of perceptual quality and alignment.

3.2. Data Processing

In case of visual fatigue, we split the database into $g \in [0, 39]$ groups including $M = 500$ images each, while limiting the experiment time to half an hour. After collecting 21$\times$20,000=42,000 quality ratings, we compute the Spearman Rank-order Correlation Coefficient (SRoCC) between

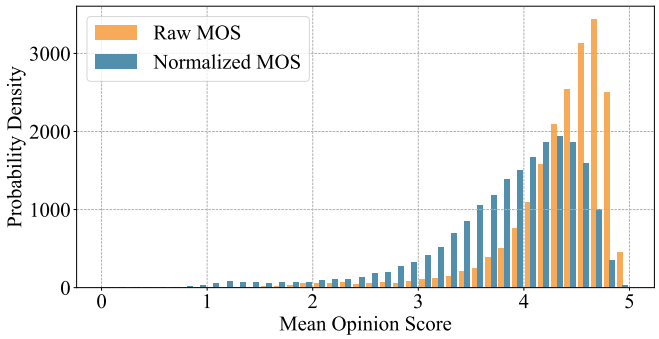


Figure 5. Distribution of raw and logarithmic normalized MOSs, where the logarithmic function unifies the entire distribution.

them and the global average and remove the outliers with SRoCC lower than 0.6. Then we normalize the average score s for between each session to avoid inter-session scoring differences as:

$$s_{ij}(g) = r_{ij}(g) - \frac{1}{M}\sum_{i=0}^{g\cdot M-1} r_{ij} + 2.5, \tag{1}$$

where (i, j) represent the index of the image and viewer and r stands for raw score. Then subjective scores are converted to Z-scores z_{ij} by:

$$z_{ij} = \frac{s_{ij} - \mu_j}{\sigma_j}, \tag{2}$$

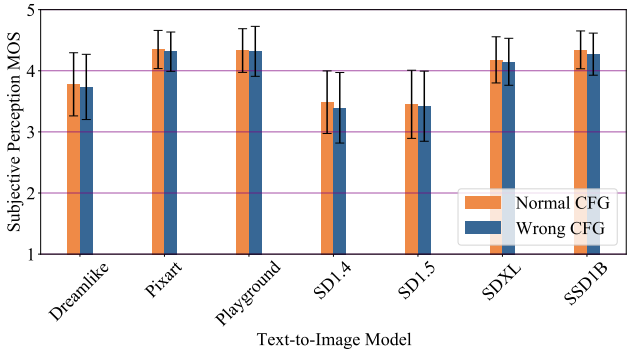
where $\mu_j = \frac{1}{N}\sum_{i=0}^{N-1} s_{ij}$, $\sigma_j = \sqrt{\frac{1}{N-1}\sum_{i=0}^{N-1}(s_{ij} - \mu_i)^2}$ and $N = 40$ is the number of subjects. Finally, the MOS of image j is computed with the following formula:

$$\begin{cases} MOS_i = \log(\frac{1}{N}\sum_{j=0}^{N-1}(z_{ij}) + 1) \\ MOS = 5 \cdot \mathrm{norm}(MOS), \end{cases} \tag{3}$$

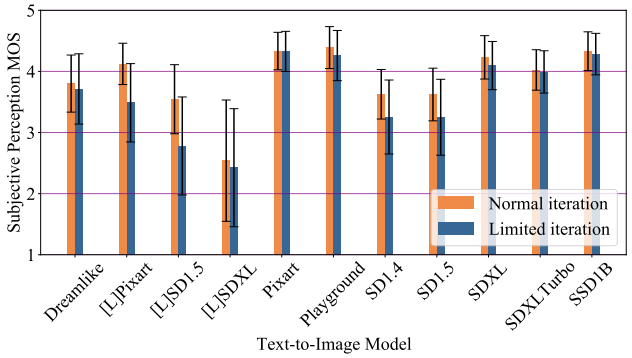
where $\mathrm{norm}(\cdot)$ indicates 0-1 normalization is a traditional data-processing technique, but logarithmic function $\log(\cdot)$ is specially designed for AIGI. As shown in Figure 5, the raw MOS data shows severe right deviation, almost all of which are concentrated in the 4 to 5 interval. However, after logarithmic processing, the distribution of MOS becomes more uniform. This highly differentiated score is more suitable for IQA tasks.

3.3. Data Analysis

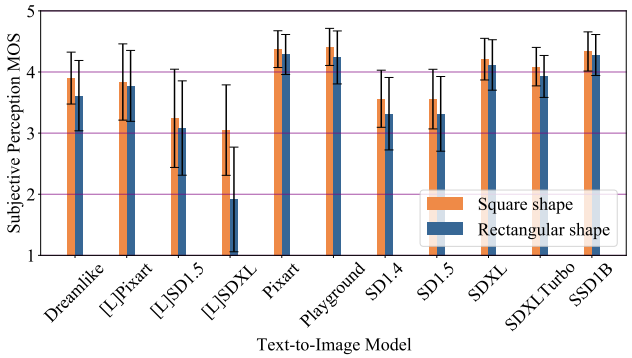
With the explosion of T2I models, their generative quality has become an unresolved issue. Compared to objective indicators, subjective evaluation indicators can better reflect real human preferences. Based on the large-scale and fine-grained subjective quality ratings in the AIGIQA-20K database, we conducted an in-depth analysis of this issue and summarized the influencing factors of AIGI subjective quality as follows:



(a) Images with default and wrong CFG



(b) Images with default and limited iterations



(c) Images with default (square) and rectangular resolutions

Figure 6. Subjective quality score of images with default and abnormal hyper-parameters. For all T2I models with abnormal hyper-parameters, the subjective quality decreases to a certain extent compared to default. ({L} for LCM)

- T2I model: Generative models themselves are the primary determinant of AIGI quality. Under the same input prompt, the generation quality of different models varies greatly. The lower limit of the latest models even outperforms the upper limit of old models.
- Prompt: The prompt has a certain impact on the quality of AIGI. Different models apply their own text encoders, some are good at generating short prompts,

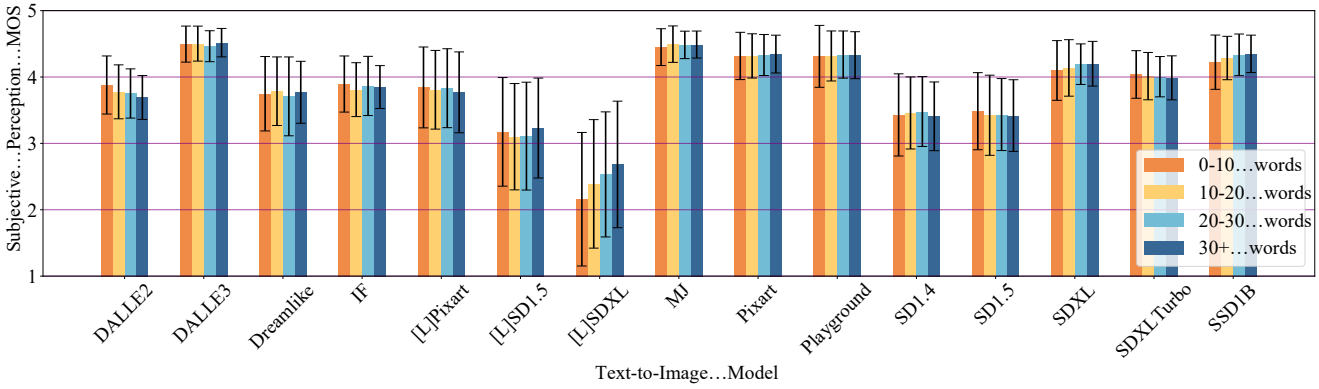


Figure 7. Subjective quality score of images with different T2I models and prompt length. ({L} for LCM)

while others are suitable for long prompts.

- Hyper-parameters: The internal parameters of a model can profoundly affect the quality of AIGI. As Sec 2.1 listed, CFG, number of iterations, and resolution can all cause AIGI quality fluctuations.

For the influence of hyperparameters, we divided the entire AIGIQA-20K database into two parts based on CFG, iteration, and resolution, and compared their subjective quality distribution under normal (default value) and abnormal (configuration in Sec 2.1) states, as shown in Figure 6. Firstly, for CFG, there has been a slight decrease in quality after adjustment compared to the default value. This is because CFG reflects the trade-off between perceived quality and fit. The larger the value, the more the model values alignment, and vice versa, the more emphasis is placed on perception. However, regardless of which side reduces and which side increases, the quality score obtained by considering both factors will inevitably decrease. This demonstrates the validity of the default CFG values in each T2I model and it is not recommended to adjust them arbitrarily. Secondly, for the number of iterations, the quality with limited iterations has decreased to varying degrees against full iterations. This indicates when iterations are insufficient, the AIGIs may lack certain details, leading to a decrease in quality. Compared to various models, the most advanced Pixart, Playground, SDXL, SDXL Turbo, and SSD1B [3, 9, 30, 34, 35] have the strongest robustness to this descent. For models with the LCM acceleration mechanism, the number of iterations is already as low as 4, and further reducing iterations will cause significant quality damage. Thirdly, for shapes, the quality of generating irregular shapes is also lower than that of squares. Since the target outputs during model training are all squares, it is expected that generating rectangular images is not ideal. Horizontally comparing others, LCM Pixart, Pixart, and SSD1B [3, 9, 26] have the strongest robustness to such descent. Overall, the newer the model, the better its support for non-square outputs.

For the T2I models themselves and prompts, Figure 7 lists the subjective quality of all 15 models at different prompt lengths. All hyper-parameters are set by default for a fair comparison. Firstly, by comparing the various models at the level of human preferences, the most advanced models currently available are DALLE3, Midjourney, Pixart, and Playground [3, 10, 30, 32]; and all other models have certain quality defects. For the acceleration mechanism, SDXL Turbo [35] is the most successful, as it reduces the iterations by 10 times at a quality cost below 0.2; In contrast, after 10 times acceleration, the output quality of the LCM [26] model is far inferior to the original, especially the acceleration effect on SDXL is extremely poor. Secondly, except for LCM SDXL, the T2I model has slightly better quality in generating short prompts than long texts. This comes from the limitation of the number of embedding tokens in the model text encoder. For example, CLIP only supports 77 tokens in absolute terms; Even worse, existing research [46] indicates that it gradually fails even from the 20th token onwards. Therefore, the defect in alignment resulted in a decrease in overall scores. In summary, to improve the output quality of the T2I model. Users should set appropriate hyper-parameters while developers need to design more powerful models and enhance their support for long text encoding and multiple resolutions.

4. Experiment

4.1. Experiment Settings

We first randomly split the AIGIQA-20K into training/validation/test sets according to the ratio of 7:1:2, with 14,000/2,000/4,000 AIGIs respectively. To benchmark the performance of quality metrics, three global indicators, including SRoCC, Kendall Rank-order Correlation Coefficient (KRoCC), and Pearson Linear Correlation Coefficient (PLCC) are applied to evaluate the consistency between the objective quality score and the subjective MOS, among

which the SRoCC and KRoCC represent the prediction monotonicity while the PLCC measures the prediction accuracy. To map the objective predicted scores to subjective MOSs, a standard five-parameter logistic function is applied as follows:

$$\hat{X} = \alpha_1 \left(0.5 - \frac{1}{1 + e^{\alpha_2 (X - \alpha_3)}}\right) + \alpha_4 X + \alpha_5, \quad (4)$$

where $\alpha_{1\sim5}$ represent the parameters for fitting, X and $\hat{X}$ stand for predicted and fitted scores respectively.

4.2. Benchmark Models

We apply 16 mainstream AIGI quality benchmarks for comparison, including both perception and alignment metrics. For perception, 12 IQA metrics are selected in the experiment. Including brisque [27], clipiqa [38], cnniqa [13], dbcnn [47], hypetiqa [36], liqe [49], musiq [14], niqe [28], qalign [42], topiq [2], unique [48], and wadiqam [1]. For the lower-better indexes (brisque, niqe), their score is reversed. These indicators mainly focus on the image itself, in the absence of the absence of text prompts. For alignment, we take 4 advanced metrics for AIGI quality. The clip [31] mainly considers T2I alignment between AIGIs and prompts while the hps [44], imagereward [45], and picscore [16] also take the perceptual quality as an auxiliary indicator. Most of them are validated as zero-shot models while cnniqa, clipiqa, and dbcnn [13, 38, 47] are trained/validated on the target set (repeating 10 times with the average result as final performance), and all the experimental results are from the testing test. The Adam optimizer [15] (with an initial learning rate of 0.00001 and batch size 128) is used for 100-epochs finetune training on an NVIDIA RTX A6000 GPU.

4.3. Performance Discussion

Table 2 shows the performance of different IQA methods across the entire AIGIQA-20K database. The most powerful quality metric currently based on multimodal large language models, q-align, is used as the baseline. We use three classic quality metrics based on deep learning, finetuning them on the training set before testing, and comparing them with the baseline. Experimental data shows that fine-tuning is of great significance for AIGC quality assessment. The performance of the three methods after training has significantly improved, with SRoCC and PLCC improving by 0.4 and KRoCC at 0.3. The performance of clipiqa and dbcnn after training has exceeded the baseline. This is because most IQA models are designed for NSIs, and when they are migrated to AIGIs, it is necessary to update the internal parameters of the model to ensure good performance. Although the fine-tuning effect is good, for promoting the application of IQA in the whole AIGC community, this indicator cannot be completely dependent on fine-tuning training. A powerful zero-shot model like qalign needs further development in the future.

Table 2. Performance results on the AIGIQA-20K database using zero-shot or finetuned metrics. The zero-shot qalign [42] is set as the baseline. [Key: **Best**, **Second Best**]

Metric	SRoCC	KRoCC	PLCC
qalign [42]	0.7461	0.5511	**0.7416**
clipiqa [38]	0.3311	0.2257	0.4829
clipiqa+finetune	**0.7863**	**0.5828**	0.7117
cnniqa [13]	0.3299	0.2244	0.3666
cnniqa+finetune	0.5968	0.4183	0.5913
dbcnn [47]	0.4710	0.3244	0.5120
dbcnn+finetune	**0.8506**	**0.6617**	**0.8688**

Table 3 further lists the performance of 12 perceived qualities and 4 fit models on the sub-database of AIGIQA-20K. According to CFG, iterations are divided into normal (default) and abnormal (adjusted) resolutions. Overall, qalign remains the most accurate indicator of AIGC quality, despite ignoring the consistency of information between images and text. The three correlation indicators rank first in all sub-databases and lead the second by about 0.05. Among the other methods, picscore, imagereward, and hps, which take the T2I alignment into account, show a leading gap. Except for hps with certain defects in PLCC, all other models have acceptable performance and can be preliminarily used to predict the quality of AIGI. The zero-shot performance of other models is not ideal, and they must undergo fine-tuning similar to Table 2, which limits their universality. Vertically comparing various sub-databases, we found that all models had more accurate AIGI evaluation results for the default CFG, but they performed better on abnormal data in terms of iteration times and resolution. Based on the error bar analysis in Figure 6, for the vast majority of T2I generative models, the limited iterations and compared to all iterations have a wider range of quality distribution compared to square resolution and rectangular resolution. Under larger quality differences, the accuracy of the evaluation will also further improve. As for CFG, there is no significant difference in the range of error bars. At this point, the more CFG deviates from the normal value, the more unnatural the generated results (such as AI artifacts, multi-finger content, etc.). Considering such distortion doesn't exist in NSIs, the zero-shot model alignment is not sensitive. Therefore, the more abnormal the CFG, the worse the performance of the evaluation.

5. Conclusion

In this paper, we establish the largest AIGI fine-grain quality database to date, AIGIQA-20K. We first se-

Table 3. Performance results on different AIGIQA-20K sub-database using zero-shot perceptual quality or alignment metrics. The data is split by default/abnormal CFG, iteration, and resolution. [Key: **Best**, **Second Best**]

Group		Default CFG			Default iteration			Default resolution		
Type	**Metric**	SRoCC	KRoCC	PLCC	SRoCC	KRoCC	PLCC	SRoCC	KRoCC	PLCC
Perce -ption	brisque [27]	0.2755	0.1874	0.2933	0.2259	0.1526	0.2772	0.1883	0.1261	0.1900
	clipiqa [38]	0.3889	0.2677	0.5375	0.2522	0.1719	0.4312	0.2964	0.2015	0.4041
	cnniqa [13]	0.3289	0.2238	0.3691	0.3350	0.2274	0.3605	0.3102	0.2088	0.2889
	dbcnn [47]	0.5051	0.3489	0.5304	0.4492	0.3085	0.4673	0.4380	0.2976	0.4307
	hyperiqa [36]	0.4390	0.2990	0.4928	0.3528	0.2393	0.4073	0.3730	0.2526	0.4096
	liqe [49]	0.4925	0.3391	0.5554	0.3675	0.2513	0.4445	0.4030	0.2746	0.4441
	musiq [14]	0.5111	0.3546	0.5848	0.3833	0.2616	0.4418	0.4146	0.2837	0.4889
	niqe [28]	0.1900	0.1266	0.3120	0.1999	0.1348	0.2977	0.0769	0.0516	0.1737
	qalign [42]	**0.7721**	**0.5764**	**0.7629**	**0.7145**	**0.5206**	**0.6813**	**0.7333**	**0.5383**	**0.7178**
	topiq [2]	0.5064	0.3491	0.5292	0.4374	0.2998	0.4487	0.4706	0.3228	0.4663
	unique [48]	0.3038	0.2041	0.3843	0.1595	0.1075	0.2127	0.2245	0.1510	0.2974
	wadiqam [1]	0.2821	0.1905	0.2855	0.2847	0.1916	0.2907	0.2516	0.1690	0.2351
Align -ment	clip [31]	0.4701	0.3656	0.5341	0.3804	0.2969	0.4733	0.3309	0.2580	0.2673
	hps [44]	0.6749	0.4899	0.6111	0.6288	0.4514	0.5052	0.6214	0.4434	0.4865
	imagereward [45]	0.6597	0.4767	0.7162	0.6150	0.4387	**0.6691**	0.6098	0.4316	**0.6579**
	picscore [16]	**0.7009**	**0.5093**	**0.7201**	**0.6474**	**0.4642**	0.6638	**0.6458**	**0.4617**	0.6474

Group		Wrong CFG			Limited iteration			Rectangular resolution		
Type	**Metric**	SRoCC	KRoCC	PLCC	SRoCC	KRoCC	PLCC	SRoCC	KRoCC	PLCC
Perce -ption	brisque [27]	0.1853	0.1239	0.2040	0.2447	0.1661	0.2394	0.2968	0.2028	0.3326
	clipiqa [38]	0.2134	0.1427	0.3343	0.3961	0.2688	0.4996	0.3500	0.2379	0.5254
	cnniqa [13]	0.3397	0.2309	0.3910	0.3429	0.2338	0.3806	0.3178	0.2168	0.3857
	dbcnn [47]	0.4034	0.2773	0.4763	0.4990	0.3434	0.5424	0.4945	0.3442	0.5601
	hyperiqa [36]	0.3352	0.2295	0.4268	0.4510	0.3069	0.5071	0.4318	0.2946	0.4984
	liqe [49]	0.3431	0.2333	0.4029	0.5061	0.3455	0.5395	0.4966	0.3417	0.5664
	musiq [14]	0.3673	0.2504	0.4515	0.5415	0.3751	0.6082	0.5144	0.3575	0.5968
	niqe [28]	0.0033	0.0020	0.0854	0.0326	0.0166	0.1767	0.1861	0.1220	0.3004
	qalign [42]	**0.6914**	**0.4986**	**0.6841**	**0.7767**	**0.5768**	**0.7708**	**0.7526**	**0.5556**	**0.7486**
	topiq [2]	0.4265	0.2950	0.4757	0.5191	0.3580	0.5479	0.4847	0.3356	0.5435
	unique [48]	0.1389	0.0940	0.2346	0.3198	0.2134	0.3987	0.2944	0.1980	0.3991
	wadiqam [1]	0.2819	0.1891	0.3346	0.2983	0.2019	0.3166	0.2876	0.1953	0.3237
Align -ment	clip [31]	0.3267	0.2531	0.3165	0.4735	0.3671	0.5155	0.4991	0.3851	0.5855
	hps [44]	**0.6388**	**0.4583**	0.4697	0.6971	0.5072	0.6370	0.6983	0.5099	0.5991
	imagereward [45]	0.6247	0.4452	**0.6421**	0.6755	0.4891	0.7098	0.6722	0.4896	0.7039
	picscore [16]	0.6372	0.4566	0.6368	**0.7090**	**0.5154**	**0.7124**	**0.7019**	**0.5109**	**0.7116**

lect 15 mainstream T2I generation models and made dynamic adjustments on CFG, iteration, and resolution hyperparameters for the first time. From this, 20,000 AIGIs are generated with different qualities to characterize the common images in today's AIGC community. Then, subjective quality labels are processed as the golden truth of quality. Finally, benchmark experiments are conducted to verify the performance of the current AIGI quality evaluator, including IQA and T2I alignment methods. Experimental results indicate that the universal zero-shot quality model is not yet complete and requires further development based on comprehensive subjective labels in this database.

Acknowledgment

The work was supported in part by the National Natural Science Foundation of China under Grant 62371283, 62271312, 62301310; in part by the Shanghai Pujiang Program under Grant 22PJ1406800; in part by the China Postdoctoral Science Foundation under Grant 2023TQ0212; and in part by the Open Project of the Key Laboratory of Media Audio & Video (Communication University of China), Ministry of Education.

References

[1] 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. 7, 8

[2] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment, 2023. 7, 8

[3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. 2310.00426, 2023. 2, 3, 6

[4] Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Exploring the naturalness of ai-generated images, 2024. 2

[5] DeepFloyd. If-i-xl-v1.0. https://www.deepfloyd.ai, 2023. 2, 3

[6] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 2

[7] Dreamlike-art. dreamlike-photoreal-2.0. https://dreamlike.art, 2023. 2, 3

[8] Yixuan Gao, Yuqin Cao, Tengchuan Kou, Wei Sun, Yunlong Dong, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Vdpve: Vqa dataset for perceptual video enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1474–1483, 2023. 2

[9] Yatharth Gupta, Vishnu V. Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss, 2024. 2, 3, 6

[10] David Holz. Midjourney. https://www.midjourney.com, 2023. 3, 6

[11] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *2017 Ninth international conference on quality of multimedia experience*, pages 1–6. IEEE, 2017. 4

[12] Xinhui Huang, Chunyi Li, Abdelhak Bentaleb, Roger Zimmermann, and Guangtao Zhai. Xgc-vqa: A unified video quality assessment model for user, professionally, and occupationally-generated content. In *IEEE International Conference on Multimedia and Expo Workshops*, 2023. 1

[13] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014. 7, 8

[14] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 7, 8

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. 7

[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, 7, 8

[17] Tengchuan Kou, Xiaohong Liu, Wei Sun, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Stablevqa: A deep no-reference quality assessment model for video stability. In *31st ACM International Conference on Multimedia*, pages 1066–1076, 2023. 2

[18] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dataset and metric for text-to-video quality assessment, 2024. 2

[19] Chunyi Li, Haoyang Li, Ning Yang, and Dazhi He. A pbch reception algorithm in 5g broadcasting. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2022. 1

[20] Chunyi Li, May Lim, Abdelhak Bentaleb, and Roger Zimmermann. A real-time blind quality-of-experience assessment metric for http adaptive streaming. In *IEEE International Conference on Multimedia and Expo*, 2023. 1

[21] Chunyi Li, Guo Lu, Donghui Feng, Haoning Wu, Zicheng Zhang, Xiaohong Liu, Guangtao Zhai, Weisi Lin, and Wenjun Zhang. Misc: Ultra-low bitrate image semantic compression driven by large multimodal model, 2024. 4

[22] Chunyi Li, Haoning Wu, Zicheng Zhang, Hongkun Hao, Kaiwei Zhang, Lei Bai, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. Q-refine: A perceptual quality refiner for ai-generated image, 2024. 1

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

[24] 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, 4

[25] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte, et al. NTIRE 2024 quality assessment of AI-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[26] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module, 2023. 3, 6

[27] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 7, 8

[28] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7, 8

[29] OpenAI. Gpt-4 technical report, 2023. 3

[30] PlaygroundAI. playground-v2-1024px-aesthetic. https://playground.com, 2023. 2, 3, 6

[31] 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. 7, 8

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 3, 6

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[34] Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models, 2022. 3, 6

[35] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation, 2023. 3, 6

[36] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 7, 8

[37] I. T. Union. Methodology for the subjective assessment of the quality of television pictures. *ITU-R Recommendation BT. 500-11*, 2002. 4

[38] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563, 2023. 7, 8

[39] 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

[40] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision, 2023. 1

[41] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models, 2023. 1

[42] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels, 2023. 7, 8

[43] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, Xiaohong Liu, Guangtao Zhai, Shiqi Wang, and Weisi Lin. Towards open-ended visual quality comparison, 2024. 1

[44] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 2, 7, 8

[45] 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, 7, 8

[46] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip, 2024. 6

[47] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 7, 8

[48] 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. 7, 8

[49] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 7, 8

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

[51] Zicheng Zhang, Wei Sun, Haoning Wu, Yingjie Zhou, Chunyi Li, Zijian Chen, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Gms-3dqa: Projection-based grid mini-patch sampling for 3d model quality assessment, 2023. 2

[52] Zicheng Zhang, Wei Sun, Yingjie Zhou, Haoning Wu, Chunyi Li, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Advancing zero-shot digital human quality assessment through text-prompted evaluation, 2023. 2

[53] Zicheng Zhang, Haoning Wu, Zhongpeng Ji, Chunyi Li, Erli Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Fengyu Sun, Shangling Jui, et al. Q-boost: On visual quality assessment ability of low-level multi-modality foundation models, 2023. 1

[54] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Kang Fu, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A reduced-reference quality assessment metric for textured mesh digital humans. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024. 2

[55] Zicheng Zhang, Yingjie Zhou, Long Teng, Wei Sun, Chunyi Li, Xiongkuo Min, Xiao-Ping Zhang, and Guangtao Zhai. Quality-of-experience evaluation for digital twins in 6g network environments. *IEEE Transactions on Broadcasting*, 2024. 1