LMM4LMM: Benchmarking and Evaluating Large-multimodal Image Generation with LMMs

Jiarui Wang[1], Huiyu Duan[1], Yu Zhao[1], Juntong Wang[1], Guangtao Zhai[1], Xiongkuo Min[1]*,
[1]Shanghai Jiao Tong University, Shanghai, China

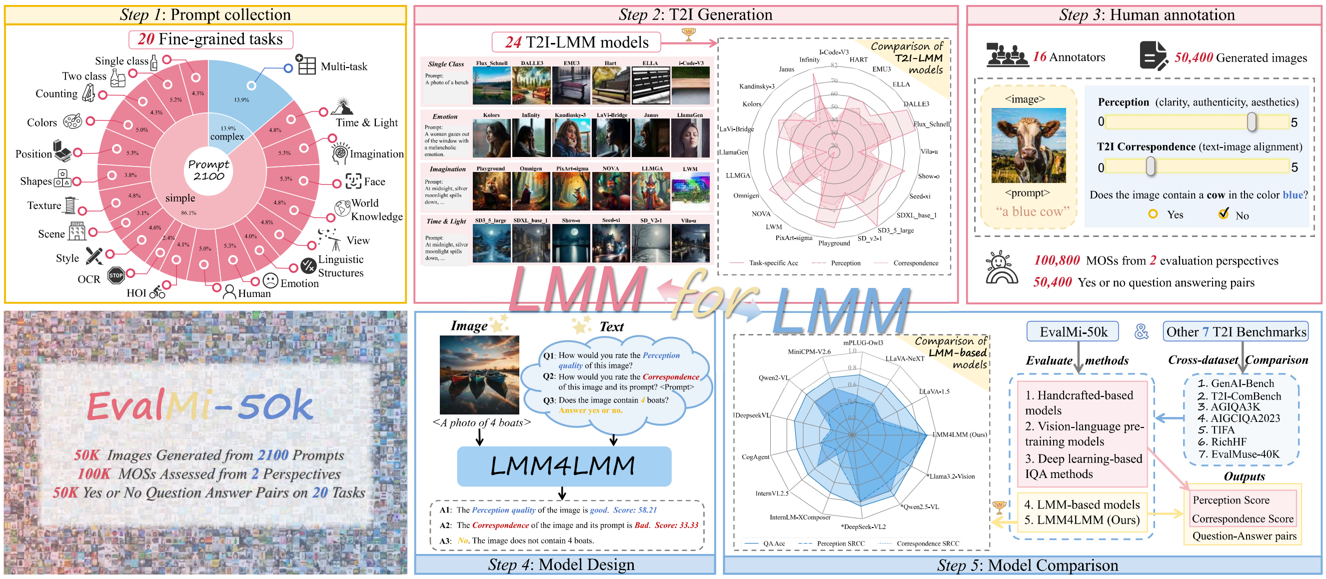

Figure 1. We present the large multimodal image generation evaluation database and model, termed EvalMi-50K and LMM4LMM, respectively. (a) We first collect 2100 comprehensive prompts across 20 fine-grained tasks. (b) Then 24 LMM-T2I models are applied to generate 50K images. (c) 100K MOSs and 50K question-answering pairs are acquired from 16 annotators. (d) We design LMM4LMM to evaluate LMM-T2I models. (e) We conduct model comparisons on EvalMi-50K and the other 7 benchmarks.

Abstract

Recent breakthroughs in large multimodal models (LMMs) have significantly advanced both text-to-image (T2I) generation and image-to-text (I2T) interpretation. However, many generated images still suffer from issues related to perceptual quality and text-image alignment. Given the high cost and inefficiency of manual evaluation, an automatic metric that aligns with human preferences is desirable. To this end, we present ***EvalMi-50K****, a comprehensive dataset and benchmark for evaluating large-multimodal image generation, which features* ***(i) comprehensive tasks****, encompassing 2,100 extensive prompts across 20 fine-grained task dimensions, and* ***(ii) large-scale human-preference annotations****, including 100K mean-opinion scores (MOSs) and 50K question-answering (QA) pairs annotated on 50,400 images generated from 24 T2I models. Based on EvalMi-50K, we propose* ***LMM4LMM****, an LMM-based metric for evaluating large multimodal T2I generation from multiple dimensions including perception, text-image correspondence, and task-specific accuracy. Extensive experimental results show that LMM4LMM achieves state-of-the-art performance on EvalMi-50K, and exhibits strong generalization ability on other AI-generated image evaluation benchmark datasets, manifesting the generality of both the EvalMi-50K dataset and LMM4LMM metric. Both EvalMi-50K and LMM4LMM will be released at https://github.com/IntMeGroup/LMM4LMM.*

*Corresponding Author

This work was supported in part by the National Natural Science Foundation of China under Grants 62271312, 62401365, 62225112, 62132006, U24A20220, and in part by the China Postdoctoral Science Foundation under Grant Number BX20250411, 2025M773473.

1. Introduction

The rapid advancement of large multimodal models (LMMs) has revolutionized the fields of both text-to-image (T2I) generation [4, 76, 77] and image-to-text (I2T) interpretation [7, 42, 43, 92], leading to high-quality AI-generated images (AIGIs) and comprehensive multimodal understanding capabilities. However, state-of-the-art T2I

Table 1. Comparision of text-to-image model evaluation benchmarks and image quality evaluation databases.

Database	MOS Granularity	Images	Annotations	Models	T2I Tasks	People per MOS	Dimensions	QA Pairs
HPD [72]	No MOS	98,807	98,807	1	3	N/A	Human Preference	✗
Pick-A-Pic [28]	No MOS	10,000	500,000	6	4	N/A	Human Preference	✗
TIFA [24]	Coarse-MOS	800	1,600	5	12	2	T2I Correspondence	✓
GenEval [13]	Coarse-MOS	1,200	6,000	6	6	5	T2I Correspondence	✓
T2I-CompBench [25]	Coarse-MOS	2,400	7,200	6	8	3	T2I Correspondence	✗
GenAIBench [34]	Coarse-MOS	9,600	40,000	6	8	3	T2I Correspondence	✓
RichHF [41]	Coarse-MOS	18,000	216,000	4	1	3	Plausibility, Alignment, Aesthetics, and Overall	✗
EvalMuse-40K [18]	Coarse-MOS	40,000	1,000,000	20	12	3-6	T2I Correspondence	✓
AGIQA-1K [90]	Fine-MOS	1,080	23,760	2	4	22	Overall	✗
AGIQA-3K [36]	Fine-MOS	2,982	125,244	6	5	21	Perception and Alignment	✗
AIGIQA-20K [37]	Fine-MOS	20,000	420,000	15	1	21	Overall	✗
AIGCIQA2023 [62]	Fine-MOS	2,400	48,000	6	10	20	Quality, Authenticity and Correspondence	✗
EvalMi-50K (Ours)	**Fine-MOS**	**50,400**	**2,419,200**	**24**	**20**	**16**	**Perception and T2I Correspondence**	✓

models may still generate images struggling with perceptual quality and text-image correspondence, thus failing to satisfy human preferences [6, 18, 41, 62, 80–83]. Since human evaluation is expensive and inefficient, it is of great significance to develop reliable evaluation metrics that align well with human perception and preference.

Traditional image quality assessment (IQA) methods [26, 51, 56, 58] generally focus on natural images with in-the-wild distortions such as noise, blur, compression [10, 49, 49, 78, 86], *etc.*, while ignoring the unique distortions in AIGIs including unrealistic structures, unnatural textures, and text-image inconsistencies [41, 62–67]. AIGI evaluation metrics such as Inception Score (IS) [15] and Fréchet Inception Distance (FID) [20] cannot evaluate the authenticity of a single image, and cannot take prompts into consideration [63]. Other common metrics such as CLIPScore [19] show less alignment with human preferences [13]. As shown in Table 1, some recent works such as AGIQA [36] and AIGCIQA2023 [62] have studied fine-grained mean opinion score (MOS) evaluation for AIGIs, however, the dataset scale or dimension scale is still relatively small. In addition, the text-image correspondence scores in these works may be affected by the perceptual quality, while they lack task-specific accuracy annotations, which are essential for benchmarking T2I models [13]. Other studies such as GenEval [13] and EvalMuse-40K [18] have T2I correspondence or task-specific accuracy annotations, but they lack consideration of the perceptual quality dimension and provide limited score annotations per image (about 3-6 per image), which may limit the model generality.

In this paper, we present **EvalMi-50K**, a large-scale dataset and benchmark towards better **eval**uation of large-**m**ultimodal **i**mage generation, which includes 50,400 images generated by 24 state-of-the-art T2I models using 2,100 diverse prompts across 20 task-specific challenges. As shown in Figure 1, we collect **2M+** human annotations from the perception, text-image correspondence, and task-specific accuracy, respectively, and finally obtain 100,800 MOSs and 50,400 question-answering (QA) pairs. Based on EvalMi-50K, we propose **LMM4LMM**, a LMM-based metric for evaluating large multimodal T2I generation from multiple dimensions including perceptual quality, text-image correspondence, and task-specific accuracy, respectively. Specifically, LMM4LMM adopts an LMM as the backbone and leverages instruction tuning [42] techniques by training the visual-language projector to give the right answers. To extract quality related and text-image aligned features and further refine these features, we apply LoRA adaptation [22] to both the vision encoder and the large language model, respectively. Through extensive experimental validation, we demonstrate that LMM4LMM achieves state-of-the-art performance on the EvalMi-50K dataset and manifests strong zero-shot generalization ability on other benchmarks. The main highlights of this work include:

- We introduce EvalMi-50K, a large-scale dataset that contains 50,400 multimodal generated images with 2M+ subjective ratings from the perception, text-image correspondence, and task-specific accuracy, respectively.
- We also use EvalMi-50K to benchmark the ability of LMMs in evaluating the generated images. EvalMi-50K can not only be used to evaluate the *generation ability* of large multimodal (LMM) T2I models, but also the *interpretation ability* of large multimodal models (LMM).
- We propose LMM4LMM, a novel LMM-based evaluation model capable of both AIGI perception quality evaluation and T2I correspondence attribution.
- Extensive experimental results on EvalMi-50K and other AIGI benchmarks manifest the state-of-the-art performance and strong generalization ability of LMM4LMM.

It should be noted that LMM4LMM also conveys the concept that we can use LMM interpretation to assess LMM image generation ability, and vice versa use LMM image generation to assess LMM interpretation ability.

2. Related Works

2.1. Benchmarks for T2I Generation

As shown in Table 1, the development of T2I generation has spawned many T2I model evaluation benchmarks and AIGI IQA databases, which can be categorized into three groups based on the presence and granularity of the human Mean Opinion Scores (MOS). No-MOS and coarse-MOS databases contain large datasets, with a limited number of annotators. Fine-MOS databases offer more reliable assessments derived from more than 15 annotators, following the guidelines of ITU-R BT.500 [55]. HPD [72] and Pick-A-Pic [28] focus on image pairs comparison, but lack precise quality assessment for each AIGI. While TIFA [24], GenAIBench [34], and T2I-CompBench [25] focus on T2I correspondence, they overlook AIGI's visual perception. While AGIQA-3K [36] and AIGCIQA2023 [62] consider both perceptual quality and T2I correspondence, they lack

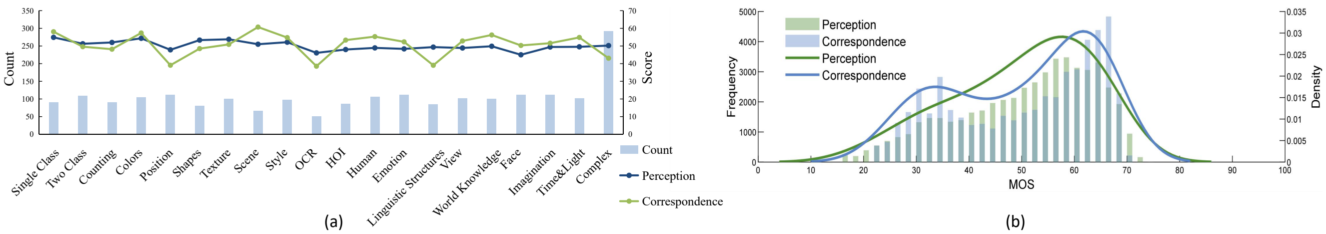

Figure 2. (a) Distribution of task counts and scores across different tasks. (b) Distribution of perception and correspondence MOSs.

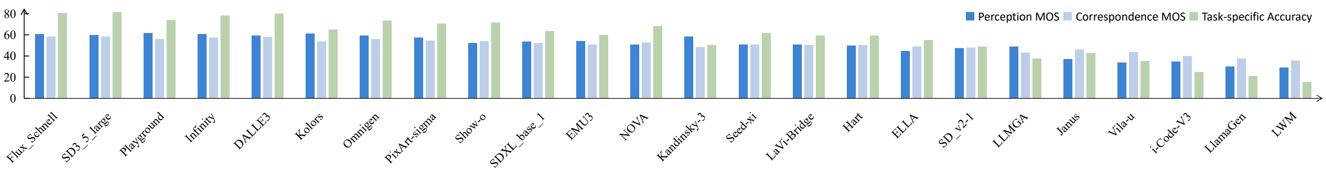

Figure 3. Comparison of T2I generation models regarding the perception MOSs, correspondence MOSs, and task-specific accuracy.

task-specific QA pairs, limiting their ability to assess T2I generation across diverse tasks. EvalMi-50K stands out by providing fine-grained MOSs across both perceptual quality and T2I correspondence, along with task-specific QA pairs.

2.2. Evaluation Metrics for T2I Generation

Many image quality assessment models have been proposed in the literature, including handcrafted IQA models (*e.g.*, NIQE [51], QAC [82], BRISQUE [50]) and deep learning-based IQA models (*e.g.*, CNNIQA [26], DBCNN [88], HyperIQA [56]). These models characterize quality-aware information to predict IQA scores but can not evaluate T2I correspondence, which is crucial for assessing the relationship between the generated image and its corresponding text prompt. CLIPScore [19], PickScore [28], and VQAScore [35] improve the evaluation of the T2I correspondence, but they struggle to assess the quality of image perception. LMMs with visual understanding capabilities perform well in QA tasks but their ability to assess image perceptual quality remains limited and often fail to give precise quality scores. HEIM [33] uses separate metrics for different evaluation perspectives. GenEval [13] and T2I-CompBench [25] employ various detection models for task-specific accuracy, but is quite complex. Our method stands out by the largest dataset with most AIGIs, annotations and latest T2I models compared to [2, 8, 29, 30], and an ***all-in-one*** manner.

3. EvalMi-50K Dataset & Benchmark

3.1. Data Collection

Our prompt design focuses on 20 different tasks as shown in Figure 1(a). The complex tasks are designed by combining simpler task components, such as color, counting, and shape, into more complex challenges. The prompts are initially crafted based on the requirements of each task and then further refined using DeepSeek R1 [16] to expand and modify them, ensuring clarity and diversity. In total, we collect 2,100 prompts, each corresponding to a specific task. To generate the AIGIs, we utilize 24 of the latest LMM-T2I models, as shown in Figure 1(b). We leverage open-source website APIs or the default weights of these models to generate images. For each prompt, each model generates a subset of images, and one of them is randomly selected from each model's output. With 2,100 distinct prompts, this process results in a total of 50,400 images (24 models × 2,100 prompts). More details of the database can be found in the *supplementary material*.

3.2. Subjective Experiment Setup and Procedure

Due to the unique distortions in AIGIs and varying elements determined by different text prompts, relying solely on an overall score for evaluation is inadequate. In this paper, we propose to evaluate AIGIs across two dimensions. (1) **Perceptual quality** focuses on visual perception, evaluating factors such as detail richness, color vibrancy, distortion levels, and authenticity. (2) **Text-image correspondence** evaluates how accurately the generated image reflects the objects, scenes, styles, and details described in the text prompt. We use a 1-5 Likert scale to score the images based on the perception and T2I correspondence. For the correspondence evaluation, in addition to the rating, annotators are instructed to answer 20 task-specific yes/no questions to determine whether the image consistently aligns with the prompt. Finally, we obtain a total of 2,419,200 human annotations including 1,612,800 reliable score ratings (16 annotators $\times$ 2 dimensions $\times$ 50,400 images), and 806,400 task-specific QA pairs (16 annotators $\times$ 50,400 images).

3.3. Subjective Data Processing

In order to obtain the MOS for an AIGI, we first convert the raw ratings into Z-scores, and then linearly scale them to the range $[0, 100]$ as follows:

$$z_{ij} = \frac{r_{ij} - \mu_j}{\sigma_i}, \quad z'_{ij} = \frac{100(z_{ij} + 3)}{6},$$

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} r_{ij}, \quad \sigma_i = \sqrt{\frac{1}{N_i - 1} \sum_{j=1}^{N_i} (r_{ij} - \mu_{ij})^2},$$

where r_{ij} is the raw rating given by the i-th subject to the j-th image. N_i is the number of images judged by subject i.

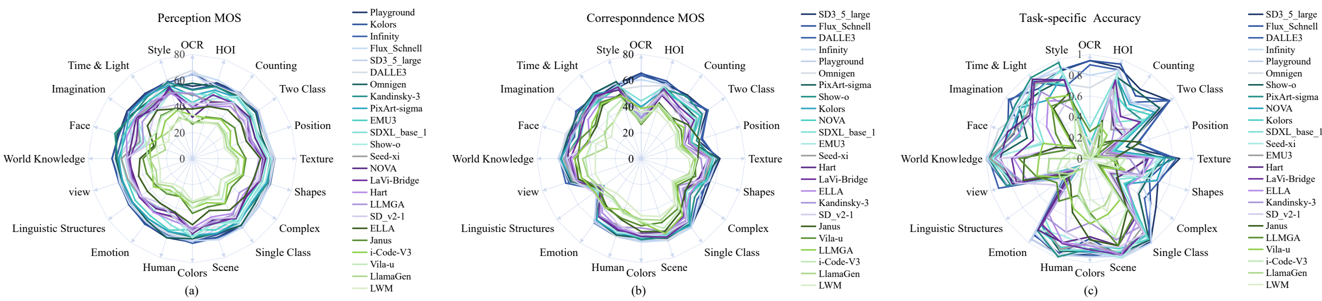

Figure 4. Comparison of MOSs and task-specific accuracy of 24 generation models across 20 tasks with descending order arranged in legend. (a) Results across perception MOSs. (b) Results across correspondence MOSs. (c) Results across task-specific accuracy.

Next, the MOS of the j-th image is computed by averaging the rescaled z-scores across all subjects as follows:

$$MOS_j = \frac{1}{M}\sum_{i=1}^{M} z'_{ij},$$

where MOS_j indicates the MOS for the j-th AIGI, M is the number of subjects, and z'_{ij} are the rescaled z-scores. The task-specific accuracy is determined by the most votes. Therefore, a total of 100,800 MOSs (2 dimensions × 50,400 images) and 50,400 question answering pairs are obtained.

3.4. Subjective Data Analysis

Figure 2(a) demonstrates the distribution of task counts and scores, highlighting the diversity and performance variations across different tasks. Figure 2(b) illustrates the distribution of MOSs for both perceptual quality and T2I correspondence. We launch comparisons of LMM-T2I generation models based on perceptual quality MOSs, T2I correspondence MOSs, and task-specific accuracy, as shown in Figure 3. Kandinsky-3 [1] excels in perceptual quality but performs poorly in correspondence, while NOVA [9] exhibits the opposite trend. This contrast highlights the necessity of evaluating the perception and correspondence as separate dimensions. We further analyze the MOSs and task-specific accuracies across different prompt categories. As shown in Figure 4(a), perception MOS is particularly sensitive to tasks such as optical character recognition (OCR) and face, as high-quality images are crucial for accurately recognizing characters and face identifications. Figure 4(b) and (c) display similar trends in correspondence evaluations, with task-specific accuracy results exhibiting sharper distinctions. While task-specific accuracy provides binary (0/1) assessments, MOS offers continuous scoring, enabling more granular evaluation of T2I correspondence. For tasks involving linguistic structures, most models perform poorly, suggesting that T2I models struggle to understand words such as "without" or "no". Additionally, models show weak performance in tasks requiring position understanding, indicating that these models may not fully grasp spatial relationships or the positioning of objects within the scene.

4. The LMM4LMM Approach

In this section, we introduce our ***all-in-one*** image quality assessment method, **LMM4LMM**, towards giving text-defined quality levels, predicting perception and T2I correspondence scores, and providing visual question answers for its correspondence assessments, depicting quality attributes from 20 task-specific challenges using one model.

4.1. Model Structure

Visual Encoding. As shown in Figure 5, the visual encoding part includes an image encoder for feature extraction and a projector for feature alignment between the image features and the input of the large language model (LLM). To enhance scalability for processing high-resolution images, we employ a pixel unshuffle operation, which reduces the number of visual tokens to one-quarter of the original size. Specifically, for an input AIGI I, we first resize it to 1024×1024 and then divide images into tiles of 448×448 pixels based on the aspect ratio and resolution of the input images. The image encoder E_I is built on a pre-trained vision transformer (ViT), *i.e.*, InternViT [7], which is pre-trained on the LAION-en dataset [54] using text-image contrastive learning. To align the extracted features with the input space of the LLM, a projector P_I with two multilayer perceptron (MLP) layers is applied. The process can be formulated as:

$$T_i = P_I(E_I(I)), \tag{1}$$

where T_i is the mapped image feature tokens.

Feature Fusion and Quality Regression. We utilize the LMM (InternVL2.5-8B [7]) to integrate the visual tokens and text instruction tokens to perform the following two tasks. (1) Quality level descriptions: the model generates a descriptive quality level evaluation of the input image, such as "*The perception quality of the image is (bad, poor, fair, good, excellent).*" Since LLMs have a better understanding of textual data than numerical data, this initial categorization provides a preliminary classification of the image's quality, which is valuable for guiding subsequent quality regression tasks. (2) Regression score output: the model takes the quality representations from the last hidden states of the

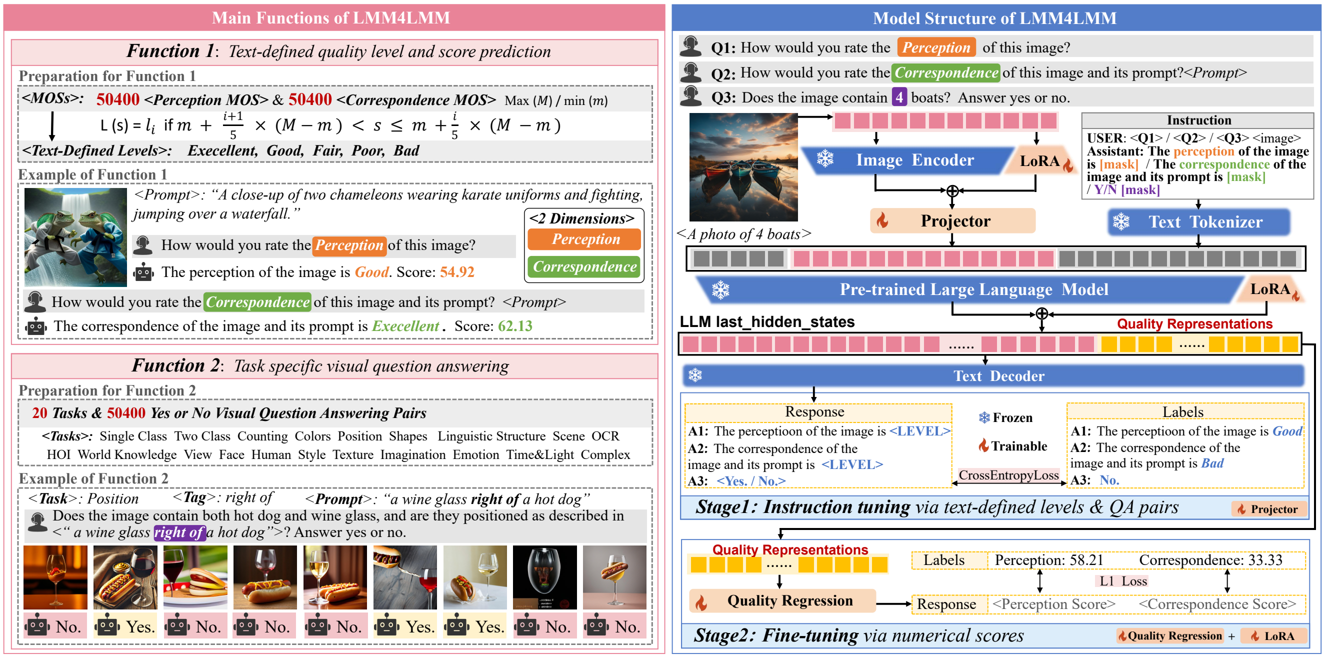

Figure 5. Overview of the LMM4LMM architecture. The model includes two functions: (1) text-defined quality level and score prediction, (2) task-specific visual question answering. The training process consists of two stages: instruction tuning of the model via text-defined levels, and then fine-tuning the vision encoder and LLM via numerical scores. The model incorporates an image encoder and a text encoder for extracting visual and textual features, which are fed into a pre-trained LLM to generate results. LoRA [22] weights are introduced to the pre-trained image encoder and the LLM to adapt the models to perception quality evaluation and T2I correspondence attribution tasks.

LLM to perform a regression task through a quality regression module, outputting numerical quality scores.

4.2. Training and Fine-tuning Strategy

The training process of LMM4LMM follows a two-stage approach to address two tasks: (1) perception quality and text-image correspondence score prediction, (2) task-specific visual question answering. We first perform instruction tuning via text-defined quality levels and QA pairs. We then fine-tune the vision encoder and LLM with LoRA [22], and train the quality regression module via numerical scores to enable accurate score generation.

Instruction Tuning. Achieving an ***all-in-one*** image quality assessment model is of great significance for enabling multi-dimensional quality evaluation in a single model. Benefiting from the generalization ability of LMMs, our model verifies the effectiveness of using the instruction tuning strategy for all-in-one task-specific question answering. We train the projector to align textual and visual semantics for joint reasoning and then use language loss during the instruction tuning phase. As a result, LMM4LMM can give visual question answers across the 20 task-specific challenges using one model weight. For score prediction, since LMMs have a better understanding of textual data than numerical data, directly generating numerical scores might be challenging for LMMs. Therefore, we first convert the continuous scores into categorical text-based quality levels. Specifically, we uniformly divide the range between the highest score (M) and the lowest score (m) into five distinct intervals, assigning the scores in each interval to respective levels:

$$L(s) = l_i \text{ if } \mathrm{m}+\frac{i-1}{5}\times(\mathrm{M}-\mathrm{m}) < s \leq \mathrm{m}+\frac{i}{5}\times(\mathrm{M}-\mathrm{m}), \tag{2}$$

where $\{l_i|_{i=1}^5\} = \{bad, poor, fair, good, excellent\}$ are the standard text rating levels as defined by ITU [55]. This step provides the LMM with a more accessible way to grasp the concept of image quality by initially framing it in terms of text-defined quality levels.

Quality Regression Fine-tuning. To further improve the performance of LMM4LMM and enable it to produce more precise quality scores, we introduce a quality regression module, which takes the last-hidden-state features from the LMM as input and generates scores from both perception quality and T2I correspondence perspectives. Fine-tuning LMMs is generally resource consuming but can lead to better performance. To make the fine-tuning process more efficient, we adopt the LoRA technique [22]. The LoRA-based approach ensures that the model adapts effectively to the regression task with numerical scores to adjust the model's predictions and produce more accurate, fine-grained IQA results. During the fine-tuning stage, we employ L1 loss for the quality regression task to minimize the difference between the predicted scores and the groundtruth values.

5. Experiments

In this section, we conduct extensive experiments to evaluate the performance of our proposed model. We first present the experimental setups in detail. Then we launch experiments to evaluate the performance of our model compared to current state-of-the-art IQA and LMM-based models in predicting scores and task-specific visual question answering based on EvalMi-50K and other seven AI-generated image evaluation datasets. We launch further cross-dataset experiments to verify the generalizability of the proposed model. Finally, we conduct ablation experiments to evaluate the efficiency of our proposed components.

5.1. Experiment Setup

To evaluate the correlation between the predicted scores and the ground-truth MOSs, we utilize three evaluation criteria: Spearman Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), and Kendall's Rank Correlation Coefficient (KRCC). For visual question answering, we adopt the average accuracy as the metric. Traditional handcrafted IQA models are directly evaluated on the corresponding databases. We load the pre-trained weights for inference for vision-language pre-training and LLM-based models. We fine-tuned three of the LLM-based models using the same fine-tuning approach as our model's backbone. For deep learning-based models, we use the same training and testing split (4:1) as the previous literature. The models are implemented with PyTorch and trained on a 40GB NVIDIA RTX A6000 GPU with batch size of 8. The initial learning rate is set to 1e-5, and decreased using the cosine annealing strategy. We employ Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. During pre-training, the number of training epochs is set to 1. For fine-tuning, the number of training epochs is set to 5. All experiments for each method are averaged using 5-fold cross-validation.

5.2. Evaluation on the EvalMi-50K Database

As shown in Table 2, handcrafted IQA models such as NIQE [51] and QAC [82], show poor performance, indicating their features handcrafted mainly for natural images are ineffective for evaluating AIGIs. Vision-language pre-training models such as CLIPScore [19] and PickScore [28] perform poorly in perception quality due to their focus on T2I correspondence and overlook AIGI's visual perception. While LMM-based models are effective in handling complex visual question-answering tasks, their interpretation of image perception quality remains insufficient. Deep learning-based IQA methods achieve relatively better results, but still fall short in the T2I correspondence dimension. Table 3 and Figure 6 compare the performances of LMM-based models across the 20 task-specific challenges derived from the EvalMi-50K dataset. LMMs excel in tasks that require the interpretation of complex visual-textual interactions, such as OCR and World Knowledge, but they struggle with low-level quality features, such as texture and style, as their focus is on semantic understanding rather than perceptual quality. When fine-tuned using our proposed methods, their performance improves significantly, which verifies the effectiveness of our approach in enhancing the evaluation and interpretation capabilities of the LMMs. Our model achieves superior performance in both score prediction and visual question answering, making it a more comprehensive method for evaluating AIGIs.

Table 2. Performance comparisons of the state-of-the-art quality evaluation methods on the EvalMi-50K from perspectives of perception and T2I correspondence. ♠ Handcrafted IQA models, ♢ vision-language pre-training models, ♣ LMM-based models, ♡ deep learning-based IQA models. *Refers to finetuned models.

Dimension	Perception			Correspondence		
Methods / Metrics	**SRCC**	**PLCC**	**KRCC**	**SRCC**	**PLCC**	**KRCC**
♠ **NIQE** [51]	0.3818	0.3885	0.2589	0.2430	0.2505	0.1643
♠ **QAC** [82]	0.0376	0.0855	0.0246	0.0511	0.0680	0.0337
♠ **BRISQUE** [50]	0.0157	0.0334	0.0104	0.0467	0.0543	0.0313
♠ **BPRI** [47]	0.0329	0.0196	0.0207	0.0068	0.0022	0.0045
♠ **HOSA** [79]	0.1480	0.1690	0.0985	0.1355	0.1471	0.0905
♠ **BMPRI** [48]	0.1519	0.1245	0.1011	0.0611	0.0415	0.0410
♠ **Higrade-2** [31]	0.0393	0.0260	0.0275	0.0326	0.0224	0.0223
♢ **CLIPScore** [19]	0.2031	0.2561	0.1369	0.2607	0.3072	0.1772
♢ **BLIPScore** [40]	0.1575	0.2166	0.1060	0.2900	0.3468	0.1970
♢ **ImageReward** [80]	0.4105	0.4676	0.2815	0.4991	0.5523	0.3470
♢ **PickScore** [28]	0.5623	0.5905	0.3939	0.4611	0.4692	0.3214
♢ **HPSv2** [72]	0.6404	0.6751	0.4556	0.5336	0.5525	0.3747
♢ **VQAScore** [35]	0.3314	0.3172	0.2253	0.6062	0.6118	0.4304
♢ **FGA-BLIP2** [18]	0.5275	0.5604	0.3694	0.6755	0.6916	0.4901
♣ **LLaVA-1.5 (7B)** [43]	0.3372	0.3525	0.2577	0.3887	0.3716	0.3149
♣ **LLaVA-NeXT (8B)** [39]	0.4333	0.4164	0.3442	0.4568	0.4803	0.3535
♣ **mPLUG-Owl3 (7B)** [85]	0.3918	0.3569	0.3018	0.4744	0.5430	0.3657
♣ **MiniCPM-V2.6 (8B)** [84]	0.3733	0.1053	0.2839	0.5916	0.5971	0.4597
♣ **Qwen2-VL (7B)** [68]	0.3760	0.3625	0.3061	0.5899	0.5954	0.4658
♣ **DeepSeekVL (7B)** [74]	0.2611	0.3010	0.1988	0.2356	0.3457	0.1872
♣ **CogAgent (18B)** [21]	0.3861	0.4235	0.2927	0.3575	0.3601	0.2888
♣ **InternVL2.5 (8B)** [7]	0.2597	0.3669	0.1859	0.5511	0.5908	0.4039
♣ **InternLM-XComposer (7B)** [87]	0.3918	0.3569	0.3018	0.1728	0.1659	0.1401
♣ **DeepSeekVL2 (1B)*** [74]	0.7899	0.8253	0.6511	0.7817	0.7991	0.6457
♣ **Qwen2.5-VL (8B)*** [3]	0.6990	0.7495	0.5715	**0.8008**	**0.8219**	**0.6657**
♣ **Llma3.2-Vision (11B)*** [46]	0.7555	0.7891	0.6155	0.6403	0.6461	0.5168
♡ **CNNIQA*** [26]	0.4348	0.5583	0.3383	0.1186	0.0791	0.1067
♡ **DBCNN*** [88]	0.5525	0.6181	0.3802	0.3301	0.3515	0.2216
♡ **HyperIQA*** [56]	0.5872	0.6768	0.4335	0.5348	0.5447	0.3742
♡ **TReS*** [14]	0.3935	0.4301	0.2695	0.1406	0.1520	0.0946
♡ **MUSIQ*** [27]	0.7985	0.8379	0.6032	0.5310	0.5510	0.3789
♡ **StairIQA*** [58]	0.8268	**0.8645**	0.6346	0.5890	0.6089	0.4199
♡ **Q-Align*** [71]	**0.8311**	0.8505	**0.6383**	0.4547	0.4640	0. 3096
♡ **LIQE*** [89]	0.8106	0.8268	0.6163	0.5617	0.5777	0.4013
LMM4LMM (Ours)	**0.8863**	**0.9094**	**0.7137**	**0.8969**	**0.9162**	**0.7332**
Improvement	+5.5%	+4.49%	+7.54%	+9.61%	+9.43%	+6.75%

5.3. Evaluation on T2I Model Performance

We further conduct comparisons of the alignment between different metric results and human annotations in evaluating T2I model performance, as shown in Table 4. Our model achieves the highest SRCC with human ratings and the lowest relative Root Mean Square Error (RMSE) in score differences. This demonstrates our model's ability to accurately assess and rank the performance of T2I generative models closest to human judgment. We also provide examples with model prediction scores at the image level. As shown in Figure 7, LMM4LMM generates scores that are more consistent with human annotations and achieves the highest accuracy in question-answering, which further demonstrates its effectiveness in both image perception evaluation and task-specific T2I correspondence attribution.

Table 3. Performance comparisons of LMMs on the EvalMi-50K across different task-specific challenges. We report the correlation between automatic evaluation metrics and human groundtruth annotations in terms of perception quality SRCC (ρ_p), correspondence SRCC (ρ_c), and QA accuracy (Acc%). The best results are marked in RED and the second-best in BLUE. *Refers to finetuned models.

Dimension	Single Class			Two Class			Counting			Colors			Position			Shapes			Texture		
Methods / Metrics	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑
LLaVA-1.5 (7B) [43]	0.295	0.232	84.2	0.371	0.544	78.1	0.130	0.387	61.4	0.302	0.322	84.9	0.372	0.474	45.7	0.357	0.183	53.2	0.190	0.215	63.8
LLaVA-NeXT (8B) [39]	0.393	0.244	83.5	0.459	0.451	79.1	0.319	0.460	68.4	0.308	0.424	83.5	0.434	0.408	46.3	0.403	0.433	59.7	0.369	0.494	62.7
mPLUG-Owl3 (7B) [85]	0.468	0.238	85.1	0.430	0.526	81.1	0.433	0.581	82.3	0.449	0.360	84.3	0.426	0.498	52.4	0.414	0.320	58.0	0.400	0.289	67.6
MiniCPM-V2.6 (8B) [84]	0.423	0.350	85.1	0.311	0.697	77.9	0.361	0.711	83.9	0.353	0.547	85.5	0.445	0.607	55.6	0.452	0.533	69.2	0.301	0.545	74.6
Qwen2-VL (7B) [68]	0.425	0.163	84.4	0.200	0.673	79.1	0.314	0.697	78.2	0.357	0.441	84.1	0.224	0.552	59.7	0.265	0.500	59.0	0.234	0.405	66.1
Qwen2.5-VL(7B) [3]	0.507	0.394	18.6	0.495	0.716	44.6	0.493	0.705	53.0	0.524	0.505	21.9	0.569	0.648	74.9	0.602	0.484	51.7	0.426	0.592	46.0
Llama3.2-Vision (11B) [46]	0.265	0.275	85.1	0.197	0.185	73.7	0.402	0.284	73.6	0.213	0.316	87.9	0.233	0.154	48.9	0.252	0.301	66.6	0.257	0.359	72.3
DeepseekVL (7B) [45]	0.137	0.043	82.5	0.192	0.447	74.1	0.222	0.194	78.9	0.094	0.134	80.3	0.275	0.332	59.7	0.213	0.030	61.2	0.111	0.100	71.7
DeepseekVL2 (1B) [74]	0.140	0.032	18.6	0.157	0.051	44.6	0.035	0.028	53.0	0.070	0.046	21.9	0.048	0.049	74.9	0.136	0.035	51.7	0.078	0.038	46.0
CogAgent (18B) [21]	0.341	0.316	85.3	0.292	0.536	81.3	0.319	0.389	71.1	0.433	0.378	84.3	0.407	0.410	35.2	0.492	0.317	57.5	0.309	0.303	62.3
InternVL2.5 (8B) [7]	0.233	0.225	83.5	0.253	0.625	79.1	0.205	0.574	71.6	0.185	0.355	83.9	0.306	0.565	56.2	0.197	0.436	58.5	0.162	0.497	69.6
InternLM-XComposer (7B) [87]	0.467	0.137	85.1	0.430	0.134	81.1	0.433	0.337	82.3	0.449	0.152	84.3	0.426	0.205	52.4	0.414	0.039	58.0	0.400	0.031	67.6
*DeepseekVL2 (1B) [74]	0.772	0.658	89.1	0.784	0.825	87.8	0.766	0.820	86.9	0.774	75.1	89.5	0.795	0.604	84.8	0.799	0.604	76.9	0.738	0.701	81.5
*Qwen2.5-VL (7B) [3]	0.708	0.609	89.5	0.705	0.828	89.2	0.699	0.763	87.4	0.681	0.661	89.2	0.690	0.777	88.3	0.725	0.788	81.6	0.585	0.788	86.2
*Llama3.2-Vision (11B) [46]	0.706	0.558	84.2	0.734	0.613	72.6	0.729	0.510	66.1	0.723	0.460	80.9	0.747	0.357	77.9	0.732	0.518	70.2	0.711	0.570	70.1
LMM4LMM (Ours)	0.850	0.799	89.5	0.861	0.899	89.3	0.868	0.867	87.5	0.860	0.826	89.5	0.851	0.841	88.8	0.863	0.817	81.8	0.805	0.852	87.1

Dimension	Scene			Style			OCR			HOI			Human			Emotion			Linguistic Structure		
Methods / Metrics	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑
LLaVA-1.5 (7B) [43]	0.337	0.298	92.2	0.445	0.209	74.0	0.454	0.666	83.4	0.457	0.428	71.8	0.447	0.428	83.9	0.288	0.423	63.4	0.427	0.278	77.2
LLaVA-NeXT (8B) [39]	0.517	0.336	87.3	0.502	0.326	77.6	0.560	0.629	92.6	0.521	0.596	75.8	0.486	0.310	85.3	0.444	0.502	55.9	0.565	0.565	80.7
mPLUG-Owl3 (7B) [85]	0.448	0.013	87.3	0.209	0.202	76.7	0.361	0.669	86.0	0.460	0.498	75.3	0.460	0.443	83.2	0.333	0.500	62.9	0.499	0.555	84.6
MiniCPM-V2.6 (8B) [84]	0.369	0.126	87.3	0.475	0.467	77.6	0.525	0.727	85.6	0.403	0.545	74.6	0.450	0.478	85.7	0.313	0.515	64.1	0.436	0.657	82.5
Qwen2-VL (7B) [68]	0.398	0.297	87.7	0.525	0.439	75.6	0.511	0.720	92.1	0.293	0.567	75.1	0.250	0.574	82.6	0.417	0.574	62.3	0.485	0.693	82.5
Qwen2.5-VL (7B) [3]	0.548	0.435	91.9	0.537	0.424	74.4	0.658	0.773	90.4	0.519	0.543	77.1	0.560	0.454	80.1	0.441	0.493	51.0	0.610	0.701	84.4
Llama3.2-Vision (11B) [46]	0.355	0.047	92.2	0.409	0.342	75.1	0.178	0.258	88.8	0.277	0.086	72.3	0.196	0.162	69.2	0.237	0.128	59.0	0.322	0.442	59.0
DeepseekVL (7B) [45]	0.333	0.109	89.0	0.301	0.012	75.8	0.352	0.199	74.2	0.360	0.286	70.3	0.456	0.141	78.5	0.249	0.330	66.0	0.444	0.502	83.9
DeepseekVL2 (1B) [74]	0.140	0.031	16.9	0.157	0.006	26.3	0.070	0.067	68.1	0.048	0.015	31.4	0.136	0.029	32.5	0.078	0.057	33.7	0.257	0.013	76.5
CogAgent (18B) [21]	0.357	0.024	87.3	0.586	0.252	77.4	0.493	0.383	65.1	0.460	0.307	71.6	0.356	0.163	80.3	0.298	0.256	61.6	0.482	0.404	83.0
InternVL2.5 (8B) [7]	0.218	0.216	89.9	0.344	0.406	74.4	0.339	0.666	72.1	0.355	0.473	73.6	0.249	0.446	82.0	0.237	0.517	58.2	0.357	0.613	74.6
InternLM-XComposer (7B) [87]	0.448	0.034	87.3	0.209	0.026	76.7	0.361	0.060	86.0	0.460	0.275	75.3	0.460	0.080	83.2	0.333	0.157	62.9	0.499	0.346	84.6
*DeepseekVL2 (1B) [74]	0.763	0.591	93.2	0.730	0.677	83.8	0.855	0.825	91.2	0.800	0.667	80.3	0.846	0.719	88.4	0.790	0.670	79.8	0.812	0.375	87.2
*Qwen2.5-VL (7B) [3]	0.680	0.521	93.2	0.668	0.642	83.7	0.881	0.850	92.9	0.751	0.611	83.5	0.693	0.562	92.0	0.633	0.690	82.9	0.754	0.795	88.3
*Llama3.2-Vision (11B) [46]	0.760	0.456	92.5	0.720	0.481	79.0	0.842	0.526	80.4	0.790	0.469	74.6	0.817	0.635	84.5	0.759	0.463	77.5	0.801	0.273	73.7
LMM4LMM (Ours)	0.856	0.755	93.3	0.860	0.804	86.1	0.938	0.882	93.0	0.921	0.864	83.5	0.916	0.851	92.0	0.907	0.864	84.7	0.878	0.837	88.4

Dimension	View			World Knowledge			Face			Imagination			Time & Light			Complex			Overall		
Methods / Metrics	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑	ρ_p↑	ρ_c↑	Acc↑
LLaVA-1.5 (7B) [43]	0.389	0.255	73.6	0.309	0.450	80.1	0.401	0.375	65.6	0.395	0.504	68.4	0.396	0.413	73.1	0.336	0.471	66.3	0.337	0.389	71.0
LLaVA-NeXT (8B) [39]	0.534	0.284	70.5	0.522	0.313	84.7	0.475	0.438	64.7	0.579	0.568	69.0	0.504	0.448	70.5	0.403	0.413	59.7	0.433	0.457	70.7
mPLUG-Owl3 (7B) [85]	0.448	0.399	71.3	0.458	0.369	80.3	0.077	0.340	68.0	0.354	0.493	66.8	0.385	0.539	69.2	0.402	0.517	64.7	0.392	0.474	72.7
MiniCPM-V2.6 (8B) [84]	0.407	0.404	76.1	0.382	0.468	61.7	0.323	0.510	70.7	0.475	0.625	71.5	0.451	0.464	65.7	0.333	0.538	66.1	0.373	0.592	73.4
Qwen2-VL (7B) [68]	0.518	0.488	77.1	0.590	0.540	79.7	0.492	0.577	69.1	0.385	0.581	67.9	0.517	0.543	67.4	0.350	0.605	62.2	0.376	0.590	72.6
Qwen2.5-VL (7B) [3]	0.537	0.457	76.1	0.562	0.496	70.5	0.626	0.549	72.1	0.598	0.602	72.6	0.556	0.504	66.7	0.549	0.685	74.7	0.528	0.640	76.2
Llama3.2-Vision (11B) [46]	0.311	0.146	73.0	0.235	0.140	66.4	0.274	0.087	71.3	0.233	0.227	70.4	0.370	0.130	61.5	0.284	0.155	60.9	0.293	0.260	70.8
DeepseekVL (7B) [45]	0.357	0.228	70.1	0.355	0.233	76.5	0.293	0.177	65.2	0.311	0.502	70.1	0.400	0.252	67.8	0.205	0.325	56.8	0.261	0.236	70.7
DeepseekVL2 (1B) [74]	0.200	0.084	30.5	0.175	0.013	16.9	0.168	0.039	41.2	0.172	0.021	39.8	0.179	0.001	33.9	0.048	0.015	62.2	0.125	0.032	42.7
CogAgent (18B) [21]	0.349	0.308	70.9	0.435	0.312	79.3	0.488	0.366	68.2	0.395	0.385	62.2	0.428	0.314	67.3	0.413	0.280	48.0	0.386	0.358	67.5
InternVL2.5 (8B) [7]	0.307	0.359	70.3	0.315	0.426	78.5	0.301	0.490	66.6	0.347	0.489	71.4	0.333	0.373	56.4	0.228	0.590	60.8	0.260	0.551	70.1
InternLM-XComposer (7B) [87]	0.448	0.227	71.3	0.458	0.205	80.3	0.077	0.113	68.0	0.354	0.188	66.8	0.385	0.020	69.2	0.402	0.170	64.7	0.392	0.173	72.7
*DeepseekVL2 (1B) [74]	0.771	0.670	83.9	0.815	0.678	89.8	0.837	0.689	77.2	0.767	0.764	81.8	0.771	0.666	82.6	0.769	0.778	83.7	0.790	0.782	84.9
*Qwen2.5-VL (7B) [3]	0.712	0.782	85.1	0.734	0.713	90.4	0.760	0.648	82.2	0.594	0.736	88.0	0.707	0.718	82.7	0.668	0.811	89.2	0.699	0.801	87.2
*Llama3.2-Vision (11B) [46]	0.718	0.258	81.4	0.759	0.658	88.0	0.837	0.598	77.3	0.723	0.624	79.6	0.745	0.653	82.2	0.714	0.532	75.9	0.756	0.640	78.1
LMM4LMM (Ours)	0.870	0.814	85.0	0.885	0.814	90.5	0.949	0.906	82.4	0.886	0.882	88.5	0.878	0.829	83.0	0.877	0.901	89.2	0.886	0.895	87.9

Table 4. Comparisons of the alignment between different metric results and human annotations in evaluating T2I model performance.

Dimension	Perception Score					Correspondence Score					Question Answering Accuracy (%)					Overall Rank	
Models	Human	Ours	Q-Align	StairIQA	LIQE	Human	Ours	Q-Align	FGA	VQAScore	Human	Ours	Qwen2.5	Llama3.2	Deepseek2	Human	Ours
Flux_schnell [32]	60.63	61.51	92.69	61.45	4.38	58.10	58.48	80.08	3.50	81.79	80.29	78.64	77.23	76.53	71.36	1	1
SD3_5_large [11]	59.50	59.77	88.78	59.00	4.34	58.35	59.04	74.80	3.58	85.28	81.43	82.18	82.98	77.39	77.93	2	3
Playground [38]	61.64	62.89	96.40	62.12	4.59	56.06	57.46	85.27	3.56	82.50	73.86	74.13	73.88	78.61	70.65	3	2
Infinity [17]	60.86	61.50	95.73	61.38	4.56	57.43	58.17	84.37	3.45	81.93	78.10	77.32	77.05	78.45	73.30	4	4
DALLE3 [4]	59.35	60.27	94.72	60.02	4.40	57.97	58.38	85.87	3.52	81.82	80.24	79.15	80.05	82.29	77.80	5	5
Kolors [61]	61.14	62.29	95.70	62.44	4.78	53.53	55.07	85.30	3.24	77.09	65.05	69.00	83.18	78.73	66.29	6	6
Omnigen [76]	59.12	60.47	91.04	59.92	4.44	55.81	57.00	79.78	3.38	80.10	73.29	72.75	73.97	80.05	68.86	7	7
PixArt-sigma [5]	57.43	59.19	91.11	59.97	4.04	54.72	56.07	80.90	3.39	79.98	70.71	70.49	71.90	74.24	66.28	8	8
Show-o [77]	52.31	52.74	83.61	54.48	3.32	54.21	54.54	72.91	3.38	80.58	71.71	71.74	71.96	79.69	67.55	9	9
SDXL_base_1 [52]	53.50	54.45	87.28	54.59	3.60	52.23	53.51	75.71	3.29	81.45	63.67	65.82	65.82	72.15	62.03	10	10
EMU3 [69]	54.29	54.86	87.58	57.78	3.54	50.97	52.61	78.50	3.12	76.53	59.90	61.56	57.67	67.05	58.35	11	12
NOVA [9]	50.69	51.35	79.61	53.16	3.27	52.73	52.77	71.39	3.29	78.17	68.19	66.89	62.93	73.65	61.26	12	14
Kandinsky-3 [1]	58.21	58.74	93.58	60.58	4.21	48.37	51.60	79.72	2.84	72.79	50.14	57.60	61.27	69.36	55.15	13	13
Seed-xi [12]	50.73	51.49	79.74	53.59	3.07	50.96	53.93	70.08	3.18	81.95	61.43	66.28	66.28	72.94	60.32	14	11
LaVi-Bridge [91]	50.56	51.16	66.27	50.66	3.09	50.19	51.01	60.83	3.08	69.22	59.10	62.40	57.70	73.63	55.09	15	15
Hart [59]	49.80	49.85	88.87	53.75	3.20	50.30	53.04	81.24	3.14	76.10	59.29	61.99	67.07	72.40	60.53	16	16
ELLA [23]	44.61	45.17	57.68	44.30	2.24	49.07	50.14	54.29	3.10	75.19	54.90	56.71	58.35	71.29	49.65	17	18
SD_v2-1 [53]	47.68	49.23	75.27	50.71	2.69	47.96	50.41	64.80	3.02	77.42	48.86	54.39	60.33	65.80	52.49	18	17
LLMGA [75]	48.67	50.54	81.63	51.16	2.90	43.43	46.21	73.96	2.59	59.66	37.67	44.91	40.04	65.27	43.58	19	19
Janus [70]	36.98	37.34	41.82	37.81	1.55	45.94	47.16	41.31	2.82	78.62	42.95	48.67	49.39	48.18	36.56	20	20
Vila-u [73]	33.80	33.18	19.54	33.80	1.23	43.47	44.32	33.75	2.61	71.08	35.24	35.85	35.85	37.32	28.05	21	21
i-Code-V3 [60]	34.70	35.14	20.76	32.62	2.41	39.80	40.49	31.80	1.68	60.11	25.00	30.98	26.70	26.95	21.66	22	22
LlamaGen [57]	29.96	30.74	12.18	29.80	1.25	37.73	39.09	27.46	2.31	61.35	21.19	22.88	22.88	20.82	19.54	23	23
LWM [44]	28.88	29.26	11.54	29.14	1.45	35.46	36.52	24.42	2.09	58.52	15.48	15.88	18.12	17.45	13.20	24	24
SRCC to human ↑	-	0.979	0.940	0.959	0.978	-	0.983	0.777	0.982	0.695	-	0.993	0.924	0.915	0.985	-	0.992
RMSE to human ↓	-	1.24	28.90	2.01	47.80	-	1.60	21.85	47.50	26.52	-	3.34	5.75	10.73	4.44	-	0.866

5.4. Zero-shot Cross-dataset Evaluation

As shown in Table 5, we present zero-shot cross-dataset performance comparisons on multiple benchmark. LMM4LMM achieves the best performance on EvalMi-50K and other 7 AIGI evaluation benchmarks. To further validate the generalization capability of our approach, we fine-tuned our method on EvalMuse-40K [18]. The results demonstrate that fine-tuning on EvalMuse-40K yields slightly lower generalization, likely due to the scoring in EvalMuse-40K is coarser compared to our dataset, which

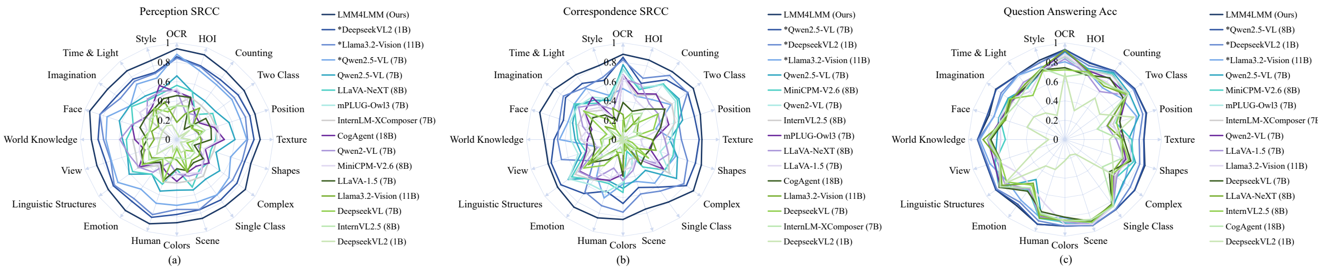

Figure 6. Comparison of MOSs and QA accuracy of different LMM models across different prompt challenges. (a) Results across perceptual quality MOS. (b) Results across T2I correspondence MOS. (c) Results across question answering accuracy.

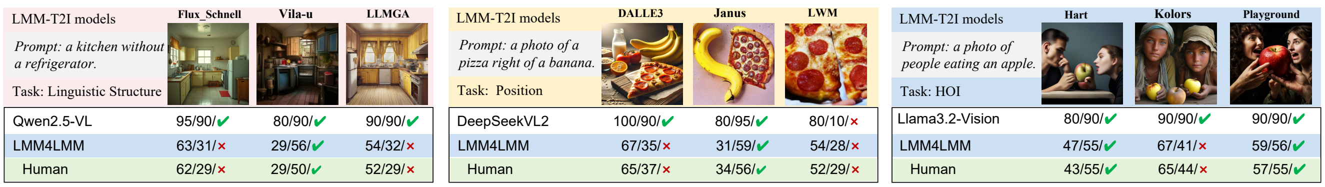

Figure 7. Visualization of the Perception/Correspondence/QA prediction from different methods compared to human annotation.

Table 5. Zero-shot cross-dataset performance comparison on multiple benchmarks. We finetune our model on EvalMi-50K/EvalMuse-40K respectively. FGA-BLIP2 [18] is finetuned on EvalMuse-40K [18]. *Refers to scores finetuned on the specific dataset.

Method	**EvalMi-50K (Ours)**		EvalMuse [18]		GenAI-Bench [34]		TIFA [24]		RichHF [41]		AGIQA3K [41]		AIGCIQA [41]		CompBench [25]	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
CLIPScore [19]	0.2607	0.3072	0.2993	0.2933	0.1676	0.2030	0.3003	0.3086	0.0570	0.3024	0.5972	0.6839	0.2337	0.6839	0.2044	0.1944
BLIPScore [40]	0.2900	0.3468	0.3583	0.3348	0.2734	0.2979	0.4287	0.4543	0.1425	0.3105	0.6230	0.7380	0.3784	0.2576	0.3967	0.3940
ImageReward [80]	0.4991	0.5523	0.4655	0.4585	0.3400	0.3786	0.6211	0.6336	0.2747	0.3291	0.7298	0.7862	0.5870	0.5911	0.4367	0.4307
PickScore [28]	0.4611	0.4692	0.4399	0.4328	0.3541	0.3631	0.4279	0.4342	0.3916	0.4133	0.6977	0.7633	0.5045	0.4998	0.1115	0.0955
HPSv2 [72]	0.5336	0.5525	0.3745	0.3657	0.1371	0.1693	0.3647	0.3804	0.1871	0.2577	0.6349	0.7000	0.6068	0.5989	0.2844	0.2761
VQAScore [35]	0.6062	0.6118	0.4877	0.4841	0.5534	0.5175	0.6951	0.6585	0.4826	0.4094	0.6273	0.6677	0.6394	0.5869	0.5832	0.5328
FGA-BLIP2 [18]	0.6755	0.6916	**0.7723***	**0.7716***	0.5638	0.5684	**0.7657**	**0.7508**	0.4576	0.4967	0.7793	0.8042	**0.7432**	**0.7367**	**0.6231**	**0.6007**
Ours (Train on EvalMi)	**0.8702***	**0.8924***	0.6560	0.6503	**0.7082**	**0.7019**	**0.7734**	**0.7604**	**0.6231**	**0.6259**	**0.8011**	**0.8205**	**0.7514**	**0.7473**	**0.6911**	**0.6726**
Ours (Train on EvalMuse)	**0.6764**	**0.6928**	**0.7852***	**0.7958***	**0.6523**	**0.6363**	0.7390	0.7264	**0.5836**	**0.5972**	**0.7797**	**0.8118**	0.6823	0.6782	0.5090	0.5020

Table 6. Ablation study on the quality-level initialization, LoRA fine-tuning strategy, and the different backbone of LMM4LMM.

	Backbone & Strategy				Quality (ours)			Correspondence (ours)			QA	GenAI-Bench			AGIQA3K		
No.	Backbone	quality level	$\text{LoRA}_{r=8}$ (vision)	$\text{LoRA}_{r=8}$ (llm)	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC	Acc	SRCC	PLCC	KRCC	SRCC	PLCC	KRCC
(1)	InternVL2.5 (8B)	✔			0.828	0.857	0.700	0.870	0.892	0.742	86.1%	0.660	0.653	0.535	0.757	0.741	0.613
(2)	InternVL2.5 (8B)		✔	✔	0.865	0.895	0.687	0.888	0.906	0.721	87.9%	0.707	0.701	0.530	0.799	0.817	0.605
(3)	InternVL2.5 (8B)	✔		✔	0.872	0.900	0.695	0.897	0.911	0729	87.3%	0.689	0.680	0.515	0.790	0.768	0.607
(4)	InternVL2.5 (8B)	✔	✔		0.871	0.900	0.694	0.893	0.913	0.723	86.9%	0.688	0.680	0.514	0.778	0.810	0.593
(5)	InternVL2.5 (8B)	✔	✔	✔	0.886	0.909	0.714	0.897	0.916	0.733	87.9%	0.708	0.702	0.532	0.801	0.821	0.608
(6)	InternVL2.5 (26B)	✔			0.834	0.867	0.704	0.848	0.866	0.718	86.6%	0.671	0.663	0.550	0.770	0.793	0.634
(7)	InternVL2.5 (26B)	✔	✔	✔	0.882	0.906	0.709	0.897	0.906	0.727	86.9%	0.726	0.741	0.548	0.811	0.814	0.627
(8)	DeepseekVL2 (1B)	✔	✔	✔	0.790	0.825	0.651	0.782	0.799	0.646	84.9%	0.613	0.616	0.500	0.782	0.712	0.558
(9)	Qwen2.5VL (8B)	✔	✔	✔	0.699	0.750	0.572	0.801	0.822	0.666	87.2%	0.626	0.616	0.505	0.767	0.786	0.619
(10)	Llama3.2VL (11B)	✔	✔	✔	0.756	0.789	0.616	0.640	0.646	0.517	78.1%	0.397	0.418	0.315	0.678	0.747	0.5500

highlights the importance of fine-grained MOS annotations in improving the model's generalization abilities.

5.5. Ablation Study

To validate the contributions of the different modules in LMM4LMM, we conduct comprehensive ablation studies, with results summarized in Table 6. Our analysis reveals three key findings: First, experiments (1), (2), and (5) demonstrate the effectiveness of quality-level initialization in model performance. Second, through experiments (3)-(7), we validate the significant performance gains achieved by LoRA fine-tuning. Third, experiments (7)-(10) compare different backbone architectures, confirming the effectiveness of our combined approach, which leverages the right balance of modules and model architecture to achieve state-of-the-art performance in IQA.

6. Conclusion

In this paper, we introduce EvalMi-50K, a large-scale dataset and benchmark consisting of 50,400 images generated by 24 T2I models using 2,100 prompts across 20 task-specific challenges and 2M+ subjective ratings from the perception, text-image correspondence, and task-specific accuracy, respectively. We use EvalMi-50K to benchmark and evaluate both the generation ability of T2I models and the interpretation ability of LMMs. Based on EvalMi-50K, we propose LMM4LMM, an LMM-based evaluation model that leverages instruction tuning and LoRA adaptation to achieve AIGI perceptual quality evaluation and T2I correspondence attribution. Extensive experiments demonstrate that LMM4LMM achieves state-of-the-art performance on the EvalMi-50K dataset and manifests strong zero-shot generalization ability on the other seven benchmarks.

References

[1] Vladimir Arkhipkin, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Anton Bukashkin, Konstantin Kulikov, et al. Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. *arXiv preprint arXiv:2410.21061*, 2024. 4, 7

[2] Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models. *arXiv preprint arXiv:2406.10429*, 2024. 3

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 7

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1, 7

[5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 74–91, 2024. 7

[6] Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Jiarui Wang, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 40111–40144, 2024. 2

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 4, 6, 7

[8] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*, 2023. 3

[9] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024. 4, 7

[10] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, and Guangtao Zhai. Finevq: Fine-grained user generated content video quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 3206–3217, 2025. 2

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 7

[12] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 7

[13] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 52132–52152, 2023. 2, 3

[14] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3209–3218, 2022. 6

[15] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[17] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024. 7

[18] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation. *arXiv preprint arXiv:2412.18150*, 2024. 2, 6, 7, 8

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2, 3, 6, 8

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[21] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2024. 6, 7

[22] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-

rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, page 3, 2022. 2, 5

[23] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 7

[24] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20406–20417, 2023. 2, 8

[25] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 78723–78747, 2023. 2, 3, 8

[26] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1733–1740, 2014. 2, 3, 6

[27] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5148–5157, 2021. 6

[28] 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. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 36652–36663, 2023. 2, 3, 6, 8

[29] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2024. 3

[30] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 3

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

[32] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. 7

[33] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 69981–70011, 2023. 3

[34] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. 2, 8

[35] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5290–5301, 2024. 3, 6, 8

[36] 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 (TCSVT)*, 34(8):6833–6846, 2023. 2

[37] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. Aigiqa-20k: A large database for ai-generated image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6327–6336, 2024. 2

[38] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 7

[39] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 6, 7

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International conference on machine learning (ICML)*, pages 12888–12900, 2022. 6, 8

[41] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19401–19411, 2024. 2, 8

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 34892–34916, 2023. 1, 2

[43] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 1, 6, 7

[44] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 7

[45] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 7

[46] AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024, 2024. 6, 7

[47] Xiongkuo Min, Ke Gu, Guangtao Zhai, Jing Liu, Xiaokang Yang, and Chang Wen Chen. Blind quality assessment based on pseudo-reference image. *IEEE Transactions on Multimedia (TMM)*, 20(8):2049–2062, 2017. 6

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

[49] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yucheng Zhu, Jiantao Zhou, Guodong Guo, Xiaokang Yang, Xinping Guan, and Wenjun Zhang. Quality evaluation of image dehazing methods using synthetic hazy images. *IEEE Transactions on Multimedia (TMM)*, 21(9):2319–2333, 2019. 2

[50] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing (TIP)*, 21(12):4695–4708, 2012. 3, 6

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

[52] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 7

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 7

[54] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 25278–25294, 2022. 4

[55] BT Series. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT*, pages 500–13, 2012. 2, 5

[56] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3667–3676, 2020. 2, 3, 6

[57] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 7

[58] Wei Sun, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, 17(6):1178–1192, 2023. 2, 6

[59] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024. 7

[60] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 16083–16099, 2023. 7

[61] Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024. 7

[62] 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 *Proceedings of the CAAI International Conference on Artificial Intelligence (CICAI)*, pages 46–57, 2023. 2

[63] Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Quality assessment for ai generated images with instruction tuning. *arXiv preprint arXiv:2405.07346*, 2024. 2

[64] Jiarui Wang, Huiyu Duan, Ziheng Jia, Yu Zhao, Woo Yi Yang, Zicheng Zhang, Zijian Chen, Juntong Wang, Yuke Xing, Guangtao Zhai, and Xiongkuo Min. Love: Benchmarking and evaluating text-to-video generation and video-to-text interpretation. *arXiv preprint arXiv:2505.12098*, 2025.

[65] Jiarui Wang, Huiyu Duan, Juntong Wang, Ziheng Jia, Woo Yi Yang, Xiaorong Zhu, Yu Zhao, Jiaying Qian, Yuke Xing, Guangtao Zhai, and Xiongkuo Min. Dfbench: Benchmarking deepfake image detection capability of large multimodal models. *arXiv preprint arXiv:2506.03007*, 2025.

[66] Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. Aigv-assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with lmm. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 18869–18880, 2025.

[67] Juntong Wang, Jiarui Wang, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Tdve-assessor: Benchmarking and evaluating the quality of text-driven video editing with lmms. *arXiv preprint arXiv:2505.19535*, 2025. 2

[68] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6, 7

[69] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 7

[70] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 7

[71] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 6

[72] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2096–2105, 2023. 2, 6, 8

[73] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 7

[74] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 6, 7

[75] Bin Xia, Shiyin Wang, Yingfan Tao, Yitong Wang, and Jiaya Jia. Llmga: Multimodal large language model based generation assistant. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 389–406, 2024. 7

[76] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 1, 7

[77] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 1, 7

[78] Yuke Xing, Jiarui Wang, Peizhi Niu, Wenjie Huang, Guangtao Zhai, and Yiling Xu. 3dgs-ieval-15k: A large-scale image quality evaluation database for 3d gaussian-splatting. *arXiv preprint arXiv:2506.14642*, 2025. 2

[79] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing (TIP)*, 25(9):4444–4457, 2016. 6

[80] 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. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 15903–15935, 2023. 2, 6, 8

[81] Zitong Xu, Huiyu Duan, Guangji Ma, Liu Yang, Jiarui Wang, Qingbo Wu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. Harmonyiqa: Pioneering benchmark and model for image harmonization quality assessment. *arXiv preprint arXiv:2501.01116*, 2025.

[82] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 995–1002, 2013. 3, 6

[83] Woo Yi Yang, Jiarui Wang, Sijing Wu, Huiyu Duan, Yuxin Zhu, Liu Yang, Kang Fu, Guangtao Zhai, and Xiongkuo Min. Lmme3dhf: Benchmarking and evaluating multimodal 3d human face generation with lmms. *arXiv preprint arXiv:2504.20466*, 2025. 2

[84] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6, 7

[85] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multimodal large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 6, 7

[86] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63:1–52, 2020. 2

[87] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 6, 7

[88] 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 (TCSVT)*, 30(1):36–47, 2018. 3, 6

[89] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14071–14081, 2023. 6

[90] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A perceptual quality assessment exploration for aigc images. *arXiv preprint arXiv:2303.12618*, 2023. 2

[91] Shihao Zhao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, and Kwan-Yee K Wong. Bridging different language models and generative vision models for text-to-image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–86, 2024. 7

[92] Xiaorong Zhu, Ziheng Jia, Jiarui Wang, Xiangyu Zhao, Haodong Duan, Xiongkuo Min, Jia Wang, Zicheng Zhang, and Guangtao Zhai. Gobench: Benchmarking geometric optics generation and understanding of mllms. *arXiv preprint arXiv:2506.00991*, 2025. 1