A-Bench: Are LMMs Masters at Evaluating AI-generated Images?

Zicheng Zhang[1*] **Haoning Wu**[2*] **Chunyi Li**[1*] **Yingjie Zhou**[1] **Wei Sun**[1]
Xiongkuo Min[1] **Zijian Chen**[1] **Xiaohong Liu**[1] **Weisi Lin**[2] **Guangtao Zhai**[1†]
[1]Shanghai Jiao Tong University [2]Nanyang Technological University
zzc1998@sjtu.edu.cn, haoning001@e.ntu.edu.sg

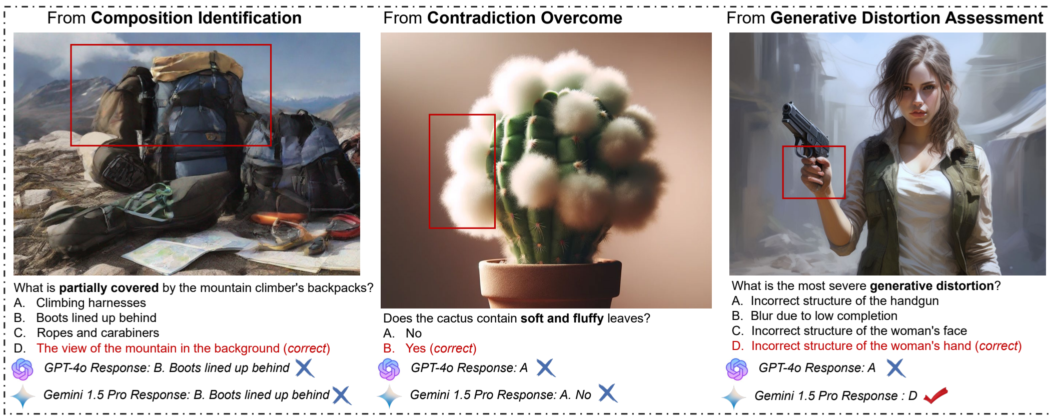


Figure 1: Error cases from the **A-Bench**.

Abstract

How to accurately and efficiently assess AI-generated images (AIGIs) remains a critical challenge for generative models. Given the high costs and extensive time commitments required for user studies, many researchers have turned towards employing large multi-modal models (LMMs) as AIGI evaluators, the precision and validity of which are still questionable. Furthermore, traditional benchmarks often utilize mostly natural-captured content rather than AIGIs to test the abilities of LMMs, leading to a noticeable gap for AIGIs. Therefore, we introduce **A-Bench** in this paper, a benchmark designed to diagnose *whether LMMs are masters at evaluating AIGIs*. Specifically, **A-Bench** is organized under two key principles: 1) Emphasizing both high-level semantic understanding and low-level visual quality perception to address the intricate demands of AIGIs. 2) Various generative models are utilized for AIGI creation, and various LMMs are employed for evaluation, which ensures a comprehensive validation scope. Ultimately, 2,864 AIGIs from 16 text-to-image models are sampled, each paired with question-answers annotated by human experts, and tested across 18 leading LMMs. We hope that **A-Bench** will significantly enhance the evaluation process and promote the generation quality for AIGIs. The benchmark is available at https://github.com/Q-Future/A-Bench.

[*]Equal contribution.
[†]Corresponding author.

Preprint. Under review.

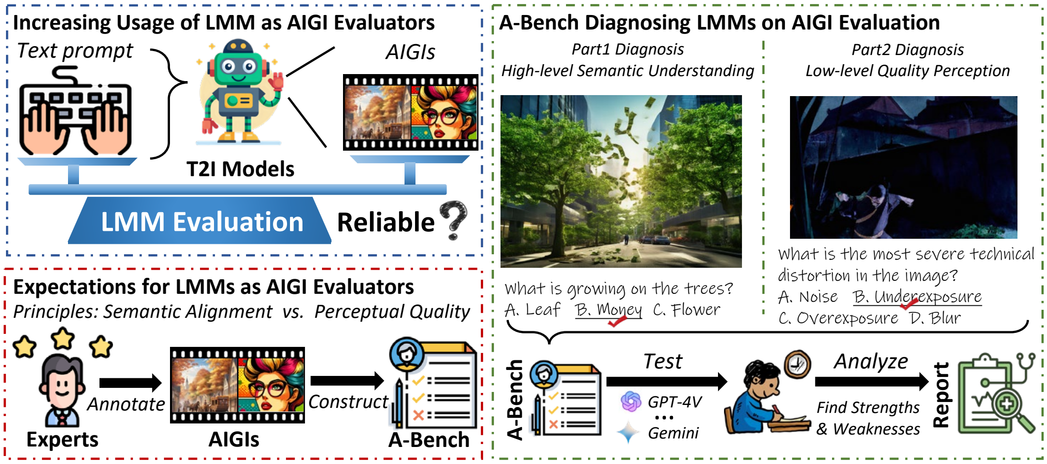


Figure 2: The proposed **A-Bench** is designed to find out whether LMMs are reliable for T2I AIGI evaluation. Instead of directly assessing the performance of LMM-based metrics, we evaluate the **LMMs themselves behind** by examining whether the *fundamental questions regarding semantic understanding and quality perception can be correctly answered.* Based on the benchmark results, we can then 'diagnose' the strengths and weaknesses across various LMMs.

1 Introduction

One look is worth a thousand words. Inspired by this age-old adage, numerous researchers dedicate their efforts to developing text-to-image (T2I) models that vividly bring text to life through imagery. These T2I models, driven by free-form text prompts, aim to create images that **accurately align with the text and showcase high perceptual quality.** Innovations such as AlignDRAW [Mansimov et al., 2015] and the text-conditional GAN [Reed et al., 2016] have introduced differential architecture for image generation. The field continues to advance with the development of stable diffusion models [Saharia et al., 2022, Rombach et al., 2022a], significantly propelling T2I technology forward. On the commercial front, major corporations leverage vast-scale data to launch stunningly effective T2I models, such as DALL-E [Ramesh et al., 2022], Midjourney [Holz, 2023], Parti [Yu et al., 2022], etc. However, despite their diversity and widespread adoption, all these advanced T2I models occasionally *face issues of low alignment with prompts and low perceptual quality* in creating AI-generated images (AIGIs), necessitating careful evaluation and improvement.

The alignment and quality evaluation of AIGIs present significant challenges that *small expert models* attempt to address. Although these *small expert models* offer some solutions, they possess inherent drawbacks and often fail to meet contemporary demands. Specifically, for alignment assessment, CLIP-based similarity models struggle with accurately judging alignment, particularly with complex text prompts [Radford et al., 2021a]. When it comes to quality evaluation, traditional image quality/aesthetic assessment methods (IQA/IAA) are not capable of identifying AIGI-generative distortions [Wu et al., 2023a,b], rendering them unsuitable for this specialized task.

Many researchers are increasingly relying on large language models (LLMs) and large multi-modal models (LMMs) for their human-like processing capabilities, which are presumed to enable accurate judgments of alignment and quality in generated content. Consequently, a variety of LMM-based evaluators have been developed, including VIE-Score [Ku et al., 2023], Prometheus [Kim et al., 2023], GPT4V-Eval [Lin et al., 2024], TIFA [Hu et al., 2023], and Davidsonian Graph [Cho et al., 2023], among others. Despite these advancements, a fundamental question remains:

Are LMMs reliable for evaluating T2I AIGIs?

These LMM-based metrics traditionally employ evaluation criteria such as SRCC/PLCC to determine their reliability. However, this approach only reveals *how well the metrics perform, without shedding light on their specific strengths and weaknesses.* To address this gap, we propose conducting a detailed and comprehensive *'diagnostic'* benchmark → **A-Bench**, focusing on LMMs' capabilities in semantic understanding and quality assessment. Rather than directly evaluating these LMM-based metrics, we focus on *studying the LMMs themselves behind*. We move away from computing SRCC/PLCC criteria and instead examine *whether the fundamental perceptual questions can be correctly answered*, which is the core basis of all LMM-based evaluators. To initiate our exploration on the AIGI evaluation abilities of LMMs, we first construct the **A-Bench** centered on a pivotal question:

What do we expect from LMMs as AIGI evaluators?

The answer lies in the capabilities of **semantic alignment and quality evaluation**. Then We define two key diagnostic subsets: **A-Bench**^{P1}$\rightarrow$*high-level semantic understanding*, and **A-Bench**^{P2}$\rightarrow$*low-level quality perception*. For *high-level semantic understanding*, **A-Bench**^{P1} targets three critical areas: *Basic Recognition*, *Bag-of-Words Pitfalls Discrimination*, and *Outside Knowledge Realization*, which are designed to progressively test the LMM's capability in AIGI semantic understanding, moving from simple to complex prompt-related content. For *low-level quality perception*, **A-Bench**^{P2} concentrates on *Technical Quality Perception*, *Aesthetic Quality Evaluation*, and *Generative Distortion Assessment*, which are designed to cover the common quality issues and AIGI-specific quality problems. The selection of focus areas is meticulously designed to encompass the most prevalent application scenarios. Specifically, a comprehensive dataset of 2,864 AIGIs sourced from various T2I models is compiled, including 1,408 AIGIs for **A-Bench**^{P1} and 1,456 for **A-Bench**^{P2}. Each AIGI is paired with a question-answer set annotated by human experts. Then we test 18 prominent LMMs, including both *open-source* and *closed-source* models, on the **A-Bench**. From the results that the best LMM still falls behind humans by a large margin, we can derive the following conclusion:

LMMs are still not masters at evaluating AIGIs.

All LMMs lag behind even the poorest human performance on **A-Bench**, and there is a substantial disparity between *open-source* LMMs and *closed-source* LMMs. The performance across different subcategories fluctuates for both **A-Bench**^{P1} and **A-Bench**^{P2}, indicating that LMMs are not yet robust for different evaluation scenarios for AIGIs. There remains a considerable gap and significant room for improvement before LMMs can be considered masters of evaluating AIGIs.

In summary, we systematically explore the capabilities of LMMs in semantic understanding and quality perception, both crucial for their role as AIGI evaluators. These two essential capabilities constitute the core of the proposed **A-Bench**, the first 'diagnostic' benchmark specifically designed for LMM assessment in AIGI evaluation. Our contributions are summarized as follows:

- We carry out the **first 'diagnostic' benchmark** on AIGI evaluation for LMMs, which consists of 2,864 AIGIs (sampled from various T2I models) paired with question-answer sets on both high-level semantic understanding and low-level quality perception.
- A detailed discussion is made about **what to 'diagnose'**. Semantic understanding is subdivided into *Basic Recognition, Bag-of-Words Pitfalls Discrimination*, and *Outside Knowledge Realization* while quality perception is broken down into *Technical Quality Perception, Aesthetic Quality Evaluation*, and *Generative Distortion Assessment*.
- From the benchmark results, several insights are gleaned, which can enable us to diagnose various issues with different LMMs and assist in their improvement for AIGI evaluation.

2 Related Works

2.1 Large Muli-modal Models

Large language models (LLMs), such as GPT-4 [OpenAI, 2023], T5 [Chung et al., 2022], and LLaMA [Touvron et al., 2023], exhibit exceptional linguistic capabilities in general human knowledge domains. By integrating visual input via CLIP [Radford et al., 2021b] and additional adaptation modules, large multi-modal models (LMMs) [Li et al., 2023a, Gao et al., 2023, Liu et al., 2023a, Dai et al., 2023, Zhang et al., 2023] are capable of addressing diverse multi-modal tasks, including image captioning, visual question answering, visual segmentation, visual classification, visual reasoning, etc. Namely, OpenFlamingo [Awadalla et al., 2023] initially integrates several gated cross-attention dense blocks into the pretrained language encoder layers. InstructBLIP [Dai et al., 2023] extends BLIP-2 [Li et al., 2023b] by incorporating vision-language instruction tuning. To further develop *open-source* LMMs, many works have employed GPT-4 [OpenAI, 2023] to create data for vision-language tuning, such as LLaVA series [Liu et al., 2023a,b, 2024]. However, whether these LMMs are masters at evaluating T2I AIGIs is still questionable, which needs further investigation.

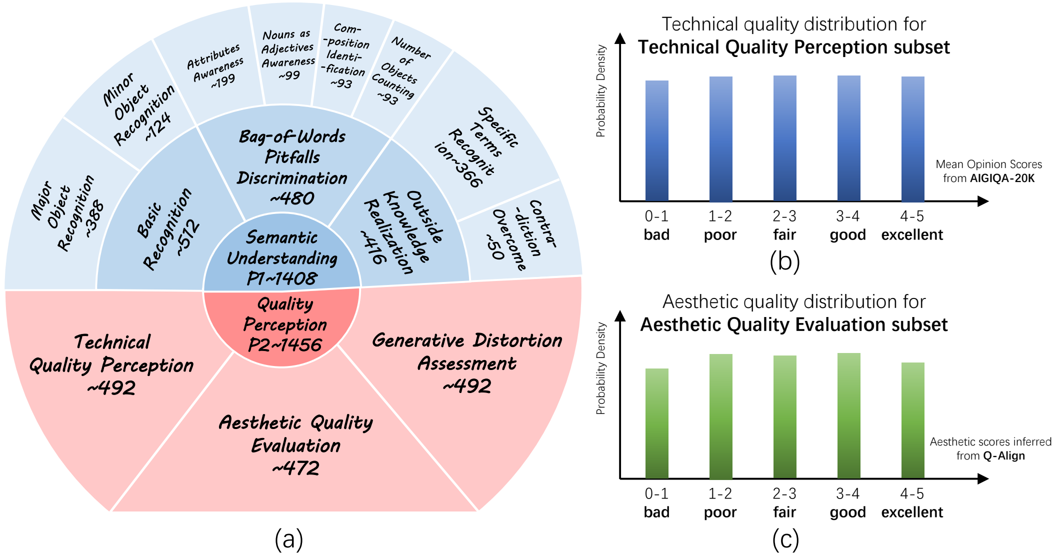


Figure 3: Illustration of focused aspects and corresponding quality distributions for **A-Bench**. The focused aspects and the amount of AIGIs employed are shown in (a). The quality scores of AIGIs sampled for **Technical Quality Perception** and **Aesthetic Quality Evaluation** subsets are obtained from AIGIQA-20K [Li et al., 2024] and predicted from Q-Align [Wu et al., 2023c] respectively.

2.2 Multi-modal Benchmarks

Benchmarks such as COCO Caption [Chen et al., 2015] and Nocaps [Agrawal et al., 2019] evaluate the capability of models to generate textual descriptions for images. Subsequently, benchmarks like GQA [Hudson and Manning, 2019] and OK-VQA [Marino et al., 2019] focus on visual question answering, assessing multi-modal models' visual perception and reasoning abilities. Further complexities are added in benchmarks such as TextVQA [Singh et al., 2019] and ScienceQA [Lu et al., 2022], which incorporate OCR tasks and commonsense reasoning, respectively. MME [Fu et al., 2023] and MMbench [Liu et al., 2023c] provide comprehensive evaluations of LMMs across various subtasks. Additionally, MMMU [Yue et al., 2023] targets extensive multi-disciplinary tasks that require college-level knowledge and sophisticated reasoning. More recently, Q-Bench [Wu et al., 2023a] focuses specifically on assessing the low-level visual perception capabilities of LMMs. Despite these efforts, there is still a gap in systematic benchmarks for assessing the abilities of LMMs in AIGI evaluation, prompting the development of **A-Bench** to address this shortfall.

3 Constructing the A-Bench

3.1 Key Principles

Covering High-level and Low-level Attributes. The demand for generating images has become increasingly stringent, with requirements for *not only accurate adherence to prompt specifications but also high visual quality of AIGIs*. To ascertain whether LMMs can effectively evaluate whether AIGIs meet these criteria, it is essential to assess their capabilities in both **high-level semantic understanding** and **low-level quality perception**. High-level semantic understanding encompasses basic recognition and the integration of external knowledge, whereas low-level quality perception involves the identification of technical quality, aesthetic appeal, and generative distortions. The detailed focused aspects can be overviewed in Fig. 3 (a).

Ensuring Diverse AIGI Scope. Considering the variety of current generative models and their application scenarios, we have selected a broad range of mainstream text-to-image (T2I) generation models to produce AI-generated images (AIGIs). To assess high-level semantic understanding, we design prompts rich in content to ensure diversity among the generated images. For evaluating low-level quality perception, we employ uniform sampling to encompass a wide spectrum of visual quality and the corresponding quality distributions are illustrated in Fig 3 (b) and (c). Throughout

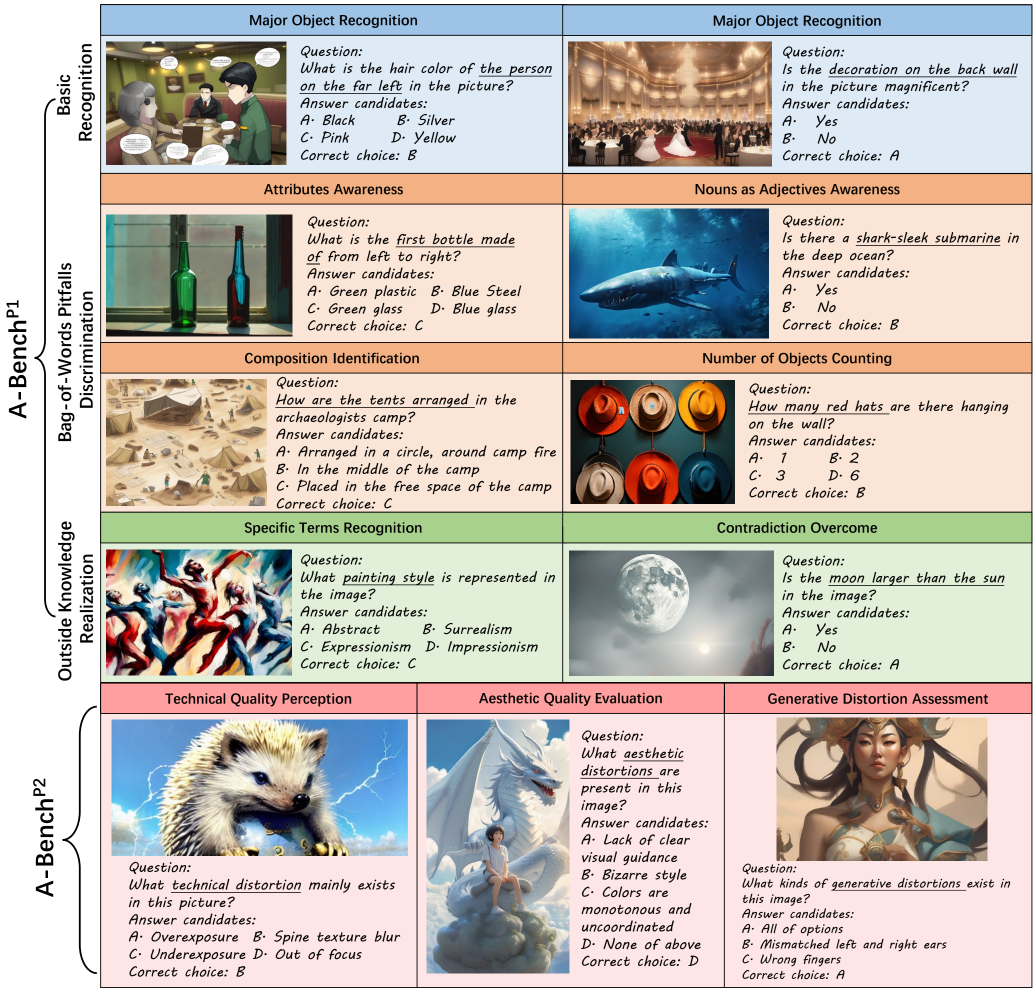


Figure 4: Examples of **A-Bench**. Each AIGI is accompanied by a question-answer pair.

the benchmarking process, we test multiple *open-source* and *closed-source* LMMs to guarantee a comprehensive evaluation. These measures ensure that our proposed **A-Bench** encompasses a diverse and extensive scope. More details about AIGIs collection can be referred to in Sec. A.1.

3.2 Focused Aspects

The key evaluation aspects of T2I models involve image-text alignment and image visual quality, which correspond to high-level semantic understanding and low-level quality perception abilities. Some representative examples regarding the subcategories discussed below are exhibited in Fig. 4.

3.2.1 High-level Semantic Understanding

To evaluate whether LMMs can effectively assess image-text alignment, we implement the **A-Bench**^{P1}, which consists of 1,408 challenging multi-modal question-answer pairs that focus on high-level semantic understanding for AIGIs. The high-level semantic understanding can be divided into the following subcategories, moving from simple to complex prompt-related content:

Basic Recognition. This aspect concentrates on the fundamental semantic understanding of AIGIs, which can be subdivided into two distinct areas: 1) **Major Object Recognition**, which involves *recognizing the primary objects in the image*, such as humans or objects depicted in the foreground. 2) **Minor Object Recognition**, which *pertains to the identification of less-prominent objects within the image*, such as background elements or secondary characters.

Bag-of-Words Pitfalls Discrimination. This dimension focuses on the discriminative semantic understanding of AIGIs crafted with *Bag-of-words prompts* (encompassing rich descriptive attributes or complex object relationships [Qu et al., 2024]). This can be subdivided into the following aspects: 1) **Attributes Awareness**, defined as *the capability to accurately identify the attributes of objects in AIGIs*. 2) Additionally, given that T2I models may incorrectly interpret nouns as adjectives, resulting in *the unwanted generation of objects instead of the intended attributes*, we have also introduced a dimension called **Nouns as Adjectives Awareness** to address this issue. 3) **Composition Identification**, recognized as the ability to *correctly comprehend the compositional relationships* such as orientation, occlusion, size comparison, and spatial arrangement. 4) **Number of Objects Counting**, regarded as *the ability to accurately count the specified objects in the image*, which is crucial for assessing whether the AIGI aligns with the numerical specifications of the prompt.

Outside Knowledge Realization. This aspect emphasizes the reasoning ability to utilize external knowledge not directly depicted in the images [Schwenk et al., 2022], and can be broken down into the following dimensions: 1) **Specific Terms Recognition**: This involves *identifying specific scenes and objects* related to distinct domains such as geography, sports, science, materials, food, everyday life, creatures, brands, and styles. 2) **Contradiction Overcome**, recognized as the *ability to correctly interpret AIGIs even when their content contradicts established world knowledge*, which is particularly crucial for evaluating AIGIs generated from controversial prompts.

3.2.2 Low-level Quality Perception

Conversely, to determine the ability of LMMs on image visual quality, we conduct the **A-Bench**^{P2}, comprising 1,456 challenging multi-modal question-answer pairs centered on low-level quality perception for AIGIs, which can be categorized into the following aspects: 1) **Technical Quality Perception** This indicates the *low-level characteristics that directly degrade the quality of images*, such as blur, noise, exposure, etc [Su et al., 2021, Ying et al., 2020]. 2) **Aesthetic Quality Evaluation** This indicates the *attributes that affect the aesthetic appeal of AIGIs and evoke varied human feelings*, which include color, lighting, etc [Huang et al., 2024]. 3) **Generative Distortion Assessment** This indicates the *unexpected AIGI-specific distortions* [Chen et al., 2023a, Li et al., 2023c, 2024], such as generative blur caused by low completion, confusing geometry structure, unnaturalness, etc.

3.3 Question Collection

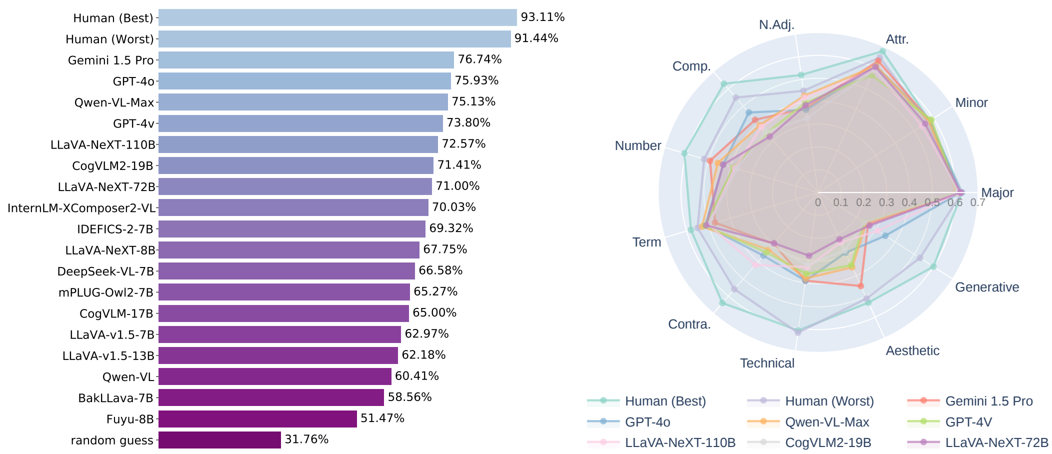
Question Type In the **A-Bench**, two types of question formats are utilized, including *Yes-or-No* questions and *What* questions. The *Yes-or-No* questions (accounting for 25.9%) are used to evaluate the fundamental judgment abilities of LMMs while the *What* questions (accounting for 74.1%) are more complicated and require LMMs to gain a more comprehensive understanding of the AIGIs.

Human Expert Annotation We have assembled a team of 15 human annotators, each with expert experience in AIGI evaluation, to develop questions for **A-Bench**. This annotation process is conducted in a controlled laboratory environment, ensuring consistency and reliability. Annotators are tasked with designing questions specific to the sub-categories of the AIGIs under review, utilizing their extensive knowledge to determine the content and format of each question. To ensure the highest quality and suitability, each question undergoes a rigorous review process, with at least three other expert annotators double-checking it. More details can be acquired in Sec. A.3.

Question Response Specifically, the example input query to LMMs can be exemplified as:

#User: What painting style is represented in the image? <|IMAGE_TOKEN|>
A. Abstract B. Surrealism C. Expressionism D. Impressionism
Answer with the option's letter from the given choices directly.

The answer candidates and correct answers are shuffled during the evaluation process. Since the responses from LMMs can be in various forms (if the correct choice is C) such as '*C*', '*Expressionism*', '*The painting style of image is expressionism*', etc., we employ a GPT-assisted choice evaluation technique proposed in [Liu et al., 2023c, Wu et al., 2023a] to validate the correctness of LMMs responses. More details are shown in Sec. A.4.



(a) Overall results of **A-Bench**. (b) Detailed results of **A-Bench**.

Figure 5: *A Quick Look* of the **A-Bench** outcomes. (a) showcases a comparative analysis of the overall accuracy between humans, 18 selected LMMs (both *closed-source* and *open-source*), and *random guess*. (b) displays a radar chart that details the accuracy performance (**subtracting the accuracy** of *random guess*) of the **top-7** LMMs across various subcategories of **A-Bench**.

4 Experiment Results

4.1 Benchmark Candidates

To ensure the results are comprehensive and up-to-date, we select the latest and widely used LMMs for benchmarking. The **Proprietary LMMs** (*closed-source*) include Gemini 1.5 Pro [Reid et al., 2024], GPT-4v [OpenAI, 2023], GPT-4o [OpenAI, 2024], and Qwen-VL-Max [Bai et al., 2023]. The **Open-source LMMs** include CogVLM2-19B (*Llama3-8B*) [Wang et al., 2023], IDEFICS-2 (*Mistral-7B-Instruct-v0.2*) [Huggingface, 2023], DeepSeek-VL-7B [Lu et al., 2024], InternLM-XComposer2-VL [Dong et al., 2024], LLaVA-NeXT (*Llama3-8B*), LLaVA-NeXT (*Qwen-72B*), LLaVA-NeXT (*Qwen-110B*) [Liu et al., 2024], mPLUG-Owl2 *(LLaMA-7B)* [Ye et al., 2023], LLaVA-v1.5 (*Vicuna-v1.5-7B*), LLaVA-v1.5 (*Vicuna-v1.5-13B*) [Liu et al., 2023b], CogVLM-17B (*Vicuna-v1.5-7B*) [Wang et al., 2023], Qwen-VL *(Qwen-7B)* [Bai et al., 2023], BakLLava (*Mistral-7B*) [Liu et al., 2023a], and Fuyu-8B (*Persimmon-8B*) [Adept, 2023]. All LMMs are tested with zero-shot setting.

For human performance on **A-Bench**, we conduct a user-study experiment with five ordinary people in a controlled laboratory setting. Initially, participants familiarize themselves with the tasks through exposure to similar cases. Subsequently, they select the appropriate responses for the questions posed in the **A-Bench**. To maintain consistency with the conditions experienced by LMMs, the order of questions is randomized, and participants receive no additional information beyond the AIGIs, questions, and answer options. The *best* and *worst* performance is included for comparison.

4.2 Findings of A-Bench

General Observation: Human > Proprietary LMMs > Open-source LMMs A concise overview of the **A-Bench** results is provided in Fig. 5, revealing several general insights: 1) All LMMs significantly outperform the *random guess*, indicating their capabilities in handling AIGI evaluation, with Gemini 1.5 Pro leading, closely followed by GPT-4o and Qwen-VL-Max. Notably, among the *open-source* LMMs, which are preferred for AIGI evaluations due to their accessibility and modifiability, LLaVA-NeXT (*Qwen-110B*) stands out, though it still significantly lags behind *closed-source* competitors. 2) Even the lowest performance by humans surpasses that of all LMMs, with a noticeable 14.70% gap compared to the top-performing LMM, Gemini 1.5 Pro, indicating that LMMs are still far from adequately performing AIGI evaluation as humans. 3) A closer examination of the radar chart in Fig. 5 (b) shows that top LMMs exhibit varied performances across different sub-categories, suggesting a lack of robustness, while humans show more consistent and balanced performance across these categories, highlighting areas where LMMs need further improvement.

Table 1: Benchmark results on the **A-Bench**^{P1} subset, which reveal the high-level semantic understanding abilities across LMMs. The best performance is marked in **bold** and the second performance is underlined for both proprietary and open-source LMMs respectively.

Categories	Basic Recognition		Bag-of-Words				Outside Knowledge		*Overall*↑
LMM *(LLM)*	*Major*↑	*Minor*↑	*Attr.*↑	*N. Adj.*↑	*Comp.*↑	*Number*↑	*Term*↑	*Contra.*↑	
HUMAN (WORST)	95.20%	94.27%	96.83%	88.64%	85.54%	82.50%	81.79%	88.89%	92.39%
HUMAN (BEST)	95.42%	95.18%	99.46%	95.12%	93.42%	91.67%	84.23%	96.00%	94.00%
Proprietary LMMs:									
GEMINI 1.5 PRO	93.80%	95.17%	**94.33%**	80.31%	72.11%	**79.31%**	73.00%	61.76%	84.69%
GPT-4V	92.97%	**95.97%**	87.43%	82.63%	64.44%	68.78%	77.58%	66.71%	83.64%
GPT-4O	**94.33%**	95.16%	91.96%	79.59%	**76.34%**	73.33%	77.53%	**68.57%**	**85.42%**
QWEN-VL-MAX	92.57%	94.77%	91.97%	**85.76%**	68.97%	75.78%	**78.94%**	65.14%	84.54%
Open-source LMMs:									
CogVLM2-19B (*Llama3-8B*)	93.30%	92.74%	89.95%	75.51%	64.52%	66.67%	75.96%	61.43%	82.47%
IDEFICS-2 (*Mistral-7B-Instruct-v0.2*)	89.95%	91.94%	86.43%	75.51%	61.29%	71.11%	73.22%	62.86%	80.00%
DeepSeek-VL-7B	91.49%	91.13%	82.41%	83.67%	63.44%	70.00%	75.41%	60.00%	81.24%
InternLM-XComposer2-VL (*InternLM2*)	92.78%	**95.16%**	86.43%	82.65%	**68.82%**	72.22%	70.77%	**64.29%**	81.90%
LLaVA-NeXT (*Llama3-8B*)	92.75%	92.37%	91.12%	83.65%	61.00%	67.02%	76.20%	62.97%	82.91%
LLaVA-NeXT (*Qwen-72B*)	**94.33%**	92.74%	**91.46%**	81.63%	62.37%	**73.33%**	**77.05%**	61.43%	**83.98%**
LLaVA-NeXT (*Qwen-110B*)	93.81%	91.13%	90.45%	**84.69%**	67.74%	67.78%	76.23%	**64.29%**	83.62%
mPLUG-Owl2 (*LLaMA-7B*)	85.31%	86.29%	83.92%	79.59%	53.76%	57.78%	71.04%	58.57%	76.44%
LLaVA-v1.5 (*Vicuna-v1.5-7B*)	87.89%	88.71%	83.92%	75.51%	61.29%	65.56%	74.86%	62.86%	79.16%
LLaVA-v1.5 (*Vicuna-v1.5-13B*)	88.40%	89.52%	86.43%	79.59%	62.37%	58.89%	74.86%	61.43%	79.67%
CogVLM-17B (*Vicuna-v1.5-7B*)	90.46%	95.14%	85.93%	77.55%	49.46%	47.78%	73.22%	61.43%	78.52%
Qwen-VL (*Qwen-7B*)	86.34%	86.29%	81.41%	77.55%	52.69%	61.11%	71.58%	57.14%	76.43%
BakLLava (*Mistral-7B*)	88.97%	81.29%	77.36%	73.87%	52.21%	62.38%	68.49%	49.12%	74.27%
Fuyu-8B (*Persimmon-8B*)	81.24%	68.07%	66.67%	57.34%	42.14%	48.46%	61.06%	29.79%	63.05%
random guess	32.28%	37.24%	31.00%	42.79%	29.81%	29.80%	26.46%	32.11%	30.87%

Findings of A-Bench^{P1}: LMMs excel at basic recognition tasks but tend to be less effective when it comes to nuanced semantic understanding. The performance results of LMMs on the **A-Bench**^{P1} subset, as detailed in Table 1, reveal several key insights: 1) Almost all LMMs show good performance in *Basic Recognition*, suggesting that they are quite adept at fundamental semantic understanding, which includes recognizing foreground and background objects in AIGIs. 2) However, their effectiveness diminishes in more complex tasks such as *Bag-of-Words*, particularly in subcategories like *Nouns as Adjectives Awareness*, *Composition Identification*, and *Number of Objects Counting*. These areas require deeper semantic understanding and reasoning, which is critical as users often employ complex prompts that include such nuanced elements. The LMMs' underperformance here indicates potential challenges in accurately aligning AIGIs with user prompts. 3) Additionally, *Outside Knowledge* poses significant challenges, with LMMs generally achieving unsatisfactory performance in the *Contradiction Overcome* subcategory, where AIGIs contain content that defies common sense, requiring LMMs to override their prior knowledge to respond correctly. The subcategory *Specific Terms* tests the knowledge base of LMMs, where proprietary LMMs generally perform better due to being trained on more recent and extensive datasets.

Findings of A-Bench^{P2}: LMMs are poor quality evaluators. The performance results of LMMs on the **A-Bench**^{P2} subset, as shown in Table 2, illustrate a notable disparity in capabilities: 1) There is a significant performance gap of approximately 20% between the top-performing LMMs and human evaluators, highlighting that LMMs lag considerably in quality perception and struggle to accurately assess the quality of AIGIs. 2) Furthermore, most LMMs exhibit their weakest performance in the *Generative Distortion Assessment* subcategory, suggesting their ineffectiveness at identifying unexpected generative distortions, such as unnatural appearances and incorrect geometric structures. 3) Interestingly, while humans generally perform better in *Technical Quality Perception* compared to *Aesthetic Quality Evaluation*, LMMs show similar performance levels in both subcategories. This difference likely stems from the more objective nature of technical quality assessments, which leads to more consistent evaluations among humans, whereas aesthetic quality, being more subjective, results in a broader range of opinions and consequently lower performance scores.

Human vs. Proprietary LMMs Proprietary (*closed-source*) LMMs are regarded as closely mirroring human perception and demonstrate superior performance, particularly in zero-shot settings for evaluating AIGI. Therefore, here we make a finer discussion about the human and proprietary LMMs. 1) Beginning with a detailed comparison of human and proprietary LMMs, we observe

Table 2: Benchmark results on the **A-Bench**^{P2} subset, which reflect the low-level quality perception abilities across LMMs. The best performance is marked in **bold** and the second performance is underlined for both proprietary and open-source LMMs respectively.

Categories / **LMM** *(LLM)*	**Technical**↑	**Aesthetic**↑	**Generative**↑	*Overall*↑
HUMAN (WORST)	94.40%	84.41%	86.30%	90.53%
HUMAN (BEST)	94.99%	86.12%	93.04%	92.25%
Proprietary LMMs:				
GEMINI 1.5 PRO	**70.97%**	**77.56%**	59.02%	**69.06%**
GPT-4V	67.68%	68.22%	57.11%	64.29%
GPT-4O	70.53%	61.65%	**67.89%**	66.76%
QWEN-VL-MAX	70.47%	69.46%	58.37%	66.04%
Open-source LMMs:				
CogVLM2-19B (*Llama3-8B*)	64.02%	61.44%	56.71%	60.71%
IDEFICS-2 (*Mistral-7B-Instruct-v0.2*)	61.18%	**68.86%**	47.36%	59.00%
DeepSeek-VL-7B	55.89%	53.81%	47.56%	52.40%
InternLM-XComposer2-VL (*InternLM2*)	62.24%	63.32%	50.31%	58.55%
LLaVA-NeXT (*Llama3-8B*)	58.54%	48.52%	52.03%	53.09%
LLaVA-NeXT (*Qwen-72B*)	59.96%	55.51%	59.76%	58.45%
LLaVA-NeXT (*Qwen-110B*)	**64.63%**	57.20%	**63.62%**	**61.88%**
mPLUG-Owl2 *(LLaMA-7B)*	57.93%	54.45%	53.86%	55.43%
LLaVA-v1.5 (*Vicuna-v1.5-7B*)	45.93%	41.31%	54.47%	47.32%
LLaVA-v1.5 (*Vicuna-v1.5-13B*)	46.14%	41.31%	48.17%	45.26%
CogVLM-17B (*Vicuna-v1.5-7B*)	54.88%	48.31%	52.44%	51.92%
Qwen-VL *(Qwen-7B)*	49.59%	34.32%	50.41%	44.92%
BakLLava (*Mistral-7B*)	47.84%	33.33%	48.44%	43.37%
Fuyu-8B (*Persimmon-8B*)	44.69%	30.28%	45.67%	40.27%
random guess	31.89%	32.91%	33.11%	32.62%

that proprietary LMMs achieve human-level performance in *Basic Recognition*, indicating their ability to correctly **assess AIGI alignment when prompts are simple**. 2) Despite this, LMMs encounter difficulties in the *Bag-of-Words* aspect, especially in identifying composition and counting objects, which highlights their **limitations in handling complex compositional relationships and specific object counts.** 3) In the *Outside Knowledge* domain, proprietary LMMs show only a slight performance gap compared to humans on *Specific Terms*, demonstrating comprehensive prior knowledge about specific terms, but they notably lag behind in identifying controversial content. While humans can easily recognize contradictory elements, proprietary LMMs often struggle due to their reliance on common sense, making accurate responses challenging. To conclude, according to the results shown in Table 1, proprietary LMMs are competent as evaluators for simple prompts in AIGI, yet they require further improvements for more complex prompts related AIGI content. 4) On the other hand, Table 2 reveals that LMMs have significant shortcomings in low-level quality perception compared to humans, with an uneven performance across different quality dimensions. Surprisingly, GPT-4o shows a distinct advantage over other proprietary LMMs in recognizing generative distortions, suggesting its superior capability in this area. However, the substantial overall difference in quality perception between proprietary LMMs and humans underlines that these models are currently **unsuitable for assessing the visual quality of AIGI**.

5 Conclusion

In conclusion, the ambition to employ LMMs for evaluating AIGIs exposes considerable deficiencies in their capabilities, as revealed by the diagnostic benchmark **A-Bench**. This benchmark scrutinizes the core capabilities of LMMs themselves, focusing on their ability to accurately address fundamental questions related to high-level semantic understanding and low-level quality perception. Our findings from **A-Bench** serve as a stark reminder of the current limitations faced by LMMs in the realm of AIGI evaluation. The results underscore that while LMMs provide valuable insights, their **evaluation capacity remains notably inferior to human performance**, especially in tasks that demand deep semantic comprehension and detailed quality assessment. By identifying specific areas for enhancement and charting a course for future research, this study not only underscores the urgent need for further development but also aids in refining the application of LMMs in AIGI evaluation tasks. Future initiatives should focus on augmenting the capabilities of LMMs to reliably match or surpass human performance in these intricate evaluation scenarios.

References

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

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

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 2204.06125, 2022.

David Holz. Midjourney. `https://www.midjourney.com`, 2023.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

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, 2021a.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. 2023a.

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. *arXiv preprint arXiv:2311.06783*, 2023b.

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*, 2023.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*, 2023.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024.

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*, pages 20406–20417, 2023.

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-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

OpenAI. Gpt-4 technical report, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021b.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023a.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

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, 2023.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2023c.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. Aigiqa-20k: A large database for ai-generated image quality assessment. *arXiv preprint arXiv:2404.03407*, 2024.

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. *arXiv preprint arXiv:2312.17090*, 2023c.

Leigang Qu, Wenjie Wang, Yongqi Li, Hanwang Zhang, Liqiang Nie, and Tat-Seng Chua. Discriminative probing and tuning for text-to-image generation. *arXiv preprint arXiv:2403.04321*, 2024.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*, 2022.

Shaolin Su, Vlad Hosu, Hanhe Lin, Yanning Zhang, and Dietmar Saupe. Koniq++ : Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects. In *The British Machine Vision Conference (BMVC)*, pages 1–12, 2021.

Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *CVPR*, 2020.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*, 2024.

Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Exploring the naturalness of ai-generated images. *arXiv preprint arXiv:2312.05476*, 2023a.

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*, 2023c.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

OpenAI. Hello gpt-4o, 2024. URL `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-05-13.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023.

Huggingface. Introducing idefics: An open reproduction of state-of-the-art visual language model, 2023. URL `https://huggingface.co/blog/idefics`.

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, 2024.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

Adept. Fuyu-8b: A multimodal architecture for ai agents, 2023. URL `https://www.adept.ai/blog/fuyu-8b`. Accessed: 2024-05-13.

dreamlike art. dreamlike-photoreal-2.0. `https://dreamlike.art`, 2023.

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, 2023b.

PlaygroundAI. playground-v2-1024px-aesthetic. `https://playground.com`, 2023.

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

Yatharth Gupta, Vishnu V. Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss. 2401.02677, 2024.

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.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation, 2023.

DeepFloyd. If-i-xl-v1.0. `https://www.deepfloyd.ai`, 2023.

A Appendix

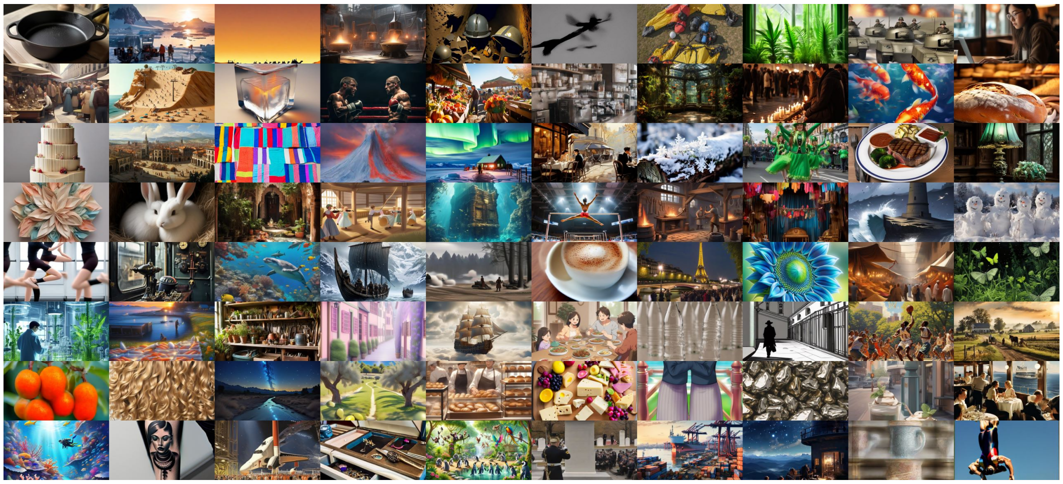


Figure 6: Overview of the AIGIs from **A-Bench**^{P1}.

A.1 AIGIs Collection

AIGI collection for A-Bench^{P1} To ensure that the AIGIs meet the specific subcategory requirements, we have gathered 2,000 manually-written prompts to serve as the textual foundation. Below, we provide examples of these prompts:

1. Basic Recognition -> Major Object Recognition: *An elaborate treehouse in a thick forest, with children playing inside, rope bridges connecting to other trees, and birds chirping around.*
2. Basic Recognition -> Minor Object Recognition: *A magical fairy ring in a moonlit forest, with tiny glowing fairies dancing and mystical plants all around.*
3. Bag-of-Words -> Attributes Awareness: *A delicate, frosty, crystal snowflake beside a warm, glowing, amber ember on a smooth, slate-gray stone.*
4. Bag-of-Words -> Nouns as Adjectives Awareness: *Shark-sleek submarine exploring ocean depths.*
5. Bag-of-Words -> Composition Identification: *A gamer's setup with consoles and controllers on a desk, multiple screens above, and game boxes and snacks partially obscured beneath the desk.*
6. Bag-of-Words -> Attributes Awareness: *Six logs in a woodpile, stacked so tightly that they seem to form a solid block.*
7. Outside Knowledge -> Specific Terms Recognition: *A barometer showing a rapid decrease in pressure.*
8. Outside Knowledge -> Contradiction Overcome: *A ship floating above the clouds, sails made of sunlight.*

Afterward, we use the collected prompts to create AIGIs. **15** text-to-image generation models are selected, which include: Dreamlike [dreamlike art, 2023], Pixart α Chen et al. [2023b], Playground v2 PlaygroundAI [2023], SD1.4 [Rombach et al., 2022b], SD1.5 [Rombach et al., 2022b], SDXL [Rombach et al., 2022b], SSD1B [Gupta et al., 2024], LCM Pixart [Luo et al., 2023], LCM SD1.5 [Luo et al., 2023], LCM SDXL [Luo et al., 2023], SDXL Turbo Sauer et al. [2023] DALLE2 [Ramesh et al., 2022], DALLE3 [Ramesh et al., 2022], IF [DeepFloyd, 2023], Midjourney v5.2 Holz [2023]. Finally, a total of 15$\times$2,000 =30,000 AIGIs are collected. To guarantee diversity, we randomly select 2,000 AIGIs, choosing one AIGI per prompt. Subsequently, we conduct a manual review of these AIGIs to remove any that failed to generate correctly or are unsuitable for annotation. This process results in the final set of AIGIs for **A-Bench**^{P1}, which can be overviewed in Fig. 6.

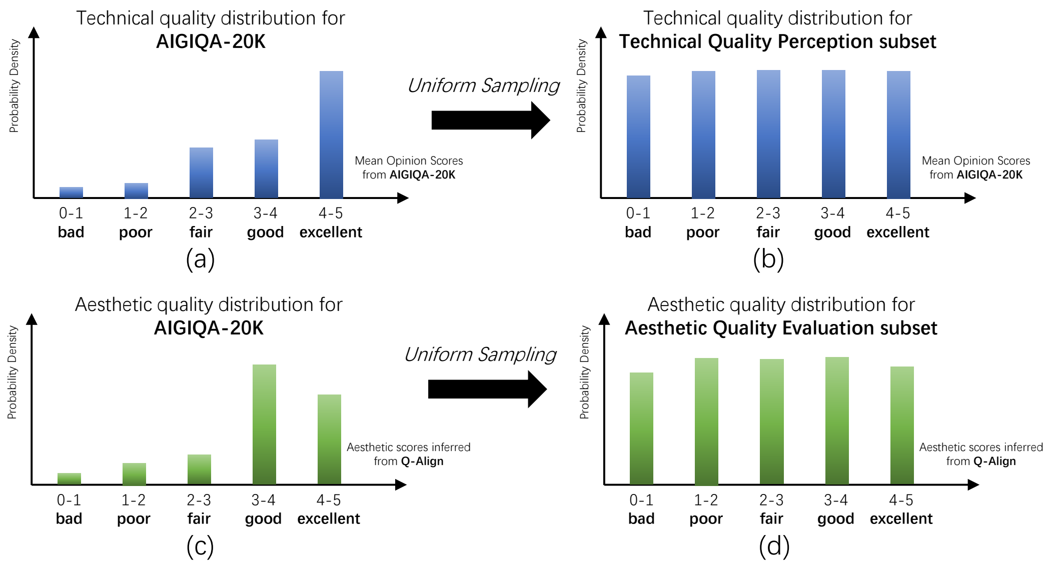


Figure 7: Illustration of the quality distribution transformation.

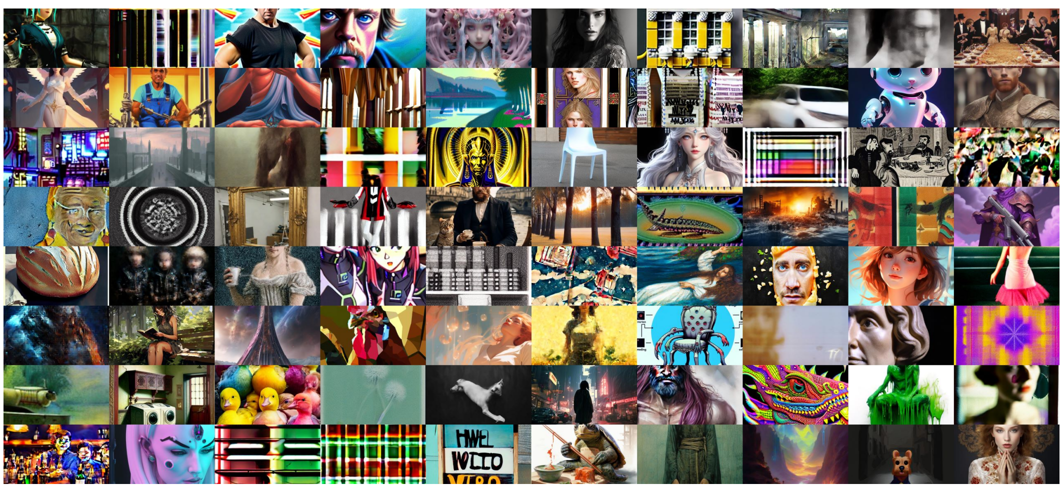


Figure 8: Overview of the AIGIs from **A-Bench**^{P2}.

AIGI collection for A-Bench^{P2} **A-Bench**^{P2} is designed for the quality evaluation of AIGIs. Consequently, it is essential to ensure that the collected AIGIs span a wide quality range to address various practical scenarios. For **Technical Quality**, we sample 500 AIGIs from the AIGIQA-20K dataset [Li et al., 2024] using a *uniform sampling* strategy. Specifically, each AIGI in the AIGIQA-20K dataset is assigned a mean opinion score (MOS) for technical quality. We apply uniform sampling to create more even distributions, as illustrated in Fig. 7. For **Aesthetic Quality**, in the absence of provided aesthetic scores, we utilize q-align [Wu et al., 2023c], an effective quality predictor, to infer the aesthetic values of AIGIs. Subsequently, we perform uniform sampling similarly to obtain 500 AIGIs for aesthetic evaluation. For **Generative Distortion**, we manually select 500 AIGIs exhibiting unexpected AIGI-specific distortions. It is important to note that there is no content overlap among the selected AIGIs, which can be overviewed in Fig. 8.

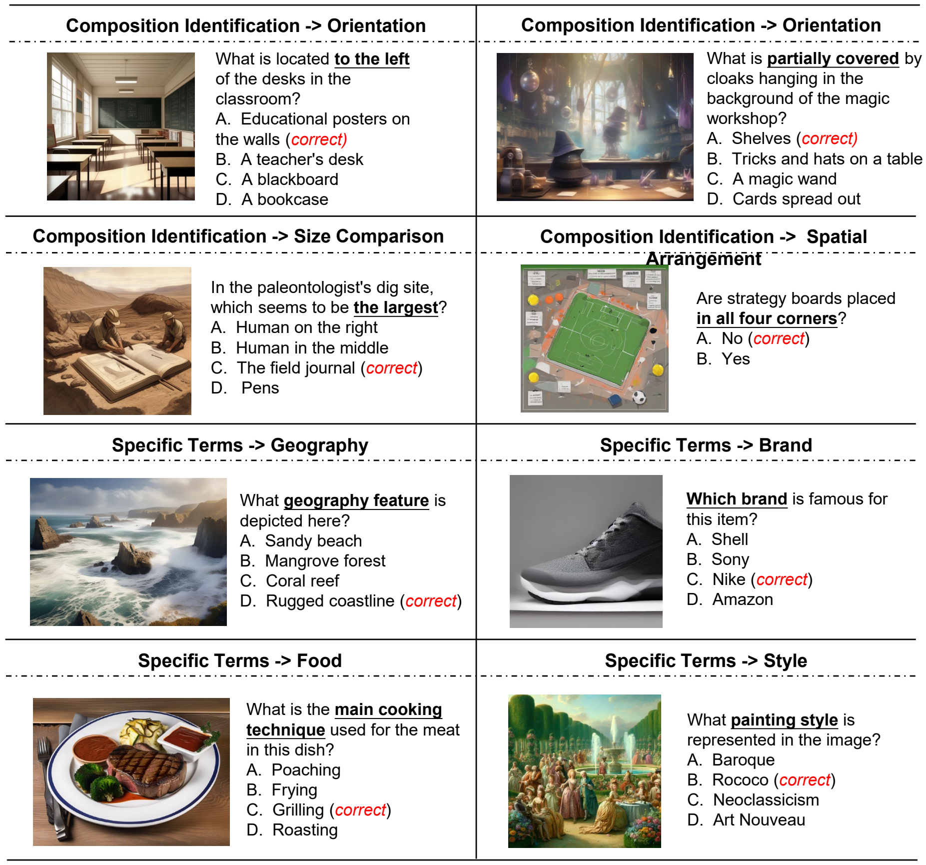


Figure 9: Some finer cases for the 'Bag-of-Words -> Composition Identification' and 'Outside Knowledge -> Specific Terms' subcategories.

A.2 Finer Explanation for some Subcategories

For certain subcategories that require additional clarification for better understanding, we provide detailed explanations here (the corresponding cases are shown in Fig. 9):

1. Bag-of-Words -> Nouns as Adjectives Awareness. The 'Noun as Adjectives' illustrates the use of nouns as adjectives to modify objects in AIGIs. Essentially, we aim for the descriptive effect, not for the nouns themselves to be visually represented in the AIGIs. For instance, as shown in Fig.4 row 2 column 2, when we describe a submarine as 'shark-sleek,' we do not intend to generate an image of an actual shark. This subcategory is designed to test whether LMMs can correctly identify such misunderstandings.

2. Bag-of-Words -> Composition Identification. We categorize composition into four distinct types: 1) Orientation, which assesses the ability to correctly determine the relative spatial positions of objects; 2) Occlusion, which involves evaluating the accuracy in discerning the overlapping relationships between objects; 3) Size Comparison, which tests the ability to accurately judge the size relationships among objects; and 4) Spatial Arrangement, which examines the ability to accurately assess the arrangement of objects within the AIGI.

3. Outside Knowledge -> Specific Terms. This subcategory covers many aspects, including geography, sports, science, materials, food, everyday life, creatures, brands, and styles. This primarily investigates whether it is possible for LMMs to infer and deduce specific knowledge within these fields based on the content of AIGIs such as identifying the exact location feature based on geographical attributes, deducing the brand from the characteristics of a product, recognizing the cooking technique of the food, etc.

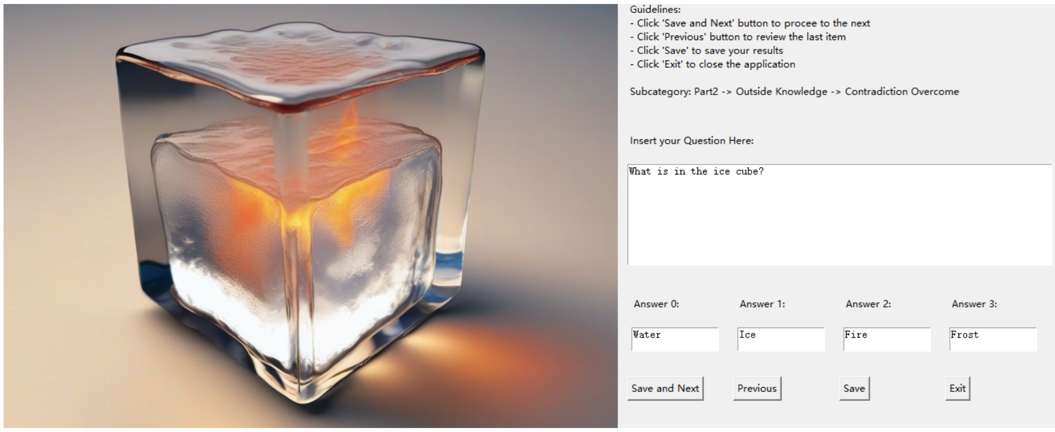


Figure 10: Illustration depicting the annotation interface, where experts are presented with the subcategory and are able to record their questions and answers.

A.3 Human Expert Annotation

A total of fifteen experts, each possessing professional skills and extensive experience in photography and AIGIs, participate in the subjective labeling experiment of **A-Bench**. All experts are informed that their annotation data will be publicly released, and they all agree to this arrangement. The hourly wage for each expert is approximately 12 US dollars, resulting in a total expense of about 2,400 US dollars for the whole subjective experiment.

The experiment takes place in a laboratory environment with standard indoor lighting. A Dell 4K monitor, supporting a resolution of 3840×2160, is used for displaying the interfaces. Screenshots of the interfaces can be referred to in Fig. 10. Each expert annotates up to 30 AIGIs per day to avoid fatigue, with every annotation carefully reviewed by at least three other experts before acceptance. This approach ensures the highest possible accuracy and rigor of the **A-Bench** labels, thereby enhancing the precision and meaningfulness of the performance testing capability of **A-Bench**.

A.4 GPT Evaluation for Choice Judgment

For some LMMs, the response to the question inquiry may vary. For example, given the correct answer *C. Blurry*' to the question *What is the most severe technical distortion of this image?*', LMMs may respond in different formats: *The image is blurry*', *There is blur in this image*', or '*low clarity*'. To address the impact of such variations on our evaluation, we've implemented a 5-round **voting** strategy [Wu et al., 2023a]. Under this strategy, we pose the same prompt, as defined in the templates, five times and determine the final outcome based on the majority of GPT's responses.

GPT Evaluation Prompt Template

#System: You are a helpful assistant that grades answers related to image perception. There are a lot of special terms or keywords related to image processing and photography.

#User: Assuming you are a grader, you will now be provided with a question [question] and a set of options [options] with option [options[0]] being the correct answer. Additionally, there will be an answer [answer] provided by a respondent. Please determine whether the respondentś answer is correct considering the context of the question. Even if the word choice is not completely the same, you can decide based on the given options and see whether the one in the answer is close enough to the given correct answer, The result is 1 if the answer is correct and else the result is 0. Please only provide the result in the following format: Result:

Example for GPT Evaluation

Question: Which is the most blurry part of this image?

Choices: ['The house on the left', 'The person in the middle', 'The background', 'The tree on the left']

LMM Answer:

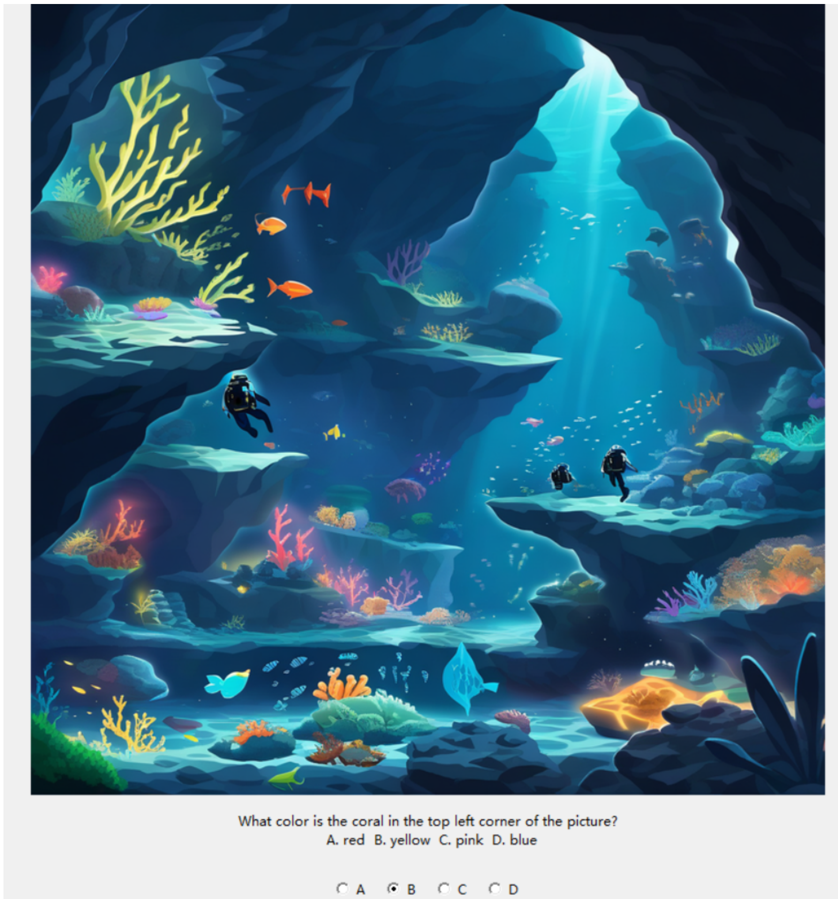


Figure 11: Illustration of the interface for the user-study.

The most blurry part in this image is the house to the left of the person.

5-Round GPT Answers:
[“Score: 1",“Score: 1",“Score: 1",“Score: 1",“Score: 1"]
$\rightarrow$ Final Correctness after Voting: ✓

A.5 User-study on A-Bench

To provide human performance on the **A-Bench**, we employ five ordinary people in a controlled laboratory setting. Initially, participants familiarize themselves with the tasks through exposure to similar cases. Subsequently, they select the appropriate responses for the questions posed in the A-Bench. The user-study interface is shown in Fig. 11.

A.6 LMM Experiment Details

The LMMs undergo testing in a zero-shot setting. Proprietary LMMs are evaluated via official APIs, whereas the *open-source* LMMs (with the exceptions of LLaVA-NeXT *Qwen-72B* and LLaVA-NeXT *Qwen-110B*) run on an NVIDIA RTX 6000 Ada with 48 GB of memory. The LLaVA-NeXT *Qwen-72B* and LLaVA-NeXT *Qwen-110B* are operated on 4 NVIDIA H100 with 320 GB of memory. All LMMs operate with default parameters, ensuring that the **A-Bench** results are readily reproducible.

A.7 Question Type Performance

We assess the performance disparity between *Yes-or-no* and *What* questions among LMMs. The *Yes-or-no* questions gauge the fundamental judgment capabilities of LMMs, whereas *What* questions

Table 3: Benchmark results on the question types. The best performance is marked in **bold** and the second performance is underlined for both proprietary and open-source LMMs respectively.

Categories	A-Bench^{P1}		A-Bench^{P2}		Overall	
LMM *(LLM)*	*Yes-or-no*↑	*What*↑	*Yes-or-no*↑	*What*↑	*Yes-or-no*↑	*What*↑
HUMAN (WORST)	91.21%	92.77%	89.45%	91.02%	91.23%	91.88%
HUMAN (BEST)	93.55%	94.25%	91.80%	92.64%	92.77%	93.39%
Proprietary LMMs:						
GEMINI 1.5 PRO	81.96%	**86.91%**	**74.08%**	**65.57%**	**76.50%**	**76.82%**
GPT-4V	82.37%	85.86%	71.11%	60.09%	75.51%	73.23%
GPT-4O	84.39%	85.76%	69.76%	65.15%	76.28%	75.81%
QWEN-VL-MAX	**86.70%**	84.02%	68.13%	64.60%	75.79%	74.91%
Open-source LMMs:						
CogVLM2-19B (*Llama3-8B*)	81.77%	83.26%	63.70%	58.65%	70.55%	71.61%
IDEFICS-2 (*Mistral-7B-Instruct-v0.2*)	78.32%	83.84%	63.87%	55.63%	68.91%	69.96%
DeepSeek-VL-7B	80.72%	82.00%	60.00%	47.15%	66.88%	66.48%
InternLM-XComposer2-VL (*InternLM2*)	82.08%	81.53%	**66.49%**	53.06%	70.90%	69.83%
LLaVA-NeXT (*Llama3-8B*)	81.17%	84.11%	52.10%	53.77%	63.89%	68.82%
LLaVA-NeXT (*Qwen-72B*)	**83.22%**	**84.31%**	57.91%	60.01%	70.22%	71.55%
LLaVA-NeXT (*Qwen-110B*)	82.99%	83.91%	59.78%	**62.87%**	**71.76%**	**73.05%**
mPLUG-Owl2 *(LLaMA-7B)*	74.92%	78.00%	56.97%	54.36%	64.38%	67.81%
LLaVA-v1.5 (*Vicuna-v1.5-7B*)	78.27%	82.74%	46.39%	47.97%	58.85%	66.21%
LLaVA-v1.5 (*Vicuna-v1.5-13B*)	79.51%	81.47%	47.23%	43.90%	61.41%	63.61%
CogVLM-17B (*Vicuna-v1.5-7B*)	76.77%	80.11%	55.13%	49.71%	64.33%	65.65%
Qwen-VL *(Qwen-7B)*	72.77%	80.95%	46.22%	44.02%	56.60%	63.39%
BakLLava (*Mistral-7B*)	71.01%	78.77%	42.11%	44.11%	55.61%	60.03%
Fuyu-8B (*Persimmon-8B*)	61.56%	64.22%	38.76%	41.66%	50.06%	52.31%

demand a more comprehensive understanding. According to the results in Table 3, it is observed that most LMMs perform better on *What* questions within **A-Bench**^{P1}, suggesting a proficiency in processing semantic content. Conversely, in **A-Bench**^{P2}, where LMMs generally show lesser performance, they exhibit limited in-depth perception, maintaining only basic evaluative capabilities without comprehensive understanding, leading to poorer performance on *What* questions. Interestingly, human performance consistently excels in *What* questions across both **A-Bench**^{P1} and **A-Bench**^{P2}, likely due to a broader range of options facilitating easier inference. However, human performance tends to be more balanced compared to LMMs, which may exhibit significant variance, such as IDEFICS-2, where there is over a 5% accuracy difference between question types, indicating less robustness.

A.8 Data Statement

The **A-Bench** dataset is released under the **CC BY 4.0** license. This includes all associated AIGIs, questions, and answer candidates. However, to prevent incorporation into the training sets of any LMMs, the correct answers remain confidential. We believe this precaution will ensure that **A-Bench** retains its long-term value as a benchmark for assessing AIGI evaluation capabilities.

A.9 Limitations and Social Impact

Limitations While **A-Bench** uses a diverse set of generative models and LMMs for evaluation, the choice and number of models might still limit the generalizability of the results. The performance of untested models or newer generative approaches might differ significantly. The rapid advancement in AI and generative models may quickly outpace the current setup of A-Bench, necessitating frequent updates or redesigns of the benchmark to stay relevant.

Social Impact By improving the evaluation metrics for AIGIs, **A-Bench** could lead to more reliable and trustworthy AI-generated content, which is crucial as these technologies increasingly intersect with areas like media, entertainment, and education. Moreover, improved benchmarks and evaluation methods can drive industry standards, potentially lowering the barrier to entry for smaller developers and promoting innovation through clearer performance targets.