MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

[1]Xiang Yue*†, [2]Yuansheng Ni*, [3]Kai Zhang*, [4]Tianyu Zheng*,
[3]Ruoqi Liu, [2]Ge Zhang, [3]Samuel Stevens, [2]Dongfu Jiang, [2]Weiming Ren, [4]Yuxuan Sun,
[2]Cong Wei, [3]Botao Yu, [5]Ruibin Yuan, [2]Renliang Sun, [7]Ming Yin,
[3]Boyuan Zheng, [4]Zhenzhu Yang, [6]Yibo Liu, [4]Wenhao Huang,
[3]Huan Sun*, [3]Yu Su*†, [2]Wenhu Chen*†

[1]IN.AI Research, [2]University of Waterloo, [3]The Ohio State University, [4]Independent,
[5]Carnegie Mellon University, [6]University of Victoria, [7]Princeton University

https://mmmu-benchmark.github.io/

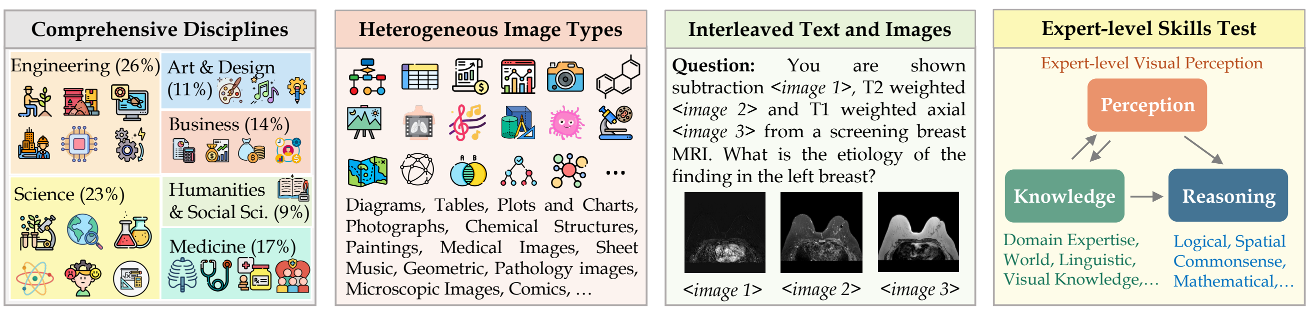


Figure 1. Overview of the MMMU dataset. MMMU presents four challenges: 1) **comprehensiveness**: 11.5K college-level problems across six broad disciplines and 30 college subjects; 2) highly **heterogeneous** image types; 3) **interleaved** text and images; 4) **expert-level** perception and reasoning rooted in deep subject knowledge.

Abstract

We introduce MMMU: a new benchmark designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. MMMU includes 11.5K meticulously collected multimodal questions from college exams, quizzes, and textbooks, covering six core disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering. These questions span 30 subjects and 183 subfields, comprising 30 highly heterogeneous image types, such as charts, diagrams, maps, tables, music sheets, and chemical structures. Unlike existing benchmarks, MMMU focuses on advanced perception and reasoning with domain-specific knowledge, challenging models to perform tasks akin to those faced by experts. The evaluation of 28 open-source LMMs as well as the proprietary GPT-4V(ision) and Gemini highlights the substantial challenges posed by MMMU. Even the advanced GPT-4V and Gemini Ultra only achieve accuracies of 56% and 59% respectively, indicating significant room for improvement. We believe MMMU will stimulate the community to build next-generation multimodal foundation models towards expert artificial general intelligence.

1. Introduction

Rapid advances in large language models (LLMs) [13, 54, 67] have sparked broad discussions on the controversial concept of artificial general intelligence (AGI), often used to describe AI systems that perform on par or surpass humans at most tasks [1, 7, 21, 29, 49, 52]. Candid and constructive discussions on AGI have been challenging due to a lack of shared operationalizable definitions. In an attempt to remedy this, Morris et al. [52] propose a leveled taxonomy for AGI that centers around both *generality* (or breadth) and *performance* (or depth). In the suggested taxonomy, Level 3, or *Expert AGI*, marks a critical milestone. It denotes an

*Core Contributors. See the Author Contribution Statement for details.
†✉: {yue.149,su.809}@osu.edu; wenhuchen@uwaterloo.ca

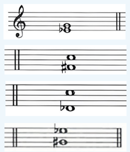
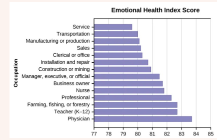
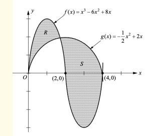
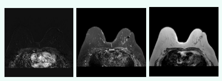
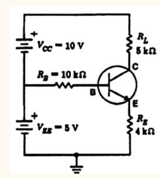
Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? **Options:** (A) Major third <*image 1*> (B) Diminished fifth <*image 2*> **(C) Minor seventh <*image 3*>** (D) Diminished sixth <*image 4*> 	**Question:** ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82? **Options:** (A) 0 (B) 0.2142 **(C) 0.3571** (D) 0.5 	**Question:** <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. **Options:** **(A)** $\int_0^{1.5}[f(x)-g(x)]dx$ (B) $\int_0^{1.5}[g(x)-f(x)]dx$ (C) $\int_0^{2}[f(x)-g(x)]dx$ (D) $\int_0^{2}[g(x)-x(x)]dx$ 
Subject: Music; **Subfield:** Music; **Image Type:** Sheet Music; **Difficulty:** Medium	**Subject:** Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts; **Difficulty:** Medium	**Subject:** Math; **Subfield:** Calculus; **Image Type:** Mathematical Notations; **Difficulty:** Easy
Health & Medicine	**Humanities & Social Science**	**Tech & Engineering**
Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast? **Options:** (A) Susceptibility artifact (B) Hematoma **(C) Fat necrosis** (D) Silicone granuloma 	**Question:** In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image 1*> **Option:** (A) Oppressor (B) Imperialist **(C) Savior** (D) Isolationist 	**Question:** Find the VCE for the circuit shown in <*image 1*>. Neglect VBE **Answer:** 3.75 **Explanation:** ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V 
Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty:** Hard	**Subject:** History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons; **Difficulty:** Easy	**Subject:** Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams; **Difficulty:** Hard

Figure 2. Sampled MMMU examples from each discipline. The questions and images need expert-level knowledge to understand and reason.

AI system that reaches "at least 90th percentile of skilled adults" in a broad range of tasks, thus starting to achieve "the substitution threshold for machine intelligence in lieu of human labor" for many industries, leading to significant risks of job displacement and economic disruption. Therefore, it is of both intellectual and societal importance to closely monitor the progress towards Expert AGI.

How to create benchmarks for measuring Expert AGI? Since the definition is based on comparison with *skilled adults*, a natural starting point is college-level exams for different disciplines, because those are designed to evaluate *skilled adults* specialized in each discipline. This strategy has been successfully adopted in benchmarks such as MMLU [24] and AGIEval [85], but only text-based questions are considered, while human experts are capable of solving multimodal problems. Meanwhile, large multimodal models (LMMs) that can understand both text and images have been making a major stride towards more general AI [9, 16, 32, 40, 73]. These LMMs have consistently excelled in existing multimodal benchmarks [3, 23, 30, 36, 43, 62, 76, 79]. For instance, CogVLM [70] achieves 85% on VQA-v2 [23], 92% on ScienceQA-IMG [46], and 93% on RefCOCO [28]. However, most existing multimodal benchmarks focus on commonsense/daily knowledge rather than expert-level domain knowledge and advanced reasoning. The closest one to our goal is ScienceQA [46]. While it covers diverse disciplines (**breadth**), the majority of the questions are at the elementary to the middle school level, thus falling short in **depth** for benchmarking Expert AGI.

To this end, we introduce MMMU: a comprehensive benchmark designed for college-level multi-discipline multimodal understanding and reasoning. It features problems sourced from college exams, quizzes, and textbooks spanning six common disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering. MMMU consists of **11.5K** carefully selected multimodal questions, which cover **30** diverse subjects and 183 subfields, thus meeting the **breadth** goal. Moreover, many problems within MMMU require expert-level reasoning, such as applying "Fourier Transform" or "Equilibrium Theory" to derive the solution, thus meeting the **depth** goal. MMMU also presents two unique challenges absent in current benchmarks (Figure 1). Firstly, it covers diverse image formats, from visual scenes like photographs and paintings to diagrams and tables, testing the perceptual capabilities of LMMs. Secondly, MMMU features interleaved text-image inputs. A model needs to jointly understand the images and text, which often requires recalling deep subject knowledge, and conducting complex reasoning based on the understanding and knowledge to reach a solution.

We evaluate 28 open-source LMMs as well as the advanced proprietary LMMs such as GPT-4V(ision) [55] on MMMU. Our key findings are summarized as follows:

- MMMU presents significant challenges; notably, GPT-4V only achieves an accuracy of 55.7%, indicating substantial room for improvement.
- There is a pronounced disparity in performance between open-source LMMs and GPT-4V. The highest-performing

open-source models, such as BLIP2-FLAN-T5-XXL and LLaVA-1.5, achieve approximately 34% in accuracy.
- LLMs augmented with optical character recognition (OCR) or generated captions do not see notable improvement, indicating that MMMU necessitates deeper joint interpretation of images and text.
- In disciplines such as Art & Design and Humanities & Social Science, where visual data is less complex, models exhibit higher performance. In contrast, Business, Science, Health & Medicine, and Tech & Engineering, which present more complex visual data and require intricate reasoning, see relatively lower model performance.
- Our error analysis on 150 error cases of GPT-4V reveals that 35% of errors are perceptual, 29% stem from a lack of knowledge, and 26% are due to flaws in the reasoning process. These findings underscore the challenges of the MMMU benchmark and point towards areas needing further research and model enhancement.

Our aim with MMMU is to push the boundaries of what LMMs can achieve. We believe it will prove instrumental in developing next-generation multimodal foundation models and monitoring the progress towards Expert AGI. We shall caution that MMMU is not a *sufficient* test for Expert AGI, as per the definition [52], because there lacks a direct mapping between performance on MMMU and "90th percentile of skilled adults," nor are college exams the only tasks an AGI shall tackle. However, we believe it should be *necessary* for an Expert AGI to achieve strong performance on MMMU to demonstrate their broad and deep subject knowledge as well as expert-level understanding and reasoning capabilities.

2. Related Work

Multimodal Pre-Training. In recent years, rapid progress has been made in multimodal pre-training, which aims to jointly encode vision and language in a fusion model. LXMERT [64], UNITER [10], VinVL [80], Oscar [34], VilBert [45], and VLP [86] are among the earliest work to train universal vision-language models to tackle many multimodal tasks. This work relies on pre-trained visual representations like Faster RCNN features [61] to minimize the training sample complexity. Later on, CLIP [60], ALIGN [27], SimVLM [71], CoCa [78], Flamingo [2], BLIP-2 [32], and Fuyu [6] (inter alia) have been proposed to train visual representation using ViT [18] from scratch with massive amount of web data. These models have achieved great success on existing VQA and captioning tasks, which require less knowledge and reasoning.

Multimodal Instruction Tuning. Inspired by open-source instruction-tuned LLMs like FLAN-T5 [14] and Vicuna [12], models like LLaVA [40, 41] and MiniGPT-4 [87] utilized open-source resources, to improve the instruction-following capabilities of LMMs. The evolutionary trajectory of LMMs has also led to subsequent advancements aimed at improving the quantity and quality of visual instruction data. Models such as LLaMA-Adapter [20, 81], mPlug-OWL [74, 75], SVIT [82], LRV-Instruction [39], and InstructBLIP [16] exemplify these developments. Another pivotal aspect of LMM research revolves around multimodal in-context learning and the management of interleaved text and image examples. This area has been explored in depth by models such as Flamingo [2] and OpenFlamingo [4], Otter [31], M3IT [33], MetaVL [51], Sparkles [25], and MMICL [83]. These models have significantly contributed to the ongoing advancements in multimodal training and instruction-following capabilities.

LMM Benchmarks. With the surge of multi-modal pre-training and instruction tuning, the prior single-task evaluation benchmarks like VQA [3, 23], OK-VQA [48], MSCOCO [36], GQA [26], etc., have become insufficient to holistically evaluate LMMs' general multimodal perception and reasoning abilities. Therefore, numerous all-round benchmarks have been established to assess different facets of LMMs. These benchmarks cover a wide spectrum of specific skills of LMMs, from Optical Character Recognition (OCR) as seen in the study by [44], to adversarial robustness [84] and hallucination [15, 38], e.g., POPE [35] and HaELM [69]. More holistic evaluations have been conducted as well, such as LAMM [76], LVLM-eHub [72], SEED [30], MMBench [43], and MM-Vet [79]. These benchmarks still largely focus on relatively basic perception abilities without requiring expert-level domain knowledge and deliberate reasoning. More recently, MathVista [47] presents a collection of visually challenging questions; however, its scope is limited exclusively to the mathematical domain. MMMU is highly different from these benchmarks by collecting more difficult expert-level problems that cover 30 different subjects and require nuanced perception, recalling domain-specific knowledge to perform step-by-step reasoning to derive the solution. In line with the motivation of our study, concurrently, GAIA [49] introduces 466 questions that test fundamental abilities of models such as reasoning, multimodality handling, or tool use.

3. The MMMU Benchmark

3.1. Overview of MMMU

We introduce the Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark, a novel benchmark meticulously curated to assess the expert-level multimodal understanding capability of foundation models across a broad scope of tasks. Covering 30 subjects across 6 disciplines, including Art, Business, Health & Medicine, Science, Humanities & Social Science, and Tech & Engineering, and over 183 subfields. The detailed subject coverage and statistics are detailed in Figure 7. The questions in our benchmark were manually collected by a team of

Statistics	Number
Total Questions	11550
Total Disciplines/Subjects/Subfields	6/30/183
Image Types	30
Dev:Validation:Test	150:900:10500
Difficulties (Easy: Medium: Hard)	28%:45%:27%
Multiple-choice Questions	10861 (94.03%)
Open Questions	689 (5.97%)
Questions with an Explanation	2035 (17.62%)
Image in the Question	11264 (97.52%)
* Images at the beginning	2006 (17.81%)
* Images in the middle	4159 (36.92%)
* Images at the end	5679 (50.42%)
Image in Options	389 (3.37%)
Example with Multiple Images	854 (7.39%)
Average question length	59.33
Average option length	9.17
Average explanation length	107.92

Table 1. Key statistics of the MMMU benchmark.

50 college students (including coauthors) from various disciplines and subjects, drawing from online sources, textbooks, and lecture materials.

MMMU, constituting 11.5K questions, is divided into a few-shot development set, a validation set, and a test set. The few-shot development set includes 5 questions per subject, and the validation set, useful for hyperparameter selection, contains approximately 900 questions, while the test set comprises 10.5K questions. MMMU is designed to measure three essential skills in LMMs: perception, knowledge, and reasoning. Our aim is to evaluate how well these models can not only perceive and understand information across different modalities but also apply reasoning with subject-specific knowledge to derive the solution.

Our MMMU benchmark introduces four key challenges to multimodal foundation models, as detailed in Figure 1. Among these, we particularly highlight the challenge stemming from the requirement for both expert-level visual perceptual abilities and deliberate reasoning with subject-specific knowledge. This challenge is vividly illustrated through our tasks, which not only demand the processing of various heterogeneous image types but also necessitate a model's adeptness in using domain-specific knowledge to deeply understand both the text and images and to reason. This goes significantly beyond basic visual perception, calling for an advanced approach that integrates advanced multimodal analysis with domain-specific knowledge.

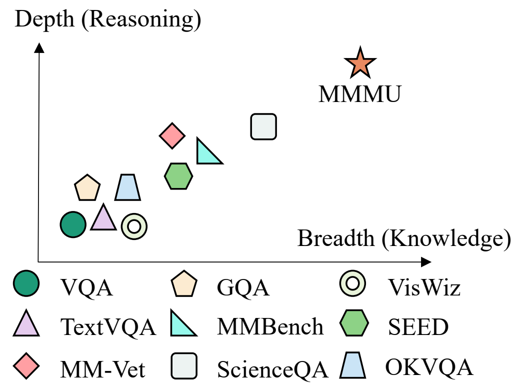
3.2. Data Curation Process

Data Collection. Our benchmark collection takes three stages. Firstly, we go through the common university majors to decide what subjects should be included in our benchmark. The selection is based on the principle that visual inputs should be commonly adopted in the subjects to provide valuable information. Through this principle, we rule out a few subjects like law and linguistics because it is difficult to find enough relevant multimodal problems in these subjects. Consequently, we select 30 subjects from six different disciplines. In the second stage, we recruit over 50 university students, including co-authors, specializing in these majors as annotators to assist in question collection. They collect multimodal questions from major textbooks and online resources, creating new questions based on their expertise where necessary. The annotators are instructed to adhere to copyright and license regulations, avoiding data from sites prohibiting copy and redistribution. Given the arising data contamination concerns of foundation models, the annotators are advised to select questions without immediately available answers, such as those with answers in separate documents or at the end of textbooks. This process results in a diverse collection of 13K questions from various sources. The detailed annotation protocol is in Appendix A.

Data Quality Control. To further control the quality of our data, we perform two steps of data cleaning. In the first stage, lexical overlap and source URL similarity are employed to identify potential duplicate problems. These suspected duplicates were then reviewed by the authors to identify and eliminate any duplications. The second stage involves distributing the problems among different co-authors for format and typo checking. This step requires authors to ensure adherence to a standardized format, undertaking necessary corrections where deviations are found. In the third and final stage, the authors categorize the problems into four difficulty levels: very easy, easy, medium, and hard. Approximately 10% of the problems, classified as very easy and not aligning with our design criteria due to their simplistic nature, are excluded from the benchmark. This rigorous process plays a crucial role in maintaining the quality and difficulty of the problem set.

3.3. Comparisons with Existing Benchmarks

To further distinguish the difference between MMMU and other existing ones, we elaborate the benchmark details in Figure 3. From the *breadth* perspective, the prior benchmarks are heavily focused on daily knowledge and common sense. The covered image format is also limited. Our benchmark aims to cover college-level knowledge with 30 image formats including diagrams, tables, charts, chemical structures, photos, paintings, geometric shapes, music sheets, medical images, etc. In the *depth* aspect, the previous benchmarks normally require commonsense knowledge or simple physical or temporal reasoning. In contrast, our benchmark requires deliberate reasoning with college-level subject knowledge.



Dataset	Size	Images	Format	Source	Answer
VQA	>1M	V	I+T	Annotated	Open
GQA	>1M	V	I+T	Synthesized	Open
VisWiz	32K	V	I+T	Annotated	Open
TextVQA	45K	OC	I+T	Annotated	MC
OKVQA	14K	V+OC	I+T	Annotated	Open
SEED	19K	V+OC	I+T	Annotated	MC
MMBench	3K	V+OC	I+T	Repurposed	MC
MM-Vet	0.2K	V+OC	I+T	Annotated	Open
ScienceQA	6K	5 Types	I+T	Textbooks	MC
MMMU	11.5K	30 Types	Interleaved	Textbooks, Internet, Annotated	Open / MC

Figure 3. The comparison between MMMU and other existing benchmarks. MMMU excels in both its breadth to cover a wide range of disciplines and its depth to test LMMs' reasoning abilities. In the image format, V means visual input, OC means optical characters, MC means multi-choice. Repurposed means the benchmark is a compilation of prior datasets.

4. Experiments

We evaluate various models including LLMs and LMMs. In each type, we consider both closed- and open-source models. Our evaluation is conducted under a *zero-shot* setting to assess the capability of models to generate accurate answers without fine-tuning or few-shot demonstrations on our benchmark. For all models, we use the default prompt provided by each model for multi-choice or open QA, if available. If models do not provide prompts for task types in MMMU, we conduct prompt engineering on the validation set and use the most effective prompt for the zero-shot setup in the main experiments. We also report the few-shot results of some selected models in the Appendix. All experiments are conducted with NVIDIA A100 GPUs.

4.1. Baselines

LMMs. We consider various large multimodal models. By default, for each model family, we use the latest, largest, and best-performing available checkpoint to date. *(i)* Kosmos2 [57] is pre-trained to ground fine-grained visual objects with texts and to follow instructions. With only 1.6B model size, Kosmos2 is able to achieve comparable or better performance with Flamingo-9B [2] on VQA and captioning tasks. *(ii)* LLaMA-Adapter2 [20] fine-tunes Llama [67] in a parameter-efficient way and utilizes visual encoder CLIP [60] and modular experts such as Optical Character Recognition (OCR) to capture more image information for later better visual understanding. *(iii)* BLIP-2 [32] introduces light-weight learnable visual queries to bridge the frozen CLIP ViT [60] and FLAN-T5 [14]. *(iv)* Starting from the parameters from BLIP-2, InstructBLIP [16] is further fine-tuned with visual instruction tuning data for better zero-shot generalization capabilities. *(v)* LLaVA-1.5 [40] linearly projects the visual embedding into word embedding space of Vicuna [12], thus equipping the LLM with visual abilities. *(vi)* As an open-source alternative to Flamingo [2], OpenFlamingo [4] has close performance on most vision-language tasks. *(vii)* CogVLM [70] concatenates image and text in the input embedding space and adds trainable visual layers in textual Transformer blocks to deeply align two modalities. It has been reported to achieve very promising performance on existing VQA benchmarks recently. *(viii)* Fuyu [6] projects the patches of the input image into text embedding space. *(ix)* Qwen-VL [5] introduces a set of trainable query embeddings and single-layer cross-attention module to bridge the modalities, supporting interleaved image-text input. *(x)* Otter [31] is fine-tuned with diverse instruction-tuning data and able to perform in-context learning. *(xi)* MiniGPT-4 [87] is built upon Vicuna [12] and designs a linear modality projection layer for visual understanding abilities. *(xii)* mPLUG-Owl2 [75] designs a modality-adaptive module to unify vision and language while preserving their distinct properties of them.

Text-only LLMs. For text-only LLMs, we consider the most capable ones including GPT-4 and several open-source LLMs, Llama2-7B [67], FLAN-T5-XXL and Vicuna-13B, which are adopted as the text encoder or decoder in the selected LMMs. To determine if an external image-to-text tool can enhance these LLMs' performance on MMMU, we deploy OCR by MMOCR[1] or captioning by LLaVA-1.5 to provide the recognized text information to text-only LLMs.

Human Experts. We involve 90 college senior students, selected to represent a wide range of experts in the corresponding 30 subjects (3 student experts per subject). These students were tasked with completing the 30 questions in their corresponding subjects (900 validation questions in total). The students were allowed to consult their textbooks

[1] https://github.com/open-mmlab/mmocr

	Validation Overall (900)	Test Overall (10,500)	Art & Design (1,163)	Business (1,428)	Science (2,426)	Health & Medicine (1,752)	Human. & Social Sci. (947)	Tech & Eng. (2,784)
Random Choice	22.1	23.9	24.1	24.9	21.6	25.3	22.8	24.8
Frequent Choice	26.8	25.8	26.7	28.4	24.0	24.4	25.2	26.5
Expert (Worst)	76.2	-	-	-	-	-	-	-
Expert (Medium)	82.6	-	-	-	-	-	-	-
Expert (Best)	88.6	-	-	-	-	-	-	-
Large Multimodal Models (LMMs): Text + Image as Input								
OpenFlamingo2-9B [4]	28.7	26.3	31.7	23.5	26.3	26.3	27.9	25.1
Kosmos2 [57]	24.4	26.6	28.8	23.7	26.6	27.2	26.3	26.8
Adept Fuyu-8B [6]	27.9	27.4	29.9	27.0	25.6	27.0	32.5	26.4
MiniGPT4-Vicuna-13B [87]	26.8	27.6	30.2	27.0	26.2	26.9	30.9	27.2
LLaMA-Adapter2-7B [81]	29.8	27.7	35.2	25.4	25.6	30.0	29.1	25.7
CogVLM [70]	32.1	30.1	38.0	25.6	25.1	31.2	41.5	28.9
Qwen-VL-7B-Chat [5]	35.9	32.9	47.7	29.8	25.6	33.6	45.3	30.2
LLaVA-1.5-13B [40]	36.4	33.6	49.8	28.2	25.9	34.9	54.7	28.3
InstructBLIP-T5-XXL [16]	35.7	33.8	48.5	30.6	27.6	33.6	49.8	29.4
BLIP-2 FLAN-T5-XXL [32]	35.4	34.0	49.2	28.6	27.3	33.7	51.5	30.4
Emu2-Chat* [63]	36.3	34.1	50.6	27.7	28.0	32.4	50.3	31.3
Yi-VL-6B* [77]	39.1	37.8	53.4	30.3	30.0	39.3	58.5	34.1
Yi-VL-34B* [77]	45.9	41.6	56.1	33.3	32.9	45.9	66.5	36.0
LLaVA-1.6-34B* [42]	51.1	44.7	58.6	**39.9**	36.0	**51.2**	**70.2**	36.3
InternVL-Chat-V1.2* [11]	**51.6**	**46.2**	**62.5**	37.6	**37.9**	49.7	70.1	**40.8**
Adept Fuyu-Heavy* [19]	48.3	-	-	-	-	-	-	-
Qwen-VL-MAX* [59]	51.4	46.8	64.2	39.8	36.3	52.5	70.4	40.7
GPT-4V(ision) (Playground) [55]	56.8	**55.7**	**65.3**	**64.3**	**48.4**	**63.5**	**76.3**	**41.7**
Claude 3 Opus* [65]	**59.4**	-	-	-	-	-	-	-
Gemini Ultra* [22]	**59.4**	-	-	-	-	-	-	-
Large Language Models (LLMs): Only Text as Input								
Llama2 7B [68]	30.1	28.7	30.7	27.2	26.7	27.7	32.6	29.8
FLAN-T5-XXL [14]	32.1	31.2	36.8	**28.9**	26.7	32.8	44.8	28.3
+ OCR	34.7	**31.9**	36.2	28.8	26.2	32.6	**50.5**	**29.7**
+ LLaVA Caption	**34.8**	**31.9**	**38.4**	27.8	**27.0**	**33.2**	49.9	28.7
Vicuna-13B [12]	33.3	31.0	35.1	**30.1**	24.7	31.4	44.8	30.1
+ OCR	**35.4**	31.9	37.1	28.6	**26.5**	32.0	49.3	30.0
+ LLaVA Caption	33.9	**32.7**	**42.0**	26.8	26.2	**33.4**	**49.4**	**31.4**
GPT-4 Text [54]	34.9	33.8	32.9	28.5	30.6	41.3	53.0	28.4

Table 2. Selected results of different models on the MMMU **validation** and **test set**. Besides reporting the performance of LMMs, we additionally add text-only LLM baselines. The best-performing model in each category is **in-bold**, and the second best is underlined. *: results provided by the authors. Due to the page limit, we show other models' results in Appendix Table 4. The live-updating leaderboard is available at: https://mmmu-benchmark.github.io/#leaderboard

but were prohibited from searching the Internet for answers.

Evaluation. We adopt micro-averaged accuracy as the evaluation metric. For both open and multiple-choice questions, we design systematic, rule-based evaluation pipelines. Specifically, to mitigate the potential influence of any intermediate generations (e.g., reasoning steps, calculations) in the long response, we construct robust regular expressions and develop response-processing workflows. These are employed to extract key phrases, such as numbers and conclusion phrases, from the long responses for accurate answer matching. If there is no valid answer in the model's response, we perform random selection as a remedy for multiple-choice questions or consider the response incorrect for open questions. For reference, we add Random Choice and Frequent Choice baselines: the former randomly selects an option, while the latter selects the most frequent option within each specific subject of the validation set, based on its frequency of occurrence in that subject.

4.2. Main Results

In this section, we present a comprehensive comparison of different LLMs and LMMs using the MMMU benchmark, detailed in Table 2. We summarize our key findings as follows:
Challenging Nature of MMMU: The benchmark poses significant challenges to current models. The Best human expert achieves a validation accuracy of 88.6%, significantly outperforming all the models reported in the table. This demonstrates the still-existing gap between human expertise and the performance of current models on the MMMU benchmark. This reflects the benchmark's rigorous standards.

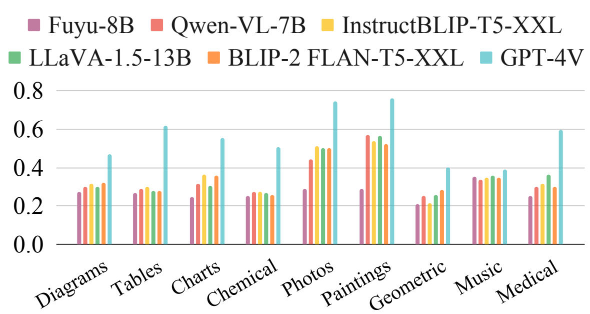


Figure 4. Performance of models on different types of images.

Disparity between Open-source Models and Closed-source models: Leading open-source models (as the paper submission) such as BLIP2-FLAN-T5-XXL and LLaVA-1.5 reach an accuracy level of approximately 34%, which is significantly lower than GPT-4V. However, it is exciting to see that open-source models have made significant strides in performance. For example, LLaVA-1.6-34B and InternVL-Chat-V1.2 achieve test accuracies of 44.7% and 46.2%, respectively, narrowing the gap with proprietary models.

Effectiveness of OCR and Captioning Enhancements: The application of OCR and captioning technologies does not yield a significant improvement in the performance of text-only LMMs. This finding suggests that the MMMU benchmark requires models that can effectively interpret and integrate both textual and visual information, underscoring the complexity of the multimodal tasks it presents.

Model Performance across Different Disciplines: In disciplines such as Art & Design and Humanities & Social Sciences, where the images tends to be more 'natural' and questions involve relatively less reasoning, models demonstrate relatively higher performance. Conversely, in fields like Science, Health & Medicine, and Technology & Engineering, where tasks often involve intricate perception and complex reasoning, models exhibit lower performance.

The MMMU benchmark underscores both the progress and the challenges in multimodal understanding and reasoning. While GPT-4V leads in performance, the overall results indicate substantial room for improvement, especially in domains with complex visual input and heavy reasoning with subject knowledge.

4.3. Analysis on Images Types and Difficulties

Different Image Types. We compare the performance of various models across top frequent image types in Figure 4. Across all types, GPT-4V consistently outperforms the other models by a huge margin. Open-source models demonstrate relatively strong performance in categories like Photos and Paintings, which are more frequently seen during training. However, for less common image categories like Geometric shapes, Music sheets and Chemical structures, all models obtain very low scores (some are close to random guesses). This indicates that the existing models are generalizing poorly towards these image types.

Models	Easy (2946)	Medium (4917)	Hard (2637)	Overall (10500)
Fuyu-8B [6]	28.9	27.0	26.4	27.4
Qwen-VL-7B [5]	39.4	31.9	27.6	32.9
LLaVA-1.5-13B [40]	41.3	32.7	26.7	33.6
InstructBLIP-T5-XXL [16]	40.3	32.3	29.4	33.8
BLIP-2 FLAN-T5-XXL [32]	41.0	32.7	28.5	34.0
GPT-4V [55]	76.1	55.6	31.2	55.7

Table 3. Result decomposition across question difficulty levels.

Different Difficulty Levels. Table 3 compares the performance of selected models across three difficulty levels. GPT-4V demonstrates a significantly higher proficiency, with a success rate of 76.1%, compared to open-source models in the "Easy" category. When it comes to the "Medium" category, while the gap narrows, GPT-4V still leads at 55.6%. The further diminishing performance gap in the "Hard" category across models indicates that as the complexity of tasks increases, the advantage of more advanced models like GPT-4V almost disappears. This might reflect a current limitation in handling expert-level challenging queries even for the most advanced models.

5. Error Analysis and Future Work

In this section, we delve into the analysis of errors by GPT-4V, a pivotal aspect for understanding its operational capabilities and limitations. This analysis serves not only to identify the model's current shortcomings but also to guide future enhancements in its design and training. We meticulously examine 150 randomly sampled error instances from GPT-4V's predictions. These instances are analyzed by expert annotators who identify the *root causes of mispredictions* based on their knowledge and the golden explanations if available. The distribution of these errors is illustrated in Figure 5, and a selection of 100 notable cases, along with detailed analyses, is included in the Appendix.

Perceptual Errors (35%): Perceptual errors, forming the bulk of the inaccuracies in the GPT-4V model, are categorized into two types: basic perceptual errors and domain-specific perceptual errors. Basic perceptual errors, as depicted in Figure 6, occur when the model accurately processes and understands the given information but fails in elementary visual interpretation, such as misjudging the sequence described as "from left to right, top to bottom." On the other hand, domain-specific perceptual errors occur due to the lack of knowledge. As we analyze the root cause, we classify such errors as lack of knowledge (see analysis below). Additionally, GPT-4V often exhibits a bias towards

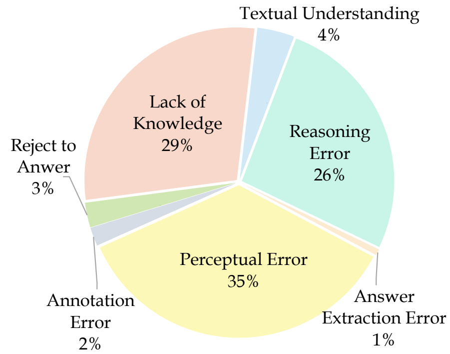


Figure 5. Error distribution over 150 annotated GPT-4V errors.

text, prioritizing textual information over visual inputs, a trend noted in recent studies [15]. A prominent example is in Figure 68, where the model incorrectly prioritizes its text-based interpretation of "imperialism" over the visual narrative in a cartoon depicting the United States as a "Savior." This underscores the need for a more balanced approach to multimodal interpretation.

Lack of Knowledge (29%): A fundamental root cause of 'domain-specific' perceptual errors in the GPT-4V model, as previously discussed, is the lack of specialized knowledge. This deficiency is exemplified in the Computer Science context illustrated in Appendix Figure 84, where the model identifies visual elements such as double circles but fails to interpret them accurately within the domain-specific context, such as their representation of an 'accept state' in Deterministic Finite Automata. Similarly, a deficit in specialized knowledge can lead to flawed reasoning, as demonstrated in the medical example in Appendix Figure 55. These instances underscore the necessity of enriching the training datasets of foundation models with a diverse range of domain-specific knowledge to improve their accuracy and general applicability in various specialized fields.

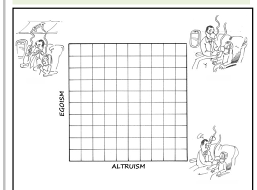
Reasoning Errors (26%): Flawed reasoning emerges as another significant cause of errors. In instances where the model correctly interprets text and images and recalls relevant knowledge, it still often fails to apply logical and mathematical reasoning skills effectively to derive accurate inferences. A notable instance of this can be observed in Appendix Figure 46, where the model neglects an essential step in a mathematical reasoning process, leading to an incorrect conclusion. Enhancing the model's reasoning capability is critical to address these shortcomings.

Other Errors: The remaining errors include Textual Understanding Error (6%), Rejection to Answer (3%), Annotation Error (2%), and Answer Extraction Error (1%). These errors are attributed to various factors such as complex text interpretation challenges, limitations in response generation, inaccuracies in data annotation, and issues in extracting precise answers from longer outputs.

Perceptual Error

Question: Look at the situation in the <image 1>. You are traveling on a plane with a small child. The cabin is depressurized and the oxygen mask falls from the ceiling. Which of these pictures shows the reconciliation of egoism and other-isms (the order is from left to right, top to bottom)?
Option: (A) 1 (B) 2 (C) 3

Ground Truth: (B) 2



<image 1>

GPT-4V: The image shows three situations on a plane when the oxygen mask falls: The adult puts on the mask first while the child watches (egoism). The adult puts the mask on the child first while struggling for breath. Both the adult and the child share the mask. The reconciliation of egoism and other-isms would be a situation where both the individual's needs and the needs of others are considered and addressed. Therefore, the correct answer is: (C) 3

Error Reason: GPT-4V recalled the right knowledge and made the right reasoning, but it failed to correctly map the IDs to the corresponding illustrations in the figure, which is not explicitly marked in the figure but is only described in text (the order is from left to right, top to bottom).

Figure 6. A basic perceptual error, easy for humans but challenging for GPT-4V. More examples can be found in the Appendix.

In summary, our error analysis underlines the challenges posed by MMMU and highlights areas for further research in visual perception, knowledge representation, reasoning abilities, and multimodal joint understanding. 1) *Interplay of language and vision*: language can aid in making visual understanding more explainable, while also leading models to hallucinate. 2) *Challenges in grounding*: tasks involving grounding or referring to specific elements within a visual input remain challenging, even for sophisticated models like GPT-4V. 3) *Complex reasoning is still challenging*: models still fail in complex reasoning scenarios involving lengthy reasoning chains or extensive calculations.

6. Conclusion

The introduction of MMMU marks a significant step towards evaluating the capabilities of LMMs in the context of Expert AGI. By assessing both basic perceptual skills and complex reasoning abilities across various professional domains, MMMU provides a comprehensive benchmark that aligns with the expectations of skilled adults in these fields.

MMMU, like any benchmark, has limitations despite its comprehensive nature. The manual curation process may carry biases, and the focus on college-level subjects might not be sufficient for testing Expert AGI [52]. However, we argue that strong performance on this benchmark should be a necessary criterion for an Expert AGI system. The challenging nature of MMMU is evident from the performance of over 30 models and human experts. To strike a balance between complexity and practicality, MMMU combines multiple-choice questions with concise open-ended questions, enabling the assessment of diverse subjects while addressing the challenges associated with evaluating open-ended responses.

References

[1] Blaise Agüera y Arcas and Peter Norvig. Artificial general intelligence is already here. *Noema Magazine*, 2023. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 3, 5

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2, 3

[4] 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. 3, 5, 6, 15, 16, 17, 18, 19, 20, 21

[5] 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. 5, 6, 7, 15, 16, 17, 18, 19, 20, 21

[6] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 3, 5, 6, 7, 15, 16, 17, 18, 19, 20, 21

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 1

[8] Bunny. Bunny-3b. `https://github.com/cappuch/Bunny-Qwen`, 2024. GitHub Repository. 15, 16, 17, 18, 19, 20, 21

[9] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 2

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120, 2020. 3

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 6, 15, 16, 17, 18, 19, 20, 21

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3, 5, 6, 15, 16, 17, 18, 19, 20, 21

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1

[14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 3, 5, 6, 15, 16, 17, 18, 19, 20, 21

[15] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*, 2023. 3, 8

[16] 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. *arXiv preprint arXiv:2305.06500*, 2023. 2, 3, 5, 6, 7, 15, 16, 17, 18, 19, 20, 21

[17] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 15, 16, 17, 18, 19, 20, 21

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[19] Adept Fuyu Team. Adept fuyu-heavy: A new multimodal model. `https://www.adept.ai/blog/adept-fuyu-heavy`, 2024. 6, 15, 16, 17, 18, 19, 20, 21

[20] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 3, 5

[21] Yingqiang Ge, Wenyue Hua, Jianchao Ji, Juntao Tan, Shuyuan Xu, and Yongfeng Zhang. Openagi: When llm meets domain experts. *arXiv preprint arXiv:2304.04370*, 2023. 1

[22] Google Gemini Team. Gemini: A family of highly capable multimodal models. `https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf`, 2023. 6, 15, 16, 17, 18, 19, 20, 21, 120

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2, 3

[24] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Mea-

suring massive multitask language understanding. In *International Conference on Learning Representations*, 2020. 2

[25] Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Collier Nigel, and Yutong Lu. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*, 2023. 3

[26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[28] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2

[29] Ehsan Latif, Gengchen Mai, Matthew Nyaaba, Xuansheng Wu, Ninghao Liu, Guoyu Lu, Sheng Li, Tianming Liu, and Xiaoming Zhai. Artificial general intelligence (agi) for education. *arXiv preprint arXiv:2304.12479*, 2023. 1

[30] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2, 3

[31] 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*, 2023. 3, 5, 15, 16, 17, 18, 19, 20, 21

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, 2023. 2, 3, 5, 6, 7, 15, 16, 17, 18, 19, 20, 21

[33] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 3

[34] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 3

[35] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[37] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 15, 16, 17, 18, 19, 20, 21

[38] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023. 3

[39] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 3

[40] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2, 3, 5, 6, 7, 15, 16, 17, 18, 19, 20, 21

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3

[42] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. 2024. 6, 15, 16, 17, 18, 19, 20, 21

[43] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 2, 3

[44] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 3

[45] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[46] Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2

[47] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 3

[48] 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. 3

[49] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023. 1, 3

[50] MiniCPM. Minicpm-v. https://github.com/OpenBMB/MiniCPM, 2024. GitHub Repository. 15, 16, 17, 18, 19, 20, 21

[51] Masoud Monajatipoor, Liunian Harold Li, Mozhdeh Rouhsedaghat, Lin F Yang, and Kai-Wei Chang. Metavl: Transferring in-context learning ability from language models to vision-language models. *arXiv preprint arXiv:2306.01311*, 2023. 3

[52] Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi: Operationalizing progress on the path to agi. *arXiv preprint arXiv:2311.02462*, 2023. 1, 3, 8

[53] OminiLMM. Ominilmm-12b. https://github.com/OpenBMB/OmniLMM, 2024. GitHub Repository. 15, 16, 17, 18, 19, 20, 21

[54] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 6, 15, 16, 17, 18, 19, 20, 21

[55] OpenAI. Gpt-4v(ision) system card, 2023. 2, 6, 7, 15, 16, 17, 18, 19, 20, 21

[56] Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, et al. Reka flash: An efficient and capable multimodal language model. https://reka.ai/reka-flash-an-efficient-and-capable-multimodal-language-model/, 2024. 15, 120

[57] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 5, 6, 15, 16, 17, 18, 19, 20, 21

[58] Qwen. Qwen-vl-plus. https://github.com/QwenLM/Qwen-VL?tab=readme-ov-file#qwen-vl-plus, 2023. GitHub Repository. 15, 16, 17, 18, 19, 20, 21

[59] Qwen. Qwen-vl-max. https://github.com/QwenLM/Qwen-VL?tab=readme-ov-file#qwen-vl-max, 2024. GitHub Repository. 6, 15, 16, 17, 18, 19, 20, 21

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5

[61] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[62] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 2

[63] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023. 6, 15, 16, 17, 18, 19, 20, 21

[64] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 3

[65] Claude Team. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family, 2024. 6, 15, 120

[66] InfiMM Team. Infimm: Advancing multimodal understanding from flamingo's legacy through diverse llm integration, 2024. 15, 16, 17, 18, 19, 20, 21

[67] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 5

[68] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6, 15, 16, 17, 18, 19, 20, 21

[69] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023. 3

[70] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2, 5, 6, 15, 16, 17, 18, 19, 20, 21

[71] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2021. 3

[72] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023. 3

[73] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 2

[74] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers

large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 3

[75] 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. 3, 5, 15, 16, 17, 18, 19, 20, 21

[76] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*, 2023. 2, 3

[77] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 6, 15, 16, 17, 18, 19, 20, 21

[78] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. 3

[79] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2, 3

[80] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 3

[81] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 3, 6, 15, 16, 17, 18, 19, 20, 21

[82] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 3, 15, 16, 17, 18, 19, 20, 21

[83] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023. 3

[84] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934*, 2023. 3

[85] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023. 2

[86] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13041–13049, 2020. 3

[87] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 5, 6, 15, 16, 17, 18, 19, 20, 21