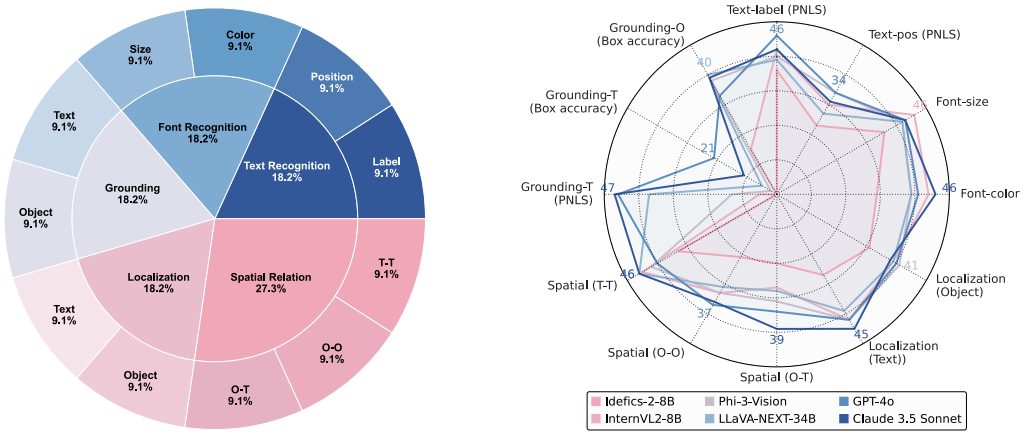
MMR: Evaluating Reading Ability of Large Multimodal Models

Jian Chen[1]*, Ruiyi Zhang[2]†, Yufan Zhou[2], Ryan Rossi[2],
Jiuxiang Gu[2], Changyou Chen[1]
[1]University at Buffalo [2]Adobe Research



(a) (b)

Figure 1: (a) Distribution of tasks in MMR Benchmark. (b) Performance of different models on MMR. Existing models show poor text grounding ability and weakness on spatial relationship reasoning.

Abstract

Large multimodal models (LMMs) have demonstrated impressive capabilities in understanding various types of image, including text-rich images. Most existing text-rich image benchmarks are simple extraction-based question answering, and many LMMs now easily achieve high scores. This means that current benchmarks fail to accurately reflect performance of different models, and a natural idea is to build a new benchmark to evaluate their complex reasoning and spatial understanding abilities. In this work, we propose the Multi-Modal Reading (MMR) benchmark[1] in 11 diverse tasks to evaluate LMMs for text-rich image understanding. MMR is the first text-rich image benchmark built on human annotations with the help of language models. By evaluating several state-of-the-art LMMs, including GPT-4o, it reveals the limited capabilities of existing LMMs underscoring the value of our benchmark.

1 Introduction

Large multimodal models have shown impressive capabilities in understanding various types of image, including text-rich images (Liu et al., 2024b; Li et al., 2023a; Li, 2023; Zhu et al., 2023; Alayrac et al., 2022; Laurençon et al., 2024b; McKinzie et al., 2024). Existing text-rich image datasets and benchmarks are composed of single-page document images (Mathew et al., 2020; Mishra et al., 2019; Mathew et al., 2022) or natural images with scene texts (Singh et al., 2019; Sidorov et al., 2020). The questions associated with these datasets typically require simple extraction rather than advanced reasoning or spatial understanding. With LMMs showing significant performance gains (Liu et al., 2024a,d), existing benchmarks have almost been solved, as evidenced by the high metric scores. This progress has made it challenging to accurately gauge the true capabilities and differentiate the performance levels of various models.

In this work, we introduce a novel Multi-Modal Reading (MMR) benchmark designed to provide a more rigorous assessment of Language Multimodal Models (LMMs) in the context of text-rich image comprehension. Specifically, MMR is built on the LAION dataset and selectively retains images that exhibit a significant presence of text. We ask human annotators to create comprehensive captions that encompass text content, visual elements, layout structures, and their inherent attributes. Leveraging advanced language models, such as GPT-4V,

*Work done at University at Buffalo.
†Corresponding Author
[1]Project page: https://llavar.github.io/mmr/

we generate a challenging Visual Question Answering (VQA) benchamrk. This benchmark surpasses the complexity of existing text-rich image VQA datasets and encompasses 11 distinct tasks. For each of these tasks, we have refined the metrics to better suit the evaluation of LMMs.

Our comprehensive evaluation of open-source and proprietary models, varying in size, reveals the current limitations of LMMs. Specifically, we still see a gap between open-source and proprietary models. These open-source models usually do not follow the instructions provided and output in the desired format, mainly due to the limited size of the instruction finetuning dataset. LLaVA-Next-34B (Liu et al., 2024a) shows the best performance in the object grounding task. Phi-3-Vision (Abdin et al., 2024) shows impressive performance even with compact size, further demonstrating the importance of data quality. These observations further show that open-source models can perform better than proprietary models in specific skills. All models show poor performance on text grounding tasks, which is an important skill to improve for future LMMs. We anticipate that the MMR benchmark can provide valuable insights for the research community and encourage further advances in the nuanced field of complex visual text understanding.

2 Related Work

Classical VQA Benchmarks TextCap (Sidorov et al., 2020) is the first text-rich image captioning dataset. Text-OCR (Singh et al., 2021) aims to comprehend text in the context of an image, which is similar to our motivation, but focuses more on text recognition in images rather than understanding. ST-VQA (Furkan Biten et al., 2019) uses spatial and textual information to answer visually grounded questions, effectively integrating visual and textual cues. OCR-VQA (Mishra et al., 2019) focuses on incorporating optical character recognition (OCR) into visual question answering (VQA), which operates primarily on text within images. TextVQA (Singh et al., 2019) also takes advantage of the textual information present in the images to answer questions, but with an emphasis on open questions. DocVQA (Mathew et al., 2021) takes this one step further by applying VQA to document images, handling a variety of layouts and formats. InfoVQA (Mathew et al., 2022) and ChartQA (Masry et al., 2022) focus on specific subdomains and aim to answer questions about information graphics and chart images, respectively. All of these benchmarks are mostly composed of extractive questions, while MMR provides complex reasoning evaluations for multimodal LLMs.

Large Multimodal Model Benchmarks Recent advancements in large multimodal models have led to the development of various benchmarks aimed at evaluating their capabilities across different tasks (Fu et al., 2023; Li et al., 2023b). MMBench (Liu et al., 2023c) and MM-Vet (Yu et al., 2023) offer comprehensive assessments of multimodal model efficacy on recognition-based tasks. More recently, the BLINK (Fu et al., 2024b) benchmark was proposed for evaluating a model's nuanced perception abilities beyond recognition. Tong et al. (2024) presents CV-Bench, which adapts existing vision benchmarks (Brazil et al., 2023; Lin et al., 2015; Zhou et al., 2019) to formulate natural language questions aimed at testing the spatial comprehension abilities of models. MM-UPD Bench (Miyai et al., 2024) on the other hand, tests a model's ability to recognize and refrain from answering unsolvable VQA problems in the multiple-choice setting. Benchmarks beyond single-image understanding have been designed to assess the ability to understand multiple images (Li et al., 2024; Wang et al., 2024b).

Benchmarks have also been proposed to evaluate more specific abilities of multimodal models. MathVista (Lu et al., 2023) focus on mathematical reasoning and SciFIBench (Roberts et al., 2024) focused on scientific figure interpretation. Recent work introduced the MMMU benchmark (Yue et al., 2024) that provides multi-discipline tasks for evaluation of large multimodal models, that require college-level knowledge about specific subject matters and deliberate reasoning capabilities. Multipanel VQA (Fan et al., 2024) introduced the multipanel visual question answering task, which involves interpreting multiple image panels arranged as a layout in a single image, such as posters and website screenshots. VisualWebBench (Liu et al., 2024c) assesses the capabilities of LLMs across a variety of web tasks. MuirBench (Wang et al., 2024a) is designed to assess the ability to comprehend multiple images simultaneously. There have also been recent work focused on benchmarking multi-modal LLMs for video analysis called Video-MME (Fu et al., 2024a). These benchmarks collectively push the boundaries of what large multimodal models can achieve, fostering continuous

improvement and innovation.

3 Multimodal Reading Benchmark

For existing text-rich image benchmarks, most of them (Singh et al., 2019, 2021; Mathew et al., 2020) focus on information extraction, such as DocVQA (Mathew et al., 2021) and TextVQA (Singh et al., 2019). The recently released OCRBench creates a new benchmark by carefully selecting questions from existing benchmarks, and MT-VQA (Tang et al., 2024) expands multilingual VQA pairs in text-rich images. To address this problem and provide a better evaluation of LMMs in text-rich images, we propose the MMR benchmark to evaluate multimodal reading ability, including spatial understanding, text recognition, and complex reasoning. Figure 1(a) shows the distribution of different tasks in MMR and 1(b) shows the performance of representative models.

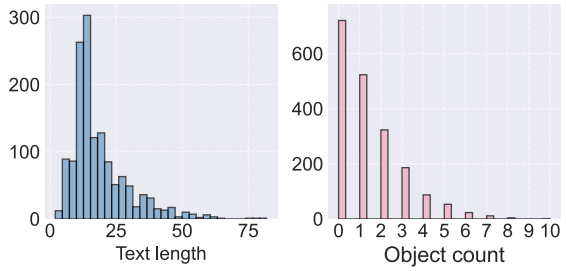


Figure 2: Text and Object length distribution in MMR.

Figure 3: Wordcloud of text (Left) and object tags (Right) of MMR Benchmark.

3.1 Statistics

MMR benchmark comprises pairs of 11 visual question answering tasks on text-rich images. The dataset is constructed based on 408 images selected from a total of 1,931 text-rich images. We included example questions for each task in Figure 5 for better understanding. We categorized all questions into five main classes, which are further divided into 11 more specific types. To ensure a fair evaluation, we manually curated 50 question-answer pairs for each type of question.

Due to the high diversity of objects and visual text in the benchmark, we do not classify them into a predefined set of labels. Instead, we use an OCR model to detect text, and RAM++ (Huang et al., 2023) to generate open-set object tags and display rough content distribution using word clouds. Figure 2 shows the distribution of OCR words and the number of objects detected on the MMR benchmark, with means of 18.45 and 1.35. Figure 3 shows the word cloud of text and object tags.

Considering the poor performance of different models on the text grounding tasks shown in Figure 1, we create an additional 900 question-answer pairs in text grounding for a comprehensive evaluation and the development of new methods.

3.2 Data collection and Human Annotations

We first build a dataset of text-rich images based on LAION-5B [2] (Schuhmann et al., 2022) with carefully designed heuristics and machine learning models. Then we ask annotators to give detailed captions for each text-rich image. In addition, we design various prompts and use human annotations to help GPT-4V (Yang et al., 2023) generate question-answer pairs. After that, human annotators are asked to verify these QA pairs and make corrections as benchmarks.

Machine-Assisted Image Selection We filter and maintain text-rich images from LAION-5B dataset. To differentiate between text-intensive document images and natural images, we first compile a binary classifier, a DiT (Li et al., 2022) base model, which was further refined using the RVL-CDIP dataset (Harley et al., 2015), to determine the presence of text in an image. Then we use PaddleOCR to extract all words from the selected images and keep images with more than 20 words and less than 100 words, which eliminates most text-intensive document images. The final step uses semantic information to select the desired images. A random sample of 20,000 images from the filtered LAION-5B is clustered into 50 groups based on CLIP-ViT-B/32 visual features. After inspecting the clustering results, two clusters are chosen as text-rich images. This cluster model then serves as the filtering mechanism for collecting images that comprise the MMR dataset.

Human Annotated Dense Captions Following the scheme of human-annotated captions from

[2] https://huggingface.co/datasets/laion/laion-high-resolution

Figure 4: Examples of human annotated dense captions. All text elements are annotated in detail, such as color, position, and contents. Detailed descriptions of visual elements and layout information are provided as well.

TRINS (Zhang et al., 2024a), we provide comprehensive annotation instructions and examples to annotators and ask them to (*i*) provide detailed descriptions of visual components, and (*ii*) describe the location, attributes, and exact words of the texts in annotations. Our goal is to better translate a text-rich image into text descriptions with minimum information loss. Considering the unstable ability of multimodal understanding and great performance on text-only tasks, this process can provide a reliable source for question-answer pairs generation. Figure 4 shows examples of annotated examples.

Human-Machine Hybrid QA Annotations We provide image annotations, including human annotations, OCR results, object detection results, and the images themselves, to GPT-4V to construct QA pairs that test the visual understanding abilities of the vision-language model on text-rich images. These questions are divided into two categories: reading and understanding visual text, and spatial position detection and understanding of visual elements (objects and text). For text recognition and position detection, we prompt the model to output the detected text and bounding-box coordinates in a fixed format. For more complex questions, we create multiple choice questions and require the model to output the index of the correct option to facilitate quantitative evaluation.

When constructing questions, we use the Azure OCR tool [3] to recognize visual text and detect their bounding boxes, and we use Grounding DINO (Liu et al., 2023a) to detect objects. These results serve as the ground truth for certain questions and are provided to GPT-4V to enhance the reliability of the answers to generated questions. All question-answer pairs are manually inspected for accuracy.

3.3 Benchmark Tasks

MMR benchmark encompasses 11 distinct tasks in texts, fonts, visual elements, bounding boxes, spatial relations, and grounding, as demonstrated in Figure 5. These tasks can be categorized into text recognition, spatial relationships, localization, and grounding, and all are essential skills to evaluate the reading ability of large multimodal models.

Text Recognition For text recognition, we ask a model to retrieve text for a given label, such as title or author name. This requires a model to extract text for a specified label. We also ask a model to retrieve text strings based on their position on the canvas or relative to other elements, testing its localized reading ability.

Font Recognition We use multiple-choice questions about font size and text color to assess the model's ability of visual text understanding. These questions are created based on OCR results, human

[3] https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/how-to/call-read-api

Figure 5: Example questions from MMR to evaluate reading capabilities.

annotations, and the image itself. They ask about the color of a specified text and compare the font sizes between two texts within the given image.

Element Localization We construct multiple-choice questions for both text and object localization, asking the model to choose the correct region for the target element, such as the bottom or top left corner of the image. This task aims to roughly localize the target text without requiring complicated format requirements for the output.

Spatial Relationship Understanding We construct multiple-choice questions to test the model's ability to comprehend pairwise spatial relationships between elements. The questions are categorized into three types: object-text, object-object, and text-text pairs. These questions are generated based on images, human annotations, element bounding boxes, and optionally OCR results.

Object Grounding We prompt the model to output relative bounding-box coordinates within a range of 0 to 1 in a Python list format, which requires high object localization precision and instruction-following ability. The ground truth is generated by combining human and model-based annotations.

Text Grounding We construct text grounding questions that require a model to output both the text string and the bounding box coordinates simultaneously, in a specified format that concatenates a string with a Python list for auto-extraction. The target text is specified by its rough location on the 3x3 grid and the corresponding text labels. This task demands high text localization precision and even higher instruction-following ability compared to object grounding, which only asks for box coordinates.

3.4 Quality Control

To ensure the correctness and quality of the QA pairs in our data set, we combine human- and

model-based annotations to select ideal images for each task. We then manually curate 50 questions per task to assemble a high-quality QA dataset.

Bounding Box For text grounding and recognition, we extract text content from human annotations (Zhang et al., 2024a) and match it with OCR results, discarding images where OCR fails to capture all annotated texts. The OCR model provides quadrilateral bounding boxes, which we filter using a 30-degree threshold on the average horizontal tilt angle of the upper and lower edges to exclude sloping text unsuitable for rectangular box detection. For object grounding, we utilize RAM++ (Huang et al., 2023) to exclude images without objects. We then manually label object tags and feed these tags into Grounding-DINO to detect rectangular bounding boxes for specified objects.

Spatial Location We extract the center coordinates of each bounding box and assign a rough position using a 3x3 grid on the canvas. This rough positioning is used as a condition for text recognition and text grounding tasks.

Multiple Choices For multiple-choice questions, we prompt GPT-4V to generate a question with four choices and the true answer in one response. Multiple choice provides an easy way to perform automatic evaluation, avoiding the usage of hard-matching metric scores (Papineni et al., 2002; Lin, 2004) or the involvement of large models in the evaluation process (Liu et al., 2023b).

Human Verification We ask human annotators to review the MMR benchamrk. Both questions and answers are manually checked, and annotators will correct the answer or rewrite the question if it does not belong to the target task.

4 Experimental Results

We evaluate the performance of popular vision-language models on the MMR benchmark. Our assessment includes seven open-source vision-language models of different sizes: Monkey-Chat (Li et al., 2023c), Idefics (Laurençon et al., 2024a), Idefics-2 (Laurençon et al., 2024b), LLaVA-v1.5 (Liu et al., 2024b), LLaVA-NEXT (Liu et al., 2024a), Phi-3-Vision (Abdin et al., 2024), and InternVL2 (Chen et al., 2023). Additionally, we evaluate five proprietary vision-language models: Qwen-vl-plus, Qwen-vl-max (Bai et al., 2023), Claude 3.5 Sonnet (Anthropic, 2024), GPT-4V (Yang et al., 2023), and GPT-4o. We include the prompts used for our experiment in Appendix A. All experiments were performed on a single A100-80GB GPU.

4.1 Evaluation metrics

We evaluate the model's performance on all tasks using three metrics tailored to the output type. For multiple-choice questions, performance is measured by the number of correct choices made by the models. For the grounding task, we assess the quality of detected bounding boxes using the IoU score, and for text recognition, we propose a new metric, PNLS, to compare text strings. To facilitate the computation of a total score across all tasks for comparing the overall performance of models, we convert the continuous IoU and PNLS metrics to binary scores using a threshold. In our benchmark, we set the thresholds to 0.3 for IoU and 0.9 for PNLS, respectively. The two metrics are explained as follows:

PNLS For text recognition tasks, we propose Partial Normalized Levenshtein Similarity (PNLS), a variant of normalized levenshtein similarity (NLS) (Biten et al., 2019). PNLS adapts the global alignment algorithm into a local-global version (Sellers, 1980), which avoids penalizing extra prefix or suffix characters. This makes it more effective for evaluating text recognition results from language models, as these models often produce verbose outputs to improve user experience.

Compared to the normalized Levenshtein similarity (NLS), PNLS uses the length of the region aligned with the true text string as the normalization factor. This aligned region is determined through dynamic programming. The score still ranges from 0 to 1 and positively correlates with performance. The motivation behind this design is to avoid penalizing extra prefixes or suffixes in a model's output. AccANLS (Zhang et al., 2024b) was proposed for the same purpose. However, it only spares penalties on prefixes and suffixes when there is an exact match of the true text string in the model's output.

The PNLS metric is formally defined as follows: String $\mathcal{T}_{1,m} = t_1 \dots t_m$ represents the true answer and $\mathcal{S}_{1,n} = s_1 \dots s_n$ is a model generated string. We first identify the sub-string of $\mathcal{S}$ that has the minimum edit distance to $\mathcal{T}$. Specifically, we first construct a scoring matrix $\mathbf{F}$ of size $(m+1) \times (n+1)$, where $F_{i,j}$ stores the smallest edit distance

		Text		Font		Localization		Spatial Relation			Grounding			Total
Models	**Size**	**Label**	**Pos.**	**Size**	**Color**	**Obj.**	**Text**	**O-T**	**O-O**	**T-T**	**O-Box**	**T-PNLS**	**T-Box**	
InternVL2	1B	35	29	32	24	28	25	17	27	19	0	1	0	237
Phi-3-Vision	4B	40	**34**	42	39	**41**	42	31	33	42	38	13	2	397
Monkey-Chat	7B	36	22	33	27	26	16	9	18	27	0	0	0	214
Idefics-2	8B	36	23	36	29	31	27	20	21	33	0	0	0	256
InternVL2	8B	42	30	**46**	44	39	42	27	33	45	15	5	0	368
LLaVA 1.5	13B	30	10	25	20	32	17	16	24	26	33	0	4	243
LLaVA-NEXT	13B	36	27	37	33	38	38	23	31	37	39	2	0	335
LLaVA-NEXT	34B	39	27	42	39	39	39	28	31	**46**	**40**	37	5	412
Idefics	80B	0	1	21	20	21	17	20	19	20	0	0	0	139
Qwen-vl-plus	-	38	23	32	35	26	23	24	23	27	34	22	3	310
Qwen-vl-max	-	39	27	41	36	34	33	26	32	37	24	32	5	366
GPT-4V	-	43	33	43	40	37	38	26	26	45	26	**48**	10	415
GPT-4o	-	**46**	**34**	43	41	40	42	34	**37**	40	33	46	**21**	457
Claude 3.5 Sonnet	-	42	31	43	**46**	38	**45**	**39**	36	**46**	39	47	11	**463**

Table 1: Empirical results of different models on 11 tasks of MMR Benchmark. The blue columns show PNLS scores and the red columns show box matching scores. The upper and lower halves list open-source and proprietary models, respectively. The highest score for each task is highlighted in bold font.

between the i-prefix $\mathcal{T}_{1,i}$ and any sub-string $\mathcal{S}_{x,j}$, $\forall x \in \{1, \ldots, j-1\}$ that ends at position j. The scoring matrix can be computed recursively

$$F_{i,j} = \begin{cases} 0 & \text{if } i = 0 \\ m & \text{if } j = 0 \\ \min \begin{pmatrix} F_{i-1,j-1} + c(t_i, s_j) \\ F_{i-1,j} + 1 \\ F_{i,j-1} + 1 \end{pmatrix} & \text{otherwise,} \end{cases}$$

where c is the substitution cost that takes a value of 0 if $t_i = s_j$ and 1 otherwise. Once $\mathbf{F}$ is computed, the minimum value in the last row is the optimal edit distance and the end index of the matched sub-string $j' = \arg\min_j(F_{m+1,j})$. The start index i' can be found by tracing back the the computation of Eq.(4.1). Finally, the PNLS is computed as: $m/(m + j' - i' + 1)$.

IoU Scores For object and text grounding tasks, we use the mean Intersection over Union (IoU) score to evaluate the model's accuracy. We also report the number of valid outputs that follow the required format, evaluating the instruction following ability, and allowing a script to automatically extract the coordinates and text strings.

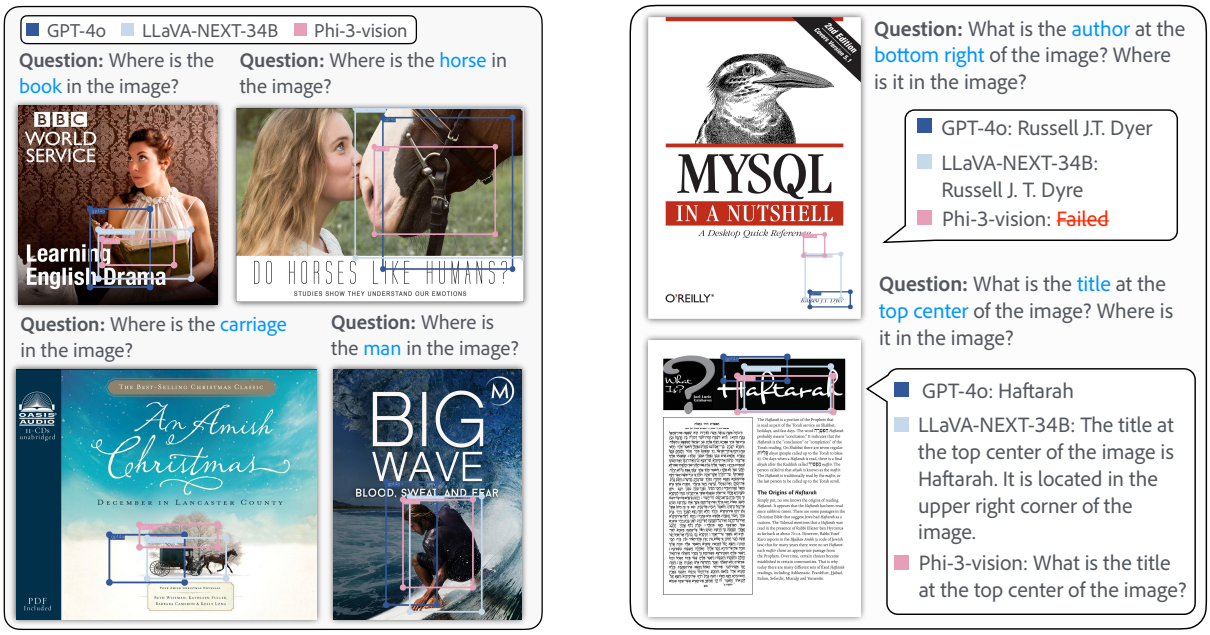
4.2 Quantitative results

Table 1 summarizes the performance of eleven models on all tasks, including counting, PNLS, and IoU scores, as introduced in Section 4.1. The text grounding task output both text and bounding box, thus are evaluated by two metrics.

We observe that GPT-4o (launched on May 13, 2024) and Claude 3.5 Sonnet (June 20, 2024) demonstrate superior overall performance, as indicated by the total score and the area covered in the radar chart. They generally outperform GPT-4V (March 14, 2023), highlighting the recent progress of proprietary models. However, we find that some open-source models can occasionally outperform GPT-4 models despite their smaller size.

Model Size v.s. Data Quality In our experiment, the performance of most models shows a positive correlation with model size. For example, LLaVA-NEXT-34B surpasses LLaVA-NEXT-13B. However, Phi-3-vision demonstrates impressive performance with only 4.2B parameters, surpassing larger models like Qwen-vl-plus and Qwen-vl-max, and rivaling LLaVA-NEXT-34B and GPT-4 models in many tasks. Despite its success, Phi-3-vision has a similar architecture to LLaVA (Liu et al., 2024b), suggesting that open-source models suffer from data-hungry issues. Thus, high-quality data is more essential than merely scaling up. This finding is further supported by the significant performance gap between Idefics-80B and its smaller successor, Idefics-2-8B.

In contrast to Phi-3-vision's notable performance on multiple-choice and text recognition questions, it performs poorly on text grounding tasks. A possible explanation is that this task demands high instruction-following ability for formatting longer outputs, which might require a larger model size, as we observed only larger models achieve reasonable performance in these tasks.



(a) (b)

Figure 6: Examples generated by different models on the Text (a) and Object (b) Grounding tasks.

Reading ability We use PNLS to evaluate the text reading ability of the model. Most models perform well in font recognition and text recognition by label but still struggle to match human performance. Additionally, when text is specified by its rough location, PNLS scores decrease, as these questions are more complex and require spatial understanding before recognition. The task becomes even more challenging with text grounding, where models must output both text and bounding boxes simultaneously. In these cases, smaller models like Idefics-2, LLaVA 1.5, and Monkey-Chat fail to provide valid results. Figure 6 shows examples of text bounding boxes detected by three models. We can see that LLaVA-NEXT and Phi-3-vision struggle to generate outputs in the required format, and all models, including GPT-4o, are unable to generate accurate bounding boxes. This indicates the need for improved visual text understanding in vision-language models.

Spatial understanding We also evaluate the spatial understanding ability of different models in localization, pairwise position understanding, and grounding tasks. Similar to the text grounding results discussed above, some smaller open-source models lack grounding ability and cannot provide valid responses to the questions. However, we find that LLaVA models and Phi-3-vision outperform GPT-4o and significantly outperform GPT-4V in the object grounding task, as measured by bounding box scores and illustrated in Figure 6. The excellent performance of LLaVA-NEXT models in these tasks could be attributed to their patch-wise encoding strategy. However, they are trained mainly on natural images with minimal experience in text-rich images (Zhang et al., 2023), resulting in poor performance in text grounding. This highlights the need for annotated text-rich images datasets.

5 Conclusion

In this paper, we introduce the Multi-Modal Reading (MMR) benchmark, which evaluates the reasoning and spatial understanding capabilities of LMMs in text-rich image understanding. The benchmark consists of eleven diverse tasks with carefully designed evaluation metrics. The experimental results showcase the performance of different models, giving suggestions on which model to choose in real-world applications. It also underscores the need for further research and development to bridge the gap between LMMs and human-level performance in text-rich image understanding.

6 Limitations

We only evaluated recently released models, and more models should be evaluated, which we hope can be handled by the community after the MMR benchmark is released. The evaluation metrics used in MMR still have limitations in accurate evaluation, and we have reformulated the VQA as multiple choices and provided output template

for LMMs to alleviate this issue. The questions are all proposed by the GPT-4V which may induce some model bias, while it is still difficult for human annotators to propose suitable questions with complex reasoning as they tend to ask extractive questions.

7 Acknowledge

This work is partially supported by NSF AI Institute-2229873, NSF RI-2223292, an Amazon research award, and an Adobe gift fund. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

8 Ethics Statement

Multi-Modal Reading (MMR) benchmark adheres to a set of ethical principles and guidelines to ensure responsible and ethical conduct of the study. Informed consent was obtained from all participants, who were fully informed about the purpose and nature of the research. Efforts were made to include participants from diverse backgrounds, promoting inclusivity and representation. The study was conducted with integrity of the research, adhering to scientific rigor and ethical standards. Compliance with relevant laws, regulations, and ethical guidelines was ensured throughout the research process. The research findings aim to contribute to the advancement of AI technology ethically, with a commitment to using the results for the betterment of society.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Anthropic. 2024. Claude 3.5 sonnet. Accessed: 2024-08-23.

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

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. 2023. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Xinze Guan, and Xin Eric Wang. 2024. Muffin or chihuahua? challenging large vision-language models with multipanel vqa. *arXiv preprint arXiv:2401.15847*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024b. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Minesh Mathew, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Icdar 2019 competition on scene text visual question answering. *2019 International Conference on Document Analysis and Recognition (ICDAR)*.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. *Preprint*, arXiv:1502.07058.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. 2023. Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pages arXiv–2310.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024a. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024b. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

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

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

Chunyuan Li. 2023. Large multimodal models: Notes on cvpr 2023 tutorial. *ArXiv*, abs/2306.14895.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2024. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. Dit: Self-supervised pre-training for document image transformer. *Proceedings of the 30th ACM International Conference on Multimedia*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023c. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. 2024c. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

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

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024d. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

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

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2020. Docvqa: A dataset for vqa on document images. *Preprint*, arXiv:2007.00398.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Li, Ziwei Liu, and Kiyoharu Aizawa. 2024. Unsolvable problem detection: Evaluating trustworthiness of vision language models. *arXiv preprint arXiv:2403.20331*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *arXiv preprint arXiv:2405.08807*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

Peter H Sellers. 1980. The theory and computation of evolutionary distances: pattern recognition. *Journal of algorithms*, 1(4):359–373.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024b. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Ruiyi Zhang, Yanzhe Zhang, Jian Chen, Yufan Zhou, Jiuxiang Gu, Changyou Chen, Nedim Lipka, and Tong Sun. 2024a. Trins: Towards multimodal language models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shuo Zhang, Biao Yang, Zhang Li, Zhiyin Ma, Yuliang Liu, and Xiang Bai. 2024b. Exploring the capabilities of large multimodal models on dense text. *arXiv preprint arXiv:2405.06706*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

A Prompt

This section include all prompt we use for all tasks and models.

Object Grounding prompt

```
where is the {object} in the image?
Please write the position as a
bounding box, and output the
[x_min, y_min, x_max, y_max]
coordinates in float numbers in
python list. Output the text only.
```

Text Grounding prompt

```
What is the {text label} at the
{area} of the image? Where is
it in the image? Please write
the position as a bounding box,
and output the [x_min, y_min,
x_max, y_max] coordinates in float
numbers in a python list. Output
the text and bounding box only.
For example: "Hello world" [x_min,
y_min, x_max, y_max]
```

Single Choice prompt

```
{question} Only print the index of
the  correct choice as answer,
such as 1, 2, 3, or 4.
```

Text Recognition prompt

```
{question} Only print the text; do
not include any other descriptions.
```

The prompt is inserted in the required format for each model. For example, LLaVA 1.5 requires the following format:

LLaVA 1.5 template

```
USER: <image>\n<prompt> ASSISTANT:
```

For the required input format of other models, please refer to the respective source code.

B PNLS demo

Figure B.1 provides an example.

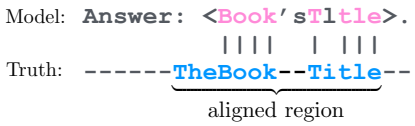


Figure B.1: Example of text similarity score. Only the 8 pink characters in the model's output "Answer: <Book'sTltle>." match the true string ("TheBookTitle" in blue). The aligned region has length 14. In this case the similarity is 8/14.