REVISITING MULTI-MODAL LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

With the advent of multi-modal large language models (MLLMs), datasets used for visual question answering (VQA) and referring expression comprehension have seen a resurgence. However, the most popular datasets used to evaluate MLLMs are some of the earliest ones created (VQAv2, GQA, TextVQA et al.) and they have many known problems, including extreme bias, spurious correlations, and an inability to permit fine-grained analysis. In this paper, we pioneer evaluating recent MLLMs (LLaVA-OneVision, MiniGemini, CogVLM, GPT-4V et al.) on datasets designed to address weaknesses in earlier ones. We assess three VQA datasets: 1) TDIUC, which permits fine-grained analysis on 12 question types; 2) TallyQA, which has simple and complex counting questions; and 3) DVQA, which requires optical character recognition for chart understanding. We also study VQDv1, a dataset that crucially requires identifying all image regions that satisfy a given query. Our experiments reveal the weaknesses of many MLLMs that have not previously been reported. Project webpage: `https://link-to-be-released`

1 INTRODUCTION

In recent years, multi-modal large language models (MLLMs) have emerged as powerful tools for tackling vision-language tasks (Li et al., 2023b; Chowdhery et al., 2022; Zhu et al., 2023; Koh et al., 2023; Liu et al., 2023b). Open source MLLMs leverage the extensive world knowledge of large language models (LLMs) and combine them with pre-trained vision encoders to process both linguistic and visual information (Liu et al., 2023b; Zhu et al., 2023; Liu et al., 2023a). These models are trained on various vision-language tasks such as visual question answering (VQA) (Goyal et al., 2017; Zhang et al., 2016), image captioning (Sharma et al., 2018), and visual conversations (sha). Their effectiveness is typically evaluated on VQA datasets (Goyal et al., 2017; Ren et al., 2015), which test the ability to produce answers to questions about images and referring expression comprehension tasks (Kazemzadeh et al., 2014), which require localizing the single object specified in the referring expression.

From 2017-2019, a series of datasets were designed to overcome the widely acknowledged weaknesses of earlier VQA and visual understand datasets (COCO, VQAv2, RefCOCO et al.) (Ren et al., 2015; Goyal et al., 2017; Mao et al., 2016), and intended to enable fine-grained analysis of visually grounded language understanding systems:

1. **VQDv1** (Acharya et al., 2019), which requires the model to produce multiple bounding boxes instead of localizing only one object, thereby testing for general query detection skills;
2. **TallyQA** (Acharya et al., 2018), which tests visual grounding through counting skills, asking questions that require intricate reasoning;
3. **TDIUC** (Kafle & Kanan, 2017), which tests versatility across 12 tasks, including object, attribute, and activity recognition, as well as overall scene understanding; and
4. **DVQA** (Kafle et al., 2018), which requires interpreting and analyzing visual data in chart form, testing for the ability to do OCR, and properly handling unusual words found in charts.

Despite this, these early datasets are now widely used to evaluate MLLMs. The most commonly used datasets, e.g. VQAv2 (Goyal et al., 2017), fail to adequately gauge visual grounding, allowing models to inflate performance by exploiting language bias without using visual information (Kafle & Kanan, 2016). Additionally, they do not categorize questions into types, preventing fine-grained analysis of abilities like attribute detection, object recognition, reasoning, and scene understanding. In contrast, TDIUC provides comprehensive evaluation across 12 diverse tasks, enabling fine-grained

analysis, while TallyQA focuses on counting, demanding intricate spatial reasoning for its complex questions. DVQA challenges models with chart interpretation, requiring OCR and handling unusual words. Referring expression datasets like RefCOCO (Mao et al., 2016) often only require localizing a single object, allowing models to exploit biases (Cirik et al., 2018; Acharya et al., 2019) and often can answer queries without even considering the sentence structures (Akula et al., 2020). In contrast, VQDv1 requires identifying multiple objects or none based on the query, making it a more rigorous test for visual grounding and reducing the ability to exploit biases.

This paper makes the following contributions:

1. We provide a robust evaluation of MLLMs on the TallyQA, TDIUC and DVQA datasets, revealing previously unreported weaknesses via fine-grained analysis across various question types and tasks.
2. Using VQDv1, we challenge MLLMs' visual grounding capabilities by requiring them to engage in complex visual reasoning to identify multiple objects beyond the limitations of single-object referring expression datasets.
3. We leverage our unique analysis to make inferences on the strengths and weaknesses of current MLLMs.

2 Multi-modal Large Language Models

Open-source MLLMs comprise a pre-trained LLM, a pre-trained vision encoder, and a learned adapter that aligns the visual and linguistic representations (Zhu et al., 2023; Liu et al., 2024b). They are usually trained in multiple stages. Initially, the adapter is trained to align the visual embeddings generated by the vision encoder with the textual embedding space of the LLM. Subsequently, the MLLM undergoes fine-tuning by adapting both the adapter and the LLM on various vision-language and instruction-tuning datasets. In our study, we consider both widely available state-of-the-art open-weight MLLMs and closed-source MLLMs.

BLIP2 (Li et al., 2023b) is a generic and compute-efficient method for vision-language pre-training that leverages frozen pre-trained image encoders and language models (LLMs). It pre-trains a lightweight Querying Transformer (Q-Former), consisting of image and text transformer sub-modules, to bridge visual and textual modalities. BLIP2, therefore, only trains a relatively light - 188M parameter transformer and achieves strong performance on VQA and image captioning tasks. We evaluate the base BLIP2 model (Li et al., 2023b), with 'blip2-flan-t5-xl' as the pretrained encoder.

iBLIP (Dai et al., 2024) (i.e., InstructBLIP), like BLIP-2, keeps the LLM and visual encoders frozen while introducing a novel instruction-aware Query Transformer that allows the model to extract informative visual features based on the textual instructions in the prompt. iBLIP is additionally trained on a much larger corpus of visual instruction tuning datasets, including knowledge-grounded image-question answering, visual reasoning, and VQA (Dai et al., 2024). This leads to improvements, including higher zero-shot performance on VQA tasks, compared to BLIP2 and larger MLLMs. We test the version that uses 'instructblip-flan-t5-xxl' as the pre-trained encoder.

LLaVA (Liu et al., 2023b) uses a visual instruction tuning dataset to fine-tune the LLM and adapter. LLaVA 1.5 enhances its vision encoder to handle higher-resolution images and replaces the linear projector layer with a multi-layer perceptron adapter. This version is trained on the VQA datasets VQAv2 and GQA datasets and a broader range of instruction-tuning data from sources like ShareGPT. These enhancements significantly improve its performance on fine-grained visual tasks, including detailed image description and complex question answering (Liu et al., 2023a). It achieves strong performance on several VQA benchmarks.

CogVLM (Wang et al., 2023) introduces a novel approach to bridging the gap between frozen pretrained language models and image encoders. Unlike shallow alignment methods, CogVLM employs a trainable visual expert module integrated into the attention and FFN layers. This deep fusion of vision-language features enables improved performance on cross-modal tasks without compromising NLP capabilities

QwenVL (Bai et al., 2023) is built upon the Qwen language model series and employs a three-stage training pipeline. It utilizes a visual receptor with a higher input resolution of 448x448 pixels, enabling more detailed image analysis. QwenVL incorporates a novel input-output interface that

supports bounding box inputs and outputs, facilitating visual grounding and text reading tasks. The model is trained on a multilingual multimodal corpus, allowing it to handle both Chinese and English inputs effectively. QwenVL demonstrates strong performance in zero-shot captioning and Chinese-language visual tasks, outperforming some larger models despite its relatively compact size

LLaVA-NeXT (Liu et al., 2024a) is an improved version of LLaVA 1.5, with a focus on enhanced visual reasoning, optical character recognition (OCR), and multi-modal document understanding. LLaVA-NeXT scales the input image resolution of input images by $4\times$, up to 1344×336 compared to 336×336 in LLaVA 1.5 to enhance its ability to grasp finer-grained visual cues. LLaVA-NeXT is also trained on a more diverse and realistic visual instruction-tuning dataset (ShareGPT-4V and LAION-GPT-V), as well as a range of OCR, document, and chart datasets. We evaluate the 7B parameter version of LLaVA-NeXT.

Mini-Gemini (Li et al., 2024b) introduces a novel framework to allow for refined image processing of the visual encoder without increasing the visual token count. To enable this it employs a dual-encoder system: one each for low-resolution and high-resolution visual embeddings, as well as a patch info mining technique to conduct patch-level mining between high-resolution regions and low-resolution visual queries. Mini-Gemini is trained on a data recipe curated to improve image comprehension and reasoning-based generation. Mini-Gemini-HD (MGM-HD) processes images at 672x672 resolution, compared to Mini-Gemini (MGM)'s 336x336 normal resolution processing. MGM-HD is claimed to enable improved performance on detail-oriented tasks like text-VQA while maintaining computational efficiency. We evaluate both the Mini-Gemini-HD (MGM-HD) and Mini-Gemini (MGM) versions at the 7B parameter scale.

LLaVA-OneVision (Li et al., 2024a) is a family of open large multimodal models that learns a single model to transfer across various modalities - single-image, multi-image, and video scenarios simultaneously. It consolidates insights from the LLaVA-NeXT series, employing a Qwen-2 language model, SigLIP vision encoder, and a 2-layer MLP projection layer. LLaVA-OneVision achieves strong transfer learning across modalities, demonstrating emerging capabilities in tasks like diagram interpretation, set-of-mark prompting, and video analysis. We evaluate the 7B parameter versions of the model.

GPT-4o/GPT-4V (Achiam et al., 2023; Yang et al., 2023) are closed-weight MLLMs created by OpenAI that enable users to leverage the capability of GPT-4 scale LLMs to analyze visual inputs. GPT-4V is a powerful generalist multi-modal model and can process arbitrarily interleaved image-text data. GPT-4V can perform many visual-language tasks well, including spatial understanding, object localization, and object counting (Yang et al., 2023). GPT-4o is reportedly an end-to-end text, vision, and audio multi-modal model, where multi-modal tokens are processed within the same network. GPT-4o has also been reported to improve linguistic and multi-modal understanding. Given that these are closed-source MLLMs, we use the API provided by OpenAI for our evaluations.

3 EXPERIMENTS

Across datasets we compute both micro performance, i.e., where every example is weighted equally, and macro performance, where we average across the mean score for different question/query types. We also generate a slim version of the datasets, by sub-sampling to maintain the long tailed distribution of the dataset while reducing the class imbalances (see Appendix Sec. C).

3.1 VISUAL QUERY DETECTION WITH VQDV1

Visual query detection (VQD) requires a model to provide bounding boxes for 0-N visual objects in response to a given query (Acharya et al., 2019). It is significantly more challenging than referring expression comprehension, which requires only localizing a single object in a scene. VQD aligns more closely with typical human referring behavior, where it is common to refer to multiple objects simultaneously. Unlike VQA, VQD requires the model to ground responses in visual inputs, providing direct evidence of task completion.

We evaluated all models on VQDv1 except for BLIP2 and iBLIP, which failed to produce bounding boxes under the zero-shot setting. All models were prompted to answer with a list of bounding boxes. We discuss details of prompt selection in Appendix D.

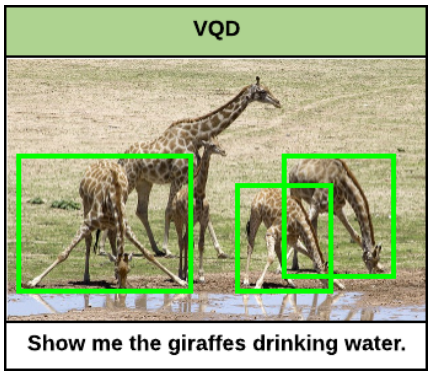


Figure 1: VQDv1 requires identifying all regions that satisfy a query.

Table 1: Performance comparison of various multi-modal large language models on VQDv1 dataset. 'L', 'MGM', 'OV' denote LLaVA, Mini-Gemini, and OneVision respectively.

Metrics/Model	L (7B)	L (13B)	QwenVL	CogVLM	GPT-4V	GPT-4o	MGM	MGM (HD)	L-NeXT	L-OV
Micro F_1	25.06	20.96	1.87	16.04	21.17	25.33	12.41	3.77	27.01	17.53
Macro F_1	19.87	16.81	2.43	20.76	16.54	21.01	15.66	4.30	21.84	21.75

VQDv1 Metrics. In (Acharya et al., 2019), average precision using an intersection over union (IoU) of 0.5 was used for evaluation; however, that requires scores for each box, which are unavailable for MLLMs. Therefore, we compute each model's micro and macro mean F_1 scores, recall, and precision. The predicted box with the highest IoU above 0.5 is considered a true positive for each ground-truth box, whereas any remaining predicted boxes are false positives. If a query has no ground truth bounding boxes, then the F_1 score is set to 1 when the model outputs no boxes. Otherwise, it is set to 0. Due to the limited number of questions with four or more bounding boxes, we grouped them.

Results for VQDv1. As presented in Table 1, all of the models struggle on VQDv1, with the best performing LLaVA-NeXT obtaining only 27.01 in terms of micro F_1 score. Fig. 2 shows the recall and precision scores across varying numbers of bounding boxes. Models struggle to ground multiple boxes, as evidenced by the recall score which decreases with an increase in the number of boxes.

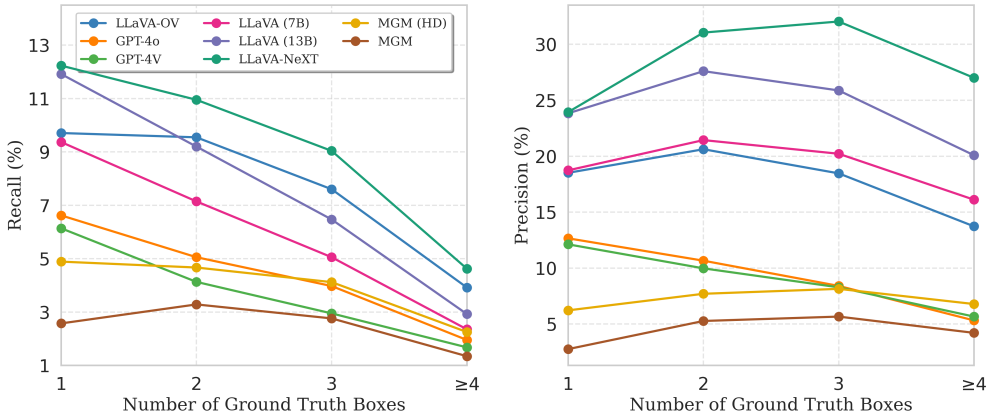
3.2 Fine-Grained VQA Assessment with TDIUC

TDIUC (Kafle & Kanan, 2017) is a VQA dataset that organizes its questions into 12 distinct types. Performance is computed for each question type. TDIUC aims to address the shortcomings of previous VQA datasets by offering a broader spectrum of question types, and it enables a comprehensive analysis of VQA capabilities for each model.

TDIUC Metrics. For TDIUC, we use micro-accuracy and macro-accuracy, where micro accuracy corresponds to the average accuracy across the 12 question types. Macro-accuracy corresponds to the mean per type metric in the original paper.

Results for TDIUC. Our main results on TDIUC are detailed in Table 2. LLaVA (13B) and LLaVA-NeXT achieve the highest micro accuracies under the asymptotic McNemar test ($p = 0.2355$). GPT-4o is the next best model, showing a statistically significant difference from LLaVA (13B) ($p = 0.0031$). BLIP2 obtains the poorest performance across question types, particularly in attribute/color recognition and counting. GPT-4V, GPT-4o, BLIP2, and iBLIP excel at absurd questions, whereas the LLaVA family performs worse, likely due to hallucinations. Compared to MuREl (Cadene et al., 2019), the best system trained on TDIUC, MLLMs greatly improve for utility affordance questions, except for BLIP2.

We note that introducing absurd questions poses an additional challenge to the model. In general, absurd questions are a test for the model's epistemic confidence in its responses.



(a) VQDv1 Recall (b) VQDv1 Precision

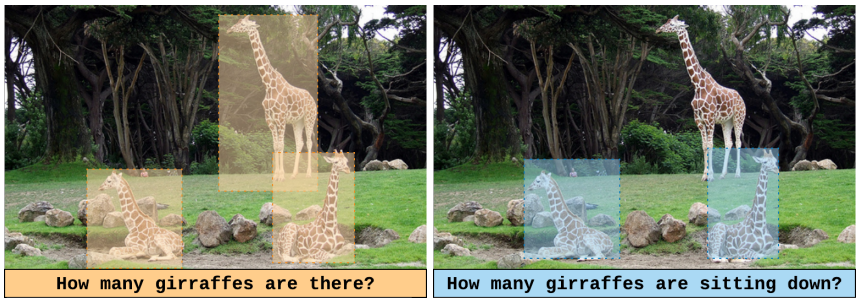
Figure 2: Recall and precision curves for queries with varying box counts.

Table 2: Accuracy on TDIUC for each question type. 'L', 'MGM' and 'OV' denote LLaVA, MiniGemini and OneVision respectively. Best performers based on paired asymptotic McNemar tests ($\alpha = 0.05$) are in bold, except for Macro. Acc., where the max is bolded. For comparison, MuRel Cadene et al. (2019) is the previous best result from training on TDIUC. Models marked as ‡ cannot be confirmed to be evaluated in a zero-shot manner.

Ques. Type	BLIP2	iBLIP	L (7B)	L (13B)	GPT-4V‡	GPT-4o‡	L-NeXT	CogVLM	QwenVL	L-OV	MGM	MGM-HD	MuRel
Absurd	**99.87**	97.44	51.48	74.73	99.04	99.45	68.14	70.13	2.34	72.29	63.27	77.99	99.80
Activity	25.00	54.00	63.50	62.00	56.50	62.50	**68.00**	0.00	55.00	**68.00**	62.50	**70.50**	63.83
Attribute	1.31	48.15	71.46	73.20	60.78	73.20	**79.08**	12.42	72.11	**80.39**	71.90	77.34	58.19
Color	5.70	62.13	77.37	80.54	69.05	78.97	81.05	23.69	**82.39**	**82.55**	77.56	**81.95**	74.43
Counting	7.15	39.24	51.95	53.27	52.36	56.14	54.93	56.48	47.83	**62.68**	49.41	55.95	61.78
Object Pres.	43.22	74.87	91.31	90.57	67.28	77.81	**92.07**	52.71	90.93	61.94	91.02	**92.40**	95.75
Object Rec.	43.74	73.79	75.03	75.29	69.30	69.30	75.23	81.88	76.27	**90.74**	76.92	76.86	89.41
Position	3.42	20.20	36.81	39.41	31.11	37.46	41.69	21.34	38.76	**53.91**	36.32	44.30	41.19
Scene	30.15	78.47	82.38	76.57	62.94	67.67	**84.29**	76.93	82.11	79.93	81.20	82.11	96.11
Sentiment	16.50	73.00	**79.50**	**82.50**	62.50	28.00	**79.50**	48.00	**80.50**	63.00	57.50	77.00	60.65
Sport	28.29	**88.45**	88.25	**89.84**	77.89	81.27	**89.24**	81.87	**88.45**	**89.64**	87.45	**89.44**	96.20
Utility/Aff.	19.88	66.67	**76.02**	**74.85**	**77.19**	**73.68**	**76.02**	25.73	70.18	**73.68**	64.33	72.51	21.43
Micro Acc.	45.07	73.38	73.86	79.07	72.19	78.30	78.91	54.77	63.09	69.75	75.92	**81.43**	-
Macro Acc.	27.02	64.70	70.42	72.73	65.49	67.12	74.10	45.93	65.57	73.23	68.28	74.86	71.56

3.3 Assessing Counting Ability with TallyQA

TallyQA (Acharya et al., 2018) tests model's ability to count visual objects accurately. Unlike earlier VQA datasets (Goyal et al., 2017), where the majority of the counting questions are straightforward and doable with simple object detection (e.g., "How many giraffes are there?"), TallyQA adds



(a) Simple counting question (b) Complex counting question

Figure 3: Examples of simple and complex counting questions in TallyQA.

additional challenges by incorporating more complex questions that necessitate detailed reasoning about the visual elements. For instance, a question such as "How many giraffes are sitting down?" requires the model to not only detect all the giraffes in the image but also to perform pose estimation to discern which giraffes are seated. This tests for enhanced capabilities including complex reasoning and specific visual analysis.

TallyQA Metrics. In addition to reporting micro accuracy, we group the questions based on their answers (0, 1, 2, 3, or 4+) and calculate the average to determine the macro accuracy.

Table 3: Results on TallyQA. For Micro Acc., best performers based on paired asymptotic McNemar tests ($\alpha = 0.05$) are in bold. For RMSE, the lowest value is bolded. For comparison, the result from SMoLA (Wu et al., 2024) is the current best on TallyQA. Models marked as † are reported to be trained with TallyQA, and thus not evaluated zero-shot. Models marked as ‡ cannot be confirmed to be zero-shot.

Model	TallyQA Test-Simple			TallyQA Test-Complex		
	Micro Acc.	Macro Acc.	RMSE	Micro Acc.	Macro Acc.	RMSE
BLIP2	64.3	43.0	3.74	27.5	24.8	1.57
iBLIP	73.1	61.7	1.22	49.3	35.6	2.15
LLaVA (7B)	75.5	66.5	1.20	64.1	45.5	2.21
LLaVA (13B)	76.6	67.3	1.01	65.6	47.8	1.93
QwenVL	62.2	65.9	1.44	41.5	40.6	5.22
CogVLM	82.9	75.7	0.62	71.6	53.9	1.42
Mini-Gemini	72.4	62.4	1.38	58.5	42.8	2.42
Mini-Gemini (HD)	78.5	69.0	0.87	66.5	48.7	1.71
LLaVA-NeXT	79.8	71.7	0.70	67.9	52.2	1.76
LLaVA-OneVision†	**83.7**	77.2	**0.56**	**73.0**	58.6	1.49
GPT-4V‡	73.6	69.0	0.86	62.6	50.4	1.58
GPT-4o‡	81.5	74.5	0.60	71.7	56.9	**1.21**
SMoLA (Wu et al., 2024)	83.3	-	-	70.7	-	-

Results for TallyQA. The results of the TallyQA analysis are displayed in Table 3. Compared to the simple counting questions, models exhibit large accuracy drops on complex counting questions, indicating deficiencies in reasoning capabilities (Hua et al., 2024). This is evident even for the top-performing GPT-4o, which experiences declines of 9.8% and 17.6% in terms of micro and macro accuracies, respectively. Additionally, as shown in Fig. 6a and 6b, the accuracy of models tend to decrease as the number of objects to be counted increases, with the accuracy dropping below 30% when the ground truth count is four or more. As shown in Figs. 6a and 6b, the BLIP models struggled to output zero, and BLIP2 always emitted a value greater than zero.

3.4 Assessing Chart Comprehension with DVQA

DVQA (Kafle et al., 2018) is a VQA dataset evaluating chart understanding. DVQA requires the model to perform grounding extensively. With synthetic charts, the model is required to handle words or formulae that are specific for that instance. This contrasts with datasets using natural images, where questions such as "What color is the sky?" are based on universal concepts, and even models that simply exploit dataset biases can obtain high accuracy by guessing that the sky is either blue or gray. In contrast, the models cannot inflate accuracy by exploiting such correlations in DVQA since the concepts correspond to arbitrary values (e.g., the labels can correspond to arbitrary bar heights and colors) (Kafle et al., 2018).

DVQA Metrics. For DVQA, we report micro and macro accuracy. DVQA has 3 question types: structural understanding, data retrieval, and reasoning. They are averaged to compute macro accuracy.

Results for DVQA. Results for DVQA are given in Table 4. LLaVA-NeXT achieved the highest micro accuracy, and under an asymptotic McNemar test all other models had a statistically significant

Table 4: Percentage (%) accuracy results on DVQA. Best performers based on paired asymptotic McNemar tests ($\alpha = 0.05$) are in bold, except for Macro. Acc., where the max is bolded. For comparison, PReFIL and Human results correspond to performance on Test-Novel Kafle et al. (2020), where PReFIL uses Improved OCR (see Kafle et al. (2020)). PReFIL is a DVQA system trained on DVQA's training set. Models marked as † are reported to be trained with DVQA, and thus not evaluated zero-shot. Models marked as ‡ cannot be confirmed to be zero-shot.

Model	Reasoning	Retrieval	Structural	Micro Acc.	Macro Acc.
BLIP2	12.79	9.38	45.78	16.17	22.65
iBLIP	15.22	14.23	48.50	19.41	25.98
LLaVA (7B)	17.76	20.22	51.40	23.10	29.79
LLaVA (13B)	19.01	22.07	57.89	25.25	32.99
CogVLM	35.89	34.53	71.88	40.33	47.44
Mini-Gemini†	31.64	39.24	84.07	41.16	51.65
Mini-Gemini (HD)†	52.64	62.66	91.37	61.08	68.89
LLaVA-NeXT†	69.14	82.73	73.47	74.06	75.11
LLaVA-OneVision†	76.72	86.65	98.19	82.80	87.19
QwenVL†	**84.65**	**92.84**	**99.41**	**89.26**	**92.30**
GPT-4V‡	33.26	61.83	88.73	49.88	61.27
GPT-4o‡	52.06	73.64	95.60	64.84	73.77
PReFIL Kafle et al. (2020)	80.73	67.13	99.57	80.04	-
Human Kafle et al. (2020)	85.83	88.70	96.19	88.18	-

difference in micro accuracy ($p < 0.0001$). Compared to other categories, all models performed best on structural questions. Structural questions include questions such as: 1) "How many bars are there?" 2) "Does the chart contain any negative values?" 3) "Are the bars horizontal?" and 4) "Is each bar a single solid color without patterns?" These questions do not require extracting textual information from the image and only require the analysis of visual features. Models were worst at reasoning questions. Our results highlight the importance of training on synthetic data, as was done in LLaVA-NeXT, for achieving strong performance. No MLLM achieves the performance of a PReFIL for reasoning questions, which was trained on DVQA's training set, or of humans (Kafle et al., 2020).

3.5 Analyzing the Strengths and Weaknesses of Today's MLLMs

We now discuss and analyse current MLLMs across a variety of criteria, based on our evaluations across DVQA, TDIUC, VQDv1 and TallyQA. We begin by evaluating the general capabilities of MLLMs across the datasets, then analyze how various MLLM development decisions, in particular - scale, architecture, model families, data recipes, and training paradigms affect the particular vision-language abilities of MLLMs we evaluate in this work.

3.5.1 Inferences on Capabilities of MLLMs

Our evaluation reveals that today's MLLMs exhibit a range of strengths and weaknesses across different vision-language tasks. Generally, MLLMs demonstrate strong performance in object recognition and scene understanding but struggle with tasks requiring complex reasoning, precise counting, and handling synthetic data representations."

On the **DVQA** dataset, which tests models on interpreting data visualizations like bar charts, we observe significant performance disparities. Open-source models like QwenVL and LLaVA-OneVision achieve high accuracies, with QwenVL attaining a Micro Accuracy of 89.26% and a Macro Accuracy of 92.30%, surpassing even the performance of models specifically trained on DVQA, such as PReFIL (Kafle et al., 2020). These models effectively interpret synthetic visual data and perform reasoning over it. In contrast, models like LLaVA (7B and 13B), BLIP2, and iBLIP show significantly lower performance, indicating challenges in handling synthetic datasets compared to natural images.

In the **TallyQA** dataset, designed to assess counting abilities, MLLMs generally perform well on simpler counting tasks but show a performance decline as the counting number increases. For instance, on the Test-Simple set, LLaVA-OneVision achieves the highest Micro Accuracy of 83.7%, but on the Test-Complex set, the accuracy drops to 73.0%. This decline suggests that while MLLMs can handle basic counting, they face difficulties in accurately detecting and enumerating multiple objects in complex scenes.

The **TDIUC** dataset provides a comprehensive evaluation across various question types. We observe that MLLMs perform differently depending on the question category. In 'Counting' questions, LLaVA-OneVision achieves the highest accuracy of 62.68%, outperforming models like GPT-4V (52.36%) and GPT-4o (56.14%). However, in categories like 'Object Recognition' and 'Sport', models such as QwenVL and LLaVA-OneVision excel, indicating proficiency in recognizing objects and scenes. Conversely, performance is lower in categories like 'Sentiment' and 'Position', highlighting limitations in understanding abstract concepts and spatial relationships.

On the **VQDv1** dataset, involving open-ended questions about images, LLaVA-NeXT outperforms other models with a Micro F1 score of 27.01% and a Macro F1 score of 21.84%. This suggests that LLaVA-NeXT has a better general understanding of visual content and can generate more accurate responses to open-ended questions.

Overall, our analysis indicates that while current MLLMs have advanced capabilities in certain areas, they still face significant challenges in tasks requiring complex reasoning, precise counting, and understanding synthetic visual representations.

3.5.2 Open vs Closed source models

We compare the performance of open-source models with that of closed-source models to understand how openness impacts model capabilities. Among the models evaluated, GPT-4V and GPT-4o are closed-source models developed by OpenAI, whereas models like LLaVA, CogVLM, and QwenVL are open-source. Strikingly, our results show that open-source models often achieve performance comparable to or even surpassing that of closed-source models.

For example, on the **VQDv1** dataset, which evaluates models on open-ended visual question answering without prior exposure, LLaVA-NeXT achieves the highest Micro F1 score of 27.01%, outperforming both GPT-4V (21.17%) and GPT-4o (25.33%). Similarly, on the **TDIUC** dataset, which provides a comprehensive evaluation across various question types, open-source models demonstrate competitive performance. For instance, consdering the 'Position' category — a task that assesses understanding of spatial relationships, LLaVA-OneVision achieves an accuracy of 53.91%, significantly outperforming GPT-4V (31.11%) and GPT-4o (37.46%). This indicates that open-source models are capable of handling spatial reasoning tasks at a level superior to closed-source models. This observation of equal or superior performance of open-source models holds across a number of abilities in TDIUC such as 'Utility/Affordance', 'Sport Recognition', 'Scene Recognition', among several others. These observations are especially striking considering the large gap in apparent model sizes between closed source and open-source models.

It's also noteworthy that on the **TallyQA** dataset, which focuses on counting objects in images, closed-source models show strong performance in certain metrics. For example, on the Test-Complex set, GPT-4o achieves the lowest Root Mean Square Error (RMSE) of 1.21, outperforming open-source models like LLaVA-NeXT (1.76) and Mini-Gemini (HD) (1.71). A lower RMSE indicates more precise counting, suggesting that closed-source models may have advantages in tasks requiring fine-grained numerical understanding.

3.5.3 Model Scale & Image Resolution

Model Scale. To assess the impact of model scale, we compare the performance of LLaVA models with different parameter sizes. The LLaVA models are available in 7B and 13B parameter versions, enabling us to evaluate how scaling affects their capabilities. Across the datasets, the 13B model generally outperforms the 7B counterpart, albeit with modest gains. On the TallyQA Test-Simple set, LLaVA (13B) achieves a Micro Accuracy of 76.6%, slightly higher than the 7B model's 75.5%. On the TDIUC dataset, the 13B model shows improved performance in several question categories, such as 'Counting' (53.27% vs. 51.95%) and 'Attribute' (73.20% vs. 71.46%). However, the

incremental improvements suggest that increasing model size from 7B to 13B does not lead to substantial performance boosts in vision-language tasks.

Image Resolution. We analyze the impact of input image resolution on MLLM performance on our evaluated datasets that enable fine-grained analysis. Models like LLaVA-NeXT and Mini-Gemini (HD) process images at higher resolutions, that are claimed to capture finer visual details. While previous studies have supported the benefit, it is interesting to note how this effect applies to visual understanding tasks that specifically de-bias language and visual biases and provide fine-grained visual analysis.

Comparing LLaVA-NeXT with LLaVA (13B) on the TallyQA Test-Complex set, LLaVA-NeXT achieves a Micro Accuracy of 67.9%, slightly higher than LLaVA (13B)'s 65.6%. Similarly, Mini-Gemini (HD) achieves a Micro Accuracy of 66.5%, outperforming significantly the standard Mini-Gemini's 58.5%. This suggests that higher resolution enables better counting performance in complex scenes. Additionally, on the DVQA dataset, Mini-Gemini (HD) achieves a significantly higher Micro Accuracy of 61.08%, than the standard model's 41.16%, with large improvements across 'Reasoning' (52.64% vs 31.64%), 'Retrieval' (62.66% vs 39.24%) and 'Structural' (91.37% vs 84.07%) capabilities. This suggests higher resolution processing improves detailed chart understanding across all reasoning, retrieval and structural analysis capabilities.

These findings suggest that incorporating higher-resolution images appears to benefit visual understanding across complex counting, visual reasoning and retrieval tasks. An important caveat to this observation is that in the multi-localization task of VQDv1, Mini-Gemini (HD) struggles significantly in comparison to the standard Mini-Gemini with 4.30 Macro F_1 compared to 15.66 Macro F_1 respectively. While LLaVA-NeXT also shows a mild improvement over L(7B). This indicates that on multiple object localization tasks such as VQDv1, higher resolution may not directly convey a universal benefit and can even hamper performance, suggesting that resolution gains do not straightforwardly translate to better localization accuracy.

4 Related Work

Problems with Widely Used Datasets. With the advent of large foundation models, datasets for training, fine-tuning, and validation have become increasingly important (Liang et al., 2022). These datasets are pivotal in reflecting a model's performance across different aspects. Notably, many recent MLLMs rely on some of the earliest established datasets (Goyal et al., 2017; Kazemzadeh et al., 2014; Ren et al., 2015), which, while foundational, are increasingly recognized for their constraints and biases. Existing VQA datasets have several well-known issues. Most fail to properly assess grounding capabilities—linking specific parts of an image to corresponding textual elements in questions. For example, on some datasets, models can achieve approximately 50% accuracy even when blinded to the image, relying solely on the questions (Kafle & Kanan, 2016). This indicates that many questions do not depend on grounding capabilities, allowing models to exploit learned biases rather than visual evidence. Moreover, popular VQA datasets focus narrowly on specific question types, limiting the assessment of models' generalization abilities. Most questions (69.84%) ask about objects in the image, hindering the model's ability to handle abstract reasoning, complex visual cues, or nuanced human interactions. Additionally, MLLMs often are not evaluated on synthetic datasets, missing opportunities to reveal limitations not observed with natural images. Mainstream referring expression recognition datasets like RefCOCO typically assume each referring expression refers to a single object, oversimplifying the task. In RefCOCOg (Mao et al., 2016), it was shown (Cirik et al., 2018) that randomly permuting words in the referring expressions only reduced performance by 5%, and models could achieve 71.2% precision for the top-2 predictions using only the image. This suggests that models exploit dataset quirks and biases rather than utilizing linguistic cues for grounding. The imbalance in target object selection and the simplistic design of referring expressions, with only one associated bounding box, further exacerbate this issue.

Related Efforts to Improve MLLM Evaluation. Recent works highlight challenges in evaluating MLLMs. In (Yuksekgonul et al., 2022), the ARO benchmark was introduced to assess models' understanding of complex compositional elements, and models evaluated on it performed poorly for like "the grass is eating the horse" versus "the horse is eating grass." Similarly, the Winoground datasets (Thrush et al., 2022) require models to match images with captions that use identical words in

different orders to assess their comprehension of linguistic composition concerning visual information. In (Shah et al., 2019), a cycle-consistency framework is proposed, evaluating models' ability to understand semantically similar questions. These studies complement ours and reveal other biases and limitations in MLLMs.

5 Discussion

In this work, we performed a detailed examination of modern MLLMs on diverse tasks that expose biases, demand finer reasoning, and require more holistic visual grounding. Our evaluations revealed several notable insights.

Our TallyQA results highlight the necessity of incorporating more complex counting questions to reflect models' counting capabilities better. The LLaVA family demonstrates robustness to complex counting questions that demand sophisticated reasoning. In contrast, other models, like QwenVL and BLIP2, perform poorly on these complex questions despite performing adequately on easy counting questions compared to LLaVA. Relying solely on easy counting questions can lead to inflated scores, which can be misleading.

Results from VQDv1 show that traditional single-object referring expressions are more accessible for models to handle. However, introducing more targets in referring expressions presents a significant challenge, as performance drops when more objects are involved. Examining VQDv1 and TallyQA, they are complementary in evaluating models. In VQDv1, the model must generate one or more bounding boxes around objects described in the question, serving as an improved version of counting questions by requiring models to justify their answers. In TallyQA, models perform well when accounting for fewer objects, but performance drops significantly as the number of objects increases, indicating poor generalization abilities. This aligns with findings from VQDv1, where models struggle with multiple bounding boxes but perform well with a single bounding box. VQDv1 and TallyQA offer a comprehensive evaluation of a model's ability to justify its answers and handle varying numbers of objects, highlighting weaknesses in object detection and counting abilities.

Results from TDIUC provide insight into models' generalization across different question types. Most perform poorly on positional reasoning, an essential skill for complex counting questions and referring expressions. TDIUC also includes counting questions, and similar to TallyQA, models show a significant drop in macro accuracy. However, these results also show that Utility/Affordance questions benefit greatly from MLLMs compared to models trained on TDIUC.

All models perform poorly on DVQA, indicating that MLLMs struggle with parsing chart information, especially in reasoning and data retrieval questions. LLaVA-NeXT (and One-Vision family) improve significantly over other open-source MLLMs on DVQA, likely due to its training on documents and diagrams. The DVQA dataset highlights the challenges presented by synthetic images.

6 Conclusions

In this paper, we conducted comprehensive, skill-specific evaluations of MLLMs released in 2023–2024. Our analysis revealed several weaknesses that are not apparent when using mainstream datasets alone. First, we found that while current MLLMs excel at simpler visual queries and common question patterns, they face substantial difficulties in tasks that deviate from typical MLLM training distributions—such as multi-object localization, intricate counting questions, or synthetic chart interpretation. Second, analyzing tasks like TDIUC highlighted that many models still struggle with aspects of positional reasoning and scene-centric questions, despite strong performances on more basic recognition tasks. Third, contrary to widespread belief that higher resolution invariably improves visual performance, our findings suggest this is task-dependent: certain tasks (e.g., complex counting or chart reasoning) benefit significantly, whereas multi-object localization might not. To enhance accessibility for researchers and facilitate benchmark comparisons, we have integrated these datasets into a fork of the widely used LAVIS framework (Li et al., 2023a), and we will work with the LAVIS team to merge our version into the main trunk or release it as a separate entity, if necessary.

REFERENCES

ShareGPT: Share your wildest ChatGPT conversations with one click. — sharegpt.com. https://sharegpt.com/. [Accessed 16-05-2024].

Manoj Acharya, Kushal Kafle, et al. Tallyqa: Answering complex counting questions. *arXiv preprint arXiv:1810.12440*, 2018. URL https://arxiv.org/abs/1810.12440.

Manoj Acharya, Karan Jariwala, et al. Vqd: Visual query detection in natural scenes. *arXiv preprint arXiv:1904.02794*, 2019. URL https://arxiv.org/abs/1904.02794.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Arjun R Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. *arXiv preprint arXiv:2005.01655*, 2020.

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. 2023.

Remi Cadene, Hamid Ben-younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. *arXiv.org*, February 2019. URL https://arxiv.org/abs/1902.09487.

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.

Volkan Cirik et al. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018. doi: 10.18653/v1/n18-2123. URL https://doi.org/10.18653/v1/n18-2123.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Yash Goyal et al. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6333. IEEE, July 2017. doi: 10.1109/CVPR.2017.670.

Hang Hua, Yulong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*, 2024.

Kushal Kafle and Christopher Kanan. Answer-type prediction for visual question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. doi: 10.1109/CVPR.2016.538. URL https://doi.org/10.1109/cvpr.2016.538.

Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, 2017.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5648–5656, 2018.

Kushal Kafle, Robik Shrestha, Scott Cohen, Brian Price, and Christopher Kanan. Answering questions about data visualizations using efficient bimodal fusion. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pp. 1498–1507, 2020.

Sahar Kazemzadeh et al. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. doi: 10.3115/v1/d14-1086.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. *ICML*, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 31–41, Toronto, Canada, July 2023a. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-demo.3.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024b.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b.

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

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

Junhua Mao et al. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. doi: 10.1109/CVPR.2016.9. URL https://doi.org/10.1109/cvpr.2016.9.

Menglin Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, 2015.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering, Feb 2019. Available at https://arxiv.org/abs/1902.05660.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, Apr 2022. Available at https://arxiv.org/abs/2204.03162.

Princeton University. *WordNet: An Electronic Lexical Database*. Princeton University, 2010. https://wordnet.princeton.edu.

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.

Junnan Wu, Xun Hu, Yongyi Wang, Bo Pang, and Radu Soricut. Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts. *arXiv.org*, April 2 2024. https://arxiv.org/abs/2312.00968.

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*, 9(1):1, 2023.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it. *OpenReview*, Sep 2022. Available at https://openreview.net/forum?id=KRLUvxh8uaX.

Peng Zhang et al. Yin and yang: Balancing and answering binary visual questions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. doi: 10.1109/CVPR.2016.542.

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.

Appendix

A Computational Resources

The evaluations of open-source MLLMs were conducted on a single A100 GPU with 40GB of RAM, which required approximately 200 hours on our university-wide computing infrastructure. To evaluate GPT-4V/GPT-4o, which are closed-source, we used the paid ChatGPT API provided by OpenAI and spent $922 for GPT-4V and $451 for GPT-4o, including runs to tune prompts.

B Additional Dataset Details

In this section, we present additional dataset details.

B.1 TallyQA

The counting questions in TallyQA are classified into complex and simple counting questions (Acharya et al., 2018). Simple counting questions were imported from existing datasets like VQA2 and Visual Genome. Complex questions were collected using Amazon Mechanical Turk (AMT) to gather 19,500 complex questions for 17,545 unique images. The images were sourced from both COCO and Visual Genome to ensure variety. The testing set of TallyQA contains 38,589 questions, which is a reasonable size. Therefore, we evaluated models on the entire original test set. The distribution of unique answers is given in Table 5. TallyQA is provided under the terms of the Apache License Version 2.0, January 2004: `http://www.apache.org/licenses/`

Table 5: The distribution of unique answers in TallyQA.

Answer	Complex	Simple
zero	4335	637
one	6853	12308
two	2479	5636
three	901	2034
four	453	1101
five	195	435
six	133	319
seven	70	152
eight	69	145
nine	31	84
ten	33	48
eleven	12	30
twelve	25	33
thirteen	7	13
fourteen	6	9
fifteen	6	7

B.2 VQDv1

VQDv1 (Acharya et al., 2019) was created synthetically using annotations from Visual Genome, COCO, and COCO Panoptic. This synthetic generation approach helps combat certain biases. The queries are generated using multiple templates for each type, allowing for diverse queries. The annotations used to generate these questions are derived from a combination of COCO's object annotations and Visual Genome's attribute and relationship information.

For VQDv1, almost 90% of the queries have less than two ground truth bounding boxes. In our subset, we retained all queries with more than one ground truth bounding box, and we sampled 10% of the queries with zero or one ground truth bounding box. Table 6 provides the distribution of ground truth boxes across queries. The VQDv1 dataset is provided under the terms of the Creative Commons

Table 6: Bounding box distribution for the original and modified versions of VQDv1.

Bounding Box Count	Original Version	Our Version
0	80025 (42.08%)	8001 (21.59%)
1	90101 (47.38%)	9008 (24.31%)
2	10127 (5.33%)	10127 (27.33%)
3	3200 (1.68%)	3200 (8.64%)
4	1894 (1.00%)	1894 (5.11%)
5	1334 (0.70%)	1334 (3.60%)
6	700 (0.37%)	700 (1.89%)
7	533 (0.28%)	533 (1.44%)
8	366 (0.19%)	366 (0.99%)
9	305 (0.16%)	305 (0.82%)
10	276 (0.15%)	276 (0.74%)
11	193 (0.10%)	193 (0.52%)
12	194 (0.10%)	194 (0.52%)
13	255 (0.13%)	255 (0.69%)
14	618 (0.32%)	618 (1.67%)
15	26 (0.01%)	26 (0.07%)
16	5 (0.00%)	5 (0.01%)
17	7 (0.00%)	7 (0.02%)
18	7 (0.00%)	7 (0.02%)
19	3 (0.00%)	3 (0.01%)
20	1 (0.00%)	1 (0.00%)
23	2 (0.00%)	2 (0.01%)
25	1 (0.00%)	1 (0.00%)
26	1 (0.00%)	1 (0.00%)

B.3 DVQA

The DVQA dataset was created by synthetically generating bar charts to test multiple aspects of bar chart understanding. This automatic generation process allows precise control over the visual elements' positions and appearances, and provides access to meta-data about the elements in the image, which is not available with real data (Kafle et al., 2018).

The original version of DVQA had two test sets: Test-Familiar and Test-Novel. The critical difference between these two sets is that every bar chart in Test-Familiar has labels in DVQA's training set, whereas Test-Novel does not. Given that we are conducting zero-shot evaluations, these two sets can be treated equivalently. Therefore, we sample the same number of questions from both. Table 7 shows the question distributions of our subset version of DVQA. The DVQA dataset is provided under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0): https://creativecommons.org/licenses/by/4.0/legalcode

Table 7: Distribution of question types in DVQA.

Question Type	Test-Familiar Version	Test-Novel Version	Our Version
Data	185356 (31.93%)	185452 (31.90%)	9269 (31.91%)
Reasoning	316923 (54.59%)	316881 (54.51%)	15844 (54.55%)
Structure	78278 (13.48%)	78988 (13.59%)	3930 (13.53%)

B.4 TDIUC

The TDIUC dataset was created by incorporating questions from three sources: existing datasets, questions generated based on image annotations, and human annotators. Questions were imported

from COCO-VQA and Visual Genome datasets, with templates and regular expressions used to classify and generate questions(Kafle & Kanan, 2017). Additionally, questions were generated using COCO's semantic segmentation annotations and Visual Genome's objects and attribute annotations(Kafle & Kanan, 2017). For certain question types like sentiment understanding and object utility/affordance, trained volunteers performed manual annotation using a web-based tool(Kafle & Kanan, 2017). We sample proportionately from 12 question types in TDIUC. Table 8 shows our subset of TDIUC question distributions. TDIUC is a public dataset but does not mention a particular license.`https://kushalkafle.com/projects/tdiuc.html`

Table 8: Distribution of question types in TDIUC.

Question Type	**Original Version**	**Our Version**
Absurd	120411 (22.35%)	6844 (25.00%)
Activity Recognition	2682 (0.50%)	77 (0.28%)
Attribute	9200 (1.71%)	296 (1.08%)
Color	62490 (11.60%)	2142 (7.82%)
Counting	52905 (9.82%)	2262 (8.26%)
Object Presence	215324 (39.96%)	11884 (43.41%)
Object Recognition	30693 (5.70%)	1646 (6.01%)
Positional Reasoning	12284 (2.28%)	523 (1.91%)
Scene Recognition	22032 (4.09%)	1188 (4.34%)
Sentiment Understanding	634 (0.12%)	27 (0.10%)
Sport Recognition	10042 (1.86%)	478 (1.75%)
Utility Affordance	171 (0.03%)	12 (0.04%)

C Creating "Slim" Evaluation Sets

We evaluate MLLMs on the entire validation set of TallyQA, which contains 38,589 questions. However, the other datasets are much larger, which makes it challenging to quickly and inexpensively evaluate MLLMs on them. To address this, we sample subsets from these datasets for evaluation. A uniform random sampling is suboptimal as these datasets have long-tailed distributions and sampling uniformly would result in discarding examples from the tail. Therefore, we adopt a stratified sampling approach for DVQA and TDIUC, where we also maintain as much answer variety as possible. Specifically, we first categorize the questions into fine-grained groups, defined by both the pre-defined types in the datasets (e.g., question types or difficulty levels) and their corresponding answers. We define r as the sampling ratio and k as the minimum number of samples from each group. For any large group, we uniformly sample an r proportion of the entries. For smaller groups, if the size m is such that $m \cdot r$ is less than k, we sample k entries. For groups even smaller than k, we use the entire group. The number of samples m' to be taken from group $|g_i| = m$ can be represented as follows:

$$m'_i = \begin{cases} m_i & \texttt{if } m_i \leq k \\ k & \texttt{if } m_i \cdot r < k \wedge m_i > k \\ \lceil m_i \cdot r \rceil & \texttt{if } m_i \cdot r \geq k \end{cases}$$

VQDv1 has a long-tail distribution regarding the number of bounding boxes per query, where queries with 0 or 1 box comprise almost 90% of the dataset. Our goal is to evaluate the MLLM's ability to generate a variable number of bounding boxes – extending the evaluation scope beyond traditional referring expression comprehension datasets such as RefCOCO (Mao et al., 2016), where all referring expressions are associated with only one bounding box. Therefore, we retained all the questions with more than one bounding box and randomly sampled queries corresponding to 0 or 1 bounding box. As seen in Table 6, this method effectively increases the ratio of questions with multiple bounding boxes.

Our sampling method preserves the most challenges samples present in the original dataset, ensuring a comprehensive evaluation while significantly reducing computational overhead. Summary statistics for the datasets are given in Table 9.

Table 9: SS: Supplementary Summary statistics for the VQA and VQD datasets we study.

Dataset Name	# of Categories	# of Unique Answers	Original Size	Sampled Size
TDIUC Kafle & Kanan (2017)	12	562	538,868	27,336
TallyQA Acharya et al. (2018)	2	16	38,589	38,589
DVQA Kafle et al. (2018)	3	2113	580,557	29,025
VQDv1 Acharya et al. (2019)	5	24	190,174	37,057

D PROMPT ENGINEERING

To make the model performance comparison as fair as possible, we endeavored to keep the prompts consistent across different models. However, this was challenging due to variations in the models' ability to process the prompts. For example, BLIP2 and iBLIP failed when prompted to answer using a template such as "My answer is ¡answer¿." Inspired by Liu et al. (2023a), for TDIUC, DVQA, and TallyQA, we prompt the models to answer as concisely as possible instead of asking them to generate entire sentences. These prompts are given in Fig. 4.

Figure 4: Prompts used for TallyQA, DVQA, and TDIUC.

- TallyQA: Please answer the question in one word.
- DVQA: Please answer the question in one word
- TDIUC: Please answer in one word. Answer 'doesnotapply' if the question is not related to the image or cannot be answered.

Despite much effort, for VQDv1, we were unable to identify a universal prompt for generating multiple bounding boxes that worked well across models. For example, as shown in Table 10, LLaVA (7B) repeatedly generated the same bounding boxes until the maximum token limit was reached when this prompt was used. We believe this occurs because the model is confused by the instruction to generate multiple bounding boxes, even when only one object is detected. This may explain why it repeatedly generates the same bounding box. While we considered non-maximal suppression or eliminating redundant boxes, our goal is to fairly evaluate MLLMs without excessively post-processing their outputs. Therefore, we fine-tuned the prompts for different models. The results reported in the paper represent the best outcomes from our evaluations. The best-identified prompts for each model on VQDv1 are given in Table 11.

Table 10: LLaVA (7B) struggled with some prompts for VQDv1.

Prompt with Query	Where is the motorcycle? Instruction: **Generate a list of bounding box coordinates around the objects that the prompt mentioned if they exist in the image. Even if the question uses a singular verb like 'is', you should still generate multiple bounding boxes if there are multiple objects that satisfy the prompt. The bounding box list should be in the following format: [[x_min, y_min, x_max, y_max], [x_min, y_min, x_max, y_max]]**
LLaVA (7B)	predicted_answer: "[[0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01, 0.0, 0.28, 0.24], [0.01,"

Table 11: The best identified VQDv1 prompts for each model.

Model	Prompt
GPT-4o/GPT-4V	Generate a list of bounding box coordinates around the objects mentioned in the prompt if they exist in the image. Even if the prompt uses a singular verb like 'is', generate multiple bounding boxes if multiple objects satisfy the query. The bounding box list should be formatted as: [[x_min, y_min, x_max, y_max]], and it can contain zero or more bounding boxes. Only provide the bounding box list, without any additional descriptions.
LLaVA-NeXT	Please generate a list of bounding boxes coordinates of the region this query describes. Use the format [[x_min,y_min,x_max,y_max]....]. Do not respond in sentences, and only generate the bounding boxes. Respond with an empty list [[]], if no such region exists in the image.
LLaVA (7B)/(13B)	Please answer the question by generating a list of bounding box coordinates around the objects the question is asking, and if no such object exists in the image, answer: [[]]

E Model Details

In this paper, all the open source MLLMs are loaded directly from HuggingFace, the detail models are below:

Table 12: MLLM Model Repository Paths

Model	Repository Path
LLaVA-NeXT	llava-hf/llava-v1.6-mistral-7b-hf
InstructBlip	Salesforce/instructblip-flan-t5-xxl
BLIP2	Salesforce/blip2-flan-t5-xl
LLaVA1.5-7b	llava-hf/llava-1.5-7b-hf
LLaVA1.5-13b	llava-hf/llava-1.5-13b-hf

GPT-4v/4o are not open sourced, therefore we are unable to identify the models. We utilize the API released by OpenAI to evaluate four datasets on GPT-4v/4o.

F Additional Evaluation Details

Root Mean Squared Error (RMSE) Computation. For TallyQA, besides Micro and Macro Accuracy, we also compute RMSE. However, we observed that due to the unpredictability of the MLLMs, the models occasionally output unreasonably large numbers as their predicted object counts. For instance, LLaVA-NeXT predicts an unreasonably large object count of 150 for one of the questions. Such outliers significantly inflate the models' overall average RMSE across all questions. As shown in the distribution of TallyQA questions, all counting numbers are between 0 and 15. Therefore, we apply a simple cutoff technique: an upper bound of 15 and a lower bound of 0 is applied to all predicted counts. This adjustment ensures that the RMSE remains meaningful and useful for analysis.

Match Answer with Ground Truth.
For TallyQA, the model is tasked with generating object counts. If the model correspondingly generates a number enclosed within a string, such as "2", we directly convert it to int type by type conversion. For the case where the model generates a word, we map the word to its corresponding number using the mappings shown in table 5. Occasionally, the model generates answers that, while not numerical, still make sense. For example, the model might generate 'none' or 'no,' which we interpret as zero. We manually account for these cases and add additional mappings accordingly. While we acknowledge that even with these steps, we may still miss some unpredictable answers

from the models, such as when the model responds with 'a few,' which is completely uninterpretable, we map these to None.

Word	Number	Word	Number	Word	Number
zero	0	four	4	eight	8
none	0	five	5	nine	9
no	0	six	6	ten	10
one	1	seven	7	eleven	11
two	2	twelve	12	fourteen	14
three	3	thirteen	13	fifteen	15

Figure 5: Mapping of words to numbers in TallyQA

In datasets like TallyQA, DVQA, and VQDv1, synonymous answers are rarely an issue due to the specific nature of each task. For example, TallyQA typically expects numerical answers that are definitive and unambiguous (numbers seldom have synonyms). The main exception is when 'none,' 'no,' and 'zero' are all interpreted as 0. In DVQA, which focuses on chart understanding, questions such as 'Which bar has the highest number?' require the model to read and provide the exact text from the graph, minimizing the possibility of synonymous answers. Similarly, VQDv1 involves generating bounding boxes and computing the Intersection over Union (IoU) to determine if the ground truth is correctly matched. The evaluation uses Recall and Precision metrics, which are not binary and therefore do not penalize synonymous answers.

In contrast, tasks in TDIUC are more likely to involve more interpretative answers. For example, the answers 'phone' and 'telephone' should be considered semantically similar and should both be acceptable if the ground truth is one of them. To minimize penalizing synonymous answers like the case above, we leverage WordNet(University, 2010), a lexical database for the English language that is specifically designed for natural language processing. Specifically, we retrieve the sets of synonyms for each word from WordNet (using the synsets function) and compare these sets. If there is any overlap in the synsets, two words are considered synonyms, and we use this to evaluate if the predicted word(s) matches the ground truth(s).

G Additional Results

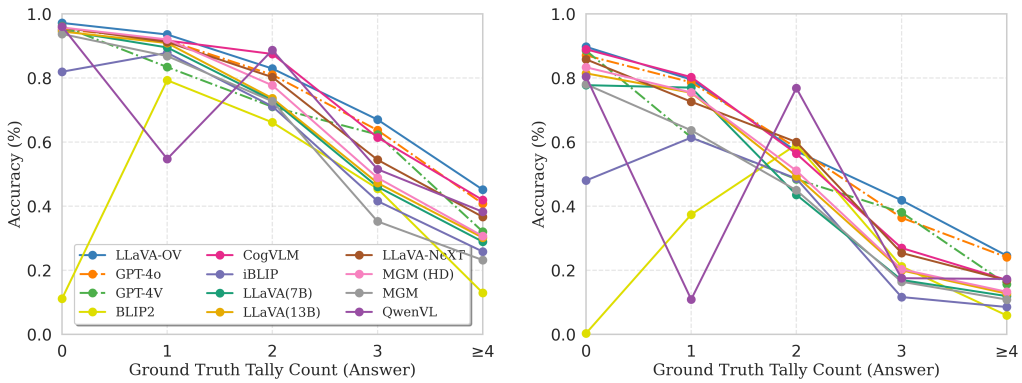
G.1 TallyQA

For TallyQA, we found that the performance of most models decreases as the correct number to output increases, as shown in Figs. 6a and 6b. Across counts, models perform much better at answering simple questions than complex questions.

G.2 VQDv1

Alternative Prompts. All models performed poorly on VQDv1. As mentioned earlier, it was challenging to identify the best prompt for each model. We hypothesized that given the verbosity of GPT-4o, it would benefit from being allowed to provide more extended responses where it reasons 'aloud.' However, this performed worse than the prompts used in our main results. In Table 13, we provide alternative prompts that we tried, where the results are given in Table 14.

Qualitative Examples. Among all the datasets we evaluated, all models consistently performed poorly on VQDv1. Consequently, we provide qualitative examples from VQDv1 in the figures below, using the prompts employed in our main results. These visualizations demonstrate the challenges models face when required to detect multiple objects.



(a) Simple counting questions in TallyQA.

(b) Complex counting questions in TallyQA.

Figure 6: Accuracy as a function of the correct answer for simple and complex counting questions in TallyQA.

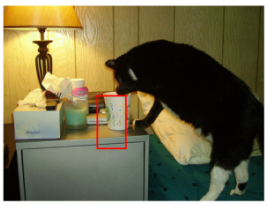
Table 13: Alternative prompts studied for VQDv1.

Model	**Prompt**
GPT-4o	Please generate a list of bounding boxes coordinates for regions that match what is described in the query. Bounding boxes should use the format: [[x_min,y_min,x_max,y_max], ..], where (x_min,y_min) is top left coordinate,(x_max,y_max) is bottom right coordinate. If there are no objects in the image that the query describes, please respond with an empty list. You can explain your answers if necessary, but end your response with the format: The bounding boxes coordinates are ¡box¿[[x_min,y_min,x_max,y_max],..,..]¡box¿. Please keep the special token ¡box¿ in your response.
LLaVA-NeXT	Generate a list of bounding box coordinates around the objects mentioned in the query, if they exist in the image. Even if the query uses a singular verb like 'is', generate multiple bounding boxes if multiple objects satisfy the query. The bounding box list should be formatted as: [[x_min, y_min, x_max, y_max]], and it can contain zero or more bounding boxes. Only provide the bounding box list, without any additional descriptions.
LLaVA (7B)/(13B)	Generate a list of bounding box coordinates around the objects that the prompt mentioned if they exist in the image. Even if the query uses a singular verb like 'is', you should still generate multiple bounding boxes if multiple objects satisfy the prompt. The bounding box list should be in the following format: [[x_min, y_min, x_max, y_max], [x_min,y_min, x_max, y_max]].

Table 14: MLLM performance on VQDv1 using the alternative prompts from Table 13.

Model	LLaVA (7B)	LLaVA (13B)	LLaVA-NeXT	GPT-4o
Micro F_1	4.27%	8.90%	14.66%	23.81%

Question: Show the cup in the picture.



model: GPT-4o

model: LLaVA-NeXT

model: LLaVA-v1.5-7b

model: GPT-4V

model: LLaVA-v1.5-13b

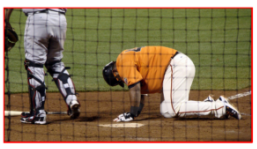
Ground truth

Question: Where is the person in the picture?

model: GPT-4o

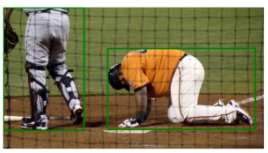
model: LLaVA-NeXT

model: LLaVA-v1.5-7b



model: GPT-4V

model: LLaVA-v1.5-13b

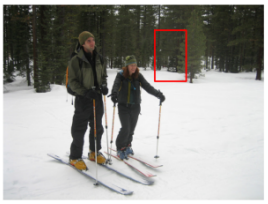


Ground truth

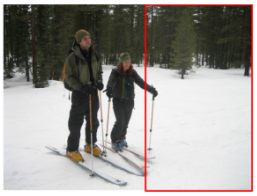
Question: Which tree is green in color?

model: GPT-4o

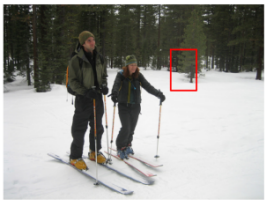
model: LLaVA-NeXT



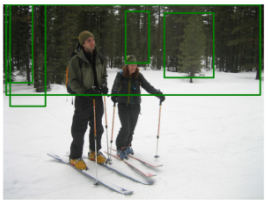
model: LLaVA-v1.5-7b



model: GPT-4V



model: LLaVA-v1.5-13b



Ground truth