MMCOMPOSITION: Revisiting the Compositionality of Pre-trained Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

The advent of large Vision-Language Models (VLMs) has significantly advanced multimodal understanding, enabling more sophisticated and accurate integration of visual and textual information across various tasks, including image and video captioning, visual question answering, and cross-modal retrieval. Despite VLMs' superior capabilities, researchers lack a comprehensive understanding of their compositionality – the ability to understand and produce novel combinations of known visual and textual components. Prior benchmarks provide only a relatively rough compositionality evaluation from the perspectives of objects, relations, and attributes while neglecting deeper reasoning about object interactions, counting, and complex compositions. However, compositionality is a critical ability that facilitates coherent reasoning and understanding across modalities for VLMs. To address this limitation, we propose **MMComposition**, a novel human-annotated benchmark for comprehensively and accurately evaluating VLMs' compositionality. With MMComposition, we can quantify and explore the compositionality of the mainstream VLMs. Surprisingly, we find GPT-4o's compositionality inferior to the best open-source model, and we analyze the underlying reasons. Our experimental analysis reveals the limitations of VLMs in fine-grained compositional perception and reasoning, and points to areas for improvement in VLM design and training. [1]

1 Introduction

Pre-trained vision-language models, such as GPT-4o (Achiam et al., 2023), LLaVA (Liu et al., 2024b), InternVL (Chen et al., 2024b), and VILA (Lin et al., 2024), have demonstrated impressive capabilities in complex reasoning, and have achieved remarkable results in various vision-language (VL) tasks. Despite these advancements, contemporary state-of-the-art VLMs still struggle with understanding fine-grained multimodal compositional information (Yuksekgonul et al., 2022; Thrush et al., 2022). For instance, VLMs often fail at counting objects in images, especially when the objects are mixed with other items or occluded, while humans can handle this task easily. This reveals a compositionality gap between humans and models. However, *compositionality* is recognized as a core capability for VLMs (Yuksekgonul et al., 2022), referring to the ability to understand and produce a potentially infinite number of novel combinations of known visual and textual components, i.e., to make "infinite use of finite means" (Chomsky, 2014). Compositionality is essential for tackling challenging questions in image captioning, visual question answering (VQA), and scene understanding, where complex interactions between objects and attributes need to be communicated in natural language.

In recent years, there has been a growing focus on evaluating the comprehensive capabilities of large VL models, such as MMBench (Liu et al., 2023b), MMMU (Yue et al., 2023), MMVet (Yu et al., 2024a;b), MME (Fu et al., 2023), Seed-bench (Li et al., 2023a), MMStar (Chen et al., 2024a), MathVista (Lu et al., 2023), and LLaVA-Bench (Liu et al., 2024b). These benchmarks evaluate VLMs' capabilities in recognition, OCR, knowledge, language generation, spatial awareness, and mathematical reasoning. While some of these benchmarks include visual compositional question-answering (QA) pairs (Fu et al., 2024; Li et al., 2023a; Tong et al., 2024b), none are specifically designed to comprehensively evaluate the models' fine-grained VL

[1] All data and code will be released upon publication of this paper.

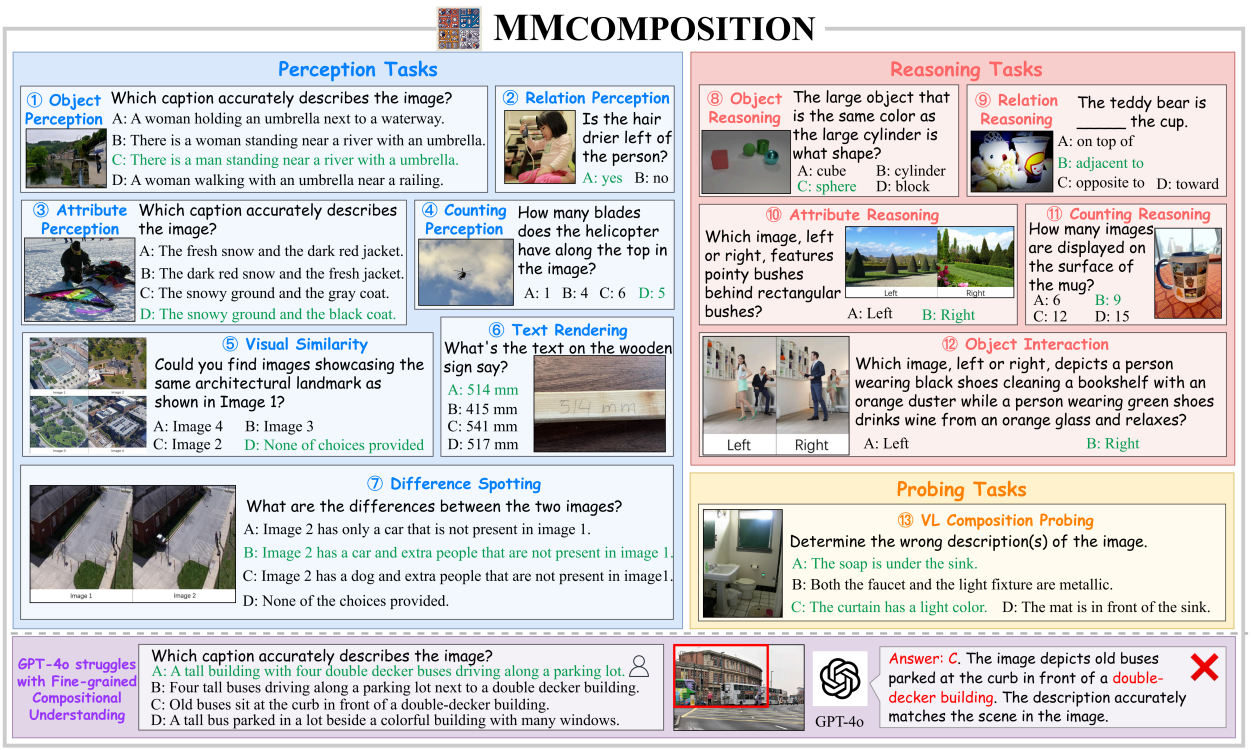

Figure 1: MMComposition comprises 13 categories of high-quality VL composition QA pairs, covering a wide range of complex compositions. In the example, GPT-4o failed to understand the compositional aspects of the visual and textual components, misidentifying a three-story building as a double-decker structure. This misinterpretation highlights the limitations of current VLMs.

compositional perception and reasoning abilities. Additionally, some existing benchmarks (Yuksekgonul et al., 2022; Hsieh et al., 2024; Zhao et al., 2022; Thrush et al., 2022; Ray et al., 2023; Ma et al., 2023) evaluate models' compositionality roughly from the perspective of attribute, relation, and object perception. These benchmarks have limitations in evaluating fine-grained visual composition and reasoning. They mainly focus on image-to-text retrieval tasks, assessing basic object, relation, and attribute recognition but neglecting deeper reasoning about object interactions, counting, and complex compositions. As a result, researchers currently have an incomplete understanding of VLMs' compositionality.

To address these issues, we propose MMComposition, a novel, human-annotated, high-quality benchmark for the comprehensive evaluation of VLMs' compositionality. MMComposition evaluates the compositionality of VLMs in three main dimensions: VL compositional perception, reasoning, and probing, which are further divided into 13 distinct categories of questions, as illustrated in Figure 1. While previous evaluation benchmarks have primarily focused on text-to-image retrieval, single-choice questions, and open-ended text generation, MMComposition introduces a more diverse and challenging set of tasks. The benchmark encompasses 4,122 questions, covering both single-image and multi-image scenarios, as well as single-choice and indefinite-choice formats. This expanded range of tasks is designed to evaluate the complex interplay between vision and language in VLMs more effectively. By incorporating a wider variety of complex composition questions, MMComposition provides a more comprehensive and in-depth assessment of models' capabilities in cross-modal compositionality, surpassing the evaluations offered by earlier benchmarks like ARO (Yuksekgonul et al., 2022) and Winoground (Thrush et al., 2022). Table 8 highlights the differences between MMComposition and other existing datasets that focus on VL compositionality.

In addition to the new benchmark, we also provide a comprehensive analysis of the models' capabilities in fine-grained VL compositional perception and reasoning. Our experiments show that most SOTA VLMs exhibit deficiencies in compositional understanding. Even GPT-4o, despite its advanced capabilities, struggles with

Table 1: Comparison with related VL compositional benchmarks: "Yes/No Ratio" refers to the proportion of yes/no questions, "Fine-grained" indicates whether the data provide detailed breakdowns of VL compositional information, and "IT Mismatch Detec." means "Image Text Mismatch Detection".

Dataset	Yes/No Ratio	Size	Human Annotation	Multi-Image	Indefinite-Choice	Task	Fine-grained
Winoground (Thrush et al., 2022)	-	400	✓	✓	-	Compositional Reasoning	✗
ARO (Yuksekgonul et al., 2022)	-	50k	✗	✗	-	T2I Retrieval	✗
Sugarcrepe (Hsieh et al., 2024)	-	7,512	✗	✗	-	T2I Retrieval	✗
VL-Checklist (Zhao et al., 2022)	-	410k	✗	✗	-	T2I Retrieval	✗
Cola (Ray et al., 2023)	-	1,200	✗	✗	-	T2I Retrieval	✗
FineMatch (Hua et al., 2024a)	-	49.9k	✓	✗	-	IT Mismatch Detec.	✓
GQA (Hudson & Manning, 2019)	0.7774	22M	✗	✗	✗	Compositional QA	✓
MMComposition (ours)	0.0483	4,122	✓	✓	✓	Compositional QA	✓

tasks requiring nuanced compositional reasoning. These findings highlight the need for further research and development to enhance the compositional abilities of VLMs. Our benchmark serves as a tool for identifying these gaps and inspiring future improvements in VLM design and training. Moreover, we analyze the critical factors in VLM architecture and training that may influence the compositionality of VLMs. According to the empirical results, we reach three findings: **(1) Visual Encoder Design**: While a mixture-of-encoder architecture can enhance compositionality, adding more encoders does not necessarily improve performance. Moreover, models that encode images with minimal degradation of image quality – preserving the original high resolution and aspect ratio – exhibit superior compositionality compared to those that utilize downsampling during the encoding process. **(2) Language Decoder Size**: Larger language decoders are associated with improved compositionality. **(3) The Volume of Training Data**: Fine-tuning models on more diverse datasets helps mitigate some compositionality limitations, driving more robust compositional understanding. In addition, although GPT-4o includes a powerful language model, we find that **for relatively simple QA tasks, only a small portion of its language capabilities are utilized** (compared to the models outperform GPT-4o, whose language model size is only 70B). **Once the language decoder size reaches a certain threshold (e.g., 34B, 70B), the visual encoder has a more significant impact on the model's compositionality**. We demonstrate in Figure 14 that the downsampling image processing in GPT-4o contributes to its inferior performance. Our experimental analysis highlights the limitations of large-scale VLMs in fine-grained compositional perception and reasoning. Our empirical analysis provides a systematic framework for evaluating and enhancing models' capability, pinpointing areas where large models still struggle.

Our main contributions are three-fold:

- We introduce **MMComposition**, a novel, human-annotated, high-quality benchmark designed to evaluate the compositionality of pre-trained VLMs. MMComposition assesses compositionality across three dimensions: compositional perception, reasoning, and probing, which are further divided into 13 distinct categories of questions. The benchmark comprises 4,122 questions, including 2,371 multi-hop reasoning questions, spanning both single-image and multi-image scenarios, as well as single-choice and indefinite-choice formats. This broad coverage ensures a comprehensive and robust evaluation framework for VLMs.

- We comprehensively evaluate 77 well-known VLMs with MMComposition. The empirical results highlight the challenging nature of MMComposition, as the highest model accuracy reached only 68.16%, compared to 90.31% for human performance. This evaluation reveals a **substantial gap** between state-of-the-art VLMs and human capabilities and provides insights into the limitations of current VLMs.

- We systematically analyze critical factors in VLM architecture that may influence the compositionality of VLMs, including the size of language decoders, the volume of training data, and the visual encoder design. Furthermore, we provide an interpretable analysis of models' limitations in complex compositional understanding. This analysis identifies critical areas for model improvement and suggests directions for future advancements.

2 Related Work

VLM Evaluation Benchmarks. The advent of large-scale VLMs has led to the development of numerous benchmarks designed to evaluate various model capabilities. Among the most commonly evaluated are image captioning (Lin et al.; Onoe et al., 2024; Masry et al., 2022), which tests a VLM's ability to generate natural language descriptions of images; VQA (Antol et al., 2015; Marino et al., 2019; Mathew et al., 2020), which assesses the model's capacity to answer image-based questions by integrating visual perception with language understanding or external knowledge; and Visual Reasoning (Johnson et al., 2017; Suhr et al., 2017), which evaluates a model's understanding of spatial relationships and logical reasoning based on visual input. In recent years, researchers have built benchmarks that aim to evaluate the comprehensive capabilities of VLMs (Li et al., 2023a; Liu et al., 2023b; Yue et al., 2023; Fu et al., 2023; Yu et al., 2024a; Lu et al., 2023; Guan et al., 2024). Although some benchmarks include QA pairs related to compositional reasoning, such as BLINK (Fu et al., 2024), MMVP (Tong et al., 2024b), and Seed-bench (Li et al., 2023a), these are often mixed with other types of QA pairs, making it challenging to assess a model's compositionality precisely. In contrast, MMComposition consolidates and refines existing categories of VL compositionality, offering a diverse set of compositional QA pairs that provide a more precise evaluation of model performance.

Compositionality for Vision-Language Models. Compositional understanding of images and text is a critical capability for VLMs. Research indicates that VLMs struggle to distinguish hard negative examples, i.e., image-text pairs that mismatch in at least one aspect (e.g., attribute, relation, object), as there is little incentive for them to learn compositionality during contrastive pre-training (Yuksekgonul et al., 2022). Hsieh et al. (2024) illustrate that contrastive pre-training with generated hard negative examples can improve models' performance on downstream tasks. Various benchmarks have been proposed to assess the capabilities of VLMs in compositional vision-language perception, including VL-Checklist (Zhao et al., 2022), ARO (Yuksekgonul et al., 2022), FineMatch (Hua et al., 2024a), Sugarcrepe (Hsieh et al., 2024), Crepe (Ma et al., 2023), Cola (Ray et al., 2023), CheckList (Zhao et al., 2022), etc. However, these benchmarks often evaluate models' capabilities from limited perspectives, such as object, attribute, and relation perception, and primarily focus on simple tasks like binary image-to-text retrieval, where models need to select the correct caption from pairs containing a correct and a hard negative caption. Moreover, the aforementioned benchmarks often contain a limited range of relations or attributes (e.g., ARO includes 48 relations and 117 attributes). GQA (Hudson & Manning, 2019) includes a diverse set of QA pairs focused on compositional reasoning, but the majority of the questions (77.74%) are simple Yes/No format. In contrast, MMComposition offers a more comprehensive assessment with various compositional scenarios, including multi-image and indefinite choice questions, providing a more comprehensive assessment. Furthermore, MMComposition evaluates the robustness in detecting complex relationships, including subtle scene composition, object interactions, and higher-order concepts beyond basic perception.

Pre-trained Vision-Language Models. Vision-language models (Radford et al., 2021; Liu et al., 2024a; Hua et al., 2024b; Ye et al., 2023; Tang et al., 2024; Chen et al., 2024b; Bi et al., 2024; Li et al., 2022; Tong et al., 2024a) aim to achieve multimodal intelligence by jointly understanding and generating visual and language information. Inspired by the remarkable success of recent large language models (LLMs) (Touvron et al., 2023; Chiang et al., 2023; Hua et al., 2021), researchers are now exploring large VLMs that combine pre-trained visual encoders and language decoders to tackle complex multimodal tasks. Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al.) are two of the early works that explore the integration of LLMs into vision-language pre-training. These models are trained as VL foundation models. Beginning with LLaVA (Liu et al., 2024a), researchers have used LLM-synthesized instruction-following chat data in VQA format for instruction tuning, achieving significantly improved results (Hua et al., 2024a). Subsequent studies have expanded to explore the broader capabilities of multimodal LLMs (Hu et al., 2023; Guan et al., 2024; Lin et al., 2023; Yu et al., 2024c; Tang et al., 2023). However, these efforts place less emphasis on improving the models' ability to fine-grained compositional perception and reasoning.

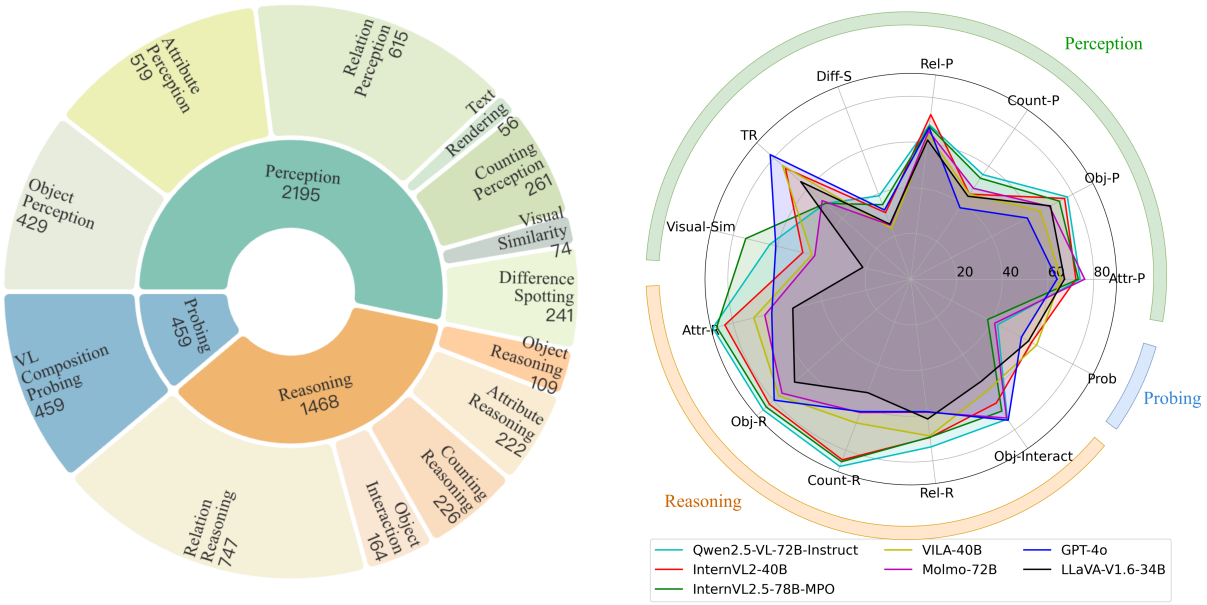

Figure 2: The statistics of 13 distinct categories of QA pairs in MMCOMPOSITION and some models' performance on each category.

3 MMComposition

3.1 Data Curation

To ensure a comprehensive and high-quality benchmark, we develop an efficient pipeline for curating VQA data that accurately reflects compositional information.

Data Collection. We use various datasets with the potential to construct VL compositional QA pairs as our seed data. This collection includes datasets that contain the description of objects, attributes, relations, and counting, such as VL-CheckList (Zhao et al., 2022), Sugar-Crepe (Hsieh et al., 2024), ARO (Yuksekgonul et al., 2022), Crepe (Ma et al., 2023), and DOCCI (Onoe et al., 2024). Additionally, we incorporate sources that are well-suited for constructing VL compositional reasoning QA pairs, including SVO-Probes (Hendricks & Nematzadeh, 2021), VSR (Liu et al., 2023a), BLINK (Fu et al., 2024), GQA (Hudson & Manning, 2019), Visual Genome (Krishna et al., 2016), and CLVER (Johnson et al., 2017). It also contains datasets with multiple images in each sample, such as Winoground (Thrush et al., 2022), MuriBench (Wang et al., 2024) and NLVR2 (Suhr et al., 2017).

Question and Answer Construction. We obtain QA pairs from the seed data in through several methodologies:

For the seed data that only contain positive and negative captions (e.g., ARC (Yuksekgonul et al., 2022)), we first generate sentence embeddings for each caption using Sentence-BERT (Reimers & Gurevych, 2019). We then utilize these embeddings to retrieve the most similar captions from the Visual Genome (Krishna et al., 2016) dataset. This process results in four captions per image in each sample, forming four answer options per question.

For data samples containing multiple images – such as those in the image difference spotting task, which includes two images per question – we concatenate the two images side by side and label them *Left* and *Right* beneath each sub-image. This setup allows for two types of question-answer options: *Left* and *Right* for questions asking which sub-image is described by a caption, and *True* and *False* for questions determining the accuracy of a caption describing the image difference. For tasks that include more than two images per question (e.g., visual similarity assessments), we concatenate all images into a single composite image and label each sub-image as $\text{Image}_1, ..., \text{Image}_i$.

For the probing task, we select several captions from the dense captions in Visual Genome (Krishna et al., 2016) as the correct options and write the misaligned captions manually for the image. Then, we randomly select $x \in \{1, 2, 3, 4\}$ captions from the set of accurate captions for a given image and complement these with $4 - x$ incorrect options drawn from a set of conflict captions. With this approach, we can obtain the indefinite-choice QA pairs.

Data Filtering and Difficulty Classification. We divide the data into different difficulty levels: easy, medium, hard, and super hard. To achieve this, we use a voting system with six models, ranging from weaker to stronger, including LLaVA-1.5-13B (Liu et al., 2024b), LLaVA-1.6-Mistral-7B (Liu et al., 2024a), LLaVA-1.6-Vicuna-13B (Liu et al., 2024a), Phi-3-Vision-128K-Instruct (Abdin et al., 2024), InternVL-Chat-V1.5 (Chen et al., 2024b), and Qwen-VL-Chat (Bai et al., 2023). Based on the accuracy of model predictions for each question, questions are categorized into different difficulty levels. Questions with zero correct predictions are classified as super hard, those with one or two correct predictions are labeled as hard, questions with three or four correct predictions are considered medium, and those with more than five correct predictions are categorized as easy. The overall difficulty of the dataset is then controlled by adjusting the ratio of questions at each difficulty level.

Human Annotation. All QA pairs in the benchmark are human-annotated. Annotators first assess image quality to ensure it meets the required standards. For human-created data, annotators are first trained with detailed instructions to develop a thorough understanding of the compositional aspects in our dataset. During annotation, they generate QA pairs based on the provided aspect prompts. For GPT-synthesized data sourced from DOCCI, annotators verify whether the question accurately reflects the compositional information in the image and whether the answer appropriately corresponds to the question.

3.2 Evaluation Metric

Let $\mathcal{D} = \{\mathcal{D}_m = \{\mathcal{T}_t\}_{t=1}^{T_d}\}_{m=1}^{|D|}$ denotes our dataset , where each catagory $\mathcal{D}_m$ consists of $\mathcal{T}_d$ subtasks. For each subtask, we calculate the accuracy across all annotations. For each question $q \in \mathcal{D}$, let $\mathcal{A}_q$ be the set of correct options, $\mathcal{P}_q$ be the set of predicted (selected) options. The score for question q, denoted as s_q, is calculated as:

$$s_q = \begin{cases} 1, & \text{if } \mathcal{P}_q = \mathcal{A}_q \\ \frac{|\mathcal{P}_q|}{|\mathcal{A}_q|}, & \text{if } \mathcal{P}_q \subset \mathcal{A}_q \\ 0, & \text{otherwise} \end{cases}$$

Here, $|\cdot|$ denotes the number of options selected by the participant and the number of correct options, $\mathcal{P}_q \subset \mathcal{A}_q$ means all selected options are correct, but some correct options are missing (under-selection). The "otherwise" case covers instances where incorrect options are selected (wrong or over-selection). This equation applies to both the single-choice and indefinite-choice questions. The final weighted average accuracy across all categories is calculated as $\text{ACC} = \sum_{m=1}^{|D|} \sum_{t=1}^{T_d} s_q \times |\mathcal{T}_t|/|\mathcal{D}_d|$, where $|\cdot|$ is the question number in one set.

3.3 Quantitative Analysis

MMComposition evaluates VL compositional capabilities across three overarching and complementary hypertasks: (1) **Perception**, including Attribute Perception (**Attr-P**), Object Perception (**Obj-P**), Counting Perception (**Count-P**), Relation Perception (**Rel-P**), Difference Spotting (**Diff-S**), Text Rendering (**TR**), and Visual Similarity (**Visual-Sim**); (2) **Reasoning**, comprising Attribute Reasoning (**Attr-R**), Object Reasoning (**Obj-R**), Counting Reasoning (**Count-R**), Relation Reasoning (**Rel-R**), and Object Interaction (**Obj-Interact**); and (3) **Probing**, consisting of Compositional Probing (**Prob**). We use GPT-4o to label each question category via in-context learning, followed by manual verification for accuracy. Figure 5 illustrates the difficulty distribution of MMComposition, highlighting the challenging nature of our dataset. Figure 6 depicts the distribution of option counts per question, with over half of the data containing more than four options. To analyze the impact of input resolution on model performance, we further display the resolution distribution of images in Figure 7, which reflects the image quality of our data. For textual analysis, we visualize the phrase distribution of questions using a word cloud diagram in Figure 3, clearly depicting the

word frequency and distribution across the questions. We also provide a detailed explanation for these 13 categories in Section A.3.

Table 2: The comprehensive performance of 38 representative VLMs on Acc, including open source models and API-based models. The **best** and second best results are in bold and underlined, respectively. The full table containing comprehensive performance of 77 VLMs is shown in Appendix Table 13.

Method	Perception↑							Reasoning↑					Probing↑	Overall ↑
	Attr-P	Obj-P	Count-P	Rel-P	Diff-S	TR	Visual-Sim	Attr-R	Obj-R	Count-R	Rel-R	Obj-Interact	Prob	
Human	97.94	98.04	93.06	92.00	79.02	85.71	86.54	91.20	78.83	100.00	77.35	88.00	91.84	90.31
Qwen2.5-VL-72B-Instruct (team, 2024)	**74.05**	**77.39**	**55.56**	68.01	39.00	50.00	63.51	**89.19**	**86.24**	**87.61**	**73.90**	74.39	42.92	**68.16**
InternVL2-40B (Chen et al., 2024b)	72.22	75.99	45.21	**72.53**	31.12	73.21	48.65	83.78	82.57	84.51	69.75	65.85	59.59	67.60
Qwen2-VL-72B-Instruct (team, 2024)	59.57	63.87	52.49	62.52	**45.23**	**82.14**	67.57	87.84	84.40	84.51	71.49	70.12	**69.57**	66.86
InternVL2-76B (Chen et al., 2024b)	70.65	75.52	48.28	70.00	19.09	78.57	48.65	85.14	83.49	85.40	70.01	67.07	58.46	66.65
InternVL2.5-78B-MPO (Chen et al., 2024b)	73.28	73.43	53.64	67.25	34.85	50.00	**74.32**	87.39	84.40	85.40	69.61	70.12	37.98	65.61
InternVL2.5-78B (Chen et al., 2024b)	70.07	66.90	47.13	64.23	32.37	48.21	60.81	85.59	82.57	80.09	68.27	73.17	37.58	62.64
Qwen2-VL-7B-Instruct (team, 2024)	68.30	71.79	41.38	64.63	32.37	39.29	52.70	81.08	76.15	80.53	67.34	69.51	41.43	62.09
VILA-40B (Lin et al., 2024)	65.70	64.10	45.21	63.65	23.65	75.00	44.59	70.72	77.06	67.26	69.08	59.15	62.16	61.83
InternVL2.5-38B (Chen et al., 2024b)	66.51	67.60	46.74	60.28	30.29	53.57	59.46	84.23	83.49	80.97	65.19	71.95	41.68	61.43
Ovis1.6-Gemma2-27B (Lu et al., 2024)	66.25	61.07	49.04	60.13	28.22	42.86	54.05	81.53	80.73	80.53	68.81	57.93	41.14	60.27
Qwen2.5-VL-7B-Instruct (team, 2024)	69.91	66.90	43.30	59.74	22.41	41.07	48.65	82.88	81.65	80.09	64.39	68.29	40.41	60.06
Molmo-72B (Deitke et al., 2024)	76.08	68.53	48.28	65.20	25.31	51.79	43.24	65.77	75.23	62.39	58.63	73.78	41.47	59.59
GPT-4o (Achiam et al., 2023)	63.97	57.58	37.93	66.76	32.37	**82.14**	60.81	62.61	79.82	61.95	58.37	**75.00**	54.65	59.03
POINTS1.5-7B-Chat (Liu et al., 2024c)	70.13	61.54	39.46	60.39	24.90	46.43	44.59	76.13	77.06	76.11	60.24	69.51	45.21	58.66
InternVL2.5-8B-MPO (Chen et al., 2024b)	65.64	66.20	45.21	58.12	21.99	41.07	56.76	78.83	80.73	76.55	65.19	59.76	37.44	58.49
InternVL2-8B (Chen et al., 2024b)	62.68	59.21	31.80	59.54	25.31	73.21	33.78	78.83	75.23	73.89	60.37	62.20	54.10	57.76
Llama-3.2-90B-Vision-Instruct	68.85	69.46	39.85	62.87	23.65	53.57	41.89	64.86	69.72	54.87	57.56	64.63	46.84	57.23
MiniCPM-V2.6 (Yao et al., 2024)	65.19	58.04	41.00	61.80	21.99	73.21	37.84	63.96	73.39	68.14	52.07	60.98	54.43	56.07
InternLM-XComposer2-4KHD-7B (Dong et al., 2024b)	62.24	55.24	39.08	58.36	23.65	67.86	27.03	70.72	74.31	60.18	55.82	59.15	60.02	55.35
Qwen-VL-Max (Bai et al., 2023)	53.76	53.15	36.40	58.67	22.82	80.36	41.89	53.60	65.14	53.98	60.91	62.80	63.87	54.75
Hunyuan-Vision	61.95	61.31	37.16	58.58	26.97	76.79	36.49	61.26	72.48	56.19	52.21	59.15	45.03	53.67
Gemini-1.5-Pro (Reid et al., 2024)	55.30	53.50	39.46	57.11	24.48	67.86	55.41	59.91	74.31	50.44	56.29	65.24	49.60	53.27
Mini-Gemini-34B (Li et al., 2023b)	58.35	55.01	37.93	53.70	25.31	73.21	39.19	54.50	73.39	58.41	55.82	61.59	41.79	51.96
Molmo-7B-D (Deitke et al., 2024)	68.02	55.71	37.16	52.40	24.90	48.21	40.54	56.76	67.89	46.02	53.41	60.98	42.70	51.61
MiniCPM-Llama3-V2.5 (Yao et al., 2024)	51.93	50.12	36.40	49.88	19.92	76.79	20.27	69.37	77.06	68.14	56.49	62.20	41.79	50.95
Bunny-Llama-3-8B-V (He et al., 2024)	58.16	51.05	34.87	54.07	21.58	50.00	12.16	45.95	66.06	53.10	48.73	57.32	59.44	49.93
Mini-Monkey (Huang et al., 2024)	52.25	56.64	26.82	52.53	26.56	73.21	18.92	68.92	65.14	59.29	50.60	50.00	42.37	49.46
Phi3.5-Vision-Instruct (Abdin et al., 2024)	55.01	45.69	30.27	52.61	21.16	66.07	31.08	45.05	63.30	53.10	53.95	53.66	54.65	49.12
Yi-VL-34B (AI et al., 2024)	53.02	38.23	30.27	50.33	26.14	64.29	17.57	50.45	56.88	55.31	51.00	52.44	53.88	47.38
Step-1V-32K	46.11	39.86	26.44	46.25	25.31	67.86	43.24	66.67	66.97	62.83	50.60	59.76	45.46	47.12
ConvLLaVA-1024-7B (Ge et al., 2024)	51.73	44.29	32.57	44.96	28.22	69.64	21.62	55.41	65.14	53.10	49.53	54.88	40.89	46.21
LLaVA-HR-13B (Luo et al., 2024)	50.32	41.26	35.25	39.81	32.37	66.07	27.03	45.50	60.55	45.58	48.46	57.32	48.80	45.12
Monkey-Chat (Li et al., 2024)	49.20	47.55	24.14	47.13	16.60	69.64	13.51	51.35	58.72	44.25	44.18	51.22	48.91	44.10
SliME-7B (Zhang et al., 2024b)	45.70	44.52	28.74	40.76	31.12	62.50	20.27	43.24	59.63	48.23	50.74	53.05	30.03	42.52
INF-LLaVA* (Ma et al., 2024)	43.19	44.76	32.95	41.92	24.48	57.14	20.27	50.00	66.06	55.31	45.65	54.27	31.41	42.41
DeepStack-L-HD-Vicuna-7B (Meng et al., 2024)	43.29	34.97	28.74	35.74	18.67	60.71	17.57	46.85	60.55	45.13	42.97	59.15	35.88	39.21
mPLUG-Owl2 (Ye et al., 2024)	40.04	36.83	28.74	42.93	26.97	30.36	12.16	41.89	60.55	38.94	44.58	50.61	30.36	38.77
InstructBLIP-13B (Dai et al., 2023)	39.21	40.56	22.99	38.86	24.07	35.71	33.78	40.54	48.62	37.17	41.23	51.83	25.24	36.76
Random Choice	23.12	24.01	21.84	25.85	29.46	35.71	25.68	36.94	46.79	38.50	34.00	47.65	28.61	29.90

4 Revisiting the Compositionality of Pre-trained Vision-Language Models

In this section, we quantify and explore the compositionality of state-of-the-art VLMs and provide a comprehensive evaluation of VLMs. For all experiments, we use a consistent prompt template and the official default hyperparameters for each model.

Overall Performance. The overall performance indicates that models struggle with perceiving and reasoning about fine-grained VL compositional information. The best human expert achieves an accuracy of 90.31%, significantly outperforming all the models reported in the table. This demonstrates the still existing gap between human expertise and the performance of current models on the MMCOMPOSITION benchmark. This reflects the benchmark's rigorous standards. The open-source InternVL2 (Chen et al., 2024c) series models secured first and second place on the leaderboard. InternVL2-40B performs better than InternVL2-76B. Among the API-based models, Qwen2-VL and GPT-4o achieved the best and second best performance. The superior performance of open-source models with relatively smaller language models compared to GPT-4o, which has a larger language model, is due to their more effective visual encoders. The mean accuracy of 7B and 13B open-source VLMs hovers around 36–38%. For reference, we provide the random guess accuracy (29.90%) as a lower bound for the benchmark.

The tasks where VLMs exhibit relative strengths and weaknesses. From Table 2, we observe that VLMs perform relatively better on tasks such as Attribute, Object, and Relation Perception, as well as Attribute, Object, and Count Reasoning, where they perform much better than other categories. However, they struggle with tasks such as Count Perception, Difference Spotting, Visual Similarity, and Probing (see illustrations in Fig. 1). These tasks often involve multiple images, some with extreme aspect ratios, and the probing tasks include indefinite-choice questions, which pose additional challenges for the models. GPT-4o

performs relatively weaker on Obj-P, Count-P, Attr-R, Count-R, and Rel-R tasks compared to smaller models that outperform it, aligning with the limitations outlined in the official GPT-4o documentation. Overall, the models perform relatively well on mid-level perception and reasoning tasks.

5 Diagnostic Analysis of Factors Influencing Model Compositionality

In this section, we analyze the factors that may influence the compositionality of VLMs. We focus on three dominant factors: visual encoder design, language decoder size, and training data volume.

5.1 Visual Encoder Design

High-Resolution Visual Encoders. A common strategy to enhance a model's perception of fine-grained visual content is to incorporate higher-resolution encoders. In this study, we adopt a controlled experimental setup, varying only the input resolution of the visual encoders while keeping the training data and text decoders fixed. As shown in Table 3, models equipped with higher-resolution encoders generally achieve better performance in multimodal compositional perception and reasoning. However, for the Mini-Gemini series, introducing a high-resolution encoder with a patch information mining mechanism surprisingly led to a performance drop. We attribute this to Mini-Gemini's dual encoder architecture, which combines high- and low-resolution encoders. The patch info mining mechanism fuses high-resolution features into the compressed low-resolution representation, limiting overall representational capacity. In contrast, other models benefit from longer visual token sequences enabled by higher resolutions, which enhance the expressiveness of their visual encoders.

Table 3: Performance comparison of models with and without high-resolution encoders (Avg. refers to average resolution).

Method	Resolution	Visual Tokens	Perception Avg. 1098*847	Reasoning Avg. 926*534	Probing Avg. 828*523	Overall
ConvLLaVA-768-7B (Ge et al., 2024)	768	144	36.07	51.02	37.11	41.51
ConvLLaVA-1024-7B (Ge et al., 2024)	1024	256	$42.96_{+6.89}$	$52.72_{+1.70}$	$40.89_{+3.78}$	$46.21_{+4.70}$
ConvLLaVA-1536-7B (Ge et al., 2024)	1536	576	$41.47_{+5.40}$	$52.25_{+1.23}$	$34.20_{-6.69}$	$44.50_{+2.99}$
LLaVA-1.5-13B (Liu et al., 2024a)	336	576	30.08	42.37	41.39	35.72
LLaVA-HR-13B (Luo et al., 2024)	1024	1024	$41.46_{+11.38}$	$49.46_{+7.09}$	$48.80_{+7.41}$	$45.12_{+9.40}$
DeepStack-L-Vicuna-7B (Meng et al., 2024)	672	2880	36.33	44.62	30.21	38.60
DeepStack-L-HD-Vicuna-7B (Meng et al., 2024)	1344	14400	$34.69_{-1.64}$	$47.00_{+2.38}$	$35.88_{+5.67}$	$39.21_{+0.61}$
Mini-Gemini-13B (Li et al., 2023b)	768	576	37.71	53.54	32.28	42.75
Mini-Gemini-13B-HD (Li et al., 2023b)	1536	576	$36.48_{-1.23}$	$49.73_{-3.81}$	$34.28_{+2.00}$	$40.95_{-1.80}$
Mini-Gemini-34B (Li et al., 2023b)	768	576	50.07	57.97	41.79	51.96
Mini-Gemini-34B-HD (Li et al., 2023b)	1536	576	$46.49_{-3.58}$	$60.63_{+2.66}$	$35.91_{-5.88}$	$50.35_{-1.61}$

Mixture-of-Encoder. Another approach to enhancing visual encoders is the use of a mixture-of-encoder architecture. In this setup, image features are extracted by a combination of high-resolution and low-resolution encoders, providing rich visual information to the language decoders. We analyze the relationship between the mixture-of-encoder architecture and model performance by aggregating different encoders while keeping the training data and decoders fixed. We use the LLaVA-1.5 pretraining data for stage-1 pretraining and the EAGLE 1.8M dataset (Bi et al., 2024) for stage-2 fine-tuning. The initial encoder is a CLIP model with 448 resolution (Radford et al., 2021), and the decoder is LLaMA-3-8B (Dubey et al., 2024). We scale up the encoders using: *(A)* ConvNeXt (Liu et al., 2022), *(B)* SAM (Kirillov et al., 2023), *(C)* DINOv2 (Oquab et al., 2023), and *(D)* Pix2Struct (Lee et al., 2023). The empirical results in Table 4 indicate that combining CLIP with encoder A improves the models' performance; however, as the number of visual encoders increases, the models' performance declines.

Visual encoder has a more significant impact on the model's compositionality, while GPT-4o struggles with processing higher-resolution images. By summarizing the empirical results of this study, we find that for relatively simple QA tasks, only a small portion of its language capabilities are utilized (compared to the models outperforming GPT-4o, whose language model size is only 70B). Once the language

Table 4: A comparative analysis of various mixture-of-encoder architectures in relation to model compositionality.

Method	Visual Encoders	Relolution	Perception	Reasoning	Probing	Overall
LLaVA-1.5 (Liu et al., 2024a)	CLIP	448	44.43	53.07	53.34	48.50
LLaVA-1.5+A	CLIP+A	1024	$45.34_{+0.91}$	$52.25_{-0.82}$	$56.93_{+3.59}$	$49.13_{+0.63}$
LLaVA-1.5+A+B	CLIP+A+B	1024	$45.32_{+0.95}$	$51.23_{-1.84}$	$49.02_{-4.32}$	$47.83_{-0.67}$
LLaVA-1.5+A+B+C	CLIP+A+B+C	1024	$42.88_{-1.55}$	$50.34_{-2.73}$	$54.21_{+0.87}$	$46.80_{-1.70}$
LLaVA-1.5+A+C+D	CLIP+A+C+D	1024	$46.94_{+2.51}$	$50.34_{-2.73}$	$54.86_{+1.55}$	$49.04_{+0.54}$

decoder size reaches a certain threshold (e.g., 34B, 70B), the visual encoder plays a more critical role in the models' performance. As discussed in Section A.2, Qwen2VL processes images by largely preserving their original resolution and aspect ratio. The Internvl-2 series models employ a dynamic 'any-resolution' encoding strategy: images are first mapped to an optimal aspect ratio from predefined ratios, then divided into 448×448 pixel tiles, with each tile converted into 256 image tokens. These approaches enable the encoders to handle images of any resolution and aspect ratio with minimal degradation of image quality. In contrast, GPT-4o processes images with downsampling when the image's longest side $>$ 2048px or shortest side $>$ 768px (our data contains 889 such examples), contributing to its inferior performance compared to other open-source models.

5.2 The Volume of Training Data

The volume of training data is a crucial factor influencing models' performance. In this study, we conduct a comparison analysis of this factor. In Table 5, we observe a significant performance increase when the training data is scaled up substantially. For instance, InternVL-Chat-V1.2 and InternVL-Chat-V1.2-Plus, which use 10 times more training data than the former, show significant performance improvements.

Table 5: The comparison of models with and without training data scale up.

Method	Dataset Size	Perception	Reasoning	Probing	Overall
INF-LLaVA (Ma et al., 2024)	1.25M	41.38	45.44	35.58	42.18
INF-LLaVA* (Ma et al., 2024)	2.56M	$39.45_{-1.93}$	$50.27_{+4.83}$	$31.41_{-4.17}$	$42.41_{+0.23}$
InternVL-Chat-V1.2 (Chen et al., 2024c)	1.2M	55.74	63.69	60.71	59.13
InternVL-Chat-V1.2-Plus (Chen et al., 2024c)	12M	$60.27_{+4.53}$	$70.64_{+6.95}$	$65.80_{+5.09}$	$64.58_{+5.45}$
InternVL-Chat-V1.5 (Chen et al., 2024b)	–	53.09	67.44	57.01	58.64
InternVL2-26B (Chen et al., 2024b)	–	$59.51_{+6.42}$	$69.48_{+2.04}$	$52.43_{-4.58}$	$62.27_{+3.63}$

5.3 Language Decoder Size

From Table 2, we observe that models with larger decoders demonstrate stronger performance. To analyze this relationship more accurately, we compare models with different decoder sizes while keeping the encoder and training data constant. The results are shown in Table 6, from which we can conclude that larger language decoders result in better performance.

5.4 Interpretable Analysis of Model Deficiencies

We conduct a comprehensive error analysis to better understand the models' deficiencies in fine-grained compositional understanding. In this analysis, the models are required to answer questions and provide explanations in a multi-turn dialogue format. Figures 4, 15, and 16 in Section A.5 illustrate the reasons why the models fail to predict the correct answers for each task. For example, in the Obj-P task (example 3), while the "yellow colored outline" is easily detected by humans, the models struggle to accurately identify the target objects due to the outline being mixed with numerous other characters. Additionally, the models face difficulties with fine-grained object counting, especially when several similar objects are present. In the Count-R (example 6) task, for instance, humans can precisely count the number of triangles on a wheel, but the models confuse the six irregular polygons for triangles.

Table 6: The comparison analysis of text decoder size and models' compositionality.

Method	Decoder	Perception	Reasoning	Probing	Overall
InternVL2-1B (Chen et al., 2024b)	Qwen2-0.5B-Instruct	39.05	49.18	27.89	41.41
InternVL2-2B (Chen et al., 2024b)	InternLM2-Chat-1.8B	41.57$_{+2.52}$	50.07$_{+0.89}$	38.10$_{+10.21}$	44.22$_{+2.81}$
InternVL2-4B (Chen et al., 2024b)	Phi3-Mini-128K-Instruct	45.79$_{+6.74}$	61.85$_{+12.67}$	41.18$_{+13.29}$	51.00$_{+9.59}$
InternVL2-8B (Chen et al., 2024b)	InternLM2.5-Chat-7B	52.64$_{+13.59}$	66.55$_{+17.37}$	54.10$_{+26.21}$	57.76$_{+16.35}$
InternVL2-26B (Chen et al., 2024b)	InternLM2-Chat-20B	59.51	69.48	52.43	62.27
InternVL2-40B (Chen et al., 2024b)	Nous-Hermes-2-Yi-34B	64.55$_{+5.04}$	74.66$_{+5.18}$	59.59$_{+7.16}$	67.60$_{+5.33}$
InternVL2-76B (Chen et al., 2024b)	Hermes-2-Theta-Llama-3-70B	62.56$_{+3.05}$	75.34$_{+5.86}$	58.46$_{+6.03}$	66.65$_{+4.38}$
LLaVA-V1.6-Mistral-7B (Liu et al., 2024a)	Mistral-7B-Instruct	33.58	40.94	38.24	36.72
LLaVA-V1.6-Vicuna-13B (Liu et al., 2024a)	Vicuna-13B-V1.5	30.98$_{-2.60}$	46.32$_{+5.38}$	38.16$_{-0.08}$	37.24$_{+0.52}$
LLaVA-V1.6-34B (Liu et al., 2024a)	Nous-Hermes-2-Yi-34B	56.91$_{+23.33}$	57.56$_{+16.62}$	58.17$_{+19.93}$	57.28$_{+20.56}$
Mini-Gemini-13B (Li et al., 2023b)	Vicuna-13B-V1.5	37.71	53.54	32.28	42.75
Mini-Gemini-34B (Li et al., 2023b)	Nous-Hermes-2-Yi-34B	50.07$_{+12.36}$	57.97$_{+4.43}$	41.79$_{+9.51}$	51.96$_{+9.21}$
SliME-7B (Zhang et al., 2024b)	Vicuna-7B-V1.5	40.04	50.14	30.03	42.52
SliME-8B (Zhang et al., 2024b)	Llama-3-8B-Instruct	40.00$_{-0.04}$	49.86$_{-0.28}$	29.96$_{-3.07}$	42.39$_{-0.13}$
SliME-13B (Zhang et al., 2024b)	Vicuna-13B-V1.5	39.18$_{-0.86}$	48.09$_{-2.05}$	33.55$_{+3.52}$	41.73$_{-0.79}$
LLaVA-HR-7B (Luo et al., 2024)	Vicuna-7B-V1.5	38.94	48.57	33.04	41.71
LLaVA-HR-13B (Luo et al., 2024)	Vicuna-13B-V1.5	41.46$_{+2.52}$	49.46$_{+0.89}$	48.80$_{+15.25}$	45.12$_{+3.41}$
Yi-VL-6B (AI et al., 2024)	Yi-6B-Chat	43.34	50.07	48.76	46.34
Yi-VL-34B (AI et al., 2024)	Yi-34B-Chat	42.81$_{-0.53}$	52.18$_{+2.11}$	53.88$_{+5.12}$	47.38$_{+1.04}$

6 Conclusion

This paper introduces MMComposition, a novel high-quality benchmark for evaluating VLM compositionality. With MMComposition, we comprehensively evaluate the compositionality of notable VLMs. Our evaluation reveals a significant gap between these models and human performance, providing insights into the limitations of existing VLMs. Additionally, we systematically analyze factors that may influence compositionality, including visual encoder design, training data volume, and language decoder size. We find that for relatively simple QA tasks, only a small portion of the language model's capacity is utilized (as seen in models outperforming GPT-4o, whose language model has 70B parameters). Once the language decoder reaches a certain size threshold (e.g., 34B, 70B), the visual encoder has a more pronounced impact on compositionality. In summary, our work provides a comprehensive and precise framework for evaluating the compositionality of VLMs, identifies key areas for improvement, and suggests potential directions for future advancements.

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. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

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.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, et al. Yi: Open foundation models by 01.ai, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

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.

Jing Bi, Yunlong Tang, Luchuan Song, Ali Vosoughi, Nguyen Nguyen, and Chenliang Xu. EAGLE: Egocentric AGgregated language-video engine. In *ACM Multimedia 2024*, 2024. URL https://openreview.net/forum?id=mk8p2JKdu0.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024c.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Noam Chomsky. *Aspects of the Theory of Syntax*. Number 11. MIT press, 2014.

Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

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

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024b.

Abhimanyu Dubey, Abhinav Jauhri, Pandey, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chaoyou Fu, Peixian Chen, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. URL https://api.semanticscholar.org/CorpusID:259243928.

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

Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. Convllava: Hierarchical backbones as visual encoder for large multimodal models. *ArXiv*, abs/2405.15738, 2024. URL https://api.semanticscholar.org/CorpusID:270045537.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 36, 2024.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2963–2975, 2023.

Hang Hua, Xingjian Li, Dejing Dou, Cheng-Zhong Xu, and Jiebo Luo. Noise stability regularization for improving bert fine-tuning. *arXiv preprint arXiv:2107.04835*, 2021.

Hang Hua, Jing Shi, Kushal Kafle, Simon Jenni, Daoan Zhang, John Collomosse, Scott Cohen, and Jiebo Luo. Finematch: Aspect-based fine-grained image and text mismatch detection and correction. *arXiv preprint arXiv:2404.14715*, 2024a.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *arXiv preprint arXiv:2404.12353*, 2024b.

Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Multi-scale adaptive cropping for multimodal large language models. *arXiv preprint arXiv:2408.02034*, 2024.

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

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32 – 73, 2016. URL https://api.semanticscholar.org/CorpusID:4492210.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912. PMLR, 2023.

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

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

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:2403.18814*, 2023b.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. Videoxum: Cross-modal visual and textural summarization of videos. *IEEE Transactions on Multimedia*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV 2014*. Springer.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a.

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.

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

Yuan Liu, Le Tian, Xiao Zhou, Xinyu Gao, Kavio Yu, Yang Yu, and Jie Zhou. Points1. 5: Building a vision-language model towards real world applications. *arXiv preprint arXiv:2412.08443*, 2024c.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024.

Yiwei Ma, Zhibin Wang, Xiaoshuai Sun, Weihuang Lin, Qiang Zhou, Jiayi Ji, and Rongrong Ji. Inf-llava: Dual-perspective perception for high-resolution multimodal large language model, 2024.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ArXiv*, abs/2203.10244, 2022. URL https://api.semanticscholar.org/CorpusID:247593713.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208, 2020. URL https://api.semanticscholar.org/CorpusID:220280200.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms, 2024.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: Descriptions of Connected and Contrasting Images. In *arXiv:2404.19753*, 2024.

Maxime Oquab, Timothée Darcet, Moutakanni, et al. Dinov2: Learning robust visual features without supervision, 2023.

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

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A Plummer, Ranjay Krishna, and Kate Saenko. Cola: How to adapt vision-language models to compose objects localized with attributes? *arXiv preprint arXiv:2305.03689*, 2023.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Alayrac, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://api.semanticscholar.org/CorpusID:201646309.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Annual Meeting of the Association for Computational Linguistics*, 2017. URL https://api.semanticscholar.org/CorpusID:19435386.

Yunlong Tang, Jing Bi, Siting Xu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023.

Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu. Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue. *arXiv preprint arXiv:2403.16276*, 2024.

Qwen team. Qwen2-vl. 2024.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

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

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

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

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

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

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13040–13051, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR, 2024a.

Weihao Yu, Zhengyuan Yang, Linfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024b.

Yongsheng Yu, Ziyun Zeng, Hang Hua, Jianlong Fu, and Jiebo Luo. Promptfix: You prompt and we fix the photo. *arXiv preprint arXiv:2405.16785*, 2024c.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 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? In *The Eleventh International Conference on Learning Representations*, 2022.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024a.

Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024b.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.