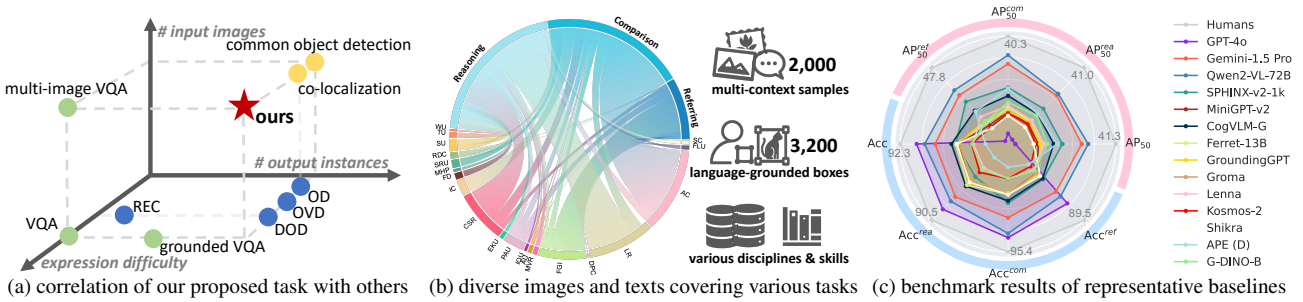
MC-Bench: A Benchmark for Multi-Context Visual Grounding in the Era of MLLMs

Yunqiu Xu[1] Linchao Zhu[1,2*] Yi Yang[1,2]
[1]ReLER Lab, CCAI, Zhejiang University
[2]The State Key Lab of Brain-Machine Intelligence, Zhejiang University
imyunqiuxu@gmail.com zhulinchao@zju.edu.cn yangyics@zju.edu.cn



(a) correlation of our proposed task with others (b) diverse images and texts covering various tasks (c) benchmark results of representative baselines

Figure 1. Multi-context visual grounding is a new task that aims at localizing target instances based on open-ended text prompts in multi-image scenarios. A new dataset MC-Bench is constructed to benchmark the MLLMs and foundation models with potential multi-context visual grounding capabilities. The benchmark results of over 20 state-of-the-art models reveal a significant performance gap between existing approaches and humans, while also suggesting potential future directions.

Abstract

While multimodal large language models (MLLMs) have demonstrated extraordinary vision-language understanding capabilities, their abilities to solve instance-level visual-language problems beyond a single image warrant further exploration. To assess these unproven abilities of MLLMs, this paper proposes a new visual grounding task called multi-context visual grounding, which aims to localize instances of interest across multiple images based on open-ended text prompts. In order to facilitate this research, we construct a new dataset MC-Bench that features 2K high-quality and manually annotated samples. Each sample consists of an instance-level labeled image pair and a corresponding text prompt that indicates the target instances in the images. These text prompts are highly open-ended and follow three distinct styles, covering 20 practical skills. We benchmark over 20 state-of-the-art MLLMs and foundation models with potential multi-context visual grounding capabilities, along with our developed simple yet effective agentic baseline and a finetuned baseline by multi-context instruction tuning. Our evaluation reveals a non-trivial performance gap between existing MLLMs and humans, along with some insightful observations that suggest potential future directions. We hope that MC-Bench and our empirical findings encourage the research community to further advance the untapped potentials of MLLMs in instance-level tasks, particularly in multi-image contexts. Project page: https://xuyunqiu.github.io/MC-Bench.

1. Introduction

Grounding visual content guided by textual inputs is a long-standing research topic involving vision-language understanding and visual localization tasks. Early works typically focus on locating instances of interest using simple textual expressions, such as object detection (OD) [8, 72, 74, 101, 102] and open-vocabulary object detection (OVD) [17, 33] based on category names, as well as referring expression comprehension (REC) [26, 44, 45, 80, 87] and describe object detection (DOD) [76, 100] with referring phrases. However, text descriptions in real-world applications are often more flexible and ambiguous. Grounding objects using free-form textual descriptions in an open world is challenging, as models must comprehend the intentions of ambiguous text inputs and grasp the overall context within the images. Recently, the development of foundation models [36, 55, 71, 78, 99] has catalyzed a shift from specialized models to general-purpose models, showcasing unprecedented generalization capabilities. Despite significant progress made by these foundation models, they still often struggle with complex text descriptions, limiting their

*Corresponding author.

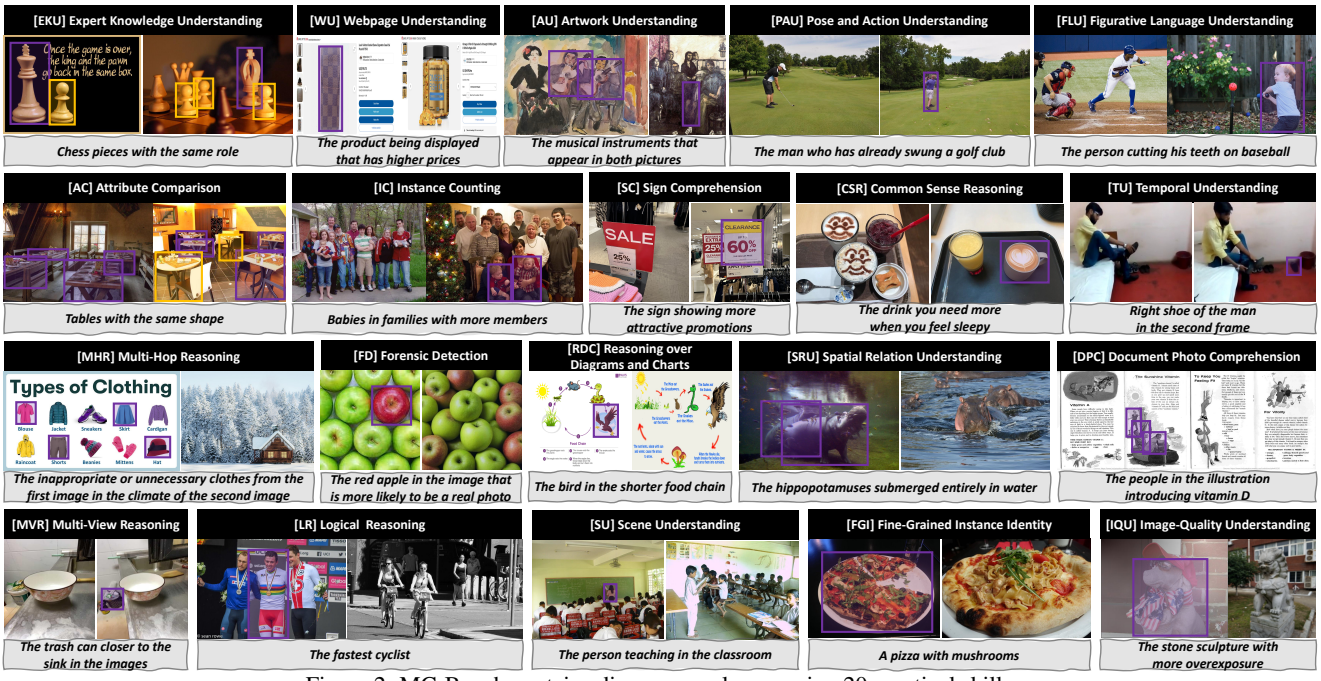


Figure 2. MC-Bench contains diverse samples covering 20 practical skills.

broader applications for real-world use.

Since the advent of multimodal large language models (MLLMs) [1, 4, 16, 22, 42, 50, 53, 86, 104, 115, 117], these models have advanced significantly, demonstrating extraordinary capabilities in understanding human language and reasoning about the visual world. Besides solving image-level visual-language tasks such as captioning [4, 29] and visual question answering (VQA) [3, 10], some recent MLLM works [7, 13, 14, 109, 113] have also explored more fine-grained tasks, showcasing promising region understanding and visual grounding capabilities. Despite their significance, we notice that, like many early visual grounding works, previous region-level MLLMs typically focus on single-image inputs, ignoring the cross-image context.

We believe that multi-image vision-language intelligence plays a pivotal role in many real-world applications, where the ability to extract and integrate contextual information from multiple images provides essential cues that enhance complex comprehension and reasoning. For instance, in autonomous driving, models [18, 77] can better understand pedestrians and vehicles in the 3D world by integrating data from multiple camera angles. In security and surveillance, models [12, 75] can enhance system understanding of the dynamic environment by integrating multiple frames from different cameras to identify and analyze the targets across different time and locations. General-purpose AI assistants (*e.g.*, chart analysis [118] and GUI agents [105]) are capable of understanding and reasoning across multiple contexts to identify correlations/discrepancies and make decisions. Although some early works investigate vision-language intelligence in multi-image scenarios, they are limited to image-level tasks [54, 82] or without complex textual descriptions [31, 83].

Driven by this intuition, this paper explores a significant yet largely overlooked scenario and introduces a practical multi-image instance-level task, namely multi-context visual grounding, to assess such unproven abilities of existing MLLMs. This new task focuses on reasoning and localizing regions of interest across multiple images based on open-ended text prompts. As illustrated in Figure 1a, compared to existing language-based visual grounding tasks [26, 49, 57, 68, 76, 80, 87, 100], multi-context visual grounding is more challenging, as it takes cross-image context into consideration and uses more nuanced and flexible textual expressions along with a greater diversity of disciplines.

To facilitate the research, we present MC-Bench, the first MLLM benchmark specifically designed for visual grounding in multi-image scenarios. MC-Bench comprises 2,000 manually labeled samples, each featuring paired images, instance-level annotations and a corresponding text prompt. The text prompts are categorized into three distinct styles (*i.e.*, referring, comparison and reasoning), covering 20 practical skills applicable to real-world scenarios (see Figure 2). Overall, we collect 3,345 diverse images from over 10 data sources, covering natural images, charts, document photos, artworks and scientific diagrams. We then carefully curate 2,000 image pairs and manually annotated 1,514 unique open-ended text prompts, along with 3,200 language-grounded bounding boxes.

We evaluate over 20 baselines with potential multi-

Table 1. Comparison to related vision-language datasets from different dimensions, *i.e.*, multi-image input, instance-level annotation, multi-domain data and text description types. ✔ in the multi-image column indicates datasets containing multi-image subsets.

Datasets	multi-image	instance-labeled	multi-domain	text description types
MS-COCO [51]	✘	✔	✘	object categories & image-level captions
RefCOCO/g/+ [35, 61]	✘	✔	✘	category/attribute/relation descriptions
RIO [68]	✘	✔	✘	sentences of intention descriptions for objects
D^3 [100]	✘	✔	✔	unrestricted descriptions for any number of instances
OmniLabel [76]	✘	✔	✔	complex object descriptions for any number of instances
ODinW [40]	✘	✔	✔	object categories & external knowledge descriptions
VQS [21]	✘	✔	✘	multi-choice QAs from the VQA dataset [3]
VizWiz-VQA-G [9]	✘	✔	✘	multi-choice QAs from the VizWiz-VQA dataset [25]
MMBench [56]	✔	✘	✔	multiple-choice QAs covering multiple ability dimensions
MMMU [107]	✔	✘	✔	multi-choice & open QAs covering diverse disciplines
SEED-Bench [39]	✔	✘	✔	multi-choice QAs spanning numerous dimensions
BLINK [20]	✔	✘	✔	multi-choice QAs on visual perception abilities
MileBench [79]	✔	✘	✔	multi-choice & open QAs on long video & image sequences
Mantis-Eval [30]	✔	✘	✔	multiple-choice & open QAs on image sequences
MICBench [96]	✔	✘	✔	multi-choice QAs on comparing image quality
Mementos [91]	✔	✘	✔	descriptions capturing unfolding events on image sequences
MC-Bench (ours)	✔	✔	✔	open-ended instance-level descriptions over multiple images

context visual grounding capabilities on MC-Bench, including advanced MLLMs and a few relevant foundation models without LLMs. The experimental results indicate that current MLLMs have significant potential for improvement. Concretely, while small-scale MLLMs (no larger than 7B) can achieve comparable instance-level performance to the foundation models [55, 78], they typically show better image-level performance. As MLLMs scale up, their performance improves significantly on all metrics. We also observe that the specialist MLLMs trained exclusively on single-image visual grounding data struggle with multi-context scenarios. In contrast, some generalist MLLMs with strong instruction-following capabilities generalize better in multi-context visual grounding, particularly those trained with multi-context data, even if those data are not instance-level labeled. Nevertheless, a simple agentic baseline that integrates the strengths of GPT-4o [1] and G-DINO [55] can easily outperform all evaluated end-to-end MLLMs by a clear margin, highlighting the potential for improvement. We also introduce a fine-tuned baseline that is trained using synthesized multi-context instruction tuning data. Moreover, we conduct human evaluations to establish an upper bound for existing MLLMs, revealing a significant performance gap between MLLMs and humans.

We hope our MC-Bench and empirical findings inspire the research community to delve deeper to discover and enhance the untapped potentials of MLLMs in instance-level tasks particularly in multi-image scenarios. The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, this work is the pioneer to explore the use of MLLMs for multi-image instance-level scenarios in open environments, and suggests a practical multi-context visual grounding task.
- We construct a new dataset, MC-Bench, featuring 2,000 manually annotated samples consisting of image pairs, text prompts, and corresponding instance-level labels. The diverse images and the open-ended prompts enable the evaluation of MLLMs from a wide range of dimensions.
- We benchmark more than 20 relevant MLLMs and foundation models on MC-Bench, revealing a non-trivial performance gap between existing MLLMs and humans. Beyond the performance scores, this work provides insightful analysis aimed at guiding improvements in MLLM development.

2. Related Work

MLLM Benchmarks. Numerous benchmarks evaluate MLLMs with single-image inputs, and the assessments of the multimodal capabilities with multiple images do not receive much attention. Only a few recent benchmarks take multi-image evaluations into consideration, where some of them focus on specific domains and tasks (*e.g.*, low-level vision [95, 96, 98] and temporal understanding [43, 46]). As summarized in Table 1, some concurrent works [20, 30, 56, 79, 107] present multi-image MLLMs benchmarks for more general purposes, covering multiple fields and disciplines. However, they are annotated for image-level perception, comprehend and reasoning tasks (*e.g.*, VQA), none of them is designed for instance-level tasks. Current MLLMs for instance-level tasks are usually evaluated on conventional benchmarks [9, 13, 21, 28, 35, 61] with limited diversity and no multi-image context.

Open-Ended Visual Content Grounding. Benefiting from the pre-trained visual-language models [69, 110], open-vocabulary object detection [23, 63, 103, 108] has received increasing attention, which localizes objects of arbitrary

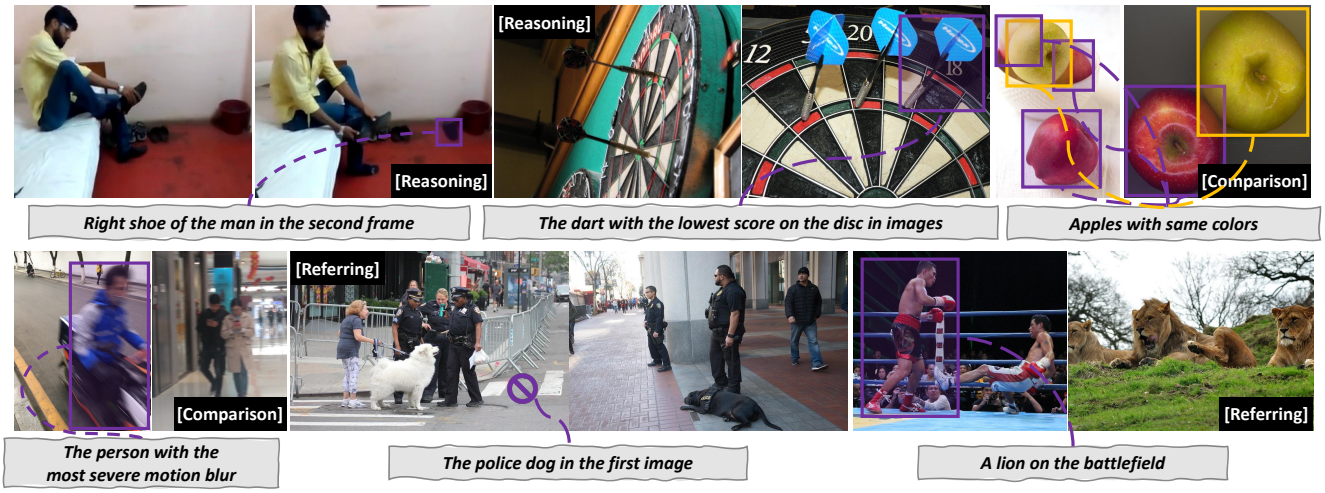


Figure 3. MC-Bench contains three distinct styles of open-ended textual descriptions, *i.e.*, referring, comparison and reasoning.

categories using language to achieve zero-shot transferability. Besides leveraging category names, another line of work [26, 68, 80, 87, 100] investigates grounding visual content using simple referring phrases or sentences that often include auxiliary cues that help distinguish specific instances from others within the same category. With the impressive success of LLMs, MLLMs have emerged as a pivotal advancement that serves to effectively connect vision and language tasks. While MLLMs [1, 4, 16, 22, 42, 53, 117] demonstrate remarkable capabilities on image-level tasks, several recent studies [14, 24, 32, 52, 59, 67, 70, 90, 106, 112, 114, 116] explore the potential of enabling MLLMs to perform region-level tasks through instruction tuning. However, most of existing works only focus on independent images and ignore multi-image context.

MLLMs with Multi-Image Context. Unlike most previous MLLLMs take single-image-text pairs as inputs, some variants of MLLMs [58, 60, 111] tailored for video tasks inherently support multiple frames and long contexts. However, these models designed to comprehend temporal sequences often face challenges when dealing with single images or multiple images that are not related temporally. Another line of work [2, 4, 5, 11, 37, 41, 43] has also noticed the importance of multiple-image capabilities for real-world applications, and takes effort for scaling the context to enable MLLMs to handle multiple and interleaved image-text inputs. Nevertheless, prior MLLMs largely neglect the multi-image instance-level scenarios, except for a few co-current works [47, 64, 85] exploring common/unique objects/parts co-localization or simple co-referring.

3. MC-Bench

3.1. Multi-Context Visual Grounding

Visual Grounding with Multi-Image Context. To meet the demands of open-ended real-world applications, this paper suggests a practical multi-image, instance-level vision-language task called multi-context visual grounding. Given a multimodal input sample, *i.e.*, multiple images and a text prompt, the models are required to localize all instances referenced in the input text description. Each image in an input sample is temporally, spatially or semantically related with others, with the text prompt linking them through shared concepts or relationships. Without loss of generality, we initially set the number of multi-images in the input samples to a pair, which maintains essential characteristics of multi-image tasks while ensuring a clear and controlled evaluation. Our evaluation pipeline and metrics can be seamlessly extended to more challenging long-context scenarios.

Visual Grounding with Open-Ended Expressions. Multi-context visual grounding aims at localizing specific instances within images using flexible and diverse text prompts, covering a broad range of practical skills. As illustrated in Figure 3, we design three distinct styles of text prompts for grounding: referring, comparison and reasoning. The referring style prompts identify instances using their category, attributes or positional information, either directly or indirectly. The comparison style prompts are slightly more challenging, requiring models to ground instances by comparing the visual content across multiple images. These comparisons can be global, based on image-level cues (*e.g.*, the quantity of objects and image quality), or local, focusing on the attributes (*e.g.*, color and shape) of objects within the images. The reasoning style prompts describe instances in a more challenging manner, where models struggle to locate instances without relying on external knowledge (*e.g.*, common sense and multi-hop reasoning skills) beyond the input itself.

Visual Grounding with One-to-Any Matching. Since the text descriptions in multi-context visual grounding are unrestricted, each positive sample includes a text prompt that may refer to one or multiple instances within the images of

that sample. In contrast, the text prompts in negative samples describe no instance within the images, and the models are encouraged to reject these negative inputs (*e.g.*, the police dog example in Figure 3). Textual expressions in the real world often exhibit high generalization and polysemy. Therefore, we also assume that the models can accurately understand the intent behind flexible prompts and group target instances accordingly. As shown in the top right of Figure 3, given images featuring apples of two colors and a prompt *'Apples of the same colors'*, the models are encouraged not only to detect all the apples but also to group them according to their colors.

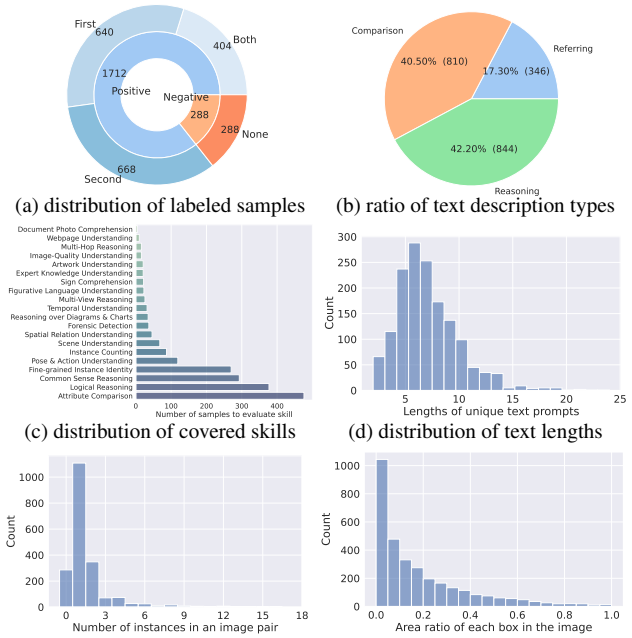
3.2. Dataset Curation

To the best of our knowledge, there is no existing dataset suitable for language-grounded cross-image instance-level tasks like multi-context visual grounding. To facilitate the research, we construct an evaluation-only dataset that effectively and faithfully evaluates the multimodal comprehend, reasoning and grounding capabilities of existing MLLMs in multi-image scenarios.

Multi-Source Image Collection. Our goal is to create a high diverse benchmark that can better simulate a variety of real-world scenarios. Guided by such goal, we first select images covering a wide range of domains and topics, *e.g.*, natural images, comics, scientific diagrams, artworks, document photos, webpage screenshots, synthesized images and *etc*. Unlike conventional benchmarks, we emphasize instance-level tasks in real-world scenarios and collect a more extensive set of scene-centric images featuring a variety of object sizes and domains. In total, we incorporate images from multiple data sources, including more than 10 existing datasets [6, 20, 30, 38, 51, 62, 65, 81, 94, 95, 97] and a few additional images crawled from the Internet. Please refer to the Appendix for more details.

Linking Images through Text Descriptions. We then repurpose the collected images and link image pairs using open-ended text descriptions. Concretely, the images are grouped into distinct subsets based on similar themes or domains. The annotators are tasked with selecting image pairs from the subsets and writing an open-ended text prompt for each selected image pair, where the text prompts are supposed to properly leverage the cross-image context and clearly identify instances. In addition, to facilitate the subsequent annotation process, annotators are asked to assign positive/negative flags to indicate whether the images contain at least one instance described by the text prompt.

Instance-Level Labeling and Cyclic Review. After labeling the text descriptions for each image pair, we distribute the triplets to other annotators for subsequent annotation. Given textual descriptions written by the text annotators, the box annotators are tasked with identifying the relevant



(a) distribution of labeled samples (b) ratio of text description types

(c) distribution of covered skills (d) distribution of text lengths

(e) distribution of instance quantity (f) distribution of instance sizes

Figure 4. Statistical analysis of the proposed MC-Bench.

instances within the positive samples and drawing bounding boxes to enclose them. Once all the samples have instance-level annotations, we reassign them to the annotators who label the text prompts, asking them to review the bounding boxes to ensure they properly encompass the target instances indicated by the written prompt. If any inconsistencies are found, the samples will be flagged for relabeling as part of the quality control process. We build an online annotation platform based on Label Studio [84], leveraging its programmable and user-friendly interface for annotating paired images (see the interface example in the Appendix).

3.3. Dataset Statistics

We gather a total of 3,345 different images from various sources, covering various domains and topics. We meticulously organize the collected images into 2,000 image pairs and provide 1,514 unique open-ended text descriptions for these image pairs. As shown in Figure 4d, the length of the text descriptions ranges from 2 to 24 words, with an average of 7.2. Each text prompt describes visual content within paired images without restriction. MC-Bench has 1,712 positive samples, with 404 containing target instances in both images, while the remaining samples having target objects in only one image (either the first or the second), as summarized in Figure 4a. Besides positive samples, we add a small proportion of negative examples to evaluate the capabilities of models for rejecting negative inputs. As illustrated in Figures 4b and 4c, MC-Bench contains three distinct styles of text expressions (*i.e.*, 346, 810 and 844 for referring, comparison and reasoning respectively) and 20

practical skills (*e.g.*, attribute comparison, logical reasoning, common sense reasoning and multi-view reasoning).

For the instance-level annotations, MC-Bench includes 3,200 language-grounded bounding boxes in total. As summarized in Figure 4e, each prompt in positive samples indicates 1 to 17 instances of 1 to 7 groups within image pairs, while there is no instance related to negative description. Unlike benchmarks for image-level tasks, we collect more challenging scene-centric images and label instances with diverse sizes. The size of the labeled bounding boxes ranges from 4e-6 to 1, as shown in the distribution in Figure 4f.

4. Experiments

4.1. Evaluation Metrics

Image-Level Metrics. For multi-context visual grounding task, we design image-level and instance-level metrics to evaluate the performance of models from different dimensions. Accuracy (Acc) is used to confirm whether the models can correctly identify which images contain the objects indicated by each text prompt, where the instance quantity and fine-grained location information is not considered.

Instance-Level Metrics. We choose average precision (AP_{50}) as the instance-level metric to verify whether the models can locate the target instances with multi-context inputs. For samples where the text prompt describes multiple groups of instances, we first apply Hungarian algorithm to match each predicted group to the most appropriate ground-truth group, ensuring that the mean intersection over union (IoU) across all predictions is maximized.

4.2. Baselines

Since the multi-context visual grounding is a new task, we implement and evaluate various advanced approaches with potential visual grounding capabilities, including latest proprietary and open-source MLLMs as well as foundation models without LLMs. Most existing methods do not support multi-image inputs, and we horizontally concatenate the images before feeding them to these models.

Specifically, we select and evaluate ❶ **the API-based generalist MLLMs**, such as GPT-4o [1] and Gemini-1.5 Pro [73], ❷ **the open-source generalist MLLMs** (*e.g.*, Qwen-VL series [5, 89], SPHINX [52], InternVL2.5 [15] and MiniGPT-v2 [11]) which are capable of performing a wide range vision-language tasks, ❸ **the open-source specialist MLLMs** (*e.g.*, Shikra [14], Kosmos-2 [66], Ferret [106], Lenna [92], Groma [59] and GroundingGPT [48]) tailored to visual grounding-related tasks and ❹ **the foundation models without LLMs**, such as ONE-PEACE [88], G-DINO [55] and APE [78]. More details (*e.g.*, model version and used prompts) are provided in the Appendix.

Apart from aforementioned end-to-end approaches, we devise and assess ❺ **an agentic baseline** that follows a simple yet effective divide-and-conquer strategy [93] and takes the advantages of MLLMs and detectors in reasoning and localization respectively. Concretely, we utilize GPT-4o as a reasoning agent and prompt it to first analyze multi-context inputs to determine which images contain the target instances described by the text phrases. This reasoning agent is then requested to generate concise and discriminative referring phrases for each individual target instance. We finally localize the target objects using G-DINO [55] along with the GPT-generated referring phrases. Some examples of our agentic baseline are showcased in Figure 5.

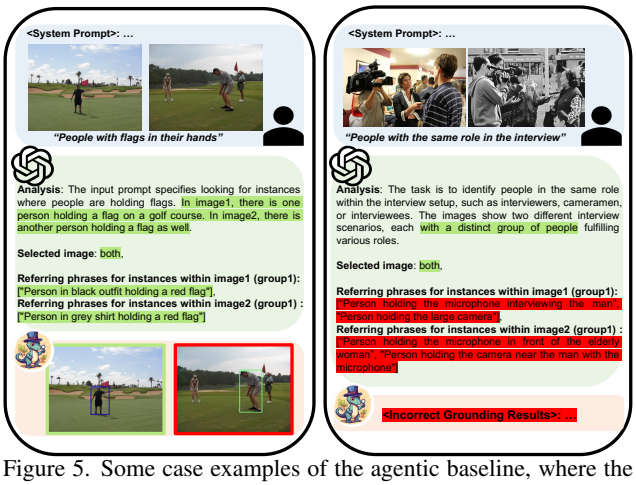


Figure 5. Some case examples of the agentic baseline, where the correct and wrong predictions are highlighted using green and red. The left case shows the detection error caused by G-DINO, while the right case demonstrates the grouping error caused by GPT-4o.

We also introduce and evaluate ❻ **a finetuned baseline** that enhances existing end-to-end MLLM (*i.e.*, Qwen2-VL-7B [89]) by multi-context instruction tuning. We construct a multi-context instruction tuning dataset with over 50K samples by collecting multi-context image-level task samples from existing datasets [19, 30] and synthesizing multi-context instance-level task samples. To accelerate the training process/maintain the generalization capabilities of the MLLM, we finetune models with LoRA [27]. Please refer to the Appendix for more training details.

We conduct ❼ **human evaluations** to establish an upper bound for the models. In total, we invite 3 volunteers who have not been exposed to annotated data to participate in the evaluation with all 2K multi-context samples. Given each text prompt, the participants are asked to draw bounding boxes for the target instances in corresponding image pairs.

4.3. Benchmark Results

We divide existing approaches into different groups and report their performance in Table 2. The proprietary generalist MLLMs [55, 73] are used through API calls and generally considered to have huge model sizes. These models inherently support image sequence inputs and show strong image-level comprehend and reasoning capabilities.

Table 2. Comparison of baselines on MC-Bench. *Sequence* indicates whether the model supports image sequences as inputs, where ✔ denotes that some intermediate steps support image sequences. The superscripts *ref*, *com* and *rea* denote the results for the three specific types respectively.

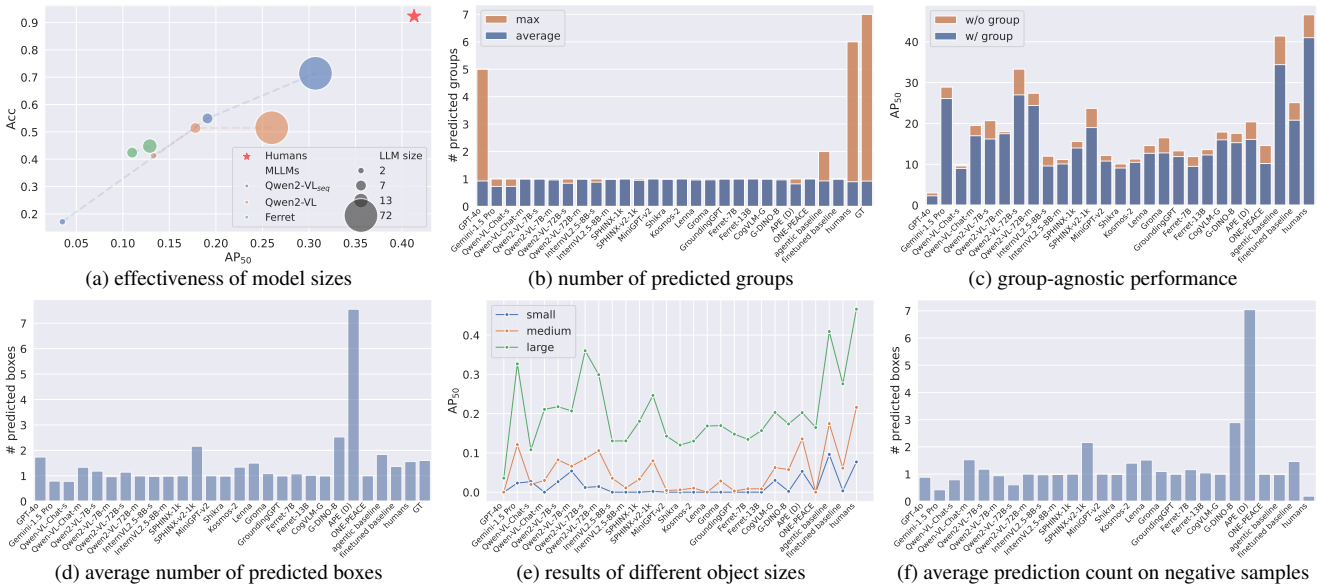
Methods	sequence	LLM size	Image-Level				Instance-Level			
			Acc^{ref}	Acc^{com}	Acc^{rea}	Acc	$\text{AP}_{50}^{\text{ref}}$	$\text{AP}_{50}^{\text{com}}$	$\text{AP}_{50}^{\text{rea}}$	AP_{50}
API-Based Generalist MLLMs										
GPT-4o [1]	✔	-	69.7	82.8	77.5	78.3	1.8	3.9	2.3	2.8
Gemini-1.5 Pro [73]	✔	-	56.1	65.1	62.7	62.5	30.6	29.9	26.1	28.2
Open-Source Generalist MLLMs										
Qwen-VL-Chat [5]	✔	7B	33.8	34.8	31.8	33.4	10.9	9.2	9.0	9.3
Qwen-VL-Chat [5]	✘	7B	36.7	47.7	45.5	44.9	21.7	17.3	17.0	17.5
Qwen2-VL [89]	✔	7B	43.9	60.1	54.3	54.9	22.5	21.3	16.2	19.1
Qwen2-VL [89]	✘	7B	43.6	52.2	53.7	51.4	19.9	18.0	17.5	17.8
Qwen2-VL [89]	✔	72B	61.6	79.1	68.0	71.4	33.7	33.2	27.0	30.7
Qwen2-VL [89]	✘	72B	43.1	53.5	52.8	51.4	29.6	26.7	24.4	26.0
InternVL2.5 [15]	✔	8B	28.0	37.7	38.5	36.4	15.7	10.9	9.6	11.1
InternVL2.5 [15]	✘	8B	38.7	54.8	53.2	51.4	12.9	11.4	10.1	10.8
SPHINX-1k [52]	✘	13B	41.9	49.6	51.1	48.9	16.2	15.8	14.0	14.9
SPHINX-v2-1k [52]	✘	13B	41.3	52.2	38.9	44.7	26.5	21.1	19.0	20.8
MiniGPT-v2 [11]	✘	7B	34.1	43.8	45.6	42.9	11.7	12.2	10.8	11.6
Open-Source Specialist MLLMs										
Shikra [14]	✘	7B	37.6	44.7	45.4	43.8	10.0	10.6	9.1	9.8
Kosmos-2 [66]	✘	1.6B	26.3	30.6	33.6	31.2	10.7	11.5	10.5	10.6
Lenna [92]	✘	7B	30.3	30.6	28.6	29.7	17.1	14.3	12.7	13.9
Groma [59]	✘	7B	34.1	44.4	42.4	41.8	17.2	15.6	12.8	14.2
GroundingGPT [48]	✘	7B	35.5	43.3	46.3	43.3	14.4	12.2	11.9	12.3
Ferret [106]	✘	7B	34.4	42.6	45.5	42.4	12.8	12.6	9.5	11.0
Ferret [106]	✘	13B	35.8	44.7	48.6	44.8	13.4	13.5	12.3	12.9
CogVLM-Grounding [90]	✘	17B	40.5	50.2	50.1	48.5	20.9	18.0	16.0	17.5
Foundation Models without LLMs										
G-DINO-B [55]	✘	✘	31.2	30.4	30.9	30.8	13.9	15.6	15.3	15.0
APE (D) [78]	✘	✘	24.0	20.6	16.2	19.3	20.4	20.8	16.1	18.8
ONE-PEACE [88]	✘	✘	32.9	42.7	42.3	40.9	17.8	15.5	10.2	13.3
Agentic Baseline$_{\textit{GPT-4o+G-DINO}}$	✔	-	66.8	84.8	75.7	77.9	41.6	37.2	34.4	36.2
Finetuned Baseline$_{\textit{Qwen2-VL-7B}}$	✔	7B	47.1	59.9	60.0	57.7	26.7	23.2	20.8	22.6
Humans	-	-	89.5	95.4	90.5	92.3	47.8	40.3	41.0	41.3

However, while Gemini-1.5 Pro [73] achieves competitive instance-level performance, GPT-4o [1] exhibits limited fine-grained localization capabilities.

For the open-source MLLMs accepting image sequence inputs (*i.e.*, Qwen-VL-Chat [5], Qwen2-VL [89] and InternVL2.5 [15]), we compare both sequence- and merge-image variants. We find that as model capabilities increase (*i.e.*, Qwen-VL to Qwen2-VL, and 7B to 72B LLM), the sequence-image variants more clearly exceed merge-image variants. Among all tested open-source MLLMs [5, 11, 15, 52, 89], Qwen2-VL-72B with image sequence inputs achieves the best results, even outperforms proprietary MLLMs on instance-level metrics.

Generally, the specialist MLLMs [14, 48, 59, 66, 90, 92, 106] are specially designed or fine-tuned for visual grounding-related tasks. However, in multi-context visual grounding, existing specialists obtain worse results in terms of both image-level and instance-level metrics. For instance, the largest specialist CogVLM-Grounding-17B [90] achieves performance comparable to some 7B generalist MLLMs (*e.g.*, Qwen-VL-Chat and Qwen2-VL). We attribute this to the limited generalization capabilities of these specialists tailored to single-image scenarios.

Compared to MLLM counterparts, the foundation models [55, 78, 88] without LLMs still perform well on instance-level metrics. However, these models tend to generate redundant low-confidence boxes within irrelevant images, leading to deteriorated Acc performance. The agentic baseline integrates extraordinary multi-modal comprehension and reasoning capabilities of GPT-4o and excellent localization capabilities of G-DINO [55], thereby achieving remarkable results and surpassing aforementioned end-to-end approaches. We also observe that after multi-context instruction tuning, the cross-image perception and localization capabilities of Qwen2-VL-7B are significantly enhanced, leading to 2.8% Acc and 3.5% AP_{50} gains. Moreover, we calculate the average results of all volunteers as the upper bound. Human evaluations outperform the agentic baseline by 14.4% in Acc and 5.1% in AP_{50}, underscoring a clear performance gap between models and humans.



(a) effectiveness of model sizes (b) number of predicted groups (c) group-agnostic performance

(d) average number of predicted boxes (e) results of different object sizes (f) average prediction count on negative samples

Figure 6. More analysis experiments on MC-Bench.

4.4. More Analysis

We conduct multiple analytical experiments to further explore MLLMs from different perspectives. For the open-source MLLMs (*e.g.*, Qwen2-VL [89] and Ferret [106]) with various model size variants, we analyze the impact of model size, as visualized in Figure 6a. Larger models show sustained performance improvement on both Acc and AP_{50}, consistent with the scaling law [34].

In multi-context visual grounding, a single text prompt may describe objects from multiple groups. As shown in Figure 6b, we observe that current approaches struggle with assigning groups, with most models predicting only one group. By replacing the standard instance-level metric with a group-agnostic one, almost all baselines achieve significantly higher AP_{50} results (see Figure 6c), indicating that while these methods correctly localize the instances, they fail to assign the correct group. Moreover, we find that most models generate only about one instance per sample on average, as illustrated in Figure 6d. These observations suggest potential for improvement in generating multiple instances and assigning groups.

Inspired by MS-COCO evaluation [51], we divide the instances into different scales (*i.e.*, small, medium and large) and analyze the performance of different object sizes in Figure 6e. We find that while existing models correctly localize large-scale instances, they usually struggle to ground medium and small objects, particularly MLLMs. The agentic baseline integrates the reasoning capabilities of GPT-4o with the localization strength of G-DNIO, demonstrating significant advantages in grounding small objects.

In order to verify the models' capabilities to reject negative samples, we calculate the average number of predictions across all negative samples, as shown in Figure 6f. We observe that most models struggle with negative samples. Gemini [73] performs the best, with 0.42 predictions per negative sample, but this is still significantly worse than human performance (0.19 predictions per negative sample).

5. Conclusion

This paper investigates a valuable yet overlooked problem in the field of MLLMs and proposes a new task, namely multi-context visual grounding. Unlike prior works that focus on single-image understanding, multi-context visual grounding aims at localizing instances in multi-image scenarios. Additionally, the text prompts used in multi-context visual grounding are more open-ended and challenging compared to those in previous language-based localization tasks. To facilitate the research, we introduce MC-Bench, a new benchmark designed for instance-level tasks in multi-context scenarios. MC-Bench contains 2,000 image pairs with diverse text prompts describing target instances in three distinct styles, covering 20 practical tasks. After benchmarking over 20 advanced MLLMs and foundation models, we found that current models typically struggle with multiple images and exhibit frustratingly low performance compared to the human upper bound. We conduct multiple analytical experiments to further investigate the issues that hinder the improvement of existing methods and to identify future directions for development. Our research advances MLLM development by highlighting weaknesses in instance-level tasks within multi-image scenarios, and MC-Bench serves as a valuable resource for further research. We hope our findings will draw attention to the application of MLLMs in instance-level tasks in multi-context scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62441617 and 62402432). This work was partially supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02032). This work was also supported in part by the China Postdoctoral Science Foundation (2024M762830).

References

[1] 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.

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

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

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

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[6] Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. Winogavil: Gamified association benchmark to challenge vision-and-language models. In *NeurIPS*, 2022.

[7] Zhixi Cai, Fucai Ke, Simindokht Jahangard, Maria Garcia de la Banda, Reza Haffari, Peter J Stuckey, and Hamid Rezatofighi. Naver: A neuro-symbolic compositional automaton for visual grounding with explicit logic reasoning. In *ICCV*, 2025.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[9] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *CVPR*, 2022.

[10] Felix Chen, Hangjie Yuan, Yunqiu Xu, Tao Feng, Jun Cen, Pengwei Liu, Zeying Huang, and Yi Yang. Mathflow: Enhancing the perceptual flow of mllms for visual mathematical problems. *arXiv preprint arXiv:2503.16549*, 2025.

[11] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[12] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *CVPR*, 2024.

[13] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S.-H. Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *CVPRW*, 2025.

[14] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[15] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

[17] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *WACV*, 2020.

[18] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *CVPR*, 2024.

[19] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: Generating fine-grained image comparisons. In *EMNLP*, 2019.

[20] 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. In *ECCV*, 2024.

[21] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *ICCV*, 2017.

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

[23] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022.

[24] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *CVPR*, 2024.

[25] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018.

[26] Zeyu Han, Fangrui Zhu, Qianru Lao, and Huaizu Jiang. Zero-shot referring expression comprehension via structural similarity between images and captions. In *CVPR*, 2024.

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[28] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

[29] Heng Jia, Yunqiu Xu, Linchao Zhu, Guang Chen, Yufei Wang, and Yi Yang. Mos2: Mixture of scale and shift experts for text-only video captioning. In *ACM MM*, 2024.

[30] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *TMLR*, 2024.

[31] Shuqiang Jiang, Sisi Liang, Chengpeng Chen, Yaohui Zhu, and Xiangyang Li. Class agnostic image common object detection. *IEEE TIP*, 2019.

[32] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Lumen: Unleashing versatile vision-centric capabilities of large multimodal models. In *NeurIPS*, 2024.

[33] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021.

[34] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[35] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *CVPR*, 2023.

[37] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *NeurIPS*, 2024.

[38] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

[39] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024.

[40] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS*, 2022.

[41] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *ICLR*, 2025.

[42] 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 *ICML*, 2023.

[43] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Empowering vision-language models to follow interleaved vision-language instructions. In *ICLR*, 2024.

[44] Kun Li, Dan Guo, and Meng Wang. Proposal-free video grounding with contextual pyramid network. In *AAAI*, 2021.

[45] Kun Li, Jiaxiu Li, Dan Guo, Xun Yang, and Meng Wang. Transformer-based visual grounding with cross-modality interaction. *TOMM*, 2023.

[46] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024.

[47] You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, et al. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. *arXiv preprint arXiv:2501.05767*, 2025.

[48] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. GroundingGPT: Language enhanced multi-modal grounding model. In *ACL*, 2024.

[49] Chen Liang, Wenguan Wang, Tianfei Zhou, Jiaxu Miao, Yawei Luo, and Yi Yang. Local-global context aware transformer for language-guided video segmentation. *IEEE TPAMI*, 2023.

[50] Chen Liang, Wenguan Wang, and Yi Yang. Towards human-like virtual beings: Simulating human behavior in 3d scenes. In *ICCV*, 2025.

[51] 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.

[52] Ziyi Lin, Dongyang Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Yu Qiao, and Hongsheng Li. Sphinx: A mixer of weights, visual embeddings and image scales for multi-modal large language models. In *ECCV*, 2024.

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

[54] Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan,

Bing Li, and Weiming Hu. Mibench: Evaluating multimodal large language models over multiple images. In *EMNLP*, 2024.

[55] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024.

[56] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024.

[57] Yu Lu, Ruijie Quan, Linchao Zhu, and Yi Yang. Zero-shot video grounding with pseudo query lookup and verification. *IEEE TIP*, 2024.

[58] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

[59] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *ECCV*, 2024.

[60] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reliable video narrator via equal distance to visual tokens. In *CVPR*, 2024.

[61] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

[62] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

[63] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, 2022.

[64] Kiet A Nguyen, Adheesh Juvekar, Tianjiao Yu, Muntasir Wahed, and Ismini Lourentzou. Calico: Part-focused semantic co-segmentation with large vision-language models. In *CVPR*, 2025.

[65] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *ICCV*, 2019.

[66] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *ICLR*, 2024.

[67] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, and Tong Zhang. Detgpt: Detect what you need via reasoning. In *EMNLP*, 2023.

[68] Mengxue Qu, Yu Wu, Wu Liu, Xiaodan Liang, Jingkuan Song, Yao Zhao, and Yunchao Wei. Rio: A benchmark for reasoning intention-oriented objects in open environments. In *NeurIPS*, 2023.

[69] 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 *ICML*, 2021.

[70] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024.

[71] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, 2025.

[72] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

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

[74] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[75] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, 2024.

[76] Samuel Schulter, Yumin Suh, Konstantinos M Dafnis, Zhixing Zhang, Shiyu Zhao, Dimitris Metaxas, et al. OmniLabel: A challenging benchmark for language-based object detection. In *ICCV*, 2023.

[77] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *CVPR*, 2024.

[78] Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *CVPR*, 2024.

[79] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. In *COLM*, 2024.

[80] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *ACL*, 2022.

[81] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019.

[82] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*, 2023.

[83] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *CVPR*, 2014.

[84] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2022. Open source software available from https://github.com/heartexlabs/label-studio.

[85] Muntasir Wahed, Kiet A Nguyen, Adheesh Sunil Juvekar, Xinzhuo Li, Xiaona Zhou, Vedant Shah, Tianjiao Yu, Pinar Yanardag, and Ismini Lourentzou. Prima: Multi-image vision-language models for reasoning segmentation. *arXiv preprint arXiv:2412.15209*, 2024.

[86] Chao Wang, Hehe Fan, Ruijie Quan, Lina Yao, and Yi Yang. Protchatgpt: Towards understanding proteins with large language models. In *SIGIR*, 2025.

[87] Hanyao Wang, Yibing Zhan, Liu Liu, Liang Ding, and Jun Yu. Balanced similarity with auxiliary prompts: Towards alleviating text-to-image retrieval bias for clip in zero-shot learning. *arXiv preprint arXiv:2402.18400*, 2024.

[88] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023.

[89] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[90] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. In *NeurIPS*, 2024.

[91] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *ACL*, 2024.

[92] Fei Wei, Xinyu Zhang, Ailing Zhang, Bo Zhang, and Xiangxiang Chu. Lenna: Language enhanced reasoning detection assistant. In *ICASSP*, 2025.

[93] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[94] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *NeurIPS*, 2021.

[95] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*, 2024.

[96] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. Towards open-ended visual quality comparison. In *ECCV*, 2024.

[97] Yixuan Wu, Zhao Zhang, Chi Xie, Feng Zhu, and Rui Zhao. Advancing referring expression segmentation beyond single image. In *ICCV*, 2023.

[98] Zhiliang Wu, Kerui Chen, Kun Li, Hehe Fan, and Yi Yang. Bvinet: Unlocking blind video inpainting with zero annotations. In *ICCV*, 2025.

[99] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024.

[100] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. In *NeurIPS*, 2023.

[101] Yunqiu Xu, Chunluan Zhou, Xin Yu, Bin Xiao, and Yi Yang. Pyramidal multiple instance detection network with mask guided self-correction for weakly supervised object detection. *IEEE TIP*, 2021.

[102] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang. H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *CVPR*, 2022.

[103] Yunqiu Xu, Chunluan Zhou, Xin Yu, and Yi Yang. Cyclic self-training with proposal weight modulation for cross-supervised object detection. *IEEE TIP*, 2023.

[104] Yunqiu Xu, Linchao Zhu, and Yi Yang. Gg-editor: Locally editing 3d avatars with multimodal large language model guidance. In *ACM MM*, 2024.

[105] An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian McAuley, Jianfeng Gao, et al. GPT-4V in wonderland: Large multimodal models for zero-shot smartphone gui navigation. *arXiv preprint arXiv:2311.07562*, 2023.

[106] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024.

[107] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

[108] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021.

[109] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *ECCV*, 2024.

[110] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *NeurIPS*, 2022.

[111] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, 2023.

[112] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, and Jianwei Yang. Llava-grounding: Grounded visual chat with large multimodal models. In *ECCV*, 2024.

[113] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. In *COLM*, 2024.

[114] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *CVPR*, 2024.

[115] Yunzhu Zhang, Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, and Linchao Zhu. Flexselect: Flexible token selection for efficient long video understanding. *arXiv preprint arXiv:2506.00993*, 2025.

[116] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023.

[117] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2023.

[118] Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. Multichartqa: Benchmarking vision-language models on multi-chart problems. In *NAACL*, 2025.