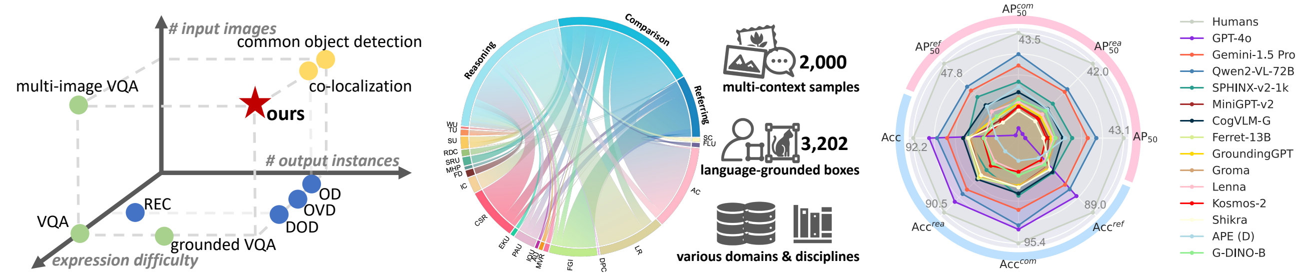
MC-Bench: A Benchmark for Multi-Context Visual Grounding in the Era of MLLMs

Yunqiu Xu
ReLER Lab, CCAI
Zhejiang University
imyunqiuxu@gmail.com

Linchao Zhu
ReLER Lab, CCAI
Zhejiang University
zhulinchao@zju.edu.cn

Yi Yang
ReLER Lab, CCAI
Zhejiang University
yangyics@zju.edu.cn



(a) correlation of ours with other tasks (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 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.

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 grouped into 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, as well as a simple yet effective stepwise 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 interesting observations that suggest potential future directions. We hope MC-Bench and our empirical findings can encourage the research community to further explore and enhance 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) [7, 60, 62] and open-vocabulary object detection (OVD) [14, 29] based on category names, as well as referring expression comprehension (REC) [24, 68, 73] and describe object detection (DOD) [64, 83] with referring phrases. However, the 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 [32, 45, 66] has catalyzed a shift from specialized models to general-purpose foundation models, showcasing unprecedented generalization capabilities. Despite significant progress made by these foundation models, they usually struggle with complex text descriptions, limiting broader their applications for real-world use.

Since the advent of multimodal large language models (MLLMs) [1, 4, 13, 19, 37, 43, 96], these models

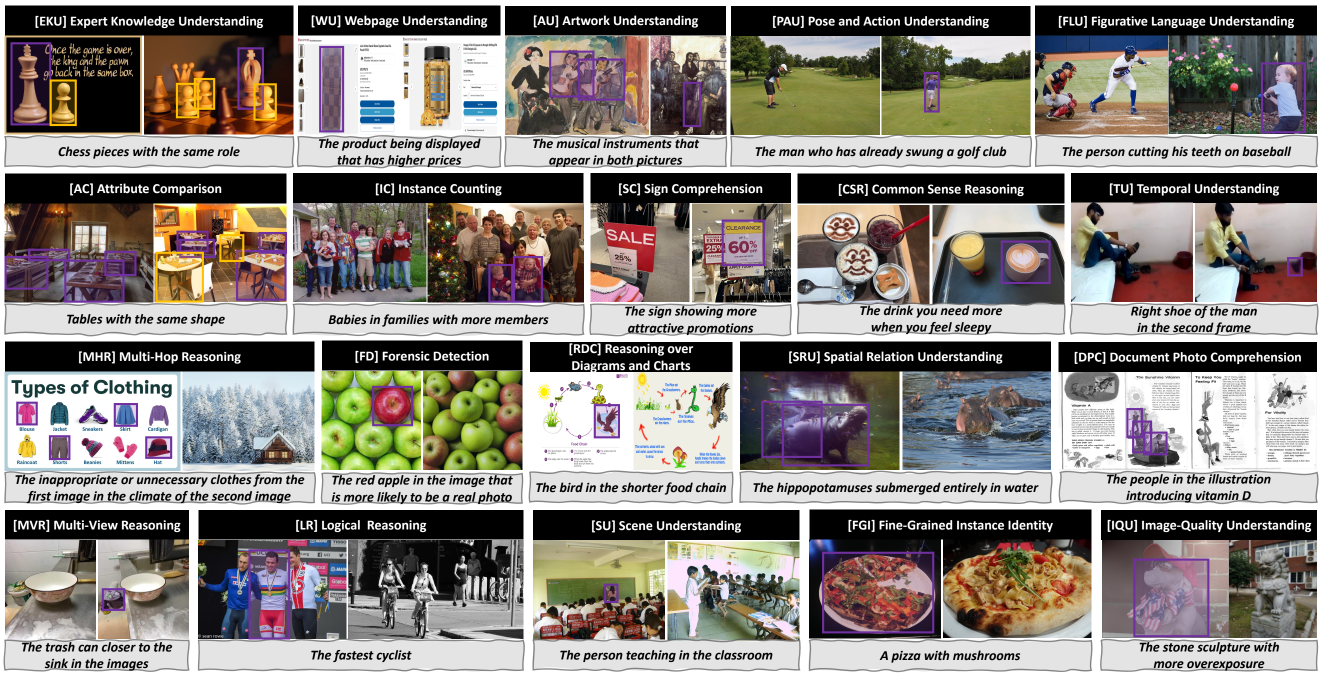


Figure 2. MC-Bench contains diverse samples covering 20 practical skills.

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 image captioning [4] and visual question answering (VQA) [3], some recent MLLM works [11, 12, 88, 92] 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 [15, 65] can better understand pedestrians and vehicles in the 3D world by integrating data from multiple camera angles. In security and surveillance, models [10, 63] 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 [97] and GUI agents [84]) 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 [44, 70] or without complex textual descriptions [28, 71].

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 [24, 57, 64, 68, 73, 83], 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 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 3,202 language-grounded bounding boxes.

We evaluate over 20 baselines with potential multi-context visual grounding capabilities on MC-Bench, including advanced MLLMs and a few relevant foundation models without LLMs. Experimental results indicate that current

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 [41]	✘	✔	✘	object categories & image-level captions
RefCOCO/g/+ [31, 51]	✘	✔	✘	category/attribute/relation descriptions
RIO [57]	✘	✔	✘	sentences of intention descriptions for objects
D^3 [83]	✘	✔	✔	unrestricted descriptions for any number of instances
OmniLabel [64]	✘	✔	✔	complex object descriptions for any number of instances
ODinW [35]	✘	✔	✔	object categories & external knowledge descriptions
VQS [18]	✘	✔	✘	multi-choice QAs from the VQA dataset [3]
VizWiz-VQA-G [8]	✘	✔	✘	multi-choice QAs from the VizWiz-VQA dataset [23]
MMBench [46]	✔✘	✘	✔	multiple-choice QAs covering multiple ability dimensions
MMMU [86]	✔✘	✘	✔	multi-choice & open QAs covering diverse disciplines
SEED-Bench [34]	✔✘	✘	✔	multi-choice QAs spanning numerous dimensions
BLINK [17]	✔✘	✘	✔	multi-choice QAs on visual perception abilities
MileBench [67]	✔	✘	✔	multi-choice & open QAs on long video & image sequences
Mantis-Eval [27]	✔	✘	✔	multiple-choice & open QAs on image sequences
MICBench [81]	✔	✘	✔	multi-choice QAs on comparing image quality
Mementos [77]	✔	✘	✔	descriptions capturing unfolding events on image sequences
MC-Bench (ours)	✔	✔	✔	open-ended instance-level descriptions over multiple images

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 [45, 66], 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 that data is not instance-level labeled. Nevertheless, a simple stepwise baseline that integrates the strengths of GPT-4o [1] and G-DINO [45] 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 can encourage 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 2K 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 Works

MLLM Benchmarks. Numerous benchmarks evaluate MLLMs with single-image inputs, and the assessments of the multimodal capabilities with multiple images does 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 [80, 81] and temporal understanding [38, 39]). As summarized in Table 1, some concurrent works [17, 27, 46, 67, 86] 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 [8, 11, 18, 26, 31, 51] with limited diversity and no multi-image context.

Open-Ended Visual Content Grounding. Benefiting from the pre-trained visual-language models [58, 89], open-vocabulary object detection [20, 53, 87] has received in-

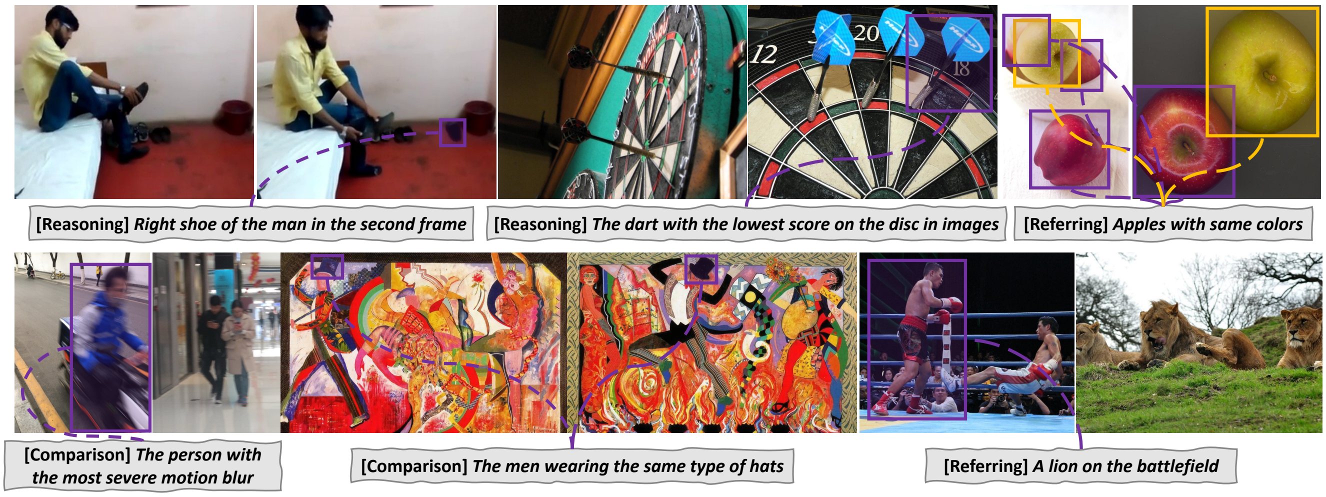


Figure 3. MC-Bench contains three distinct styles of open-ended textual descriptions, *i.e*., referring, comparison and reasoning.

creasing attention, which localizes objects of arbitrary categories using language to achieve zero-shot transferability. Besides leveraging category names, another line of work [24, 57, 68, 73, 83] 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, 13, 19, 37, 43, 96] demonstrate remarkable capabilities on image-level tasks, several recent studies [12, 21, 42, 49, 56, 59, 76, 85, 90, 93, 94] explore the potential of enabling MLLMs to perform region-level tasks through instruction tuning. However, most of prior works focus on grounding objects from independent images and ignore the multi-image context.

MLLMs with Multi-Image Context. Unlike most previous MLLLMs take single-image-text pairs as inputs, some variants of MLLMs [48, 50, 91] tailored for video tasks inherently support multiple frames and long contexts. However, these models designed to comprehend temporal sequences and often face challenges when dealing with single images or multiple images that are not related temporally. Another line of work [2, 4, 5, 9, 36, 38, 55] has also notice 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 multimodal capabilities in multi-image instance-level scenarios.

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. Without loss of generality, this work initially sets the number of multi-images in the input samples to a pair. An image pair in each input sample is temporally, spatially or semantically related, with the text prompt linking them through various shared concepts or relationships.

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. Textual expressions in the real world often exhibit high generalization and polysemy Therefore, we 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.

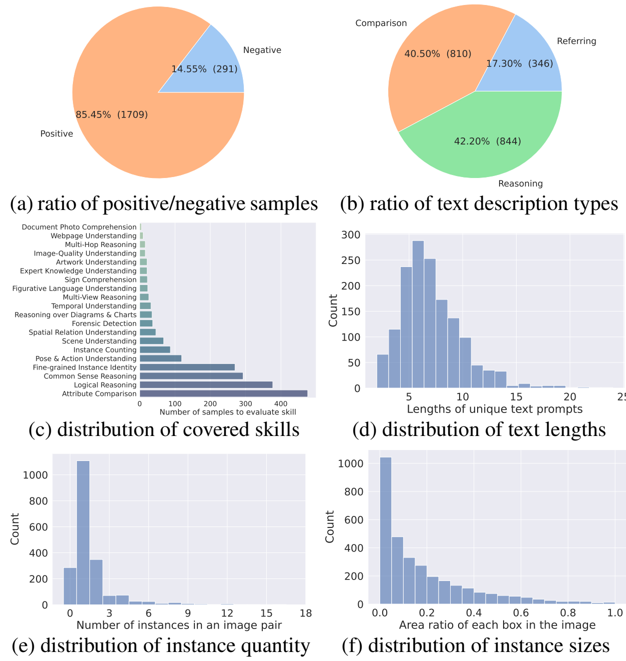
3.2. Dataset Curation

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 try to construct an evaluation-only dataset and faithfully benchmark 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, 17, 27, 33, 41, 52, 54, 69, 79, 80, 82] and a few additional images crawled from the Internet. Please refer 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 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



(a) ratio of positive/negative 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.

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 [72], 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 prompts describe visual content within paired images without restriction. Besides 1,709 positive samples, we add a small proportion of negative examples to evaluate the capabilities of models for rejecting negative inputs, as in Figure 4a. 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 *etc*.).

For the instance-level annotations, our MC-Bench includes 3,202 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 the negative descriptions. Unlike benchmarks for image-level tasks, we collect more challenging scene-centric images and label instances with divers sizes. The sizes of the labeled bounding boxes range from 4e-6 to 1, and the distribution is shown in Figure 4f.

Figure 5. Some case examples of the stepwise 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.

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 the sample where text prompt describes multiple groups of instances, we first employ group matching to align each predicted group with the appropriate ground-truth group and ensure that the mean 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 [61], ❷ **the open-source generalist MLLMs** (*e.g.*, Qwen-VL [5], Qwen2-VL [75], MiniGPT-v2 [9], SPHINX [42] and Kosmos-2 [55]) which are capable of performing a wide range vision-language tasks, ❸ **the open-source specialist MLLMs** (*e.g.*, Shikra [12], Ferret [85], Groma [49], Lenna [78] and GroundingGPT [40]) tailored to visual grounding-related tasks and ❹ **the foundation models without LLMs**, such as G-DINO [45], APE [66] and ONE-PEACE [74]. More detailed model ID information and prompts are provided in the Appendix.

Apart from aforementioned end-to-end approaches, we devise and assess ❺ **a stepwise baseline** that follows a simple yet effective divide-and-conquer strategy and takes the advantages of MLLMs and detectors in reasoning and localization respectively. Concretely, we prompt GPT-4o to analyze the multi-context inputs and select images containing instances described by the textual inputs. GPT-4o is further requested to generate concise and discriminative referring phrases for each individual target instance. We finally localize the target objects using G-DINO [45] along with the GPT-generated referring phrases. Some examples of the stepwise baseline are depicted in Figure 5.

We also introduce and evaluate ❻ **a finetuned baseline** that enhances existing end-to-end MLLM (*i.e.*, Qwen2-VL-7B [75]) 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 [16, 27] 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 [25]. Please refer 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 textual prompt, the participants are asked to draw bounding boxes to the 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 [45, 61] 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. However, while Gemini-1.5 Pro [61] 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] and Qwen2-VL [75]), we

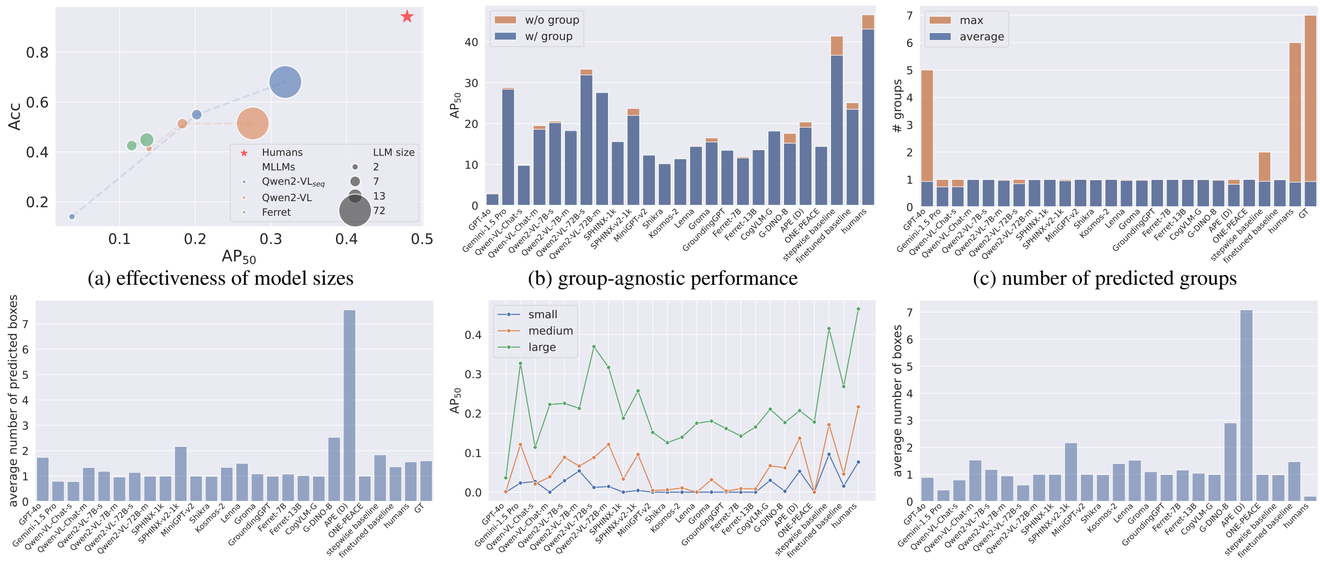
Table 2. Comparison of baselines on MC-Bench. *Sequence* indicates whether the model supports image sequences as inputs, where ✔✘ denotes that some intermediate steps support image sequences. The superscripts *ref*, *com* and *rea* denote the results for the three specific types respectively.

Methods	sequence	LLM size	Image-Level				Instance-Level			
			$\mathrm{Acc}^{\mathrm{ref}}$	$\mathrm{Acc}^{\mathrm{com}}$	$\mathrm{Acc}^{\mathrm{rea}}$	Acc	$\mathrm{AP}_{50}^{\mathrm{ref}}$	$\mathrm{AP}_{50}^{\mathrm{com}}$	$\mathrm{AP}_{50}^{\mathrm{rea}}$	AP_{50}
API-Based Generalist MLLMs										
GPT-4o [1]	✔	-	69.4	82.8	77.5	78.3	1.8	4.1	2.3	2.8
Gemini-1.5 Pro [61]	✔	-	55.8	65.1	62.7	62.5	30.5	30.0	26.1	28.4
Open-Source Generalist MLLMs										
Qwen-VL-Chat [5]	✔	7B	33.8	34.8	31.8	33.4	10.9	9.5	9.7	9.8
Qwen-VL-Chat [5]	✘	7B	36.7	47.7	45.5	44.9	21.8	19.5	17.4	18.6
Qwen2-VL [75]	✔	7B	44.2	61.1	54.3	54.9	22.7	22.4	17.2	20.2
Qwen2-VL [75]	✘	7B	43.4	52.2	53.7	51.3	19.8	18.6	18.1	18.3
Qwen2-VL [75]	✔	72B	61.3	79.1	68.0	71.3	33.0	34.7	27.6	31.9
Qwen2-VL [75]	✘	72B	43.1	53.5	52.8	51.4	30.0	29.4	25.2	27.6
SPHINX-1k [42]	✘	13B	41.9	49.6	51.1	48.9	16.2	16.8	14.3	15.6
SPHINX-v2-1k [42]	✘	13B	41.3	52.2	38.9	44.7	26.4	23.3	19.4	22.0
MiniGPT-v2 [9]	✘	7B	34.1	43.8	45.6	42.9	11.6	13.2	11.5	12.3
Open-Source Specialist MLLMs										
Shikra [12]	✘	7B	37.6	44.7	45.4	43.8	10.0	11.4	9.5	10.2
Kosmos-2 [55]	✘	1.6B	26.3	30.6	33.6	31.2	10.7	13.0	10.9	11.4
Lenna [78]	✘	7B	30.1	30.6	28.6	29.6	17.1	15.0	13.4	14.4
Groma [49]	✘	7B	34.4	44.4	42.4	41.9	17.2	16.9	13.6	15.5
GroundingGPT [40]	✘	7B	35.5	43.3	46.3	43.3	14.0	14.3	12.3	13.5
Ferret [85]	✘	7B	34.7	42.6	45.5	42.5	12.8	13.6	10.0	11.6
Ferret [85]	✘	13B	35.8	44.7	48.6	44.8	13.3	14.5	13.0	13.6
CogVLM-Grounding [76]	✘	17B	40.5	50.2	50.1	48.5	21.1	19.0	16.5	18.2
Foundation Models without LLMs										
G-DINO-B [45]	✘	✘	31.5	30.4	30.9	30.8	14.0	16.1	15.3	15.2
APE (D) [66]	✘	✘	24.0	20.6	16.1	19.3	20.3	21.3	16.3	19.1
ONE-PEACE [74]	✘	✘	32.9	42.7	42.3	40.9	17.4	17.0	11.4	14.4
Stepwise Baseline$_{\text{GPT-4o+G-DINO}}$	✔✘	-	66.5	84.8	75.7	77.8	41.1	38.3	34.1	36.7
Finetuned Baseline$_{\text{Qwen2-VL-7B}}$	✔	7B	47.1	59.9	60.0	55.7	26.5	24.9	21.9	23.6
Humans	-	-	89.0	95.4	90.5	92.2	47.8	43.5	42.0	43.1

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, 9, 42, 75], Qwen2-VL-72B [75] with image sequence inputs achieves the best performance, even outperforms proprietary MLLMs on instance-level metrics.

Generally, the specialist MLLMs [12, 40, 49, 55, 76, 78, 85] 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 [76] achieves comparable performance 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 [45, 66, 74] 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 stepwise baseline integrates extraordinary multi-modal comprehension and reasoning capabilities of GPT-4o and excellent localization capabilities of G-DINO [45], 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 0.8% Acc and 3.4% AP_{50} gains. Moreover, we calculate the average results of all volunteers as the upper bound. Human evaluations outperform the stepwise baseline by 14.4% and 6.4% on Acc and AP_{50} respectively,



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

(d) average number of predictions (e) results of different object sizes (f) predictions on negative samples

Figure 6. More analysis experiments on MC-Bench.

underscoring a clear performance gap between models and humans.

4.4. More Analysis

We conduct multiple experiments to further explore MLLMs from different perspectives. For the open-source MLLMs (*e.g.*, Qwen2-VL and Ferret) 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 [30].

In multi-context visual grounding, a single text prompt may describe objects from multiple groups. As shown in Figures 6b and 6c, we find 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, several baselines achieve higher AP scores, 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.

Following the MS-COCO [41] standard, we divide the instances into different scales (*i.e.*, small, medium and large) and visualized the results for different object sizes in Figure 6e. We observe that while existing models correctly localize large-scale instances, they usually struggle to ground medium and small objects.

To verify the models' ability to reject negative samples, we calculated the average number of predictions across all negative samples, as shown in Figure 6f. We observe that most models struggle with negative samples. Gemini [61] 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 investigate a valuable yet overlooked problem in the field of MLLMs and proposes a new task, namely multi-context visual grounding. Unlike existing works focus on single-image understanding, multi-context visual grounding aims to localize 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 within multi-context scenarios.

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. 1, 3, 4, 6, 7, 14

[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. 4

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2, 3

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1, 2, 4

[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. 4, 6, 7, 14

[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. 5, 14

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[8] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *CVPR*, 2022. 3

[9] 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. 4, 6, 7, 14

[10] 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. 2

[11] 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. *arXiv preprint arXiv:2406.16866*, 2024. 2, 3

[12] 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. 2, 4, 6, 7, 14

[13] 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. 1, 4

[14] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *WACV*, 2020. 1

[15] 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. 2

[16] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: Generating fine-grained image comparisons. In *EMNLP*, 2019. 6, 15

[17] 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. 3, 5, 14

[18] 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. 3

[19] 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. 1, 4

[20] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 3

[21] 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. 4

[22] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 15, 16

[23] 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. 3

[24] Zeyu Han, Fangrui Zhu, Qianru Lao, and Huaizu Jiang. Zero-shot referring expression comprehension via structural similarity between images and captions. In *CVPR*, 2024. 1, 2, 4

[25] 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. *arXiv preprint arXiv:2106.09685*, 2021. 6, 16

[26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 3

[27] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 3, 5, 6, 14, 15

[28] Shuqiang Jiang, Sisi Liang, Chengpeng Chen, Yaohui Zhu, and Xiangyang Li. Class agnostic image common object detection. *IEEE TIP*, 2019. 2

[29] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021. 1

[30] 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. 8

[31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 3

[32] 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. 1

[33] 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. 5, 14

[34] 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. 3

[35] 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. 3

[36] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 4

[37] 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. 1, 4

[38] 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. 3, 4

[39] 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. 3

[40] 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 multimodal grounding model. In *ACL*, 2024. 6, 7, 14

[41] 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. 3, 5, 8, 14

[42] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 4, 6, 7, 14

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 4

[44] Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*, 2024. 2

[45] 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. 1, 3, 6, 7, 14

[46] 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. 3

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 14

[48] 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. 4

[49] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *ECCV*, 2024. 4, 6, 7, 14

[50] 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. 4

[51] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 3

[52] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021. 5, 14

[53] 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. 3

[54] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *ICCV*, 2019. 5, 14

[55] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 4, 6, 7, 14

[56] 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. *arXiv preprint arXiv:2305.14167*, 2023. 4

[57] 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. 2, 3, 4

[58] 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. 3

[59] 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. 4

[60] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 1

[61] 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. 6, 7, 8, 14

[62] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[63] 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. 2

[64] 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. 1, 2, 3, 15, 16

[65] 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. 2

[66] 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. 1, 3, 6, 7

[67] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024. 3

[68] 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. 1, 2, 4

[69] 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. 5, 14

[70] 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. 2

[71] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *CVPR*, 2014. 2

[72] 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`. 5, 14

[73] 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. 1, 2, 4

[74] 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. 6, 7

[75] 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. 6, 7, 14, 15

[76] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 4, 7, 14

[77] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024. 3

[78] Fei Wei, Xinyu Zhang, Ailing Zhang, Bo Zhang, and Xiangxiang Chu. Lenna: Language enhanced reasoning detection assistant. *arXiv preprint arXiv:2312.02433*, 2023. 6, 7, 14

[79] 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. 5, 14

[80] 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. 3, 5, 14

[81] 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. *arXiv preprint arXiv:2402.16641*, 2024. 3

[82] Yixuan Wu, Zhao Zhang, Chi Xie, Feng Zhu, and Rui Zhao. Advancing referring expression segmentation beyond single image. In *ICCV*, 2023. 5, 14

[83] 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. 1, 2, 3, 4

[84] 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. 2

[85] 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. 4, 6, 7, 14

[86] 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. 3

[87] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 3

[88] 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. 2

[89] 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. 3

[90] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*, 2023. 4

[91] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, 2023. 4

[92] 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. 2

[93] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *CVPR*, 2024. 4

[94] 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. 4

[95] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *ACL*, 2024. 16

[96] 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. 1, 4

[97] Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. Multichartqa: Benchmarking vision-language models on multi-chart problems. *arXiv preprint arXiv:2410.14179*, 2024. 2

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

Supplementary Material

In this Supplementary Material, we provide additional details and results omitted in the main text:

- Appendix A: General discussions about MC-Bench, including license information, intended use, social impacts, and limitations and future works.
- Appendix B: Implementation details, including existing datasets incorporated in MC-Bench, annotation interfaces, and evaluated baselines.
- Appendix C: Additional experimental results, including ablation study for the finetuned models.
- Appendix D: The datasheets for MC-Bench, including motivation, composition, collection processes, preprocessing/cleaning/labeling, uses, distribution, and maintenance.

A. General Discussions

A.1. License

The introduced MC-Bench dataset is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). This license applies to all the images and annotations we have directly contributed. The dataset also incorporates images sourced from pre-existing collections. For these images, the original licensing terms are respected and remain applicable.

A.2. Intended Use

MC-Bench is initially constructed to facilitate a significant yet largely overlooked research problem, *i.e.*, multi-context visual grounding (grounding objects using open-ended textual prompts in multi-image scenarios). The primary purpose of MC-Bench is to function as a dynamic benchmark that continuously evolves and evaluates MLLMs for multi-context visual grounding. Preliminary benchmark results on MC-Bench not only reveal a large performance gap between current MLLMs and humans, but also identify future directions for development through multiple analytical experiments. We hope MC-Bench can encourage the research community to delve deeper to discover and enhance these untapped potentials of MLLMs in instance-level tasks particularly in multi-image scenarios.

A.3. Social Impacts

The data in MC-Bench is not expected to have specific negative impacts. As the images in MC-Bench are collected from published and publicly available sources, so there are few privacy concerns. Our text and bounding box annotations do not contain any offensive, insulting or threatening information. Although a few human annotations could be subjective, we perform cyclic review and multi-round labeling procedures to reduce the bias and ensure the annotation quality. Beyond the dataset, MC-Bench evaluates a variety of advanced MLLMs and foundation models. The generated results of these models could be biased or wrong. The related social impacts on the usage of AI-generated content may apply to our work. Overall, we consider MC-Bench exhibits minimal negative social impacts.

A.4. Limitations and Future Works

Although MC-Bench evaluates a wide spectrum of potential skills, it does not cover all possible vision-language tasks in real world and exhibits a long-tail distribution. Over time, we aim to expand MC-Bench by adding a greater variety of tasks and increasing the number of samples for the tail tasks. Meanwhile, MC-Bench currently focuses on multi-context samples consisting of two images and one corresponding text description. In the future, we aim to extend MC-Bench to accommodate a more general multi-context visual grounding task by incorporating more multi-context samples, each containing a larger number of images.

By the submission deadline, we have evaluated $\sim$20 recent representative approaches with publicly available checkpoints or APIs. Since several concurrent works have yet to release their code or checkpoints, we leave their evaluation for future work. We plan to establish a leaderboard for MC-Bench and update it as new approaches are introduced. Furthermore, in our current human evaluations, only three subjects are evaluated, and the results may vary to some extent due to differences in individual cognitive and reasoning levels, as well as the ambiguity and subjectivity of the text descriptions. We will invite more participants for human evaluations to establish a more robust upper bound, based on the average results of multiple individuals.

Benchmark results on MC-bench reveal a significant performance gap between MLLMs and humans, especially for the end-to-end models. While a few MLLMs accept image sequences as inputs, few of them are specifically designed for instance-level tasks. Our analysis experiments also show some potential areas for improvement. Driven by these observations, we plan to investigate more effective solutions for the multi-context visual grounding task in the future.

B. Implementation Details

Table A.1. Existing datasets incorporated in our MC-Bench. We collect and repurpose the images for multi-context visual grounding. The original tasks, original license information and URL links of source datasets are provided.

Source Datasets	Original Tasks	Original Licenses	URL Links
MS-COCO [41]	instance segmentation and image captioning	CC BY 4.0	URL
GRD [82]	referring expression segmentation	CC BY 4.0	URL
Q-Bench [80]	visual question answering on image quality	CC BY-NC-SA 4.0	URL
Mantis-Eval [27]	multi-image visual question answering	Apache-2.0	URL
DocVQA [52]	visual question answering on documents	N/A	URL
BLINK [17]	question answering on visual perception tasks	Apache-2.0	URL
CLEVR-Change [54]	visual question answering on scene changes	CC BY 4.0	URL
STAR [79]	visual question answering on videos	Apache-2.0	URL
NLVR2 [69]	multi-image visual question answering	N/A	URL
WinoGAViL [6]	vision-language associations	CC BY 4.0	URL
SEED-Bench2-plus [33]	visual question answering on text-rich images	CC BY 4.0	URL

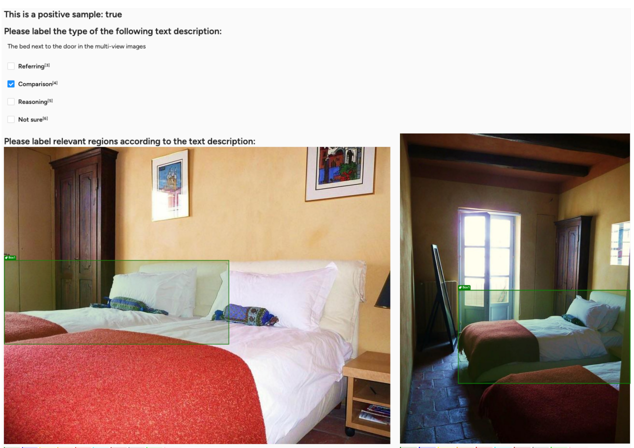


Figure B.1. The interface for collecting human annotations.

B.1. Existing Datasets Incorporated in MC-Bench

The images in MC-Bench are collected from multiple data sources. We list the used source datasets in Table A.1, and we also summarize their original tasks, original license information and URL links that may apply to future users.

B.2. Annotation Interfaces

We use the open-source annotation tool Label Studio [72] for annotations, in instance-level labeling stage. Figure B.1 shows the user-friendly interface used for collecting human annotations. The *positive sample label* in top left of the annotation interface indicates whether this sample is a positive sample. The annotators are first asked to verify whether the positive/negative sample label is correct. They are then required to categorize the style of the text prompts and draw the bounding boxes. If *positive sample label* of the sample being annotated is False, no box should be annotated.

B.3. Evaluated Baselines

Existing End-to-End Baselines. We evaluate several existing models with potential for multi-context visual grounding, including latest proprietary and open-source MLLMs as well as foundation models without LLMs. All the baselines are evaluated with the official pre-trained models and default hyper-parameters. For the proprietary API-based models, we evaluate the `gpt-4o-2024-05-13` version of GPT-4o [1] and the `gemini-1.5-pro-002` version of Gemini [61]. All experiments on open-source models were conducted on 4 NVIDIA RTX 3090 GPUs, except for Qwen2-VL-72B [75], which was excluded due to memory constraints.

For the models inherently accept multi-image inputs, we feed the image sequences to the models. For the models only supports single-image inputs, we horizontally concatenate image pairs and feed the merged images to the models. To allow the models to distinguish between image pairs, we add a thin white band between two images.

For the specialist [12, 40, 49, 55, 76, 78, 85] and a few generalist approaches [5, 9, 42] with predefined grounding prompts, we utilize their default prompts provided to localize target objects within images. As for the generalist models [1, 61, 75] without predefined grounding prompts, we carefully select the optimal prompts to generate the best results. Tables B.2 and B.3 showcase the prompts we use.

Stepwise Baseline. Following a divide-and-conquer strategy, we first leverage GPT-4o as a reasoning agent to analyze the target regions and generate some referring phrases that are easier for the detector to understand. Specifically, we use the GPT API and prompt the model of `gpt-4o-2024-05-13` version to generate the intermediate results, and the utilized prompt is presented in Table B.4.

We extract the phrase information from the JSON files generated by GPT. Then, we use the phrases as text query to localize objects from corresponding images. Concretely, the pre-trained G-DINO [45] with Swin-B [47] backbone

Table B.2. The system prompt we used for GPT-4o and Gemini-1.5-Pro. All the prompts are the same except for the second step in the key guidelines, which varies due to differences in the default coordinate system.

 # Your Role: excellent object detector

Objective
You will be provided with two images and a text describing some instances of interest in the images. Then, you will analyze all inputs and find instances / regions in the images that match the input text prompt from the images. Finally, you will output high-quality bounding box coordinates for each potential instance / region.

Key Guidelines
1. Generate one bounding box for one potential instance / region. Do not output bounding boxes covering multiple instances.
2. (for GPT-4o only) The top-left corner of the input images is coordinate [0, 0], and the bottom-right corner is [1, 1]. The output bounding box coordinate is in [x, y, w, h] format. You should also give confidence scores (range from 0 to 1) for every bounding boxes you predict.
2. (for Gemini only) The output coordinates are relative widths or heights in range [0,1], scaled by 1000 and converted to an integer. The output bounding box coordinate is in [xmin, ymin, xmax, ymax] format. You should also give confidence scores (range from 0 to 1) for every bounding boxes you predict.
3. The input textual prompt can indicate one or more instances / regions within the image pairs, or it can indicate no instance / region.
4. You should make full use of contextual information across input images to compare, analyze, and reason to find the target instances.
5. Output results strictly in accordance with the given output format, and converted into JSON format.
6. The input text prompt may describe multiple groups of instances. For example, ‘apples of the same colors’ may indicate several red apples and several green apples. In such case, you should group the red and green apple into different groups and add addition keys in the output, e.g., ‘Boxes within image1 (group 2)’. In each group, all the apples referred to is either red or green.
7. The group number depends on the inputs. The ‘Boxes within ...’ keys are not output if not applicable.

Output Format
Input prompt: [textual description for the candidate instances / regions]
Analysis: [interpret text prompt and paired images, then explain some key factors for decision making]
Positive: [answer ‘True’ if there is any relevant instance, otherwise answer ‘False’]
Selected image: [answer ‘image1’ or ‘image2’ or ‘both’ or ‘none’]
Boxes within image1 (group 1): [[box1 for instance1], [box2 for instance2], ...]
Scores within image1 (group 1) [score1 for box1, score2 for box2, ...]
Boxes within image2 (group 1): [[box3 for instance3], [box4 for instance4], ...]
Scores within image1 (group 1) [score3 for box3, score4 for box4, ...]
Boxes within image1 (group 2): [[box5 for instance5], [box6 for instance6], ...]
Scores within image1 (group 2) [score5 for box5, score6 for box6, ...]
Boxes within image2 (group 2): [[box7 for instance7], [box8 for instance8], ...]
Scores within image1 (group 2) [score7 for box7, score8 for box8, ...] (remove or add more groups if applicable)

is adopted. As each GPT-generated phrase only refers one instance within images, we selected the top-1 prediction as the final results. Moreover, a confidence threshold of 0.05 is used to filter out the less confident predictions.

Finetuned Baseline. We select the advanced Qwen2-VL-7B [75] as our baseline and construct an instruction tuning dataset for performance boosting. Concretely, the instruction tuning dataset contains two different types of data: multi-context samples for image-level tasks and instance-level tasks. We collect ∼7.5K multi-context samples from Birds-to-Words [16] and Multi-VQA [27], as datasets for multi-context image-level tasks (*e.g.*, multi-image captioning and image-level VQA) are already available. Due to the lack of multi-context instance-level task samples, we synthesize pseudo multi-image samples based on existing detection datasets (*i.e.*, LVIS [22] and OmniLabel [64]).

Table B.3. The system prompt we used for Qwen2-VL.

 # Your Role: excellent object detector

Objective
You will be provided with two images and a text describing some instances of interest in the images. Then, you will analyze all inputs and find instances / regions in the images that match the input text prompt from the images. Finally, you will output high-quality bounding box coordinates for each potential instance / region.

Key Guidelines
1. Generate one bounding box for one potential instance / region. Do not output bounding boxes covering multiple instances.
2. The input textual prompt can indicate one or more instances / regions within the image pairs, or it can indicate no instance / region.
3. You should also give confidence scores (range from 0 to 1) for every bounding boxes you predict.
4. You should make full use of contextual information across input images to compare, analyze, and reason to find the target instances.
5. Output results strictly in accordance with the given output format.

Output Format
The output format should strictly follow the examples:
1. <|img_id_start|>xx<|img_id_end|><|object_ref_start|>xxx<|object_ref_end|><|box_start|>(xx,xx),(xx,xx)<|box_end|><|score_start|>xx<|score_end|>
2. <|img_id_start|>xx<|img_id_end|><|object_ref_start|>xxx<|object_ref_end|><|box_start|>(xx,xx),(xx,xx)<|box_end|><|score_start|>xx<|score_end|><|img_id_start|>xx<|img_id_end|><|object_ref_start|>xxx<|object_ref_end|><|box_start|>(xx,xx),(xx,xx)<|box_end|><|score_start|>xx<|score_end|>...
3. xxx does not exist.

More specifically, we randomly select two images and generate instructions based on the original object category or referring annotations of the two images, such as '*Output the bounding boxes of <category name> in the first image*', '*Output the bounding boxes of <object description> in the two images*', and *etc*. We generate ∼53K synthetic multi-context instance-level task samples for training.

We finetune Qwen2-VL-7B with the LoRA [25] using LLaMA-Factory [95] framework. The model is finetuned using ∼60K training samples and bfloat16 format over 3 epochs. The learning rate is set to 1e-4 with a cosine annealing scheduler, and the global batch size is set to 32. All other settings and hyper-parameters follow the default choices of LLaMA-Factory. After the model trained, we use the prompt '*Output the bounding boxes of <input description>*' for multi-context visual grounding.

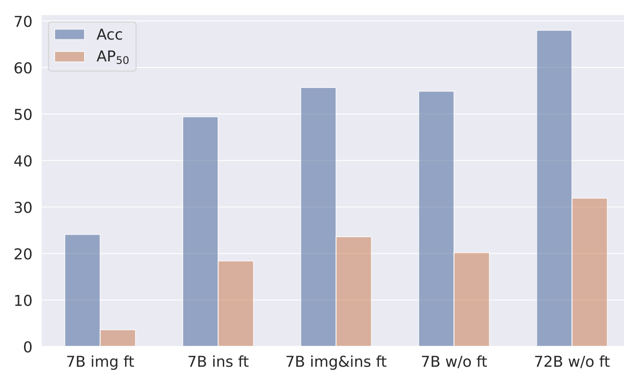


Figure B.2. Effectiveness of instruction tuning data for the finetuned baseline. The *img ft*, *ins ft* and *img&ins ft* denote the models trained with collected image-level task samples, synthetic instance-level task samples and merged samples, respectively.

C. Additional Experimental Results

Ablation Study for the Finetuned Baseline. Figure B.2 illustrates the effectiveness of instruction tuning with different data. Models finetuned with only multi-context image-level task or instance-level task samples obtain performance degradation. Particularly, the performance of model trained with only collected image-level task samples decreases significantly. The model trained with only synthetic instance-level task samples also shows sightly performance drop compared to the model without instruction tuning. We conjecture that most of the synthetic data generated based on object detection datasets [22, 64] only boosts the cross-image referring abilities and brings limited cross-image comparison and reasoning capabilities. After train-

Table B.4. The system prompt we used for grounding phrase generation in the stepwise baseline.

 # Your Role: excellent referring phrase generator

Objective
You will be provided with two images and a text describing some instances of interest in the images. Then, you will analyze all inputs and find instances / regions in the images that match the input text prompt from the images. Finally, you will output high-quality referring phrases for each potential instance / region for subsequent grounding tasks.

Key Guidelines
1. Writing a unique referring phrase for each potential instance / region. Do not output a phrase to refer to multiple instances.
2. The given referring phrases should be as concise as possible while maintaining sufficient distinctiveness, allowing for easy differentiation of an instance from the image based on the provided referring phrases.
3. The input textual prompt can indicate one or more instances / regions within the image pairs, or it can indicate no instance / region.
4. The given referring phrase could include the appearance, category and context information of the candidate instances / regions. Any other clues that can better differentiate and identify candidate areas/objects are acceptable.
5. The given referring phrase cannot contain cross-image information.
6. Output results strictly in accordance with the given output format, and converted into JSON format.
7. The input text prompt may describe multiple groups of instances. For example, 'apples of the same colors' may indicate several red apples and several green apples. In such case, you should group the red and green apple into different groups and add addition keys in the output, e.g., 'Referring phrases for instances within image1 (group 2)'. In each group, all the apples referred to is either red or green.
8. The group number depends on the inputs. The 'Referring phrases ...' keys are not output if not applicable.

Output Format
Input prompt: [textual description for the candidate instances / regions]
Analysis: [interpret text prompt and paired images, then explain some key factors for decision making]
Positive: [answer 'True' if there is any relevant instance, otherwise answer 'False']
Selected image: [answer 'image1' or 'image2' or 'both' or 'none']
Referring phrases for instances within image1 (group 1): ['phrase1 for instance1', 'phrase2 for instance2', ...]
Referring phrases for instances within image2 (group 1): ['phrase3 for instance3', 'phrase4 for instance4', ...]
Referring phrases for instances within image1 (group 2): ['phrase5 for instance5', 'phrase6 for instance6', ...]
Referring phrases for instances within image2 (group 2): ['phrase7 for instance7', 'phrase8 for instance8', ...] (remove or add more groups if applicable)

ing model with merged data, the finetuned baseline achieves the best performance across image-level and instance-level metrics, surpassing the pre-trained Qwen2-VL-7B by a non-trivial margin. We also notice that a clear performance gap remains when compared to the 72B model.

D. Datasheets for MC-Bench

D.1. Motivation

1. **For what purpose was the dataset created?** (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)
   The primary purpose of MC-Bench is to function as a dynamic benchmark that continuously evolves and evaluates MLLMs for open-ended visual grounding in multi-image scenarios. This dataset first explores a significant yet largely overlooked research problem, *i.e.*, grounding objects from multi-image inputs based on open-ended textual prompts. The benchmark results on MC-Bench show a large performance gap between existing MLLMs and humans, as illustrated in Table 2 in the main text.
2. **Who created this dataset (*e.g.*, which team, research group) and on behalf of which entity (*e.g.*, company, institution, organization)?**
   This dataset was created by the authors of this paper.
3. **Who funded the creation of the dataset?** (If there is an associated grant, please provide the name of the grantor and the grant name and number.)
   The institute of the authors funded the creation of the dataset.

4. **Any other comments?**
None.

D.2. Composition

1. **What do the instances that comprise the dataset represent (*e.g*., documents, photos, people, countries)?** (Are there multiple types of instances (*e.g*., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)
An instance of our dataset represent the multimodal triplet (*i.e*., an image pair and a textual prompt describing the regions/objects within the images). More detailed descriptions are provided in our paper.
2. **How many instances are there in total (of each type, if appropriate)?**
Our dataset owns 2,000 samples (*i.e*., paired images and corresponding text descriptions). We provide more detailed dataset statistics in our paper.
3. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** (If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (*e.g*., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (*e.g*., to cover a more diverse range of instances, because instances were withheld or unavailable).)
The dataset cannot contain all possible instances, as the dataset is designed for open-ended visual grounding evaluation. We try to covering diverse range of image domains, disciplines and skills, but we can't guarantee a full sampling of them as discussed in §A.4.
4. **What data does each instance consist of?** "Raw" data (*e.g*., unprocessed text or images) or features? In either case, please provide a description
Each instance of our dataset represent an image pair and a textual prompt describing the regions/objects within the images.
5. **Is there a label or target associated with each instance?** If so, please provide a description.
Yes. We provide the bounding box annotations covering the regions described by the textual prompts. More detailed descriptions are provide in our paper.
6. **Is any information missing from individual instances?** (If so, please provide a description, explaining why this information is missing (*e.g*., because it was unavailable). This does not include intentionally removed information, but might include, *e.g*., redacted text.)
No. All necessary information has been provided.
7. **Are relationships between individual instances made explicit (*e.g*., users' movie ratings, social network links)?** (If so, please describe how these relationships are made explicit.)
Yes. Instances are categorized into three groups (*i.e*., referring, comparison and reasoning) based on the text prompt style of each instance.
8. **Are there recommended data splits (*e.g*., training, development/validation, testing)?** (If so, please provide a description of these splits, explaining the rationale behind them.)
Yes. As MC-Bench is an evaluate-only dataset, all samples belong to the testing split.
9. **Are there any errors, sources of noise, or redundancies in the dataset?** (If so, please provide a description.)
Yes. We try our best to improve the quality of annotations, but the dataset might still contain a few missing labeled objects or subjectivity inconsistencies.
10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (*e.g*., websites, tweets, other datasets)?** (If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (*i.e*., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (*e.g*., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)
Images in MC-Bench are from other publicly available datasets or self-contained. We repurpose these images for multi-context visual grounding. These external datasets are commonly used and long-term exist. We use the official archival versions of them. More detailed descriptions of all external resources are provided in §B.1.
11. **Does the dataset contain data that might be considered confidential (*e.g*., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?** (If so, please provide a description.)
No.
12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** (If so, please describe why.)
Yes. Some of the scenes may bring anxiety to some people, *e.g*., photos of car accidents and hospital surgeries. However, we consider our dataset's offensiveness to be limited, since the source images are collected from prior public datasets.
13. **Does the dataset identify any subpopulations (*e.g*., by age, gender)?** (If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.)
No. This is not explicitly identified.
14. **Is it possible to identify individuals (*i.e*., one or more**

natural persons), either directly or indirectly (*i.e.*, in combination with other data) from the dataset? (If so, please describe how.)
Yes. Some samples are about referring expression understanding, where models are required to localize some individuals from images based on the textual descriptions.

15. **Does the dataset contain data that might be considered sensitive in any way (*e.g.*, data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** (If so, please provide a description.)
No. There are no sensitive data used.
16. **Any other comments?**
None.

D.3. Collection Process

1. **How was the data associated with each instance acquired?** (Was the data directly observable (*e.g.*, raw text, movie ratings), reported by subjects (*e.g.*, survey responses), or indirectly inferred/derived from other data (*e.g.*, part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)
The images are collected from existing public data sources. The text descriptions of image pairs are written by the annotators, based on the content of the image pairs.
2. **What mechanisms or procedures were used to collect the data (*e.g.*, hardware apparatus or sensor, manual human curation, software program, software API)?** (How were these mechanisms or procedures validated?)
Software program and manual human curation.
3. **If the dataset is a sample from a larger set, what was the sampling strategy (*e.g.*, deterministic, probabilistic with specific sampling probabilities)?**
The image are randomly selected from other datasets with specific topics.
4. **Who was involved in the data collection process (*e.g.*, students, crowdworkers, contractors) and how were they compensated (*e.g.*, how much were crowdworkers paid)?**
The data are collected by the authors and students. The involved students are paid nicely.
5. **Over what timeframe was the data collected?** (Does this timeframe match the creation timeframe of the data associated with the instances (*e.g.*, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.)
The dataset was collected in the Spring of 2024, which does not necessarily reflect the timeframe of the data collected.
6. **Were any ethical review processes conducted (*e.g.*, by an institutional review board)?** (If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.)
No ethical review processes were conducted, since the source images are collected from other public datasets.
7. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (*e.g.*, websites)?**
The images are collected from other sources (*i.e.*, repurpose published datasets), while the text descriptions and bounding boxes are labeled by our annotators.
8. **Were the individuals in question notified about the data collection?** (If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.)
N/A.
9. **Did the individuals in question consent to the collection and use of their data?** (If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.)
N/A.
10. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** (If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).)
N/A.
11. **Has an analysis of the potential impact of the dataset and its use on data subjects (*e.g.*, a data protection impact analysis) been conducted?** (If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.)
N/A.
12. **Any other comments?**
None.

D.4. Preprocessing/cleaning/labeling

1. **Was any preprocessing/cleaning/labeling of the data done (*e.g.*, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** (If so, please provide a description. If not, you may skip the remainder of the questions in this section.)

Yes. We reorganized images collected from existing datasets and introduced extra annotations. Specifically, we provided textual prompts for each image pair describing some objects within the images, and we also labeled the language-grounded regions using bounding boxes.

2. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (*e.g.*, to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.
   Yes. MC-Bench itself contains partial the raw data (*i.e.*, textual descriptions and bounding box annotations). The rest of raw data (*i.e.*, images) were collected from other published datasets (see §B.1) and we did not modify the images.
3. **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.
   We leverage the open-source annotation tool, Label Studio (`https://github.com/HumanSignal/label-studio`), in both text and box annotation stages, owing to its programmable and user-friendly interface for annotating paired images.
4. **Any other comments?**
   None.

D.5. Uses

1. **Has the dataset been used for any tasks already?** (If so, please provide a description.)
   The images of MC-Bench are collected from published datasets for other tasks. In contrast, the textual prompts and bounding box annotations in MC-Bench are newly introduced and have not used for any other tasks.
2. **Is there a repository that links to any or all papers or systems that use the dataset?** (If so, please provide a link or other access point.)
   Yes. We are going to maintain a leaderboard for MC-Bench on the project page (`https://xuyunqiu.github.io/MC-Bench/`). The links of all the evaluated methods will be provided.
3. **What (other) tasks could the dataset be used for?**
   There are many more, such as multi-image VQA and common object detection.
4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (*e.g.*, stereotyping, quality of service issues) or other undesirable harms (*e.g.*, financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?)
   No.
5. **Are there tasks for which the dataset should not be used?** (If so, please provide a description.)
   No.
6. **Any other comments?**
   None.

D.6. Distribution

1. **Will the dataset be distributed to third parties outside of the entity (*e.g.*, company, institution, organization) on behalf of which the dataset was created?** (If so, please provide a description.)
   Yes, the dataset is publicly available on the Internet.
2. **How will the dataset will be distributed (*e.g.*, tarball on website, API, GitHub)?** (Does the dataset have a digital object identifier (DOI)?)
   On our GitHub project page (`https://xuyunqiu.github.io/MC-Bench/`).
3. **When will the dataset be distributed?**
   The dataset was first released in June 2024.
4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** (If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.)
   The dataset is licensed under a CC license. More detailed license information is provided in §A.1.
5. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.)
   As far as we know, no.
6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.)
   As far as we know, no.
7. **Any other comments?**
   None.

D.7. Maintenance

1. **Who is supporting/hosting/maintaining the dataset?**
   The authors.
2. **How can the owner/curator/manager of the dataset be contacted (*e.g.*, email address)?**
   The dataset owner can be contacted through the authors' email address.
3. **Is there an erratum?** (If so, please provide a link or other access point.)

Currently, no. As errors are encountered, future versions of the dataset may be released (but will be versioned).

4. **Will the dataset be updated (*e.g.*, to correct labeling errors, add new instances, delete instances')?** (If so, please describe how often, by whom, and how updates will be communicated to users (*e.g.*, mailing list, GitHub)?)
   Yes. The dataset will be update by the dataset owner. The update information will be posted on the project page.
5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (*e.g.*, were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** (If so, please describe these limits and explain how they will be enforced.)
   No.
6. **Will older versions of the dataset continue to be supported/hosted/maintained?** (If so, please describe how. If not, please describe how its obsolescence will be communicated to users.)
   Yes. The older versions of the dataset will be provided in the same webpage.
7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** (If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.)
   Yes. Others may do so and should contact the original authors about incorporating fixes/extensions.
8. **Any other comments?**
   None.