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

Supplementary Material

A. General Discussions

A.1. License

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

A.2. Intended Use

We believe that images in the real world are not isolated, but are inherently linked via spatial, temporal or semantic context. 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).

MC-Bench is highly relevant and beneficial to various downstream applications. For spatial-relevant context evaluations, MC-Bench assesses multi-view reasoning, which is particularly valuable for robotic navigation and manipulation (illustrated in Figure 7). In addition to these spatial-aware applications, there is a broader range of potential real-world applications that benefit from leveraging temporal and semantic context (*e.g.*, animal/traffic surveillance, sports/food analysis and GUI agents).

The primary purpose of MC-Bench is to function as a dynamic benchmark that continuously evolves and evaluates MLLMs for multi-context visual grounding. The preliminary 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.

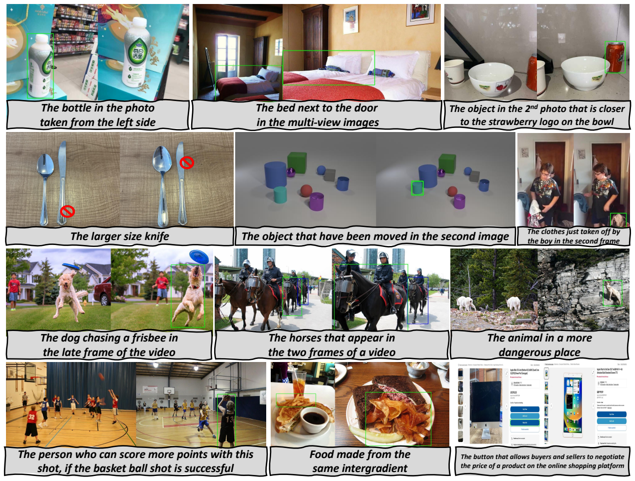

The bottle in the photo taken from the left side
The bed next to the door in the multi-view images
The object in the 2nd photo that is closer to the strawberry logo on the bowl
The larger size knife
The object that have been moved in the second image
The clothes just taken off by the boy in the second frame
The dog chasing a frisbee in the late frame of the video
The horses that appear in the two frames of a video
The animal in a more dangerous place
The person who can score more points with this shot, if the basket ball shot is successful
Food made from the same intergradient
The button that allows buyers and sellers to negotiate the price of a product on the online shopping platform

Figure 7. MC-Bench includes spatial-, temporal- or semantic-relevant context evaluations, which is highly relevant and beneficial to various downstream real-world applications (*e.g.*, robotics, surveillance systems and general-purpose assistants).

A.4. Limitations and Future Works

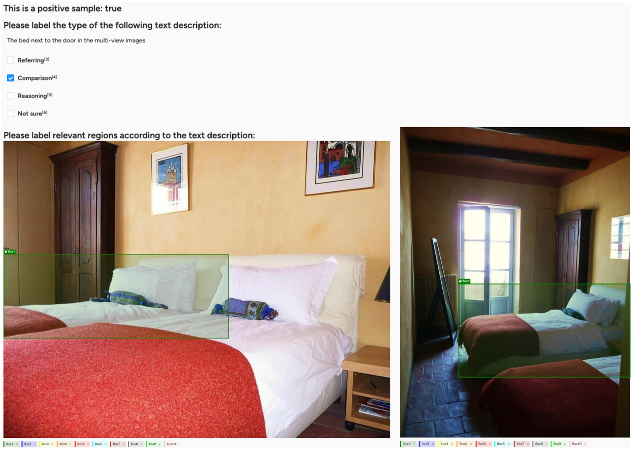
Although MC-Bench evaluates a wide spectrum of potential skills, it does not cover all possible vision-language tasks in real world and exhibits a long-tail distribution. Over time, we aim to expand MC-Bench by adding a greater variety of tasks and increasing the number of samples for the tail tasks. 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 ICCV submission deadline, we have evaluated ∼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 and update it as new approaches are introduced.

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

Table 3. 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 [13]	instance segmentation and image captioning	CC BY 4.0	URL
GRD [30]	referring expression segmentation	CC BY 4.0	URL
Q-Bench [29]	visual question answering on image quality	CC BY-NC-SA 4.0	URL
Mantis-Eval [10]	multi-image visual question answering	Apache-2.0	URL
DocVQA [18]	visual question answering on documents	N/A	URL
BLINK [7]	question answering on visual perception tasks	Apache-2.0	URL
CLEVR-Change [19]	visual question answering on scene changes	CC BY 4.0	URL
STAR [28]	visual question answering on videos	Apache-2.0	URL
NLVR2 [23]	multi-image visual question answering	N/A	URL
WinoGAViL [3]	vision-language associations	CC BY 4.0	URL
SEED-Bench2-plus [11]	visual question answering on text-rich images	CC BY 4.0	URL


This is a positive sample: true

Please label the type of the following text description:

The bed next to the door in the multi-view images

□ Referring[3]

☑ Comparison[4]

□ Reasoning[5]

□ Not sure[6]

Please label relevant regions according to the text description:

Figure 8. The interface for collecting human annotations.

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

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 3, and we also summarize their original tasks, original license information and URL links that may apply to future users. For datasets released under licenses other than CC BY (*e.g.*, Q-Bench [29]), we obtain permission to include their images in our MC-Bench.

B.2. Annotation Interfaces

We use the open-source annotation tool Label Studio [24] for annotations, in instance-level labeling stage. Figure 8 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.

Table 4. Annotation format of MC-Bench.

```
{
    "info"             : info,
    "images"           : [image],
    "annotations"      : [annotation],
    "descriptions"     : [description],
    "categories"       : categories
}

description{
    "id"               : int,
    "images_id"        : [int],
    "text"             : str,
    "positive_sample"  : bool,
    "text_style"       : str
}

image{
    "id"               : int,
    "text_id"          : int,
    "inter_img_id"     : int,
    "file_name"        : str,
    "height"           : int,
    "width"            : int
}

annotation{
    "id"              : int,
    "image_id"        : int,
    "text_id"         : int,
    "category_id"     : int,
    "area"            : int,
    "bbox"            : [x,y,w,h],
    "iscrowd"         : 0 or 1
}
```

B.3. Annotation Format

The annotation format of our MC-Bench is similar to MS-COCO [13]. As illustrated in Table 4, the main contents

Table 5. The prompt we used for GPT-4o.

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. 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.
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)

User: <input description><image 1><image 2>
GPT: ...... (bounding box coordinates following given format)

are saved in `description`, `image` and `annotation`, including the text prompts, image and instance-level annotation information. The annotations are stored using JSON file. Our MC-Bench API can be used to access and manipulate annotations.

B.4. 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 [21]. All experiments on open-source models were conducted on 4 NVIDIA RTX 3090 GPUs, except for Qwen2-VL-72B [25], which was excluded due to memory

Table 6. The multi-round prompt we used for Gemini-1.5-Pro.

User: *Which picture contains the instance described by <input description>?<image 1><image 2>*
Gemini: ...... (unstructured results containing image referents and concise analysis)

User: *Please answer using one of* {*both pictures, the first picture, the second picture, no existence*}.
Gemini: ...... (image referents selected from the given template)

User: *Return bounding boxes for <input description>, using [ymin, xmin, ymax, xmax] format.<image n>*
Gemini: ...... (bounding box coordinates following given format)

Table 7. The prompt we used for Qwen2-VL.

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. 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.

User: <image 1><image 2><input description>
Qwen: ...... (bounding box coordinates following given format)

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 [5, 12, 17, 20, 26, 27, 31] and a few generalist approaches [2, 4, 14] with predefined grounding prompts, we utilize their default prompts provided to localize target objects within images. As for the generalist models [1, 21, 25] without predefined grounding prompts, we carefully select the optimal prompts to generate the best results. Tables 5-7 showcase the prompts we use.

Agentic 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 8.

We extract the phrase information from the JSON files

Table 8. The prompt we used for grounding phrase generation in the agentic baseline.

System:
# 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)

User: <input description><image 1><image 2>
GPT: ...... (generated phrases for subsequent grounding)

generated by GPT. Then, we use the phrases as text query to localize objects from corresponding images. Concretely, the pre-trained G-DINO [15] with Swin-B [16] backbone 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 [25] 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 [6] and Multi-VQA [10], 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 [8] and OmniLabel [22]). 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*

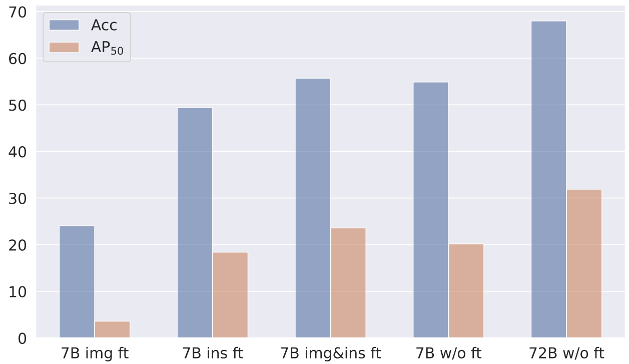


Figure 9. Effectiveness of instruction tuning data for the finetuned baseline. The *img ft*, *ins ft* and *img&ins ft* denote the models trained with collected image-level task samples, synthetic instance-level task samples and merged samples, respectively.

two images', and *etc*. We generate ∼53K synthetic multi-context instance-level task samples for training.

We finetune Qwen2-VL-7B with the LoRA [9] using LLaMA-Factory [32] 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.

C. Additional Experimental Results

Ablation Study for the Finetuned Baseline. Figure 9 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 [8, 22] only boosts the cross-image referring abilities and brings limited cross-image comparison and reasoning capabilities. After training 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.

Human Evaluation. In our current human evaluations, the results of three subjects vary to some extent due to differences in individual cognitive and reasoning levels, as well as the ambiguity and subjectivity of text descriptions. We report more detailed human evaluation results in Table 9. We find that the worst performing volunteer still outperforms existing end-to-end MLLMs and achieves on par AP_{50} results to the agentic baseline.

Table 9. Detailed human evaluation results of three volunteers on MC-Bench.

#	Acc^{ref}	Acc^{com}	Acc^{rea}	Acc	AP_{50}^{ref}	AP_{50}^{com}	AP_{50}^{rea}	AP_{50}
1	92.2	96.5	93.0	94.3	52.9	44.0	46.4	46.0
2	89.6	95.6	89.9	92.2	52.6	40.7	44.1	42.8
3	86.7	94.2	88.5	90.5	37.9	36.2	32.4	35.0

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 §3.3 in the main text.
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. The updated dataset (MC-Bench v0.5) will be released alongside the ICCV camera-ready version in late July 2025.
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 updated 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.

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] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.