

# ◆ ChatterBox: Multi-round Multimodal Referring and Grounding

Yunjie Tian\*  
UCAS

tianyunjie19@mails.ucas.ac.cn

Tianren Ma\*  
UCAS

matianren18@mails.ucas.ac.cn

Lingxi Xie  
Huawei Inc.

198808xc@gmail.com

Jihao Qiu  
UCAS

qiujihao19@mails.ucas.ac.cn

Xi Tang  
UCAS

tangxi19@mails.ucas.ac.cn

Yuan Zhang  
UCAS

zhangyuan192@mails.ucas.ac.cn

Jianbin Jiao  
UCAS

jiaojb@ucas.ac.cn

Qi Tian  
Huawei Inc.

tianqi1@huawei.com

Qixiang Ye  
UCAS

qxye@ucas.ac.cn

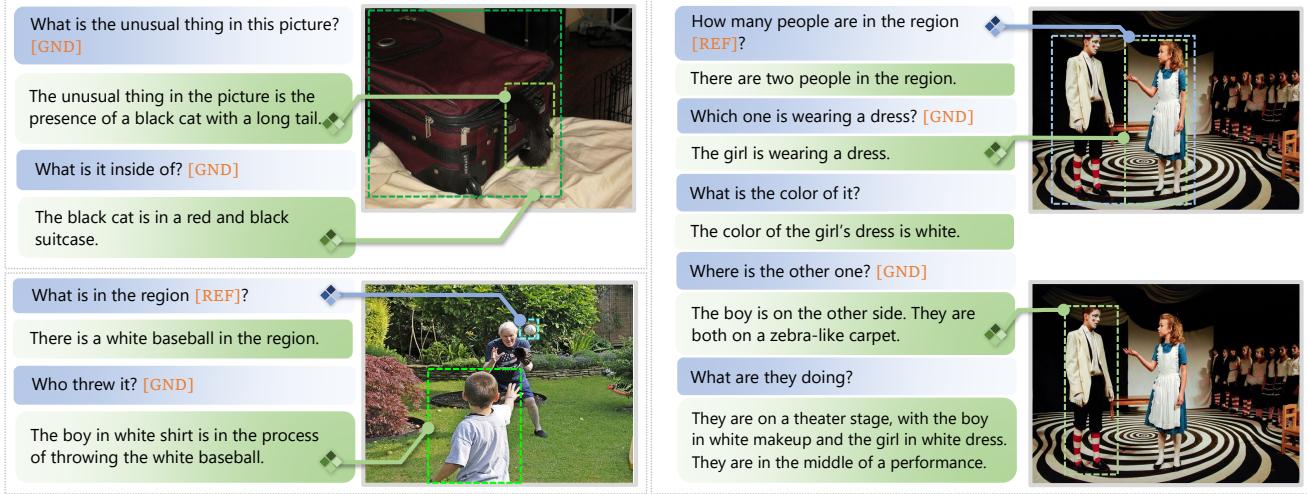


Figure 1. A showcase of the multi-round referring and grounding (MRG) task. During the dialogue, the agent can receive either a [REF] token for referring expressions or a [GND] token for visual grounding; without these tokens, the task becomes generic visual question answering. All the answers are generated by the ChatterBox agent, demonstrating its strong ability in visual recognition. In particular, ChatterBox can understand **logically related** questions and incorporate contextual information to provide answers. For instance, in the right-hand thread, the question ‘Where is the other one?’ necessitates the agent to recognize that ‘one’ refers to a person and then locate the ‘other’ person distinct from the one mentioned earlier.

## Abstract

*In this study, we establish a baseline for a new task named multimodal multi-round referring and grounding (MRG), opening up a promising direction for instance-level*

*multimodal dialogues. We present a new benchmark and an efficient vision-language model for this purpose. The new benchmark, named CB-300K, spans challenges including multi-round dialogue, complex spatial relationships among multiple instances, and consistent reasoning, which are beyond those shown in existing benchmarks. The proposed model, named ChatterBox, utilizes a two-branch architec-*

\*Equal contribution.

ture to collaboratively handle vision and language tasks. By tokenizing instance regions, the language branch acquires the ability to perceive referential information. Meanwhile, ChatterBox feeds a query embedding in the vision branch to a token receiver for visual grounding. A two-stage optimization strategy is devised, making use of both CB-300K and auxiliary external data to improve the model’s stability and capacity for instance-level understanding. Experiments show that ChatterBox outperforms existing models in MRG both quantitatively and qualitatively, paving a new path towards multimodal dialogue scenarios with complicated and precise interactions. Code, data, and model are available at: <https://github.com/sunsmarterjie/ChatterBox>.

## 1. Introduction

Large language models (LLMs) have shown impressive capabilities across a wide range of natural language tasks [2]. In the computer vision community, researchers have integrated LLMs with images and videos, creating a series of multimodal large language models (MLLMs) [1, 19, 20, 24]. Recently, with the success of instruction tuning, comprehensive image-text instruction datasets have been constructed [18, 24, 41, 48], empowering the ability of MLLMs in a wide range of multimodal understanding tasks.

We argue that a powerful multimodal agent should have the ability to understand logically related questions and perform basic vision-aware tasks such as referring and grounding (see Figure 1). However, few existing models were equipped with all these abilities. To fill up the empty, we propose the multi-round multimodal referring and grounding (MRG) task, where the MLLM is expected to engage in referring (*i.e.*, recognizing a designated region or object) and grounding (*i.e.*, locating a region or object from the image) tasks at any time in a multi-round dialogue, meanwhile retaining the consistent logic of the entire conversation. MRG is similar to the behavior in which humans interact with agents.

We make two-fold contributions to enable MRG. Firstly, we establish a new benchmark named CB-300K, which comprises the first-ever image-text dataset for MRG and an evaluation metric that takes the accuracy of both visual and linguistic understanding into consideration. The construction of CB-300K mainly builds upon Visual Genome [17], where we feed the metadata into GPT-4 [29] and prompt it to generate multi-round dialogues with referring and grounding requests. Post-processing is then performed to guarantee the correctness of dialogues and organize them to subsets for various purposes.

Secondly, we propose an MLLM named ChatterBox to solve the challenging task. The key lies in the integration of visual and linguistic information. To fulfill this purpose, we design a two-branch architecture, where the language

Table 1. A comparison of ChatterBox to recent studies *w.r.t.* the abilities to perform multi-round dialogues (including region-level referring and visual grounding), the proposal of new data ( $\dagger$ : it involves generating new dialogue data rather than simply reorganizing existing data), and training costs. N/R: not reported.

Method	Multi-Round	Region Referring	Visual Grounding	New Data $\dagger$	Training (G-days)
LLaVA [24]	✓	✗	✗	✓	4.7
InstructBLIP [9]	✓	✗	✗	✗	24
VisionLLM [40]	✗	✓	✓	✗	N/R
Kosmos-2 [31]	✓	✓	✓	✓	256
GPT4RoI [48]	✓	✓	✗	✗	N/R
LISA [18]	✓	✗	✓	✗	8
<b>ChatterBox</b>	✓	✓	✓	✓	15

branch understands the logic of the question, and the vision branch plays the role of visual feature extraction and recognition (*e.g.*, grounding). The ChatterBox design can be easily understood and optimized, requiring only 15 GPU-days for a two-stage optimization process that involves both CB-300K and auxiliary external data (*e.g.*, RefCOCO [14] and LLaVA-Instruction-150K [24]).

We conduct both quantitative and qualitative studies on the CB-300K benchmark and validate ChatterBox’s superiority over existing models in MRG. Some examples of ChatterBox performing MRG are displayed in Figure 1. ChatterBox also transfers to easier tasks (*e.g.*, single-round dialogue, referring, grounding) seamlessly. Our research advocates that delicate and precise interactions are strongly required to enhance the ability of multimodal dialogue as well as artificial general intelligence systems.

We compare our work to recent studies on multimodal dialogue in Table 1 and summarize our contributions below:

- We introduce a new task named multi-round multimodal referring and grounding (MRG).
- We propose a data construction scheme and establish the CB-300K benchmark to facilitate the research in MRG.
- We present ChatterBox, a vision-language model that injects explicit vision modules into an MLLM, providing an agile and effective solution of MRG.

## 2. Related Work

### 2.1. Multimodal Large Language Models

Large language models [2, 6–8, 10, 37, 39, 45, 47] have opened a new era of AI, demonstrating the potential to create a generalist model that can even cover different modalities. The computer vision community has witnessed a trend of unifying vision and language data using multimodal large language models [1, 19, 20, 24]. The pioneering efforts involved projecting vision and language data into the same

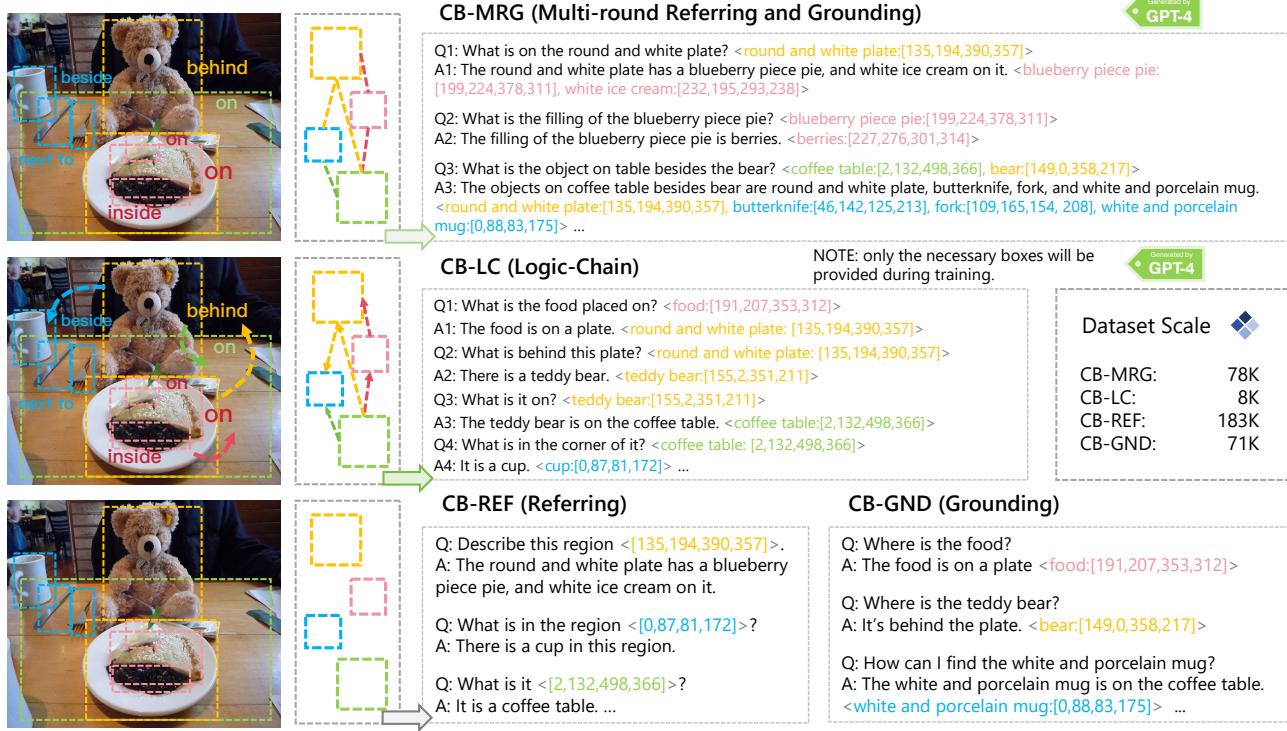


Figure 2. The CB-300K data contains four subsets for different purposes. The images and metadata (object locations and descriptions) are inherited from Visual Genome. The same image can appear in different subsets. The former two subsets, CB-MRG and CB-LC, are obtained by prompting GPT-4 to read the metadata and generate questions and answers. The latter two subsets, CB-REF and CB-GND, are produced using manually designed rules and then polished by GPT-3.5. *This figure is best viewed in color.*

feature space [1, 33]. Attempts to adapt a LLM to visual tasks have been made internally or externally: Flamingo [1] interleaves cross-attention blocks within a LLM for visual-language alignment. BLIP-2 [20] proposes Q-former, an external block that uses multiple vision-language losses to align queried visual features with text.

## 2.2. Multimodal Instruction Data

Later, inspired by the instruction tuning mechanism [30] of the GPT series, MLLMs started collecting instruction data from various sources. One of the early efforts was visual instruction tuning [24] which provides a novel method for data construction. By feeding external metadata including bounding box annotations and textual descriptions into GPT-4 with designed prompts, comprehensive detailed conversations about the images can be generated without vision access. The idea was followed by other works [4, 5] to harvest various types of instruction data. In another approach, the image feature was fed into an MLLM and prompted for instruction data [52]. Additionally, richer information (*e.g.*, phrase grounding) was also collected [31] with the assistance of external vision-language models, such as GLIP [21]). The new data and learning strategy enabled more abilities to emerge via multimodal dia-

logue [1, 11, 18, 24, 41, 43].

## 2.3. Instance-Level Understanding of MLLMs

MLLMs can be largely enhanced by the ability of instance-level understanding, *i.e.*, the models can (1) respond to questions targeted at specified regions of the image and (2) find regions that correspond to the contents in the dialogue. We address these two abilities as visual referring [4, 48] and visual grounding [25, 31], respectively, and they have been well studied in the computer vision community under various task designations and settings. Integrating them into MLLMs, however, remains a challenge. There are two main approaches to integration, differing in whether to encode the position information explicitly or not. Explicit methods [31, 40] are easier to optimize and explain by introducing location tokens, while implicit methods [5, 41, 42] offer greater flexibility. Nonetheless, these multimodal methods are not well-suited for handling instance-level multi-round dialogues, which can lead to superficial interaction.

## 3. The CB-300K Benchmark

We establish a benchmark named ChatterBox-300K with the aim to enhance the ability of multimodal dialogue sys-

tems in multi-round referring and grounding. Similar to previous work [24], the dataset is mainly constructed upon context-rich image data (where we use the Visual Genome dataset [17] for the richness in the annotation of object-level relationship) and assisted by a large language model (where we use GPT-4 as an off-the-shelf model for understanding). We also define a metric for evaluating the models in this new scenario.

### 3.1. Data Collection

When an image is sampled from Visual Genome, we refer to the annotation data which mainly has three parts: (1) objects with bounding boxes (*e.g.*, there is a man at  $[x_1, y_1, w_1, h_1]$  and a computer at  $[x_2, y_2, w_2, h_2]$ ), (2) the relationship between objects (*e.g.*, the man is operating the computer), (3) auxiliary attributes of objects (*e.g.*, the man is in black). We summarize all the information (in pure texts) as contexts, feed them to GPT-4, and ask it to generate question-and-answer pairs of different aspects. An additional request for GPT-4 is that, in each sequence of consecutive question-and-answer pairs, the latter questions should build on the former ones, so that we can fully test the model’s ability to engage in multi-round dialogue.

In summary, there are four subsets in CB-300K including a generic subset for MRG and three specifically designed subsets, as illustrated in Figure 2. We elaborate the design principles as follows and leave the technical details (*e.g.*, the full prompts for GPT-4, and the Python code for data generation and filtering) in the supplementary material.

- **CB-MRG** is a generic data collection for **MRG**. We ask GPT-4 to generate questions and answers that focus on instance-level relationship. Provided with related instances, GPT-4 shall write dialogues using their relationship. Meanwhile, their location (as a bounding box, formed  $[x, y, w, h]$ ) are also attached in each question and answer, retaining the maximum spatial information for further referring and grounding tasks. We also write additional prompts, asking GPT-4 to create multi-round referring and grounding requests, which is achieved by asking about further relationships upon the instance(s) that appeared in the former answers.
- **CB-LC** is a subset that extends the ability for **logic-chain** MRG. The key differences include (1) adding strict restrictions upon the prompt we used to generate CB-MRG (*e.g.*, each question must be built upon exactly one aforementioned relationship), (2) deleting invalid question-and-answer pairs using manually-designed rules (*e.g.*, logic-deficient dialogues, mismatched or missing boxes, *etc.*; more details are provided in the supplementary material), and (3) calling GPT-4 again to check the entire thread, cleaning up incorrect descriptions and contradictions. The strict filtering procedure makes CB-LC a high-quality subset for logic-chain MRG, but the size is rela-

Table 2. The number of threads and the number of question-and-answer pairs of each individual subset and the entire benchmark.

Set	# threads	# Q&A pairs
CB-MRG	77,814	437,229
CB-LC	7,834	25,617
CB-REF	183,446	183,446
CB-GND	70,783	70,783
CB-300K	339,877	717,075

tively small<sup>\*</sup>.

- **CB-REF** adds more dialogues for **referring expression**. These dialogues are generated by a manually designed rule that extracts the annotated bounding box(es) as referential inputs and the description of the region as answers. Various Q&A styles are applied and GPT-3.5 is used to polish the grammatical issues.
- **CB-GND** adds more dialogues for **visual grounding**. These dialogues are generated manually using the reversed rules of CB-REF.

During the generation procedure, the description and bounding boxes are available for each instance in each question. To facilitate multi-round referring, we post-process the data, using the pronoun ‘it’ to replace the whole unit. The processing details are different between training and testing and will be elaborated in the following sections.

Table 2 displays the statistics of CB-300K. We randomly split the dataset by leaving out 800 and 200 threads from CB-MRG and CB-LC for testing and using the remaining data for training. We guarantee that different occurrences of the same image (in different subsets) do not appear in the training and testing splits simultaneously.

**Comparison to other datasets.** The CB-300K dataset differs from existing datasets for multimodal dialogues, such as LLaVA-Instruction-150K [24], in the following aspects.

- CB-300K focuses on recognizing instance-level information, reflecting in a large amount of visual referring and grounding requests. With a bounding box available for each instance, CB-300K offers the ability to localize and describe instances more accurately, which improves the granularity of dialogues.
- CB-300K constructs a large number of threads for MRG, forcing the model to gain the ability to perform visual recognition based on logic chains.
- CB-300K is a versatile dataset, which implies that it can be used for various purposes. A typical example lies in using different subsets to train individual yet complementary abilities (*e.g.*, referring, grounding, multi-round un-

<sup>\*</sup>This is partly caused by the limited ability of GPT-4 in understanding the prompt and generating correct question-and-answer pairs (it is a probabilistic LLM that can make errors). We expect that more logic-chain data can be generated in the future with stronger LLMs as the AI assistant.

derstanding, *etc.*) and then combining them into a strong MRG system.

### 3.2. Evaluation Metric

Evaluation is an important issue for multimodal dialogues, particularly for this study which involves logic-chain multifaceted abilities including multi-round reasoning, question answering, and visual grounding. We notice that recent studies (*e.g.*, [43]) have applied state-of-the-art LLMs for benchmarking, but it incurs two-fold risks: (1) GPT-4’s speculative sampling strategy and online update brings unstable randomness for evaluation; meanwhile, (2) GPT-4 developed certain preferences during instruction tuning, making its quantitative decisions biased.

Within a single round, the requirement is similar to the task of grounded image captioning [22, 51], where we employ two scores for evaluation. The first term focuses on the language part (*i.e.*, whether the answer is linguistically correct). We refer to the BERT score [49] to compute the similarity between the model’s output and the ground-truth answer (we use the RoBERTa-large model [26]). The second term focuses on the visual grounding part (*i.e.*, whether the detected bounding box is accurate), and the IoU between the detected and ground-truth boxes is naturally taken into consideration.

In summary, if there is no request for grounding (*i.e.*,  $M = 0$ ), the single-round score equals  $\text{BERT}(\mathbf{a}_m, \mathbf{a}_m^*)$ , where  $\text{BERT}(\cdot, \cdot)$  denotes the BERT score function, and  $\mathbf{a}_m$  and  $\mathbf{a}_m^*$  are the output and ground-truth answer texts; otherwise, it is computed by

$$t = \lambda \cdot \text{BERT}(\mathbf{a}_m, \mathbf{a}_m^*) + (1 - \lambda) \cdot \frac{1}{M} \sum_{m=1}^M \text{IoU}(\mathbf{b}_m, \mathbf{b}_m^*), \quad (1)$$

where  $\mathbf{b}_m$  and  $\mathbf{b}_m^*$  are the detected and ground-truth bounding boxes for the  $m$ -th object.  $\lambda$  is a hyper-parameter that balances the linguistic and visual scores, which is set to be 0.3 by default.

For multi-round evaluation, due to the logical relationship between subsequent rounds, if the answer in a former round is incorrect (*e.g.*, having detected an incorrect object), the task in a latter round (*e.g.*, asking about the attribute of the object) is no longer meaningful. To reflect this mechanism, we introduce a set of hyper-parameters named truncation thresholds,  $\{\tau_n\}_{n=1}^N$ , throughout the entire thread, where  $N$  is the number of rounds. For any  $n$ , if  $t_n$  computed by (1) is smaller than  $\tau_n$  (0.3 by default), we immediately terminate the thread and set all scores in the later rounds to be 0. The overall multi-round score is the average of all rounds, *i.e.*,  $T = \frac{1}{N} \sum_{n=1}^N t_n$ .

## 4. The ChatterBox Model

The overall architecture of the ChatterBox model is presented in Figure 3. The input data (image and text) is fed to two branches for visual feature extraction and language-related processing. The text output (answer) is directly produced by the multimodal branch. Moreover, when there is a request for localizing visual objects and/or regions, a separate embedding designed for querying is integrated with the visual features and fed to a standalone visual grounding module. We will now describe each module in detail. We denote the input image as  $\mathbf{x}_{\text{img}}$  and the input text as  $\mathbf{x}_{\text{txt}}$ .

### 4.1. Individual Modules

**Visual feature extraction.** We resize  $\mathbf{x}_{\text{img}}$  into  $512 \times 512$  and feed it into an iTPN-B model [38] that is pre-trained on Object365 [34], which takes HiViT [50] as backbone. The output is a set of features with resolutions of  $128 \times 128$ ,  $64 \times 64$ ,  $32 \times 32$ , and  $16 \times 16$ , respectively, denoted as  $\{\mathbf{f}_{\text{img}}\}$ . As we shall see later,  $\{\mathbf{f}_{\text{img}}\}$  is only used for visual grounding.

**Multimodal feature extraction.** We feed  $\mathbf{x}_{\text{txt}}$  to the language branch of the CLIP-L/14 model [33], and  $\mathbf{x}_{\text{img}}$  (resized into  $224 \times 224$ ) into the vision branch of the same CLIP-L/14 model. The outputs are a set of language tokens, denoted as  $\mathbf{f}_{\text{txt}}$ , and a set of  $16 \times 16$  vision tokens, denoted as  $\mathbf{f}'_{\text{img}}$ . To process referring, we follow GPT4RoI [48] to insert a special language token [BBOX] as a placeholder. The token embedding is then replaced by the features extracted from the corresponding region, for which the RoIAlign [12, 48] operation is performed on the same CLIP-L/14 model.

**Multimodal understanding.** A multimodal model is trained. It takes  $\mathbf{f}_{\text{txt}}$  and  $\mathbf{f}'_{\text{img}}$  as input and output two-fold embeddings. The first set is simply decoded into the text answer, denoted as  $\mathbf{z}_{\text{ans}}$ . The second set corresponds to the queries of visual grounding, denoted as  $\mathbf{q}_{\text{gnd}}$ , which is only produced when the multimodal model detects a request for localization in the question. The multimodal model is inherited from LLaVA [24], and we apply the LoRA algorithm [13] for fine-tuning.

**Visual grounding.** We use  $\mathbf{q}_{\text{gnd}}$  to query the multi-scale feature set  $\{\mathbf{f}_{\text{img}}\}$  for visual grounding. The module follows an enhanced DETR [3] object detector named DINO [46]. Differently, to facilitate communication between them, we design a two-stage querying mechanism. In the first stage, we perform cross-attention between  $\mathbf{q}_{\text{gnd}}$  and  $\{\mathbf{f}_{\text{img}}\}$  to generate some mixed tokens and propagate them through a few self-attention layers (*a.k.a.* the encoder) followed by a query selection module. In the second stage,  $\mathbf{q}_{\text{gnd}}$  is expanded in dimension and directly added to the queries generated in the first stage (both the label queries and box queries are generated by the DINO encoder) and the obtained queries are then propagated through a few attention

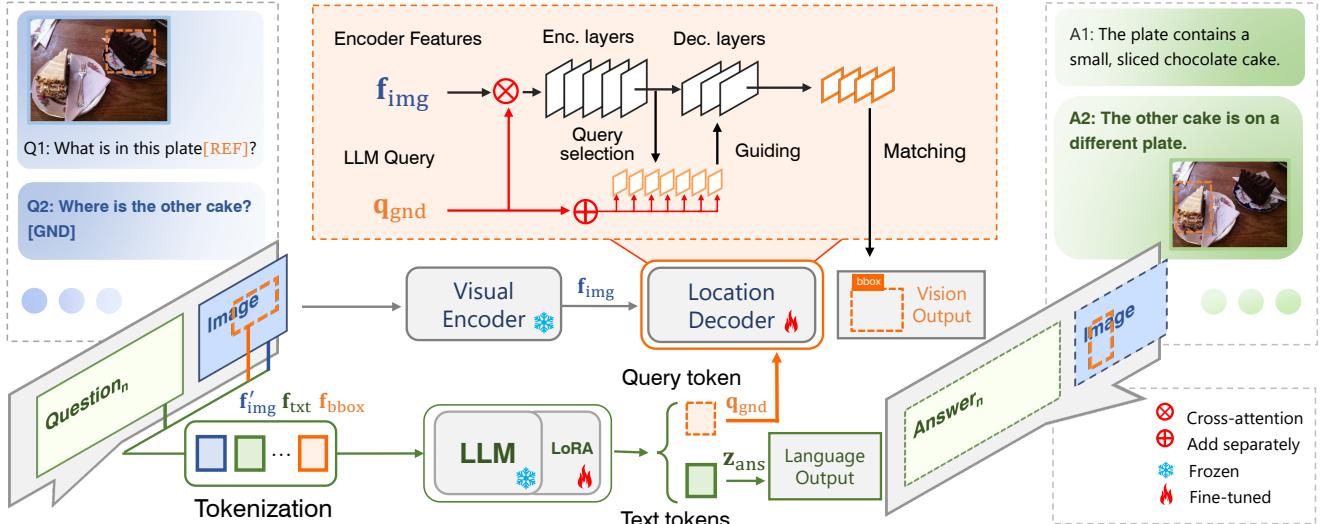


Figure 3. The architecture of the ChatterBox model. It receives the image and the current question with dialogue history as input, and produces language output and, if necessary, vision output (*i.e.*, visual grounding results). The location decoder is magnified to illustrate the interaction between the query token and visual features. *This figure is best viewed in color.*

layers (*a.k.a.* the decoder) to produce the set of box proposals and eventually the bounding boxes,  $\mathcal{B}_{\text{gnd}}$ . Please refer to the supplementary material for details.

## 4.2. Data Pre-processing and Organization

We organize the CB-300K dataset with some external multimodal dialogue data into the following groups. Full details are provided in the supplementary material.

- **Group A: visual question answering.** It involves the Q&A pair without locations (*i.e.*, bounding boxes) in both the questions and answers. We pre-process CB-MRG and CB-LC by removing all locations from the texts. we combine the two sets (CB-MRG, CB-LC) with LLaVA-Instruction-150K [24]. These data are sampled at a ratio of 3 : 2 : 5.
- **Group B: referring expression.** It involves the Q&A pairs with locations in the question but not in the answer. Besides CB-REF, we pre-process CB-MRG and CB-LC by choosing the ones with locations in the question, replacing the descriptions to the locations by ‘it’ or ‘the/this region’, and removing all locations from the answer. Then, we combine the three sets (CB-MRG, CB-LC, CB-REF, orderly) with public datasets. These data are sampled at a ratio of 2 : 3 : 5 : 5.
- **Group C: visual grounding.** It involves the Q&A pairs with locations in the answer. Besides CB-GND, we filter CB-MRG and CB-LC by choosing the ones with locations in the answer. Then, we combine the three sets (CB-MRG, CB-LC, CB-GND, orderly) with public datasets. These data are sampled at a ratio of 3 : 1 : 2 : 5.

## 4.3. Optimization

There are two sources of supervision. For the text output, we compute the auto-regressive cross-entropy loss between  $\mathbf{z}_{\text{txt}}$  and the ground-truth answer, denoted by  $\mathcal{L}_{\text{txt}}$ . For the grounding output (if present), we compute the localization loss (the same as in DINO) between  $\mathcal{B}_{\text{gnd}}$  and the ground-truth set of bounding boxes, denoted as  $\mathcal{L}_{\text{gnd}}$ . The overall loss is then written as

$$\mathcal{L}_{\text{overall}} = \lambda_{\text{txt}} \cdot \mathcal{L}_{\text{txt}} + \lambda_{\text{gnd}} \cdot \mathcal{L}_{\text{gnd}}, \quad (2)$$

where  $\lambda_{\text{txt}}$  and  $\lambda_{\text{gnd}}$  are coefficients and both of them are set to 1.0 by default.

In practice, we find that the visual grounding module is a bit difficult to optimize, so we warm up the training procedure by only using the data in Group C in the early stage. After the grounding loss  $\mathcal{L}_{\text{gnd}}$  becomes small, we add the data in Groups A and B for training.

## 4.4. Discussions of the Design Principle

The design principle of the ChatterBox model is to reflect the idea of decomposition, *i.e.*, the LLM serves as a logic controller to understand user’s intention, and the visual understanding and recognition ability is offered by external modules including the feature extractor and the visual grounding architecture. This is related to a few prior works (*e.g.*, ViperGPT [36], HuggingGPT [35], Chameleon [44], *etc.*).

Compared to another research line (*e.g.*, Kosmos-2 [31], [5], [43], *etc.*) that trained a universal tokenizer for vision-language understanding, our methodology claims both advantages and disadvantages. On the one hand, Our model

is easily implemented and diagnosed with relatively light computational overhead. On the other hand, extending ChatterBox to a uni-model to support tasks beyond referring and grounding requires heavier engineering efforts.

## 5. Experiments

### 5.1. Implementation Details

**Model architecture.** In the vision branch, we inherit a hierarchical transformer pyramid network (iTPN-B) [38] pre-trained on the Objects365 datasets [34] as the visual encoder, and the DINO detector [46] as the location decoder which incorporates 300 queries by default. DINO itself includes an encoder-decoder architecture with 6 blocks for each part. In the language (multimodal) branch, we use a LLaVA-13B model [24], an MLLM based on LLaMA [39] and tuned on visual instruction corpus. To fuse the visual features with the query token produced by the LLM, we follow SAM [15] to employ a cross-attention operation with a two-way transformer. The individual modules can be replaced by other choices as long as they offer the desired functionality, *e.g.*, vision/language encoding and grounding.

**Training configurations.** Please refer to Sec. 4.2 for the details of data preparation and mixture. We utilize  $8 \times$  NVIDIA A800 GPUs (80GB) for training, making use of DeepSpeed to improve computational efficiency. In the first stage, we employ the AdamW optimizer [27] with a learning rate of 0.00005, zero weight decay, a batch size of 6, and a gradient accumulation step of 5. We integrate the WarmupDecayLR learning rate scheduler initialized with a warm-up iteration count of 50. In the second stage, the learning rate is adjusted to 0.00003, while the other training parameters remain unchanged. The data from Groups A, B, and C are sampled at a ratio of 2 : 1 : 10, which aims to maximally preserve the ability of visual grounding that we have established in the first stage. The two stages take approximately 1.5 and 0.5 days, respectively, and the total training cost is around 15 GPU-days.

### 5.2. Results

**Multi-round Dialogue.** We first evaluate the entire task, MRG, using the metrics defined in Section 3.2. A comparison against prior works is summarized in Table 3. We curate all threads in the test set of CB-LC into three question-and-answer pairs, where each round (except for the first one) is logically related to the previous rounds, thereby the difficulty increases round by round.

In terms of the linguistic output, ChatterBox produces better BERT( $\cdot$ ) scores than GPT4RoI [48], Kosmos-2 [31], and LISA [18], and the advantage becomes more significant in the latter two rounds, implying its stronger ability in dealing with multi-round dialogues. ChatterBox is slightly

inferior to LLaVA [24] in the first round but surpasses it in the latter two rounds.

Regarding the visual output, only Kosmos-2 is compared since LLaVA and GPT4RoI cannot perform visual grounding and LISA is unstable in localization\*. Similarly, ChatterBox achieves the best  $\overline{\text{IoU}}(\cdot, \cdot)$  scores throughout the entire thread, and the advantage is even larger than that of the BERT( $\cdot$ ) scores. This is because the grounding quest, calling for the integration of vision and language, is more challenging. Combining the high quality of linguistic and visual output yields the better MRG scores (*i.e.*,  $\{t_n\}$  and  $T$ ).

We qualitatively compare ChatterBox to Kosmos-2 and LISA (two models equipped with visual grounding) in Figure 4). Thanks to the specifically collected data for MRG and the explicit vision modules, ChatterBox shows a stronger ability to accomplish logically complex quests, while the competitors can run into failure. More examples are provided in the supplementary material.

**Single-round referring expression.** We show that our model trained for MRG also enjoys the expected ability of single-round referring expression. We evaluate it on Ref-COCOg [14] and compare it against GPT4RoI and Kosmos-2. Table 4 summarizes the results in terms of the METEOR, CIDEr and BERT scores computed upon the captions on given regions. ChatterBox shows preferable performance. Evaluation details and some examples are provided in the supplementary material.

**Single-round visual grounding.** Similarly, ChatterBox can be used for single-round visual grounding. We compare it against Kosmos-2 on the COCO [23] 2017 test set. Table 5 summarizes the box-level IoU, success rate (IoU is at least 0.5), and mean IoU of successful cases. Since the MLLMs are sensitive to the prompt, we examine three types of prompts, including (1) ‘Where is the [name]?’, (2) ‘Can you find the [name]?’, and (3) ‘Can you tell the position of the [name]?’ , with [name] replaced by the name of object. We report the result of the best prompt for ChatterBox and Kosmos-2. As shown, ChatterBox surpasses Kosmos-2 in all metrics. Additionally, ChatterBox also shows stronger robustness, as the lowest success rate over three prompts is about 0.6, while the number is around 0.2 for Kosmos-2. These results are impressive considering that the grounding data is  $180 \times$  fewer (500K vs. 90M). We owe the ability to the explicit vision module. Some examples are provided in the supplementary material.

**Diagnostic studies.** The first part involves not using the CB-300K data for training. Comparing the first two rows of Table 6, we find that the collected data consistently improves model’s ability of MRG; similarly, the gain is

---

\*LISA’s lack of support for explicit instructions (*e.g.*, the [GND] token) makes it unable to produce stable localization results. Meanwhile, LISA produces a segmentation mask which sometimes contains outliers that may deteriorate the box-level IoU.

Table 3. A quantitative comparison of the MRG metrics (see Section 3.2) between ChatterBox (our work) and prior works.

Method	Round #1			Round #2			Round #3			$T$
	BERT( $\cdot$ )	$\overline{\text{IoU}}(\cdot, \cdot)$	$t$	BERT( $\cdot$ )	$\overline{\text{IoU}}(\cdot, \cdot)$	$t$	BERT( $\cdot$ )	$\overline{\text{IoU}}(\cdot, \cdot)$	$t$	
LLaVA [24]	<b>0.9353</b>	—	—	0.9122	—	—	0.9002	—	—	—
GPT4RoI [48]	0.9157	—	—	0.8818	—	—	0.8673	—	—	—
Kosmos-2 [31]	0.9023	0.282	0.468	0.8871	0.244	0.437	0.8712	0.137	0.357	0.421
LISA [18]	0.9171	—	—	0.8822	—	—	0.8708	—	—	—
ChatterBox (ours)	0.9303	<b>0.401</b>	<b>0.560</b>	<b>0.9184</b>	<b>0.377</b>	<b>0.539</b>	<b>0.9082</b>	<b>0.306</b>	<b>0.487</b>	<b>0.529</b>

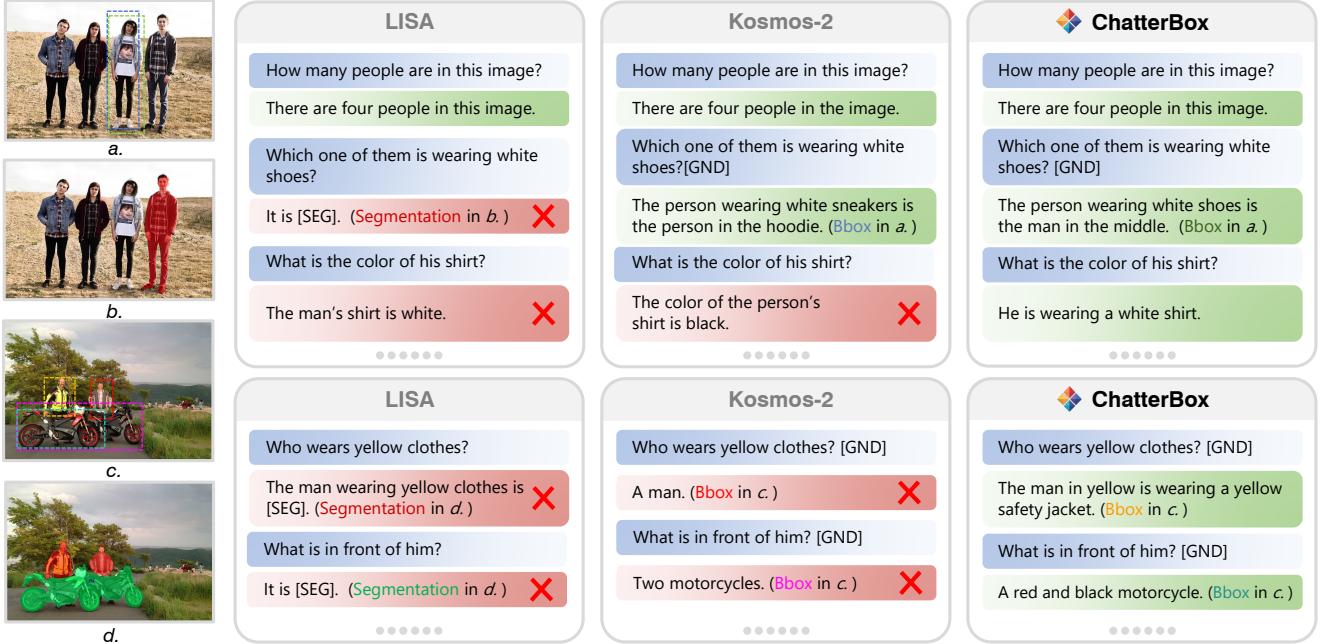


Figure 4. A qualitative comparison of multi-round dialogue between LISA [18], Kosmos-2 [31], and ChatterBox (ours). Our model demonstrates a superior ability to understand multi-round dialogues and perform reasoning (please also refer to the examples in Figure 1). We stress that the stronger ability of visual recognition is brought by the explicit vision modules.

Table 4. A quantitative comparison of single-round referring expression on the RefCOCOg dataset.

Method	METEOR	BERT
GPT4RoI [48]	9.7	0.873
Kosmos-2 [31]	11.5	0.871
ChatterBox	<b>14.5</b>	<b>0.880</b>

larger in the second and third rounds. We will release the CB-300K data to facilitate the research in this direction. The second part involves not replacing the concrete object names with pronouns (*e.g.*, ‘it’ or ‘the object’) in the second and third rounds, which degenerates MRG into single-round dialogues because the understanding does not rely on the former rounds. Not surprisingly, the model reports similar scores in all three rounds. This indicates that MRG indeed

Table 5. A quantitative comparison of single-round visual grounding on the COCO [23] 2017 test set. Please refer to the main text for the details of prompts and metrics.

Method	mIoU	Succ. Rate	mIoU @ Succ.
Kosmos-2 [31]	0.627	0.688	0.854
ChatterBox	<b>0.710</b>	<b>0.762</b>	<b>0.904</b>

increases the difficulty of dialogues so that we believe it is a promising direction for MLLMs.

## 6. Conclusions

We establish a baseline for multi-round multimodal referring and grounding (MRG), opening up a promising direction for instance-level multimodal dialogues. In specific, we

Table 6. Diagnostic results in terms of the BERT score and the  $T$  score. **CB-300K**: whether CB-300K is used for training. **Ref. Words**: whether pronouns (*e.g.*, ‘it’ or ‘the object’, instead of concrete object names) are used in the inference stage. Note: the third row is **not** a fair comparison because it is easier than MRG.

CB-300K	Ref. Words	Round #1	Round #2	Round #3	$T$
✗	✓	0.9291	0.9042	0.8946	0.478
✓	✓	0.9303	0.9184	0.9082	0.529
✓	✗	0.9303	0.9237	0.9210	0.547

present a new benchmark and an efficient vision-language model. The new benchmark, CB-300K, spans challenges including multi-round dialogue, complex spatial relationships among multiple instances, and consistent reasoning, which are beyond those shown in existing benchmarks. The proposed model, ChatterBox, with well-defined feature extraction and optimization strategies, is validated to be very effective in performing multi-round referring and grounding. With the flexibility to complex instance relationships, the robustness to multiple instances, and the plug-and-play architecture, ChatterBox has the potential to significantly advance the multimodal dialogue tasks that involve complicated and precise interactions.

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. [2](#), [3](#)
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. [2](#)
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. [5](#)
- [4] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models. *arXiv preprint arXiv:2308.13437*, 2023. [3](#)
- [5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing Multi-modal LLM’s Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*, 2023. [3](#), [6](#)
- [6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. *See <https://vicuna.lmsys.org> (accessed 14 April 2023)*, 2023. [2](#)
- [7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. [2](#)
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. [2](#)
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. [2](#)
- [11] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. [3](#)
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. [5](#)
- [13] 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. [5](#)
- [14] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. [2](#), [7](#), [13](#)
- [15] 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. *arXiv preprint arXiv:2304.02643*, 2023. [7](#)
- [16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *arXiv preprint arXiv:1602.07332*, 2016. [11](#), [13](#)
- [17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. [2](#), [4](#)
- [18] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model. *arXiv preprint arXiv:2308.00692*, 2023. [2](#), [3](#), [7](#), [8](#)

- [19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2
- [20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3
- [21] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded Language-Image Pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10955–10965, 2022. 3
- [22] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 5
- [23] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7, 8, 13
- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 3, 4, 5, 6, 7, 8, 13
- [25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3
- [26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5
- [27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7
- [28] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 13
- [29] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. 2
- [30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 3
- [31] 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. 2, 3, 6, 7, 8, 13, 17
- [32] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 13
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5
- [34] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 5, 7
- [35] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Huggingpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 6
- [36] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergrpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023. 6
- [37] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. 2
- [38] Yunjie Tian, Lingxi Xie, Zhaozhi Wang, Longhui Wei, Xiaopeng Zhang, Jianbin Jiao, Yaowei Wang, Qi Tian, and Qixiang Ye. Integrally Pre-Trained Transformer Pyramid Networks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18610–18620. IEEE, 2023. 5, 7
- [39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 7
- [40] Wenhui Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 2, 3
- [41] 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. 2, 3
- [42] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the Power of Referential Comprehension for Multi-modal LLMs. *arXiv preprint arXiv:2310.00582*, 2023. 3
- [43] 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. *arXiv preprint arXiv:2310.07704*, 2023. 3, 5, 6

- [44] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023. 6
- [45] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 2
- [46] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 5, 7
- [47] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2
- [48] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. *arXiv preprint arXiv:2307.03601*, 2023. 2, 3, 5, 7, 8
- [49] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 5
- [50] Xiaosong Zhang, Yunjie Tian, Lingxi Xie, Wei Huang, Qi Dai, Qixiang Ye, and Qi Tian. Hivit: A simpler and more efficient design of hierarchical vision transformer. In *The Eleventh International Conference on Learning Representations*, 2022. 5
- [51] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4777–4786, 2020. 5
- [52] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

## A. CB-300K Generation Details

In this section, we provide a detailed description of the generation procedure of CB-MRG and CB-LC, which are the main parts of CB-300K. We start with introducing the data cleaning procedure.

### A.1. Data Cleaning

The data cleaning process is mainly implemented on the Visual Genome [16] dataset. We primarily encounter two potential situations that may lead to errors. **Firstly**, some instances in VG have multiple corresponding bounding boxes, making GPT-4 mistakenly interpret them as distinct instances. To mitigate this, we apply NMS to instances with the same name. Additionally, during the NMS process, we

assign unique identifiers to boxes that exceed the threshold to avoid ambiguity. **Secondly**, there are cases in the dataset where a single object has multiple names due to different annotators for the same image. This situation can also lead to errors in the generated dialogues. To address this, we apply NMS to instances with the same name across the entire dataset, discarding instances that exceed a pre-defined threshold.

### A.2. Prompt for CB-MRG

The prompt for CB-MRG aims to transform object relationships (*e.g.*, a cat [x11, x12, y11, y12] on a table [x21, x22, y21, y22]) into dialogues. The main structure of the prompt includes:

- **Input format explanation.** Each object in the provided description has corresponding coordinates. For example: ‘black and nice motorcycle [60, 77, 462, 307] on street [4, 65, 496, 329]’
- **Task definition.** GPT-4 shall propose questions and answer pairs based on the input relationship information. In a dialogue, each following question should be based on the previous answer. Questions can be about object relationships, object actions, object attributes, object status, object types, etc.
- **Question constraints:** Questions can only be proposed when their answer can be verified as correct with the given description.
- **Indexing objects:** While describing multiple relationship details of an image, objects of the same type are differentiated by adding indices in a ‘name\_number’ format.
- **Output format requirements:** Coordinates of objects mentioned in the sentences shall be appended after each question and answer, in the format <the name of object1: [x1, y1, x2, y2], the name of object2: [x3, y3, x4, y4]>. For example:

*Question-1: What is the color of the shirt of the man?*  
*<man: [0.1, 0.1, 0.3, 0.5], shirt: [0.1, 0.2, 0.3, 0.4]>*

*Answer-1: The color is red.*

*<shirt: [0.1, 0.2, 0.3, 0.4]>*

*Question-2: Where is the man sitting?*

*<man: [0.1, 0.1, 0.3, 0.5]>*

*Answer-2: The man is sitting on a chair.*

*<chair: [0.1, 0.4, 0.3, 0.6]>*

The generated Q&A pairs are not yet ready to be used. On the one hand, the question sometimes gives too strong hints for the answer or still includes requests that are beyond the given information. On the other hand, the answers might lack the corresponding object’s location or provide redundant coordinates. These flaws can be harmful to model training and evaluation, so additional instructions are introduced for refinement:

- If the question contains information that is part of the answer, remove the redundant description. For example, re-

place ‘Where is the white car parking on the street?’ with ‘Where is the white car?’

- Only the coordinates of objects explicitly mentioned in the answer, instead of all objects mentioned in the response sentence, shall be provided. For example, Answer: No, the white cup is on the left of the man. <cup: [25, 100, 51, 124]>
- If multiple objects are part of the answer, list all these objects. Note that certain objects and the answer may have indirect relationships.
- If the answer is not related to the image, there is no need to provide coordinates.
- Ensure consistency between relationship information and answer statements. For example, given relationship information: ‘tan dirt [0, 86, 500, 498] on feet [189, 457, 304, 492],’ if the question is ‘What is the object on the tan dirt?’ the answer should be ‘I don’t know about it since the information is not given,’ not ‘The object on the tan dirt is feet.’
- Do not make any location inferences based on coordinate information.

**The full text of the prompt is very long and is provided in prompt\_CB\_MRG.txt.**

### A.3. Prompt for CB-LC

Firstly, we prompt GPT-4 (see prompt\_CB\_LC.txt) to construct dialogues. This prompt is similar to the one used for CB-MRG but exhibits some differences:

- Relationship chains are defined as ‘[[object1 relationship1 object2], [object2 relationship2 object]], etc. (e.g., ‘[papers below full shelf], [full shelf has white books]’, etc.).
- Each dialogue is based on one relationship chain.
- Questions must follow the sequence of objects in the relationship chain.
- Each question must include the subject (object from the previous answer).
- Here are some examples for GPT-4:

‘relationships: [[a man [0.1,0.3,0.4,0.7] holding a cup [0.1,0.4,0.15,0.45]], [a cup [0.1,0.4,0.15,0.45] has water [0.1,0.41,0.15,0.45]]];

*Question-1: What is the man holding?*

<man: [0.1,0.3,0.4,0.7]>

*Answer-1: The man is holding a cup.*

<cup: [0.1,0.4,0.15,0.45]>

*Question-2: What does the cup have?*

<cup: [0.1,0.4,0.15,0.45]>

*Answer-2: There is water in the cup.*

<water: [0.1,0.41,0.15,0.45]>,

- There are some excluded dialogue examples for GPT-4: ‘relationships: [[a man [0.1,0.3,0.4,0.7] holding a cup [0.1,0.4,0.15,0.45]]];

*Question-1: Where is the cup?*

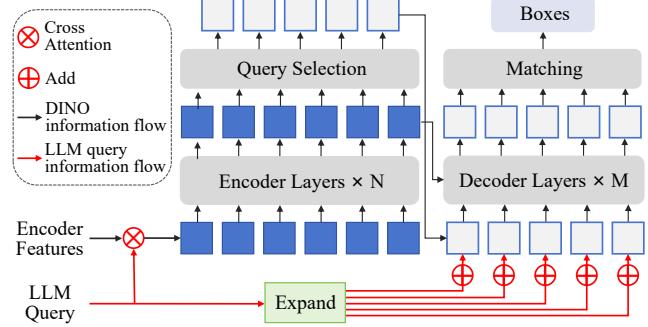


Figure 5. The architectural visualization of target-guided query.

<cup: [0.1,0.4,0.15,0.45]>  
*Answer-1: The cup is held by the man.*  
<man: [0.1,0.3,0.4,0.7]>”

**Secondly**, we conduct some manual cleaning operations (see code at cb\_lc\_cleaning.py) to filter out the dialogues that violate prompt requirements, for example:

- Objects appear in questions or subjects appear in answers.
- Objects in the current question do not appear in the previous answer.
- Generated dialogues without corresponding object coordinates.

**Finally**, we use GPT-4 to clean up the dialogues, deleting responses that do not meet the requirements of the questions.

- GPT-4 shall determine(see cb\_lc\_cleaning\_1.txt) whether the answer addresses the question and delete dialogues where any round fails to meet the prompt requirements.
- GPT-4 shall determine(see cb\_lc\_cleaning\_2.txt) whether the dialogue is contradictory to the given description and delete dialogues where any round contradicts the relationship information (mainly checking for position errors).

Please note that these steps aim to generate logically consistent multi-round dialogues based on the given relationship information while ensuring compliance with prompt requirements. Each step involves manual and automated cleaning to refine the generated dialogues.

## B. Experimental Details

In this section, we provide additional clarification on the LLM query on the DINO detector and the used public datasets.

### B.1. Target-guided Query

To facilitate communication between LLM query and vision features, we design a two-stage querying mechanism (named target-guided query, or the ‘tgt’ query for short). The visualization of the target-guided query is shown in

Figure 5. To transfer the LLM query to the visual detector, we propose the target-guided query to enable the ChatterBox to control the visual detector to predict boxes. The target-guided query contains a two-stage querying mechanism. First, we use a cross-attention operation between the query token of LLM and the visual features of the visual encoder, the new visual features are formed with LLM information for subsequent modules. Second, the query token of the LLM is first expanded into two queries with dimensions of  $(bs \times N_{\text{queries}} \times D)$  (where  $N_{\text{queries}}$  denotes the query number of DINO detector, and  $D$  is the query dimension), one directly added onto the label queries of the DINO detector, and the other onto the box queries.

## B.2. Datasets

We organize the CB-300K dataset with some modified public image-text datasets into the following groups.

- **Group A: visual question answering.** It involves the Q&A pair without locations (*i.e.*, bounding boxes) in both the question-and-answer. We pre-process CB-MRG and CB-LC by removing all locations from the texts. We combine the two sets (CB-MRG, CB-LC) with LLaVA-Instruction-150K [24]. These data are sampled at a ratio of 3 : 2 : 5.
- **Group B: referring expression.** It involves the Q&A pairs with locations in the question but not in the answer. Besides CB-REF, we pre-process CB-MRG and CB-LC by choosing the ones with locations in the question, replacing the descriptions to the locations by ‘it’ or ‘the/this region’, and removing all locations from the answer. Then, we combine the three sets (CB-MRG, CB-LC, CB-REF, orderly) with external datasets including COCO [23], RefCOCO [14], RefCOCO+ [14], RefCOCOg [28], Flickr30K [32], and Visual Genome [16]. These data are sampled at a ratio of 2 : 3 : 5 : 1 : 1 : 1 : 1.
- **Group C: visual grounding.** It involves the Q&A pairs with locations in the answer (the locations may or may not appear in the question). Besides CB-GND, we filter CB-MRG and CB-LC by choosing the ones with locations in the answer. Then, we combine the three sets (CB-MRG, CB-LC, CB-GND, orderly) with COCO [23], RefCOCO [14], RefCOCO+ [14], and RefCOCOg [28]. These data are sampled at a ratio of 3 : 1 : 2 : 2 : 1 : 1 : 1.

## B.3. Evaluation

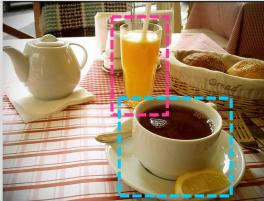
As current MLLMs’ answers may contain extra or unrelated parts that, although do no harm to human understanding, interference the evaluation program’s process. We conducted a text cleaning work to make MLLMs’ answers easier to be evaluated. Words like ‘it is’, ‘there is’, ‘region0’ that do not appear in the evaluation program’s ground truth are removed by our cleaning scripts (see `cb_eval_clean.py`).

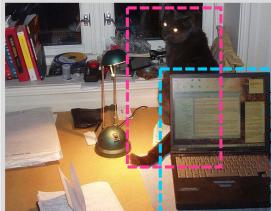
## C. Visualization Examples

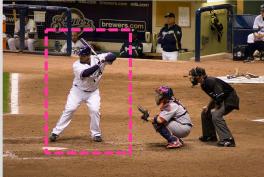
In this section, we provide more visualization examples. These examples are generated by the ChatterBox agent. We test special ability in terms of referring expression (Figure 6), visual grounding (Figure 7), multi-round dialogues (Figure 8), and conduct a comparison with Kosmos-2 [31] (Figure 9).

**ChatterBox**

**a.**  What is next to it (REF in a.) ?  
There is a donut on a plate.  
.....

**b.**  What is it (REF in b.) ?  
It is a glass of orange juice.  
What is that (REF in b.) ?  
There is a mug of iced tea.  
.....

**c.**  What is in this region (REF in c.) ?  
There is a black cat in this region.  
What is in that region (REF in c.) ?  
There is a black laptop in the region.  
What is the relation between them?  
The black cat is standing in front of the black laptop.  
.....

**d.**  Describe this region(REF in d.) ?  
There is a man in a white baseball uniform, holding a baseball bat. He is the batter in the game.  
.....

**e.**  What is the man (REF in e.) wearing on his head?  
The man is wearing a yellow and green hat.  
What does it look like?  
The man is wearing a yellow and green hat that is shaped like a sunflower.  
.....

Figure 6. Examples of referring expression. *This figure is best viewed in color.*

**ChatterBox**

Where is the earring? [GND]

The earring is on the woman's ear. It is a dangling earring. (Bbox in a.)

Where is the person holding a white and black cow? [GND]

The person holding a white and black cow is on a yard. (Bbox in b.)

What can I use to input text? [GND]

You can use the on-board keyboard of the laptop. (Bbox in c.)

What is the catcher using to hold the ball? [GND]

The catcher is using a baseball glove to hold the ball. (Bbox in d.)

Find the unusual thing in the picture and tell me why. [GND]

The unusual thing in the picture is a pizza sitting on a roof, as it is typically served indoors or outdoors in a more conventional setting, not on a rooftop. (Bbox in e.)

Where is the dog? [GND]

The dog is in a backyard. (Bbox in f.)

What is in its mouth? [GND]

The dog is holding a frisbee in its mouth. (Bbox in f.)

**a.**

**b.**

**c.**

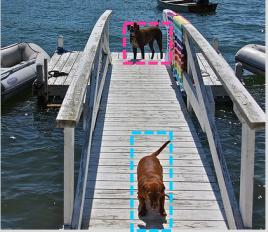
**d.**

**e.**

**f.**

Figure 7. Examples of visual grounding. *This figure is best viewed in color.*

**ChatterBox**



a.

What is it? (REF in a.)

It is a dog.

Where is the other dog? [GND]

It is on the other side of the wooden walkway. (Bbox in a.)

.....

What is in this region(REF in b.) ?

There is a cat in this region.

What is it on? [GND]

The cat is on a laptop. (Bbox in b.)

.....

What is in this region (REF in c.) ?

There is a sandwich in this region.

What is next to it? [GND]

There is a glass of red wine next to the sandwich. (Bbox in c.)

.....

How many people are in this region (REF in d.) ?

There is a man in this region.

What is the color of his jeans?

The man is wearing blue jeans.

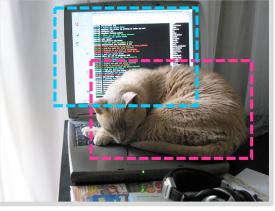
How many people are in this region (REF in d.) ?

There are two people in this region.

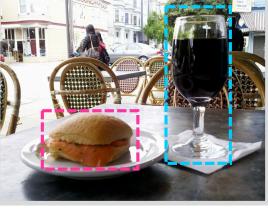
which one of them is in red? [GND]

The woman in the red is in this region. (Bbox in d.)

.....



b.



c.



d.

Figure 8. Examples of multi-round dialogues. *This figure is best viewed in color.*

## Kosmos-2

What can I use to input text? [GND]

A mouse. (Bbox in a.)

.....

Where is the sauce bottle? [GND]

A restaurant in the mountains. The man is enjoying a meal with a pizza and wine. (Bbox in b.)

.....

What is in this region (REF in c.)?

A window.

What is in that region (REF in c.)?

A laptop.

What is the relation between them?

The cat is sitting on the laptop.

.....

Where is the person holding a white and black cow? [GND]

The person holding the white cow is on a yard. (Bbox in d.)

.....

## ChatterBox

What can I use to input text? [GND]

You can use the on-board keyboard of the laptop. (Bbox in a.)

.....

Where is the sauce bottle? [GND]

The sauce bottle is on the table. (Bbox in b.)

.....

What is in this region (REF in c.)?

There is a black cat in this region.

What is in that region (REF in c.)?

There is a black laptop in the region.

What is the relation between them?

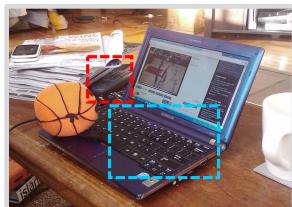
The black cat is standing in front of the black laptop.

.....

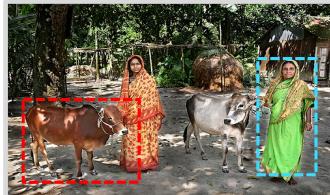
Where is the person holding a white and black cow? [GND]

The person holding a white and black cow is on a yard. (Bbox in d.)

.....



a. b.



c. d.

Figure 9. The comparison examples with Kosmos-2 [31]. This figure is best viewed in color.