ChatterBox: Multimodal Referring and Grounding with Chain-of-Questions

Yunjie Tian[1]*, **Tianren Ma**[1]*, **Lingxi Xie**[2], **Qixiang Ye**[1]†

[1]University of Chinese Academy of Sciences
[2]Huawei Inc.
{tianyunjie19, matianren18}@mails.ucas.ac.cn, 198808xc@gmail.com, qxye@ucas.ac.cn

Abstract

In this study, we establish a benchmark and a baseline approach for **M**ultimodal referring and grounding with **C**hain-of-**Q**uestions (**MCQ**), opening up a promising direction for 'logical' multimodal dialogues. The newly collected dataset, named CB-300K, spans challenges including probing dialogues with spatial relationship among multiple objects, consistent reasoning, and complex question chains. The baseline approach, termed **ChatterBox**, involves a modularized design and a referent feedback mechanism to ensure logical coherence in continuous referring and grounding tasks. This design reduces the risk of referential confusion, simplifies the training process, and presents validity in retaining the language model's generation ability. Experiments show that ChatterBox demonstrates superiority in MCQ both quantitatively and qualitatively, paving a new path towards multimodal dialogue scenarios with logical interactions.

Code — https://github.com/sunsmarterjie/ChatterBox

Introduction

Large language models (LLMs) have shown impressive capabilities across a wide range of natural language tasks (Brown et al. 2020). In the computer vision community, researchers have integrated LLMs with images and videos to create Multimodal Large Language Models (MLLMs), enabling them to understand and handle visual information (Alayrac et al. 2022; Li et al. 2023b; Liu et al. 2023a; Li et al. 2022a). Multimodal dialogue with referring and grounding tasks, which explores MLLM's visual understanding and interaction ability (Peng et al. 2023; Chen et al. 2023b; You et al. 2023), has gained widespread attention.

Despite the progress achieved, understanding referential objects in continuous dialogues remains challenging. In pure natural language dialogues, determining nominal substitution (*e.g.*, 'it', 'that', 'all', 'the former', *etc.*) may be ordinary, as the contextual information is usually provided explicitly. However, in a multimodal dialogue, visual information is implicitly encoded, and relationship among visual objects might be overlooked or confused. When asked to generate answers based on visual reasoning, MLLMs may struggle to associate words with its visual entity, leading to difficulties with referential issues. To address this problem, we introduce **M**ultimodal referring and grounding with **C**hain-of-**Q**uestions (MCQ), opening up a promising direction for logically continuous multimodal dialogues. MCQ encompasses a sequence of logically related questions, where the answer to each question is derived from the exact understanding of the foregoing information. Specifically, we demonstrate an example in Figure 1, where each question is intricately linked to the previous one, implying that a referential confusion could result in a series of subsequent inaccuracies.

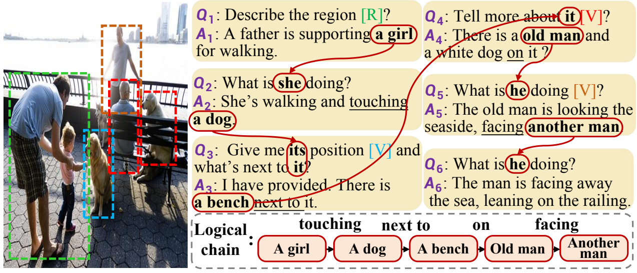

Q_1: Describe the region [R]?
A_1: A father is supporting **a girl** for walking.
Q_2: What is **she** doing?
A_2: She's walking and touching **a dog**.
Q_3: Give me **its** position [V] and what's next to **it**?
A_3: I have provided. There is **a bench** next to it.
Q_4: Tell more about **it** [V]?
A_4: There is a **old man** and a white dog on it ?
Q_5: What is **he** doing [V]?
A_5: The old man is looking the seaside, facing **another man**.
Q_6: What is **he** doing?
A_6: The man is facing away the sea, leaning on the railing.

Figure 1: An example of MCQ. Colored [R] and [V] are trigger words that tell the model to accept referring or give grounding information. From Q_1 to Q_6, these questions constitute a logical chain. A referential confusion can lead to subsequent inaccuracies, which brings challenges to MLLMs.

Our contributions to advancing MCQ are two-fold. **(1)** We establish a new benchmark named CB-300K, which comprises the first-ever instance-level image-text dataset for MCQ and an evaluation metric that considers the accuracy of both visual and linguistic understanding in chain-of-questions. **(2)** We set up a baseline named ChatterBox to solve the challenging task. The main difference from the existing MLLMs (Peng et al. 2023; Chen et al. 2023b; You et al. 2023) lies in the modularized design and the referent feedback mechanism. The modularized design introduces independent vision modules (for referring and grounding tasks) and connects them with the LLM by feeding/producing a specified token for each task. The referent feedback mechanism embeds the preceding visual cues into the sub-

*Equal contribution.
†Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

sequent question to assist reference. While answering questions, this strategy avoids ambiguity in language descriptions and increases the probability of finding the correct objects.

We conduct both quantitative and qualitative studies, affirming ChatterBox's superiority over existing models in MCQ. ChatterBox also transfers to easier tasks (*e.g.*, single-round dialogues, 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 summarize the contributions of our work as follows:

- We introduce a new task named Multimodal referring and grounding with Chain-of-Questions (MCQ), aimed at fostering more natural and seamless interactions with multimodal dialogue systems.
- We propose a data construction scheme and establish the CB-300K benchmark to facilitate the research in MCQ. Besides chain-of-questions, CB-300K also involves instance-level multi-round dialogues with complex spatial relationships and consistent reasoning.
- We present ChatterBox with a modularized design that minimizes the impact on the text generation ability of the language model and a referent feedback mechanism that mitigates the referring confusion. Both techniques contribute to the better performance of MCQ.

Related Work

Multimodal Large Language Model. Large language models (Devlin et al. 2018; Touvron et al. 2023; Brown et al. 2020; Chiang et al. 2023; Chung et al. 2022; Zeng et al. 2022; Thoppilan et al. 2022; Chowdhery et al. 2022; Zhang et al. 2022b) 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 (Li et al. 2022a, 2023b; Alayrac et al. 2022; Liu et al. 2023a). Pioneering efforts have focused on aligning vision and language data into the same feature space (Radford et al. 2021; Alayrac et al. 2022) and adapting an LLM to visual tasks have been made internally or externally, with cross-attention (Alayrac et al. 2022) or Q-former (Li et al. 2023b) modules.

Multimodal Instruction Data. Inspired by the instruction tuning mechanism (Ouyang et al. 2022) of the GPT series, MLLMs started collecting instruction data from various sources. One of the early efforts was visual instruction tuning (Liu et al. 2023a) which provides a novel method for data construction by feeding external metadata into GPT-4 to generate detailed conversations. The idea was followed by other works (Chen et al. 2023a,b) to harvest various types of instruction data. In another approach, the image feature was fed to an MLLM and prompted for instruction data (Zhu et al. 2023). Moreover, richer information, such as phrase grounding, has been collected (Peng et al. 2023) with the assistance of external vision-language models, such as GLIP (Li et al. 2022b). The new data and learning strategy enabled more abilities to emerge via multimodal dialogue (Liu et al. 2023a; Gong et al. 2023; Alayrac et al. 2022; Wang et al. 2023b; You et al. 2023; Lai et al. 2023).

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

Table 1: The number of threads and the number of question-and-answer pairs of each individual subset and the entire benchmark.

MLLMs with Instance-Level Understanding. MLLMs can be largely enhanced with 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 (Zhang et al. 2023; Qiu et al. 2024; Chen et al. 2023a; Ma et al. 2024) and visual grounding (Peng et al. 2023; Liu et al. 2023b), respectively. There are two main approaches to integrate them together, differing in whether to encode the position information explicitly or not. Explicit methods (Peng et al. 2023; Wang et al. 2023a) are easier to optimize and explain by introducing location tokens, while implicit methods (Chen et al. 2023b; Wang et al. 2023b; Xuan et al. 2023) offer greater flexibility.

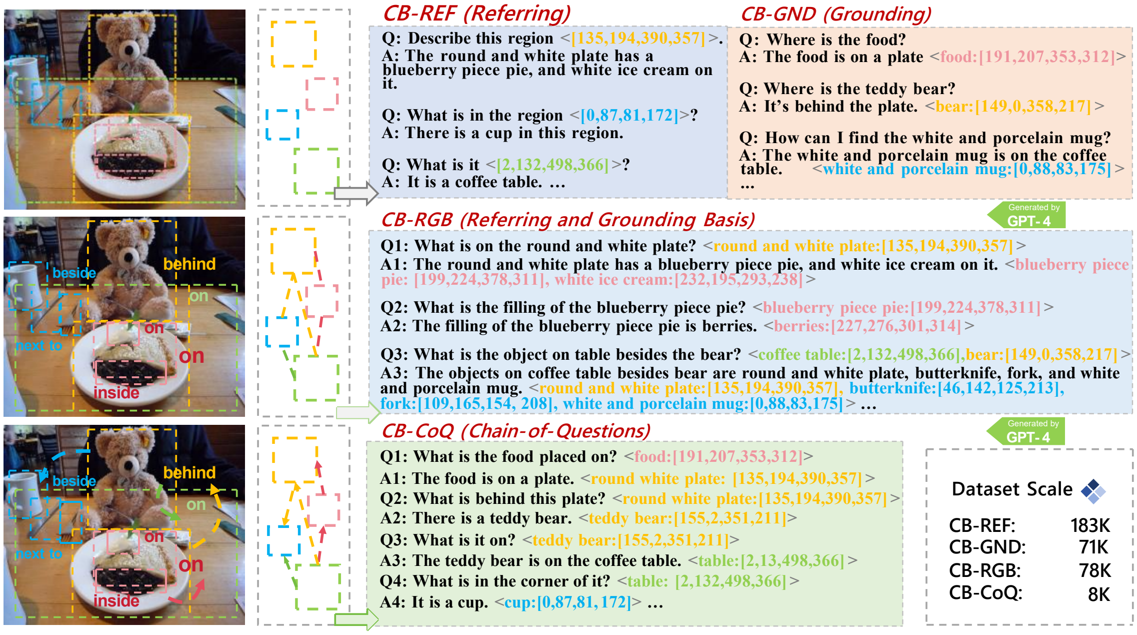
CB-300K: A New Benchmark with Chain-of-Questions

We establish a benchmark named ChatterBox-300K (CB-300K for short) with the aim to advance multimodal dialogue systems. Different from previous work's dataset curation (Liu et al. 2023a; Peng et al. 2023; You et al. 2023), CB-300K incorporates multimodal chain-of-questions about complex spatial relationships and consistent reasoning among multiple instances. To construct CB-300K, we leverage the Visual Genome dataset (Krishna et al. 2017) due to its richness of instance-level relationship annotations, and get assistance from GPT-4. A new metric is also proposed to evaluate MLLMs in this new setting.

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 this information into pure text contexts, which are then fed to GPT-4. We instruct GPT-4 to generate Q&A pairs of different aspects. An additional request for GPT-4 is that, in each sequence of consecutive Q&A pairs, the latter questions should be built on the former ones to construct the generic multi-round dialogues.

In summary, there are four subsets in CB-300K including a generic subset, a chain-of-question subset, and two specif-



CB-REF (Referring)

Q: Describe this region <[135,194,390,357]>.
A: The round and white plate has a blueberry piece pie, and white ice cream on it.

Q: What is in the region <[0,87,81,172]>?
A: There is a cup in this region.

Q: What is it <[2,132,498,366]>?
A: It is a coffee table. …

CB-GND (Grounding)

Q: Where is the food?
A: The food is on a plate <food:[191,207,353,312]>

Q: Where is the teddy bear?
A: It's behind the plate. <bear:[149,0,358,217]>

Q: How can I find the white and porcelain mug?
A: The white and porcelain mug is on the coffee table. <white and porcelain mug:[0,88,83,175]>
…

CB-RGB (Referring and Grounding Basis)

Q1: What is on the round and white plate? <round and white plate:[135,194,390,357]>
A1: The round and white plate has a blueberry piece pie, and white ice cream on it. <blueberry piece pie: [199,224,378,311], white ice cream:[232,195,293,238]>

Q2: What is the filling of the blueberry piece pie? <blueberry piece pie:[199,224,378,311]>
A2: The filling of the blueberry piece pie is berries. <berries:[227,276,301,314]>

Q3: What is the object on table besides the bear? <coffee table:[2,132,498,366],bear:[149,0,358,217]>
A3: The objects on coffee table besides bear are round and white plate, butterknife, fork, and white and porcelain mug. <round and white plate:[135,194,390,357], butterknife:[46,142,125,213], fork:[109,165,154, 208], white and porcelain mug:[0,88,83,175]> …

CB-CoQ (Chain-of-Questions)

Q1: What is the food placed on? <food:[191,207,353,312]>
A1: The food is on a plate. <round white plate: [135,194,390,357]>
Q2: What is behind this plate? <round white plate:[135,194,390,357]>
A2: There is a teddy bear. <teddy bear:[155,2,351,211]>
Q3: What is it on? <teddy bear:[155,2,351,211]>
A3: The teddy bear is on the coffee table. <table:[2,13,498,366]>
Q4: What is in the corner of it? <table: [2,132,498,366]>
A4: It is a cup. <cup:[0,87,81, 172]> …

Dataset Scale

CB-REF:	183K
CB-GND:	71K
CB-RGB:	78K
CB-CoQ:	8K

Figure 2: The CB-300K benchmark comprises four subsets. The former two subsets, CB-REF and CB-GND, are produced using manually designed rules and then polished by GPT-3.5. The latter two subsets, CB-RGB and CB-CoQ, are obtained by prompting GPT-4 to read the metadata and generate questions and answers. *This figure is best viewed in color.*

ically designed subsets, as illustrated in Figure 2. Among them, **CB-CoQ** is the core subset of CB-300K. It contains chain-of-question (MCQ) dialogues, and its construction is rather challenging: (1) adding strict restrictions upon the prompt we used to generate CB-RGB (*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.*), and (3) calling GPT-4 again to check the entire thread, cleaning up incorrect descriptions and contradictions. The filtering procedure guarantees CB-CoQ's high quality, but the strict rules and GPT-4's limited ability makes its size relatively small.

Table 1 displays the statistics of CB-300K. We extracted 800 threads in CB-RGB and 200 threads in CB-CoQ for testing, and the remaining threads are used for training. The CB-300K dataset differs from existing multimodal dialogue datasets (LLaVA-Instruction-150K (Liu et al. 2023a), Shikra-RD (Chen et al. 2023b), *etc.*), showcasing advancements in the following aspects. **1)** CB-300K constructs a precious subset for MCQ. By structuring logical coherence between multimodal dialogues, CB-CoQ provides a foundation for MLLMs to conduct thorough visual understanding based on chain-of-questions. **2)** CB-300K focuses on deeply excavating instance-level information. Abundant annotations are provided for visual referring and grounding requests. With a bounding box assigned to each instance in the image, the granularity of multimodal dialogues can be enhanced effectively. **3)** CB-300K is an integrated and versatile dataset. Users can use different subsets to train individual yet complementary abilities and combine them into a strong interaction system.

Evaluation Metric

For the relationship understanding tasks, existing metrics revolves around the model's performance within a single response (Li et al. 2023a; Yu et al. 2023b). This makes it more important to develop a metric that can evaluate the model's response to a series of upcoming questions.

We have noticed that some recent studies have applied state-of-the-art LLMs for evaluation (You et al. 2023; Bai et al. 2023). However, this approach carries certain risks: (1) GPT-4's speculative sampling strategy and online update brings unstable randomness during evaluation; (2) GPT-4 may develop certain preferences during instruction tuning, making its quantitative decisions biased. Therefore, we use RoBERTa-large (Liu et al. 2019) to calculate the BERT score (Zhang et al. 2019) that can evaluate the similarity between the model's output and the ground-truth answer.

Within each round, the setting is similar to grounded image captioning (Zhou et al. 2020; Li et al. 2023c). We employ two scores for evaluation: The first term utilizes the BERT score to evaluate the language part. The second term focuses on the visual grounding precision, taking into account the IoU between the detected and ground-truth bounding boxes.

If there is no request for grounding (*i.e.*, $M = 0$), the single-round score equals $\mathrm{BERT}(\mathbf{a}_m, \mathbf{a}_m^*)$, where $\mathrm{BERT}(\cdot, \cdot)$ 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

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

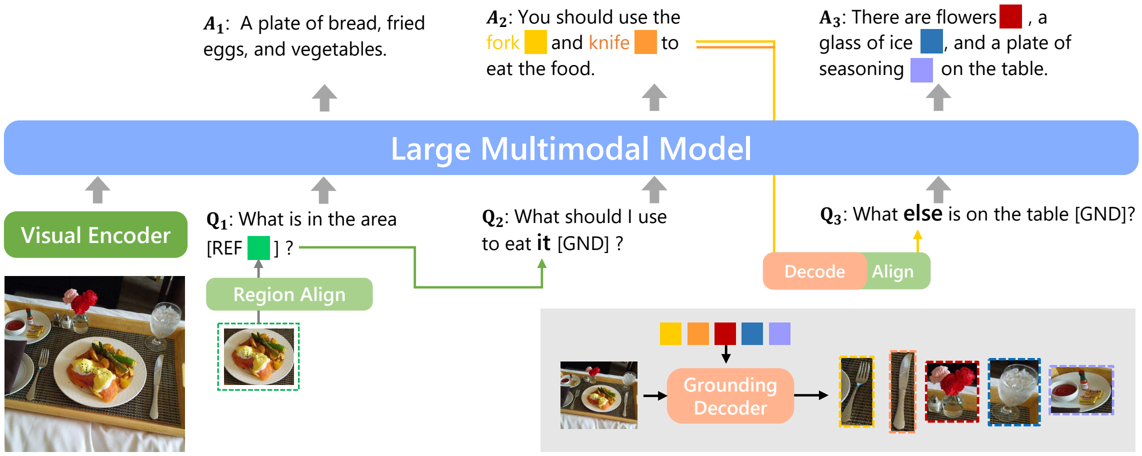

A_1: A plate of bread, fried eggs, and vegetables.
A_2: You should use the fork 🟨 and knife 🟧 to eat the food.
A_3: There are flowers 🟥 , a glass of ice 🟦, and a plate of seasoning 🟪 on the table.
Large Multimodal Model
Visual Encoder
Q_1: What is in the area [REF 🟩] ?
Region Align
Q_2: What should I use to eat it [GND] ?
Q_3: What else is on the table [GND]?
Decode
Align
Grounding Decoder

Figure 3: The architecture of the ChatterBox model. ChatterBox adopts a modularized design (LLM for language output, Region Align for referring, and Grounding Decoder for grounding) and referent feedback mechanism to resolve MCQ. ChatterBox takes image, referential region (if triggered with `[REF]`) and language instruction with preceding information as input, generates visual grounding results (if triggered with `[GND]`) and language answers. *This figure is best viewed in color.*

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

In the context of chain-of-questions, adherence to logical coherence is crucial. If a referential confusion happens in the preceding round (*e.g.*, incorrectly identifying an object), the subsequent questions (*e.g.*, inquiries regarding the attributes of the object) lose their significance. 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 Eq. 1 is smaller than τ_n, we immediately terminate the thread and set all scores in the later rounds to be 0. The overall evaluation score is the average of all rounds, *i.e.*, $T = \frac{1}{N}\sum_{n=1}^N t_n$.

ChatterBox: Understanding Chains-of-Questions

The overall architecture of the ChatterBox model is presented in Figure 3. We start the section by introducing the detailed modularized design. We then present the referent feedback mechanism for MCQ. Finally, we provide the training process and a discussion of the ChatterBox design.

Modularized Design

In contrast to existing approaches that employ coordinates to address visual referring and grounding tasks (Peng et al. 2023; Chen et al. 2023b; You et al. 2023), ChatterBox introduces a modularized design that ensures certain autonomy for both language and vision modules. The language model of ChatterBox exchanges a single token with the vision module for internal communication. This design simplifies the optimization process, accelerating ChatterBox's training procedure ($\sim$ 15 GPU days). Moreover, the inherited modules are minimally affected so that the grounding module produces boxes with high mIoU (Table 4) while the language model demonstrates text generation capabilities comparable to the baseline (Table 2).

Multimodal understanding. ChatterBox employs a large multimodal model that takes images ($\mathbf{x}_{\mathrm{img}}$) and texts ($\mathbf{x}_{\mathrm{txt}}$) as input and outputs text answers and query tokens to guide the grounding module. We feed $\mathbf{x}_{\mathrm{txt}}$ to the language branch of the CLIP-L/14 model (Radford et al. 2021), and $\mathbf{x}_{\mathrm{img}}$ (resized into 224×224) into the vision branch of the same CLIP-L/14 model. The outputs are a set of language tokens, denoted as $\mathbf{f}_{\mathrm{txt}}$, and a set of 16×16 vision tokens denoted as $\mathbf{f}'_{\mathrm{img}}$. It takes $\mathbf{f}_{\mathrm{txt}}$ and $\mathbf{f}'_{\mathrm{img}}$ as input to the large language model and output two-fold embeddings. The first set is simply decoded into the text answer, denoted as $\mathbf{z}_{\mathrm{ans}}$. The second set corresponds to the queries of visual grounding, denoted as $\mathbf{q}_{\mathrm{gnd}}$, which is only produced when the multimodal model detects a request for localization in the question. The multimodal model is inherited from LLaVA (Liu et al. 2023a), and fine-tuning is performed using the LoRA algorithm (Hu et al. 2021).

Visual grounding. The Grounding Decoder in Figure 3 is the module for visual grounding. In detail, we resize $\mathbf{x}_{\mathrm{img}}$ into 512×512 and feed it to an ViT model (Tian et al. 2023) which is pre-trained on Object365 (Shao et al. 2019). The output is a set of features with resolutions of 128×128, 64×64, 32×32, and 16×16, respectively, denoted as $\{\mathbf{f}_{\mathrm{img}}\}$. We use $\mathbf{q}_{\mathrm{gnd}}$ to query the multi-scale feature set $\{\mathbf{f}_{\mathrm{img}}\}$ for visual grounding. The module follows an enhanced DETR (Carion et al. 2020) object detector named DINO (Zhang et al. 2022a). Differently, to facilitate communication between them, we design a two-stage querying mechanism. In the first stage, we perform cross-attention between $\mathbf{q}_{\mathrm{gnd}}$ and $\{\mathbf{f}_{\mathrm{img}}\}$ 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, $\mathbf{q}_{\mathrm{gnd}}$ 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 en-

Method	Round #1			Round #2			Round #3			T
	$\mathrm{BERT}(\cdot)$	$\overline{\mathrm{IoU}}(\cdot,\cdot)$	t	$\mathrm{BERT}(\cdot)$	$\overline{\mathrm{IoU}}(\cdot,\cdot)$	t	$\mathrm{BERT}(\cdot)$	$\overline{\mathrm{IoU}}(\cdot,\cdot)$	t	
LLaVA (Liu et al. 2023a)	**0.935**	–	–	0.912	–	–	0.900	–	–	–
GPT4RoI (Zhang et al. 2023)	0.915	–	–	0.881	–	–	0.867	–	–	–
Kosmos-2 (Peng et al. 2023)	0.902	0.282	0.468	0.887	0.244	0.437	0.871	0.137	0.357	0.421
Shikra (Chen et al. 2023b)	0.913	0.272	0.464	0.891	0.231	0.429	0.870	0.132	0.353	0.415
LISA (Lai et al. 2023)	0.917	–	–	0.882	–	–	0.870	–	–	–
ChatterBox w/o RF	0.930	**0.401**	**0.560**	**0.918**	**0.377**	**0.539**	**0.908**	**0.306**	**0.487**	**0.529**
ChatterBox w/ RF	0.930	**0.401**	**0.560**	**0.920**	**0.379**	**0.541**	**0.915**	**0.310**	**0.492**	**0.532**

Table 2: A quantitative comparison of the MCQ metrics between ChatterBox and prior works.

coder) and the obtained queries are then propagated through a few attention layers (*a.k.a.* the decoder) to produce the set of box proposals and eventually the bounding boxes, $\mathcal{B}_{\mathrm{gnd}}$.

Visual referring. To build a referring module (Region Align in Figure 3), we follow GPT4RoI (Zhang et al. 2023) 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 (He et al. 2017; Zhang et al. 2023) operation is performed on the same CLIP-L/14 model. During referring, ChatterBox embeds a region into a token using the referring module and positions the resulting token next to the region in the text, as illustrated in figure 3.

Referent Feedback

The referent feedback mechanism (RF for short) is crafted to aid the multimodal model in comprehending referential instances within continuous dialogues. It functions by feeding back the grounded box from the previous question as a referring token (termed as feedback token) into the subsequent questions. Unlike the visual referring task of ChatterBox, where the referring token is placed within the question next to the referred instance, the feedback token is inserted at the end of the question to distinguish the referring task. Specifically, RF uses the referring module to embed the grounded region generated in the previous question into a token embedding. This token embedding is then inserted at the end of the next question. The feedback token can originate from the referred region, similar to the first-round dialogue in Figure 3. We note that we only utilize the RF mechanism when the confidence score of the grounded box is high enough (>0.8) and the next question uses referential words (such as 'it', 'this', 'he', *etc.*).

Figure 3 depicts two examples of the referent feedback mechanism in a three-round continuous dialogue. In the initial round, a visual referring request triggers the referring module to process the referred region (green region), generating a referring token positioned next to the referred instance ('area'). In the subsequent request with the referential word 'it', the referring token serves as a feedback token and is placed at the end of the question. This token carries region information, assisting the language model in understanding that 'it' refers to the green region. In the third request involving the referential word 'else', which queries the output boxes in the second-round answer, these boxes become feedback tokens after being processed by the referring module. Subsequently, these tokens are then positioned at the end of the question to help the model understand that the word 'else' implies the model should output answers excluding the information in these feedback tokens. These feedback tokens guide the language model, preventing confusion between similar objects.

Training and Discussion

There are two sources of supervision. For the text output, we compute the auto-regressive cross-entropy loss between $\mathbf{z}_{\mathrm{txt}}$ and the ground-truth answer, denoted by $\mathcal{L}_{\mathrm{txt}}$. For the grounding output (if present), we compute the localization loss between $\mathcal{B}_{\mathrm{gnd}}$ and the ground-truth set of bounding boxes, denoted as $\mathcal{L}_{\mathrm{gnd}}$. The overall loss is then written as $\mathcal{L}_{\mathrm{overall}} = \lambda_{\mathrm{txt}} \cdot \mathcal{L}_{\mathrm{txt}} + \lambda_{\mathrm{gnd}} \cdot \mathcal{L}_{\mathrm{gnd}}$, where λ_{txt} and λ_{gnd} are coefficients and both of them are set to 1.0 by default.

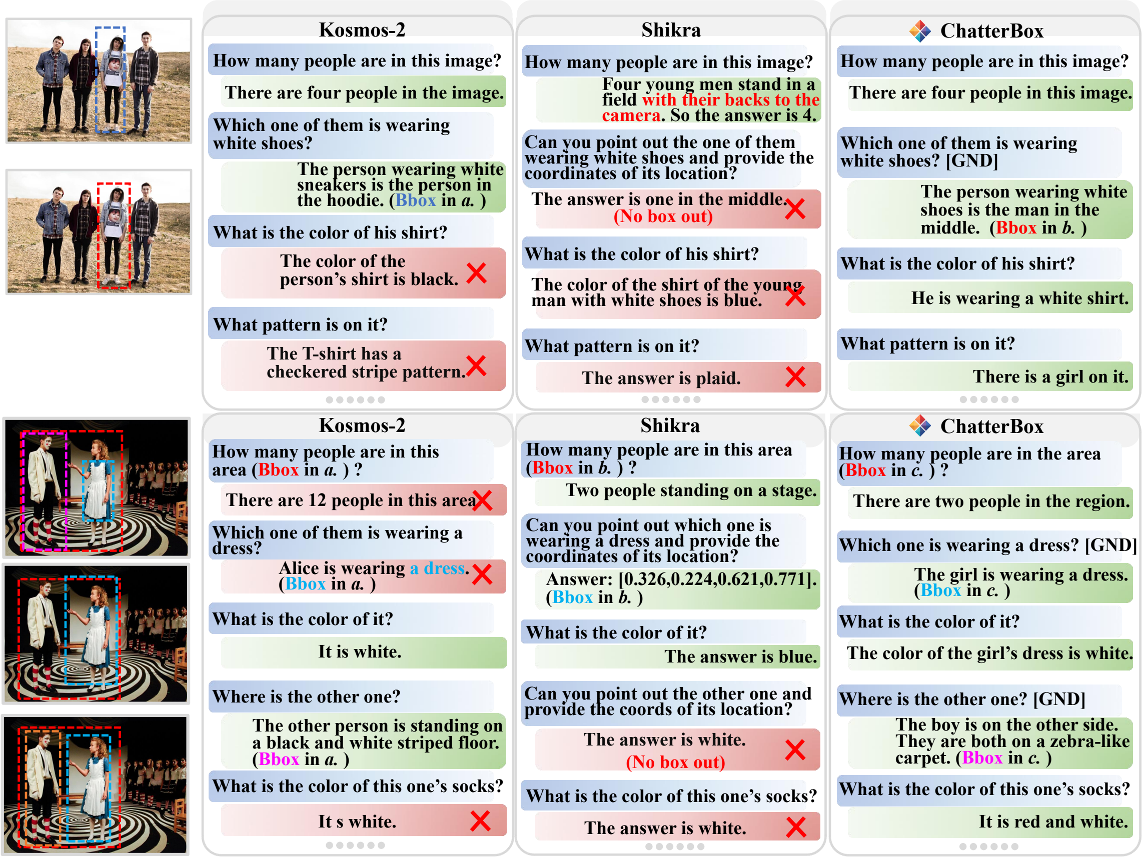
In practice, we have found that the visual grounding module can be challenging to train. Therefore, we adopt a two-stage training approach to mitigate this difficulty. During the initial stage, we warm up the training procedure by exclusively using the data related to grounding. Once the grounding loss $\mathcal{L}_{\mathrm{gnd}}$ reaches a sufficiently small value, we transition into the second stage, incorporating all available training data. In the first stage, we train the multimodal understanding ability using LoRA to fine-tune the LLaVA model. Then, in the second stage, we fine-tune all learnable parameters including LoRA and visual modules.

Discussion. The design principle of the ChatterBox model is to reflect the idea of decomposition, which has been reflected in prior works (*e.g.*, ViperGPT (Surís, Menon, and Vondrick 2023), HuggingGPT (Shen et al. 2023), Chameleon (Yu et al. 2023a), *etc.*). Under such principles, the LLM serves as the logic controller to understand the user's intention, and the additional ability is implemented by collaborating with external modules.

Experiment

Experimental Settings

Model architecture. In the grounding module, we employ a hierarchical transformer pyramid network (iTPN-B) (Tian et al. 2023, 2024) pre-trained on the Objects365 datasets (Shao et al. 2019) as the visual encoder. For the location decoder, we employ DINO detector (Zhang et al. 2022a), which by default incorporates 300 queries. DINO itself includes an encoder-decoder architecture with 6 blocks



Kosmos-2

How many people are in this image?

There are four people in the image.

Which one of them is wearing white shoes?

The person wearing white sneakers is the person in the hoodie. (Bbox in *a.*)

What is the color of his shirt?

The color of the person's shirt is black. ✗

What pattern is on it?

The T-shirt has a checkered stripe pattern. ✗

Shikra

How many people are in this image?

Four young men stand in a field with their backs to the camera. So the answer is 4.

Can you point out the one of them wearing white shoes and provide the coordinates of its location?

The answer is one in the middle. (No box out) ✗

What is the color of his shirt?

The color of the shirt of the young man with white shoes is blue. ✗

What pattern is on it?

The answer is plaid. ✗

ChatterBox

How many people are in this image?

There are four people in this image.

Which one of them is wearing white shoes? [GND]

The person wearing white shoes is the man in the middle. (Bbox in *b.*)

What is the color of his shirt?

He is wearing a white shirt.

What pattern is on it?

There is a girl on it.

Kosmos-2

How many people are in this area (Bbox in *a.*) ?

There are 12 people in this area ✗

Which one of them is wearing a dress?

Alice is wearing a dress. (Bbox in *a.*) ✗

What is the color of it?

It is white.

Where is the other one?

The other person is standing on a black and white striped floor. (Bbox in *a.*)

What is the color of this one's socks?

It s white. ✗

Shikra

How many people are in this area (Bbox in *b.*) ?

Two people standing on a stage.

Can you point out which one is wearing a dress and provide the coordinates of its location?

Answer: [0.326,0.224,0.621,0.771]. (Bbox in *b.*)

What is the color of it?

The answer is blue.

Can you point out the other one and provide the coords of its location?

The answer is white. (No box out) ✗

What is the color of this one's socks?

The answer is white. ✗

ChatterBox

How many people are in the area (Bbox in *c.*) ?

There are two people in the region.

Which one is wearing a dress? [GND]

The girl is wearing a dress. (Bbox in *c.*)

What is the color of it?

The color of the girl's dress is white.

Where is the other one? [GND]

The boy is on the other side. They are both on a zebra-like carpet. (Bbox in *c.*)

What is the color of this one's socks?

It is red and white.

Figure 4: A qualitative comparison in the continuous dialogues among Kosmos-2 (Peng et al. 2023), Shikra (Chen et al. 2023b), and ChatterBox (ours). In the dialogues above, ChatterBox exceeds its counterparts with no confusion or mistakes in its responses, showcasing its superior ability in continuous dialogues and reasoning.

for each part. In the language (multimodal) module, we use a LLaVA-13B model (Liu et al. 2023a), which is an MLLM based on LLaMA (Touvron et al. 2023) and fine-tuned on a visual instruction corpus. To fuse the visual features with the query token produced by the LLM, we utilize a cross-attention operation with a two-way transformer, following the SAM approach (Kirillov et al. 2023). 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. We utilize 8× NVIDIA A800 GPUs (80GB) for training, making use of DeepSpeed to improve computational efficiency. In the first stage, we employ the AdamW optimizer (Loshchilov and Hutter 2017) 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, with the total training cost being around 15 GPU-days. Due to limited space, we have moved the data pre-processing and organization to the appendix.

Evaluating the MCQ Task

We first evaluate our model in the MCQ setting using the metrics defined in Section , and a comparison with prior works is summarized in Table 2. We curate all threads of CB-CoQ's test set into three Q&A pairs, and each round (except for the first one) is logically related to the previous rounds.

In terms of the language output, ChatterBox produces better BERT scores than GPT4RoI (Zhang et al. 2023), Kosmos-2 (Peng et al. 2023), Shikra (Chen et al. 2023b), and LISA (Lai et al. 2023) (GPT4RoI, LISA, and ChatterBox all use LLaVA-13B model), and the advantage becomes more significant in the latter two rounds, implying its strong ability in dealing with continuous dialogues. Compared to the baseline (LLaVA (Liu et al. 2023a)), ChatterBox only exhibits a slight decrease (0.935→0.930) in BERT score, and surpasses it in the latter two rounds without any suspense. This result also shows that modularized design has minimal interference with ChatterBox's language ability.

Regarding the visual output, only Kosmos-2 and Shikra are compared, since LLaVA and GPT4RoI cannot perform

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val.	testA	testB	val.	testA	testB	val.	test
GPT4RoI (Zhang et al. 2023)	10.8	8.7	13.5	11.3	8.8	13.4	10.9	11.1
Kosmos-2 (Peng et al. 2023)	10.3	9.5	10.4	11.8	11.3	11.1	12.3	12.2
Shikra (Chen et al. 2023b)	–	–	–	–	–	–	15.0	15.2
ChatterBox	**13.6**	**13.3**	**13.6**	**15.1**	**15.1**	**14.3**	**16.7**	**16.6**

Table 3: A quantitative comparison of single-round referring on the RefCOCO/+/g datasets. All results are evaluated using the METEOR metric.

Method	mIoU	Succ. Rate	mIoU @ Succ.
Kosmos-2 (Peng et al. 2023)	0.627	0.688	0.854
Shikra (Chen et al. 2023b)	0.606	0.498	0.668
ChatterBox	**0.710**	**0.762**	**0.904**

Table 4: A quantitative comparison of single-round visual grounding on the COCO (Lin et al. 2014) 2017 test set. Please refer to the main text for the details of prompts and metrics.

visual grounding and LISA is unstable in localization [1]. Similarly, ChatterBox achieves the best $\overline{\mathrm{IoU}}(\cdot,\cdot)$ scores throughout the entire thread. Combining the high quality of language and visual output yields the better MCQ scores (*i.e.*, $\{t_n\}$ and T). This capability highlights the effectiveness of our model in handling grounding tasks.

We conduct another comparison with Kosmos-2 and Shikra in Figure 4. ChatterBox's visual module demonstrates its effectiveness once again, showing a stronger ability to accomplish continuous quests, while the competitors may run into failures.

Diagnostic Studies

Single-round referring. Our model, trained for MCQ, exhibits the anticipated proficiency in single-round referring tasks. We assess its performance on RefCOCO/RefCOCO+/RefCOCOg (Kazemzadeh et al. 2014) and compare it with GPT4RoI (Zhang et al. 2023), Kosmos-2 (Peng et al. 2023) and Shikra (Chen et al. 2023b). The result summarized in Table 3 shows that ChatterBox outperforms the competitors.

Single-round grounding. Similarly, ChatterBox can be used for single-round visual grounding. We compare it with Kosmos-2 (Peng et al. 2023) and Shikra (Chen et al. 2023b) on the COCO (Lin et al. 2014) 2017 test set. Table 4 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 all the models. As shown, ChatterBox surpasses Kosmos-2 and Shikra in these 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 (0.5 for Shikra). These results are impressive considering that the grounding data is $180\times$ fewer ($500\mathrm{K}$ vs. $90\mathrm{M}$). Additionally, We note that the successful cases of ChatterBox exhibit much higher IoU (0.904) compared to other methods. This is attributed to the precise box outputs facilitated by ChatterBox's independent grounding module.

CB-300K	Ref. Words	Round #1	Round #2	Round #3	*T*
✗	✓	0.929	0.904	0.895	0.478
✓	✓	0.930	0.918	0.908	0.529
✓	✗	0.930	0.924	0.921	0.547

Table 5: Diagnostic results in terms of the BERT score and the T score. **CB-300K**: whether CB-300K is used for training. **Ref. Words**: whether nominal substitution (*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 MCQ.

Benefit brought by the CB-300K dataset. We test the effectiveness of the proposed CB-300K in this part and summarize the results in Table 5. The first part involves not using the CB-300K data for training. Comparing the first two rows of Table 5, we find that the collected data consistently improves the model's ability of MCQ; similarly, the gain is 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 MCQ 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 MCQ indeed increases the difficulty of dialogues, so we believe it is a promising direction for MLLMs.

Conclusion

Multimodal large language models (MLLMs) are easily confused in handling complex referential questions especially when the logic forms a long chain. To address this issue, we propose a challenging task named Multimodal referring and grounding with Chain-of-Questions (MCQ), opening up an important direction for enhancing multimodal dialogue systems with strong logical coherence. We establish the CB-300K benchmark and equip it with an evaluation metric. CB-300K offers a large corpus of referring and grounding quests, many of which require complex logic understanding at the instance level. We also set up a baseline approach, ChatterBox, to tackle this problem. It is a modularized vision-language model equipped with referent feedback, and its effectiveness is validated by dealing with the MCQ task in the CB-300K benchmark. With the flexibility to handle complex instance relationships and the stability in logically continuous dialogues, ChatterBox has the potential to significantly advance the multimodal dialogue tasks with complicated and precise interactions.

[1] LISA does not support an explicit trigger(*e.g.*, the `[GND]` token), and its segmentation mask may contain outliers that deteriorate the box-level IoU.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 62225208 and 62450046 and CAS Project for Young Scientists in Basic Research No. YSBR-117.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Bai, S.; Yang, S.; Bai, J.; Wang, P.; Zhang, X.; Lin, J.; Wang, X.; Zhou, C.; and Zhou, J. 2023. TouchStone: Evaluating Vision-Language Models by Language Models. *arXiv preprint arXiv:2308.16890*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, C.; Qin, R.; Luo, F.; Mi, X.; Li, P.; Sun, M.; and Liu, Y. 2023a. Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models. *arXiv preprint arXiv:2308.13437*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023b. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; and Chen, K. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. LISA: Reasoning Segmentation via Large Language Model. *arXiv preprint arXiv:2308.00692*.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023a. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint arXiv:2307.16125*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; Chang, K.-W.; and Gao, J. 2022b. Grounded Language-Image Pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10955–10965.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023c. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, T.; Xie, L.; Tian, Y.; Yang, B.; Zhang, Y.; Doermann, D.; and Ye, Q. 2024. ClawMachine: Fetching Visual Tokens as An Entity for Referring and Grounding. *arXiv preprint arXiv:2406.11327*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824*.

Qiu, J.; Zhang, Y.; Tang, X.; Xie, L.; Ma, T.; Yan, P.; Doermann, D.; Ye, Q.; and Tian, Y. 2024. Artemis: Towards Referential Understanding in Complex Videos. *arXiv preprint arXiv:2406.00258*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Surís, D.; Menon, S.; and Vondrick, C. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Tian, Y.; Xie, L.; Qiu, J.; Jiao, J.; Wang, Y.; Tian, Q.; and Ye, Q. 2024. Fast-iTPN: Integrally pre-trained transformer pyramid network with token migration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tian, Y.; Xie, L.; Wang, Z.; Wei, L.; Zhang, X.; Jiao, J.; Wang, Y.; Tian, Q.; and Ye, Q. 2023. Integrally pre-trained transformer pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18610–18620.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; et al. 2023a. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023b. CogVLM: Visual Expert for Pretrained Language Models. *arXiv preprint arXiv:2311.03079*.

Xuan, S.; Guo, Q.; Yang, M.; and Zhang, S. 2023. Pink: Unveiling the Power of Referential Comprehension for Multimodal LLMs. *arXiv preprint arXiv:2310.00582*.

You, H.; Zhang, H.; Gan, Z.; Du, X.; Zhang, B.; Wang, Z.; Cao, L.; Chang, S.-F.; and Yang, Y. 2023. Ferret: Refer and Ground Anything Anywhere at Any Granularity. *arXiv preprint arXiv:2310.07704*.

Yu, L.; Shi, B.; Pasunuru, R.; Muller, B.; Golovneva, O.; Wang, T.; Babu, A.; Tang, B.; Karrer, B.; Sheynin, S.; et al. 2023a. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023b. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. *arXiv preprint arXiv:2308.02490*.

Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022a. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, S.; Sun, P.; Chen, S.; Xiao, M.; Shao, W.; Zhang, W.; Chen, K.; and Luo, P. 2023. GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. *arXiv preprint arXiv:2307.03601*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhou, Y.; Wang, M.; Liu, D.; Hu, Z.; and Zhang, H. 2020. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4777–4786.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.