Visual Intention Grounding for Egocentric Assistants

Pengzhan Sun[1] Junbin Xiao[1] * Tze Ho Elden Tse[1] Yicong Li[1]
Arjun Akula[2] Angela Yao[1]
[1]National University of Singapore [2]Google DeepMind
{pengzhan, junbin, eldentse, ayao}@comp.nus.edu.sg,
liyicong@u.nus.edu, arjunakula@google.com

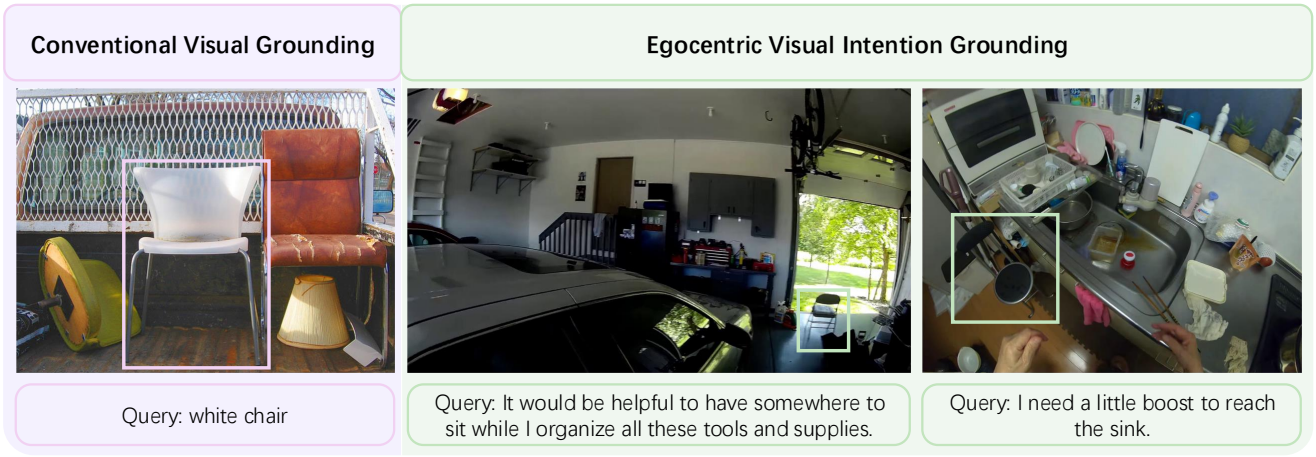

Figure 1. Traditional visual grounding (left) vs. egocentric visual intention understanding (center and right). Traditional grounding identifies the "white chair" by detecting specific objects from third-person perspectives. Egocentric visual intention understanding must infer user needs in complex, first-person scenarios, *e.g.*, seating in a workshop (center) or using a chair to reach the sink (right).

Abstract

Visual grounding associates textual descriptions with objects in an image. Conventional methods target third-person image inputs and named object queries. In applications such as AI assistants, the perspective shifts – inputs are egocentric, and objects may be referred to implicitly through needs and intentions. To bridge this gap, we introduce EgoIntention, the first dataset for egocentric visual intention grounding. EgoIntention challenges multimodal LLMs to 1) understand and ignore unintended contextual objects and 2) reason about uncommon object functionalities. Benchmark results show that current models misidentify context objects and lack affordance understanding in egocentric views. We also propose Reason-to-Ground (RoG) instruction tuning; it enables hybrid training with normal descriptions and egocentric intentions with a chained intention reasoning and object grounding mechanism. RoG significantly outperforms naive finetuning and hybrid training on EgoIntention, while maintaining or slightly improving naive description grounding. This advancement enables unified visual grounding for egocentric and exocentric visual inputs while handling explicit object queries and implicit human intentions. Our code and model are available at https://github.com/pengzhansun/EgoIntention.

1. Introduction

Consider the following scenarios: a person is looking for a place to sit down for organizing tools in a messy workshop, or a kid is trying to reach the sink in a kitchen (refer to Figure 1). A wearable artificial intelligent (AI) assistant could enhance these tasks by identifying contextually relevant objects (*e.g.*, a chair) without explicit object references. To achieve this, such an assistant must possess strong egocentric visual perception capabilities [11, 37, 43, 51]. This would significantly improve task efficiency, reduce cognitive load, and support hands-free, context-aware interaction in dynamic environments.

Building on this vision, we introduce the egocentric visual intention grounding task. Given an egocentric visual

*Corresponding author

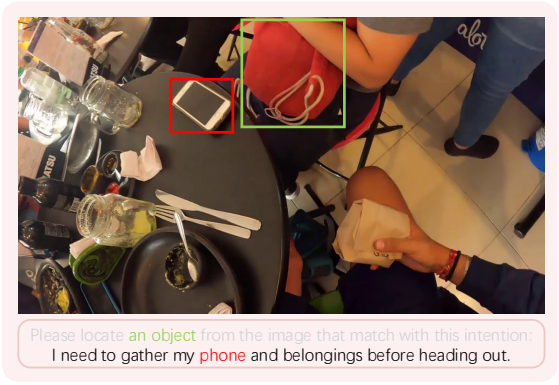

Figure 2. Challenge of Visual Intention Grounding. The model must infer the intended object from the full intention sentence, rather than simply detecting explicitly mentioned objects. In this example, "gather my phone and belongings" explicitly mentions "phone" (highlighted in red) , which often misleads existing visual grounding models to identify the wrong object (red box). The correct target, a handbag (green box), is only implied.

input and a human intention query, a model must accurately localize the intended object within the scene. This task supports real-world applications where users locate objects based on their needs. The object may not be directly named but can be inferred from the user's intention. Additionally, egocentric AI assistants operate from a first-person perspective, introducing challenges such as occlusions and dynamic viewpoints not present in conventional third-person vision systems. Unlike conventional visual grounding tasks [27, 30, 47], such as referring expression comprehension [9, 10, 16, 28, 29, 38, 39], this task requires reasoning about object affordance beyond explicit mentions of the object. As shown in Figure 2, existing models often misinterpret explicit mentions and fail to infer the actual intended object, highlighting the need for contextual reasoning in visual intention grounding.

Despite recent advancements in vision-language models [4, 6, 8, 17, 19, 31, 35, 36, 46, 48, 50], existing methods [1, 3, 5, 26, 40, 41] struggle to solve visual intention grounding in diverse real-world scenarios. First, solving the visual intention grounding task using two separate models, such as a language model (*e.g.*, GPT-4 [1]) for reasoning and an object detector (*e.g.*, GroundingDINO [23]) for localization, yields suboptimal results. These off-the-shelf models process visual and language information independently, often resulting in inconsistent associations between the two modalities. Consequently, a model may hallucinate objects by incorrectly identifying items that are not present in the scene. Secondly, while multimodal large language models (MLLMs) offer a unified approach, existing methods are primarily designed for third-person visual grounding tasks. Without datasets connecting egocentric visual data to intention sentences, these models struggle to adapt to and perform well on first-person intention grounding.

To bridge the gap in egocentric visual intention grounding, we introduce EgoIntention. EgoIntention is a comprehensive dataset built upon Ego4D [11], the largest real-world egocentric vision dataset. We inherit object bounding boxes annotation from PACO-Ego4D [33], a dataset that annotates object parts and attributes, and augment them with carefully curated human intention descriptions. Additionally, we address the inherent subjectivity in intention-object relationships by incorporating supplementary bounding box annotations for alternative objects that could reasonably fulfill each stated intention. Our EgoIntention dataset addresses two main gaps in visual grounding research: 1) the lack of egocentric data, and 2) the limited coverage of complex intentions arising from first-person viewpoints. Comprising 26,384 images, 52,768 human intention descriptions, and 89,841 annotated target object bounding boxes, this comprehensive dataset establishes a robust foundation for advancing visual intention understanding in real-world egocentric applications.

Despite providing a rich benchmark for egocentric intention grounding, EgoIntention also exposes significant challenges for existing models. We identify two major limitations. First, multimodal models rely on explicit object-centric prompts (*e.g.*, "Where is the white chair?"), directly mapping object names to their locations. However, intention-based queries require implicit reasoning. For example, when assisting in a workshop, the model must infer that the user needs a chair for sitting, rather than simply detecting a chair. Second, models must reason beyond direct object matches, recognizing alternative objects based on affordances. In the case of a child trying to reach the sink, the model should identify the chair as a suitable support, even though the query does not explicitly mention "chair." This requires a deeper understanding of object functionality and context, which is a challenge for current MLLMs.

To address the above challenges, we propose a model-agnostic Reason-to-Ground (RoG) instruction tuning approach. RoG disentangles intention understanding from object grounding to enhance multimodal models' reasoning capabilities. By doing so, RoG reduces spurious correlations between unintended object mentions and the actual intended object locations. On EgoIntention, our RoG instruction tuning improves MiniGPTv2's performance by 3.9 Precision@0.5 compared to naive finetuning and significantly outperforms the off-the-shelf GPT-4 [1] + GroundingDINO pipeline by 12.2 Precision@0.5.

Our contributions can be summarized as follows:

1. We construct the EgoIntention dataset for visual intention grounding. The dataset is the first egocentric visual grounding dataset with multiple intention queries.
2. We benchmark and reveal that existing MLLMs struggle with intention reasoning and egocentric visual grounding.

Table 1. Comparison of intention-related visual grounding datasets.

Dataset	#Images	Language Query	Ego-view	Multi-intention annotations
ADE-Aff [7]	10,000	Verb	✗	✗
PAD [25]	4,002	Verb	✗	✗
COCO-Tasks [34]	39,724	Phrase	✗	✗
RIO [32]	40,214	Template language	✗	✓
IntentionVG [42]	101,648	Free-form language	✓	✗
EgoIntention	26,384	Free-form language	✓	✓

These models misinterpret explicitly mentioned objects as targets or hallucinate objects not present in the scene.

3. We propose Reason-to-Grounding (RoG) instruction tuning, a model-agnostic training approach to enhance MLLMs for egocentric intention grounding while retaining their performances for normal visual grounding.

2. Related Work

2.1. Visual Grounding Datasets

Visual grounding [27, 30, 47] is a multimodal task that locates a target object in an image based on a given language query. Early works focused on referring expression comprehension [9, 10, 16, 28, 29, 38, 39], which matches descriptive phrases to objects within an image. Datasets such as RefCOCO [16], RefCOCO+ [47], and RefCOCOg [27] have played key roles in advancing this field. More recently, the scope of language input has been widened to encompass descriptions of object affordance [2, 18, 45]. This evolution led to datasets using verbs or phrases, (*e.g.*, ADE-Aff [7], PAD [25], COCO-Tasks [34]) and full sentences in datasets like RIO [32] and IntentionVG [42].

Compared to IntentionVG, where egocentric images are captured with a fixed viewpoint centered on objects, our dataset is constructed from Ego4D [11], introducing greater visual challenges such as motion blur, small object sizes, and perspective distortions inherent to first-person vision. For language queries, our dataset provides multiple intention sentences per object, reflecting the diverse ways an object can be used to fulfill different needs. Similar to the RIO dataset, we annotate each sample with both a context sentence, describing an object's typical use in its expected environment, and an uncommon sentence, which represents a less conventional use case requiring creative object substitution.

2.2. Visual Grounding Models

Traditional visual grounding methods [20, 21, 24, 27, 44] are specialized models explicitly trained to map language queries to object locations. Models such as MDETR [14], SeqTR [52], and Polyformer [22] leverage Transformer-based architectures to enhance this association. GroundingDINO [23], a recent advancement in open-set object detection, extends DINO [49] with grounded pre-training, allowing it to detect arbitrary objects given category names or referring expressions. Recently, multimodal large language models (MLLMs) have emerged as the dominant paradigm for vision-language tasks [1, 3, 5, 15, 26, 40, 41]. By leveraging vast amounts of image-text data and instruction tuning, these models achieve impressive generalization across various multimodal benchmarks. While effective in conventional visual grounding tasks, both specialist models and MLLMs primarily perform word-by-word detection on the input language query, rather than truly understanding the underlying human intention. To address these challenges, we propose Reason-to-Ground (RoG), a method that explicitly disentangles intention reasoning from object localization, as detailed in method Section 4.

3. Intention Grounding & EgoIntention Dataset

3.1. Task Description

We introduce visual intention grounding, a novel paradigm that establishes a direct mapping between human intentions and target objects in visual scenes. This approach bridges the gap between natural language understanding and visual object grounding. Formally, given an egocentric image I and a human intention query Q, the task requires the model to comprehend the underlying intention and localize the target object O that satisfies the user's need by predicting its bounding box coordinates (x_1, y_1, x_2, y_2).

3.2. Dataset Collection

Our dataset, EgoIntention, sources its images from the Ego4D dataset [11] and its associated object bounding box annotations from PACO [33]. The key contribution of EgoIntention is the collection of multiple intention queries per object, reflecting the diverse ways objects are used in real-world scenarios. Our systematic data collection pipeline consists of three stages, as illustrated in Figure 3.

The initial stage generates intention descriptions. We use GPT-4 [1]'s multimodal capabilities to analyze egocentric visual inputs and generate contextually relevant human intention sentences. To capture diverse real-world scenarios, we design two types of intention descriptions:

- **Context-aware intentions:** These sentences reflect complex object relationships and environmental cues from a first-person perspective. For instance, a user might think, "I noticed a wobbly table leg that needs fixing," expressing the need for a hammer.
- **Uncommon intentions:** These describe atypical object uses, where users repurpose objects based on necessity. For example, a backpack might be used as an improvised umbrella during unexpected rain.

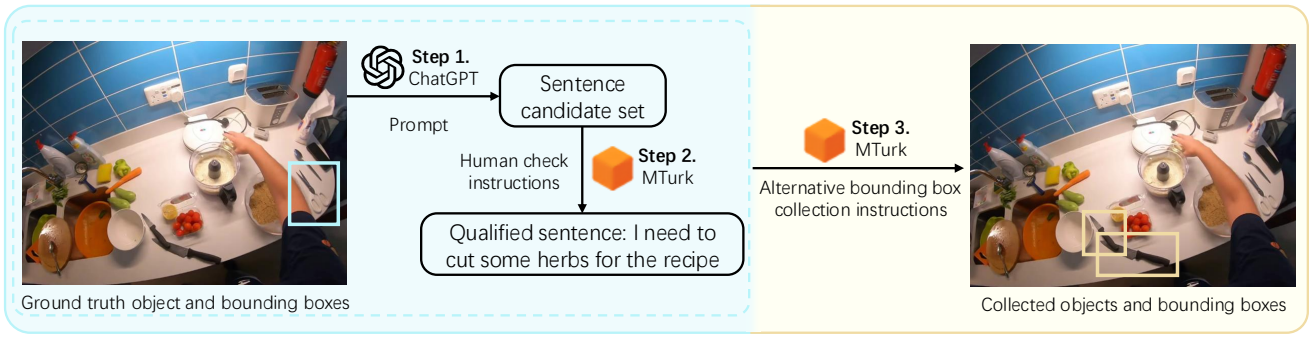

Figure 3. Dataset collection pipeline for EgoIntention. Our data collection process consists of three key stages. (1) Intention Sentence Generation: We use GPT-4 to generate egocentric human intention sentences based on visual input, covering both context-aware and uncommon intention scenarios. (2) Human Validation via MTurk: Annotators assess the semantic validity and real-world applicability of generated sentences, filtering out low-quality or ambiguous descriptions. (3) Alternative Object and Bounding Box Collection: Given the inherent subjectivity of human intentions, additional valid object candidates are identified by human annotators, supplementing the original ground truth annotations with alternative bounding boxes.

These sentences are generated using carefully crafted prompts through an in-context learning approach as shown in Appendix.

The second stage involves human validation via Amazon Mechanical Turk (MTurk). Annotators assess each generated sentence using detailed guidelines to ensure semantic correctness and real-world plausibility. Since multiple valid intentions may exist for a given scenario, we do not use inter-annotator agreement as a filtering criterion. Instead, one annotator selects and, if necessary, refines a suitable sentence. This selection is then verified by a second annotator and a GPT-4-based checker, both of which independently assess whether the sentence expresses a genuine need for the target object. Only those passing both checks are retained.

To account for the subjectivity of human intentions, we include additional object annotations. A single intention may be satisfied by various objects (*e.g.*, flower pots, bottles, or cups for desktop decoration). Annotators identify such alternatives, which are added as supplementary bounding boxes. As before, only those passing both human and GPT-4 verification are included.

We observed that GPT-4 performs more reliably as a verifier than as a generator. While context-aware sentences passed human checks at 97.2%, uncommon intentions had a lower pass rate of 74.1%, often due to generic outputs or object name leakage. However, as a verifier, GPT-4 achieved 92% agreement with human judgments in a 500-sample study, supporting its role as an auxiliary checker. To further validate our pipeline, we conducted an inter-annotator study on 500 verified samples, yielding agreement rates of 98.6% for context-aware and 91.2% for uncommon intentions. The maintained or slightly improved RefCOCO performance after joint training further supports the quality of our annotations.

3.3. Dataset Statistics

	Image	Context BBox	Uncommon BBox
Train	15,667	25,772	25,933
Val	825	1,402	1,366
Test	9,892	17,699	17,669

Table 2. Number of images and bounding boxes (BBox) in the EgoIntention dataset. Context BBox refers to bounding boxes associated with objects commonly used in the given scene, aligning with expected human intentions based on environmental cues. Uncommon BBox represents bounding boxes for objects used in unconventional ways, where the intended action requires creative or atypical object usage.

Our EgoIntention dataset builds upon PACO's image splits, comprising 15,667 training, 825 validation, and 9,892 test samples. For each image, we annotate two types of intention queries: context-aware intentions that leverage environmental cues, and uncommon intentions that capture alternative object uses. Following our multi-stage annotation pipeline, we enrich the dataset with supplementary bounding box annotations to accommodate the inherent diversity of object choices for each intention. The distribution of bounding box annotations across different splits and intention types is summarized in Table 2.

4. Method

This section begins with an observational study of off-the-shelf hybrid models in Section 4.1, highlighting key limitations in their reasoning and detection pipeline. Based on these findings, we propose Reason-to-Ground (RoG) to enhance multimodal large language models for visual intention

grounding in Section 4.2. We then detail our supervised fine-tuning approach in Section 4.3.

4.1. Observation of Off-the-Shelf Hybrid Models

A straightforward approach to visual intention grounding is to leverage two off-the-shelf models separately for reasoning and detection. Specifically, we use ChatGPT for intention reasoning and GroundingDINO for object detection. From our observation study (detailed in Section 5), we identify two key findings: Performing reasoning first improves accuracy. Narrowing the search space before detection leads to more precise object localization. Hybrid off-the-shelf models suffer from inconsistent representations. These models operate in distinct visual and language spaces, causing discrepancies that hinder effective grounding. Motivated by these findings, our main method introduces Reason-to-Ground, an instruction tuning approach for multimodal large language models.

4.2. Reason-to-Ground Instruction Tuning

We propose a novel training strategy, Reason-to-Ground Instruction Tuning (RoG). Our method decomposes the task into two essential components: human intention understanding and visual grounding. Existing approaches [3, 5] directly feed implicit intention sentences with a task-specific token `<ref>`, which prompts the MLLM to output the bounding box corresponding to the input language query. In contrast, RoG employs a two-stage process. In the first stage, we facilitate human intention understanding by querying the MLLM with a `<reason>` token followed by the implicit intention sentence. This prompts the model to output the target object category. In the second stage, we perform `<ref>` token and the explicit object description derived from the first stage. These two stages are illustrated with an example in Figure 4.

By disentangling **intention reasoning** and **object grounding**, RoG prevents the model from naively associating explicitly mentioned objects with target bounding boxes, leading to more accurate intention-driven visual grounding.

4.3. Supervised Fine-tuning

We propose a unified grounding framework that can process exo- and egocentric visual inputs while handling explicit object queries and implicit human intentions. To achieve this comprehensive capability, we leverage traditional exocentric datasets (RefCOCO [16], RefCOCOg [27], and RefCOCO+ [47]) alongside our proposed EgoIntention dataset during training. Our method RoG effectively extends MLLMs' visual grounding capabilities to accommodate egocentric image inputs and implicit human intention queries through LoRA [12] supervised fine-tuning.

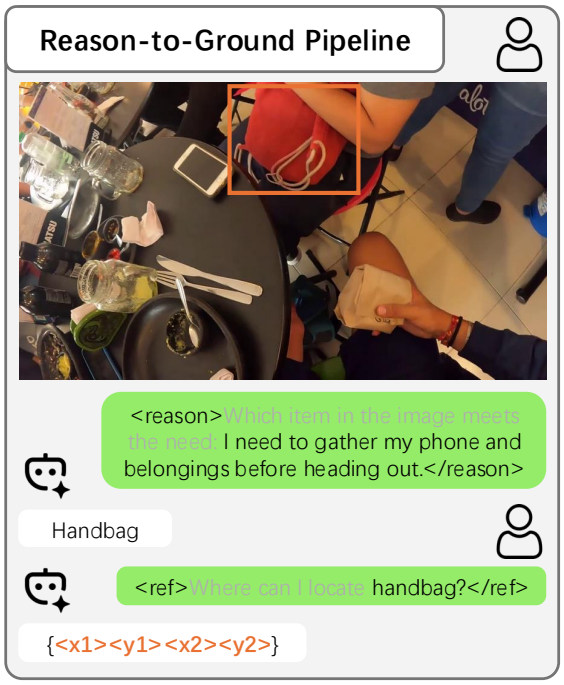

Figure 4. Overview of Reason-to-Ground Instruction Tuning (RoG): The model first infers an explicit object category from an implicit intention sentence (intention reasoning), then localizes the object in the image (object grounding).

5. Experiments

5.1. Implementation Details

We conducted supervised finetuning using a comprehensive dataset that combines traditional referring expression comprehension datasets (RefCOCO, RefCOCO+, and RefCOCOg) with our proposed visual intention grounding dataset, EgoIntention. For our experiments, we evaluated five state-of-the-art multimodal large language models (MLLMs), all with 7B parameters: MiniGPTv2, Groma, CogVLM-grounding, and Qwen-VL. All model training was performed on 4 NVIDIA A40 GPUs. To maintain computational efficiency while preserving model performance, we employed Low-Rank Adaptation (LoRA) [12] for parameter-efficient finetuning.

5.2. Models

5.2.1. Hybrid Model Baselines

A straightforward approach to visual intention grounding is to use off-the-shelf state-of-the-art models to separately address the vision and language components of the task. For vision, we adopt GroundingDINO, a popular open-set object detector that accepts a set of object category queries and returns the corresponding bounding boxes. For language understanding, we use ChatGPT, a large-scale language model

developed by OpenAI, to reason about the user's intention and infer the target object category.

As detailed in Section 4, we implement this two-stage framework in two variants: (1) first GPT-4 then GroundingDINO (Reasoning–Detection baseline, R-D), and (2) first GroundingDINO then GPT-4 (Detection–Reasoning baseline, D-R).

D-R baseline. We begin by applying GroundingDINO to detect a comprehensive set of candidate object bounding boxes in the egocentric image. To ensure broad coverage, we use all object categories provided by the PACO dataset as text queries. GroundingDINO returns bounding boxes along with associated logits and predicted phrases. We then pass these predicted phrases, along with the human intention sentence, to GPT-4. GPT-4 performs reasoning over the textual descriptions to identify the object category most aligned with the expressed intention.

R-D baseline. In contrast, this variant begins with GPT-4. Given the egocentric image and a prompt listing all candidate object categories, GPT-4 reasons about the intention and outputs the predicted object category. This output is then used to filter the query space for GroundingDINO, which subsequently predicts the bounding box for the inferred object by focusing only on the categories identified through GPT-4's reasoning.

5.2.2. Multimodal Large Language Models

We evaluate our approach on several state-of-the-art MLLMs, categorized into two groups based on their specialized capabilities:

Grounding Specialist MLLMs excel at visual grounding tasks including grounded captioning, referring expression generation, referring expression comprehension, and grounded visual question answering:

- **CogVLM-grounding** incorporates a trainable visual expert module within the attention and FFN layers, effectively bridging the gap between the frozen pretrained language model and image encoder.
- **Groma** demonstrates exceptional region-level grounding capability by integrating region tokens into both user instructions and model responses, enabling precise localization of described language queries.

Compared to grounding specialists, **Generalist MLLMs** exhibit enhanced reasoning abilities—a critical capability for solving visual intention grounding tasks:

- **MiniGPT-v2** provides a unified interface for numerous vision-language tasks, employing unique task identifiers during model training to facilitate multi-task learning.
- **Qwen-VL**, built upon the Qwen-LM foundation, transcends conventional image description and question-answering capabilities by incorporating robust visual grounding functionality.

5.3. Evaluation Metrics

Our evaluation accounts for multiple alternative ground-truth boxes. We calculate each sample's score by computing the IoU between the predicted box and all ground truth boxes and then take the maximum IoU. A prediction is correct if its IoU with any ground truth box exceeds a threshold of 0.3 or 0.5 (Precision@0.3 or @0.5).

5.4. Zero-shot Evaluation

We first test all the 6 above mentioned methods in a zero-shot setting in Table 3. Our experiments reveal that **the order of using the reasoner and detector significantly impacts results**, despite the D-R and R-D baselines using the same models. For visual context-aware intention understanding, the D-R baseline correctly understands and accurately detects only 21.1% of samples. In contrast, the R-D baseline improves P@0.5 accuracy to 46.6%. We attribute this difference to GroundingDINO's limitations when processing numerous language queries. When used first, it must handle all the object categories from the EgoIntention dataset and omits many object candidates. However, by using GPT-4 for initial reasoning, we narrow the target objects to one or two categories. This focused input allows GroundingDINO to perform 25% better than in the D-R baseline. For uncommon intention understanding, we observe a similar trend. The R-D baseline (23.6%) outperforms the D-R baseline (14.6%), but P@0.5 improves by only 9%. This smaller improvement stems from GPT-4's lower reasoning accuracy for uncommon intentions, failing to provide GroundingDINO with the correct object category for detection.

Grounding-specialist MLLMs, such as CogVLM-grounding and Groma, perform poorly on EgoIntentions. This can be attributed to human intention reasoning being crucial for solving the visual intention grounding task. However, these specialist models are primarily aligned with grounding-related tasks and lack the necessary reasoning capability to infer which object should be grounded. Consequently, they fail to identify the intended object, leading to suboptimal results. In contrast, **generalist MLLMs such as MiniGPT-v2 and Qwen-VL demonstrate more reasonable grounding performance** when given an intention query. Despite reasoning capabilities learned from pretraining and alignment, generalist MLLMs' overall performance remains limited. MiniGPT-v2 achieves an overall Precision@0.5 score of 19.9%, while Qwen-VL reaches 21.7%; both are significantly lower than the R-D GPT4-GroundingDINO baseline (35.1%). This performance gap highlights the importance of integrating strong intention reasoning with visual grounding for improved results.

5.5. RoG Supervised Finetuning Results

In addition to the EgoIntention training set, we incorporate RefCOCO, RefCOCO+, and RefCOCOg (RefCOCO/+/g) to

Table 3. Benchmark comparison on EgoIntention across various methods, including two-stage pipeline approaches, grounding-specialist MLLMs, and generalist MLLMs. Performance is evaluated using Precision@0.3, Precision@0.5, and mIoU, reported for both the Context Split and Uncommon Split. The Overall P@0.5 metric summarizes the general performance across splits.

Method	Context Split			Uncommon Split			Overall P@0.5
	P@0.3	P@0.5	mIoU	P@0.3	P@0.5	mIoU	
D-R GroundingDINO-GPT4	30.1	21.1	0.217	20.6	14.6	0.150	17.8
R-D GPT4-GroundingDINO	54.3	46.6	0.402	31.7	23.6	0.242	35.1
CogVLM-grounding	8.0	5.9	0.057	6.3	4.9	0.042	5.4
Groma	9.6	7.4	0.074	8.9	6.9	0.070	7.2
MiniGPT-v2	31.3	21.7	0.224	25.5	18.0	0.186	19.9
Qwen-VL	27.9	21.4	0.225	27.3	22.0	0.228	21.7

Table 4. Comparison of instruction tuning methods on RefCOCO, RefCOCO+, RefCOCOg, and EgoIntention datasets. Naive SFT refers to LoRA-based supervised fine-tuning, while RoG SFT represents LoRA-based fine-tuning with our Reason-to-Ground Instruction Tuning (RoG) instruction tuning approach. We report Precision@0.5 as the evaluation metric across different validation and test splits.

Model	RefCOCO			RefCOCO+			RefCOCOg		EgoIntention		
	val	testA	testB	val	testA	testB	val	test	context	uncommon	overall
Zero-shot MiniGPTv2	87.4	91.3	83.7	79.0	85.1	72.8	83.5	84.1	21.7	18.0	19.9
Naive SFT	86.6	91.0	83.0	79.0	84.9	72.0	82.6	84.2	46.0	40.9	43.4
RoG SFT	87.8	91.4	84.0	79.8	85.4	73.8	84.3	85.2	**49.9**	**44.7**	**47.3**
Zero-shot Qwen-VL	89.3	92.4	85.4	83.2	88.2	77.2	85.3	85.6	21.4	22.0	21.7
Naive SFT	89.5	92.8	85.7	83.4	88.8	77.8	85.9	86.3	32.1	26.1	29.1
RoG SFT	89.3	92.5	85.3	83.3	88.8	77.4	86.2	86.4	**35.5**	**31.7**	**33.6**

maintain model performance on referring expression comprehension while also leveraging these datasets to enhance intention grounding performance. As a result, our training data consists of three components: RefCOCO/+/g, RefCOCOIntention/+/g, and EgoIntention. The RefCOCOIntention/+/g dataset is generated automatically using GPT-4, applying the same prompt used for collecting EgoIntention, to create human intention queries.

While Naive SFT improves EgoIntention performance, it slightly degrades the model's capability in referring expression comprehension as shown in Table 4. In contrast, **our RoG instruction tuning strategy not only further enhances performance on EgoIntention but also leads to improved results on the RefCOCO series datasets.** After fine-tuning with our RoG strategy, the generalist MLLM MiniGPT-v2 surpasses the best off-the-shelf two-stage method (R-D GPT4-GroundingDINO) by 12.2% according to Precision@0.5, demonstrating the effectiveness of our approach in bridging intention reasoning and grounding. Unlike R-D GPT4-GroundingDINO, which treats these as two independent sub-tasks, our method jointly models the reasoning and grounding processes, leading to more coherent and accurate results. Additionally, we observe that RoG SFT performs better on MiniGPT-v2 than on Qwen-VL. This discrepancy arises because MiniGPT-v2 exhibits stronger adaptability to new task-specific tokens, enabling it to integrate hierarchical reasoning instructions more effectively. In contrast, Qwen-VL demonstrates weaker instruction-following capabilities during supervised fine-tuning, limiting its performance gains under the same strategy. See visualization examples in Appendix.

5.6. Ablation Study

5.6.1. Impact of Supervised Finetuning Datasets

We analyze the performance gains from different datasets used during the SFT stage in Table 5. Finetuning exclusively on EgoIntention hurts generalization, dropping RefCOCO validation performance from 87.4% to 66.5%. Combining RefCOCO/+/g with EgoIntention maintains REC performance while improving EgoIntention metrics. Specifically, context intention performance improves from 42.8% to 45.9%, and uncommon intention performance increases from 39.2% to 40.8% (see Table 5). This suggests that conventional REC datasets contribute additional gains, particularly in object grounding.

We explore whether adding human intention annotations

Table 5. Ablation study of training datasets used for MiniGPT-v2 fine-tuning and their impact on visual grounding datasets.

SFT Datasets			Method	RefCOCO			RefCOCO+			RefCOCOg		EgoIntention			
RC/+/g	RCInt./+/g	EgoInt.		Val	TestA	TestB	Val	TestA	TestB	Val	Test	Con	Unco	Ave.	Obj.
-			0-shot	87.4	91.3	83.7	79.0	85.1	72.8	83.5	84.1	21.7	18.0	19.9	40.8
✓			Naive SFT	87.6	91.3	84.4	**80.0**	85.3	**73.8**	**84.8**	85.2	23.7	19.4	21.5	38.1
		✓	Naive SFT	66.5	71.5	60.4	60.2	66.6	51.8	64.6	65.8	42.8	39.2	41.0	46.2
✓		✓	Naive SFT	87.5	**91.5**	**84.6**	79.9	**85.6**	73.5	84.7	**85.4**	45.9	40.8	43.3	48.6
✓	✓	✓	Naive SFT	86.6	91.0	83.0	79.0	85.0	72.0	82.6	84.2	46.0	40.9	43.4	51.3
✓	✓	✓	RoG SFT	**87.8**	91.4	84.0	79.8	85.4	73.8	84.3	85.2	49.9	44.7	47.3	52.2

to REC datasets improves EgoIntention performance. We create RefCOCOInt/+/g by collecting intention sentences for RefCOCO/+/g training sets. This augmentation yields only slight improvements. Without human verification, data quality remains a bottleneck. These limited gains suggest high-quality annotations are crucial for advancing intention grounding.

The best overall performance comes from applying our RoG strategy across all training datasets, demonstrating the effectiveness of our approach in jointly improving REC task and visual intention grounding task.

5.6.2. RoG Improves Explicit Object Grounding

As shown in Table 5, we further evaluate the model's performance on EgoIntention with explicit object queries, as indicated in the last column labeled "object". Compared to Naive SFT, RoG SFT disentangles visual intention understanding from object grounding, preventing the model from being misled by explicitly mentioned but unintended objects in the intention query. **Our approach also leads to a significant improvement in egocentric visual grounding with explicit object queries**, boosting the accuracy from 51.3% with Naive SFT to 52.2% with RoG SFT.

5.6.3. Hallucination and Misleading Errors.

Failure cases are due to *hallucination* (object is neither mentioned in the query nor present in the image) and object reasoning. The latter can be divided into *misleading language* (object mentioned in the query but not intended) and *misleading vision* (object appears in the image but is not intended). We use RAM++ [13] to extract all object tags in the image to check for misleading vision errors. RoG fine-tuned MiniGPTv2 achieves a lower error rate across three categories as shown in the table 6.

Table 6. Error rates (%) for hallucination and reasoning failures across context and uncommon splits.

Model	Context Split			Uncommon Split		
	Language Mislead	Vision Mislead	Hallucination	Language Mislead	Vision Mislead	Hallucination
Zero-shot	3.4	8.4	30.1	6.9	7.5	58.2
Naive SFT	**1.4**	6.4	23.4	3.9	7.5	44.0
RoG SFT	2.1	**3.8**	**12.0**	**2.4**	**6.6**	**24.3**

5.6.4. Uncommon Intention Coverage

Uncommon intentions are inherently subjective and underexplored. Prior work like RIO [32] offers limited uncommon cases with 4,826 test samples only for testing and none for training. We curated five diverse, uncommon samples per object category as in-context learning GPT-4 prompts to construct the first large-scale training set for this setting. Table 7 shows that removing our uncommon split from training leads to a clear drop in Precision@0.5 on EgoIntention uncommon test set. Training with our uncommon intention data also improves zero-shot performance on RIO uncommon set for both naive and RoG SFT models.

Table 7. Impact of training with uncommon intentions on EgoIntention and RIO benchmarks.

Methods	Training w/ EgoInt. Uncommon	EgoInt. Uncommon	Zero-shot on RIO Uncommon
Naive SFT	–	33.0	21.4
	✓	**40.9**	**22.1**
RoG SFT	–	33.7	23.8
	✓	**44.7**	**26.0**

6. Conclusion

We introduce egocentric visual intention grounding, where AI assistants infer and localize intended objects based on implicit human intentions. To support this research, we construct EgoIntention and benchmark state-of-the-art large vision-language models (LVLMs). Our results reveal that LVLMs struggle with implicit intention inference and egocentric visual grounding. For improvements, we propose Reason-to-Ground Instruction Tuning (RoG), a model-agnostic approach that disentangles intention reasoning from visual grounding, reducing spurious correlations and improving alignment with human intent. By applying RoG in supervised fine-tuning with hybrid data from normal visual grounding and intention grounding tasks, LVLMs retain strong performance on conventional visual grounding while achieving significant improvements in egocentric visual intention grounding, offering a promising approach for both object and intention queries in exo- and ego-centric visual environments.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3

[2] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023. 3

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 2, 3, 5

[4] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26540–26550, 2024. 2

[5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2, 3, 5

[6] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2

[7] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018. 3

[8] Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. *arXiv preprint arXiv:2504.20930*, 2025. 2

[9] Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. Learning distributions over logical forms for referring expression generation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1914–1925, 2013. 2, 3

[10] Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419, 2010. 2, 3

[11] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 3

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[13] Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. *arXiv preprint arXiv:2310.15200*, 2023. 8

[14] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 3

[15] Liu Kangcheng, Xiao Junbin, Zhang Rui, Lv Hanqi, and Du Zidong. Bottom-up and top-down thoughts for visual intention grounding. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, page 889–898. Association for Computing Machinery, 2025. 3

[16] 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, 3, 5

[17] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2928–2937, 2022. 2

[18] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14251–14260, 2024. 3

[19] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024. 2

[20] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020. 3

[21] 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. 3

[22] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023. 3

[23] 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. 2, 3

[24] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative

network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 3

[25] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot affordance detection. *arXiv preprint arXiv:2106.14747*, 2021. 3

[26] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2025. 2, 3

[27] 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. 2, 3, 5

[28] Margaret Mitchell, Kees van Deemter, and Ehud Reiter. Natural reference to objects in a visual domain. In *Proceedings of the 6th international natural language generation conference*, 2010. 2, 3

[29] Margaret Mitchell, Kees Van Deemter, and Ehud Reiter. Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1184, 2013. 2, 3

[30] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 2, 3

[31] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14076–14088, 2024. 2

[32] Mengxue Qu, Yu Wu, Wu Liu, Xiaodan Liang, Jingkuan Song, Yao Zhao, and Yunchao Wei. Rio: A benchmark for reasoning intention-oriented objects in open environments. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 8

[33] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 2, 3

[34] Johann Sawatzky, Yaser Souri, Christian Grund, and Jurgen Gall. What object should i use?-task driven object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7605–7614, 2019. 3

[35] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv e-prints*, pages arXiv–2403, 2024. 2

[36] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024. 2

[37] Pengzhan Sun, Bo Wu, Xunsong Li, Wen Li, Lixin Duan, and Chuang Gan. Counterfactual debiasing inference for compositional action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3220–3228, 2021. 1

[38] Kees van Deemter, Ielka van der Sluis, and Albert Gatt. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the fourth international natural language generation conference*, pages 130–132, 2006. 2, 3

[39] Jette Viethen and Robert Dale. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67, 2008. 2, 3

[40] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 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] Wenxuan Wang, Yisi Zhang, Xingjian He, Yichen Yan, Zijia Zhao, Xinlong Wang, and Jing Liu. Beyond literal descriptions: Understanding and locating open-world objects aligned with human intentions. *arXiv preprint arXiv:2402.11265*, 2024. 3

[43] Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. Egoblind: Towards egocentric visual assistance for the blind people. *arXiv preprint arXiv:2503.08221*, 2025. 1

[44] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4644–4653, 2019. 3

[45] Tomoya Yoshida, Shuhei Kurita, Taichi Nishimura, and Shinsuke Mori. Text-driven affordance learning from egocentric vision. *arXiv preprint arXiv:2404.02523*, 2024. 3

[46] 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. 2

[47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2, 3, 5

[48] Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023. 2

[49] 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. 3

[50] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 2

[51] Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3363–3373, 2025. 1

[52] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022. 3