Beyond Literal Descriptions: Understanding and Locating Open-World Objects Aligned with Human Intentions

Anonymous ACL submission

Abstract

Visual grounding (VG) aims at locating the foreground entities that match the given natural language expressions. Previous datasets and methods for classic VG task mainly rely on the prior assumption that the given expression must literally refer to the target object, which greatly impedes the practical deployment of agents in real-world scenarios. Since users usually prefer to provide intention-based expression for the desired object instead of covering all the details, it is necessary for the agents to interpret the intention-driven instructions. Thus, in this work, we take a step further to the intention-driven visual-language (V-L) understanding. To promote classic VG towards human intention interpretation, we propose a new intention-driven visual grounding (IVG) task and build a large-scale IVG dataset termed IntentionVG with free-form intention expressions. Considering that practical agents need to move and find specific targets among various scenarios to realize the grounding task, our IVG task and IntentionVG dataset have taken the crucial properties of both multi-scenario perception and egocentric view into consideration. Besides, various types of models are set up as the baselines to realize our IVG task. Extensive experiments on our IntentionVG dataset and baselines demonstrate the necessity and efficacy of our method for the V-L field. To foster future research in this direction, our newly built dataset and baselines will be publicly available.

1 Introduction

Recently, the research community has witnessed the rapid advancement of multimodal embodied intelligence (Ahn et al., 2022; Reed et al., 2022; Driess et al., 2023; Shah et al., 2023; Gao et al., 2023; Brohan et al., 2023). For an intelligent agent, the capability of locating the target objects in the unpredictable open-world scenarios based on natural language expressions is crucial, underscoring the importance of visual grounding (VG) task within the broader context. Notably, instructions provided by users often encapsulate their genuine needs through nuanced intention-driven expressions, which are usually not literal or explicit with much details as classic VG task. This nuance brings to light the critical role of VG based on user intention expressions, where challenge lies in interpreting and responding to user commands in a way that truly reflect their underlying desires, transcending surface-level expressions to foster more flexible and understanding human-machine interactions.

Contrary to the classic VG task, realizing intention-based grounding involves several unique aspects that require consideration. 1) *Intention-Driven Descriptions*: Since humans tend to provide intention-based expressions to get the desired objects rather than detailing every aspect, it is imperative for intelligent agents to interpret these intention-driven instructions and act accordingly, focusing on the semantic core of the requests rather than their literal descriptions. However, previous studies in this field have primarily concentrated on literal textual descriptions, with scant attention to understanding user intentions. 2) *Egocentric Perspective*: As explored in prior studies (Qi et al., 2020; Kurita et al., 2023; Zhu et al., 2023a; Lee et al., 2023), the practical agents actually receive all visual information from a first-person view. However, most classic VG datasets are predominantly collected from third-person perspective, which greatly deviate from the application contexts of an embodied agent. 3) *Multi-Scene Perception*: Given that agents are expected to navigate and identify specific targets across diverse scenarios in the real world, the capability to perceive and interact within dynamic multi-scene environments is crucial for accurately accomplishing visual grounding task. Yet, most prior research on VG has overlooked this critical aspect, focusing mainly on static, single-scene visual input. In summary, while intention-based VG is highly meaningful for multimodal em-

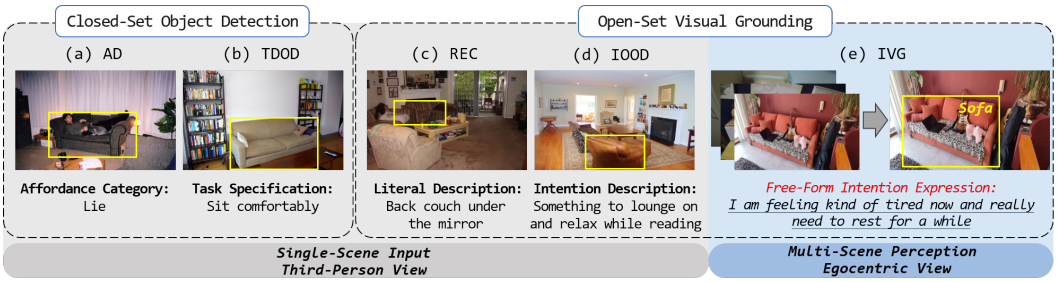

Figure 1: Task Comparison between Affordance Detection (AD), Task-Driven Object Detection (TDOD), Referring Expression Comprehension (REC), Intention-Oriented Object Detection (IOOD) and Intention-driven VG (IVG).

bodied intelligence, there is a notable scarcity of research on visual grounding based on human intentions, and the current lack of relevant data further compounds the challenge of this task.

Therefore, in this work, we attempt to fill this important blank space that has been neglected before and move towards intention-oriented visual grounding. Specifically, based on classic VG, we propose a new intention-driven visual grounding (IVG) task to push towards intention-oriented vision-language understanding, which requires the models to identify the corresponding scene and target object that match the intent expressions from the given multi-scene input. To solve the data scarcity problem of intention-oriented grounding task, we build a large-scale grounding dataset termed IntentionVG which is also the first grounding dataset to support free-form intention expressions. Besides, we also construct several baseline models as straightforward solutions to our proposed IVG task, including both zero-shot & fine-tuning settings and integrated & end-to-end model types. The constructed baselines set the new state-of-the-art (SOTA) performance on our IntentionVG benchmark dataset for IVG task, which leaves further room for achieving performance improvement by future research.

Overall, our main contributions of this work can be summarized as follows:

- We propose a new IVG task (as presented in Fig. 1) and introduce a new setting based on egocentric viewpoint with multi-scene perception to better evaluate the embodied agents' perceiving ability, transcending the classic VG task towards better understanding of human desires in the open-world scenarios.
- We build an intention-oriented grounding benchmark dataset named IntentionVG, which to the best of our knowledge is the largest-scale grounding dataset and also the first grounding dataset that supports free-form intention-based vision-language annotations.
- We develop a series of baseline models under both zero-shot and fine-tuning settings to effectively realize precise vision-language understanding for our IVG task, setting new SOTA performance on our IntentionVG benchmark dataset.

2 Related Work

Classic Visual Grounding is to locate the target object corresponding to the given natural language expression in an image. The two fundamental VG tasks are distinguished by their output form. Referring Expression Comprehension (REC) (Mao et al., 2016; Chen et al., 2018; Deng et al., 2021, 2023; Bai et al., 2023; Zhu et al., 2023b; Chen et al., 2023a,b; You et al., 2023) and Referring Expression Segmentation (RES) (Hu et al., 2016; Ye et al., 2019; Ding et al., 2021; Yang et al., 2022; Wang et al., 2022b; Lai et al., 2023; Zou et al., 2023b,a; Wang et al., 2024) have been well studied by previous works, among which REC is the main focus of this work given its heightened significance. However, previous studies in this field are mainly stuck on the literal description based grounding with a single input image, and the images among previous datasets are typically collected in a third-person perspective. In this work, we pioneeringly propose a new IVG task and an intention-based grounding dataset IntentionVG based on egocentric view and multi-scene perception, pushing towards human intention understanding in the practical scenarios.

Vision-Language Complex Reasoning aims at understanding intricate textual-visual input and accomplishing the vision-language (V-L) tasks based on explicit reasoning, in which broader knowledge and strong expression comprehension ability are essential thus posing a greater challenge compared with conventional V-L tasks. Due to the reasoning ability and rich prior knowledge in large language models (LLMs), numerous methods such as (Pi et al., 2023; Zhao et al., 2023; You et al., 2023; Li et al., 2024; Chen et al., 2024) leverage LLMs to

understand complex instructions. However, prior works mainly focus on literal description reasoning, falling short in understanding potential user intentions. Recently, LISA (Lai et al., 2023) introduces a challenging RES benchmark that incorporates complex language expressions, and RIO (Qu et al., 2023) proposes a new IOOD dataset that includes the specific affordance of the objects. The format of RIO's sentences is *"Something can be used..."*, which fails to describe the intentions directly from the user's perspective but simply describes the affordance of the target objects. Therefore, in this work, starting from a perspective that aligns more closely with real-world scenarios, we for the first time integrate egocentric viewpoint of data collection, multi-scene perception, and free-form expression of human intentions to construct a new IVG task with a corresponding IntentionVG benchmark dataset, thereby enhancing LLMs' reasoning capabilities to better understand human intentions.

3 IVG Task & IntentionVG Dataset

In this section, we first introduce our IVG task's definition (Sec. 3.1) and present collection pipeline for IntentionVG data (Sec. 3.2). Then, the specific details and evaluation metrics about our IntentionVG are provided (Sec. 3.3 and Sec. 3.4).

3.1 IVG Task Description

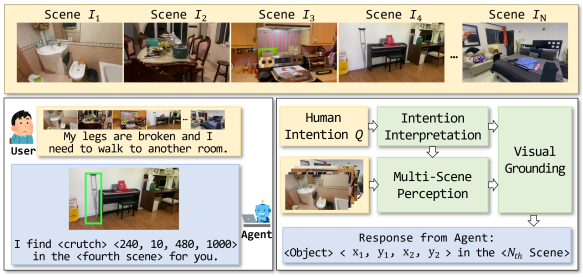

Figure 2: The illustration about the overall pipeline of our intention-driven visual grounding task, which mainly comprises intention interpretation, multi-scene perception and the subsequent visual grounding.

As presented in Fig. 2, the visual-textual input and corresponding ground truth consist of a human intention query Q, a set of scene candidates $I_1, ..., I_N$, a positive scene index N_{th}, and a target bounding box (x_1, y_1, x_2, y_2) together with its object category <Object> in the positive scene image $I_{N_{th}}$. The overall pipeline of our proposed IVG task can be decomposed into two stages. The first stage (i.e., intention interpretation and multi-scenario perception) is aimed at identifying the target scene image $I_{N_{th}}$ from a predefined set of potential scenes that aligns mostly with the given intention expressions, based upon query Q made by users. In this phase, it is imperative for the models to comprehend the textual queries posed by humans in conjunction with the observed visual scenes, and subsequently make correct judgments by returning the correct scene index N_{th}. The second stage (i.e., visual grounding) involves the localization of specific object within the chosen scene image, returning the target bounding box (x_1, y_1, x_2, y_2) and its category tag <Object>. In essence, our new IVG task necessitates the model's proficiency in concurrently understanding both user intention-based requests and multi-scene visual inputs, as well as the models' capability to perform scene selection and visual localization in alignment with the underlying human intentions. The complete response can be organized into the following format:

$$<Object>(x_1, y_1, x_2, y_2) \ in \ <N_{th} \ Scene> \quad (1)$$

3.2 Data Collection Engine

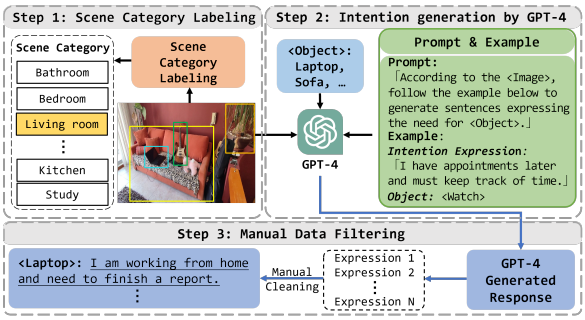

Figure 3: The illustration of data collection engine for IntentionVG. We start by inheriting EgoObjects (Zhu et al., 2023a) data and conduct scene category labeling for each image. Then we feed GPT-4 with V-L input to generate the draft of intention-driven response. At last, we conduct data filtering by manually selecting the well matched expression-bounding box (bbox) pairs.

As shown in Fig. 3, we build our IntentionVG dataset based on the egocentric grounding dataset EgoObjects (Zhu et al., 2023a), inheriting the annotations of object categories and bounding boxes. The entire data collection process involves three steps. Since the practical applications require the agents capable of conducting visual perception among diverse scenes, we first conduct scene category labeling of each image. Specifically, the inherited EgoObjects data is manually annotated with indoor scene categories, resulting in a total of 10 scene classes. In the second step, with the rapid developments of multimodal large language models, we take advantage of GPT-4 (Achiam et al., 2023) to adeptly comprehend the visual relationship between object categories and given images,

generating their associated intention expressions. Utilizing the collected images and object category information, we craft well-designed prompts with examples (i.e., *"According to the <Image>, follow the examples below to generate sentences expressing the need for <object>. Example: [I have appointments later and must keep track of time] for <watch> ..."*) to query GPT-4 for the expected outputs. At last, we manually review and refine the response generated by GPT-4, making sure that the intention expression of each object is objectively aligned with the scene category of the image. Subsequently, based on the proportional distribution within each scene category, we partition our IntentionVG data into training and testing sets with 98,269 and 3,379 images, respectively.

3.3 IntentionVG Dataset Details

Datasets	#Imgs	#Labels	Intentions	#Cats	#Avg Len
Classic Visual Grounding					
ReferIt	20K	97K	–	238	3.2
RefCOCO	20K	50K	–	80	3.6
RefCOCO+	20K	49K	–	80	3.5
RefCOCOg	26K	54K	–	80	8.4
GRES	20K	60K	–	80	3.7
Affordance Detection					
ADE-Aff	10K	26K	Verbs	150	/
PAD	4K	4K	Verbs	72	/
PADV2	30K	30K	Verbs	103	/
COCO-Tasks	40K	64K	Phrases	49	2.6
Intention-Oriented Object Detection					
RIO	40K	130K	Template	69	15.7
Intention-Driven Visual Grounding					
IntentionVG	**100K**	**500K**	**Free-Form**	**1096**	11.2

Table 1: Comparison with classic VG (Kazemzadeh et al., 2014; Yu et al., 2016; Nagaraja et al., 2016; Liu et al., 2023a), AD (Chuang et al., 2018; Luo et al., 2021; Zhai et al., 2022; Sawatzky et al., 2019) and RIO (Qu et al., 2023) datasets. # denotes the specific number, where Intentions, Cats and Avg Len denote the intention expression types, object/affordance categories and average length of expressions. "-", "/" denote the intention-based and non-verb expressions are unavailable.

As detailed in Table 1, prevailing datasets, such as the widely-used benchmark RefCOCO (Yu et al., 2016), exhibit limitations in terms of small data scale and a scarcity of natural language expressions that reflect human intentions. We compare the built IntentionVG dataset with existing benchmark (including the datasets for classic VG, AD, IOOD and our IVG tasks) to highlight the distinct and significant properties of our dataset, as outlined in Table 1. Furthermore, Fig. 8 in Appendix Sec. A.4 showcases the examples from our IntentionVG dataset, while the dataset statistics of our IntentionVG are presented in Fig. 4, Fig. 5 and Fig. 7 in Appendix Sec. A.4, illustrating its data diversity and great potential for practical applications.

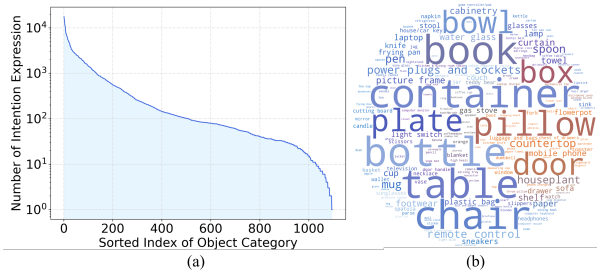

Figure 4: IntentionVG dataset statistics. (a) the number of referring expressions per object's category in the log scale. (b) the word cloud highlights the head categories.

Figure 5: Word clouds of partial categories from our IntentionVG benchmark dataset.

Intention-Driven Descriptions. In comparison to the previous grounding counterparts, our newly built IntentionVG is the first visual grounding dataset covering free-form intention-oriented expressions for each object in the provided images. Compared with the affordance detection counterparts, our IntentionVG dataset transcends the closed-set affordance categories, providing informative and unique intention-driven language expressions for each bbox.

Egocentric View & Multi-Scene Perception. Unlike most existing grounding datasets, our IntentionVG is the first work to incorporate the egocentric perspective and multi-scene perception that are critically needed by multimodal agents in real-world scenarios. The grounding data within IntentionVG are annotated with scene categories, with its training and testing sets specifically designed to support model training and evaluation under multi-scene setting, which moves beyond the typical single-image input of classic VG and more closely aligns with application scenarios.

Breakable Data Scales & Object Categories. To the best of our knowledge, our IntentionVG stands as the largest-scale dataset within the

grounding research community to date. In terms of the number of images, object instances, and referential tokens, IntentionVG significantly outpaces the previous largest classic grounding dataset, RefCOCOg (Nagaraja et al., 2016), multiplying its scale by nearly 4, 9, and 12 times respectively. Meanwhile, it encompasses intention-based expression counts that exceed the largest existing AD dataset COCO-Tasks (Sawatzky et al., 2019) and IOOD dataset RIO (Qu et al., 2023) by 8, 4 times separately. Featuring 1096 object categories and nearly 500K expressions about user intentions, IntentionVG spans a broader spectrum of multimodal knowledge, marking a pivotal advancement in the pursuit of open-world intention understanding.

More Complex and Free-Form References. Benefiting from the powerful GPT-4 (Achiam et al., 2023) and our carefully crafted prompt template, the reference expressions of IntentionVG dataset are enriched with visual context to capture human intentions more effectively. Without sticking to a rigid template (e.g., [Something to ...] format in RIO (Qu et al., 2023)), our IntentionVG allows for the diverse intentions behind interacting with various target entities to be articulated and emphasized through flexible natural language expressions.

3.4 Evaluation Metrics

To advance the IVG task's applicability in practical scenarios, we have introduced two grounding task settings based on the quantity of images provided: single-scene and multi-scene grounding. We have also tailored evaluation metrics for each setting, enabling a thorough assessment of model performance across different contexts.

Single-Scene. In this setting, the model's grounding capability is assessed with just one provided image. We use Precision@0.5 (P@0.5) as the metric to evaluate models' grounding performance. This measure reflects the model's ability to correctly identify the target object in alignment with the user's intention with its top one prediction. A prediction is considered correct if the Intersection over Union (IoU) between the predicted and the ground truth (GT) bbox exceeds a threshold of 0.5 (i.e., $threshold > 0.5$), indicating a significant overlap and, hence, an accurate localization result.

Multi-Scene. Since the intelligent agents need to move and search for the expected targets among different scenarios in practice, it is vital for embodied agents to accomplish our IVG task based on multi-scenario perception in the first-person view. To assess the accuracy of models in multi-scene perception and the subsequent VG, we utilize the metrics of Recall@1 (R@1) and Precision@0.5 (P@0.5). Recall@1 measures the ratio of cases accurately identified within the top-1 perception result to the overall count of cases in the test set, reflecting the model's precision in pinpointing the most intention-relevant image from a multitude of scenes. Additionally, we introduce Precision@0.5 given Recall@1 correct cases (P@0.5|R@1) as an evaluation metric to gauge the grounding performance of models specifically for those cases correctly identified in the multi-scene perception phase. These employed metrics ensure a nuanced understanding of the models' effectiveness in accurately grounding objects across multiple scenes.

4 Baseline Construction

To realize our IVG task based on multi-scene perception and egocentric viewpoint, we formulate two different kinds of baseline models below. More illustrations of the baseline structures can be found in Appendix Sec. A.3 for better understanding.

4.1 Zero-shot Setting

For zero-shot setting, it is intuitively to follow a two-step integrated approach to realize the multi-scene perception and VG for IVG task. We initially employ EVA-CLIP (Sun et al., 2023) as perceiver to extract multi-scene representations and features of intention expressions, followed by feature similarity matching to select the scene that best matches the textual expression. Subsequently, the well matched single scene is fed to the grounding model with the corresponding intention expression to deduce the associated bounding box. The adopted grounding models can be categorized into two types: specialists and generalists. Specialist baselines are the SOTA methods designed for classic VG task, including MDETR (Kamath et al., 2021), SeqTR (Zhu et al., 2022) and Polyformer (Liu et al., 2023b), which are trained on VG datasets. Generalist baselines comprise models capable of handling various V-L tasks that are trained on large-scale visual question answering (VQA) and VG datasets, such as LLM-based method Shikra (Chen et al., 2023b) and Mini-GPTv2 (Chen et al., 2023a). We also incorporate LLMs (e.g., GPT-4 (Achiam et al., 2023)) as the optional interpreter which could translate intention expressions into explicit object

descriptions, helping the following perceiver and grounding model better solve our IVG task.

4.2 Fine-tuning Setting

For fine-tuned baselines, we incorporate both integrated and end-to-end models. For the two-step integrated baselines, we first utilize contrastive learning loss to fine-tune EVA-CLIP (Sun et al., 2023) with the annotations of scene categories, enhancing its capability of discerning the similarity between scenes and intention expressions. Then we perform fine-tuning on generalist grounding models using their respective training prompts for grounding task, followed by putting together the fine-tuned scene perceiver and grounding models.

The end-to-end baseline model is built upon generalist Qwen-VL (Bai et al., 2023). Since only Qwen-VL possesses the capability to accommodate multiple images as inputs, we introduce a well-designed prompt that facilitates the simultaneous input of multiple scenes and intention expressions and build the end-to-end baseline upon Qwen-VL. Through instruction tuning under cross entropy loss, our fine-tuned Qwen-VL can concurrently conduct scene perception and grounding.

To be noticed, since target object matching the intent description may appear in multiple scenes simultaneously, to ensure that only the positive scene contains the corresponding object, we impose a strong constraint on the input scenes to force that only one of them exists the object. Besides, two hyper-parameters including input scene number N and multi-scene occurrence rate α are introduced during fine-tuning. A higher number N of input scenes implies that the baseline models need to identify the most relevant scene and target object from a larger set of images during fine-tuning, thus increasing the difficulty of the training objective. Besides, a higher α value means that the baseline models are more likely to encounter multi-scene input samples during fine-tuning, as opposed to the classic VG's typical single-image input. α=0 or 1.0 represents extreme circumstances during fine-tuning, where the model is exclusively fed either single-scene or multi-scene input samples.

5 Experiments

To evaluate the effectiveness and the designing rationale of our data and baseline models, comprehensive experiments are conducted on our built IntentionVG dataset for the new IVG task.

5.1 Implementation Details

Our work is implemented based on Pytorch (Paszke et al., 2019) and trained with 8 NVIDIA A800 GPUs. The original model weights of all the adopted baselines are inherited and integrated for the subsequent fine-tuning and evaluations under the zero-shot setting. For the baseline constructions under fine-tuning setting, we either directly fine-tune the end-to-end baseline model (i.e., Qwen-VL (Bai et al., 2023)) or put together the separately fine-tuned intention interpreter, multi-scene perceiver and grounding models for the integrated baseline construction. Taking EVA-CLIP (Sun et al., 2023) as scene perceiver, we introduce the annotations of bounding boxes and scene categories to respectively tune the grounding models and EVA-CLIP. Training details about fine-tuning are presented in Table 7 in Appendix Sec. A.1.

5.2 Main Results and Analysis

To comprehensively evaluate the intention-oriented grounding performance of all the constructed baseline models for our new IVG task, we conduct experimental comparison on the newly built IntentionVG testing set. As illustrated in Table 2, the baseline models can be categorized into two settings based on the evaluation manner: zero-shot and fine-tuning, and into two types based on the framework structure: integrated and end-to-end. For a fair comparison, we re-implement these SOTA methods and report their performance on our IntentionVG testing set. It is evident from Table 2 that the zero-shot setting baselines, having not been exposed to our intention-driven grounding data, struggle to comprehend user intentions and identify corresponding targets, resulting in generally lower performance compared to the fine-tuned baselines. With EVA-CLIP serving as the multi-scene perceiver, neither specialists nor generalists classic VG SOTA methods can effectively address our IVG task directly, particularly in the multi-scene setting. In contrast, baselines under the fine-tuning setting, both integrated and end-to-end types, have witnessed significant performance improvements on our IVG task with the support of IntentionVG data. Additionally, the introduction of the LLM-based intention interpreter has substantially enhanced the performance of zero-shot baselines on the IVG task, underscoring the critical importance of genuinely comprehending user intentions for accurately accomplishing the intention-driven grounding task.

Methods	Framework Type	Intention-Driven Visual Grounding			
		Single-Scene	Multi-Scene		
		P@0.5	R@1	P@0.5|R@1	P@0.5
Zero-shot Setting					
Perceiver + MDETR (Kamath et al., 2021)	Integrated	14.23	54.00	16.48	8.90
Perceiver + SeqTR (Zhu et al., 2022)	Integrated	9.38	54.00	9.38	5.07
Perceiver + Polyformer (Liu et al., 2023b)	Integrated	16.50	54.00	18.64	10.07
Perceiver + OFA (Wang et al., 2022a)	Integrated	18.57	54.00	18.57	10.03
Perceiver + Shikra (Chen et al., 2023b)	Integrated	22.45	54.00	24.44	13.20
Perceiver + Ferret (You et al., 2023)	Integrated	21.80	54.00	24.19	13.06
Perceiver + LISA (Lai et al., 2023)	Integrated	15.32	54.00	17.73	11.08
Perceiver + Qwen-VL (Bai et al., 2023)	Integrated	20.72	54.00	22.71	12.27
Perceiver + MiniGPT-v2 (Chen et al., 2023a)	Integrated	13.04	54.00	14.51	7.84
Interpreter + Perceiver + Shikra (Chen et al., 2023b)	Integrated	37.55	62.63	42.35	26.45
Interpreter + Perceiver + MiniGPT-v2 (Chen et al., 2023a)	Integrated	42.57	62.63	45.80	28.68
Fine-tuning Setting					
Perceiver + Shikra (Chen et al., 2023b)	Integrated	47.19	62.22	50.43	31.38
Perceiver + MiniGPT-v2 (Chen et al., 2023a)	Integrated	44.18	62.22	49.92	31.06
Qwen-VL (Bai et al., 2023)	End-to-End	**50.58**	**74.13**	**53.02**	**39.30**

Table 2: Comparisons with the classic VG SOTA approaches on our ItentionVG testing set. "Perceiver" and "Interpreter" separately denote the EVA-CLIP model and GPT-4 employed for multi-scene perception and user intention understanding. Besides, "Integrated" and "End-to-End" respectively refer to the baseline is a grounding model combined with Perceiver or Interpreter and the single grounding model as an end-to-end structure.

Due to the significantly higher difficulty of our IVG task's multi-scene setting compared to single-scene counterpart, the multi-scene accuracy value is correspondingly lower than the other one. This further emphasizes the vital importance of researching intention-oriented visual grounding where previous SOTA methods have fallen short.

5.3 Ablation Study

To justify the efficacy of our IntentionVG's data, we conduct extensive ablation experiments on IntentionVG testing set. As illustrated in 3.4, the tables below involve both the traditional single-scene and harder multi-scene settings. For all ablations, Qwen-VL (Bai et al., 2023) stays employed as our baseline and the multi-scene input is adopted as 5 images for fair evaluation. Except for the 1st ablation, 10% of our IntentionVG data is used to fine-tune the baseline for ablations. Noticeably, more ablations can be found in Appendix Sec. A.2.

Effect of Data Scale. First, we investigate the effect of different percentages of introduced training samples in our IntentionVG dataset. The results are shown in Table 3. It is clear in Table 3 that model performance under both settings for our intention-oriented grounding task is consistently improved with more and more training samples, validating the efficacy and high quality property of our collected data. Since the original Qwen-VL can not handle the multi-scene grounding task at all, when

Ratios	Single-Scene	Multi-Scene		
	P@0.5	R@1	P@0.5|R@1	P@0.5
0%	20.72	0.00	0.00	0.00
10%	46.01	70.64	48.96	34.58
25%	47.63	73.46	50.67	37.22
50%	48.75	73.82	51.66	38.14
100%	**50.58**	**74.13**	**53.02**	**39.30**

Table 3: Ablation study on effect of data scale.

ratios=0 (i.e., without employing our intention-based data) the corresponding performance is much poor. As the ratios of employed training samples continually rise, there's no sign of diminishing performance gains, suggesting that our dataset has great potential to help multimodal large language models better understand human intentions with consistently scaled up training data.

Scenes (N)	Single-Scene	Multi-Scene		
	P@0.5	R@1	P@0.5|R@1	P@0.5
1	45.51	19.37	1.34	0.26
3	45.35	51.17	37.52	19.20
5	46.01	**70.64**	**48.96**	**34.58**
8	**46.14**	70.32	40.90	28.76
10	45.99	60.97	32.66	19.91

Table 4: Ablation study on effect of scene number N.

Effect of Scene Number. Furthermore, we explore the impact of varying the input scene number during training. Scene number N implies that the baseline model need to identify the most relevant scene and target object from a set of given images during fine-tuning. As shown in Table 4, the model achieves optimal performance with an intermediate scene number of 5 under both settings. As

scene number gradually increases from 1 to 10, the impact on the model's single-scene performance remains minimal. However, under multi-scene setting, the model's ability to perform multi-scene perception and subsequent grounding initially improves and then diminishes. We believe this pattern occurs because, with few input scenes at the start, the model's learned capability in perception and grounding during tuning phase is weak. As the number of scenes increases, the model's related abilities enhance. Yet, beyond the sweet spot N=5, the model becomes overwhelmed due to the excessive number of input scenes, leading to confusion and an inability to learn the effective representations to accurately complete our IVG task.

Rates (α)	Single-Scene	Multi-Scene		
	P@0.5	R@1	P@0.5|R@1	P@0.5
0	45.70	19.71	1.48	0.29
0.25	44.57	46.41	38.81	18.02
0.5	45.54	70.61	48.58	34.31
0.75	45.92	68.84	48.43	33.34
0.9	**46.01**	70.64	**48.96**	34.58
1.0	42.77	**71.68**	48.45	**34.73**

Table 5: Ablation study on the effect of multi-scene Occurrence rate during fine-tuning.

Effect of Multi-Scene Occurrence Rate. Additionally, we delve into the impact of the multi-scene occurrence rate, denoted as α. α means that the possibility of baseline models to encounter multi-scene input samples during fine-tuning. It is evident from Table 5 that when α=0, the model performs poorly on our multi-scene IVG task due to the lack of multi-scene samples during tuning. A larger α helps the model achieve better grounding performance based on multi-scene perception. However, an excessively high α value, specifically α=1.0, significantly diminishes the performance in intention-driven grounding tasks with single-image inputs. Therefore, α=0.9 is set as our default setting.

Supervision Types		Single-Scene	Multi-Scene		
Scene	Object	P@0.5	R@1	P@0.5|R@1	P@0.5
✘	✘	45.89	19.84	1.63	0.32
✔	✘	44.76	59.46	45.65	27.15
✘	✔	45.70	19.37	1.34	0.26
✔	✔	**46.01**	**70.64**	**48.96**	**34.58**

Table 6: Ablation study on effect of GT formality.

Effect of Supervision Formality. At last, we exploit the effect of the GT formality for supervision during training. The supervision signal includes four different settings, which are only grounding bbox, bbox + scene category, bbox + object tag and bbox + scene category + object tag (the 1st to 4th row in Table 6). As shown in Table 6, with the bbox, scene category, and object tag serving as the most comprehensive supervision targets, the model gains access to the fullest extent of information and establishes explicit associations between inputs and GTs, thereby achieving best performance. The inclusion of scene category and object tag as part of the supervision signal respectively enhances the model's multi-scene perception and open-domain object recognition capabilities. Thus, removing either of these two parts leads to performance decline on our IVG task. From the 1st and 3rd rows we can observe that introducing the object tag as an additional part of GT on top of bbox does not result in performance improvement. We believe this is because, without the guidance of the scene category, the model becomes very confused during fine-tuning and fails to learn a clear mapping from multi-scene inputs to the target scene and subsequent grounding results. Consequently, the performance of these two rows remains poor on the multi-scene grounding setting.

6 Conclusion and Broader Impact

In this paper, we move beyond previous works that focused solely on literal description based grounding and take a step further to intention-driven V-L understanding. Specifically, by considering that the practical agents need to move and search for expected targets among different scenarios, we put forward a new IVG task. The IVG task requires agents to interpret user intentions and locate specific targets based on egocentric view and multi-scene perception. To enable existing models to better accomplish our IVG task and assess their capabilities on it, we build the first also the largest-scale intention-driven grounding dataset termed IntentionVG with free-form expressions and develop a series of baseline to accomplish the IVG task.

Comprehensive experiments conducted on our Intention dataset demonstrate that most previous methods struggle to directly understand users' non-literal intent expressions and locate the expected targets. With the aid of our data, there is a significant enhancement in the ability to comprehend intentions for IVG task, but there remains substantial room for improvement, warranting further investigations by the research community. Aspiring to foster future research into intention-oriented grounding and inspire new research in this direction, we plan to release our newly built IntentionVG dataset and baseline models to the community.

Limitations

While this work significantly advances the classic visual grounding task towards user intention-based grounding that aligns more closely with real-world applications, it is not without limitations, which leaves opportunities and room for future research. One potential limitation of this work is the scale of the IntentionVG benchmark dataset and the capacity of the developed baseline models could both be expanded and boosted to enhance performance on our proposed intention-driven visual grounding task. Additionally, this research primarily focuses on the referring expression comprehension and generating bounding boxes as grounding output results, indicating that while the built baseline frameworks showcase new intention-driven grounding capabilities, they currently can not generate dense segmentation masks for the target objects. However, integrating with visual foundation models, such as the Segment Anything Model (Kirillov et al., 2023) and its variations, could equip our framework with the ability to produce the dense referring segmentation masks, effectively addressing this shortcoming. This possibility opens a new avenue for research, aiming to create a more versatile and powerful framework that leverages both large language models and visual foundation models to interact with user-provided intention-driven natural language expressions and visual perception inputs.

References

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

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee K Wong. 2024. Mapgpt: Map-guided prompting for unified vision-and-language navigation. *arXiv preprint arXiv:2401.07314*.

Jun Chen, Deyao Zhu1 Xiaoqian Shen1 Xiang Li, Zechun Liu2 Pengchuan Zhang, Raghuraman Krishnamoorthi2 Vikas Chandra2 Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

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

Xinpeng Chen, Lin Ma, Jingyuan Chen, Zequn Jie, Wei Liu, and Jiebo Luo. 2018. Real-time referring expression comprehension by single-stage grounding network. *arXiv preprint arXiv:1812.03426*.

Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. 2018. 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.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779.

Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yanyong Zhang, Houqiang Li, and Wanli Ouyang. 2023. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. 2021. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. 2023. Physically grounded vision-language models for robotic manipulation. *arXiv preprint arXiv:2309.02561*.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. 2016. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. 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.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Shuhei Kurita, Naoki Katsura, and Eri Onami. 2023. Refego: Referring expression comprehension dataset from first-person perception of ego4d. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15214–15224.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Clarence Lee, M Ganesh Kumar, and Cheston Tan. 2023. Determinet: A large-scale diagnostic dataset for complex visually-grounded referencing using determiners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20019–20028.

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

Chang Liu, Henghui Ding, and Xudong Jiang. 2023a. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601.

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

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

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. 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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. 2023. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991.

Mengxue Qu, Yu Wu, Wu Liu, Xiaodan Liang, Jingkuan Song, Yao Zhao, and Yunchao Wei. 2023. Rio: A benchmark for reasoning intention-oriented objects in open environments. *arXiv preprint arXiv:2310.17290*.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175*.

Johann Sawatzky, Yaser Souri, Christian Grund, and Jurgen Gall. 2019. 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.

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, pages 492–504. PMLR.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Wenxuan Wang, Xingjian He, Yisi Zhang, Longteng Guo, Jiachen Shen, Jiangyun Li, and Jing Liu. 2024. Cm-masksd: Cross-modality masked self-distillation for referring image segmentation. *IEEE Transactions on Multimedia*.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022b. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. 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.

Wei Zhai, Hongchen Luo, Jing Zhang, Yang Cao, and Dacheng Tao. 2022. One-shot object affordance detection in the wild. *International Journal of Computer Vision*, 130(10):2472–2500.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. 2022. Seqtr: A simple yet universal network for visual grounding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 598–615. Springer.

Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. 2023a. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20110–20120.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. 2023a. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. 2023b. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.

A Appendix

In this appendix, we provide the following items:

- (Sec. 1) More implementation details on the built IntentionVG dataset for IVG task.
- (Sec. 2) More ablation studies on our newly built IntentionVG dataset for our IVG task.
- (Sec. 3) More illustrations about the constructed baselines for our IVG task.
- (Sec. 4) More visualizations of the dataset statistics and samples from our IntentionVG dataset.

A.1 Implementation Details

Configuration	Fine-tuning	
	Perceiver	Grounding Model
Optimizer	LAMB	AdamW
Base Lr	0.0005	0.00001
Weight Decay	0.05	0.1
Batch Size	2048	32
Lr Decay Schedule	cosine	cosine
Training Epochs	25	1

Table 7: Training settings on our IntentionVG dataset.

The specific training hyper-parameter configurations for fine-tuning all the baseline models on our newly constructed IntentionVG dataset can be found in Table 7.

A.2 More Ablation Studies

Expression Forms	Single-Scene	Multi-Scene		
	P@0.5	R@1	P@0.5	R@1 P@0.5
Object Tag	14.39	30.71	21.57	6.62
[Something to...]	44.70	66.24	46.74	30.96
[I want to...]	43.80	47.24	40.10	18.95
Free Form	**46.01**	**70.64**	**48.96**	**34.58**

Table 8: Ablation study on the effect of expression form in our IntentionVG dataset.

Effect of Expression Form. We also probe into the effect of our grounding data's expression form. As presented in Table 8, the baseline model achieves the best result under both settings, by learning associations between more freely enriched expressions and target objects. During the fine-tuning phase, relying solely on the target object's tag as textual description can lead to an excessive dependence on direct target description, significantly reducing accuracy on our IVG task. Besides, either changing the expression form of our IntentionVG data from free formality to the fixed counterpart (i.e., following the fixed template [Something to...] or [I want to...] to describe affordance or intention) will result in an considerable accuracy decrease across the two settings, which greatly impedes the potential for practical applications.

A.3 Baseline Structures

We have also provided more illustrations about the constructed baselines of two types (i.e., integrated and end-to-end) in Fig. 6.

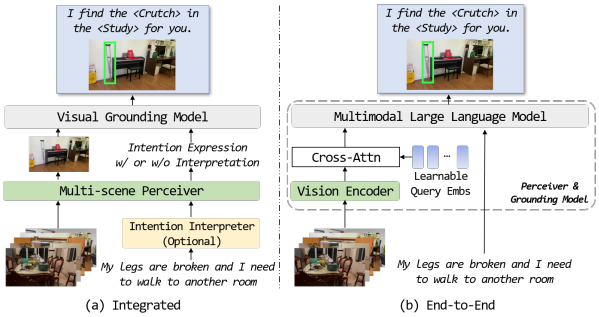

Figure 6: The illustration about the overall pipeline of our built baseline models with both integrated and end-to-end structures for the proposed IVG task.

A.4 More Visualization Results

More IntentionVG Dataset Statistics. More data statistics information about our newly built IntentionVG dataset are presented in Fig. 7.

Samples of IntentionVG Dataset. A few examples in our newly built IntentionVG dataset for intention-driven visual grounding task are presented in Fig. 8.

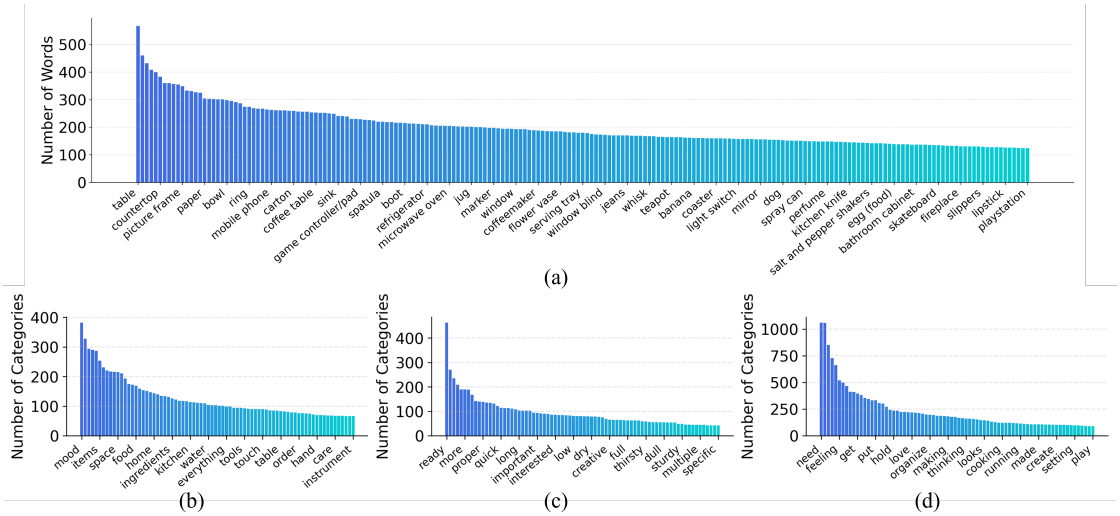

Figure 7: Our IntentionVG dataset statistics. (a) shows the statistics of the word diversity in intention descriptions for each category, and (b), (c), (d) separately present the occurrence frequency of a noun, adjective, verb in different categories of intention description. The horizontal coordinates for (a), (b), (c) and (d) are respectively the examples of the specific categories, nouns, adjectives and verbs with the ranked top 200, 75, 75 and 75 highest vertical values.

Figure 8: Visualizations of samples from our IntentionVG benchmark dataset.