2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Towards Unified Interactive Visual Grounding in The Wild

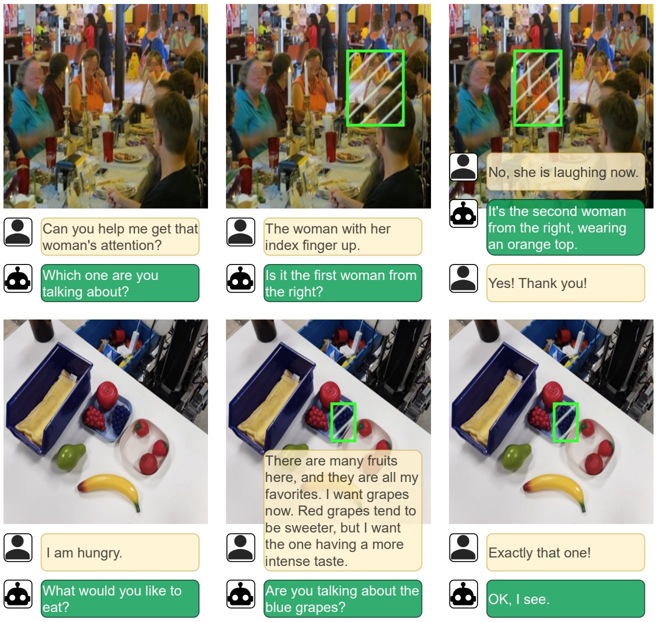
Jie Xu†[1,2], Hanbo Zhang*[2], Qingyi Si[2], Yifeng Li[2], Xuguang Lan*[1], and Tao Kong[2]

***Abstract*— Interactive visual grounding in Human-Robot Interaction (HRI) is challenging yet practical due to the inevitable ambiguity in natural languages. It requires robots to disambiguate the user's input by active information gathering. Previous approaches often rely on predefined templates to ask disambiguation questions, resulting in performance reduction in realistic interactive scenarios. In this paper, we propose TiO, an end-to-end system for interactive visual grounding in human-robot interaction. Benefiting from a unified formulation of visual dialog and grounding, our method can be trained on a joint of extensive public data, and show superior generality to diversified and challenging open-world scenarios. In the experiments, we validate TiO on GuessWhat?! and InViG benchmarks, setting new state-of-the-art performance by a clear margin. Moreover, we conduct HRI experiments on the carefully selected 150 challenging scenes as well as real-robot platforms. Results show that our method demonstrates superior generality to diversified visual and language inputs with a high success rate. Codes and demos are available on https://jxu124.github.io/TiO/.**

I. INTRODUCTION

Robots are increasingly entering our daily lives. To work with non-expert users, they are required to understand the world visually and interact in languages. Nevertheless, ambiguities and uncertainties are ubiquitous. Therefore, robots need to actively seek help from humans to make informed decisions. This paper focuses on interactive visual grounding tasks [1], in which robots ask questions to disambiguate and locate targets by interaction, as shown in Fig. 1. It poses challenges of 1) how to generate informative questions to gather information based on dialog history and visual observations, and 2) how to ground target objects given much longer dialogues than visual grounding tasks [2], [3]. Previous approaches usually rely on predefined interactive templates for disambiguation [1], [4], [5], which usually requires users to follow rules to interact and are less flexible and robust against open-world scenarios.

To address this issue, we present Three-in-One (TiO), a unified model towards end-to-end and robust interactive visual grounding for HRI in the open world. TiO is trained on a joint of extensive public data, and show superior generality to diversified and challenging open-world scenarios. To do so, our model differentiates sub-tasks with different prompts and a shared vocabulary including (object-based) caption generation, (multi-turn) visual question-answering, (multi-turn) visual question generation, visual dialog, and interactive visual grounding. On this basis, all sub-tasks are unified in a sequence-to-sequence formulation, i.e., outputs are auto-regressively predicted with inputs and previous outputs. As a result, TiO can respond to different prompts accordingly for different tasks.


Can you help me get that woman's attention?

Which one are you talking about?

The woman with her index finger up.

Is it the first woman from the right?

No, she is laughing now.

It's the second woman from the right, wearing an orange top.

Yes! Thank you!

I am hungry.

What would you like to eat?

There are many fruits here, and they are all my favorites. I want grapes now. Red grapes tend to be sweeter, but I want the one having a more intense taste.

Are you talking about the blue grapes?

Exactly that one!

OK, I see.

Fig. 1. TiO in the wild (top row) and the realistic interactive robot manipulation tasks (bottom row). Each image along with the corresponding round of in the dialog shows that TiO can ask informative questions based on previous dialog history and complex observations, while maintaining explainable internal states to evaluate the grounded candidates (green box) of the target.

We evaluate TiO from three aspects: 1) on standard interactive visual grounding benchmarks, 2) on human-robot interaction with challenging and diversified scenes, and 3) on the real-robot manipulation platform. For 1), we perform experiments on GuessWhat?! [8] and the more challenging InViG [9], where the images are posed with ambiguous instructions and robots are required to ask questions for disambiguation. Results show that TiO achieves new state-of-the-art performance of 65.5% and 78.1% accuracy, respectively. Besides, we comprehensively and empirically analyze the superiority of TiO by extensive ablation studies. For 2), we set up a new challenging benchmark including 150 examples for HRI, to evaluation TiO in the aspect of diversified visual inputs, understanding attributes and behaviors of humans, and open-ended instructions. We show that TiO outperforms all baseline methods by a clear margin, and demonstrate superior generality to open-world applications. For 3), we deploy TiO on two real-robot manipulation platforms, one for desktop

---

[1] Xi'an Jiaotong University, [2] ByteDance Research

† This work was done during an internship at ByteDance.

* Correspondence to: Hanbo Zhang {zhb@bytedance.com} and Xuguang Lan {xglan@mail.xjtu.edu.cn}

979-8-3503-8457-4/24/$31.00 ©2024 IEEE

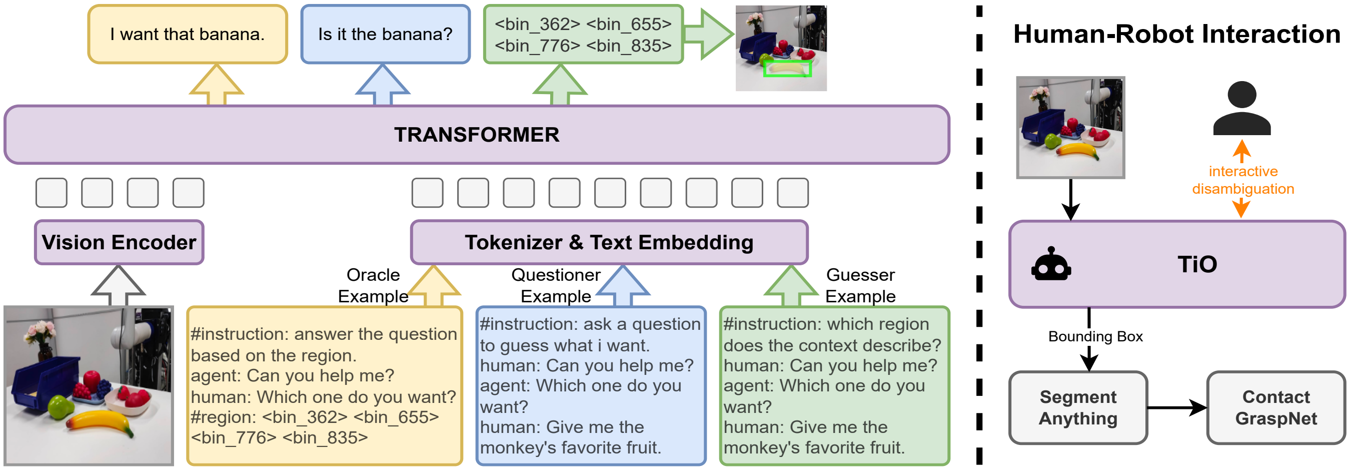


Fig. 2. Overview of TiO. Left: TiO network, which is a visual-language interactive disambiguation model that can interact with humans through natural language for disambiguation. It unifies the Questioner, Oracle, and Guesser in a single transformer with different instructions. Right: TiO deployed on interactive manipulation robots. In our interactive manipulation system, TiO provides the target object's bounding box based on the disambiguation by interaction with the human user then converts it into a segmentation map using Segment Anything [6]. Contact GraspNet [7] finally generates the best grasp based on the projected point clouds.

scenarios following similar settings of [1], [5], and another for mobile manipulation tasks. Results illustrate that TiO achieves robust performance on real-robot applications.

II. RELATED WORKS

A. *Interactive Visual Grounding*

For decision-making in the wild, it is crucial for robots to interact actively when necessary and avoid ambiguity-induced mistakes. Classical methods in HRI often rely on predefined corpora to interact actively [10]–[14]. These methods are usually designed for specific tasks and scenarios, limiting their application in real-world interactive scenes. As deep learning arises, recent advances [15]–[17] have shown promise to generate open-ended free-form texts for HRI, which have led to investigations [1], [4], [5], [18] of open-ended interaction scenarios in real-world settings. Nevertheless, the interactions still exhibit limited diversity and naturalness due to a restricted number of templates, making interaction less flexible and even confusing. To address these issues, recent research has introduced large-scale interactive datasets for more natural HRI [9], [19], [20]. Specifically, GuessWhat?! [20] focuses on interactive visual disambiguation, where the robot has a visually grounded dialog with humans to identify a target object by iteratively asking judgment questions. Nevertheless, methods based on GuessWhat?! still suffer from limited interaction patters, setting restrictions for the users to only answer 'yes' or 'no'. Recently, InViG [9] offers more challenging scenarios with 21k open-ended free-form human-human dialogues for interactive visual grounding, which have paved the way for more natural and flexible interactive visual grounding in open-world scenarios.

B. *End-to-End and Unified HRI*

With developments in deep learning, visual-language models (VLMs) [21]–[31] are more and more prevalent to be the solution of visual-language tasks including Visual Question-Answering (VQA) [32], Visual Grounding [2], Image Captioning [33], etc. Nevertheless, these models usually cannot adapt to multi-turn interaction directly. Recently, large language models have shown promise in language interaction, with impressive few-shot and zero-shot generality to open-ended dialogues [34]–[41]. Taking advantage of these models, recent works have shown promising results in adapting them to multi-modal inputs [28], [42]–[46]. Nevertheless, these models are still insufficient to address multi-turn dialogues and follow open-ended instructions in interactive visual grounding, which requires the model to reason from dialog history and visual observations to generate informative questions, and moreover, guess the target objects.

III. PRELIMINARIES

Interactive visual grounding means grounding targets visually with ambiguous instructions by iteratively asking questions to gather information. Formally, a dataset of interactive visual grounding can be denoted by $D = \{I_i, H_i, G_i\}_{i=1}^{S}$, where I_i, H_i and G_i represent the input image, dialog history, and target bounding box of $i-$th sample, respectively. S is the number of samples in total. For each sample, G_i is often represented using a 4-d tuple $< x_1, y_1, x_2, y_2 >$ indicating the coordinates of upper-left and bottom-right corners. The dialog history $H_i = \{Q_i^j, O_i^j\}_{j=1}^{J}$ is a J-round conversation between the Questioner and the Oracle. The Questioner is simultaneously the Guesser, who is to ask questions to guess the target of the Oracle. The Oracle is the one who knows the intended object and is responsible for response (with O_i^j) to the utterances of the Questioner.

During the inference of a model for interactive visual grounding, previous works [1], [4], [5] usually rely on a real person to play the role of the Oracle and evaluate the success rate of visual grounding through multi-turn interaction. Alternatively, an Oracle model can be trained on the same dataset [8] for automatic evaluation.

IV. Method

As shown in Fig. 2, our TiO is a unified transformer for all visual-language sub-tasks that ensemble interactive visual grounding. To do so, we 1) unify training on datasets from image captioning, visual question-answering (VQA), visual grounding (VG), and visual question generation (VQG); 2) unify prompts and predictions for multiple tasks; 3) unify the encoding and decoding of texts and bounding box coordinates in the tokenizer. Therefore, during inference, with the corresponding prompt as inputs, TiO can play the role of the Guessor, Oracle, or Questioner, with superior performance compared to baseline methods. Combined with robot grasping models (e.g. Segment Anything [6] + Contact GraspNet [7]), TiO can be deployed on the real-robot platform for interactive manipulation tasks robustly with natural language inputs.

A. *TiO Network*

a) Backbone: Considering the successful practice of encoder-decoder Transformer architecture in generative tasks (e.g., dialog generation, visual grounding) [26], [30], [47], [48], we have chosen it as the backbone architecture for TiO, as shown in Fig. 2. Our encoder-decoder transformer model within the vision encoder has a parameter capacity of ∼930M. It consists of 24 encoder layers and 12 decoder layers. The input of the encoder includes both visual and linguistic tokens from vision encoder and text tokenizer. The decoder is the predictor for all tasks by auto-regressively generating tokens.

b) Vision Embedding: The scaled image (target resolution is 512×512) is directly used as the input of the vision encoder, then converted into patch features with a 32×32 grid. Through a learnable linear projection layer, it is then converted into 1024 image embeddings in the hidden space, which are directly fed into the transformers concatenated with text embeddings.

c) Text Input/Output: Text tokenization uses BartTokenizer along with adding 1000 location tokens for bounding box prediction. Formally, each location token is represented by "<bin_i>", where $i \in \{0, 1, ..., 999\}$. The tokenization of object bounding boxes is as follows: each bounding box (x_1, y_1, x_2, y_2) is firstly normalized and then mapped into the range of $[0, 1000)$. For example, $(0.0, 0.12, 0.3, 0.4)$ can be converted to "<bin_0> <bin_120> <bin_400> <bin_300>", and then encoded as location tokens. We utilize the identical position encoding as [30]. Additionally, we combine the image embeddings (red blocks in Fig. 2) and text embeddings (blue blocks in Fig. 2) together through a simple concatenation. It serves as the input for the transformer's encoder. The transformer is a sequential-to-sequential architecture, and its decoder predicts one token at a time auto-regressively. To perform the generation, we employ the beam-search method with a beam width equal to 5.

B. *Training*

a) Datasets: We have shown the datasets involved in training TiO in Table I. Apart from GuessWhat?! [8] and InViG [9] that are inherently suitable for interactive visual grounding, we also collect datasets related to each of the sub-tasks, including Image Captioning (Caption), VQA, Visual Question Generation (VQG), and Visual Grounding (VG). For abbreviation, we use Dialog to represent the tasks VQA+VQG+Caption+VG. Concretely, for visual grounding, we collect data from RefCOCO [2], and OpenImages [51]. All target objects are discretized into integer tokens introduced in Section IV-A. For image captioning, we collect the Mini-GPT4 Caption [49], [50], which consists of 3.4K long captions for images. For dialog, we use LLaVA [44] instructions, and Visual Dialog [19]. They comprise more than 350K high-quality dialogues and are useful for learning multi-turn language modeling. Totally, our data includes 955K unique samples based on 135K unique images during the training phase.

TABLE I

STATISTICS ON THE DATASETS OF TRAINING TASKS

Source	Task	#Image	#Sample
SBU Captions [49], [50]	Caption	3.4K	3.4K
LLaVA [44]	VQA	195.4K	257.2K
VisDial [19]	Dialog	103.4K	103.4K
GuessWhat?! [8]	VG + Dialog	40.1K	68.7K
InViG [9]	VG + Dialog	17.7K	18.1K
RefCOCO [2]	VG + Caption	12.7K	31.7K
RefCOCOg [2]	VG + Caption	17.4K	33.1K
RefCOCO+ [2]	VG + Caption	12.7K	31.6K
OpenImages [51]	VG	28.0K	28.0K

Notably, one important issue is to train the model to stop conversation after rounds of interaction when information is gathered enough. To train the model to automatically generate stop signs, we introduce the question "Can you specify the target object?" during training manually. The training data is randomly sampled from GuessWhat?! and InViG. If the sampled dialogue is the last round, the answer will be "yes", otherwise, "No". During the inference, we will iteratively and automatically ask the model to answer this question and stops the conversation if the answer is "Yes".

b) Unified Multi-Task Formulation: Benefiting from the unified transformer architecture introduced in Section IV-A, it is straightforward to unify all sub-tasks and datasets during training and inference. Concretely, we differentiate sub-tasks using different prompts as shown in Fig. 2. All tasks are formulated with images and texts as inputs and text-only predictions. Therefore, our network is trained end-to-end simply using Cross-Entropy loss to maximize the likelihood of the next ground truth token:

$$loss = \sum_{l=1}^{L} - \log \left[P(w_l | w_{<l}) \right] \quad (1)$$

where w_l is l-th ground truth token for the prediction, $w_{<l}$ is the ground truth token sequence before w_l, and L is the length of the ground truth sequence of labels.

C. *Interactive Grasping System*

We deploy TiO with real-robot manipulation tasks. In our system, TiO is combined with a robot grasping model for interactive visual grounding and grasping. The system takes

TABLE II

RESULTS OF SELF-PLAY EVALUATION ON INVIG BENCHMARK

Oracle	Guesser	Questioner	SR(%)
XVLM-Oracle [9], [29]	Vilbert-Guesser [52]	Vilbert-Questioner [52]	35.3
XVLM-Oracle [9], [29]	XVLM-Guesser [9], [29]	XVLM-Questioner [9], [29]	40.1
XVLM-Oracle [9], [29]	**TiO** (ours)		46.1
TiO (ours)	Vilbert-Guesser [52]	Vilbert-Questioner [52]	51.8
TiO (ours)	XVLM-Guesser [9], [29]	XVLM-Questioner [9], [29]	61.1
TiO (ours)			**78.1**

TABLE III

RESULTS OF SELF-PLAY EVALUATION ON GUESSWHAT?! BENCHMARK

Oracle	Guesser	Questioner	SR(%)
Baseline Oracle [8]	Baseline Guesser [53]	Baseline Questioner [53]	44.6
Baseline Oracle [8]	Baseline Guesser [53]	VDST [54]	45.9
Baseline Oracle [8]	GDSE-SL [55]	GDSE-SL [55]	47.8
Baseline Oracle [8]	Guesser(MN) [56]	TPG [56]	48.8
Baseline Oracle [8]	GST [57]	VDST [54]	50.6
XVLM-Oracle [9], [29]	XVLM-Guesser [9], [29]	XVLM-Questioner [9], [29]	53.0
Vilbert-Oracle [52]	Vilbert-Guesser [52]	Vilbert-Questioner [52]	62.8
TiO (ours)			**65.5**

TABLE IV

ABLATION STUDY ON INVIG BENCHMARK

Model	Self-play	Guesser
TiO	78.1%	77.1%
TiO-Small	71.5%	74.9%
Only GuessWhat?!	49.1%	56.3%
Only InViG	56.9%	71.2%
w/o Dialog	68.9%	75.5%
w/o VG	65.8%	73.7%
w/o IVG	< 5%	< 5%

TABLE V

ABLATION STUDY ON GUESSWHAT?! BENCHMARK

Model	Self-play	Guesser	Oracle
TiO	65.5%	74.6%	88.6%
TiO-Small	61.3%	72.1%	86.3%
Only GuessWhat?!	61.8%	71.8%	86.4%
Only InViG	37.2%	46.9%	39.1%
w/o Dialog	61.3%	72.6%	86.1%
w/o VG	56.2%	71.1%	84.7%
w/o IVG	< 5%	< 5%	< 5% (62.2%)

raw images and ambiguous text instructions as inputs. TiO then plays the role of questioner first to interact with humans and gather the necessary information. When the model stops the conversation actively (see IV-B for details), it turns to the role of the Guesser and generates the bounding box of the target object given all dialogue history, which will be fed into the grasping model as the input. The grasping model consists of two parts: 1) Segment Anything [6] which is used to get the mask of the target object on 2-D images from the bounding box from TiO, and 2) a Contact GraspNet [7] to detect grasps on the point cloud of the target. The best grasp with the highest confidence score and no collisions will be selected to execute.

V. EXPERIMENTS

In order to widely evaluate the performance of TiO and other interactive disambiguation methods, this section investigates the following three points:

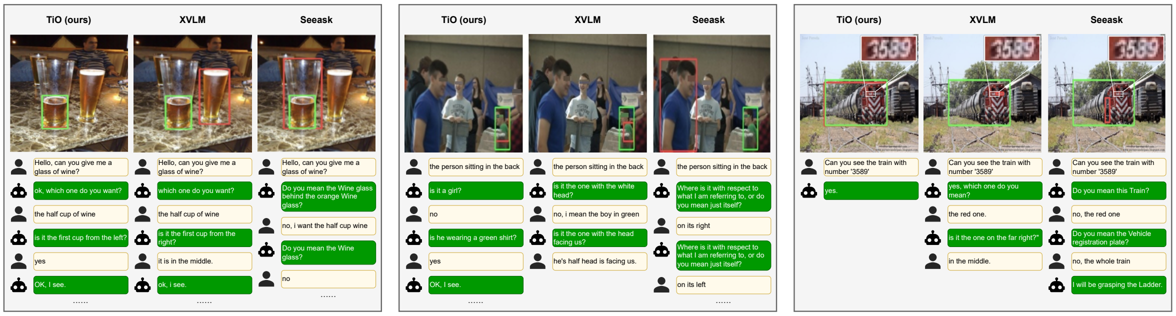
- The performance of TiO on standard benchmarks
- Open-ended interactions with humans
- The generality to interactive real-robot manipulation systems

A. Evaluation on InViG and GuessWhat?!

a) Experiment Setup: Firstly, we evaluate TiO and baselines based on *Self-play Success Rate* on two benchmarks: InViG and GuessWhat?!. InViG [9] is an interactive visual grounding dataset that contains 21k open-ended free-form human-human dialogues and 17.7k images with several ambiguous objects in each image. GuessWhat?! [8] is a simpler dataset that only contains close-ended questions with answers limited to "yes", "no", and "n/a". It consists of 155k dialogues on 66k images, totaling 821k question-answer pairs. In the self-play evaluation, the input contains an image of the scene and an ambiguous language instruction. The interactive model will simultaneously play the role of Questioner, Oracle, and Guesser, to iteratively ask informative questions, answer the questions by itself, and finally stop the conversation and guess the target from the dialogue history. A prediction is considered to be correct if it has an Intersection over Union (IoU) larger than 0.5 with the ground truth. Self-play offers chances for large-scale evaluation of thousands of interactive examples without the participation of humans, while being challenging and comprehensive to simultaneously evaluate the joint performance of all sub-tasks.

b) Compared to Baselines: The performance comparison on the InViG benchmark is shown in Table II. When XVLM-Oracle is as an Oracle, TiO surpassed Vilbert-Guesser/Questioner by 10.8% in disambiguation success rate and surpassed XVLM-Guesser/Questioner by 6.0%. When TiO is as an Oracle, TiO surpassed Vilbert-Guesser/Questioner by 26.3% in disambiguation success rate and surpassed XVLM-Guesser/Questioner by 17.0%. The self-play disambiguation success rates of TiO (as an Oracle) relative to XVLM-Oracle increased by 16.5%, 21.0%, and 32.0% for Vilbert, XVLM, and TiO, respectively. The performance comparison on the GuessWhat?! benchmark is shown in Table III. TiO's self-play success rate surpasses the best baseline by 2.7%. Note, the answer is limited to "yes", "no", and "n/a" on GuessWhat?!, and there is still a certain gap with the dialogue in the real scene. These results demonstrate that the joint training strategy of TiO can comprehensively improve the visual language alignment ability and the context understanding ability of the Oracle, Questioner, and Guesser within the scope of these two benchmarks. We also test OFA [30] on InViG and GuessWhat?!. Since it lacks visual grounding and questioning capabilities based on dialogue, it cannot achieve reasonable performance (Success Rate < 5%).

c) Ablation Studies: We also conduct ablation studies on GuessWhat?! and InViG to figure out the contributions of unified training. To be specific, we first implement TiO with a smaller capacity (TiO-Small) with $\sim$470M parameters,



TiO (ours):
Hello, can you give me a glass of wine?
ok, which one do you want?
the half cup of wine
is it the first cup from the left?
yes
OK, I see.
......

XVLM:
Hello, can you give me a glass of wine?
which one do you want?
the half cup of wine
is it the first cup from the right?
it is in the middle.
ok, i see.
......

Seeask:
Hello, can you give me a glass of wine?
Do you mean the Wine glass behind the orange Wine glass?
no, i want the half cup wine
Do you mean the Wine glass?
no
......

TiO (ours):
the person sitting in the back
is it a girl?
no
is he wearing a green shirt?
yes
OK, I see.
......

XVLM:
the person sitting in the back
is it the one with the white head?
no, i mean the boy in green
is it the one with the head facing us?
he's half head is facing us.

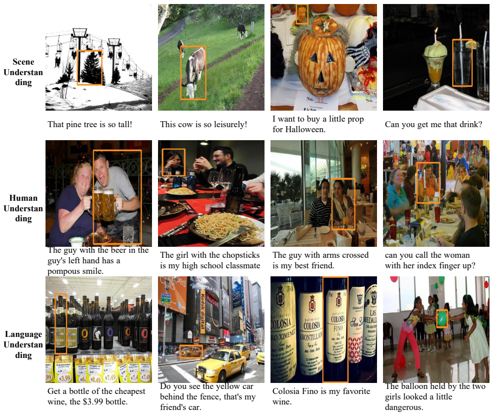
Seeask:
the person sitting in the back
Where is it with respect to what I am referring to, or do you mean just itself?
on its right
Where is it with respect to what I am referring to, or do you mean just itself?
on its left
......

TiO (ours):
Can you see the train with number '3589'
yes.

XVLM:
Can you see the train with number '3589'
yes, which one do you mean?
the red one.
is it the one on the far right?"
in the middle.

Seeask:
Can you see the train with number '3589'
Do you mean this Train?
no, the red one
Do you mean the Vehicle registration plate?
no, the whole train
I will be grasping the Ladder.

Fig. 3. Qualitative results of different interactive visual grounding methods on our 3 HRI evaluation sets. Left: *scene understanding*. Middle: *human understanding*. Right: *language understanding*. The green box denotes the target object by the human user, and the red box denotes the prediction after interaction.



That pine tree is so tall!
This cow is so leisurely!
I want to buy a little prop for Halloween.
Can you get me that drink?
The guy with the beer in the guy's left hand has a pompous smile.
The girl with the chopsticks is my high school classmate
The guy with arms crossed is my best friend.
can you call the woman with her index finger up?
Get a bottle of the cheapest wine, the $3.99 bottle.
Do you see the yellow car behind the fence, that's my friend's car.
Colosia Fino is my favorite wine.
The balloon held by the two girls looked a little dangerous.

Fig. 4. Examples of our evaluation benchmark for HRI experiments. Top row: *Scene Understanding*. Middle row: *Human Understanding*. Bottom row: *Language Understanding*.

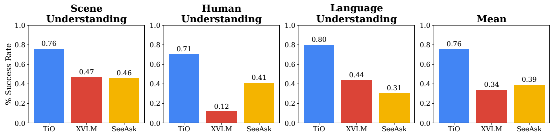


Fig. 5. Interactive visual grounding success rate of HRI on 3 evaluation sets. Our approach achieves the highest performance on the more challenging interactive scenarios.

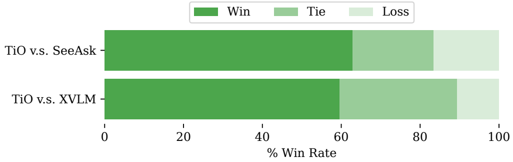


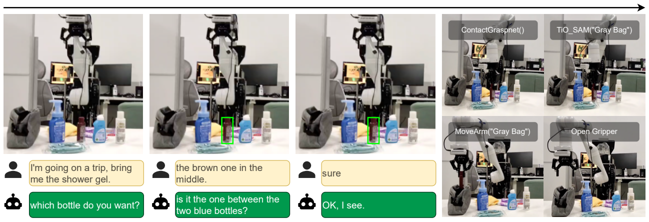
Fig. 6. Human scoring of TiO against baselines.

about half of the full version. As shown in Table IV and Table V, we can conclude that model capacity matters. Larger models consistently improve the performance of all sub-tasks. We also ablate different data sources with TiO-Small to validate their effects on the performance. Concretely, *Only GuessWhat?!* and *Only InViG* mean training on a single interactive dataset. *w/o Dialog* indicates without VisDial and LLaVA. *w/o VG* means without RefCOCO and OpenImages. *w/o IVG* means without both GuessWhat?! and InViG. We can conclude that: 1) More sub-tasks improve performance significantly; 2) The model trained solely on GuessWhat?! or InViG performs relatively well in its own domain, but generalizes worse than the full version. 3) Dialog and VG data can further improve performance when combined with the interactive visual grounding data. 4) Interactive visual grounding (IVG) data itself is indispensable. Models cannot achieve reasonable performance without such data. We meticulously design instructions and decoding methods to further evaluate *w/o IVG* setting. While it also lacks the ability of visual grounding based on dialogue, but it shows some grounded answer capability (shown in brackets), thanks to VG data. This indicates that IVG data is crucial for interactive disambiguation.

B. Evaluation on Human-Robot Interaction

a) Experimental Setup: In order to more comprehensively evaluate the disambiguation ability of TiO, we propose a challenging disambiguation evaluation set, which contains 150 images from the test set of the InViG dataset, *OpenImages* [51], and *objects365* [58], 50 of which are sampled from the images containing *human*-related categories. Then, we divide it into 3 parts aim to evaluate the performance of models on understanding diversified visual concepts (*Scene Understanding*), human attributes & behaviors (*Human Understanding*), and language expressions (*Language Understanding*), which are usually required in open-ended HRI applications. For each one of them, we select 50 images and re-label the instructions. Based on the evaluation set, we then recruit 10 volunteers to interact with TiO and baselines.

b) Compared to Baselines: The success rate of interactive visual grounding with the volunteers on the 150 samples set for HRI is shown in Fig. 5. Our method outperforms XVLM [59] and SeeAsk [18] on all 3 evaluation subsets. We achieved success rates of 76.0%, 70.6%, and 80.0% on the sets of Scene Understanding, Human Understanding, and Langauge Understanding, respectively. By contrast, XVLM performs significantly worse on the *human understanding* set due to the scarcity of human-related data in InViG dataset (<1%). SeeAsk, being a rule-based method, demonstrates



I'm going on a trip, bring me the shower gel.

which bottle do you want?

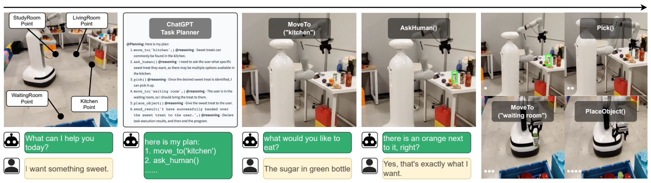
the brown one in the middle.

is it the one between the two blue bottles?

sure

OK, I see.

Fig. 7. Showcase of the interactive visual grounding in robotic manipulation tasks on the desktop.



@Planning: Here is my plan:
1. move_to('kitchen',) @reasoning - Sweet treats can commonly be found in the kitchen.
2. ask_human() @reasoning - I need to ask the user what specific sweet treat they want, as there may be multiple options available in the kitchen.
3. pick() @reasoning - Once the desired sweet treat is identified, I can pick it up.
4. move_to('waiting room',) @reasoning - The user is in the waiting room, so I should bring the treat to them.
5. place_object() @reasoning - Give the sweet treat to the user.
6. send_result('I have successfully handed over the sweet treat to the user.',) @reasoning - Declare task execution results, and then end the program.

What can I help you today?

I want something sweet.

here is my plan:
1. move_to('kitchen')
2. ask_human()
......

what would you like to eat?

The sugar in green bottle

there is an orange next to it, right?

Yes, that's exactly what I want.

Fig. 8. Showcase of the interactive visual grounding in robotic manipulation tasks on the mobile platform.

TABLE VI

DISAMBIGUATION AND GRASPING SUCCESS RATE IN DESKTOP SCENES.

Method	Grounding Success Rate	Grasp Success Rate
XVLM	40% (6/15)	40% (6/15)
SeeAsk	46% (7/15)	46% (7/15)
TiO	**86% (13/15)**	**73% (11/15)**

relatively stable performance on the 3 evaluation subsets. However, its performance is relatively low on the Language Understanding subset, revealing the drawbacks of rule-based methods on understanding diversified and open-ended language inputs. A human scoring based on the evaluation of the disambiguation dialog is shown in Fig. 6. It can be found that, the dialog between humans and TiO is more preferred than 2 baselines in most cases.

c) Qualitative Demonstrations: As shown in Fig. 3, we can see that TiO exhibits strong generalization to diversified inputs and open-world settings. It can understand fine-grained linguistic concepts, including “half cup of wine” against “cup” in the first example, and “with the number 3589” in the third example. Besides, TiO can generate informative and natural questions to exclude disturbance efficiently. By contrast, SeeAsk can only follow question templates.

C. Real-World Evaluation

a) Experimental Setup: In this phase, we have deployed TiO on two real-robot platforms to evaluate the performance of interactive robotic manipulation. We apply a Kinova arm for manipulation and a RealSense camera for RGB and point cloud observation. The mobile platform is developed by our own team.

b) Disambiguation on Desktop Robot: Before the interaction, the robot observes the scene from above the desktop. After receiving the instruction, it interacts with the user to gather information, after which it invokes the grasping model (Section IV-C) for grasp execution. To evaluate the effectiveness of the disambiguation process, we measure the success rate of each method in correctly understanding the user’s intent (Table VI). Our method, TiO, outperforms both XVLM and SeeAsk by a large margin, achieving the highest interactive grounding success rate of 86%. This result demonstrates the robustness and accuracy of TiO in understanding and disambiguating the user’s ambiguous request. We also compare the grasp success rates of TiO, XVLM, and SeeAsk in the desktop scene. Remarkably, TiO exhibited the highest grasp success rate among the three methods. It highlights the efficacy of our system in enabling the robot to successfully interact with objects physically in the desktop environment.

c) Disambiguation on Mobile Robot: By integrating TiO into mobile robotic platforms, we showcase its performance and adaptability in more complex tasks. As shown in Fig. 8, In this phase, TiO is combined with a task planner for navigation in a house. In detail, after receiving the instruction from the human user, the task planner (ChatGPT in our case) will plan for a macro-action sequence given the instruction and a list of macro-action candidates (e.g. move_to, ask_human, grasp). The robot will follow the plan to navigate to the first place. The TiO is responsible for the multi-modal interaction and disambiguation with humans. Concretely, it first observes the scene and detects whether possible candidates exist. If the answer is “Yes”, it continues to interact with humans. Otherwise, the task planner will be invoked again to re-plan for the next place. The existence detection of candidates is trained with the question “Is there a [OBJ]?”, and the training data is generated directly from OpenImages. Our results indicate that TiO can successfully facilitate the disambiguation process, enabling the robot to accurately interpret human intentions and execute actions in dynamic scenarios.

VI. CONCLUSION

This paper proposes TiO that unifies three agents Questioner, Oracle, and Guesser into one single transformer, for interactive visual grounding. Compared with baselines including XVLM [59] and SeeAsk [18], TiO sets the new state-of-the-art both on standard benchmarks of GuessWhat?! and InViG, and moreover, the interaction with humans, outperforming baselines by a large margin. We also deploy TiO on real-robot platforms and validate its effectiveness in real-world interactive robotic manipulation tasks. Currently, restricted by the training data, TiO can only generate simple sentences for interaction. In the future, it is promising to improve the text generation ability by using text datasets. Besides, it is also important to adapt to more expressions of robotic manipulation (e.g. the grounding of actions), which is important for interactive manipulation systems.

ACKNOWLEDGMENT

This work was supported in part by National Key R&D Program of China under grant No. 2021ZD0112700, NSFC under grant No.62125305, No. U23A20339, No.62088102, No. 62203348.

REFERENCES

[1] Mohit Shridhar, Dixant Mittal, and David Hsu. Ingress: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research*, 39(2-3):217–232, 2020.

[2] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016.

[3] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[4] Hanbo Zhang, Yunfan Lu, Cunjun Yu, David Hsu, Xuguang Lan, and Nanning Zheng. INVIGORATE: Interactive Visual Grounding and Grasping in Clutter. In *Proceedings of Robotics: Science and Systems*, Virtual, July 2021.

[5] Yang Yang, Xibai Lou, and Changhyun Choi. Interactive robotic grasping with attribute-guided disambiguation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8914–8920. IEEE, 2022.

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

[7] Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13438–13444. IEEE, 2021.

[8] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] Hanbo Zhang, Jie Xu, Yuchen Mo, and Tao Kong. Invig: Benchmarking interactive visual grounding with 500k human-robot interactions, 2023.

[10] Robin Deits, Stefanie Tellex, Pratiksha Thaker, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, 2(2):58–79, 2013.

[11] Sachithra Hemachandra and Matthew R Walter. Information-theoretic dialog to improve spatial-semantic representations. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5115–5121. IEEE, 2015.

[12] Geert-Jan M Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I Christensen. Clarification dialogues in human-augmented mapping. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 282–289, 2006.

[13] Stephanie Rosenthal, Joydeep Biswas, and Manuela M Veloso. An effective personal mobile robot agent through symbiotic human-robot interaction. In *AAMAS*, volume 10, pages 915–922, 2010.

[14] Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. *Robotics: Science and Systems Foundation*, 2014.

[15] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016.

[16] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864, 2017.

[17] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2193–2202, 2017.

[18] Yuchen Mo, Hanbo Zhang, and Tao Kong. Towards open-world interactive disambiguation for robotic grasping. In *2023 International Conference on Robotics and Automation (ICRA)*, 2023.

[19] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog, 2017.

[20] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[22] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[23] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[24] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020.

[25] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[26] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[28] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

[29] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th international conference on Machine learning*, 2022.

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

[31] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks, 2022.

[32] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[33] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[34] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[35] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[36] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[37] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[40] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[41] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[42] Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2307.02469*, 2023.

[43] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[45] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[46] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[47] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[48] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[49] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

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

[51] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

[52] Tao Tu, Qing Ping, Govindarajan Thattai, Gokhan Tur, and Prem Natarajan. Learning better visual dialog agents with pretrained visual-linguistic representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5622–5631, 2021.

[53] Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. *arXiv preprint arXiv:1703.05423*, 2017.

[54] Wei Pang and Xiaojie Wang. Visual dialogue state tracking for question generation, 2019.

[55] Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. Beyond task success: A closer look at jointly learning to see, ask, and guesswhat. *arXiv preprint arXiv:1809.03408*, 2018.

[56] Rui Zhao and Volker Tresp. Improving goal-oriented visual dialog agents via advanced recurrent nets with tempered policy gradient. In *LaCATODA@ IJCAI*, pages 1–7, 2018.

[57] Wei Pang and Xiaojie Wang. Guessing state tracking for visual dialogue. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 683–698. Springer, 2020.

[58] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.

[59] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.