Extending Phrase Grounding with Pronouns in Visual Dialogues

Panzhong Lu[1], **Xin Zhang**[1], **Meishan Zhang**[2*], **Min Zhang**[2]
[1]School of New Media and Communication, Tianjin University
[2]Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen)
{panzhong171,hsinz}@tju.edu.cn, {zhangmeishan,zhangmin2021}@hit.edu.cn

Abstract

Conventional phrase grounding aims to localize noun phrases mentioned in a given caption to their corresponding image regions, which has achieved great success recently. Apparently, sole noun phrase grounding is not enough for cross-modal visual language understanding. Here we extend the task by considering pronouns as well. First, we construct a dataset of phrase grounding with both noun phrases and pronouns to image regions. Based on the dataset, we test the performance of phrase grounding by using a state-of-the-art literature model of this line. Then, we enhance the baseline grounding model with coreference information which should help our task potentially, modeling the coreference structures with graph convolutional networks. Experiments on our dataset, interestingly, show that pronouns are easier to ground than noun phrases, where the possible reason might be that these pronouns are much less ambiguous. Additionally, our final model with coreference information can significantly boost the grounding performance of both noun phrases and pronouns.

1 Introduction

Grounded language learning has been prevailing for decades in many fields (Chandu et al., 2021), generally aiming to learn the real-world meaning of textual units (e.g., words or phrases) by conjointly leveraging the perception data (e.g., images or videos). Bisk et al. (2020) advocate that we cannot overlook the physical world that language describes when doing language understanding research from a novel perspective. In particular, with the stimulation of modeling techniques and multi-modal data collection paradigms, the task has made excellent progress in the downstream tasks, which involves multi-modal question answering (Agrawal et al., 2017; Chang et al., 2022), video-text alignment (Yang et al., 2021) and robot navigation (Roman Roman et al., 2020; Gu et al., 2022).

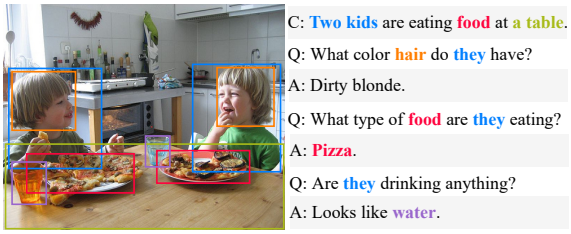

Figure 1: An example of grounding noun phrases and pronouns referred in the caption and dialogue (partly) to the associated image regions. With an image described by a caption, two people are discussing what they can see. Here, we annotate the same object with the same color, respectively. And obviously, the same object mentioned in the text naturally forms a coreference chain.

Typically, as one branch of grounded language learning, phrase grounding, first proposed by Plummer et al. (2015), also plays a key role in visual language understanding. Its goal is to ground the phrases in a given caption to the corresponding image regions. Recently, many researchers have attempted varied approaches to explore this task. Mu et al. (2021) propose a novel graph learning framework for phrase grounding to distinguish the diversity of context among phrases and image regions. Wang et al. (2020) develop a multimodal alignment framework to utilize the caption-image datasets under weak supervision. Kamath et al. (2021) advance phrase grounding with their end-to-end modulated pre-trained network named MDETR. Overall, the natural language processing (NLP) and computer vision (CV) communities have seen huge achievements in the task of phrase grounding.

In spite of its apparent success, there remains a worth-thinking weakness. Almost all previous works mainly focus on the noun phrases/words, which can derive their meanings by the expressional forms to some extent. There is little work that takes account into pronouns. As shown in Fig-

*Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7614–7625
December 7-11, 2022

ure 1, pronouns definitely have underlying effects on the performance of visual grounding, which should be carefully examined (Yu et al., 2019). As a result, here we shift our eyes from the common (almost noun) phrase grounding with the extension of pronouns for the first time.

In this paper, we present the first work for investigating phrase grounding that includes pronouns, and explore how coreference chains can have an effect on the performance of our task. We annotate an initial dataset based on visual dialogue (Das et al., 2017), as shown in Figure 1. For the model, we can directly apply MDETR (Kamath et al., 2021), which is an end-to-end modulated detector. However, the model does offer much information to understand pronouns. Thus, we enhance the vanilla model with coreference information from the dialogue end, where a graph neural network is adopted to encode the graph-style coreference knowledge.

Finally, we conduct experiments on our constructed dataset to benchmark the extended phrase grounding task. According to the results, we find that interestingly, pronouns are easier to ground by MDETR than phrases. The underlying reason might be that the pronouns are always more important during dialogue, leading to less ambiguity in speech communication. In addition, our final model can be significantly enhanced by adding the gold graph-style coreference knowledge; however, the model fails to obtain any positive gain when the coreference information is sourced from a state-of-the-art machine learning model. We conduct several in-depth analyses for comprehensive understanding of our task as well as the model.

In summary, our contributions are as follows:

- We extend the task of phrase grounding by taking account of pronouns, and correspondingly establish a new dataset manually, named VD-Ref, which is the first dataset with ground-truth mappings from both noun phrases and pronouns to image regions.
- We benchmark the extended phrase grounding task by a state-of-the-art model, and also investigate our task with the coreference knowledge of the text, which should benefit our task straightforwardly.
- We observe several unexpected results by our empirical verification, and to understand these results, in-depth analyses are offered to illustrate them, which might be useful for the future investigation of phrase grounding.

Sect.	#Img	#Pronoun	#Phrase	#Box	#Coref
Train	6199	18600	35118	16559	14582
Dev	1063	3256	5739	3074	2503
Test	1595	4033	7941	4347	3754

Table 1: Data statistics of our constructed dataset. #Box means the number of bounding boxes in the image. #Coref means the number of coreference chains.

2 Our Task and The VD-Ref Dataset

2.1 Task Description

The phrase grounding task's general purpose is to map multiple noun phrases to the image regions, however, in this paper, we take the challenge a step further by grounding various noun phrases and pronouns from the given dialogue to the appropriate regions of an image. Take Figure 1 for example, with all the expressions mentioned in the dialogue, like the coreference chain that includes "Two kids" and "they", the task needs to predict the corresponding regions of the object "kids" using bounding boxes in image.

Formally, we define the task as follows: given an image I and the corresponding ground-truth dialogue D, we denote $M = \{N, P\}$ as all the language expressions, typically, N is the oun phrases and P is the pronouns, the prime objective of the task is to predict a bounding box (or bounding boxes) B for each expression.

2.2 Data Collection

With the aim to build a high-quality dataset that includes sufficient pronouns, we adopt the large-scaled VisDialog dataset (Das et al., 2017) which contains 120k images from the COCO (Lin et al., 2014), where each image is associated with a dialogue[1] around to the image. We randomly choose a set of 10k complete sets from the VisDialog dataset, and use the StanfordCoreNLP (Manning et al., 2014) tool to tokenize the sentences, making it proper for the succeeding human annotation.

2.3 Annotation Process

The whole annotation workflow is divided into three stages as follows: (i) developing a convenient online tool for the user annotation; (ii) setting up a standard annotation guideline according to our task purpose; (iii) Recruiting sufficient expert users to annotate the dataset and ensuring each instance

[1]If not specified, the following dialogues that are discussed all contain a caption.

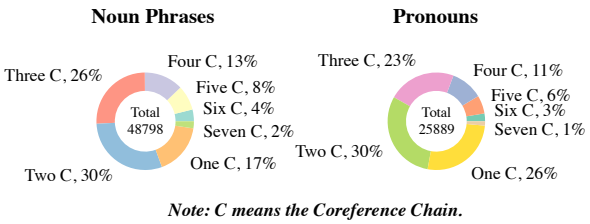

Figure 2: The proportion of noun phrases and pronouns in different number coreference chains.

with three annotations. Firstly, we adopt the label-studio platform (Tkachenko et al., 2020-2022) as the basis to design a user-friend interface targeted to our task, where the concreted interface is shown in Appendix A. Then, we let three people with the visual grounding research experience previously as our experts. They annotate 100 data-pairs together as examples, and establish an annotation guideline based on their consensus after several discussions.

Next, we recruit a number of college students who are expertised at English skills to annotate our dataset. Before starting our task, the students are asked to read the guideline of the annotation process carefully and attempt to annotate some test sets of data, during this period, we examine these students and choose 20 of them to do the following annotation task. In the annotation of each datapoint, the prepared data is split into micro-tasks so that each one consists of 500 dialogues. We assign three workers to each micro-task, and their identities are remained hidden from each other. After all annotation tasks are finished, we let our experts check the results and make corrections of the unconsistent annotations as well.

Finally, we establish the `VD-Ref` dataset, which is annotated manually with the noun phrases and pronouns that naturally form the coreference chains as well as the relevant bounding boxes in images.

2.4 Statistics of the `VD-Ref` Dataset

Totally, we collect 74,687 entity mentions and 23,980 objects from 8,857 VisDialog datasets, where the mentions include 48,798 noun phrases and 25,889 pronouns, on average, a dialogue consists of 5.51 noun phrases and 2.92 pronouns. On the contrary, the existing datasets for phrase grounding hardly consider the pronouns. The ReferItGame dataset (Kazemzadeh et al., 2014) only involves in the noun and noun phrases, while the Flickr30k Entities dataset does not label the corresponding bounding boxes in images, although it annotate the pronouns in captions.

Alternatively, because of the diversity of our dataset, the number of coreference chains varies. As Figure 2 shows, the pie charts display the distinctive distributions of noun phrases and pronouns in the `VD-Ref` dataset. It is clear that whether for the noun phrases or the pronouns, the dialogues that have no more than three coreference chains account for the major proportion, up to 70%, accordingly, the dialogues that have more than three coreference chains constitute the rest proportion.

Moreover, as the mentions of the coreference chains and bounding boxes come in pairs, we can define the coreference chain into four types:

- **one mention vs. one box**: This type contains only one mention and one corresponding box, indicating that the chain exclude pronoun.
- **one mention vs. boxes**: As the referred object is separated into several regions, more than one box is needed to annotate.
- **mentions vs. one box**: In this coreference chain, all noun phrases and pronouns refer to the same single box on the image.
- **mentions vs. boxes**: This type contains several mentions that have noun phrases and pronouns and associated multiple boxes.

Finally, the train, validation and test sets contain 6,199 (70.00%), 1,063 (12.00%) and 1,595 (18.00%) image-dialogue pairs, respectively. We report other statistics in Table 1 as well.

3 Method

Recent works (Kamath et al., 2021; Li et al., 2022) bring the successful vision-language transformer architecture and the pre-train-then-finetune paradigm to the phrase grounding task, achieving state-of-the-art performance. To explore our constructed dataset, we adopt the representative MDETR (Kamath et al., 2021) model. Meanwhile, we propose to enhance the textual representations with the natural coreference chains in texts by Relational Graph Convolutional Networks (R-GCN) (Schlichtkrull et al., 2018). Bellow, we briefly describe how MDETR learns and grounds, and then present our suggested coreference graph encoding.

3.1 Grounding Model

As depicted in Figure 3, for a given image-text pair, MDETR first use an image-encoder (Tan and Le, 2019, EfficientNet) to extract visual features. Then, the features are projected to the image-text shared embedding space by a conv layer, flattened to a se-

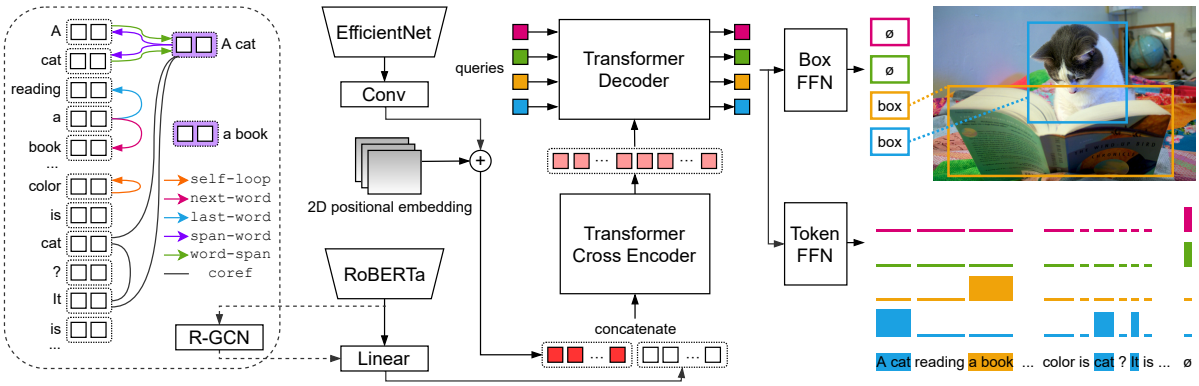

Figure 3: MDETR Model (right) and our suggested coreference graph encoding (left, dashed). Here, we use R-GCN encode the coreference graph into the roberta representation and directly fed the output into the linear layer.

quence, and added with 2-D positional embeddings. Similiarly, a text-encoder (Liu et al., 2019, Roberta) and a linear layer are used to extract and project textual features, respectively. Next, we concatenate vectors of two modalities into one sequence, encoding it by a transformer (Vaswani et al., 2017) encoder. We set N queries and apply a transformer decoder to cross attend the encoded sequence.

Finally, for each one of N hidden states from the decoder, two feedforward networks (FFN) separately predict the object box and a distribution over all *token positions* that correspond to the object, which named **soft token prediction**. Figure 3 shows an example that a query predict the box of the cat and a distribution where tokens refer to this cat are with highest values.

Training. MDETR uses the bipartite matching (Carion et al., 2020; Tan et al., 2021) to find the best match between the predicted boxes and the gold-standard objects then computes the box losses (L1 & GIoU). The soft token prediction is supervised by a soft-cross-entropy between the predicted distribution and a uniform distribution, where tokens refered to the matched gold box have equal probabilities and sum to 1. In addition, the matching cost consists of this grounding loss and the box L1 & GIoU losses. The final loss for the MDETR training is the weighted sum of the above losses and a extra contrastive alignment loss[2].

Inference. For each referring expression, we rank all N proposed boxes by scores of the max over scores assigned to the tokens in this expression, and output the top 10 boxes for the evaluation.

[2]This loss is able to align the query hidden state from the decoder and it's matched referring tokens, please refer to the §2.2.2 of Kamath et al.'s (2021) paper for more details.

3.2 Coreference Graph Encoding

By carefully examining the MDETR model in our extended task, we find that it actually predicts the coreferenced expressions for each detected object to some extend. We guess that explicitly injecting the text coreference information into the representations could boost the model performance to some extent. Thus we propose to encode a simple coreference graph via R-GCN.

Graph Construction. Following the previous graph-based NLP studies (Sahu et al., 2019; Wu et al., 2021; Hu et al., 2020, 2021), we construct our coreference graph in two steps. For the node building, we first initiate the `word` nodes by the input text embeddings. To represent the multi-word mention in text, we generate a virtual `span` node[3] and setup the embedding by the mean embedding of all words in it.

Based on the above two node types (i.e., `word` & `span`), we build the coreference graph with the following six edge types:

- `self-loop`: include the information of itself.
- `next-word`: to keep the sequential information, we link a word to its next word.
- `last-word`: likewise link a word to the last.
- `span-word`: we link a span node to its words for the graph message passing.
- `word-span`: likewise link a word to its span.
- `coref`: we use this undirected edge to connect words or spans refered to the same object.

R-GCN Encoding. We compute node representations on this edge-labeled graph by the R-GCN (Schlichtkrull et al., 2018). Formally, we denote

[3]We offer an ablation study (§4.4) to verify the effectiveness of this scheme.

Model	Coref F1	Recall@1 Overall	Recall@1 Pronoun	Recall@1 Phrase	Recall@5 Overall	Recall@5 Pronoun	Recall@5 Phrase	Recall@10 Overall	Recall@10 Pronoun	Recall@10 Phrase
ANY-BOX-PROTOCOL										
MDETR	-	43.35	50.15	39.94	57.18	67.35	52.13	65.04	75.41	60.29
MDETR + NeuralCoref	37.6	42.04	49.39	38.36	53.72	63.50	48.84	61.91	72.60	56.60
MDETR + C2f-SpanBERT	66.0	42.36	49.45	38.87	54.80	64.07	50.28	63.58	73.48	58.61
MDETR + Gold†	100	**47.54**	**58.79**	**41.91**	**59.30**	**70.52**	**53.69**	**66.67**	**76.83**	**61.59**
MERGED-BOX-PROTOCOL										
MDETR	-	51.98	60.86	47.59	62.86	71.98	58.40	68.03	77.20	63.61
MDETR + NeuralCoref	37.6	51.04	62.14	45.60	62.01	72.45	56.79	67.03	77.03	61.99
MDETR + C2f-SpanBERT	66.0	51.96	62.45	46.71	62.44	72.46	57.43	67.55	76.90	62.33
MDETR + Gold†	100	**55.43**	**65.86**	**50.26**	**65.72**	**74.73**	**61.23**	**70.52**	**78.64**	**66.50**

Table 2: Test results. † means the result is statistically significant compared with MDETR.

the hidden representation of node i in the l-th R-GCN layer as $\boldsymbol{x}_i^{(l)} \in \mathbb{R}^{d^{(l)}}$, where $d^{(l)}$ is the hidden dimension. The message-passing framework of R-GCN is defined as follows:

$$\boldsymbol{x}_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{|\mathcal{N}_i^r|}\boldsymbol{W}_r^{(l)}\boldsymbol{x}_j^{(l)} + \boldsymbol{W}_0^{(l)}\boldsymbol{x}_i^{(l)}\right), \tag{1}$$

where $\mathcal{R}$ denotes the relation set except the `self-loop`. $\mathcal{N}_i^r$ is the set of neighbouring nodes under the relation $r \in \mathcal{R}$. $\boldsymbol{W}_r$ is the feature transformation matrix for relation r, while $\boldsymbol{W}_o$ corresponds to the `self-loop`.

In the end, we re-construct the sequence from all `word` node representations of the last layer. It is worth to note that this coreference graph encoding is general and could be applied to any grounding models. In our experiments, the output of the R-GCN is directly fed to the Linear layer.

4 Experiment

4.1 Settings

Implementation. We use the pretrained MDETR with Roberta-base and EfficientNet-B3.[4] We employ 2-layer R-GCN[5] to encode the Roberta representations. We use the AdamW (Loshchilov and Hutter, 2019) to update model parameters with lr $1e^{-5}$, weight decay $1e^{-4}$, and batch size 16. The lr of the re-initiated MDETR soft token prediction head and R-GCN module is set to $1e^{-4}$. The 2-norms of gradients are clipped to a maximum of 0.1 to avoid the gradient explosion problem. All experiments are implemented with AllenNLP (Gardner et al., 2018) and conducted on a RTX 3090 GPU.

[4] https://github.com/ashkamath/mdetr

[5] We use the R-GCN implemetation from PyTorch Geometric (Fey and Lenssen, 2019).

Coreference. We consider three ways to obtain coreference chains for our graph-encoding:

- `Gold`: the gold-standard coreference chains annotated in our dataset.
- `NeuralCoref`: the off-the-shelf coreference resolution toolbox based-on SpaCy from HuggingFace (2019), we load the "en-core-web-md" model for SpaCy.
- `C2f-SpanBERT`: the widely used span-based coarse-to-fine model (Lee et al., 2018) with a pretrained SpanBERT-large-cased (Joshi et al., 2020).[6] We train it with the gold coreferences and perform an 5-fold cross-validation to get predictions of the whole dataset.

In our main results, we train the MDETR + `NeuralCoref` or `C2f-SpanBERT` with only pseudo coreferences, which would fit the real scenario. We will investigate the recent state-of-the-art works of text coreference resolution(Wu et al., 2020) and update the results in the future version paper.

Evaluation. Following previous studies, we compute the Recall@k to measure whether the model is able to give the "correct" box in top k predictions, where a box is treated as "correct" if the Intersection-over-Union (IoU) between it and a ground-truth box is above a threshold of 0.5. For each text mention, we consider $k \in \{1, 5, 10\}$. We conduct experiments on both `Any-Box` and `Merged-Box` protocol, where the former decides a proposed box is correct to a mention when it has an Iou > 0.5 with *any* of the gold boxes of this mention, while the latter merges all ground-truth boxes of a mention into one smallest enclosing box.

We use the best-performing model on the devset to evaluate the performance of the testset. We run

[6] https://github.com/allenai/allennlp-models

Model	Coref F1	Protocol	Miss Overall	Miss Pronoun	Miss Phrase	Part Overall	Part Pronoun	Part Phrase	Correct Overall	Correct Pronoun	Correct Phrase
	#Mention			498	5507		3335	2107		125	300
+ NeuralCoref	37.6	Any-Box	34.52	45.78	33.50	48.44	48.19	48.84	39.76	45.60	37.33
		Merged-Box	42.10	54.22	41.00	60.69	63.27	56.62	47.29	64.00	40.33
	#Mention			410	2921		2497	3452		1051	1541
+ C2f-SpanBERT	66.0	Any-Box	26.06	45.12	23.38	47.96	47.94	47.97	45.99	46.72	45.49
		Merged-Box	32.81	49.02	30.54	60.30	63.84	57.73	57.48	64.41	52.76

Table 3: Test recall@1 of MDETR with `NeuralCoref` or `C2f-SpanBERT`, by the mention prediction types, where `Miss` means a mention is not extracted by coref models, `Part` (resp. `Correct`) denote a mention is extracted with the incorrect (resp. correct) coreference cluster. We also provide the number of each type, i.e., the #Mention row.

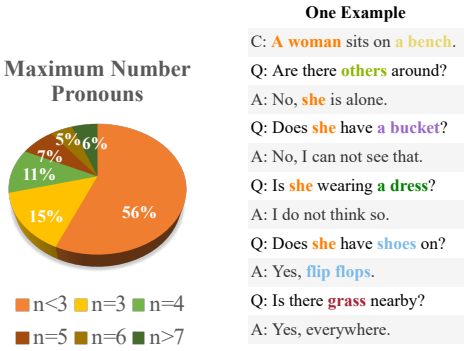

Figure 4: The assessment of our whole dataset on the maximum number of pronouns (in one coreference chain) for every dialogue(left) and one example (right).

each setting by 5 different random seeds, and the average test scores are reported. We regard a result as statistically significant when the p-value is below 0.05 by the paired t-test with baseline MDETR.

4.2 Main Results

Table 2 summarizes the main results of phrase grounding experiments on our VD-Ref dataset. We group the models into two settings, `Any-Box` and `Merged-Box` protocols, and report the performance of grounding pronouns and phrases in terms of Recall@k ($k = 1, 5, 10$). In details, we have the following intriguing findings:

Among all the models, we find that pronouns are easier to ground than phrases, no matter the protocol setting. The possible reason is that as an essential part of the sentence in dialogue, pronouns are straightforward and appear more frequently, containing richer details in context, thus promoting the performance to be grounded.

Besides, after comparing the results of the MDETR with gold (MDETR + Gold) and without (MDETR), we see that adding the gold graph-style coreference knowledge can also considerably improve the model's performance. This empirically supports the value of introducing coreference knowledge. Noticeably, the Recall@10 is generally utilized to evaluate the best recall performance, and at this point, MDETR would reach its limit on this task, making it hard to be improved to some extent, while the addition of the gold graph-style coreference increase that by more than 1%, which further proof the significance of coreference knowledge.

However, we still observe that performance declines when we apply machine learning models (e.g., `NeuralCoref` and `C2f-SpanBERT`) to obtain the coreference chains for our graph structure representations. One possible reason is that these models do not do so well in dialogues, making investigating the more thorough sense worthwhile.

4.3 Analysis

Pronouns outperform phrases. To dig into the in-deep reasons for this performance, we count the maximum number of pronouns (in one coreference chain) for every dialogue, and select one annotated dialogue as an example (see Figure 4). Here, we find that pronouns frequently occur in dialogues, and the maximum number of pronouns larger than three accounts for 44% in our dataset, indicating the importance of pronouns in dialogues. Besides, take Figure 4 (right) for an example. Four pronouns refer to "woman" in the dialogue. The reason behind this is that expressing pronouns are more concise to refer to the specific object, reducing ambiguity in communication.

Detailed Comparison of Non-Gold Methods. To find reasons for the unexpected failure of MDETR with pseudo coreferences from `NeuralCoref` and `C2f-SpanBERT`, we split testset mentions by the coreference cluster prediction of each of them is failed/partially correct/correct. Detailed R@1 values in the three types are in Table 3. When the prediction fails (`Miss`), model perfor-

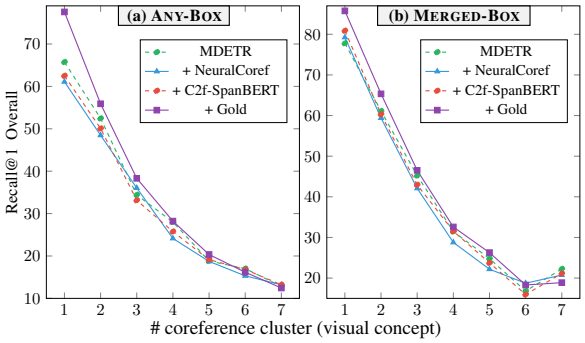

Figure 5: The test Recall@1 (overall) scores grouped by #cluster, which act as the number of visual concepts and represent the difficulty of a data point.

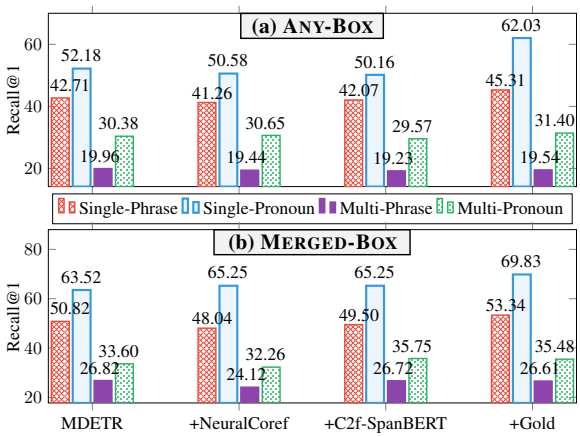

Figure 6: The test Recall@1 of one phrase/pronoun referring to Single/Multi visual objects.

mances are significantly lower than the average, which hurts the overall performances much since these mentions took considerable portions. Surprisingly, the partially correct scores are above the average, which means that even with the defective coreference knowledge, models could precisely ground to a certain extent. Improving the coref model recall could be an effective way to promote grounding performance of suggested method.

Understanding Complex Scenarios. Generally, models would perform worse in a complex scene than in a simple one. We design analysis to validate it in practice to evaluate model abilities in complex scenarios. We measure the difficulty of a scenario (data point) by the number of *coreference clusters*, which represents the number of *visual concepts* that need identifying, grouping testset into different parts. As shown in Figure 5, performances of all methods decline as the clusters increase. The `Gold` offers notable improvements in the simple data (#cluster ≤ 3). All methods perform poorly in complex scenarios, which would be one major limitation of phrase grounding models currently.

Grounding Single/Multi-Object Mentions. As discussed by Kamath et al. (2021), the `Any-Box` and `Merged-Box` protocols are used to handle that the recall@k implies each mention referring to single object. Here we divide mentions into two types, single and multi reference (e.g., the multi reference "two kids" referred two boxes in Figure 1), and compare the performances. In Figure 6, indeed the multiple reference are much challenging, showing shocking gaps to the single. That is, except for the challenges in complex scenarios (instance-level), the model ability on multi-object mentions (prediction-level) also need to be upgraded. Besides, the performance of pronouns is consistently better than that of noun phrases as expected.

Objective	Model	R@1	R@5	R@10
	ANY-BOX-PROTOCOL			
Phrase	MDETR	38.98	52.01	60.43
	MDETR + Gold	41.43	53.15	61.54
Pronoun	MDETR	59.98	69.77	74.93
	MDETR + Gold	62.31	70.51	75.71
	MERGED-BOX-PROTOCOL			
Phrase	MDETR	47.34	57.75	63.60
	MDETR + Gold	48.48	58.03	63.89
Pronoun	MDETR	65.75	73.62	76.69
	MDETR + Gold	67.03	73.81	76.98

Table 4: Results of different grounding objective, where for the pronoun (resp. noun phrase) task, the noun phrase (resp. pronoun) is not trained and evaluated.

Single Grounding Objective. Our extended task grounds the noun phrases and pronouns simultaneously, as illustrated in §4.2. We now evaluate the MDETR and `Gold` in single grounding objectives, i.e., ground only noun phrases or pronouns. Table 4 lists the test scores. First, our suggested coreference graph encoding with `Gold` annotations could consistently boost performance in both subtasks. Then, all methods in phrase grounding only task exists notable performance gap to our extended task results (Table 2), e.g., in ANY-BOX recall@1, 39.94 (extended) $\rightarrow$ 38.98 (phrase only) of MDETR, 41.91 $\rightarrow$ 41.43 of `Gold`. This shows that learning to ground both pronouns and noun phrases could promote the dialog and scenery understanding, thus improving the phrase grounding.

Model	Recall@1		
	Overall	Pronoun	Phrase
ANY-BOX-PROTOCOL			
MDETR + Gold	47.54	58.79	41.91
w/o coref	44.22	54.48	39.58
w/o Virtual Span	46.21	55.61	41.52
w/o coref & Virtual Span	43.56	52.60	39.04
MDETR	43.35	50.15	39.94
MERGED-BOX-PROTOCOL			
MDETR + Gold	55.43	65.86	50.26
w/o coref	52.71	63.74	47.19
w/o Virtual Span	54.31	64.41	48.63
w/o coref & Virtual Span	51.96	62.46	46.71
MDETR	51.98	60.86	47.59

Table 5: Ablation study of graph encoding.

4.4 Ablation Study

To verify the effectiveness of our designed coreference graph, we conduct ablation experiments in the gold coreference setting.

Coreference Edge. We first drop coref edges to show the importance of text coreference knowledge. As presented in Table 5, obviously, the model performance decrease dramatically in both protocols. However, the graph with virtual spans provides mild improvements, we investigate this at the last.

Virtual Span Node. The virtual spans are used to represent the multi-word text mentions. Here we remove them and the span-word & word-span edges, then densely connect every words with each other in one coreference cluster by the coref edge. As shown, the model performance is degraded to a certain extent. Thus, the virtual span scheme is with not only conceptual advantages but also better performance. In addition, we can directly use the span node features when applying to other span-based models, like Liu and Hockenmaier (2019).

Only Word Node and Relation. In the end, we remove both coref and virtual span, keeping only next-word, last-word, and self-loop edges. In Table 5, we can see that this model is only comparable to the baseline. First, this corroborates, without virtual spans, the coreference is still effective (the above paragraph). Moreover, virtual span nodes alone act as the span indicator could improve the model as well (the first paragraph).

5 Related work

Visual Grounding. General visual grounding, also known as referring expression comprehension (Deng et al., 2021; Qiao et al., 2021), is akin to phrase grounding to some extent, since they all aim to study the mapping from the expressions to the specific image regions. The main difference between them is that the visual grounding particularly focuses on one single expression, while the phrase grounding is more general and can be applied to multiple expressions.

Phrase Grounding. A wealth of prior work (Yu et al., 2020; Dogan et al., 2019; Wang et al., 2020; Kamath et al., 2021) on phrase grounding has achieved promising results. Typically, Bajaj et al. (2019) present an end-to-end framework with a separate graph neural network to explore phrase grounding, and Liu et al. (2020) enhance this task by proposing a language-guided graph representation to capture the global context of grounding entities and their relations. In this work, we first propose that grounding pronouns is indispensable, then follow the foundation of using graph structures to our task, positing that the extra coreference knowledge in texts are positive and useful.

Visual Coreference Resolution. It is true that our proposed task has some similarities with the visual coreference resolution task. Yu et al. (2019) formalizes visual-aware pronoun coreference resolution (PCR), builds a dataset for PCR in visual-supported dialogues, and then presents a PCR model with image features. In other words, It solves the pronoun coreference at the text side with the help of visual information. In contrast, our task tackles coreference across the text and image, and in addition, we are also concerned about noun phrases, not only the pronouns. Additionally, Kottur et al. (2018) indeed presents visual coreference resolution (VCR) very similar to ours, with only a difference in the coreference direction (image-to-text v.s. ours text-to-image). As it targets visual question answering, the work does not build a dataset for VCR nor evaluate it. Moreover, it handles VCR at the sentence level for each question in the visual dialogue. In our work, we focus on VCR directly, with a released benchmark dataset, initial models as well as benchmark results.

Related Datasets. The usual visual grounding datasets (Yu et al., 2016), RefCOCO, RefCOCO+ and RefCOCOg, only include one simple expression without pronouns. There are several benchmark datasets (Lin et al., 2014; Krishna et al., 2017) for phrase grounding, and the most well-known is

Flickr30k Entities dataset (Plummer et al., 2015). Nevertheless, since these datasets are among the first to build the relations between the noun phrases mentioned in a sentence and the specific localization of a corresponding image, they may ignore the pronouns, which can also be grounded and assistant to visual language understanding.

6 Conclusion

In this work, we proposed to extend phrase grounding task with pronouns, additionally, we established our dataset, VD-Ref, the first dataset which contains ground-truth mappings from both noun phrases and pronouns to image regions. Furthermore, we took the state-of-the-art model MDETR as our baseline and introduced extra coreference knowledge with graph neural networks. Experiments on our dataset showed the exciting phenomenon that pronouns are more accessible grounded than phrases and demonstrated the significance of coreference knowledge in visual language understanding. To this end, we conducted in-depth analyses of our results. In the future, we would expand more sophisticated dataset, and do more richer experiments on our dataset.

Our dataset and baseline code are avaliable at https://github.com/izhx/Phrase-Grounding-with-Pronoun.

Limitations

In this work, we collect our dataset and extend phrase grounding with pronouns by a series of explored experiments. Admittedly, due to the uneven distribution of raw data and complex annotation process, the main limitation is that our dataset only considers the visual phrases and pronouns, while lacking the annotations on non-visual textual expressions, and giving no insight into the scenery regions as well, which could restrict the research on more sophisticated conditions with varied coreference chains. Future work should be undertaken to expand a more complicated dataset and do more abundant experiments with coreference chains.

Ethical Statement

We build the dataset VD-Ref to go on our researches, aiming to extend the phrase grounding task with pronouns, and study the performance where the coreference chains impact on.In the data annotation process, we adhere to a certain code of conduct on ethical consideration. When recruiting annotators for our task, we claim that all the potential annotators are free to choose whether they want to participate, and they can withdraw from the study anytime without any negative repercussions. Additionally, the whole annotation tasks are anonymized, totally agnostic to any private information of annotators. Furthermore, the annotation results and dataset do not involve any sensitive information that may harm others. Overall, the establishment of our dataset is compliant with ethics.

Acknowledgments

We thank all reviewers for their hard work. This research is supported by grants from the National Natural Science Foundation of China (No. 62176180).

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. VQA: visual question answering - www.visualqa.org. *Int. J. Comput. Vis.*, 123(1):4–31.

Mohit Bajaj, Lanjun Wang, and Leonid Sigal. 2019. G3raphground: Graph-based language grounding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4280–4289. IEEE.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proc. of the EMNLP*, pages 8718–8735, Online. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the ACL-IJCNLP*, pages 4283–4305, Online. Association for Computational Linguistics.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16495–16504.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1749–1759. IEEE.

Pelin Dogan, Leonid Sigal, and Markus H. Gross. 2019. Neural sequential phrase grounding (seqground). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4175–4184. Computer Vision Foundation / IEEE.

Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proc. of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. 2022. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proc. of the ACL*, pages 7606–7623, Dublin, Ireland. Association for Computational Linguistics.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. 2020. SelfORE: Self-supervised relational feature learning for open relation extraction. In *Proc. of the EMNLP*, pages 3673–3682, Online.

Xuming Hu, Chenwei Zhang, Yawen Yang, Xiaohe Li, Li Lin, Lijie Wen, and Philip S. Yu. 2021. Gradient imitation reinforcement learning for low resource relation extraction. In *Proc. of the EMNLP*, pages 2737–2746, Online and Punta Cana, Dominican Republic.

HuggingFace. 2019. NeuralCoref 4.0: Coreference resolution in spacy with neural networks. Https://github.com/huggingface/neuralcoref.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1760–1770. IEEE.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proc. of the EMNLP*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, volume 11219 of *Lecture Notes in Computer Science*, pages 160–178. Springer.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proc. of the NAACL-HLT, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. Grounded language-image pre-training. In *2022 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 21-24, 2022*. IEEE Computer Society.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Jiacheng Liu and Julia Hockenmaier. 2019. Phrase grounding by soft-label chain conditional random field. In *Proc. of the EMNLP-IJCNLP*, pages 5112–5122, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. 2020. Learning cross-modal context graph for visual grounding. In *AAAI, New York, NY, USA, February 7-12, 2020*, pages 11645–11652. AAAI Press.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. of the ACL: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Zongshen Mu, Siliang Tang, Jie Tan, Qiang Yu, and Yueting Zhuang. 2021. Disentangled motif-aware graph learning for phrase grounding. In *AAAI, Virtual Event, February 2-9, 2021*, pages 13587–13594. AAAI Press.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2021. Referring expression comprehension: A survey of methods and datasets. *IEEE Trans. Multim.*, 23:4426–4440.

Homero Roman Roman, Yonatan Bisk, Jesse Thomason, Asli Celikyilmaz, and Jianfeng Gao. 2020. RMM: A recursive mental model for dialogue navigation. In *Findings of the EMNLP*, pages 1732–1745, Online. Association for Computational Linguistics.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proc. of the ACL*, pages 4309–4316, Florence, Italy. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Zeqi Tan, Yongliang Shen, Shuai Zhang, Weiming Lu, and Yueting Zhuang. 2021. A sequence-to-set network for nested named entity recognition. In *Proc. of the IJCAI, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3936–3942. ijcai.org.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Qinxin Wang, Hao Tan, Sheng Shen, Michael Mahoney, and Zhewei Yao. 2020. MAF: Multimodal alignment framework for weakly-supervised phrase grounding. In *Proc. of the EMNLP*, pages 2030–2038, Online. Association for Computational Linguistics.

Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, and Bo Long. 2021. Graph neural networks for natural language processing: A survey. *CoRR*, abs/2106.06090.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proc. of the ACL*, pages 6953–6963, Online. Association for Computational Linguistics.

Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11542–11552. IEEE.

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*, volume 9906 of *Lecture Notes in Computer Science*, pages 69–85. Springer.

Tianyu Yu, Tianrui Hui, Zhihao Yu, Yue Liao, Sansi Yu, Faxi Zhang, and Si Liu. 2020. Cross-modal omni interaction modeling for phrase grounding. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1725–1734. ACM.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *Proc. of the EMNLP-IJCNLP*, pages 5123–5132, Hong Kong, China. Association for Computational Linguistics.

A Annotation Interface

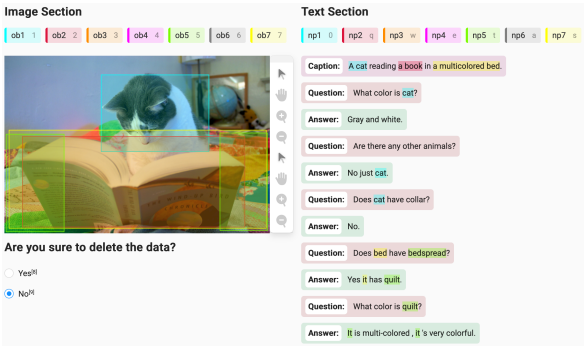

Figure 7: The designed interface of our annotation platform using label-studio tool.

We designed the annotation interface. As illustrated in Figure 7, the left panel is the image section, while the right panel is the text section. They both have several boxes with distinct colors, which are used to annotate image regions and textual expressions. Moreover, the interface provides seven colors to choose from since the number of objects in the dialogue does not exceed 7 as a precondition. Notably, there is one option, "Are you sure to delete the data?", for the annotators and reviewers to remove vague and ambiguous datasets, where the dialogue contains too much irrelevant content or the image is incomplete, making it challenging to be recongnized.