Segment-Level and Category-Oriented Network for Knowledge-Based Referring Expression Comprehension

Yuqi Bu[1,2]*, **Xin Wu**[1,2]*, **Liuwu Li**[1,2], **Yi Cai**[1,2]†, **Qiong Liu**[1], **Qingbao Huang**[3,4]
[1] School of Software Engineering, South China University of Technology
[2] Key Laboratory of Big Data and Intelligent Robot (SCUT), MOE of China
[3] School of Electrical Engineering, Guangxi University
[4] Guangxi Key Laboratory of Multimedia Communications and Network Technology
{seyqbu,sexinw}@mail.scut.edu.cn, liuwu.li@outlook.com,
{ycai,liuqiong}@scut.edu.cn, qbhuang@gxu.edu.cn

Abstract

Knowledge-based referring expression comprehension (KB-REC) aims to identify visual objects referred to by expressions that incorporate knowledge. Existing methods employ sentence-level retrieval and fusion methods, which may lead to issues of similarity bias and interference from irrelevant information in unstructured knowledge sentences. To address these limitations, we propose a segment-level and category-oriented network (SLCO). Our approach includes a segment-level and prompt-based knowledge retrieval method to mitigate the similarity bias problem and a category-based grounding method to alleviate interference from irrelevant information in knowledge sentences. Experimental results show that our SLCO can eliminate interference and improve the overall performance of the KB-REC task. ‡

1 Introduction

Referring expression comprehension (REC), a.k.a. visual grounding, aims to identify a visual object referred to by a referring expression that disambiguates multiple objects (Cirik et al., 2018; Qiao et al., 2021). As a core task of language-vision fields, REC benefits many downstream multimodal tasks, e.g., robotics (Berg et al., 2020; Wang et al., 2022) and vision-and-language navigation (Qi et al., 2020; Gao et al., 2021).

To explore a broader domain of knowledge, Wang et al. extend the REC task to knowledge-based referring expression comprehension (KB-REC), and propose a baseline model and benchmark (Wang et al., 2020). This task requires the use of external knowledge (e.g., commonsense and encyclopedia) to refer to objects. This necessitates the model's ability to retrieve knowledge related to expressions and associate it with image and expression, enabling localization of the referent.

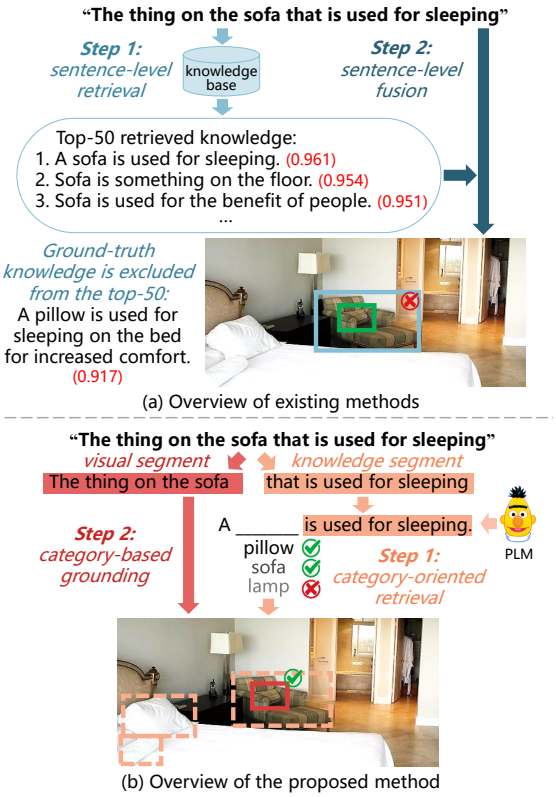


Figure 1: Comparison of the proposed method's scheme with existing methods. The red numbers at the end of the knowledge sentences represent the similarity score.

The existing method ECIFA (Wang et al., 2020) retrieves and fuses knowledge in a sentence-level framework. They utilize sentence-level similarity to retrieve the most similar unstructured knowledge sentences from external knowledge bases, e.g., descriptive sentences from Wikipedia. Then they fuse all these retrieved knowledge sentences with expressions to locate the referent. However, ECIFA still has two limitations. Firstly, the sentence-level similarity method exhibits a similarity bias problem (Bogatu et al., 2022). This means that although the retrieved knowledge sentences are lexically similar to the query expression, they may not be the intended knowledge for understanding the expres-

*Equal contribution.
†Corresponding author.
‡The code is available at https://github.com/Buki2/SLCO.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8745–8757
July 9-14, 2023

sion. Consequently, the irrelevant knowledge may result in an incorrect localization of the referent due to error propagation. As shown in Fig. 1(a), the intended knowledge for this expression is about pillows, while existing methods retrieve knowledge all about sofas due to sentence similarity with the expression. As a result, the lack of knowledge about pillows leads to localizing the incorrect object, the sofa. Secondly, the retrieved unstructured knowledge sentences may contain a large amount of information that is unrelated to the referent, leading to interference in object localization. As shown in Fig. 1(a), the irrelevant information "on the floor" in the second knowledge sentence may mislead the model to focus on objects located on the ground, rather than the intended focus of "on the sofa" in the expression, resulting in an incorrect localization of the sofa on the floor. Even the irrelevant information "on the bed" in the ground-truth knowledge sentence may potentially mislead the model to localize the incorrect object on the bed.

Based on statistical analysis, we find that most knowledge-based referring expressions can be divided into two segments according to the information contained: (1) Visual segments (e.g., "on the sofa" in Fig. 1(b)), which can be interpreted based on visual content, such as color, shape, and relative position of objects; (2) Knowledge segments (e.g., "used for sleeping" in Fig. 1(b)), which require additional knowledge beyond the visual content to be understood, such as function and non-visual object attributes. Distinguishing these two types of segments and discarding the visual segment during knowledge retrieval can help to solve the similarity bias problem. Moreover, for grounding, it is only necessary to know the category of objects corresponding to the knowledge segment, and detailed descriptive knowledge about the object is not required. Therefore, we employ a category-oriented method to retrieve knowledge categories and fuse them with the visual segment for object localization, which can avoid irrelevant information from knowledge sentences. For example, in Fig. 1(b), the knowledge segment can identify which object categories are used for sleeping and narrows the target to pillows and sofas. Then, these categories associated with the visual segment distinguish multiple instances of pillows and accurately locates the referent one on the sofa.

In this paper, we propose a segment-level and category-oriented network (SLCO), which utilizes knowledge segments to retrieve knowledge categories and delegate them to visual segments for grounding target objects. It consists of three modules: a segment detection module, a prompt-based retrieval module, and a category-based grounding module. Firstly, the segment detection module identifies visual and knowledge segments. Then, inspired by the excellent knowledge retrieval ability of prompt learning (Shin et al., 2020; Zhong et al., 2021), we present a prompt-based retrieval module that uses knowledge segments as hints to elicit knowledge categories from generic language models. Finally, the category-based grounding module associates the retrieved knowledge categories with visual segments for target object localization.

The contributions can be summarized as follows:

- We propose a segment-level and prompt-based retrieval method that can retrieve object categories corresponding to knowledge segments, thereby addressing the similarity bias problem and reducing incorrect knowledge retrieval.
- We propose a category-based method to associate knowledge categories with visual segments for object localization, thereby alleviating the interference from irrelevant information in knowledge sentences.
- Experimental results on the KB-Ref dataset show that our SLCO can eliminate interference and improve the overall performance.

2 Related Work

2.1 Referring Expression Comprehension

REC is a fundamental task in the multimodal field. Existing methods are twofold based on the alignment pattern used. Two-stage methods (Wang et al., 2019; Yu et al., 2018) involve an initial stage for detecting boxes, followed by a second stage for ranking these boxes based on an expression. In contrast, one-stage methods (Yang et al., 2020; Huang et al., 2021; Li et al., 2021; Deng et al., 2021; Yang et al., 2022) integrate both visual and textual features to directly regress the bounding box.

To extend this task to a broader knowledge domain, (Wang et al., 2020) introduced a knowledge-based REC task with a benchmark dataset and a two-stage model. This model retrieves knowledge by sentence similarity and fuses it with expression and image for object ranking and selection. However, this model has the problems of similarity bias

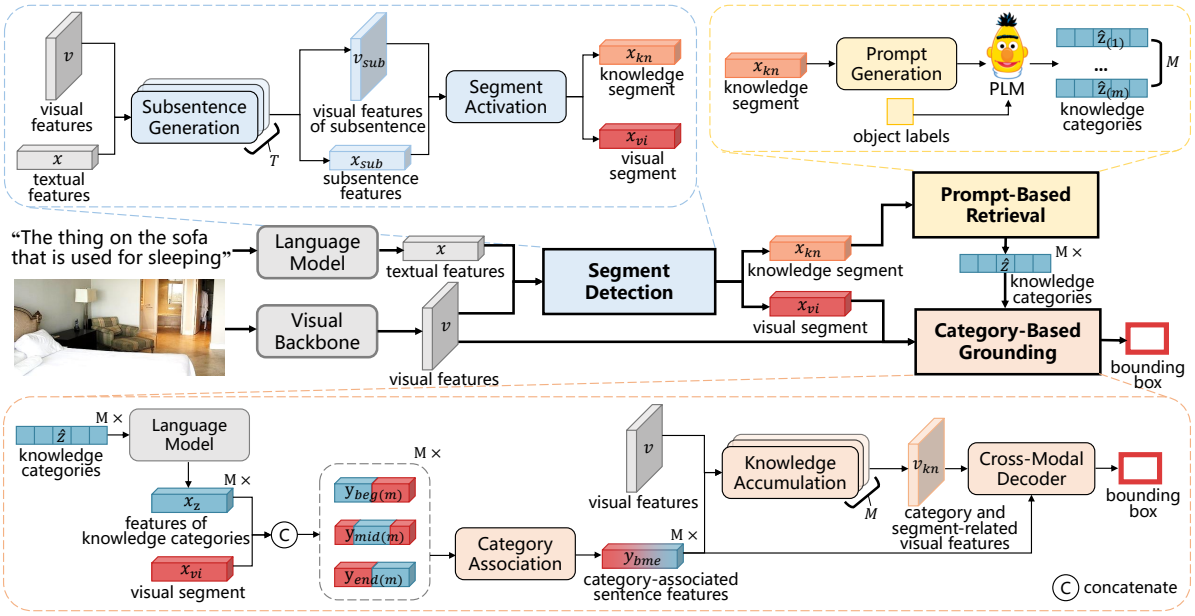


Figure 2: Overview of the proposed segment-level and category-oriented network.

and interference from knowledge sentences. To solve these problems, we try to identify parts of expressions that require knowledge and retrieve it in a segment-level and category-oriented manner.

2.2 Knowledge Retrieval

Early attempts in knowledge retrieval primarily relied on similarity measures and matching methods. However, these methods lack the flexibility to handle the variability of natural language. Recently, prompt-based methods (Petroni et al., 2019; Liu et al., 2021) have been shown to possess superior knowledge retrieval capabilities. Many studies (Shin et al., 2020; Qin and Eisner, 2021; Zhong et al., 2021) have focused on prompt engineering to identify effective templates for knowledge retrieval. However, existing methods typically assume that the subject/object and relation are known, and the task is to identify the corresponding object/subject. Nevertheless, in real-world scenes, it is more valuable to analyze which parts of a sentence require knowledge, rather than solely relying on the prespecified subject/object and relation. In this paper, we propose a method for detecting parts of sentences that require knowledge and automatically generating prompts for knowledge retrieval.

3 Proposed Method

An illustration in Fig. 2, SLCO contains three main modules: (1) A segment detection module, which identifies knowledge segments and visual segments in an expression; (2) A prompt-based retrieval module, which employs knowledge segments as hints to elicit knowledge category from language models; and (3) A category-based grounding module, which associates knowledge category and visual segments for object localization.

3.1 Segment Detection

The main idea of this module is to identify parts of expressions that cannot be inferred solely from images as knowledge segments, and parts with corresponding visual features as visual segments. Given a referring expression $x = \{x_1, ..., x_n\}$ comprising n tokens, a knowledge segment can be defined as a subset of the expression $x_{kn} \subseteq x$, which consists of tokens associated with external knowledge.

Given an image and a referring expression, we first encode visual features v with a convolutional backbone and encode textual features x using a pretrained language model. Due to the difficulties in aligning long expressions with visual features all at once, inspired by sentence decomposition (Yang et al., 2020; Li et al., 2021), we perform a subsentence generation to break the expressions into subsentences by T iterations. The features of subsentence $x_{sub(t)}$ at the t-th iteration are:

$$x_{sub(t)} = s_{sub(t)} \cdot x, \quad (1)$$

$$s_{sub(t)} = Conv(v_{sub(t-1)} \cdot x \cdot s_{sub(t-1)}), \quad (2)$$

where $s_{sub(t)}$ is a score to determine the position of subsentence, with a matrix of ones as an initial

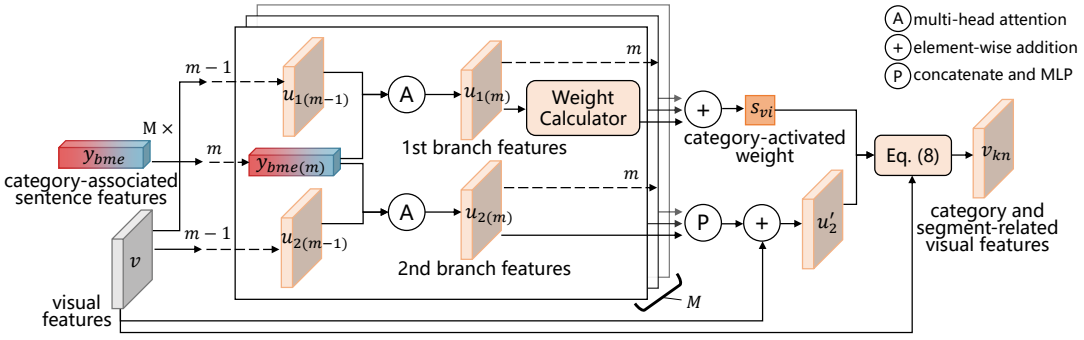


Figure 3: The iterative process of knowledge accumulation method.

value. In iteration, we use a FiLM method (Perez et al., 2018) for feature projection and obtain visual features of the subsentences $v_{sub(t)}$, as follows:

$$v_{sub(t)} = ReLU(v_{sub(t-1)} \odot Linear(x_{sub(t)}) + Linear(x_{sub(t)})), \quad (3)$$

where $\odot$ represents element-wise multiplication.

To determine which parts of expressions have corresponding visual features, we present a segment activation method that attends visual features to textual features, in contrast to the text-to-image alignment scheme of most REC models. Concretely, we first concatenate the results of T subsentence features to obtain x_{sub} and the visual features of T subsentences to obtain v_{sub}, respectively. Then, to identify activated parts of expressions, we use cross-modal attention to attend the visual features v_{sub} to the textual features x_{sub}. The vision-dependent subsentence features x'_{sub} are:

$$x'_{sub} = \sigma(x_{sub} \cdot v_{sub}^{\top}) \cdot v_{sub}, \quad (4)$$

where σ represents a softmax function. Finally, we project these features into two scores via MLP: one representing the parts of the sentence activated by visual context and the other representing the unactivated parts. We then obtain visual segment x_{vi} and knowledge segment x_{kn} by multiplying the sentence features with these scores, as follows:

$$x_{kn} = \sigma(Linear(x'_{sub})) \odot x + x, \quad (5)$$

and $x_{vi} = 1 - x_{kn}$.

During the early stages of model training, the visual features and textual features obtained by single-modal encoders are relatively independent, which makes it challenging to align their representations. To tackle this issue, we introduce a pseudo-supervision strategy to supervise the detection process. In particular, to generate pseudo-annotations, we extract common substrings between expressions and their corresponding reference knowledge, as well as extract knowledge guide words (e.g., "used for" and "made up of") along with the following words. Subsequently, the extracted knowledge-related words are merged and converted into tokens-level scores, where 1 indicates knowledge segments and 0 indicates visual segments. Finally, we use a mean squared error (MSE) loss $\mathcal{L}_{seg}$ to reduce the discrepancy between pseudo-annotations and the predicted scores of knowledge segments.

3.2 Prompt-Based Retrieval

After the detection of knowledge segments, this module conducts segment-level knowledge retrieval of object categories to which the intended knowledge in expressions pertains.

Firstly, we use words in the expression that have a higher score than the median of the knowledge segment scores to regenerate knowledge segments for prompts. Considering that longer knowledge segments are primarily descriptive statements, while shorter segments pertain to similar objects or synonyms, we devise two prompt templates for these scenes. They are in the forms of "A ___ is x_{kn}" and "___ is a kind of x_{kn}", where x_{kn} is a knowledge segment in the input slot. When given a prompt, a pretrained language model f_{PLM} predicts the probability of different tokens $z \in \mathcal{Z}$ that could potentially fill the answer slot. These predictions are then filtered through labels of objects in an image to narrow down potential answers. The top-M highest-scoring tokens $\hat{z}$ are:

$$\hat{z} = \underset{z \in \mathcal{Z}}{\operatorname{argmax}} P(f_{PLM}(z)|x_{prompt}). \quad (6)$$

To enhance the category-oriented retrieval ability of this module, we expand the pretrained language

model's knowledge by incorporating the knowledge bases on which KB-Ref is based. Specifically, inspired by entity-level masking (Sun et al., 2019), we mask knowledge categories in knowledge sentences for fine-tuning the language model.

3.3 Category-Based Grounding

In this module, we associate the retrieved knowledge categories with the detected visual segments for visual grounding. According to (Akula et al., 2020), the position of knowledge categories in a sentence may significantly impact its meaning. Thus, we consider three association forms that can accommodate most situations. In particular, features of each candidate knowledge category $x_{z(m)}$ are integrated into the beginning, middle, and end of a visual segment to obtain $y_{beg(m)}$, $y_{mid(m)}$, and $y_{end(m)}$, respectively. Then, we perform category association to concatenate and linear project these features into category-associated sentence features $y_{bme(m)}$ for the m-th knowledge category.

During iteration, features of top-ranked category-associated sentences with high probability value tend to largely accumulate, facilitating the model to learn important category information. Therefore, we present a knowledge accumulation method to iteratively incorporate textual information of M category-associated sentences into visual features using multi-head attention, as shown in Fig. 3. Specifically, there are two parallel branches. At each step m of the iteration, the first branch fuses the m-th category-associated sentences $y_{bme(m)}$ with the $(m-1)$-th visual features $u_{1(m-1)}$ to obtain $u_{1(m)}$, as follows:

$$u_{1(m)} = \sigma\left(\frac{u_{1(m-1)} \cdot y_{bme(m)}^{\top}}{\sqrt{d_y}}\right) \cdot y_{bme(m)}, \quad (7)$$

where d_y is the dimension of $y_{bme(m)}$. Additionally, the $u_{1(m)}$ is used to compute the weight of each sentence concerning the visual features. We follow the calculation of the verification score in (Yang et al., 2022) to compute this weight. After M iterations, the weights are summed element-wise to obtain category-activated weight s_{vi}. The second branch also fuses $y_{bme(m)}$ with the $(m-1)$-th visual features $u_{2(m-1)}$ to obtain $u_{2(m)}$ in each step of iteration. After M iterations, these features $u_{2(m)}$ are concatenated and projected via MLP, then added to the original visual features, obtaining u_2'. Finally, the category and segment-related visual features v_{kn} are obtained as follows:

$$v_{kn} = \sigma\left(\frac{u_2' \cdot u_2'^{\top}}{\sqrt{d_u}}\right) \cdot v \odot s_{vi} + v, \quad (8)$$

where d_u is the dimension of u_2'.

After activating objects on visual features with knowledge categories and visual segments, we employ a variant of the Transformer decoder (Vaswani et al., 2017) to further distinguish multiple instances of similar objects. It comprises 6 layers of multi-head attention and point-wise fully connected. In each layer, two self-attentions are replaced with cross-modal attention. In one instance, visual features serve as query, text features serve as key and value, and the other is reversed. The resulting features are then projected into four-dimensional object coordinates via MLP.

3.4 Training Objective

The proposed SLCO is trained by a joint loss containing an MSE loss $\mathcal{L}_{seg}$ and a diversity loss (Yang et al., 2020) $\mathcal{L}_{div}$ for segment detection, as well as a smooth L1 loss $\mathcal{L}_{l1}$ and a GIoU loss (Rezatofighi et al., 2019) $\mathcal{L}_{giou}$ for grounding, as follows:

$$\mathcal{L} = \lambda_{seg}\mathcal{L}_{seg} + \lambda_{div}\mathcal{L}_{div} + \lambda_{l1}\mathcal{L}_{l1} + \lambda_{giou}\mathcal{L}_{giou}, \quad (9)$$

where λ are trade-off factors.

4 Experiment

4.1 Experimental Setup

Dataset. KB-Ref (Wang et al., 2020) is the first and currently the only dataset for the KB-REC task. It includes 43,284 knowledge-based referring expressions for objects in 16,917 images from Visual Genome (Krishna et al., 2017). The knowledge involved is derived from Wikipedia, ConceptNet, and WebChild, which are reformed into unstructured sentences. We follow the official data splits.

Evaluation Metrics. Following (Wang et al., 2020), accuracy is an average of the number of predictions with IoU greater than 0.50. Given the absence of ground-truth boxes in practical applications, the model inputs available for experiments are images and expressions, without ground-truth boxes. For knowledge retrieval, accuracy is an average of the number of predictions that the ground-truth knowledge category is within the top-M results, i.e., Acc@M. We obtain the ground-truth category corresponding to the ground-truth knowledge from the KB-Ref dataset.

Model	Visual Backbone	Language Model	Val	Test
Two-stage alignment methods				
LGRANs (Wang et al., 2019)	VGG-16	LSTM	21.72	21.37
MAttNet (Yu et al., 2018)	VGG-16	LSTM	22.04	21.73
ECIFA (Wang et al., 2020)	VGG-16	LSTM	24.11	23.82
One-stage alignment methods				
LBYLNet (Huang et al., 2021)	DarkNet-53	LSTM	22.65	22.41
ReSC (Yang et al., 2020)	DarkNet-53	BERT	27.56	26.88
BBA (Li et al., 2021)	DarkNet-53	BERT	28.28	27.08
TransVG (Deng et al., 2021)	ResNet-101 w/ DETR	BERT	25.03	24.53
VLTVG (Yang et al., 2022)	ResNet-101 w/ DETR	BERT	29.23	28.96
SLCO (Ours)	ResNet-101 w/ DETR	BERT	**32.15**	**30.44**

Table 1: Comparison with state-of-the-art methods on the KB-Ref dataset. All two-stage methods use object detection results as candidate bounding boxes. Bold values indicate the best performance.

Implementation. During training, we finetune the knowledge retrieval module on a 2080Ti GPU and then end-to-end optimize the remaining modules on two P100 GPUs. The height and width of the input image are resized to 640 and the max length of the expression is set to 40. We use a ResNet-101 (He et al., 2016) initialized with weights from DETR (Carion et al., 2020) as the visual backbone, and BERT (Devlin et al., 2019) as the language model. We then follow the preprocessing of (Yang et al., 2022). For training, we use the AdamW optimizer to train SLCO with a batch size of 28 and a total of 90 epochs. The initial learning rate for feature encoders is set to 10^{-5}, and the other modules to 10^{-4}. For the first 10 epochs, we freeze the weights of feature encoders.

In the prompt-based retrieval module, we use a LAMA framework (Petroni et al., 2019) and an uncased BERT large model as f_{PLM} based on (Shin et al., 2020) which suggests its effectiveness for knowledge retrieval among different PLMs. We employ a Faster R-CNN (Ren et al., 2017) pretrained on Visual Genome (Krishna et al., 2017) to identify object labels in an image. The number of retrieved knowledge M is set to 3.

We follow (Yang et al., 2022) to set smooth L1 and GIoU loss for object localization. The trade-off between these loss parameters has been tuned to be 5:2. Additionally, we perform a grid search and find the optimal parameters 10 and 0.125 for the MSE loss and diversity loss in the segment detection module. Accordingly, we set λ_{seg}, λ_{div}, λ_{l1}, λ_{giou} as 10, 0.125, 5, and 2 in Eq.(9).

Baseline Models. With regards to two-stage methods, LGRANs (Wang et al., 2019) uses a graph-based attention method to infer inter-object relationships. MAttNet (Yu et al., 2018) employs three modules to handle the grounding of object appearance, location, and relationship. ECIFA (Wang et al., 2020) retrieves knowledge by cosine similarity between expressions and knowledge sentences, and then uses a stack of LSTM to fuse all the knowledge sentences with expressions. In regard to one-stage methods, LBYLNet (Huang et al., 2021) uses a landmark convolution method to encode object features. ReSC (Yang et al., 2020) utilizes a recursive framework to fuse visual and textual features. Based on it, BBA (Li et al., 2021) employs a bottom-up and bidirectional framework to align multimodal features. TransVG (Deng et al., 2021) constructs a Transformer-based grounding framework. VLTVG (Yang et al., 2022) further extracts text-conditioned discriminative visual features.

Models for the ordinary REC task lack mechanisms to acquire external knowledge and interact with multimodal information. Therefore, following (Wang et al., 2020), we train all models using their default implementations in the ordinary REC training manner on the KB-Ref dataset.

4.2 Main Results

The results in Table 1 show that ECIFA performs better than other two-stage methods, as it explicitly incorporates external knowledge from multiple knowledge bases. As for one-stage methods, BERT-based models generally outperform LSTM-based ones. The reason lies in that the implicit knowledge from the pretrained language model BERT enhances the comprehension of knowledge-based referring expressions. Moreover, the results show

Method	Val	Test
Full model	**32.15**	**30.44**
w/o Detection	29.73$_{\downarrow 2.42}$	29.73$_{\downarrow 0.71}$
w/o Retrieval	30.15$_{\downarrow 2.00}$	29.69$_{\downarrow 0.75}$
w/o Grounding	29.70$_{\downarrow 2.45}$	29.56$_{\downarrow 0.88}$

Table 2: Ablation studies on variants of SLCO architecture, evaluating the effect of three primary modules, i.e., a segment detection module, prompt-based retrieval module, and category-based grounding module.

Method	Retrieval	KB-REC	
	Acc@1	Val	Test
Parsing	44.82	29.75	28.82
Detection	**52.59**	**32.15**	**30.44**

Table 3: Results of segment detection methods on knowledge retrieval and the KB-REC task.

Method	Acc@1	Acc@2	Acc@3
A. Knowledge retrieval methods			
Sim. w/ expr.	26.40	30.63	35.63
Sim. w/ seg.	43.47	46.58	51.29
Prompt w/ expr.	45.48	56.48	61.34
Prompt w/ seg.	**52.59**	**61.85**	**65.53**
B. Fine-tuning strategies of language models			
None	35.96	46.33	52.42
Random mask	48.08	58.64	63.27
Category mask	**52.59**	**61.85**	**65.53**

Table 4: Knowledge retrieval results from different retrieval methods and different fine-tuning strategies. "Sim." refers to the cosine similarity method. "w/ expr." and "w/ seg." refer to retrieving knowledge based on expressions and knowledge segments, respectively.

that our proposed SLCO achieves a performance gain of up to 2.92%, demonstrating the effectiveness of the segment-level and category-oriented strategy. Furthermore, SLCO is the first model that is able to associate both implicit and explicit knowledge from pretrained language models.

Additionally, we conduct an evaluation of the inference time. Our model exhibits an inference time of 0.117 seconds per sample, whereas the baseline model ECIFA necessitates 0.367 seconds.

4.3 Ablation Study

We conduct a series of experiments to verify the effectiveness of three main modules (cf. Table 2).

Effectiveness of Segment Detection Module. There is an average decrease of 1.57% when we remove this module and its loss functions. This result validates the effectiveness of our segment-level method, which solves the similarity bias problem in the sentence-level method.

Effectiveness of Prompt-Based Retrieval Module. To evaluate the importance of this module, we replace the retrieved knowledge categories with empty strings. The results show that removing this module leads to an average decline of 1.38%, indicating the value of knowledge categories for visual grounding. Moreover, this experimental setup corresponds to using only implicit knowledge, which is similar to the knowledge sources employed by the state-of-the-art one-stage methods in Table 1. Nevertheless, our method has better performance than these methods.

Effectiveness of Category-Based Grounding Module. We ablate this module as well as the prompt-based retrieval module, thus the visual grounding process is performed by visual segments only. Results show that this reduces the model performance by 1.67% on average. The reason lies in that model lacks the ability to associate knowledge categories, making it hard to understand referring expressions and localize the correct referent.

4.4 Evaluation of Segment Detection Method

To explore methods for identifying knowledge and visual segments, we construct a parsing method to compare with the proposed detection method. The parsing method divides the expressions according to the constituency parsing in the syntax. Based on observation, knowledge information mostly appears in predicates or subordinate clauses of expressions. Thus, we take the first half of the parsed sentence as a visual segment and the second half as a knowledge segment.

In Table 3, it can be observed that the parsing method underperforms in both knowledge retrieval and KB-REC. The reason is that the diversity of natural language expressions poses a challenge in identifying knowledge segments based on specific rules, as they may appear in various positions within sentences. Moreover, the results demonstrate the flexibility of our segment detection method in recognizing knowledge and visual information in expressions, which improves the performance of both knowledge retrieval and KB-REC.

Method	Val	Test
A. Objects associated with knowledge		
Knowledge + expression	30.40	29.38
Knowledge + visual segment	**32.15**	**30.44**
B. Methods of associating knowledge		
Concatenation after attention	29.25	29.23
Addition after attention	29.83	29.73
Iterative attention	**32.15**	**30.44**

Table 5: Results of different objects associated with knowledge categories and results of different methods to handle multiple category-associated sentences.

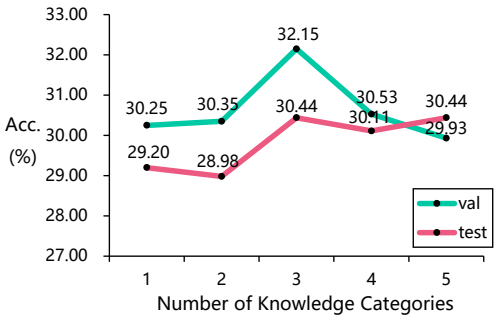


Figure 4: Results for different number of retrieved knowledge categories.

4.5 Evaluation of Knowledge Retrieval Method

In block A of Table 4, we compare the proposed method with the cosine similarity method in the baseline model ECIFA (Wang et al., 2020). Results show that replacing the entire expression in the cosine similarity method with the knowledge segments obtained by our segment detection module can significantly improve the performance of knowledge retrieval. Additionally, our prompt-based knowledge retrieval method significantly outperforms the sentence-level similarity method, indicating that our segment-level and category-oriented method can effectively alleviate the similarity bias problem. Moreover, we evaluate the inference time of different retrieval methods under the same settings. The results show that our prompt-based method demonstrates high efficiency in retrieving knowledge with 0.020s per sample, which is 70 times faster than the cosine similarity method in ECIFA which takes 1.400s.

Results in block B of Table 4 show that fine-tuning a language model using category masks improves the model's capacity to retrieve categories. It contributes to our category-oriented method and alleviates the interference of irrelevant information from unstructured knowledge sentences.

4.6 Evaluation of Knowledge-Based Grounding Method

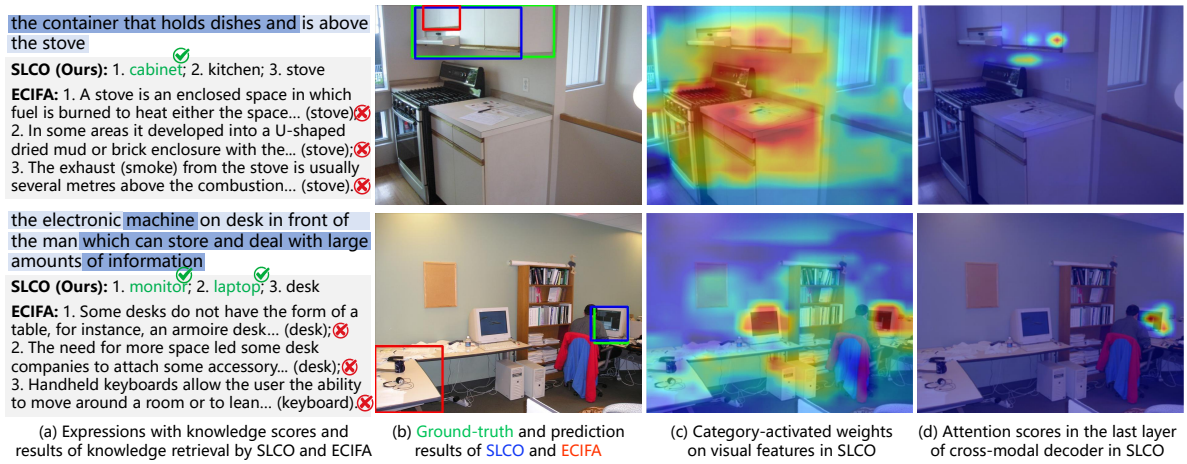
Block A of Table 5 shows comparison results for associating the textual features of knowledge categories and expressions. It can be seen that associating knowledge categories with visual segments is superior to that with the entire expressions. This is because visual segments concentrate on disambiguation at the instance level and reduce interference from irrelevant parts of expressions.

In block B of Table 5, we evaluate the performance of various methods for associating multiple category-associated sentences and visual features. There are three settings for the association: Multiple sentences and visual features are processed separately by multiple attention mechanisms, and then their results are (1) concatenated or (2) added together; and (3) the attention mechanism is iteratively applied to multiple sentences. The results indicate that the iterative method is the most effective, as it accumulates features from the top-1 category-associated sentence, which is more likely to contain accurate knowledge. In contrast, the concatenation method and the addition method treat all sentences equally, making it difficult to determine which sentences are more important.

The results in Fig. 4 show that the association of the top-3 knowledge categories with expressions achieves the best performance. As the number of knowledge decreases below three, the overall accuracy of knowledge retrieval diminishes, resulting in a degradation of grounding performance. Conversely, when the number of knowledge exceeds three, an excessive number of candidate knowledge categories may impede the model's ability to accurately associate knowledge for object localization.

4.7 Qualitative Results

As shown in Fig. 5(a) and Fig. 5(b), the baseline model is interfered with by the words "stove" and "desk" in expressions, leading to incorrect results of knowledge retrieval and object localization. In contrast, SLCO effectively avoids this problem by utilizing knowledge segments to retrieve knowledge categories. As shown in Fig. 5(a), SLCO accurately retrieves relevant knowledge categories based on the knowledge segments, and then activates multiple objects related to the retrieved categories in the visual features shown in Fig. 5(c). Then, in



(a) Expressions with knowledge scores and results of knowledge retrieval by SLCO and ECIFA

(b) Ground-truth and prediction results of SLCO and ECIFA

(c) Category-activated weights on visual features in SLCO

(d) Attention scores in the last layer of cross-modal decoder in SLCO

Figure 5: Visualization and comparison with the baseline model ECIFA. In Fig. (a), the text highlighted in dark blue indicates knowledge segments. Fig. (a) shows the top 3 out of the 50 sentences retrieved by ECIFA, and the word at the end of the sentence is the knowledge category to which the sentence belongs. Fig. (b) shows ground truth (green box) and the results of SLCO (blue box) and ECIFA (red box). Fig. (c) and Fig. (d) show the significance of weights and scores, with warmer colors representing higher.

Fig. 5(d), the decoder further refines the objects by visual segments "above the stove" and "in front of the man", thereby accurately localizing the referent. Qualitative results show that SLCO can solve the issues of similarity bias and interference by irrelevant information in knowledge sentences.

5 Conclusion

In this paper, we propose a segment-level and category-oriented network to endow the model with the ability to identify and utilize the knowledge and visual segments in a targeted manner. Specifically, the proposed method uses knowledge segments to retrieve knowledge, which addresses the similarity bias problem in the sentence-level method. Additionally, our category-oriented retrieval method can elicit knowledge categories from language models, mitigating the interference from irrelevant information in knowledge sentences. Experimental results demonstrate the effectiveness of the proposed method in addressing two limitations of the existing methods, thus improving the accuracy of the KB-REC task. In future work, we will explore more fine-grained information in expressions and combine it with knowledge and visual content.

Limitations

To better understand the limitations of the proposed method, we conducted an error analysis by randomly selecting 100 incorrect predictions and categorizing their error types. The results revealed that 32% of errors were caused by grounding issues, specifically, an inability to distinguish between multiple objects of the same category, despite having knowledge category of the referent object. The results indicate that there is a need for improvement in the ability to discriminate visual objects, especially for object categories with long-tailed distributions. Additionally, the results show that 20% of errors are due to imprecise object detection, particularly for small objects. This highlights the need for optimization of the visual encoder and loss function. Moreover, 14% of errors are attributed to incorrect knowledge retrieval. To address this, incorporating more fine-grained information in expressions for retrieval should be considered as a future research direction. Furthermore, 34% of incorrect predictions can be attributed to issues with the ground-truth annotations, which may negatively impact the model's learning process.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62076100, 61976094, and 62276072), the Guangxi Natural Science Foundation (No. 2022GXNSFAA035627), Fundamental Research Funds for the Central Universities, SCUT (x2rjD2220050), the Science and Technology Planning Project of Guangdong Province (2020B0101100002), CAAI-Huawei MindSpore Open Fund, CCF-Zhipu AI Large Model Fund, and the Open Research Fund of Guangxi Key Laboratory of Multimedia Communications and Network Technology.

References

Arjun R. Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6555–6565. Association for Computational Linguistics.

Matthew Berg, Deniz Bayazit, Rebecca Mathew, Ariel Rotter-Aboyoun, Ellie Pavlick, and Stefanie Tellex. 2020. Grounding language to landmarks in arbitrary outdoor environments. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 208–215. IEEE.

Alex Bogatu, Zili Zhou, Dónal Landers, and André Freitas. 2022. Active entailment encoding for explanation tree construction using parsimonious generation of hard negatives. *CoRR*, abs/2208.01376.

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.

Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2018. Using syntax to ground referring expressions in natural images. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6756–6764. AAAI Press.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. 2021. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3064–3073. Computer Vision Foundation / IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. 2021. Look before you leap: Learning landmark features for one-stage visual grounding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16888–16897. Computer Vision Foundation / IEEE.

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.

Liuwu Li, Yuqi Bu, and Yi Cai. 2021. Bottom-up and bidirectional alignment for referring expression comprehension. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 5167–5175. ACM.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3942–3951. AAAI Press.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. REVERIE: remote embodied

visual referring expression in real indoor environments. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9979–9988. Computer Vision Foundation / IEEE.

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

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5203–5212. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 658–666. Computer Vision Foundation / IEEE.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4222–4235. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

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.

Peng Wang, Dongyang Liu, Hui Li, and Qi Wu. 2020. Give me something to eat: Referring expression comprehension with commonsense knowledge. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 28–36. ACM.

Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1960–1968. Computer Vision Foundation / IEEE.

Yefei Wang, Kaili Wang, Yi Wang, Di Guo, Huaping Liu, and Fuchun Sun. 2022. Audio-visual grounding referring expression for robotic manipulation. In *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*, pages 9258–9264. IEEE.

Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. 2022. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9489–9498. IEEE.

Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, pages 387–404. Springer.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1307–1315. Computer Vision Foundation / IEEE Computer Society.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5017–5033. Association for Computational Linguistics.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
After the conclusion in Section 5 and before the references.

☒ A2. Did you discuss any potential risks of your work?
The task in this paper does not identify potential risks and harmful effects for the time being.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The introduction in Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

The method in Section 3 and the experiments in Section 4.

☑ B1. Did you cite the creators of artifacts you used?
The method in Section 3 and the experiments in Section 4.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In the abstract, we claim that our code will be made available to the public.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
The implementation in Section 4.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We do not propose a new dataset.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We do not have documentation of the artifacts.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
The dataset in Section 4.

C ☑ Did you run computational experiments?

The implementation in Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
The implementation in Section 4.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The implementation in Section 4.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
The experiments in Section 4 and the limitations after the conclusion.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
The implementation in Section 4.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.