McQueen: a Benchmark for Multimodal Conversational Query Rewrite

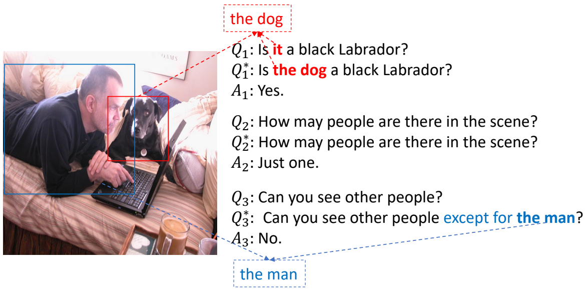
Yifei Yuan[1]*, Chen Shi[2], Runze Wang[2],
Liyi Chen[3], Feijun Jiang[2], Yuan You[2], and Wai Lam[1]
[1]The Chinese University of Hong Kong
[2]Alibaba Group
[3]Nankai University
{yfyuan,wlam}@se.cuhk.edu.hk
{deling.sc,yunze.wrz,feijun.jiangfj,youyuan.yy}@alibaba-inc.com
liyichen@mail.nankai.edu.cn

Abstract

The task of query rewrite aims to convert an in-context query to its fully-specified version where ellipsis and coreference are completed and referred-back according to the history context. Although much progress has been made, less efforts have been paid to real scenario conversations that involve drawing information from more than one modalities. In this paper, we propose the task of multimodal conversational query rewrite (McQR), which performs query rewrite under the multimodal visual conversation setting. We collect a large-scale dataset named McQueen based on manual annotation, which contains 15k visual conversations and over 80k queries where each one is associated with a fully-specified rewrite version. In addition, for entities appearing in the rewrite, we provide the corresponding image box annotation. We then use the McQueen dataset to benchmark a state-of-the-art method for effectively tackling the McQR task, which is based on a multimodal pre-trained model with pointer generator. Extensive experiments are performed to demonstrate the effectiveness of our model on this task[1].

1 Introduction

Recent years have witnessed an increasing attention in conversational-related tasks, such as conversational question answering (Choi et al., 2018; Reddy et al., 2019), visual conversation modeling (Das et al., 2017), etc. One main challenge in multi-turn conversation modeling is that information from context history is easy to be abbreviated or omitted in the follow-up queries, causing the so-called coreference and ellipsis. To address this concern, the task of query rewrite (Elgohary et al., 2019; Pan et al., 2019; Su et al., 2019) aims to reconstruct the original query to a fully specified form based on its history context. The rewrite eliminates the coreference and ellipsis in the original query without changing its semantic information, thus helping turn the more challenging multi-turn conversation modeling problem to a single-turn version.



Q_1: Is it a black Labrador?
Q_1^*: Is the dog a black Labrador?
A_1: Yes.

Q_2: How may people are there in the scene?
Q_2^*: How may people are there in the scene?
A_2: Just one.

Q_3: Can you see other people?
Q_3^*: Can you see other people except for the man?
A_3: No.

Figure 1: An example of the mulitmodal query rewrite task, where $\mathcal{Q}$, $\mathcal{Q}^*$, and $\mathcal{A}$ denote the queries, their corresponding rewrites and the answers. Red color denotes the coreference rewrite part and blue denotes the ellipsis rewrite part. Image boxes are utilized for representing the area of the entities in the rewrite.

Following this line, several attempts have been made in the query rewrite task which achieve decent performance on the natural language level. Nevertheless, conversations in real scenario tend to involve knowledge from more than one modalities, such as vision, text, speech, etc. Information from different modalities is not handled in isolation, but often integrated together to improve the quality of perception and understanding. For example, as shown in Figure 1, in the first turn of the visual conversation, with the lack of context, the user directly uses the pronoun "it" to refer to the dog in the image. In the third turn, for ellipsis that does not appear in the context history, in order to perform ellipsis completion, one also needs to find clues from the corresponding image. Rewriting the query under the circumstance where grounding

[*] Work done when Yifei Yuan was an intern at Alibaba. This work was supported by Alibaba Group through Alibaba Research Intern Program, and a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Codes: 14200620).

[1] The dataset and code of this paper are both available in https://github.com/yfyuan01/MQR

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4834–4844
December 7-11, 2022 ©2022 Association for Computational Linguistics

outside the text information is needed poses more challenges to traditional query rewrite models that based only on textual features.

In this paper, we propose the task of multimodal conversational query rewrite (McQR), which aims to perform query rewrite under the multimodal visual conversation setting. To achieve this goal, we collect a large-scale dataset called McQueen. Specifically, for each visual conversation consisting of an image and the corresponding history question-answer context, we provide manual rewrite for the query, where the coreference resolution and ellipsis completion are performed respectively. Furthermore, in order to assist downstream tasks such as coreference entity detection, for all the entities appearing in the rewrite, we annotate the image boxes for representing their corresponding image area.

We then use the McQueen dataset to benchmark a state-of-the-art method for effectively tackling the McQR task. Inspired by the big success of pre-trained models such as BERT (Devlin et al., 2019), our model is based on a multimodal pre-trained model where interactions between different modalities can be better captured. Furthermore, we enhance the model with a pointer generator specially designed for the multimodal Transformer blocks (Vaswani et al., 2017), so that the rewritten query is either generated from scratch or copied from certain contextual parts with high attention weights. Extensive experiments are conducted to compare our method with several state-of-the-art methods. Our model outperforms all the methods in both the McQR task and two subtasks. We further perform analysis to investigate the role of different modalities in this task and demonstrate that the introduction of image information provides extra guidance for the concerned query rewrite task.

In summary, the contribution of our paper lies in three folds:

- We formally define the task of multimodal conversational query rewrite (McQR), which aims to generate a fully-specified rewrite query based on both the context history and the visual image.
- We propose a large-scale dataset McQueen, containing 15k visual conversations and over 80k rewrites. For the entities appearing in the rewrites, we also annotate the image boxes for representing their corresponding image area.
- We benchmark a multimodal Transformer-based model with pointer mechanism for effectively tackling the McQR task. Extensive analysis shows the role of different modalities in our model.

2 Related Work

2.1 Query Rewrite

The task of query rewrite provides reconstructed queries based on abbreviated in-context queries without changing their semantic meaning. First introduced by (Elgohary et al., 2019; Su et al., 2019; Pan et al., 2019), most works formulate it as a standard generation task, which can be solved via a Sequence-to-Sequence model (Quan et al., 2019; Vakulenko et al., 2021; Anantha et al., 2021). Some attempts have been made to introduce a multi-task learning setup in order to enhance the training process (Rastogi et al., 2019; Song et al., 2020; Zhang et al., 2020). Some works seek to focus on query rewrite under the low-resource scenario (Yu et al., 2020; Voskarides et al., 2020; Yu et al., 2021). For modeling the linguistic knowledge in conversational context more effectively, prior knowledge is leveraged such as using semantic role labeling to provide extra guidance (Xu et al., 2020), and reducing the generation search space via sequence tagging (Hao et al., 2021). Although these works have achieved great performance on their corresponding task, query rewrite under the multimodal setting has not been explored.

2.2 Visual Coreference Resolution

Visual dialog entails answering a set of questions grounded by an image (Das et al., 2017). Based on that, visual coreference resolution involves linking the words in the text (usually nouns and pronouns) to a certain area in the image (Kong et al., 2014; Kottur et al., 2018). Following this line, Li et al. (2021) restrict coreference resolution to pronouns and resolve coreferences in visual dialog in an unsupervised way. Yu et al. (2019) define the task of visual pronoun coreference resolution where a dataset called VisPro and a model called VisCoref are benchmarked accordingly. Based on that, Yu et al. (2022) resolve pronoun coreference and propose a novel framework to improve visual dialog understanding. This task can be seen as a subtask of the McQR task where coreference resolution and ellipsis completion are both taken into account.

3 The McQueen Dataset

3.1 Dataset Overview

Our dataset is based on a visual dialog dataset called VisDial (Das et al., 2017). The original VisDial dataset consists of over 133k dialogs, each associated with an image and 10 rounds of question-answer pairs. All question-answer pairs are conducted in a conversational format and revolve around the content of the picture.

Our dataset randomly selects 15k conversations from the VisDial dataset with the total of over 80k rewrite utterances. For each query in a visual conversation, we conduct manual annotation to resolve the information omission. The query is reconstructed based on the image as well as the history context so that the coreference and ellipsis are referred-back or completed. For negative queries that do not contain any information omission, the rewrite stays the same as the original query. In addition, for all the entities appearing in the coreference and ellipsis, we annotate the image boxes for representing their corresponding image areas.

3.2 Dataset Construction

3.2.1 Text Rewrite Annotation

For manual annotation, we hire 16 annotators in total. Before the annotation starts, we provide 100 examples for all the annotators to refer to. We also provide a guideline and some tutorials by listing some typical coreference and ellipsis cases so that the bias and language style shift between individuals are minimized as possible. After that, the annotators start working on a small portion of data where query rewrite is performed. After all the results are returned and the data quality is checked, the main annotation phase begins and the rest of data is labeled. On average, each annotator is in charge with the rewrite of 5059 queries. The rewrite annotation interface can be seen in Appendix A.

3.2.2 Image Box Annotation

Besides the rewrite annotation, we also provide image annotation to assist downstream or related tasks (e.g. coreference entity detection). The image box annotation begins right after the rewrite annotation. The overall procedure also follows the (1) tutorial (2) trial phase (3) main phase pipeline. Specifically, the annotators have to extract the entities in the ellipsis and coreference part and draw the bounding boxes of them in the image. Each annotator is in charge of the image annotation of the rewrites written by him/herself. The image annotation interface can be seen in Appendix B.

3.2.3 Quality Control

After the all the annotation is finished, we re-group and shuffle the annotators to perform cross quality inspection. Each group is asked to check the annotation results of other groups. In addition, two new annotators who do not take part in the annotation phase are recruited to check the quality of all the annotation results. The annotators have to answer three questions for each query : (1) *Is the rewrite result correct or not?* (2) *Are all the coreference and ellipsis resolved in the rewrite?* (3) *Are the entities in the coreference and ellipsis correctly annotated in the image?* All the conversation rewrites must get the all "yes" result from all the annotators before official acceptance, otherwise they are collected to be revised and re-checked (the questionnaire interface is shown in Appendix C). The whole check-revise process lasts for three iterations. Considering chance agreement, we measured the Inter-Annotator Agreement (IAA) in terms of Cohen's κ (Cohen, 1960). The final κ score is 0.82, reaching the "almost perfect" level[2]. Besides, after each quality check iteration, we randomly sample 100 conversations from the dataset and manually evaluate the utterance-level precision and recall rate, where precision denotes the rate of the retrieved rewrites being correct, while the recall rate records the portion that the coreference and ellipsis being handled. The precision and recall rate in the 1st/2nd/3rd iteration are (89.0%, 87.1%)/(95.5%, 94.2%)/(98.3%, 98.2%), respectively.

3.2.4 Annotation Cost and Duration

The overall phase including the annotation and quality check spanned for 10 weeks (from March to May 2022), where the annotation guidance lasts for 2 weeks, data annotation lasts for 5 weeks, quality check lasts for 3 weeks. All the annotators are English native speakers recruited from a professional data management company Appen[3]. The annotation costs $5942 US dollars in total, with $0.31 per utterance rewrite, $0.03 per image box annotation.

3.3 Dataset Statistics

According to Table 2, 86.5% of our dataset covers positive rewrite cases that coreference or ellipsis

[2] According to (Landis and Koch, 1977), if $\kappa >= 0.81$

[3] https://appen.com/

	Task	Conv. Size	Utterance Size.	Modality	Coref.	Ellip.	Image Box
CANARD (2019)	Query Rewrite	5644	40527	t	✓	✓	✗
REWRITE (2019)	Query Rewrite	2000	40000	t	✓	✓	✗
RESTORE (2019)	Query Rewrite	203965	203965	t	✓	✓	✗
VisDial (2017)	Visual Dialog	133351	1333510	t&v	✗	✗	✗
VisPro (2019)	VCR	5000	29722	t&v	✓	✓	✗
McQueen	McQR	15000	80944	t&v	✓	✓	✓

Table 1: Comparison with existing datasets, where VCR denotes visual pronoun coreference resolution.

Characteristic	% Ratio
Queries w/ Coreference	57.2
Queries w/ Ellipsis	30.4
Queries w/o both	13.5

Table 2: Proportion of queries with coreference and ellipsis in our dataset.

Avg. Num. of Turn per Conv.:	5.40
Avg. Num. of Entity per Conv.:	3.04
Avg. Num. of Image Box per Conv.:	2.02
Avg. Length of Rewrites:	6.65
Avg. Length of Contexts:	42.13

Table 3: Statistics of the dataset. The length is calculated by the number of English words.

occurs in the query. Table 3 lists the statistics of our dataset, where each visual conversation contains 5.40 Q-A turns with 2.02 image boxes on average. Table 4 lists the number of rewrites in our dataset under different history context lengths, where most of the data contains context from 3-4 turns, and over 16k data contains context over 9 turns.

Futhermore, we compare our dataset with existing datasets from related works. According to Table 1, our dataset is: (1) *more complete* - we provide manual annotation on both image and text levels, where the rewrite query and entity boxes are both provided; (2) *more diverse* - since our dataset is designed not only for all the coreference cases, but also performs ellipsis completion in the rewrite; (3) *much larger* - compared with existing VCR dataset, the dataset has a larger scale.

4 McQR Task

4.1 Task Definition

We define the task of multimodal conversational query rewrite (McQR). Concerning a multimodal visual conversation with an image I and the corresponding text conversation D, at time step m, given the conversation history H with $m-1$ turns, the i-th turn consists of a query q_i and an answer a_i. The goal of the task is to generate a rewrite query r_m for the current query q_m based on both image I and textual history H, which can be represented as

$$r_m = Rewriter(H, I, q_m) \qquad (1)$$

0-2	3-4	5-6	7-8	9+	o/a
15143	17064	16431	16126	16180	80944

Table 4: Number of rewrites concerning different history turn lengths in McQueen.

4.2 Methods

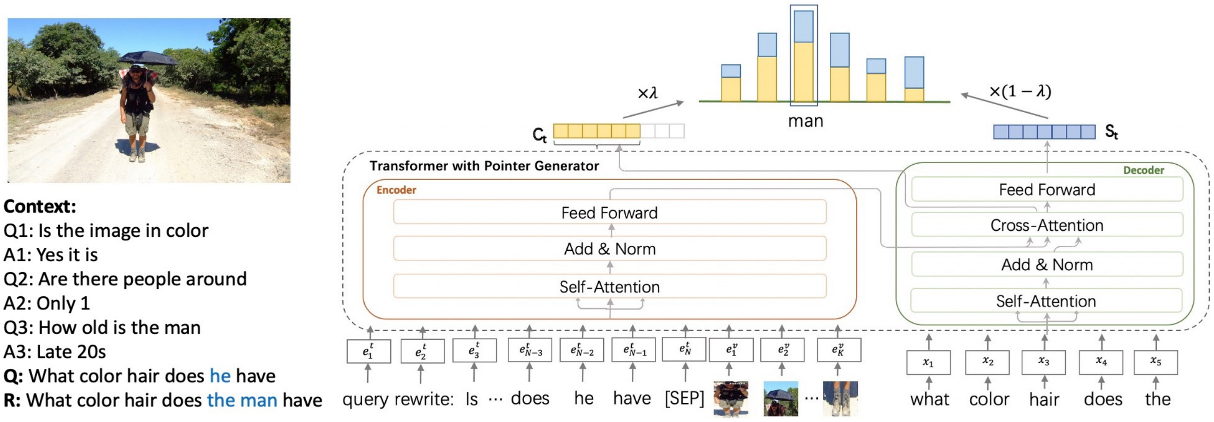
4.2.1 Model Overview

Figure 2 depicts the overall structure of our proposed model. Our model is based on a multimodal pre-trained model VL-T5 (Cho et al., 2021), where a Transformer-based encoder and decoder are jointly trained to perform generation. Besides, we add a pointer generator to the multimodal Transformer decoder so that the rewrite is either generated from scratch or copied from history context.

4.2.2 Image and Text Embeddings

Text Embedding The text input of our model includes two parts: the task prefix and the history context. The task prefix is a short prompt which aims to differentiate the concerned McQR task from other tasks in the pre-training phase (e.g. visual grounding). Specifically, we use “query rewrite:” as the task prefix. Following it, the history context contains all the utterances including all the queries and answers in previous turns. The input text is then tokenized and embedded before passed into the encoder. Following the setting in T5 (Raffel et al., 2020), we also add a relative position bias to each self-attention layer. As a result, the text input is represented as e^t.

Image Embedding To extract image features, we first detect several object regions from the image, denoted as Region of Interest (ROI). Following previous works (Cho et al., 2021; Lu et al., 2019), we also use Faster R-CNN (Ren et al., 2015) pretrained on the Visual Genome dataset (Krishna et al., 2016) to extract ROI features in our task.



Context:
Q1: Is the image in color
A1: Yes it is
Q2: Are there people around
A2: Only 1
Q3: How old is the man
A3: Late 20s
Q: What color hair does he have
R: What color hair does the man have

Transformer with Pointer Generator
Encoder
Feed Forward
Add & Norm
Self-Attention
Decoder
Feed Forward
Cross-Attention
Add & Norm
Self-Attention
c_t
s_t
×λ
×(1 − λ)
man
e_1^t e_2^t e_3^t e_{N-3}^t e_{N-2}^t e_{N-1}^t e_N^t e_1^v e_2^v e_K^v
query rewrite: Is ··· does he have [SEP] ···
x_1 x_2 x_3 x_4 x_5
what color hair does the

Figure 2: The overall architecture of our model.

For each image I, we extract $n = 36$ ROIs from it. The final visual vector can be represented as e^v.

4.2.3 Encoder-Decoder Structure

Encoder. We use a multimodal Transformer encoder-decoder structure to encode the image text features and generate the target rewrite. The multimodal encoder is a stack of L Transformers which take the image and text representations as input

$$h_0 = [e_1^t, .., e_m^t, e_1^v, ..., e_n^v] \tag{2}$$

$$h_i = FNN(MultiHead(h_{i-1}, h_{i-1}, h_{i-1})) \tag{3}$$

Decoder. Similarly, the decoder is also a stack of L Transformers. Given the decoder input x_t, the decoding step consists three phases, where the first sub-layer is a self-attention layer which can be represented as

$$d_i = MultiHead(d_{i-1}, d_{i-1}, d_{i-1}) \tag{4}$$

where $d_0 = x_t$. After that, the second sub-layer calculates the cross attention between the encoder outputs and the decoder self-attention result

$$s_i = MultiHead(d_i, h_L, h_L) \tag{5}$$

The third sub-layer is a position-wise fully connected feed-forward neural network, followed by a softmax layer to output the final probability

$$P_{vocab}(y_t) = \sum_{i:w_i \in V} \beta_{t,i} = Softmax(FNN(s_i)) \tag{6}$$

where β is the softmax score over the whole vocabulary V.

Multimodal Transformer with Pointer Generator. Additionally, following the motivation that most part of the rewrite sentence can be related to certain parts of the input context, we add a pointer generator (See et al., 2017) to the multimodal Transformer such that the rewrite is either generated from scratch or directly copied from the input. Specifically, the cross attention in the last decoder layer can be naturally taken as the context vector

$$c_i = MultiHead(d_i, h_L^U, h_L^U) \tag{7}$$

where h_L^U is the textual part of the encoding result, which is the embedding of tokens before the "[SEP]" token.

$$\alpha_{t,i} = softmax(\frac{(W_s s_t)^T W_h h_i}{\sqrt{d}}) \tag{8}$$

$$P_{copy}(y_t) = \sum_{i:w_i \in H} \alpha_{t,i} \tag{9}$$

where α is the copy distribution over the input H, P_{copy} determines where to copy at time step t.

By incorporating the pointer generator, the final probability of generating the target word y_t at time step t is represented as

$$P(y_t) = \lambda P_{vocab}(y_t) + (1 - \lambda) P_{copy}(y_t) \tag{10}$$

$$\lambda = sigmoid(w_d^T c_t + w_l^T s_t + w_a^T x_t) \tag{11}$$

where w_d, w_l, w_a are the weights to learn.

4.2.4 Training

We adopt the standard generation loss when fine-tuning the pre-trained model parameterized θ on our McQR task. At each time step t, the decoder output token is determined based on the generated rewrites of previous time steps denoted as $y_{<t}$, the input image and text encoding vector e^v and e^t. We minimize the negative log-likelihood of generating

	BLEU-2	BLEU-4	ROUGE-2	ROUGE-L	METEOR	EM
Original	57.27	48.44	54.86	73.45	38.28	- | -
AllenNLP Coref	59.72	50.21	56.09	76.07	39.87	8.97 | 96.14
L-Gen	77.14	66.31	74.44	84.68	46.09	41.84 | 57.77
T(E)-Gen	78.03	65.40	75.53	86.80	47.32	41.37 | 56.58
T(L)-Gen	79.30	67.16	76.28	88.14	49.23	42.06 | 58.25
L-Ptr	77.46	69.46	76.07	86.65	48.63	47.87 | 66.72
T(E)-Ptr	79.52	68.26	77.27	88.49	50.10	49.49 | 68.94
T(L)-Ptr	80.12	68.41	78.40	89.54	50.56	48.15 | 69.53
VLBart	90.01	84.28	88.67	93.80	58.87	64.28 | 86.89
VLT5	90.37	84.87	89.23	94.10	59.45	65.62 | 89.95
VLBart-Ptr	90.16	84.52	88.87	93.89	59.26	64.87 | 87.22
VLT5-Ptr	**90.47**	**84.94**	**89.32**	**94.45**	**59.46**	**65.67 | 90.22**

Table 5: The main experiment on our dataset, where T(E) denotes Transformer with Early Fusion, T(L) denotes Transformer with Late Fusion. For EM rate, we report both EM for positive and negative rewrite samples.

the target rewrite y given the image text input and previous generation results as

$$min{-}logP_\theta(y|e^t, e^v) = -\sum_{j=1}^{|y|} logP_\theta(y_t|e^t, e^v, y_{<t}) \tag{12}$$

5 Experiments

5.1 Compared Methods

We compare our methods with several baselines.

- Original (Elgohary et al., 2019) is the method where the rewrite is set to be the same as the input query.
- AllenNLP Coref (Gardner et al., 2018) is a deep learning based NLP tool. We utilize its coreference resolution model to generate the rewrite.
- (L/T)-Gen (Su et al., 2019) denotes the LSTM/Transformer based encoder-decoder generation model. For Transformer-based models, we report the performance of two variants: Early Fusion which utilizes the same encoder to encode image and text features, and Late Fusion which first embeds image and text into vector spaces separately and then performs fusion into a joint embedding.
- (L/T)-Ptr (See et al., 2017) adds a pointer generator to the (L/T)-Gen model.
- VL-(Bart/T5) (Cho et al., 2021) is the muliti-modal implementation of large pre-trained language model Bart/T5.
- VL-(Bart/T5)-Ptr is the model depicted in Section 4.2 that adds a pointer generator to the pre-trained model.

5.2 Experimental Settings

Our code is implemented based on Pytorch and Huggingface Transformers (Wolf et al., 2019). We use the base version of the pre-trained VL-Bart/T5 in all our experiments. Our dataset is split into the training, testing, and validation dataset following the portion of 60%, 20%, 20%, resulting in 48566 queries for the training set, 16189 queries for the testing and validation set. By default, we set the batch size as 32 and the learning rate to be 5e-5, the model is fine-tuned for 20 epochs with the random seed of 42. In the testing stage, all models decode words by beam search with beam size set to 5.

We employ several evaluation metrics to evaluate the quality of the generated rewrite. We first report the BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Denkowski and Lavie, 2014) rate. In addition, we report the Exact Match (EM) of both positive samples that involves some changes in the rewrite and negative samples that the rewrite is the same as the query. On two subtasks, we report the precision, recall, and F1 score.

5.3 Experiment Results

We first report the overall performance on the query rewrite task. Then we perform experiments on two subtasks including coreference resolution and ellipsis completion, respectively.

Table 5 demonstrates the overall performance of different models on the McQR task. We have the following observations: first of all, compared with LSTM based models (e.g. L-Gen), Transformer based models have a stronger generation ability (e.g. T-(E) Gen), where the BLEU-2 score increases from 77.14 to 78.03. In addition, with the help of the pointer generator, certain parts from the original query and the history context are able to be preserved in the rewrite, where the negative EM

	Precision	Recall	F1
AllenNLP Coref	15.76	14.48	15.10
L-Gen	28.63	28.18	28.40
T-Gen	42.12	41.45	41.78
L-Ptr	38.07	37.47	37.77
T-Ptr	43.88	43.19	43.53
MBERT	72.42	65.03	68.53
VLT5	77.00	70.84	73.79
VLBart	76.59	70.37	73.35
VLBart-Ptr	77.28	71.00	74.00
VLT5-Ptr	**77.33**	**71.05**	**74.05**

Table 6: Precision, recall, F1 score of visual coreference resolution. For Transformer based models, we record the late fusion version. All three metrics are evaluated on utterance level.

	Precision	Recall	F1
L-Gen	17.95	17.29	17.61
T-Gen	29.02	30.12	29.56
L-Ptr	24.49	23.59	24.03
T-Ptr	32.60	31.41	32.00
VLBart	51.93	50.04	50.97
VLT5	54.33	52.35	53.32
VLBart-Ptr	52.45	50.54	51.47
VLT5-Ptr	**54.59**	**52.60**	**53.58**

Table 7: Precision, recall, F1 score of visual ellipsis completion. For Transformer based models, we also report the late fusion version.

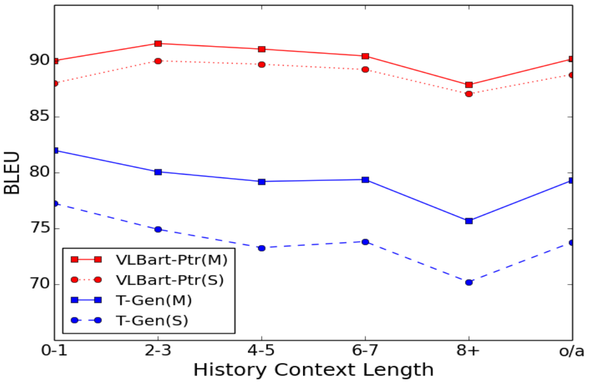


Figure 3: BLEU-2 performance under different length of turns, where (M) denotes the multimodal version, (S) denotes the text-only version, o/a denotes the overall performance over all turns.

score increases from 86.89 to 87.22 in the VLBart model. Furthermore, by leveraging the pre-trained model, a significant performance improvement can be seen where the BLEU-2 increases from 80.12 in T(L)-Ptr to 90.47 in VLT5-Ptr.

Table 6 shows the results of different models on the visual coreference resolution (VCR) subtask. We further utilize one state-of-the-art VCR method named MBERT to compare (Yu et al., 2022). According to the table, traditional text-only coreference resolution methods such as AllenNLP Coref may not have a good performance in the VCR task. The reason is that, in the conversations, many pronouns refer to entities that never appear in the history context, but can be found clues in the image. Furthermore, it can be concluded that pointer network also has a positive influence on the results since many coreferred entities can be directly copied from previous turns.

Besides VCR, we also perform tests on the visual ellipsis completion task, as shown in Table 7. The overall condition is similar to results in Table 6, while there encounters a performance degradation in this task. This may result from the fact that recovering the omitted information may not be that apparent (we will use some examples to illustrate in Section 6.3). In conclusion, our model shows superior performance on both two tasks, demonstrating the effectiveness of our model in resolving the information omission of the conversations.

6 Extensive Analysis

6.1 Image Role Analysis

In order to further investigate the role of image information in the query rewrite task, we remove the image features and compare the performance with the original model under different length of context respectively. Specifically, we divide the overall 16189 testing data into five categories by the history context length, ending up with 3136/3380/3180/3185/3308 data records with the context turn length from 0-1/2-3/4-5/6-7/8+, as shown in Figure 3. According to the figure, the performance gets improved with the help of image information. Compared with Transformer generator, our model is less sensitive to the turn length, where the model still has a decent performance when the conversation goes deep. We can also observe that, in our model, when there is no or short textual context that the information in conversation history is limited, the gap between monomodal and multimodal performance reaches to the largest, showing that the model is more dependant to the image information in this case. While when the conversation is deep, the gap decreases, showing that the model learns to copy information from the rich history context when rewriting the query.

6.2 Multimodal Attention Visualization

We utilize the self-attention weight heat map in the Transformer block of our model to visualize how

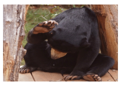
Context History		 **Q1**: Is it outdoors? **A1**: Yes. **Q2**: There is wooden floor outside? **A2**: Yes. **Q3**: Are there walls? **A3**: No. **Q4**: Is it pavilion? **A4**: Yes. **Q5**: Any tables? **A5**: No. **Q6**: Any people? **A6**: No.	**Q1**: How old is the girl? **A1:** Looks to be in her 20's. **Q2**: How old is the man? **A2:** About the same age. **Q3**: Is it a professional photo shoot? **A3:** No. **Q4**: What color is her hair? **A4:** Brown. **Q5**: What color is the brush? **A5:** White. **Q6**: What color is his hair? **A6:** Black.
Query	Is he alone?	Any animals?	Are they inside?
Ground-Truth	Is the boy alone?	Are there any animals except for the bear?	Are the girl and the man inside the room?
L-Gen	Is he alone?	Any animals?	Are the inside?
T(L)-Gen	Is he alone?	Any animals except for the people?	Are the man inside?
T(L)-Ptr	Is he alone?	Is there any animals except for the people?	Are the girl and man inside?
VLBart	Is the surfer alone?	Are there any animals except for the bear?	Are the man inside the room?
VLBart-Ptr	Is the boy alone?	Are there any animals except for the bear?	Are the girl and the man inside the room?

Figure 4: Example rewrites generated by different methods, where the red part denotes the coreference part and the blue part denotes the ellipsis part. The underline denotes the errors in the rewrites.

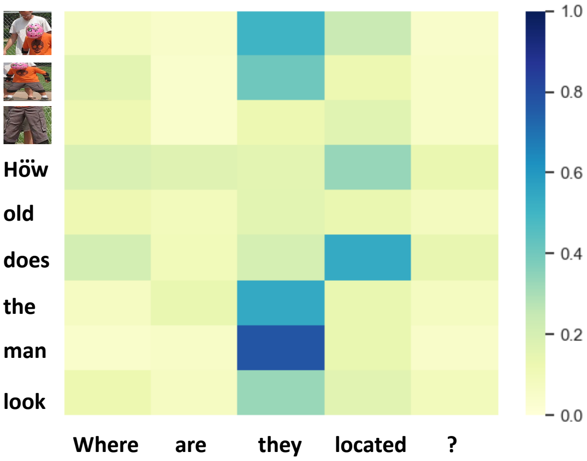


Figure 5: The heat map visualization of the self-attention weight in the Transformer block of our model. The full visual conversation example is shown in Appendix D.

the model learns the cross-modal relationship of the conversation. In the example shown in Figure 5 (the full conversation information can be found in Appendix D), the first three image segments correspond to ROI “man”, “boy”, “shorts”, respectively. According to the map, the pronoun “they” is correctly related to the “man” and “boy” ROI segments and also has a high attention weight to the text “the man” in the history context[4]. The learned relationship between visual and text representations can serve as the reason that our model shows superior performance concerning the multimodal feature incorporation in the McQR task.

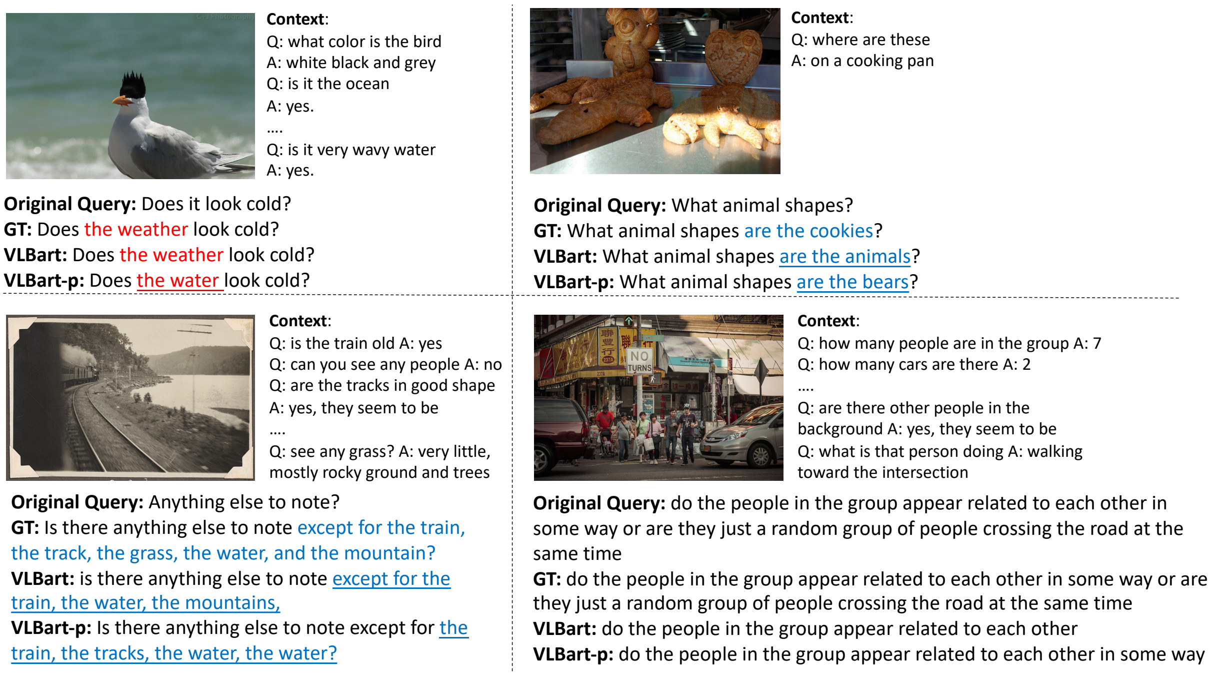
6.3 Case Study

We provide several example rewrites generated by different methods to vividly demonstrate the quality of rewrites, according to Figure 4. In the first case where the query is the first utterance where no textual context is provided, non pre-trained models (e.g. T(L)-Ptr) have difficulty resolving the pronoun “he” in the query. The results imply that with the help of prior knowledge, multimodal pre-trained models have a superior ability in incorporating image and text features, mitigating the gap between representations of multimodal heterogeneous space. The second case shows that traditional Transformer pointer network tends to copy incorrect context segments whose rewrite results may not make sense to human readers. For example, the abbreviated entity after “except for” should apparently be “the bear” instead of “the people”, while our model has a better ability of understanding the whole context and outputing more accurate results. Furthermore, when it comes to complicated cases where coreference and ellipsis are both observed, our model is both capable of copying the entities from history to solve the coreference (e.g. the red part in the third case) and generate parts that require reasoning (e.g. the blue part in the third case) from scratch.

7 Error Case Analysis

As shown in Figure 6, we demonstrate the most common error cases of our model, which can be classified into several cases including bad sentence structure, wrong coreference result, inaccurate omission completion, etc. According to the figure, we can see that pre-trained language models can generate sentences that have a proper format, but still have some problems in understanding the deep semantic correlation in the visual conversations. For example, in the top two cases, although the coreference seems simple for human readers, it poses challenges for machines other than directly making a copy from context. Specifically, for com-

[4] The pronoun “they” in the query refers to “the man and the boy” in the original conversation.



Context:
Q: what color is the bird
A: white black and grey
Q: is it the ocean
A: yes.
....
Q: is it very wavy water
A: yes.

Original Query: Does it look cold?
GT: Does the weather look cold?
VLBart: Does the weather look cold?
VLBart-p: Does the water look cold?

Context:
Q: where are these
A: on a cooking pan

Original Query: What animal shapes?
GT: What animal shapes are the cookies?
VLBart: What animal shapes are the animals?
VLBart-p: What animal shapes are the bears?

Context:
Q: is the train old A: yes
Q: can you see any people A: no
Q: are the tracks in good shape
A: yes, they seem to be
....
Q: see any grass? A: very little, mostly rocky ground and trees

Original Query: Anything else to note?
GT: Is there anything else to note except for the train, the track, the grass, the water, and the mountain?
VLBart: is there anything else to note except for the train, the water, the mountains,
VLBart-p: Is there anything else to note except for the train, the tracks, the water, the water?

Context:
Q: how many people are in the group A: 7
Q: how many cars are there A: 2
....
Q: are there other people in the background A: yes, they seem to be
Q: what is that person doing A: walking toward the intersection

Original Query: do the people in the group appear related to each other in some way or are they just a random group of people crossing the road at the same time
GT: do the people in the group appear related to each other in some way or are they just a random group of people crossing the road at the same time
VLBart: do the people in the group appear related to each other
VLBart-p: do the people in the group appear related to each other in some way

Figure 6: Common error cases of our method.

plicated cases such as when coreference and ellipsis appear at the same time, when the conversation is long, and when the omitted information requires reasoning between history entities (e.g. the bottom cases), there still remains much space to explore.

8 Conclusion and Future Work

We propose the task of multimodal conversational query rewrite (McQR), which aims to perform query rewrite under a multi-turn visual conversation. To facilitate the research, we benchmark a large-scale dataset with manual annotation which covers 15k visual conversations with more than 80k rewrites. We also provide image box annotation of entities appearing in the rewrites. Accordingly, we propose a model based on an existing multimodal pre-trained model and further enhance it with a pointer generator. Extensive experiments are performed to show the effectiveness of our model.

9 Limitations

Although our model achieves over 90% BLEU-2 rate on our dataset, there is still room for improvements in the future work, since the exact match for positive rewrite queries can be further lifted. In addition, in the future work, we will also continue to explore how the dataset and model can further benefit downstream research under this scenario. Furthermore, fine-grained image features such as the image box annotation can be leveraged for improving the performance.

10 Ethics Statement

All data records of our dataset are originally from the Visual Dialog (Das et al., 2017) dataset, where all the images are collected from the COCO (Lin et al., 2014) dataset. When annotating the data, we make sure that the annotators do not have any other rights except for the conversation information. Upon data publication, we will strictly follow the user terms of the Visual Dialog dataset.

References

R. Anantha, Svitlana Vakulenko, Zhucheng Tu, S. Longpre, Stephen G. Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *NAACL*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *ICML*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *EMNLP*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *EMNLP*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *ArXiv*, abs/1803.07640.

Jie Hao, Linfeng Song, Liwei Wang, Kun Xu, Zhaopeng Tu, and Dong Yu. 2021. Rast: Domain-robust dialogue rewriting as sequence tagging. In *EMNLP*.

Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? text-to-image coreference. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3558–3565.

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 *ECCV*.

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. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Mingxiao Li, K. U. Leuven, and Marie-Francine Moens. 2021. Modeling coreference relations in visual dialog. In *EACL*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

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 *ECCV*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Zhufeng Pan, Yan Wang, Kun Bai, Lianqiang Zhou, and Xiaojiang Liu. 2019. Improving open-domain dialogue systems via multi-turn incomplete utterance restoration. In *EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Jun Quan, Deyi Xiong, Bonnie Lynn Webber, and Changjian Hu. 2019. Gecor: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. *ArXiv*, abs/1909.12086.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Pushpendre Rastogi, Arpit Gupta, Tongfei Chen, and Lambert Mathias. 2019. Scaling multi-domain dialogue state tracking via query reformulation. In *NAACL*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149.

A. See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Shuangyong Song, Chao Wang, Qianqian Xie, Xinxing Zu, Huan Chen, and Haiqing Chen. 2020. A two-stage conversational query rewriting model with multi-task learning. *Companion Proceedings of the Web Conference 2020*.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance rewriter. In *ACL*.

Svitlana Vakulenko, S. Longpre, Zhucheng Tu, and R. Anantha. 2021. Question rewriting for conversational question answering. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Nikos Voskarides, Dan Li, Pengjie Ren, E. Kanoulas, and M. de Rijke. 2020. Query resolution for conversational search with limited supervision. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Kun Xu, Haochen Tan, Linfeng Song, Han Wu, Haisong Zhang, Linqi Song, and Dong Yu. 2020. Semantic role labeling guided multi-turn dialogue rewriter. In *EMNLP*.

Shih Yuan Yu, Jiahua Liu, Jingqing Yang, Chenyan Xiong, Paul N. Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Shih Yuan Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

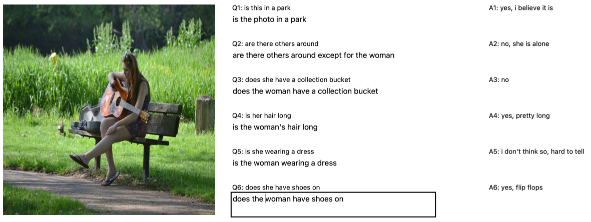
Xintong Yu, Hongming Zhang, Ruixin Hong, Yangqiu Song, and Changshui Zhang. 2022. Vd-pcr: Improving visual dialog with pronoun coreference resolution. *Pattern Recognit.*, 125:108540.

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 *EMNLP*.

Xiyuan Zhang, Chengxi Li, Dian Yu, Sam Davidson, and Zhou Yu. 2020. Filling conversation ellipsis for better social dialog understanding. In *AAAI*.

A Rewrite Annotation Interface

Figure 7 shows the annotation interface when preparing the rewrite. We assign one text box for each query in the visual conversation so that the annotators can rewrite the query according to the image and history context.


Q1: is this in a park
is the photo in a park
A1: yes, i believe it is

Q2: are there others around
are there others around except for the woman
A2: no, she is alone

Q3: does she have a collection bucket
does the woman have a collection bucket
A3: no

Q4: is her hair long
is the woman's hair long
A4: yes, pretty long

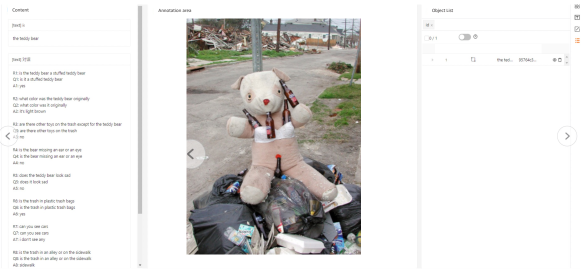
Q5: is she wearing a dress
is the woman wearing a dress
A5: i don't think so, hard to tell

Q6: does she have shoes on
does the woman have shoes on
A6: yes, flip flops

Figure 7: Rewrite annotation interface of our dataset.

B Image Box Annotation

Figure 8 shows the image box annotation interface, where the annotators extract entities from the rewrites and mark them on the image in the format of box.



[text] id
the teddy bear

[text] 对话
R1: is the teddy bear a stuffed teddy bear
Q1: is it a stuffed teddy bear
A1: yes

R2: what color was the teddy bear originally
Q2: what color was it originally
A2: it's light brown

R3: are there other toys on the trash except for the teddy bear
Q3: are there other toys on the trash
A3: no

R4: is the bear missing an ear or an eye
Q4: is the bear missing an ear or an eye
A4: no

R5: does the teddy bear look sad
Q5: does it look sad
A5: no

R6: is the trash in plastic trash bags
Q6: is the trash in plastic trash bags
A6: yes

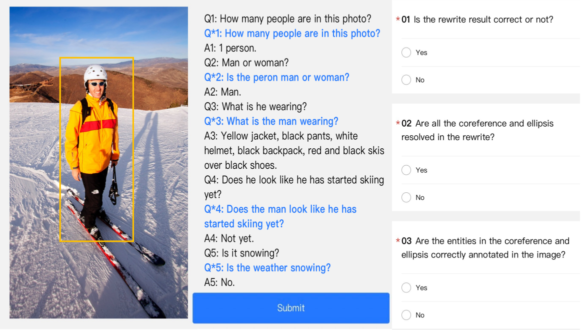
R7: can you see cars
Q7: can you see cars
A7: i don't see any

R8: is the trash in an alley or on the sidewalk
Q8: is the trash in an alley or on the sidewalk
A8: sidewalk

Figure 8: Image box annotation interface of our dataset.

C Quality Control Questionnaire

We design the quality control questionnaire as shown in Figure 9. All quality checkers have to click and answer the three questions.


Q1: How many people are in this photo?
Q*1: How many people are in this photo?
A1: 1 person.
Q2: Man or woman?
Q*2: Is the peron man or woman?
A2: Man.
Q3: What is he wearing?
Q*3: What is the man wearing?
A3: Yellow jacket, black pants, white helmet, black backpack, red and black skis over black shoes.
Q4: Does he look like he has started skiing yet?
Q*4: Does the man look like he has started skiing yet?
A4: Not yet.
Q5: Is it snowing?
Q*5: Is the weather snowing?
A5: No.

*01 Is the rewrite result correct or not?
○ Yes
○ No

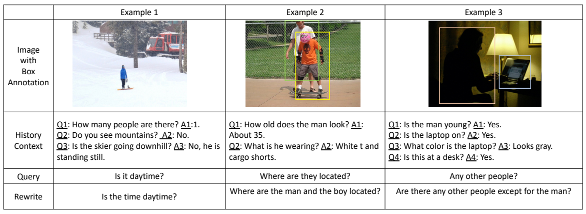
*02 Are all the coreference and ellipsis resolved in the rewrite?
○ Yes
○ No

*03 Are the entities in the coreference and ellipsis correctly annotated in the image?
○ Yes
○ No

Figure 9: The questionnaire interface in the quality control phase.

D Dataset Examples

We list some examples of our dataset to vividly demonstrate the characteristic of the task. All the examples are listed in Figure 10.



	Example 1	Example 2	Example 3
Image with Box Annotation			
History Context	Q1: How many people are there? A1:1. Q2: Do you see mountains? A2: No. Q3: Is the skier going downhill? A3: No, he is standing still.	Q1: How old does the man look? A1: About 35. Q2: What is he wearing? A2: White t and cargo shorts.	Q1: Is the man young? A1: Yes. Q2: Is the laptop on? A2: Yes. Q3: What color is the laptop? A3: Looks gray. Q4: Is this at a desk? A4: Yes.
Query	Is it daytime?	Where are they located?	Any other people?
Rewrite	Is the time daytime?	Where are the man and the boy located?	Are there any other people except for the man?

Figure 10: Example cases of our dataset.