Exploring Union and Intersection of Visual Regions for Generating Questions, Answers, and Distractors

Anonymous ACL submission

Abstract

Multiple-choice visual question answering (VQA) is to automatically choose a correct answer from a set of choices after reading an image. Existing efforts have been devoted to a separate generation of an image-related question, a correct answer, or challenge distractors. By contrast, we turn to a holistic generation and optimization of questions, answers, and distractors (QADs) in this study. This integrated generation strategy eliminates the need for human curation and guarantees information consistency. Furthermore, we first propose to put the spotlight on different image regions to diversify QADs. Accordingly, a novel framework ReBo is formulated in this paper. ReBo cyclically generates each QAD based on a recurrent multimodal encoder, and each generation is focusing on a different area of the image compared to those already concerned by the previously generated QADs. In addition to traditional VQA comparisons with state-of-the-art approaches, we also validate the capability of ReBo in generating augmented data to benefit VQA models.

1 Introduction

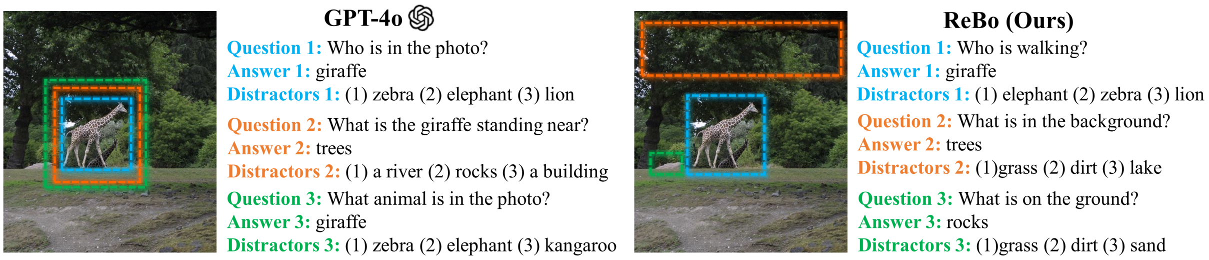
Visual Question Answering (VQA) (Antol et al., 2015; Goyal et al., 2017; Krishna et al., 2017) represents a burgeoning research domain that necessitates the development of algorithms capable of responding to arbitrary natural language questions of a given image. A specific subset of VQA, known as multiple-choice (MC) VQA (Zhu et al., 2016; Kembhavi et al., 2017; Lu et al., 2022b), involves the algorithm choosing the correct answer from a predefined list of distractors. MC-VQA, which requires vision-language understanding and cross-modality reasoning, is the representative benchmark for Large Vision-Language Models (LVLMs) (Zhu et al., 2023; Liu et al., 2024c; Dai et al., 2024). In the era of large models, the imperative for large-scale, high-quality MC-VQA datasets has become increasingly pronounced.

The traditional process of manually generating data is both labor-intensive and error-prone. Many automated methods are available today to independently generate questions (Zhang et al., 2016), answers (Li et al., 2018), and distractors (Lu et al., 2022a) (QADs) by machines based on images. However, these machine-generated QADs are often created independently, making it challenging to ensure intrinsic dependencies between them. To address this issue and enhance the capabilities of large models in vision-language understanding and cross-modality reasoning, our work focuses on the unified generation of QADs.

In the process of jointly generating QADs, how to comprehensively understand an image and diversify its generated QADs is rarely touched. As illustrated in Figure 1, the three bounding boxes focused on by GPT-4o are significantly intersected, inducing redundant questions such as “who is in the photo” and “what animal is in the photo”. In contrast, the QADs generated by our model, ReBo, are semantically rich and comprehensive for comprehending the image, as a broad union region with small intersections is concentrated on.

In the long run, addressing the above challenge come down to how to align image understanding across QADs. We tackle this issue in two folds. First, we automate the generation of QADs in a unified manner, ensuring a consistent image understanding from questions to answers and distractors. Next, we research the generation of a series of QADs by diversifying their focuses across image regions, which prevents information redundancy and provides a comprehensive understanding of the entire image.

From the methodological point of view, we introduce a **Re**current multimodal encoder to generate groups of QADs considering the **Bo**unding boxes (ReBo) of the given image. ReBo takes the



GPT-4o

Question 1: Who is in the photo?
Answer 1: giraffe
Distractors 1: (1) zebra (2) elephant (3) lion

Question 2: What is the giraffe standing near?
Answer 2: trees
Distractors 2: (1) a river (2) rocks (3) a building

Question 3: What animal is in the photo?
Answer 3: giraffe
Distractors 3: (1) zebra (2) elephant (3) kangaroo

ReBo (Ours)

Question 1: Who is walking?
Answer 1: giraffe
Distractors 1: (1) elephant (2) zebra (3) lion

Question 2: What is in the background?
Answer 2: trees
Distractors 2: (1)grass (2) dirt (3) lake

Question 3: What is on the ground?
Answer 3: rocks
Distractors 3: (1)grass (2) dirt (3) sand

Figure 1: An example of the vision regions that different QADs focus on. Compared with GPT-4o, our model generates semantically rich QADs and provides a more comprehensive understanding of the entire image.

QADs generated in previous steps as part of the input to generate QAD in the next step. In addition, ReBo considers the union and intersection of image bounding boxes, ensuring that each group of QADs focuses on diverse regions. In this way, ReBo disperses its attention on a broad area of the image and boosts the diversity of the generated QADs. We conduct extensive experiments to validate the performance of ReBo in different scenarios. Moreover, a further experimental analysis suggests that the QADs generated by ReBo can be used to promote existing VQA models in VQA tasks.

Our main contributions are listed as follows:

- We propose a recurrent multimodal encoder-based framework ReBo to jointly generate a series of QADs for an image in a unified way.
- We introduce to diversify QAD generations by broadening observation and insight for a comprehensive understanding of an image.
- We conduct quantitative and qualitative evaluations which demonstrate that ReBo can lead to excellent performance in diverse scenarios.
- We validate the superiority of our generated QADs in improving existing VQA models.

2 Related Work

Most prior research focused on generating a part or parts of QADs, that is, question, answer, or distractors. For instance, the studies of Visual Question Generation aim at generating questions related to an image or a video. Zhang et al. (2016) took images and captions as inputs to generate questions with different types. Johnson et al. (2016) introduced Densecap to produce region captions, providing additional context to steer the process of question generation. Krishna et al. (2019) formulated a visual question generation framework by optimizing the mutual information between the generated question and the pair of image and anticipated answer. Shen et al. (2020) explored a visual question generation approach based on a Double Hint strategy concerning textual answers and regions of visual interests.

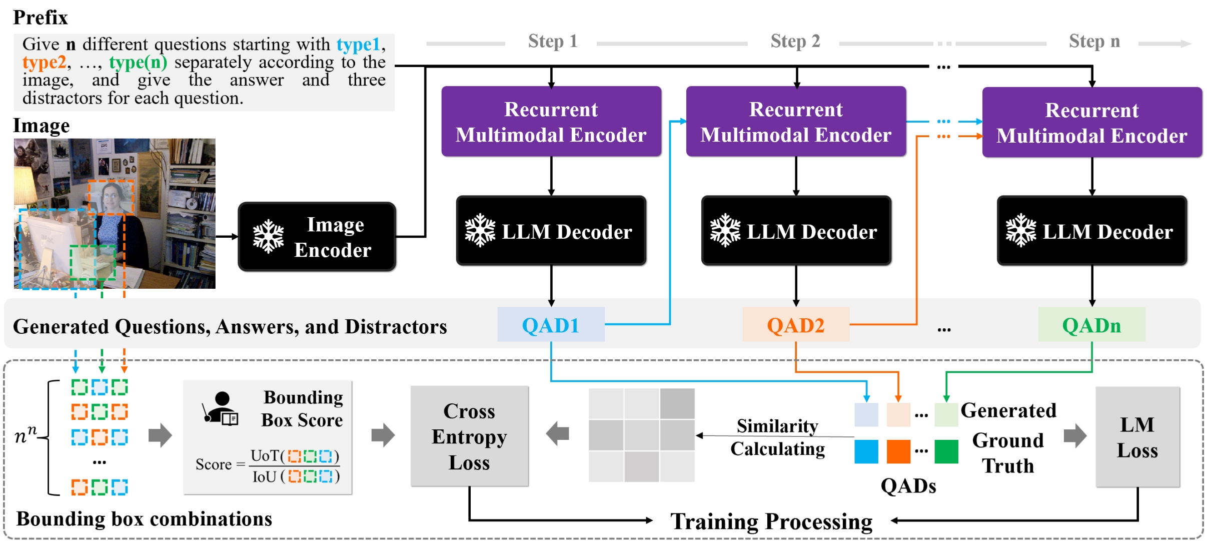
On the other hand, the studies of VQA deploy attention on generating correct answers by understanding images, questions, and their interactions. For example, Li et al. (2018) proposed iQAN by taking Visual Question Generation as a dual task to improve VQA performance. Xiong and Wu (2020) designed question-generating mechanisms and encouraged collaborative learning interactions among question-answering agents. Changpinyo et al. (2022) used neural models to generate textual questions and question answering. In recent years, some research has broken into the joint generation of question-answer pairs. Yang et al. (2021) employed variational inference to generate question-answer pairs considering diversity and consistency. Su et al. (2021) presented an end-to-end Generator-Pretester Network, which generated question-answer pairs from videos.

In contrast to Visual Question Generation and VQA, Visual Distractors Generation is a newly rising research field, which targets to generate challenging distractors according to the image, question, and answer. For example, Lu et al. (2022a) introduced a reinforcement learning approach to generate distractors in the context of visual images.

In this study, we explore a joint generation of groups of QADs as well as take into account their diversified discriminative correlations. Our proposed framework is capable of capturing the information from a broad region of the image, thereby enhancing the diversity and contextuality of the generated QADs.

3 Our Method: ReBo

We propose the unified framework ReBo to generate QADs as diverse as possible. In this section, we first introduce the model architecture in Section 3.1. Then, we describe the recurrent multimodal



Prefix
Give n different questions starting with type1, type2, …, type(n) separately according to the image, and give the answer and three distractors for each question.
Image
Image Encoder
Step 1
Step 2
Step n
Recurrent Multimodal Encoder
Recurrent Multimodal Encoder
Recurrent Multimodal Encoder
LLM Decoder
LLM Decoder
LLM Decoder
Generated Questions, Answers, and Distractors
QAD1
QAD2
...
QADn
n^n
Bounding Box Score
Score = UoT( ) / IoU( )
Cross Entropy Loss
Similarity Calculating
QADs
Generated Ground Truth
LM Loss
Training Processing
Bounding box combinations

Figure 2: The model architecture of ReBo. We freeze the Image Encoder and LLM Decoder and introduce a Recurrent Multimodal Encoder to generate various QADs. The Recurrent Multimodal Encoder module takes the prefix and previously generated QADs as text inputs and helps the LLM decoder to generate QADs in each step. We also use IoU and UoT of to guide the generation. The training processing will be removed during inference.

encoder in Section 3.2, followed by the details of the diversifying QAD generations in Section 3.3.

3.1 Model Architecture

Our model comprises an image encoder, a recurrent multimodal encoder, and a LLM decoder. We freeze the parameters of the image encoder and the LLM decoder, and train the recurrent multimodal encoder.

Given n groups of QADs to be generated for a given image, we divide the generation process into n steps. In each generation step, the recurrent multimodal encoder takes all of the QADs generated in previous steps as part of the text input to help the LLM decoder generate the QAD at current step. At each step, the generated QAD will focus on a different area of the image. After n steps, the Rebo model will generate QADs considering the union and intersection of diverse visual regions.

As shown in Figure 2, an image is fed into the frozen image encoder to obtain its visual representation. On the other hand, the text representation is composed of two elements: a fixed prefix and the ground truth QADs. The fixed prefix contains the number of QADs and the type information of each question, and the ground truth QADs comprise all of the QADs in previous steps. In specific, the input text in step i is the concatenation of the fixed prefix and all of the ground-truth QADs in previous $i-1$ steps. The recurrent multimodal encoder takes both the visual representation and text representation as inputs, and the frozen LLM decoder predicts one single QAD in each step.

We record the language modeling loss in each step and accumulate them as the total language modeling loss. An additional cross-entropy loss is introduced to optimize the predicted QADs, and its combination with the total language modeling loss is taken as the final loss function of ReBo.

To ensure that the generated QADs have a comprehensive understanding of the total image and share less redundant information, we present a novel mechanism to analyze the union and intersection of regions of interest in the image focused on by various QADs, which will be introduced in Section 3.3.

3.2 Recurrent Multimodal Encoder

For a global optimum, simultaneously generating and optimizing n groups of QADs is suggested. A straightforward solution is to use only one decoder to generate a unified representation of all groups of QADs. However, this method cannot model the specific representation of each individual QAD as well as their inherent correlations. These are crucial for generating an informative and comprehensive QADs combination, as will be analyzed in Section 3.3. Therefore, we design a recurrent multimodal encoder module to cyclically generate each group of QADs from a single input image.

To generate n groups of QADs for a given image, we divide the generation process into n steps. In each step, we recurrently utilize the recurrent multimodal encoder to help the LLM decoder generate different QADs. To be more specific, the recurrent multimodal encoder takes the image feature of this image as the visual input, and the text input in each step is formed by concatenating the prefix and all of the previous ground-truth QADs in the training process. As portrayed in Figure 2, the text input in step 1 is merely the prefix, that in step 2 is the prefix and the ground-truth QAD1, and that is the prefix, ground-truth QAD1, and ground-truth QAD2 in step 3. In contrast, the output of the LLM decoder in each step is a single group of QAD. All groups of QADs will be generated cyclically according to the recurrent multimodal encoder and LLM decoder for the given image. During the inference process, we replace the ground truth with the predict result of the LLM decoder in each step.

3.3 Diversifying QAD Generations

One bounding box can help induce a group of QAD, and we can obtain n groups of QADs for the given image with n bounding boxes. To make the generated QADs focus on diversified image regions, we evaluate the scores of different bounding boxes combinations of and employ these scores to supervise the QADs generation, as illustrated in Figure 2.

Given an image with n bounding boxes and R_i representing the i-th one, we can obtain its bounding box combination set C as follows:

$$C = R^n = R \times ... \times R, R = \{R_i\}_{i=1}^{n}, \quad (1)$$

where R^n denotes the n-fold Cartesian product of the bounding box set R. The cardinality of C is n^n, and its each element represents a possible combination of bounding boxes based on which we can induce groups of QADs.

Then, we introduce Intersection over Union (IoU) and Union over Total (UoT) to score each element in C. The IoU of the k-th bounding box combination C_k is defined as follows:

$$IoU_k = \frac{\sum_{R_i,R_j \in C_k, i \neq j} \left(R_i \bigcap R_j / R_i \bigcup R_j \right)}{n(n-1)/2}. \quad (2)$$

IoU_k denotes the intersection region of the bounding boxes in C_k, and a higher score typically implies more redundant discriminative information provided by C_k.

In addition to reduce the intersection attention region of different QADs, we also expect to enlarge the total union attention region of all QADs to cover as much of the image area as possible. Therefore, we define the UoT of C_k as follows:

$$UoT_k = \frac{\bigcup_{R_i \in C_k} R_i}{H \times W}, \quad (3)$$

where H and W denote the height and width of the image, respectively.

Finally, we can obtain the score vector s whose each element describes the overall score of each bounding box combination as follows:

$$s = [\ s_k\ ]_{k=1}^{n^n}, s_k = \frac{UoT_k}{IoU_k}. \quad (4)$$

The score vector s can serve as the ground truth to guide ReBo in generating diverse QADs. That is, we can minimize the soft cross-entropy loss between s and the prediction probability p to generate less redundant and more comprehensive QADs. Suppose the embeddings of n predicted QADs $E = [\ e_i\ ]_{i=1}^{n}$ and the ground-truth embeddings $E^* = [\ e_j^*\ ]_{i=1}^{n}$. Their cosine similarities can be calculated as

$$sim(e_i, e_j^*) = \frac{{e_i}^{\mathrm{T}} e_j^*}{\left\| e_i \right\| \left\| e_j^* \right\|}. \quad (5)$$

A large $sim(e_i, e_j^*)$ indicates a high probability of predicting the j-th QADs as the i-th one. Then, the prediction probabilities of all of the possible bounding box combinations can be calculated as

$$p = [\ p_k\ ]_{k=1}^{n^n}, p_k = \prod_{R_i,R_j \in C_k} sim(e_i, e_j^*), \quad (6)$$

where e_i and e_j^* are the predicted embedding and ground-truth embedding of QAD_i and QAD_j induced respectivley from the region R_i and R_j.

The final loss function of ReBo is defined as

$$Loss = \sum_{i=1}^{n} LM_i + H(s, p), \quad (7)$$

where LM_i denotes the language modeling loss at the step i, s is the score vector in Eq. (4), p is the prediction probability in Eq. (6), and $H(s, p)$ represents their cross entropy.

4 Experiments

4.1 Datasets and Metrics

Visual7W. Visual7W (Zhu et al., 2016) is collected on 47,300 COCO (Lin et al., 2014) images, consisting of 327,939 QA pairs together with 1,311,756 multiple-choices. We refer to telling QA of Visual7W in our experiments and take no extra operations. Each question starts with one of six Ws, what, where, when, who, why, and how. We only select the QADs that contain bounding boxes from the dataset. To cover as many regions of the image with as few QADs as possible, for images containing QADs up to 3, we calculate the bounding box scores for all possible combinations of three bounding boxes associated with QADs. The QADs combination with the highest bounding box score is selected as the corresponding QADs for each image. We also remove the images that only have one QAD. The final dataset contains 8k/5k images and 21k/13k QADs for training and testing.

A-OKVQA. A-OKVQA (Schwenk et al., 2022) is a knowledge-based visual question-answering benchmark. A-OKVQA is an augmented successor of OK-VQA (Marino et al., 2019) and contains a diverse set of 17.1k/1.1k/6.7k questions/answer/rationale triplets for training/validation/testing. We use the A-OKVQA dataset to assess whether the generated QADs of ReBo can enhance existing VQA models.

Metrics. We employ BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015) with ground-truth QADs to evaluate the quality of the generated QADs.

4.2 Baselines

We compare ReBo with the following models:

- **VisualBert**† (Li et al., 2020) is a pre-trained vision-and-language encoder for multimodal understanding, and we add a Bert decoder to generate QADs.
- **BLIP**† (Li et al., 2022) proposes a novel dataset bootstrapping method CapFilt, a captioner capable of generating synthetic captions given noisy web images, and a filter designed to eliminate the noisy texts.
- **BLIP2**† (Li et al., 2023) adapts frozen large language models to understand visual features extracted from the frozen image encoder in image-to-text generation tasks.
- **VQADG**† (Ding et al., 2024) first presents to generate questions, answers, and distractors in a unified way. This paper also incorporates contrastive learning to improve the quality of QADs.
- **Qwen-VL**† (Bai et al., 2023b) is a large vision-language model based on language model (Bai et al., 2023a). We select Qwen-VL-Chat in this paper, which is a multimodal LLM-based AI assistant trained with human alignment techniques.

We also compare ReBo with LLMs, including Llama-2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), ChatGPT (Ouyang et al., 2022), Qwen1.5 (Team, 2024b), and Llama-3 (Team, 2024a), as well as LVLMs, involving LLaVA-1.5 (Liu et al., 2024a), CogVLM (Wang et al., 2023), and LLaVA-NeXT (Liu et al., 2024b). The implementation details can be found in Appendix B. The source code of our model will be released once acceptance.

4.3 Results and Analysis

In this section, we will introduce the performance of ReBo and validate the performance of the generated QADs in promoting existing VQA models. We will also conduct human evaluations and case studies to demonstrate the effectiveness of ReBo.

4.3.1 Main Results

For LLMs and LVLMs, we provide examples and instruct the LLMs to generate QADs, and image captions are employed. The prompts used for LLMs and LVLMs are provided in Appendix A. We retrain all of the V&L baseline models on the same dataset. We extend two variants of generation type to conduct a more comprehensive evaluation of the recurrent multimodal encoder. The concatenation generation type implies that the QADs associated with one image are generated at once in a naive manner, which means the output would be "QAD1<sep>QAD2<sep>QAD3". The recurrent generation type entails generating QADs for each step using the recurrent multimodal encoder, which means the output would be "QADi" in step i. All V&L baseline models are retrained in the concatenation generation type. We evenly partitioned the entire dataset into ten subsets and calculated the mean and variance of the results over ten runs.

The experimental results of generating QADs on the benchmark are summarized in Table 1, from which we can observe that: (1) the performance of ReBo is promising across five metrics, and (2) Llama-3, LLaVA-1.5, and Qwen-VL achieve peak performance respectively in the families of LLMs, LVLMs, and V&L models. Table 2 further summarizes the separate evaluation results for questions, answers, and distractors. We can conclude that: (1) ReBo can generate more image-related questions,

Model	FT	V&L	PLM	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
Llama-2	✗	✗	Llama-2-7B-Chat	17.02±4.28	2.52±0.42	25.41±1.57	21.73±6.27	8.65±7.14
Mistral	✗	✗	Mistral-7B-Instruct-v0.2	18.69±0	2.95±0	26.70±0	23.69±0	13.13±0
ChatGPT	✗	✗	GPT-3.5-Turbo	21.23±0.01	2.37±0	25.46±0	23.28±0.01	6.61±0
Qwen1.5	✗	✗	Qwen1.5-7B-Chat	21.55±0.01	3.93±0	27.58±0	25.38±0.01	14.03±0.03
Llama-3	✗	✗	Llama-3-8B-Instruct	24.61±0	4.77±0	28.78±0	27.84±0.01	23.09±0.09
LLaVA-NeXT	✗	✓	Mistral-7B-Instruct-v0.2	19.83	2.89	24.96	20.32	8.45
CogVLM	✗	✓	Vicuna-7B-v1.5	23.02	5.67	26.16	23.43	14.49
LLaVA-1.5	✗	✓	Vicuna-7B	27.5	6.56	28.28	27.36	22.34
VisualBert†	✓	✓	BERT	19.52±6.44	3.77±0.41	25.29±0.05	22.19±2.26	10.18±16.83
BLIP†	✓	✓	BERT	23.76±2.11	6.53±0.35	26.35±0.14	26.20±0.62	9.62±8.80
BLIP2†	✓	✓	FlanT5-XL	27.91±0.33	7.13±0.21	28.30±0.11	28.29±0.23	34.88±8.56
VQADG†	✓	✓	T5	28.72±0.83	7.20±0.15	27.22±0.04	29.73±0.23	30.89±1.59
Qwen-VL†	✓	✓	Qwen-7B-Chat	29.34±0.32	7.62±0.11	26.70±0.11	29.62±0.08	34.45±2.21
ReBo	✓	✓	FlanT5-XL	**31.19±0.63**	**9.40±0.19**	**29.52±0.08**	**31.78±0.49**	**48.28±7.60**

Table 1: Performance evaluation for different models on the Visual7W dataset. FT denotes a fine-tune model, V&L denotes a vision and language model, PLM denotes a pre-trained language model, and "†" denotes our re-implementation.

decent answers, and challenging distractors with a superiority ranging from 2-11%, and (2) the performance gap of VQADG behind ReBo indicates that simply concatenating the single part of QADs is not a promising strategy, which is consistent with the argument in Introduction.

4.3.2 Augmenting VQA models

To verify the boosting effects of ReBo over existing VQA models, we employ the QADs generated by ReBo as additional data to train the InstructBLIP on the VQA task in this section.

To ensure fairness, we use ReBo to generate QADs according to the images from the validation split dataset of the Visual7W, we then train a VQA model separately on Visual7W and Visual7W+generated dataset, and finally evaluate the accuracy on the A-OKVQA dataset. To ensure the diversity of the generated QADs, we extract three question types at a time from all six question types (e.g., "what", "where", and "when" for one iteration) for ReBo to generate QADs. 500k QADs can be yielded as training data after 300 iterations. Then, we filter high-quality QADs respectively from the views of questions and answers: (1) For questions, we select the QADs with less overlapped information with the ground truth based on their cosine similarities; (2) as to answers, we calculate the cosine similarities between our generated answers and the pseudo-answers generated by InstructBLIP, and preserve those with high similarities as the final augmented data. After filtering, the final QADs are used as the augmented data to train the VQA model InstructBLIP.

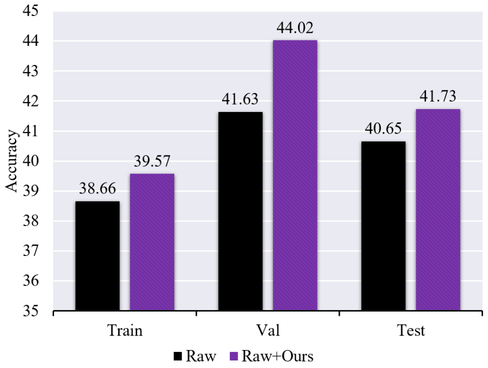


Figure 3: Augmenting existing VQA model. Raw denotes the model trained only on the raw Visual7W dataset. Raw+Ours denotes the model trained on both the raw Visual7W and the generated dataset.

To ensure the generalization of this evaluation, we employ the A-OKVQA dataset for testing in addition to the QADs generated on the Visual7W dataset for training as aforementioned. The performance is depicted in Figure 3. It can be observed that the vision-language capability of InstructBLIP is boosted by our generated QAD data over training, validation, and testing splits of A-OKVQA, ranging from 0.91 to 2.93 points. It is noteworthy that our proposed method is model-agnostic and it can be applied to any model on any benchmark.

4.3.3 Ablation Study

We conduct ablation experiments to verify the performance of the components of ReBo. We remove both bounding box combination scores (BBCS) and recurrent multimodal encoder (RME) to reformulate ReBo into the model with concatenation

Model	Question		Answer		Distractor	
	BLEU-1	CIDEr	BLEU-1	CIDEr	BLEU-1	CIDEr
Mistral	31.55±0	35.90±0	8.63±0	35.34±0	8.86±0	10.34±0
ChatGPT	32.31±0	19.01±0.2	9.02±0	7.8±0.07	9.60±0.04	8.02±0
Llama-2	36.63±2.79	41.64±53.97	7.12±0.36	24.71±12.71	7.61±0.41	7.38±0.28
Qwen1.5	37.97±0.01	45.1±0.09	10.33±0.04	39.53±0.92	9.65±0.01	9.32±0.15
Llama-3	37.19±0	51.50±1	17.41±0.04	59.27±2.23	11.47±0.02	13.58±0.08
LLaVA-NeXT	31.76	25.61	6.71	15.63	4.79	4.52
LLaVA-1.5	46.61	73.64	13.8	42.43	9.67	9.69
CogVLM	48.46	77.46	2.88	2.47	4.58	6.06
BLIP†	49.45±2.07	61.40±80.89	8.57±38.23	10.55±20.05	2.71±3.09	0.57±0.10
VisualBert†	46.68±0.54	70.96±23.55	15.05±0.62	34.38±18.44	4.63±0.50	2.30±0.52
BLIP2†	46.64±0.61	101.43±44.32	24.38±0.90	78.52±20.73	11.30±0.37	15.69±3.84
Qwen-VL†	50.69±0.56	105.96±18.36	22.23±0.61	67.67±15.65	12.88±0.13	16.35±1.69
VQADG†	**51.33±0.88**	119.55±97.17	27.26±1.12	84.06±31.54	14.58±0.93	20.07±3.83
ReBo	50.11±1.25	**128.25±37.75**	**30.63±1.61**	**95.44±24.89**	**16.16±2.44**	**22.55±10.10**

Table 2: Separate comparisons of question, answer, and distractor on the Visual7W dataset.

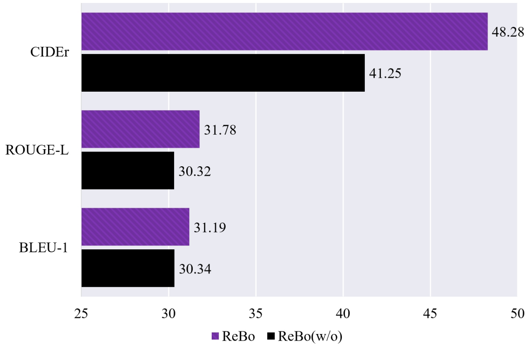


Figure 4: The ablation results for ReBo. ReBo (w/o) indicates ReBo without bounding box combination scores and the recurrent multimodal encoder.

Model	Q	A	D	I	U
BLIP2	3.68	2.79	2.87	3.15	3.26
VQADG	3.73	3.45	3.21	3.32	3.57
Qwen-VL	3.88	3.49	2.98	3.34	3.59
ReBo	**4.07**	**3.72**	**3.26**	**3.70**	**4.02**

Table 3: Human evaluation of the generated QADs. Q, A, and D denote the total quality score of questions, answers, and distractors, I denotes the intersection between different QADs, and U denotes the union score for all QADs associated with a given image.

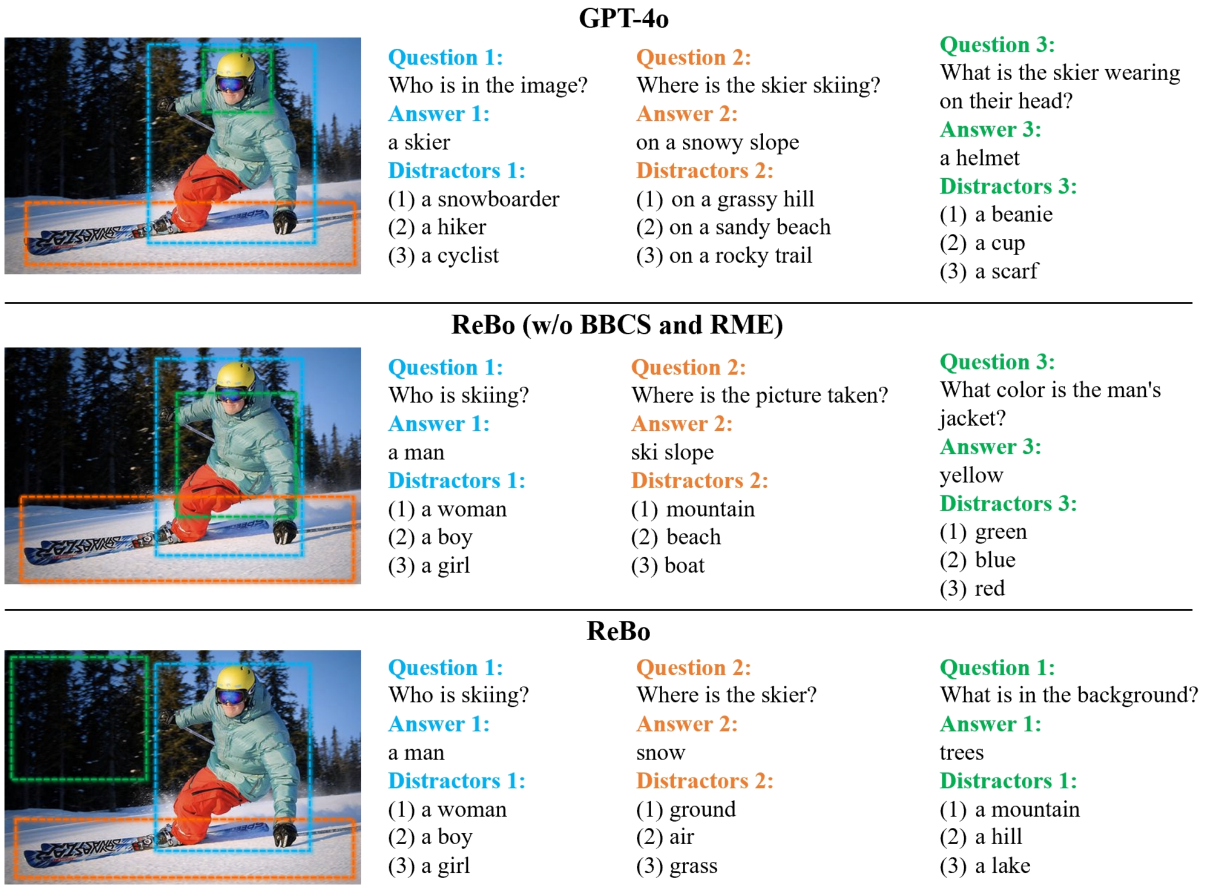
generation types. Experimental results in Figure 4 demonstrate that both modules contribute to achieving good performance for ReBo.

Excluding BBCS and RME seems not to significantly affect the BLEU-1 and ROUGE-L performance of ReBo, yet they help generate informative QADs that focus on diverse regions. More details can be found in the case studies in Figure 5 and Appendix C.

4.3.4 Human Evaluations

To further assess the effectiveness of ReBo, we conducted a human evaluation of 300 images. We generate three QADs separately using BLIP2, VQADG, Qwen, and ReBo for each image. The total human evaluation data comprises 300 images and 3600 QADs.

We recruit six annotators to rate them from 1 to 5 points on five qualitative aspects: (1) *Quality* The overall quality of the generated QADs includes question relevance, answer accuracy, and the confusion level of distractors. (2) *Intersection* The intersection score represents whether the semantic contents of generated QADs for a given image are dissimilar. (3) *Union* The union score represents whether the generated QADs can summarize the overall content of the image. A higher score implies that the model performs better. Table 3 displays the results of human evaluation, revealing that ReBo achieves the highest scores across all five metrics. Experimental results demonstrate that our recurrent multimodal encoder and bounding box scores are not only capable of generating high-quality QADs, but also facilitate the generalization of QADs with small intersections among each other and cover more information from the image.



GPT-4o

Question 1: Who is in the image?
Answer 1: a skier
Distractors 1: (1) a snowboarder (2) a hiker (3) a cyclist

Question 2: Where is the skier skiing?
Answer 2: on a snowy slope
Distractors 2: (1) on a grassy hill (2) on a sandy beach (3) on a rocky trail

Question 3: What is the skier wearing on their head?
Answer 3: a helmet
Distractors 3: (1) a beanie (2) a cup (3) a scarf

ReBo (w/o BBCS and RME)

Question 1: Who is skiing?
Answer 1: a man
Distractors 1: (1) a woman (2) a boy (3) a girl

Question 2: Where is the picture taken?
Answer 2: ski slope
Distractors 2: (1) mountain (2) beach (3) boat

Question 3: What color is the man's jacket?
Answer 3: yellow
Distractors 3: (1) green (2) blue (3) red

ReBo

Question 1: Who is skiing?
Answer 1: a man
Distractors 1: (1) a woman (2) a boy (3) a girl

Question 2: Where is the skier?
Answer 2: snow
Distractors 2: (1) ground (2) air (3) grass

Question 1: What is in the background?
Answer 1: trees
Distractors 1: (1) a mountain (2) a hill (3) a lake

Figure 5: Case studies. The focus regions of the QADs generated by different models are portrayed. Our model ReBo can generate QADs focusing on diverse image regions.

4.3.5 Case Studies

We present case studies to demonstrate the QADs generated by GPT-4o, ReBo without BBCS and RME, and ReBo in Figure 5. For GPT-4o, we design the prompt and give examples to generate questions, answers, and distractors. The prompt can be found in Appendix A. We present three groups of QADs generated by each method and highlight their focus regions.

It shows from the figure that GPT-4o and ReBo without BBCS and RME can generate complete QADs, yet they may produce some inappropriate or incorrect answers and/or distractors. For example, GPT-4o generates a distractor *"a snowboarder"*, which is almost indistinguishable from the correct answer *"a skier"*. ReBo without BBCS and RME generates an incorrect answer *"yellow"* for the question *"What color is the man's jacket?"*. Our ReBo can generate meaningful questions, correct answers, and misleading distractors. Furthermore, the QADs generated by ReBo focus on a broad region of the image, comprising the regions of people, background trees, and ground snow. In contrast, GPT-4o and ReBo without BBCS and RME disregard the semantic richness of the generated QADs and are likely to be concerned with overlapped regions. More case studies are presented in Appendix C.

5 Conclusion

In this paper, we propose a novel framework with a recurrent multimodal encoder and bounding box scores to generate a series of QADs. The multimodal encoder recurrently generates different QADs for an image, utilizing the previous QADs as part of the input to generate current QADs. The bounding box scores consider the intersection over union and the union over total image, which can facilitate the generation of QADs that attend to as large and diverse areas as possible for one image. We conduct experiments on the benchmark to demonstrate a significant advantage of our model in the evaluation metrics. Additionally, our generated QADs, as supplementary data to the original dataset, exhibit the capability to promote the performance of existing VQA models.

6 Limitations

Our focus in this study is devoted on generating diverse QADs jointly. This task is challenging as it involves learning interactions between QADs, as well as encoding, generating, and evaluating QADs. We notice that there is still large room for progress. For example, how to tailor our model specific to different types of question, answer, and distractors and how to evaluate the generated QADs in a human-like manner remain untouched and will be tackled in our future study.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In Proceedings of ICCV, pages 2425–2433.

Jinze Bai, Shuai Bai, Yunfei Chu, and et al. 2023a. Qwen technical report. arXiv preprint arXiv:2309.16609.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of ACLW, pages 65–72.

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for vqa are image captions. arXiv preprint arXiv:2205.01883.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. JMLR, 25(70):1–53.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. In Proceedings of NeurIPS, pages 49250–49267.

Wenjian Ding, Yao Zhang, Jun Wang, Adam Jatowt, and Zhenglu Yang. 2024. Can we learn question, answer, and distractors all from an image? a new task for multiple-choice visual question answering. In Proceedings of LREC-COLING, pages 2852–2863.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In Proceedings of CVPR, pages 19358–19369.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of CVPR, pages 6904–6913.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of CVPR, pages 4565–4574.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In Proceedings of CVPR, pages 4999–5007.

Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information maximizing visual question generation. In Proceedings of CVPR, pages 2008–2018.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV, 123:32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of ICML, pages 19730–19742.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of ICML, pages 12888–12900.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. Visual question generation as dual task of visual question answering. In Proceedings of CVPR, pages 6116–6124.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Proceedings of ACL, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In Proceedings of ECCV, pages 740–755.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In Proceedings of CVPR, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. In Proceedings of NeurIPS, pages 34892–34916.

Jiaying Lu, Xin Ye, Yi Ren, and Yezhou Yang. 2022a. Good, better, best: Textual distractors generation for multiple-choice visual question answering via reinforcement learning. In Proceedings of CVPR, pages 4921–4930.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022b. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Proceedings of NeurIPS, pages 2507–2521.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of CVPR, pages 3195–3204.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In Proceedings of NeuIPS, pages 27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of ACL, pages 311–318.

Colin Raffel, Noam 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. JMLR, 21(140):1–67.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In Proceedings of ECCV, pages 146–162.

Kai Shen, Lingfei Wu, Siliang Tang, Fangli Xu, Zhu Zhang, Yu Qiang, and Yueting Zhuang. 2020. Ask question with double hints: Visual question generation with answer-awareness and region-reference.

Hung-Ting Su, Chen-Hsi Chang, Po-Wei Shen, Yu-Siang Wang, Ya-Liang Chang, Yu-Cheng Chang, Pu-Jen Cheng, and Winston H Hsu. 2021. End-to-end video question-answer generation with generator-pretester network. T-CSVT, 31(11):4497–4507.

Meta LLaMA Team. 2024a. Introducing meta llama 3: The most capable openly available llm to date.

Qwen Team. 2024b. Introducing qwen1.5.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In Proceedings of CVPR, pages 4566–4575.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079.

Peixi Xiong and Ying Wu. 2020. Ta-student vqa: Multi-agents training by self-questioning. In Proceedings of CVPR, pages 10065–10075.

Sen Yang, Qingyu Zhou, Dawei Feng, Yang Liu, Chao Li, Yunbo Cao, and Dongsheng Li. 2021. Diversity and consistency: Exploring visual question-answer pair generation. In Proceedings of Findings of EMNLP, pages 1053–1066.

Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2016. Automatic generation of grounded visual questions. arXiv preprint arXiv:1612.06530.

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

Yuke Zhu, Oliver Groth, Michael Bernstein, and Fei-Fei Li. 2016. Visual7w: Grounded question answering in images. In Proceedings of CVPR, pages 4995–5004.

A Prompts for Large Languege Models

Table 4 presents the prompts used by ChatGPT, Mistral, Qwen1.5, Llama-2, and Llama-3 for QADs generation. For LLaVA-1.5, LLaVA-NeXT, CogVLM, and GPT-4o we directly use image instead of image caption.

QADs Generation Prompt Input

Image caption: The image depicts a man sitting at a desk with a laptop computer and a monitor in front of him. There is also a cup of coffee on the desk, indicating that the man is working in an office environment. There are several other items scattered around the workspace, including a pair of headphones, a pen, and a bottle of water. The man is likely working in an office environment, as he has a laptop computer and a monitor in front of him. There is also a cup of coffee on the desk, indicating that the man is working in an office environment.

Refer to the following example and based on the above image caption, generate three questions starting with ‘what’, ‘who’, and ‘where’, and generate the answer and three distractors for each question, the distractors should be seperated with numbers like (1) (2) (3).

Example:
Question 1: What is on the bookshelf?
Answer 1: books
Distractor 1: (1) small plant (2) picture frame (3) book ends

Question 2: Who is wearing a watch?
Answer 2: the lady
Distractors 2: (1) the umpire (2) the man (3) the girl

Question 3: Where is the image taken?
Answer 3: near to house
Distractors 3: (1) in the park (2) on the beach (3) on the highway

Table 4: Prompts used for QADs generation.

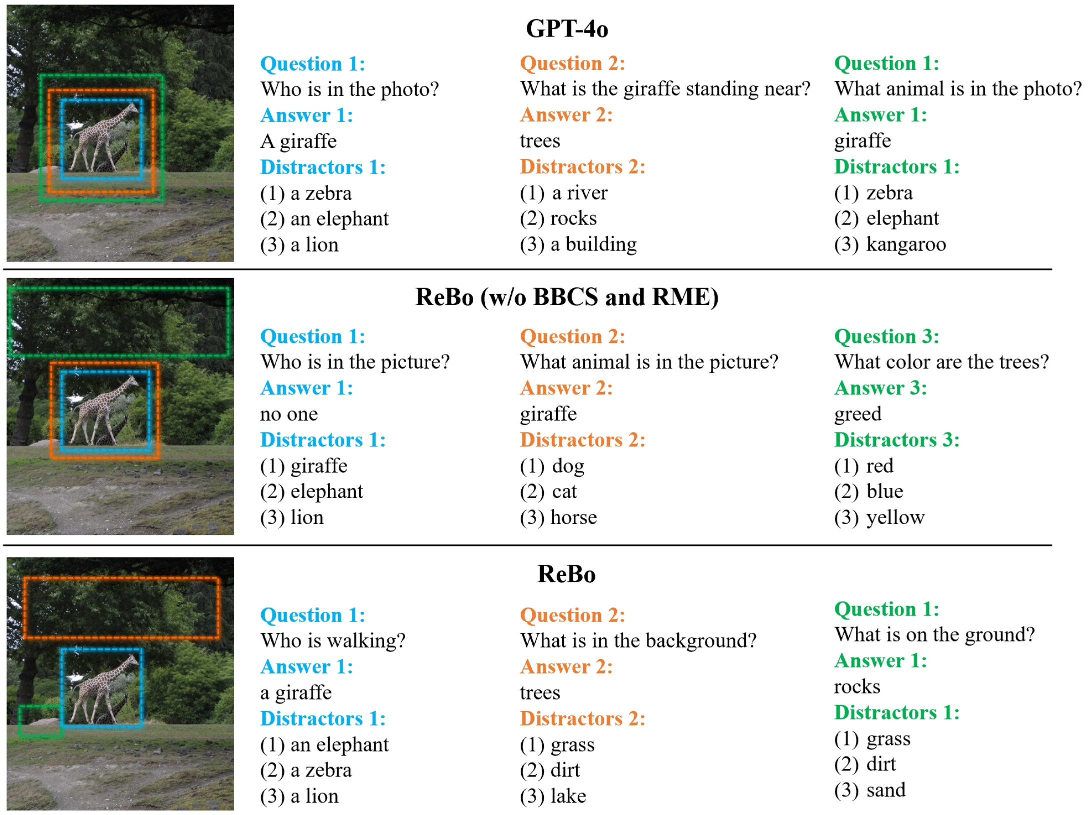
B Implementation Details

We adapt our model based on the modular architecture of InstructBLIP (Dai et al., 2024). We retain the image encoder and the LLM decoder while adapting the Q-Former into a recurrent multimodal encoder. We implement our model with the image encoder ViT-g/14 (Fang et al., 2023) and the large language model FlanT5-XL (Chung et al., 2024), which is an instruction-tuned model based on the encoder-decoder Transformer T5 (Raffel et al., 2020). We refer (Ding et al., 2024) to employ an extra contrastive learning loss function to normalize the embeddings of prediction results and ground truth. For the hyper-parameters, we set the maximum text length to 60 and the minimum text length to 20 for the recurrent generation type and 60 to 180 for the concatenation generation type. The image size in all models is resized to 224. We use the batch size 8 and 32 for training and testing and fine-tune the datasets for 10 epochs. Other parameters are set according to the original articles. For Large Language Models, we calculated the mean and variance of the results over three runs. For Large Vision-Language Models, we report only one result due to consistent outputs. For our model and all other baselines, we divided the training and testing data into ten splits and calculated the mean and variance of the results over ten runs. We use the HuggingFace[1] transformers library implementation for LLMs and LVLMs. Our experiments are run on 1 NVIDIA A40 48G GPU.

C More Case Studies

Figure 6 and Figure 7 present two additional case studies. In both two cases, ReBo generates QADs that focus on diverse regions, whereas GPT-4o and ReBo without BBCS and RME generate QADs with semantic overlap.

[1] https://huggingface.co/



GPT-4o

Question 1:
Who is in the photo?
Answer 1:
A giraffe
Distractors 1:
(1) a zebra
(2) an elephant
(3) a lion

Question 2:
What is the giraffe standing near?
Answer 2:
trees
Distractors 2:
(1) a river
(2) rocks
(3) a building

Question 1:
What animal is in the photo?
Answer 1:
giraffe
Distractors 1:
(1) zebra
(2) elephant
(3) kangaroo

ReBo (w/o BBCS and RME)

Question 1:
Who is in the picture?
Answer 1:
no one
Distractors 1:
(1) giraffe
(2) elephant
(3) lion

Question 2:
What animal is in the picture?
Answer 2:
giraffe
Distractors 2:
(1) dog
(2) cat
(3) horse

Question 3:
What color are the trees?
Answer 3:
greed
Distractors 3:
(1) red
(2) blue
(3) yellow

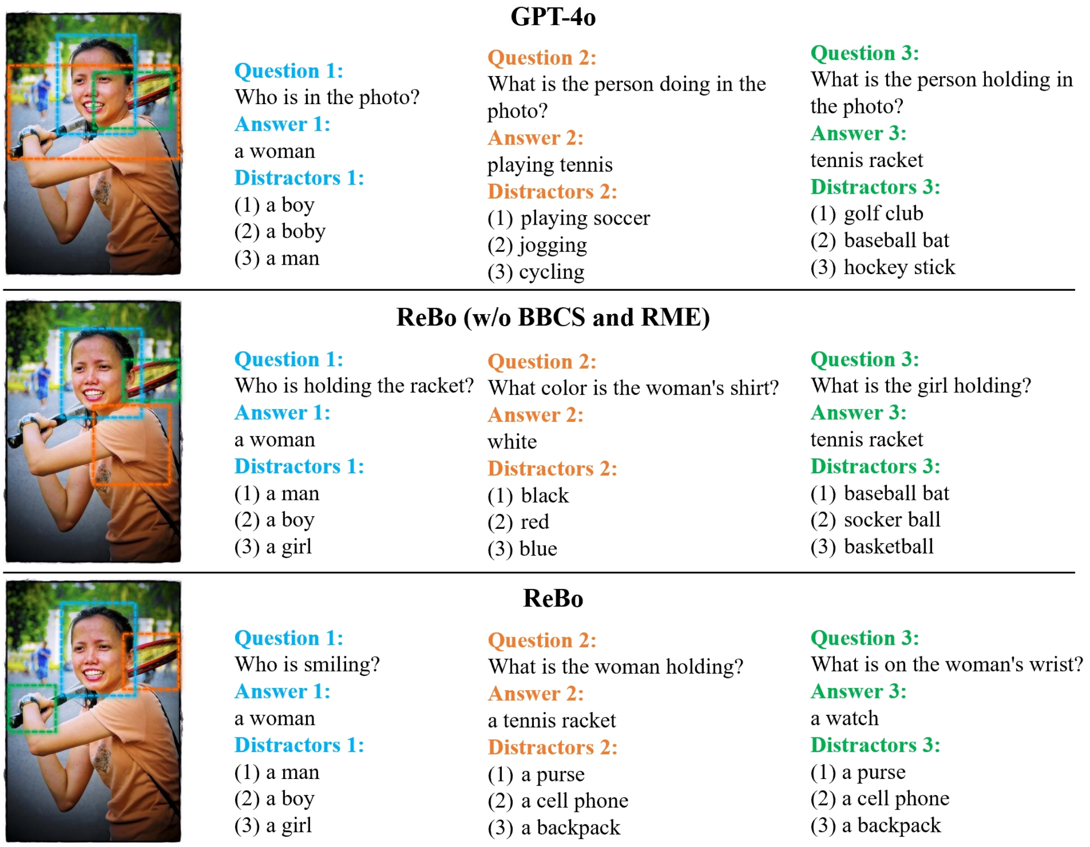
ReBo

Question 1:
Who is walking?
Answer 1:
a giraffe
Distractors 1:
(1) an elephant
(2) a zebra
(3) a lion

Question 2:
What is in the background?
Answer 2:
trees
Distractors 2:
(1) grass
(2) dirt
(3) lake

Question 1:
What is on the ground?
Answer 1:
rocks
Distractors 1:
(1) grass
(2) dirt
(3) sand

Figure 6: Case study 1. Only ReBo generates QADs focusing on different visual regions.



GPT-4o

Question 1:
Who is in the photo?
Answer 1:
a woman
Distractors 1:
(1) a boy
(2) a boby
(3) a man

Question 2:
What is the person doing in the photo?
Answer 2:
playing tennis
Distractors 2:
(1) playing soccer
(2) jogging
(3) cycling

Question 3:
What is the person holding in the photo?
Answer 3:
tennis racket
Distractors 3:
(1) golf club
(2) baseball bat
(3) hockey stick

ReBo (w/o BBCS and RME)

Question 1:
Who is holding the racket?
Answer 1:
a woman
Distractors 1:
(1) a man
(2) a boy
(3) a girl

Question 2:
What color is the woman's shirt?
Answer 2:
white
Distractors 2:
(1) black
(2) red
(3) blue

Question 3:
What is the girl holding?
Answer 3:
tennis racket
Distractors 3:
(1) baseball bat
(2) socker ball
(3) basketball

ReBo

Question 1:
Who is smiling?
Answer 1:
a woman
Distractors 1:
(1) a man
(2) a boy
(3) a girl

Question 2:
What is the woman holding?
Answer 2:
a tennis racket
Distractors 2:
(1) a purse
(2) a cell phone
(3) a backpack

Question 3:
What is on the woman's wrist?
Answer 3:
a watch
Distractors 3:
(1) a purse
(2) a cell phone
(3) a backpack

Figure 7: Case study 2. Although Rebo without BBCS and RME can still generate QADs focusing on various image regions, it unfortunately produces the incorrect answer *"white"*.