Can We Learn Question, Answer, and Distractors All from An Image? A New Task for Multiple-Choice Visual Question Answering

Wenjian Ding[1], **Yao Zhang**[2], **Jun Wang**[3], **Adam Jatowt**[4], **Zhenglu Yang**[1*]
[1]TKLNDST, CS, Nankai University, China
[2]School of Statistics and Data Science, LPMC, KLMDASR & LEBPS, Nankai University
[3]Shandong Key Laboratory of Language Resource Development and Application,
College of Mathematics and Statistics Science, Ludong University
[4]University of Innsbruck, Austria
wjding@mail.nankai.edu.cn, yaozhang@nankai.edu.cn, junwang@mail.nankai.edu.cn,
adam.jatowt@uibk.ac.at, yangzl@nankai.edu.cn

Abstract

Multiple-choice visual question answering (MC VQA) requires an answer picked from a list of distractors, based on a question and an image. This research has attracted wide interests from the fields of visual question answering, visual question generation, and visual distractor generation. However, these fields still stay in their own territories, and how to jointly generate meaningful questions, correct answers, and challenging distractors remains unexplored. In this paper, we introduce a novel task, Visual Question-Answer-Distractors Generation (VQADG), which can bridge this research gap as well as take as a cornerstone to promote existing VQA models. Specific to the VQADG task, we present a novel framework consisting of a vision-and-language model to encode the given image and generate QADs jointly, and contrastive learning to ensure the consistency of the generated question, answer, and distractors. Empirical evaluations on the benchmark dataset validate the performance of our model in the VQADG task.

Keywords: Multiple-Choice Visual Question Answering, Distractors, Text Generation

1. Introduction

Multiple-Choice Visual Question Answering (MC VQA) (Kembhavi et al., 2017; Zellers et al., 2019; Lu et al., 2022b) has become one of the current hotspots in natural language processing and computer vision research. Most of existing MC VQA studies focus on the stand-alone generation of question, answer, or distractors. In reality however, an all-in-one generation of QADs may provide a feasible solution of alleviating learning bias (Niu et al., 2021) and tackling data scarcity. As a byproduct, the generated high-quality MC VQA data can further assist in improving existing VQA models, or serve as a crucial component of pre-training data for large language models (LLM). In this paper, we investigate how to jointly generate meaningful questions, correct answers, and challenging distractors in a unified framework. Taking the example in Figure 1 (a) for explanation, our task is to generate the image-related question "What is the color of the used napkin", its answer "Green", and several distractors, e.g., "Red", in one body.

Existing studies typically concentrate on a portion of this task: Visual Question Generation (VQG) (Li et al., 2018; Krishna et al., 2019), Visual Question Answering (VQA) (Antol et al., 2015; Zellers et al., 2019; Lu et al., 2022b), or Visual Distractor Generation (VDG) (Lu et al., 2022a), as shown in Figure 1 (b). VQG aims to generate a question by comprehending the image content, which should be image-related, meaningful, and grammatically well-formed. VQA assumes all of the information can be induced from the given image and it provides a correct answer corresponding to the question. In contrast to the above two tasks, VDG has rarely been studied in MC VQA, which requires generating challenging and high-quality distractors by investigating the image, question, and answer. Existing distractors (Zhu et al., 2016) are rather simple to evaluate a model's actual cross-modality discriminative ability. For example, in Figure 1(a), it is easy to distinguish the correct answer "Green" from the distractors "Blue" and "Orange" created by (Zhu et al., 2016), while is not trivial to eliminate our generated distractor "Red". Our distractor is more deceptive since it is both content-related with the image and semantics-related with the question and answer.

The difficulty of jointly researching QADs lies in the infeasibility of simply combining VQA, VQG, and VQD. It is that three tasks are intrinsically correlated and thus, generating each component of QADs should consider image context while remaining the other two components as conditions. To address this issue, we propose **V**isual **Q**uestion-**A**nswer-**D**istractors **G**eneration (VQADG), which is the first attempt to generate QADs in a unified way. Technically, we propose a vision-and-language model with an encoder-decoder architecture. The visual image and question content

[*] Zhenglu Yang

LREC-COLING 2024, pages 2852–2863
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

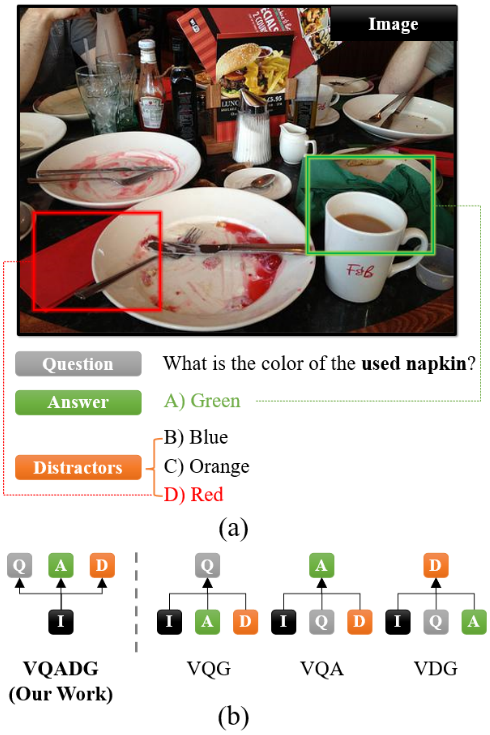


Figure 1: (a) An example of MC VQA, which consists of an image, a question, an answer, and multiple distractors. The used "green" napkin indicates the correct answer, while the unused "red" napkin is a challenging distractor generated by our work. (b) In contrast to VQG, VQA, and VDG which focus on parts of QADs, our VQADG generates QADs jointly.

are multimodally encoded and transferred to an autoregressive text decoder to generate QADs. The high-quality QADs generated by our framework are not only more deceptive than the manually created ones but also can serve as the augmented data to enhance existing VQA models. Moreover, we introduce contrastive learning to keep the generated QADs consistent. A comprehensive experimental evaluation on Visual7W (Zhu et al., 2016) validates the holistic generation capability of our model.

The main contributions of this work are listed as follows:

- We introduce visual question-answer-distractors generation which is the first attempt to jointly generate meaningful questions, correct answers, and challenging distractors from images.
- We propose a novel vision-and-language model with a multimodal mixture of encoder-decoder architecture and integrate contrastive learning to enhance the generation quality of QADs.
- Extensive experimental results on the benchmark dataset reveal that generated QADs can be used to enhance the performance of existing VQA models.

2. Related Works

2.1. Visual Question-Answer-Distractors Generation

Many recent efforts have been invested in generating QADs.

VQG targets to generate pertinent questions by considering visual and textual clues, such as ground truth answers, question types, and answer categories. Fan et al. (2018) designed a strategy to perform the learning of the distribution of question types for each image. Krishna et al. (2019) proposed a model that maximizes the mutual information among the image, the expected answer, and the generated question.

VQA takes an image and a meaningful question as input and produces a correct answer as output. VQA can be divided into open-ended VQA (Antol et al., 2015; Masry et al., 2022) and MC VQA (Zhu et al., 2016; Kembhavi et al., 2017; Lu et al., 2022b) based on the answer form. Recently, some studies focus on combining question and answer. Li et al. (2018) introduced question generation as a dual task of question answering to improve the VQA performance. Yang et al. (2021) integrated variational inference to generate various question-answer pairs.

VDG targets to generate challenging and high-quality distractors when given the context image, meaningful question, and correct answer. Lu et al. (2022a) proposed a reinforcement learning strategy to generate distractors for visual images.

In a word, QADs are separately studied by the above studies, research on generating QADs in a unified architecture has been thus far under-explored.

2.2. Vision-and-Language Pretraining

Transformers (Vaswani et al., 2017; Devlin et al., 2018; Raffel et al., 2020; Brown et al., 2020) have shown exceptional performance in the natural language processing domain. Following this success, large-scale image-text pairs have been used to improve the multimodal representation of vision-and-language models (Cho et al., 2021; Wang et al., 2022; Li et al., 2022; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Dai et al., 2023). Most of previous vision-and-language pretraining models (Su et al., 2019; Chen et al., 2020b) typically focused on discriminative tasks, while recent works have begun to turn to generative downstream tasks (Cho et al., 2021; Li et al., 2021, 2022). In this paper, we resort to VL-T5 (Cho et al., 2021), which consists of a multimodal encoder to fuse image features and textual question types, and an autoregressive decoder that generates QADs in generative tasks.

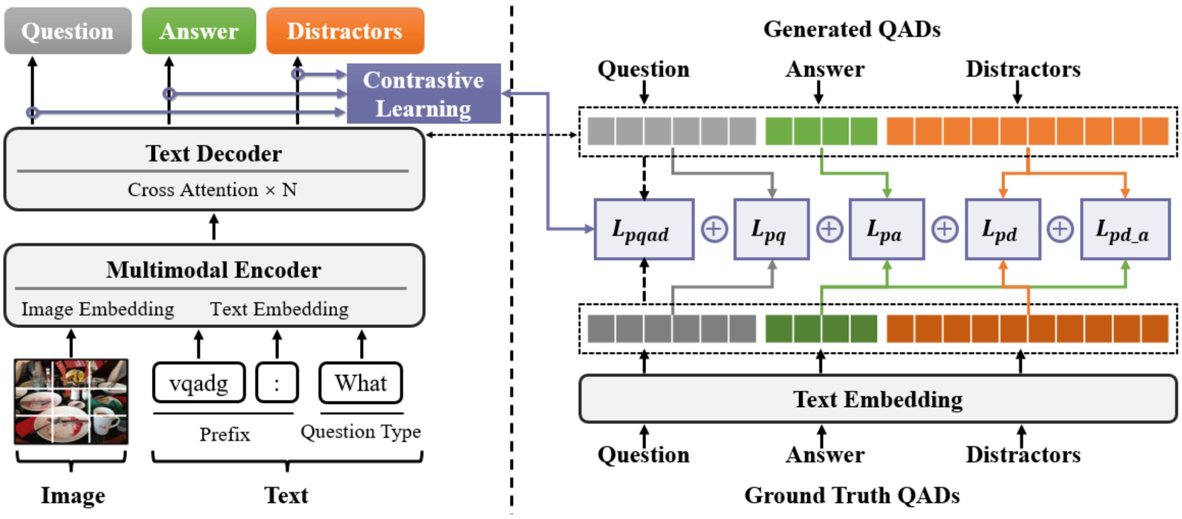


Figure 2: The model architecture of our VQADG model. The multimodal encoder takes the concatenation of image embedding and text embedding as input and outputs their contextualized joint representation. This cross-modal representation is used to guide the text decoder generates QADs jointly. We incorporate contrastive learning loss and language modeling loss to train our final VQADG model.

3. The VQADG Model

We propose VQADG, a unified vision-and-language framework with a multimodal mixture of an encoder-decoder architecture to generate QADs. Our VQADG model consists of a multimodal encoder, a text decoder, and different contrastive learning objectives. The multimodal encoder and text decoder use the pre-trained VL-T5 model (Cho et al., 2021) as the backbone model. We also design different contrastive learning loss functions to improve consistency of QADs. The overall architecture of VQADG is shown in Figure 2. This section initially presents the preliminary of the new task, and then introduces our model architecture.

3.1. Preliminary

We define the task VQADG, given:

- An informative image I, which can generate various QADs according to different textual prefixes.
- A prefix T in textual modality, which represents the instruction to generate specific QADs. $T = (t_1, t_2, ..., t_M)$ is a sequence with M tokens.

The goal of VQADG is to generate QADs according to the image I and a textual prefix including question type T in a unified framework. QADs comprise a meaningful question $Q = (q_1, q_2, ..., q_N)$, a correct answer $A = (a_1, a_2, ..., a_P)$, and challenging distractors $D = (d_1, d_2, ..., d_S)$. Q, A, and D are all textual sequences consisting of words or tokens limited by length N, P, and S. Our final objective is to train the VQADG model to obtain the best model parameter θ^*, which can maximize the likelihood of the QADs in an autoregressive manner as follows:

$$\theta^* = \arg\max P(Q, A, D|I, T; \theta). \qquad (1)$$

3.2. Multimodal Encoder

Our multimodal encoder is extended from the text encoder of T5 (Raffel et al., 2020), which consists of self-attention layers and fully-connected layers with residual connections. It takes both image and text as input and outputs a contextualized joint representation, which guides the generation of QADs in the decoding stage.

Image Embedding We employ a pre-trained object detector for image embedding extraction to represent the image I, that is, Faster R-CNN (Ren et al., 2015) trained on Visual Genome (Krishna et al., 2017). Specifically, each input image I is sliced into 36 patches. The object features and bounding box coordinates of each patch are extracted from the pre-trained image extractor. Then, the object features and bounding box coordinates are encoded with a linear layer. Following linear projection, object features are embedded as the visual feature embedding e_f, and bounding box coordinates are embedded as the visual position embedding e_p. The final image embedding is the sum of the visual feature embedding and visual position embedding, which is denoted as $e_I = e_f + e_p$.

Text Embedding The text input $T = (Prefix : Type)$ includes a prefix and the original input text, i.e., question type. The prefix $Prefix = vqadg$ is added to support the model for generating QADs[1].

[1] The prefix setting allows our model to be applied to

Introducing the prefix setting is inspired by the VL-T5 model (Cho et al., 2021)[2]. The question type refers to the type of question expected to be generated by the model, including *what, where, when, why, who,* and *how*. This augmented input text is encoded as the text embedding e_T.

The multimodal encoder provides the contextualized joint representation after receiving the concatenation of image embedding e_I and text embedding e_T as input. This joint embedding serves as the whole multimodal representation which can guide the generation of QADs in the text decoder via the cross-attention layer.

3.3. Text Decoder

Different from the original T5, which only requires a single modality input, our decoder pays attention to the textual and joint content from the multimodal encoder. A cross-attention layer takes the output of the self-attention layer and multimodal encoder as input to model vision-and-language interactions. Thus, the text decoder is a stack of transformer blocks, which comprises a self-attention layer, a cross-attention layer, and a fully-connected layer with residual connections. The final text decoder iteratively attends to previously generated tokens and the encoder outputs (via cross-attention), and then predicts the probability of the next text tokens. We jointly optimize two objectives during training, one of which is the generation-based objective that optimizes the language modeling loss and another is the understanding-based objective that optimizes the contrastive learning loss. The language modeling loss is formulated as a cross entropy loss which trains the model to minimize the negative log-likelihood of label text tokens $y = (Q, A, D)$ when given the input prefix type T and image I:

$$L_{LM} = -\sum_{j=1}^{|y|} \log P(y_j | y < j, I, T). \tag{2}$$

3.4. Contrastive Learning

We incorporate contrastive learning loss with language modeling loss to improve the consistency of generated QADs. Contrastive learning (Radford et al., 2021) can learn effective embeddings accordingly to keep the positive pairs stay closely and the negative pairs stay away in the embedding space. We leverage the embeddings of the predicted result P and ground truth G as positive pairs; the negative pairs are generated by replacing P or G with the samples selected from each mini-batch randomly. In terms of the similarity of P and G, we calculate the softmax-normalization of P to G and G to P as follows:

$$p_m^{g2p}(G) = \frac{\exp(s(G, P_m)/\tau)}{\sum_{m=1}^{M} \exp(s(G, P_m)/\tau)}, \tag{3}$$

$$p_m^{p2g}(P) = \frac{\exp(s(P, G_m)/\tau)}{\sum_{m=1}^{M} \exp(s(P, G_m)/\tau)}, \tag{4}$$

where τ is a temperature hyperparameter and $s(P, G)$ is the cosine similarity $\frac{P^T G}{|P||G|}$. Let $y^{p2g}(P)$ and $y^{g2p}(G)$ deno te the ground truth one-hot label, where $y = 1$ if G and P are positive pairs and $y = 0$ if they are negative pairs. The contrastive learning loss is defined as the cross-entropy H between p and y as:

$$\begin{aligned} L_{pg} = \frac{1}{2}[H(y^{p2g}(P), p^{p2g}(P)) + \\ H(y^{g2p}(G), p^{g2p}(G))]. \end{aligned} \tag{5}$$

We design different contrastive learning loss functions for only question, only answer, only distractors, answer and distractors, and joint QADs. For the situation with only question, we choose the question part of P as the predicted result, and we select G as ground truth. P and G of one sample are the positive pairs in contrastive learning, thus the question part of contrastive learning loss is L_{pq}. In selecting the question part from the overall embedding of P, we calculate the ratio of the question length to the length of total QADs and then use this ratio to segment the embedding. In the same manner, the contrastive learning of the answer is L_{pa}. For distractors generation, in addition to L_{pd}, we design another contrastive learning loss to improve the similarity between distractors and the correct answer to generate challenging distractors:

$$\begin{aligned} L_{pd\_a} = \frac{1}{2}[H(y^{a2d}(A), \alpha p^{a2d}(A)) + \\ H(y^{d2a}(D), \alpha p^{d2a}(A))], \end{aligned} \tag{6}$$

where α is a hyperparameter that measures the similarity between distractors and the answer. For example, $\alpha = 2$ denotes the similarity is 0.5. We also use L_{pqad} to represent the total contrastive learning loss between QADs with the overall predicted result. Finally, we accumulate the sum of the contrastive learning losses to construct the training loss L of our framework as follows:

$$L_{CL} = L_{pqad} + L_{pq} + L_{pa} + L_{pd} + L_{pd\_a}, \tag{7}$$

$$L = \beta L_{LM} + (1 - \beta) L_{CL}, \tag{8}$$

other tasks in MC VQA as well, such as $Prefix = vqg$ in VQG.

[2] (Cho et al., 2021) finds that a single prefix can successfully handle multiple VQA-related tasks without dataset-specific prefixes. Similar results were observed in text QA (Khashabi et al., 2020).

where β is a hyperparameter that measures the ratio of the language modeling loss to the contrastive learning loss. In terms of the ground truth label in the contrastive learning loss, we test image, text, and both of them in experiments; finally, using text as the ground truth label achieves the best performance.

4. Experiments

4.1. Datasets

We evaluate our model using the publicly available MC VQA dataset, Visual7W (Zhu et al., 2016) and VQAv2 dataset (Goyal et al., 2017). Visual7W is collected based on COCO (Lin et al., 2014) and consists of 47,300 images and 327,939 MC QA pairs. We extracted MC VQA data from the VQAv2 dataset, which includes 87,544 images and 187,688 MC QA pairs.

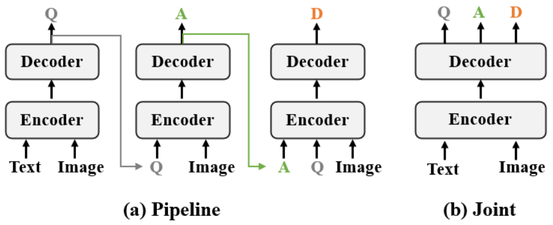
4.2. Baselines

We compare our model with the following state-of-the-art methods:

- **IQ** (Krishna et al., 2019) approaches the question generation task by maximizing the mutual information between the generated question, image, and answer or its category.
- **VisualBert**† (Li et al., 2020) is a pre-trained vision-and-language encoder for joint vision and language representation. Corresponding to the output of the VisualBert encoder, we incorporate a Bert decoder to generate QADs.
- **BLIP**† (Li et al., 2022) employs noisy image-text data to train a model designed for vision-language comprehension and generation tasks. We retrained BLIP to ensure it can generate QADs according to the images.
- **InstructBLIP** (Dai et al., 2023) is a instruction tuning framework towards generalized vision-language models. We used InstructBLIP to assess the quality of QADs.

In view of the limitations of the aforementioned state-of-the-art methods in generating distractors, we extend four variants from our proposed VQADG model to conduct a more comprehensive evaluation:

- **Pipeline**: QADs are generated in a pipeline manner, as shown in Figure 3 (a).
- **Pipeline+CL**: Contrastive learning loss functions for QADs are added to the question model, answer model, and distractors model, respectively.
- **Joint**: QADs are generated in a joint manner, as shown in Figure 3 (b). This variant only uses language modeling loss during the training stage.



(a) Pipeline

(b) Joint

Figure 3: Two types of variants of our VQADG model.

- **Joint+CL**: The complete contrastive learning loss is added to the Joint model. This variant is equivalent to our VQADG model.

4.3. Evaluation 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. Moreover, we calculate the consistency score (Yang et al., 2021) to evaluate the consistency degree between the generated question, answer, distractors, and image. The consistency score is calculated as:

$$S_t = Sigmoid \circ f(I, t), \tag{9}$$

where $t \in \{Q, A, D_1, D_2, ..., D_i\}$ and $S_t \in [0, 1]$. A high score indicates that the image has high consistency with the generated QADs.

4.4. Implementation Details

The VQADG model[3] is built on the pre-trained VL-T5 and fine-tuned on the Visual7W dataset. We set the maximum text length to 80, and α is set to 2 after a pilot study. In the case of BLIP, we adjust the image size to 224 and eliminate label smoothing for optimal performance. VisualBert and BLIP are all retrained on Visual7W to generate QADs. We train all the models with the same prefix and a batch size of 32.

4.5. Results and Analysis

4.5.1. Automatic Evaluation

We evaluate the question generation, question-answer pair generation, and QADs generation.

The performance of our models and baseline models on Visual7W are shown in Table 1. In terms of the question evaluation, the Pipeline model can

[3] The source code will be released after the paper is published.

Model		Content			BLEU-4	ROUGE-L	METEOR	CIDEr
		Q	A	D				
IQ (Krishna et al., 2019)		✓	–	–	15.40	47.12	21.40	122.18
VisualBert† (Li et al., 2020)		✓	–	–	18.72	50.78	25.61	120.13
BLIP† (Li et al., 2022)		✓	–	–	22.21	53.90	27.62	152.94
VQADG	Pipeline	✓	–	–	23.16	54.47	27.27	159.33
	Pipeline+CL	✓	–	–	23.59	54.82	27.30	161.10
	Joint	✓	–	–	23.32	54.56	27.37	157.40
	Joint+CL	✓	–	–	**24.35**	**55.35**	**27.64**	**165.71**
VisualBert† (Li et al., 2020)		✓	✓	–	12.33	40.34	22.41	58.37
BLIP† (Li et al., 2022)		✓	✓	–	17.10	**46.61**	23.40	115.75
VQADG	Pipeline	✓	✓	–	17.07	44.97	24.15	107.31
	Pipeline+CL	✓	✓	–	17.43	45.32	24.24	109.78
	Joint	✓	✓	–	17.55	45.64	24.45	111.61
	Joint+CL	✓	✓	–	**18.25**	46.57	**24.69**	**118.53**
VisualBert† (Li et al., 2020)		✓	✓	✓	6.30	25.52	30.21	18.42
BLIP† (Li et al., 2022)		✓	✓	✓	9.09	29.76	28.28	32.92
VQADG	Pipeline	✓	✓	✓	10.23	31.70	31.38	56.07
	Pipeline+CL	✓	✓	✓	10.49	32.02	31.48	58.10
	Joint	✓	✓	✓	10.65	32.38	31.58	59.12
	Joint+CL	✓	✓	✓	**11.14**	**33.34**	**31.72**	**63.76**

Table 1: The comparison results between our methods and some baselines on the traditional metrics. The performance of IQ on Visual7W has been reported in (Roy et al., 2022), VisualBert† and BLIP† are implemented by us.

Data Type	Accuracy
VQAv2	55.42
VQAv2+Generated Train	55.81
VQAv2+Generated Val	**56.09**

Table 2: Evaluation on the effect of the generated VQA pairs by VQADG as augmented data to enhance the performance of the VQA task.

Model Type	Raw	Ours
Zero-shot (Dai et al., 2023)	66.57	**63.20**
Fine-tuned (Dai et al., 2023)	68.05	**65.06**

Table 3: Evaluation on the effect of the generated distractors by VQADG. Low accuracy of the VQA task implies that the generated distractors can fool existing VQA models, which demonstrates the superior of our model.

generate questions directly. The Joint model generates QADs and we only take the question part for evaluation. For QA evaluation, the Pipeline model generates questions and answers, respectively. Therefore, we concatenate questions and answers for evaluation and take the question and answer part from the QADs generated from the Joint model. For QAD evaluation, the Joint model can generate QADs, and we concatenate the QADs generated from the Pipeline model. We leverage VisualBert† and BLIP† to generate QADs and take the same operation as the Joint model for evaluation.

On almost all evaluation metrics in three generation tasks, our VQADG model outperforms the baseline models. Specifically, in the question generation task, our four models achieve significant improvements over the traditional model (IQ) and also over the pre-trained vision-and-language model (VisualBert† and BLIP†). In the evaluation of question-answer pairs and QADs generation, our four methods outperform VisualBert† and BLIP† by a large margin, with the exception of BLIP† achieving a higher ROUGE-L score in the question-answer evaluation. Furthermore, we find that contrastive learning loss can improve the performance of the Pipeline and the Joint model. Our Joint+CL model achieves the highest performance across all generation tasks, outperforming the other three variants of our models.

Evaluation on the Generated QADs We applied the InstructBLIP (Dai et al., 2023) to independently evaluate the quality of the generated question-answer (QA) pairs and distractors, utilizing the VQAv2 dataset (Goyal et al., 2017). In the QA evaluation, we harnessed QA pairs from the newly generated data to supplement the VQA dataset. Subsequently, we trained a novel VQA

Model	BLEU-4			ROUGE-L			METEOR			CIDEr		
	Q	A	D	Q	A	D	Q	A	D	Q	A	D
Joint+CL	24.35	7.54	3.32	55.35	20.66	13.65	27.64	17.59	34.33	165.71	65.31	37.13
Joint+CL-w/o QAD	24.08	7.67	3.29	55.20	20.61	13.57	27.53	17.52	34.31	163.56	65.12	36.79
Joint+CL-w/o Q	24.23	8.08	3.32	55.25	20.73	13.74	27.60	17.67	34.34	164.50	66.43	37.51
Joint+CL-w/o A	24.25	7.64	3.25	55.27	20.66	13.65	27.54	17.51	34.26	164.82	64.24	36.80
Joint+CL-w/o D	24.32	7.77	3.32	55.28	20.30	13.44	27.60	17.51	34.37	165.21	64.85	36.63
Joint	23.32	7.59	3.13	54.56	19.44	12.85	27.37	17.47	34.19	157.40	63.80	34.86

Table 4: Ablation study towards contrastive learning, where “w/o QAD” denotes to remove the L_{pqad} loss function, “w/o Q” denotes to remove the L_{pq} loss function, “w/o A” denotes to remove the L_{pa} loss function, and “w/o D” denotes to remove the L_{pd} and L_{pd_a} loss functions.

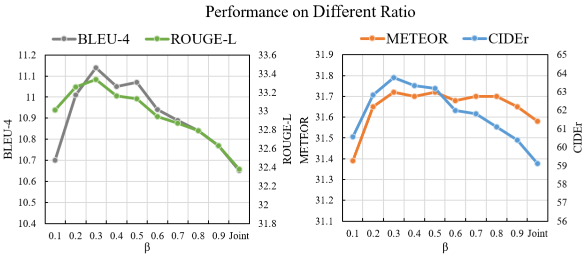


Figure 4: The performance of the Joint+CL model on generating QADs for different ratios (β) of the language modeling loss to contrastive learning loss.

Metric	Type	Content		
		Q	A	D
P(%)	Raw	57.31	83.03	78.04
	Joint	67.10	85.96	77.65
	Joint+CL	**69.43**	**86.07**	**78.51**

Table 5: Consistency evaluation for the raw data, the Joint model, and the Joint+CL model.

Model	C1	A	C2	F
Pipeline	4.23	3.44	2.6	3.02
Joint	4.28	3.85	2.84	2.96
Joint+CL	**4.53**	**4.17**	**3.30**	**3.31**

Table 6: Human evaluation for generated QADs by our models. C1 denotes to consistency score, A denotes to accuracy score, C2 denotes to confusion score, and F denotes to fluency score.

model with both the original and the generated datasets. This model was then put to the test on the original test data. For distractor evaluation, we generated new distractors using the QA pairs from the original dataset and compared them to the original distractors within the dataset.

Table 2 presents the outcomes of the QA pair evaluation. We selected multiple-choice VQA questions from the test dataset and gauged the performance of the VQA model when the generated QA pairs were employed as supplementary data. The experimental findings suggest that incorporating the generated QA pairs can potentially boost the accuracy of pre-existing VQA models. Despite the modest improvement in model performance due to our generated data, it is noteworthy that our data is comparatively easier to obtain than manually curated data and can be produced at a larger scale. Table 3 showcases the evaluation results of the distractors in the VQA model. Notably, both the zero-shot and fine-tuned models exhibit reduced accuracy when confronted with our generated distractors. This decline in performance underscores the challenging nature of our generated distractors and their potential to mislead the VQA model. The results presented in the two tables suggest that our proposed method is capable of generating high-quality QADs, which can be utilized to enhance the performance of existing models.

Ablation Study To verify the effectiveness of different components in the proposed VQADG model (i.e., the contrastive learning loss and the ratio of language modeling loss to contrastive learning loss), we conduct an ablation study in the following experiments.

Table 4 presents the ablation study of contrastive learning loss in each part of QADs. We remove the contrastive learning loss of QADs, question, answer, and distractors from the Joint+CL model, respectively, and then evaluate the QADs generation. As shown in Table 4, compared with the Joint+CL model, the performance of the Joint model would decrease when removing the QADs, question, answer, and distractors contrastive learning loss, respectively. In addition, all of the Joint models with contrastive learning loss perform better than the Joint model. This indicates that all of the four contrastive learning loss functions have boosting effects on the corresponding generation tasks.

Figure 4 illustrates the QADs performance of the Joint+CL model with different ratios of language

	Case 1: Meaningful Question	**Case 2: Correct Answer**	**Case 3: Challenging Distractor**
Input Image			
VisualBert†	**Q: What is the name of the train?** A: Randen. D1: Randall. D2: Wendall. D3: Lionel.	Q: How is the name of the keyboard? **A: Held up by a keyboard.** D1: Open. D2: Down. D3: Pressed.	Q: who is riding the bike? A: The young woman. D1: A little girl. **D2: A young boy.** D3: A man.
Joint	**Q: What is the weather like?** A: Cloudy. D1: Sunny. D2: Rainy. D3: Snowy.	Q: How many buttons are on the keyboard? **A: Four.** D1: Two. D2: Three. D3: Five.	Q: Who is riding the bike? A: A woman. D1: A man. D2: A child. **D3: A teenager.**
Joint+CL	**Q: What is the color of the train?** A: Red. D1: Blue. D2: Green. D3: Yellow.	Q: How many mouse pads are there? **A: Two.** D1: One. D2: Three. D3: Four.	Q: Who is riding the bike? A: A woman. D1: A man. D2: A child. **D3: A dog.**

Figure 5: The figure illustrates the meaningful questions, the correct answers, and the challenging distractors, generated by VisualBert†, the Joint model, and the Joint+CL model, respectively.

modeling loss to contrastive learning loss. As shown in the figure, the Joint+CL model achieves the best result when the ratio is 3:7 (i.e., $\beta = 0.3$).

We also report the percentage of consistent QAPs (denoted as P(%))[4] in Table 5. Note that for the percentage of consistent QADs in distractors, we calculate the consistency scores for the ground truth answer and three distractors, separately, and then select the one with the largest consistency score as the predicted answer. The experimental result demonstrates that our Joint+CL model outperforms the Raw data and the Joint model in the consistency evaluation.

4.5.2. Human Evaluation

To further assess the quality of the generated QADs, we conduct human evaluations on 900 QADs with 300 images generated by our models. We recruit three people to rate them between 1 to 5 points on four qualitative aspects: language fluency, consistency, accuracy, and confusion score, to measure the overall quality of the generated QADs, the question, the answer, and the distractors, respectively.

Table 6 displays the outcomes of human evaluation, revealing that the Joint+CL model achieves the highest scores across all four metrics. It is noteworthy that, for the fundamental variants of our models, the Joint model outperforms the Pipeline model, signifying that QADs generated jointly are superior to those produced in a pipeline manner, with the exception of language fluency. This finding underscores the significance of generating QADs concurrently.

4.5.3. Case Study

We conduct case study to demonstrate the quality of the QADs generated by our baselines and models. Figure 5 presents several QADs generated by VisualBert†, the Joint model, and the Joint+CL model. We can observe that in Case 1, compared with VisualBert† and our Joint model, the question generated by our Joint+CL model is meaningful and has high relevance to the image. In Case 2, our Joint+CL model can generate the correct answer, while VisualBert† and the Joint model generate the wrong answer. In Case 3, our Joint+CL model generates the distractor “Dog” which is more deceptive than those generated by VisualBert† and the Joint model.

[4] More details can be seen in (Yang et al., 2021).

5. Conclusion

In this paper, we have introduced a new task, Visual Question-Answer-Distractors Generation (VQADG). A vision-and-language model has been proposed to encode the multimodal information including the vision features and prefix containing question types. QADs are generated jointly according to an autoregressive text decoder. Furthermore, contrastive learning has been incorporated to cope with the consistency requirement. Our model has achieved superior results under major evaluation metrics on the benchmark dataset. Additionally, the generated QADs have demonstrated effectiveness in enhancing the performance of existing VQA models.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundations of China (Nos. 62306156, 62106091), Fundamental Research Funds for the Central Universities, Nankai University (No. 63231183), and Shandong Provincial Natural Science Foundation (No. ZR2021MF054).

7. Bibliographical References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of NeurIPS*, pages 23716–23736.

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.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*, pages 1877–1901.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, pages 1597–1607.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *Proceedings of ECCV*, pages 104–120.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of ICML*, pages 1931–1942.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. 2018. A question type driven framework to diversify visual question generation. In *Proceedings of IJCAI*, pages 4048–4054.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R Lyu. 2019. Generating distractors for reading comprehension questions from real examinations. In *Proceedings of AAAI*, pages 6423–6430.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of CVPR*, pages 9729–9738.

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.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. In *Proceedings of Findings of EMNLP*, pages 1896–1907.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Proceedings of NeurIPS*, pages 9694–9705.

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.

Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C Lee Giles. 2018. Distractor generation for multiple choice questions using learning to rank. In *Proceedings of ACL*, pages 284–290.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of NeurIPS*.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32.

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*.

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.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of CVPR*, pages 12700–12710.

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. *NeurIPS*, 35:27730–27744.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763.

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(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings ofNeurIPS*, volume 28.

Anurag Roy, David Johnson Ekka, Saptarshi Ghosh, and Abir Das. 2022. Few-shot visual question generation: A novel task and benchmark datasets. *arXiv preprint arXiv:2210.07076*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Shagun Uppal, Anish Madan, Sarthak Bhagat, Yi Yu, and Rajiv Ratn Shah. 2021. C3vqg: category consistent cyclic visual question generation. In *Proceedings of ACM MM*, pages 1–7.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998—6008.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of ICML*, pages 23318–23340. PMLR.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS*, 32.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of CVPR*, pages 6720–6731.

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

8. Language Resource References

Goyal, Yash and Khot, Tejas and Summers-Stay, Douglas and Batra, Dhruv and Parikh, Devi. 2017. *Making the v in vqa matter: Elevating the role of image understanding in visual question answering*. PID https://visualqa.org/.

Lin, Tsung-Yi and Maire, Michael and Belongie, Serge and Hays, James and Perona, Pietro and Ramanan, Deva and Dollár, Piotr and Zitnick, C Lawrence. 2014. *Microsoft coco: Common objects in context*. PID https://cocodataset.org/.

Zhu, Yuke and Groth, Oliver and Bernstein, Michael and Fei-Fei, Li. 2016. *Visual7w: Grounded question answering in images*. PID https://ai.stanford.edu/ yukez/visual7w/.