VTQA: Visual Text Question Answering via Entity Alignment and Cross-Media Reasoning

Kang Chen Xiangqian Wu[*]
Faculty of Computing, Harbin Institute of Technology (HIT)
Suzhou Research Institute of HIT
chenkangcs@stu.hit.edu.cn, xqwu@hit.edu.cn

Abstract

Achieving the optimal form of Visual Question Answering mandates a profound grasp of understanding, grounding, and reasoning within the intersecting domains of vision and language. Traditional VQA benchmarks have predominantly focused on simplistic tasks such as counting, visual attributes, and object detection, which do not necessitate intricate cross-modal information understanding and inference. Motivated by the need for a more comprehensive evaluation, we introduce a novel dataset comprising 23,781 questions derived from 10,124 image-text pairs. Specifically, the task of this dataset requires the model to align multimedia representations of the same entity to implement multi-hop reasoning between image and text and finally use natural language to answer the question. Furthermore, we evaluate this VTQA dataset, comparing the performance of both state-of-the-art VQA models and our proposed baseline model, the Key Entity Cross-Media Reasoning Network (KECMRN). The VTQA task poses formidable challenges for traditional VQA models, underscoring its intrinsic complexity. Conversely, KECMRN exhibits a modest improvement, signifying its potential in multimedia entity alignment and multi-step reasoning. Our analysis underscores the diversity, difficulty, and scale of the VTQA task compared to previous multimodal QA datasets. In conclusion, we anticipate that this dataset will serve as a pivotal resource for advancing and evaluating models proficient in multimedia entity alignment, multi-step reasoning, and open-ended answer generation. Our dataset and code is available at https://visual-text-qa.github.io/.

1. Introduction

A paramount objective in AI research is to endow systems with the capacity to comprehend the intricacies of the real world, akin to human understanding. Question Answering (QA) stands out as an effective task for evaluating the cognitive capabilities of AI systems. To answer questions, people need to extract information from multiple modalities, such as text, images and structured data like knowledge bases, graphs and tables. And furthermore, people need to align the information and do multi-step reasoning between different modalities.

Although Visual Question Answering (VQA) [5] has been widely researched as a multimedia QA task, VQA models only extract information from image when answering questions and focus mainly on scene recognition, counting, color and other visual detection tasks, which do not require much logical reasoning or assignment between different modalities. Only recently, there are some attempts to introduce more modal information into VQA tasks. For example: (1) FVQA [24] and KBVQA [23] combine knowledge base (KB) with VQA task, which requires the ability of knowledge understanding and multi-step reasoning. But it is difficult to construct a comprehensive KB in real world, thus limiting its ability to address open-ended questions; (2) TextbookQA [10] and ScienceQA [14] use textbook as data sources, involving texts, images, tables and other multimodal information. But most images in these datasets are manually drawn schematic diagrams, and the questions are all in a multiple-choice setting, which are far from the real world.

To address these issues, more recently, there appear some new datasets such as MultiModalQA [21] and MuMuQA [18], which involve reasoning across texts, images and tables. However, it's noteworthy that all these datasets are conducted with extracted QA. And in the MultiModalQA dataset, each image corresponds to a Wikipedia entity, simplifying image reasoning to ranking images based on these entities, which diminishes the necessity for cross-media reasoning abilities. As for MuMuQA, although it requires grounding between image and text along with multi-hop reasoning, there are still some problems: (1) the data

[*]corresponding author

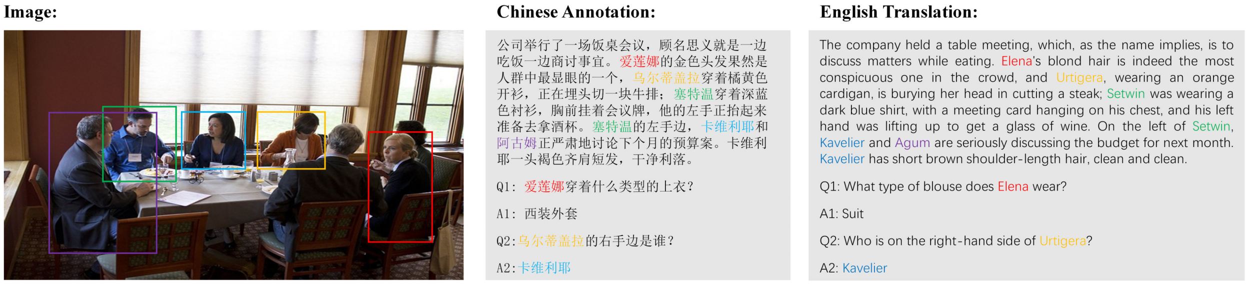


Image:

Chinese Annotation:

公司举行了一场饭桌会议，顾名思义就是一边吃饭一边商讨事宜。爱莲娜的金色头发果然是人群中最显眼的一个，乌尔蒂盖拉穿着橘黄色开衫，正在埋头切一块牛排；塞特温穿着深蓝色衬衫，胸前挂着会议牌，他的左手正抬起来准备去拿酒杯。塞特温的左手边，卡维利耶和阿古姆正严肃地讨论下个月的预算案。卡维利耶一头褐色齐肩短发，干净利落。

Q1：爱莲娜穿着什么类型的上衣？

A1：西装外套

Q2:乌尔蒂盖拉的右手边是谁？

A2:卡维利耶

English Translation:

The company held a table meeting, which, as the name implies, is to discuss matters while eating. Elena's blond hair is indeed the most conspicuous one in the crowd, and Urtigera, wearing an orange cardigan, is burying her head in cutting a steak; Setwin was wearing a dark blue shirt, with a meeting card hanging on his chest, and his left hand was lifting up to get a glass of wine. On the left of Setwin, Kavelier and Agum are seriously discussing the budget for next month. Kavelier has short brown shoulder-length hair, clean and clean.

Q1: What type of blouse does Elena wear?

A1: Suit

Q2: Who is on the right-hand side of Urtigera?

A2: Kavelier

Figure 1. Example in our dataset with the question-answer pairs and their corresponding image and text. Different representations of the same object in text and image are identified with the same color. For example, ’Elena’ in the text and the object bounding box corresponding to ’Elena’ in the image are marked red.

is all from news, resulting in most questions related to human beings; (2) questions all follow a specific pattern: first perform image entity grounding and then find the answer in the news body text; (3) there is only 1384 human-curated examples and the training data is automatically generated, which leads to poor quality and difficult to train with. Current multimedia benchmarks are still far from the real-world QA scene and cannot measure the multimedia understanding ability of AI system well.

In order to address these gaps, we propose a novel task named ‘Visual Text Question Answering’ (VTQA), accompanied by a dedicated dataset, which includes only questions that require multi-hop reasoning through both images and text. All the annotations in VTQA dataset are first marked in Chinese, and then translated into English.

To answer VTQA questions, the proposed model needs to: (1) learn to identify entities in image and text referred to the question, (2) align multimedia representations of the same entity, and (3) conduct multi-step reasoning between text and image and output open-ended answer. The VTQA dataset consists of 10,124 image-text pairs and 23,781 questions. The images are real images from MSCOCO dataset [12], containing a variety of entities. The annotators are required to first annotate relevant text according to the image, and then ask questions based on the image-text pair, and finally answer the question open-ended. To ensure textual richness, we enforce a minimum text length requirement of over 100 Chinese words. Additionally, we systematically exclude questions that can be answered solely based on either the image or the text, thereby guaranteeing the complexity of the questions.

Moreover, we conduct an evaluation on the VTQA dataset, employing both state-of-the-art VQA models and our newly proposed baseline model. The baseline model we propose is called Key Entity Cross-Media Reasoning Network (KECMRN), which follows the general paradigm for answering questions on this dataset by iteratively performing key entity extraction and alignment, and cross-modal multi-step reasoning to answer questions. Evaluation results on the VTQA dataset show that existing state-of-the-art VQA models struggle to achieve satisfactory performance on this dataset, which illustrates the challenges and potential of this dataset for cross-modal question answering tasks.

The contributions of this work can be summarized as follows: (1) we propose a new cross-modal QA dataset named VTQA. Information diversity, multimedia multi-step reasoning and open-ended answer make our dataset more challenging than the existing datasets; (2) we benchmark the state-of-the-art VQA model on our new dataset and show the performance of these models degrades drastically; (3) we propose a baseline that is capable of multimedia entity alignment and multi-step reasoning.

2. Related Work

Visual Question Answering (VQA) aims to answer a natural language question based on an image, which requires model to understand and reason in the vision-language joint space. Several datasets have been proposed in the past few years, such as DAQUAR [15], FM-IQA [4], VQA [2, 5], COCO-QA [19], Visual7W [28], Visual Genome [11], GQA [8], OKVQA [16], A-OKVQA [20], VizWizQA [6] and so on. However, the natural language questions in these datasets can be regarded as instructions to guide the model to complete visual tasks such as object detection, scene recognition, counting, etc. And these questions often only involve the surface information of the image, without considering the complexities between different modalities. understanding of the reasoning process. In order to improve the multimodal perception ability of the model, we need to design more challenging problems, requiring the model to combine the features of images and texts for cross-modal knowledge fusion and logical reasoning.

Recently, some datasets attempt to introduce more modal information into VQA tasks, such as FVQA [24], KB-VQA [23], TextbookQA [10], and ScienceQA [14]. But the information used in these datasets includes manually collected databases or manually drawn schematic diagrams in textbooks, which are far from real-world QA scenarios. The VTQA we propose is different from these, using the semantically rich real images of the COCO dataset and human-annotated text passages, which can be better generalized to real-world QA scenarios.

The most similar to our dataset is MuMuQA [18]. The MuMuQA dataset is a news-based QA dataset whose images and questions mainly involve people and events. This dataset has only a few human-labeled data, and most of the data is generated by automated methods, so there may be a lot of noise and errors. In contrast, our VTQA dataset covers a variety of genres and topics, and all texts, questions, and answers are annotated by humans, ensuring the high quality and accuracy of the data. Our VTQA dataset has 23781 items, which is richer and more diverse than MuMuQA dataset.

3. VTQA Dataset

In this section, we present the details of our Visual Text Question Answering (VTQA) dataset. We first introduce the VTQA task of our dataset (Section 3.1) and then detail how the dataset was constructed (Section 3.2), along with the statistical properties of our dataset.

3.1. VTQA Task

As illustrated in Fig. 1, given an image-text pair and a question, a system is required to answer the question by natural language. Importantly, the system needs to: (1) analyze the question and find out the key entities, (2) align the key entities between image and text, and (3) generate the answer according to the question and aligned entities. For example, in Fig. 1, the key entity of Q1 is “Elena”. By referencing the descriptor “gold hair” in the text, it is discerned that the second person from the right in the image is “Elena”. Subsequently, the answer “suit” is generated based on the visual information. As for Q2, which is a more complex question, the previous steps need to be repeated several times to answer it.

3.2. Dataset Construction

To promote the progress of this open-ended multimedia multi-hop VTQA task, we collect a new dataset. Our dataset consists of 10124 image-text pairs and 23,781 questions. We collect data through newly developed annotation interface.

In the first round of labeling, we present images from COCO and corresponding image descriptions, as well as the object detection labels, to the annotators. The annotators are required to generate a text that exceeds 100 words, which should involve the object in the image and contain information that is not included in the image description. To ensure compliance with annotation requirements, annotators can choose to skip some images (12198 images were ultimately skipped, 9830 images were selected, and some of them were selected more than once). The annotators then come up with questions based on the image-text pair and the annotation process requires that the questions cannot be answered only by image or text. Each image-text pair is labeled with 1-4 questions.

In a second round of labeling, different annotators were asked to determine whether the question could be answered solely by relying on either images or text within the corresponding text-question pair or image-question pair. Through this step, we filtered down to 23781 questions from a pool of 28919 questions. The questions that pass this step will be labeled with answers and answer categories. We set three categories for answers: (1) ***YN*** means yes-or-no answer, (2) ***E*** means that the answer is extracted from the text, and (3) ***G*** means that the answer is generated from the text-image pair. And the annotators need to label ‘yes’ or ‘no’ in English for the yes-or-no answer, since there are too many words to express ‘yes’ or ‘no’ in Chinese, for example, ‘可以’ and ‘是的’ both mean ‘yes’.

We randomly split the dataset into training, validation, test-dev and test splits and each image will only appear in one split. Each split has 11312, 1245, 2189, 9035 samples, respectively.

3.3. Dataset Analysis

Table 1 shows a comparison of VTQA and other VQA datasets. As shown in the table, VTQA is much larger than most other datasets, especially with a much larger number of multimodal questions than existing multimodal QA datasets. The text length of the VTQA dataset is shorter than that of MuMuQA, but still much longer than the length of the questions in all datasets, and the texts in VTQA are manually generated, containing more types of entities than existing texts extracted from textbooks, wikis, and news.

Images in VTQA dataset are from MSCOCO dataset, which contains multiple objects and rich contextual information. Fig. 2 presents the top-10 categories distribution of the images used in this dataset. Unlike MuMuQA, in which most images are related to people, images in our dataset contain more kinds of objects.

The texts and questions length statistics are shown in Fig. 3. It is evident that most texts range from 300 to 600 and most questions range from 9 to 21. Obviously, compared with the previous VQA datasets (usually less than 20 words), our dataset also puts forward higher requirements for text understanding to extract information from the long text.

	#I	#Q	#MMQ	Contexts	AvgQ	AvgT	Answer Type	Image Source	Text Source
VQA	200K	1.1M	-	I	6.1	-	Open/MC	COCO	-
FVQA	5826	5826	-	I+KB	9.5	-	Open	COCO	-
KBVQA	2402	2402	-	I+KB	6.8	-	Open	COCO	-
TextbookQA	3455	26260	12567	I+T	9.2	-	MC	Textbook	Textbook
ScienceQA	10332	21208	6532	I+T	12.1	-	MC	Textbook	Textbook
MultimodalQA	57058	29918	8240	I+T	18.2	66.2	Open	Wikipedia	Wikipedia
MuMuQA*	1384	1384	1384	I+T	11.8	633.5	Open	News	News
VTQA	9830	23781	23781	I+T	10	238.4	Open	COCO	annotated

Table 1. Statistics for VTQA and comparisons with existing datasets. #Q: number of questions, #I: number of images, #MMQ: number of multimodal questions, AvgQ: average question length, AvgT: average text length. *We only count the manually annotated parts of the MuMuQA dataset.

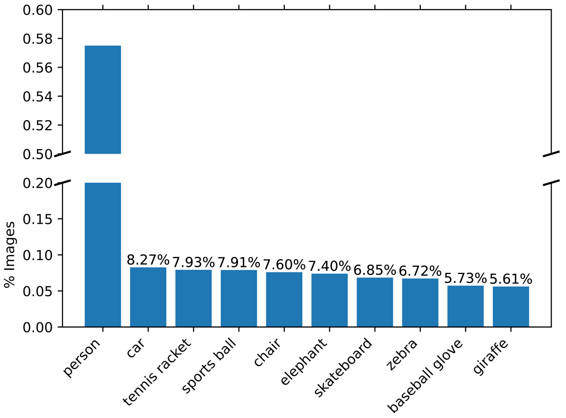


Figure 2. Top-10 categories distribution of the images used in our dataset.

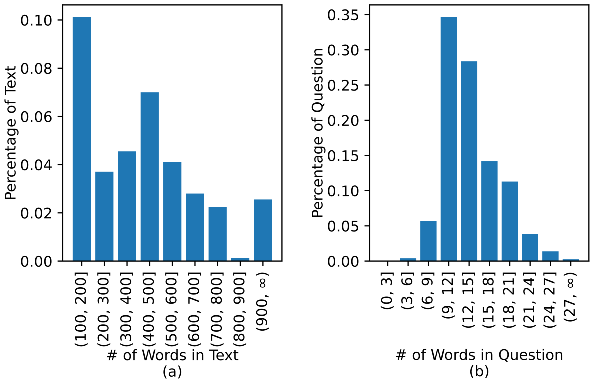


Figure 3. Percentage of questions and texts with different Chinese word lengths.

4. VTQA Model

In this section, we describe a competitive baseline method for evaluation on our benchmark, which is called Key Entity Cross-Media Reasoning Network (KECMRN). Before presenting the KECMRN, we first introduce its basic component, the KECMR layer. The KECMR layer is a modular composition which consists of one Key Entity Extract (KEE) layer and multiple Cross-Media Reason (CMR) layers. The KEE layer and CMR layer are composed of attention unit and feed-forward unit from [22]. Then we use the KECMR module with other layers to combine our KECMRN. An overview flowchart of KECMRN is shown in Fig. 4.

4.1. Attention and Feed-Forward Units

As shown in [22], the combination of attention unit and feed-forward unit has strong representational and learning ability. We use the same settings to construct our units.

Multi-Head Scaled Dot-Product Attention Unit. Given a query $q \in R^{1\times d}$, n key-value pairs (packed into a key matrix $K \in R^{n\times d}$ and a value matrix $V \in R^{n\times d}$), the attended feature is obtained as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

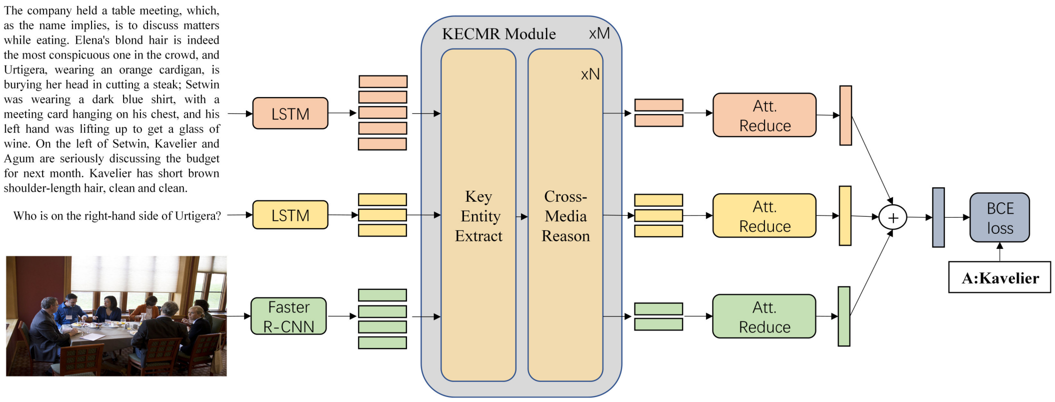
Multi-head attention divides the input into h parts and makes single attention on each part. The attended feature is given by:

$$MH(Q, K, V) = Concat(head_1, \cdots, head_h)W^O \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

where $W_i^Q, W_i^Q, W_i^Q \in R^{d\times d_h}$ and $W^O \in R^{(h\times d_h)\times d}$ are the projection matrices. d_h is the dimension of the output features from each head.

Feed-Forward Unit. The feedforward unit takes the output features of the multi-head attention layer, and further



The company held a table meeting, which, as the name implies, is to discuss matters while eating. Elena's blond hair is indeed the most conspicuous one in the crowd, and Urtigera, wearing an orange cardigan, is burying her head in cutting a steak; Setwin was wearing a dark blue shirt, with a meeting card hanging on his chest, and his left hand was lifting up to get a glass of wine. On the left of Setwin, Kavelier and Agum are seriously discussing the budget for next month. Kavelier has short brown shoulder-length hair, clean and clean.

Who is on the right-hand side of Urtigera?

Figure 4. Overall flowchart of the Key Entity Cross-Media Reasoning Network.

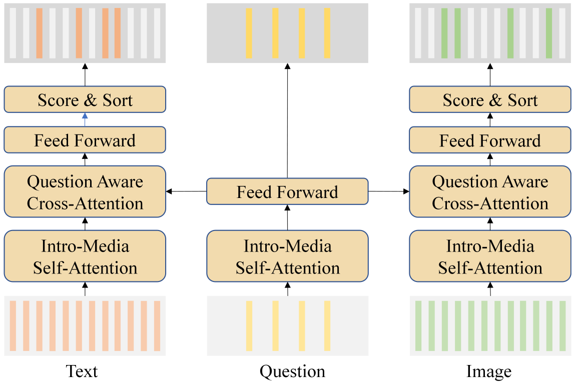

Score & Sort
Feed Forward
Question Aware Cross-Attention
Intro-Media Self-Attention
Feed Forward
Intro-Media Self-Attention
Score & Sort
Feed Forward
Question Aware Cross-Attention
Intro-Media Self-Attention
Text
Question
Image

Figure 5. Key Entity Extract Layer.

transforms them through two fully-connected layers with ReLU activation in between.

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{4}$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer.

4.2. Key Entity Extract Layer

As stated in Section 3.1, the first step to answer VTQA questions is to find out the key entity according to the questions. We compose the attention unit and feed-forward unit to integrate question information into the text and the image respectively. Then we apply a fully-connected layer to the question-aware text/image features to get the score for each feature. Finally, we extract the top-k features as key entities.

The complete KKE layer is shown in the Fig. 5. Mathematical, given input text features T, image features I and question features Q, the KEE layer can be formulated by (the processing of feature I is aligned with that of feature T):

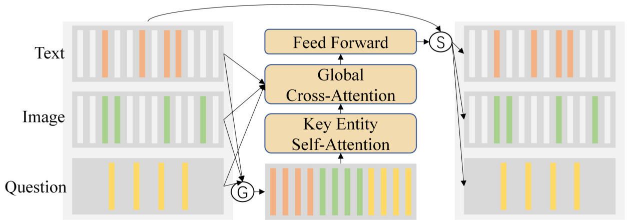


Figure 6. Cross-Media Reason Layer. 'G' means gather and 'S' means scatter.

$$Q = FFN(MH(Q, Q, Q)) \tag{5}$$

$$T = FFN(MH(MH(T, T, T), Q, Q)) \tag{6}$$

$$score_T = W_T T + b_T \tag{7}$$

The final score will be used to sort the image/text features and the top-k features are considered as image/text key entities. To unify the expression, we treat all the question features as key entities.

4.3. Cross-Media Reason Layer

The CMR layer is designed for multi-step reasoning across medias. As show in Fig. 6, we first gather the key entities as $S \in R^{l_k \times d}$ from input features, where l_k is the total number of all key entities. Then the key entities pass through self-attention, global cross-attention based on original features and feed-forward unit. Finally, the key entities are scattered to the input features. As the requirement for multi-step reasoning, we inserted multiple CMR layers in a KECMR module. The CMR layer can be formulated by:

$$S = gather([T, I, Q]) \tag{8}$$

$$S = FFN(MH(MH(S, S, S), [T, I, Q], [T, I, Q])) \tag{9}$$

$$[T, I, Q] = scatter(S, [T, I, Q]) \tag{10}$$

4.4. Key Entity Cross-Media Reason Network

We use the same network framework as [26] but replace MCA layer with our KECMR module and add an extra text stream. As show in Figure 4, The input image is represented as a set of regional visual features in a bottom-up manner [1]. The input question and text are transformed to features by passing through a one-layer LSTM network [7]. Then we use our KECMR module several times to extract key entities and conduct multi-steps cross-media reasoning. Finally, we use the attention reduce layer to fuse the multimedia features and project the fused feature into the answer probability distribution.

The attention reduce layer can be formulated by:

$$MLP(x) = W_2(max(0, xW_1 + b_1)) + b_2 \tag{11}$$

$$AttRe(X) = sum(softmax(MLP(X))X) \tag{12}$$

We can obtain the attended image feature I_a by passing image feature I through attention reduce layer, and the same applies to text feature T and question feature Q. Then we fuse the multimedia features by:

$$z = W_i I_a + W_t T_a + W_q Q_a \tag{13}$$

The fused feature z is projected into a vector $s \in R^N$followed by a sigmoid function, where N is the number of the answers in the training set. We use binary cross-entropy (BCE) as the loss function to train an N-way classifier on top of the fused feature z.

5. Experiments

In this section, we evaluate the current state-of-the-art VQA model and our KECMRN model, and provide results of them in VTQA dataset.

5.1. Evaluation Metrics

As the answers divided into three types, we use different metrics for distinct types of answers.

Exact match (EM). This metric measures the percentage of predictions that match the ground truth answer exactly and will be used in all types of answers.

(Macro-averaged) F1 score (F1). This metric measures the average overlap between the prediction and ground truth answer. We treat the prediction and ground truth as bags of tokens and compute their F1. This metric will be used for the ***E*** and ***G*** types of answers.

YN accuracy (YNAcc). This metric is only used for the ***YN*** type of answers. The answer will be transformed into 'yes' or 'no' by a pre-defined yes-or-no dictionary, which is manually collected based on the answers in our dataset. Then we calculate the accuracy just based on the yes-or-no answer.

5.2. Implementation Details

KECMRN: the hyper-parameters of our model used in the experiments are as follows. The dimensionality of input image features, input question features, input text features and fused multimodal features are 2,048, 1024, 1024, and 2,048, respectively. The latent dimensionality d in the multi-head attention is 1024, the number of heads h is set to 8. The number of CMR layer in each KECMR module and the number of KECMR module are 2 and 2. The number of key entities k is set to 16. We train our model on the train set for 13 epochs and evaluate our model on test set. The Faster-RCNN used for image feature extraction is pretrained and frozen, while all other modules are trained from scratch.

GPT4V [17]**/LLaVA** [13]**/MMICL** [27]: popular multimodal large language models (MLLM). These models demonstrated impressive multimodel chat abilities and exceptional performance across numerous intricate visual reasoning datasets, particularly excelling in zero-shot evaluations. Notably, LLaVA achieved an impressive 80.0% accuracy, while GPT4V attained 77.2% and MMICL attained 70.3% on the VQAv2 dataset, underscoring their remarkable capabilities in the realm of multimodal understanding and visual question answering. We follow the same evaluation strategy as on the VQAv2 dataset, with the sole modification of adapting the prompt to 'Search for answers in the context above, do not use additional knowledge beyond the text, and answer the question using a single word or phrase'. This adjustment facilitated enhanced performance on VTQA datasets.

BEiT3 [25]**/VLMo** [3]: visual language models directly pre-trained on image-text pairs, offering versatility in addressing a spectrum of visual-language tasks through fine-tuning on downstream tasks. Notably, on the VQAv2 dataset, BEiT3-base achieved an accuracy of 77.65%, while VLMo-base demonstrated a commendable accuracy of 76.6%. To evaluate the transferability of BEiT3 and VLMo on our VTQA task, we follow the same fine-tuning strategy as on the VQAv2 dataset, but we concatenate the text and questions into a single paragraph of text and increase the maximum number of training epochs from 10 to 50 to obtain better results, since our dataset is relatively small compared to VQAv2.

MCAN-region [26]**/MCAN-grid** [9]: models based on MCAN [26] but using different image features. They achieve SOTA results on VQAv2 dataset (70.93% for MCAN-region and 72.71% for MCAN-grid) among non-pretrained visual-language models(non-PVLM), which indicates that these models were not pretrained on visual-language data, although individual pretraining on either visual or text data is allowed. We concatenate the text and questions and increase the maximum number of input words to train MCAN in VTQA task.

Trans-CA/Trans-SA: models based on Trans-

Methods	EM	YN-ACC	E-F1	G-F1
English Version				
LLaVA	39.90	65.05	60.06	26.20
GPT4V	48.93	**75.59**	**70.24**	**48.37**
MMICL	38.76	62.90	62.32	45.96
BEiT3-base	41.02	56.51	50.18	28.14
VLMO-base	51.85	65.46	64.59	42.74
MCAN-region	56.09	72.06	67.03	40.52
MCAN-grid	55.88	70.73	65.39	39.19
Trans-CA	57.38	73.57	69.02	44.14
Trans-SA	54.39	67.89	65.95	41.11
KECMRN(ours)	**57.95**	74.99	68.91	44.25
Chinese Version				
MCAN-region	49.78	71.26	59.94	45.74
MCAN-grid	50.22	74.17	59.02	45.59
Trans-CA	49.87	74.27	59.63	47.45
Trans-SA	48.99	73.29	58.26	46.31
KECMRN(ours)	**51.32**	**77.59**	**60.52**	**51.11**

Table 2. Results of our KECMRN on test set compaered with other multimodal models.

former [22]. Trans-CA treat image, text and question features as independent input and perform Self Attention and Question Aware Cross Attention in each layer. The layers in Trans-CA resemble those of the KEE module, excluding the "score&sort" step. Conversely, Trans-SA adopts a different approach by concatenating all image, text, and question features and only applying Self Attention in each layer. The layers in Trans-SA resemble those of the CMR module, excluding the "Global Cross-Attention" step.

5.3. Results and Analysis

Table 2 provides the results of various baselines on the test set of our VTQA benchmark. VLMs, including MMICL, LLaVA, BEiT3, and VLMo, primarily pretrained on English datasets, are exclusively evaluated in the English version.

While models such as GPT4V, MMICL, and LLaVA demonstrate superior performance across a variety of visual-language datasets in zero-shot evaluation, they encounter a significant drop in performance when evaluated on the VTQA dataset. This is partly because the model lacks exposure to our training data during training, unlike the VQA dataset included in the model's VL fine-tuning data. The decline also results from the heightened complexity of our VTQA task compared to VQA, posing new challenges to the model's image-text alignment and inference capabilities.

BEiT3 and VLMo, pretrained directly on image-text pairs, achieve impressive results on various multimodal datasets through fine-tuning on downstream tasks. Despite fine-tuning on VTQA training data, BEiT3 and VLMo still exhibit a substantial performance decrease. This is partly attributed to the limitation that the pretraining data for BEiT3 and VLMo only includes image-text pairs with relatively short texts, while the texts in the VTQA dataset are longer, averaging 284 words in English version. This, once again, underscores the challenges inherent in the VTQA task.

Methods	EM	YN-ACC	E-F1	G-F1
Trans(1)	55.45	72.43	65.35	42.27
Trans(3)	55.07	70.98	65.7	43.58
Trans(6)	54.07	69.43	64.83	43.91
LSTM	57.95	74.99	68.91	44.25

Table 3. Results of different text encoders based on KECMRN.

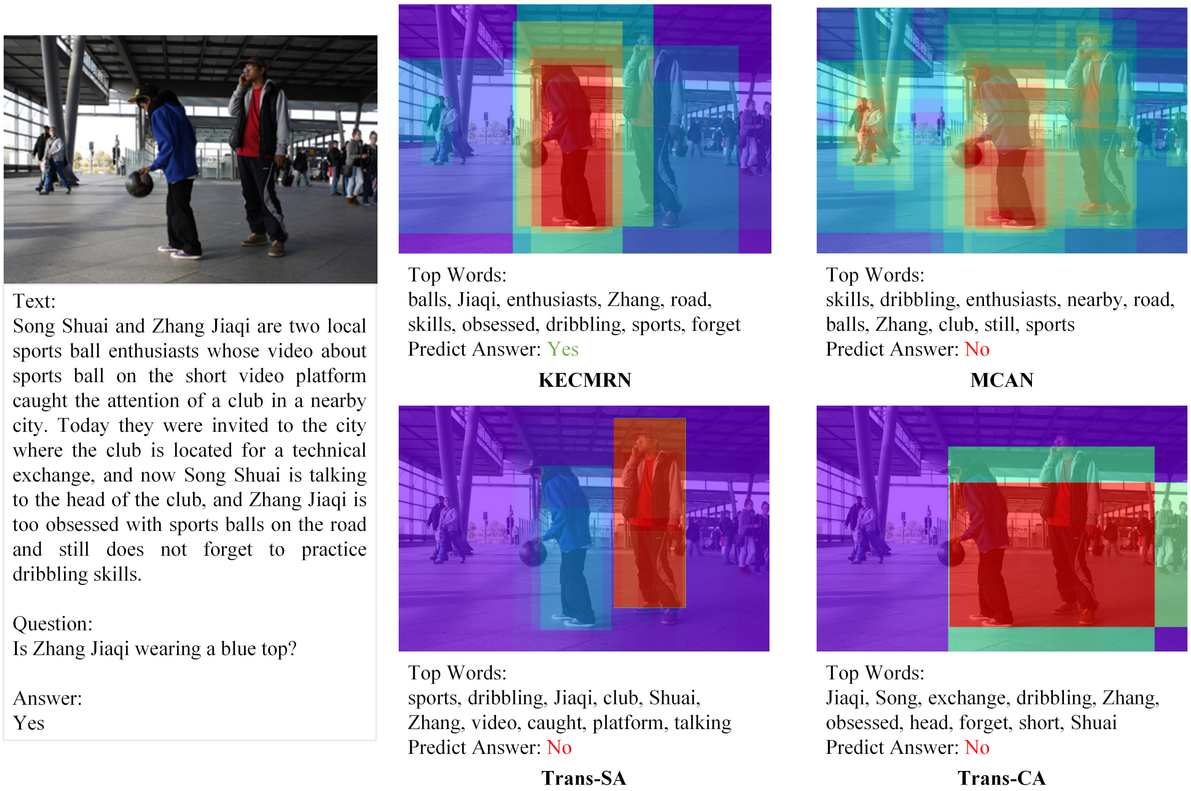
While MCAN models outperform the pre-trained models, a notable decrease is observed. In addition, Trans-CA exceeded the results of all VQA models, while the results of Trans-SA are relatively poor. This highlights the necessity for a more intricate independent processing of text, image, and question features in our VTQA dataset. Our KECMRN achieves the best result with 57.95% on the English version and 51.32% in Chinese version.

Fig. 7 illustrates the outcomes of attention visualization across different models. The attention weights from the 'Attention Reduce' module were employed to colorize image regions and organize the Top Words. Notably, our KECMRN model exhibits precise identification of *Zhang Yanqi* on the left side, engaged in ball play. This accurate recognition is based on key phrases such as *sports balls* and *dribbling skills*, leading to a correct inference for the associated question ('Yes'). In contrast, although MCAN also directs attention to the person on the left, the absence of Key-Entity Extract (KEE) causes attention to be excessively dispersed. Consequently, there is an inaccurate localization of the person's upper garment, resulting in an incorrect answer. Furthermore, Trans-SA mistakenly focuses on the person on the right, while Trans-CA struggles to discern whether *Zhang Yanqi* is positioned on the left or right, both yielding incorrect answers. The attention visualization results underscore the superior precision of our KECMRN model in focusing on the correct visual regions and words.

5.4. Ablation Study

We conducted a number of ablation experiments to investigate the reasons for the effectiveness of KECMR. All the reported results are evaluated on English version.

Text Encoders: Table 3 provides an overview of the results obtained using different text encoders in conjunction with KECMRN. In this table, Trans(k) denotes a k-layer transformer encoder model. Contrary to expectations, the performance of deeper transformer encoder models tends to be inferior, with LSTM outperforming all transformer en-



Text:
Song Shuai and Zhang Jiaqi are two local sports ball enthusiasts whose video about sports ball on the short video platform caught the attention of a club in a nearby city. Today they were invited to the city where the club is located for a technical exchange, and now Song Shuai is talking to the head of the club, and Zhang Jiaqi is too obsessed with sports balls on the road and still does not forget to practice dribbling skills.

Question:
Is Zhang Jiaqi wearing a blue top?

Answer:
Yes

Top Words:
balls, Jiaqi, enthusiasts, Zhang, road, skills, obsessed, dribbling, sports, forget
Predict Answer: Yes

KECMRN

Top Words:
skills, dribbling, enthusiasts, nearby, road, balls, Zhang, club, still, sports
Predict Answer: No

MCAN

Top Words:
sports, dribbling, Jiaqi, club, Shuai, Zhang, video, caught, platform, talking
Predict Answer: No

Trans-SA

Top Words:
Jiaqi, Song, exchange, dribbling, Zhang, obsessed, head, forget, short, Shuai
Predict Answer: No

Trans-CA

Figure 7. An example of attention visualization for different models.

Methods	EM	YN-ACC	E-F1	G-F1
Trans-CA	57.38	73.57	69.02	44.14
CA+SA	57.17	73.57	68.96	42.88
KEE+SA	57.63	74.42	69.1	43.06
KEE+CMR	57.95	74.99	68.91	44.25

Table 4. Results of the KECMRN model with different module combinations.

coder models. This phenomenon is attributed to both insufficient training data and the intricate structure of transformer models.

KEE/CA and CMR/SA: Table 4 displays the results of the KECMRN model with different module combinations. In this table, **CA+SA** denotes the alternate use of CA and SA layers, similar to our KECMRN design. **KEE+SA** follows a similar pattern. Compared to Trans-CA, replacing some CA layers with SA layers results in a slight performance decrease (↓0.21). Conversely, the inclusion of the KEE layer improves performance (↑0.46) by concentrating information on key entities. Additionally, the CMR layer further optimizes features related to key entities, leading to enhanced results (↑0.32).

6. Conclusion

We propose a novel cross-modal question answering dataset, VTQA, which necessitates models to acquire pertinent information from both text and image sources and perform complex cross-modal reasoning to answer questions. Our experimental findings reveal that even state-of-the-art pretrained models fall short of achieving satisfactory performance on VTQA, underscoring the challenges and potential inherent in the dataset for cross-modal question-answering tasks. Looking ahead, we aim to augment the dataset's size and coverage and provide additional reasoning annotations in future work to more effectively assess and catalyze the development of cross-modal reasoning models. Simultaneously, we aspire to explore more fitting evaluation metrics to surmount the limitations of EM metrics and more equitably gauge the model's question-answering proficiency.

7. Acknowledgment

This work was supported in part by the Natural Science Foundation of China under Grant 62073105, U23A20389 and 62306095, by the Natural Science Foundation of Heilongjiang Province of China under Grant ZD2022F002, and by the Heilongjiang Touyan Innovation Team Program.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society, 2018. 6

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 2

[3] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, pages 32897–32912. Curran Associates, Inc., 2022. 6

[4] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 2

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 6325–6334. IEEE Computer Society, 2017. 1, 2

[6] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. 2

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 6

[8] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[9] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10264–10273, 2020. 6

[10] Aniruddha Kembhavi, Min Joon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 5376–5384. IEEE Computer Society, 2017. 1, 2

[11] 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. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, 2017. 2

[12] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 2

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 6

[14] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022. 1, 2

[15] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1682–1690, 2014. 2

[16] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[17] OpenAI, Josh Achiam, Steven Adler, et al. Gpt-4 technical report, 2024. 6

[18] Revanth Gangi Reddy, Xilin Rui, Manling Li, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avirup Sil, Shih-Fu Chang, Alexander G. Schwing, and Heng Ji. Mumuqa: Multimedia multi-hop news question answering via cross-media knowledge extraction and grounding. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11200–11208. AAAI Press, 2022. 1, 3

[19] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 2

[20] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022*, pages 146–162, Cham, 2022. Springer Nature Switzerland. 2

[21] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: complex question answering over text, tables and images. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021. 1

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 4, 7

[23] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, page 1290–1296. AAAI Press, 2017. 1, 2

[24] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. FVQA: fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10): 2413–2427, 2018. 1, 2

[25] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186, 2023. 6

[26] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6281–6290. Computer Vision Foundation / IEEE, 2019. 6

[27] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023. 6

[28] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004, 2016. 2