

MMCoQA: Conversational Question Answering over Text, Tables, and Images

Yongqi Li¹, Wenjie Li¹, Liqiang Nie²

¹The Hong Kong Polytechnic University

²Shandong University

{liyongqi0, nieliqiang}@gmail.com cswjli@comp.polyu.edu.hk

Abstract

The rapid development of conversational assistants accelerates the study on conversational question answering (QA). However, the existing conversational QA systems usually answer users' questions with a single knowledge source, e.g., paragraphs or a knowledge graph, but overlook the important visual cues, let alone multiple knowledge sources of different modalities. In this paper, we hence define a novel research task, i.e., multimodal conversational question answering (MMCoQA), aiming to answer users' questions with multimodal knowledge sources via multi-turn conversations. This new task brings a series of research challenges, including but not limited to priority, consistency, and complementarity of multimodal knowledge. To facilitate the data-driven approaches in this area, we construct the first multimodal conversational QA dataset, named MMConvQA. Questions are fully annotated with not only natural language answers but also the corresponding evidence and valuable decontextualized self-contained questions. Meanwhile, we introduce an end-to-end baseline model, which divides this complex research task into question understanding, multi-modal evidence retrieval, and answer extraction. Moreover, we report a set of benchmarking results, and the results indicate that there is ample room for improvement.

1 Introduction

The ever-increasing variety of information leads to the current information explosion. Question answering (QA) systems play an important role in alleviating information overload by providing users brief and accurate answers. Towards this end, a great many QA systems have been developed by utilizing external knowledge sources to obtain the correct answer, including knowledge-based QA (Deng et al., 2019), document-based QA (Wang et al., 2018), and community-based QA (Fang et al., 2016). Recently, as the rapid

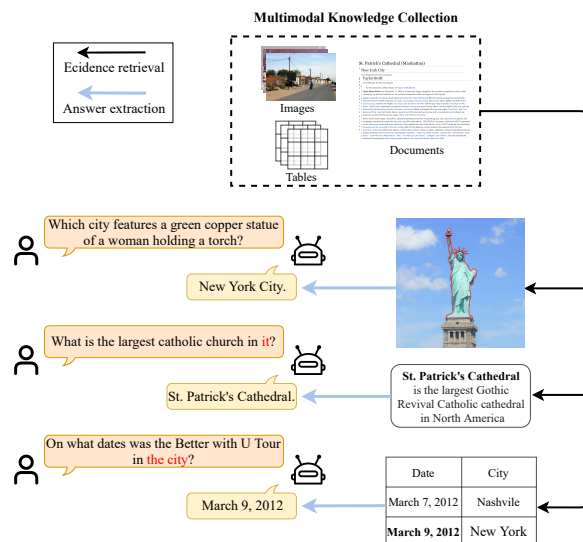


Figure 1: Illustration of multimodal conversational question answering. The user asks questions in a conversation and the QA system extracts accurate answers from the multimodal knowledge collection to satisfy users' information needs.

development of conversational assistants, there is growing interest in all matters conversational. Conversational QA, aiming to satisfy users' complex information needs via multi-turn conversations, attracts a lot of attention.

The existing conversational QA systems usually rely on a single knowledge source, e.g., paragraphs or a knowledge graph, and assume it contains enough evidence to extract answers to users' questions. However, these conversational QA systems are limited in real-world QA scenarios due to the following reasons. On the one hand, the important visual cues are overlooked in the existing conversational QA systems. As an old saying goes, "a picture is worth a thousand words", namely a picture can often vividly express a lot of information. For example, as shown in Figure 1, the question "Which city features a green copper statue of a woman holding a torch?" can be naturally answered by looking at the related picture.

On the other hand, the series of questions in a conversation may dynamically require multiple knowledge sources that encompass different modalities rather than only one constant knowledge source. As shown in Figure 1, three questions in the conversation involve images, passages, and structured tables respectively to extract the correct answers. In fact, although QA systems have been well studied thus far, conversational question answering with multiple knowledge sources of multi-modalities is still untapped. In this paper, we hence define this novel research task, i.e., multimodal conversational question answering (MMCoQA), aiming to answer users’ questions with multimodal knowledge sources via multiturn conversations.

MMCoQA is indeed non-trivial due to the following research challenges. 1) Priority of multimodal knowledge. For a specific question, one modality may be more suitable for locating its corresponding answer than the others. For example, questions about numerical inquiries like date or statistics are better answered by utilizing tables. Different from the previous conversational QA tasks, the most appropriate modality that can be used to answer the current question is not given in MMCoQA. Given the conversation context, how to correctly determine the appropriate modality for the current question is a challenge. 2) Consistency of multimodal knowledge. Different modalities may provide consistent evidence to extract the correct answer for a question. For example, for the first question in Figure 1, the visual modality provides intuitional and direct information, while the related paragraph “The Statue of Liberty ..., off the coast of New York City. She holds a torch in her raised right hand ...” also reveals a certain of cues to indicate the correct answer. How to utilize the consistency among different modalities to verify the answer is another challenge. 3) Complementarity of multimodal knowledge. Some questions may require evidences of different modalities to reason the final answer. For example, the question “Billy Slater played for the NRL team in 2006 with a character holding what on the logo?” must be answered based on both the table about Billy Slater’s career and the image of his team logo. Therefore, to answer these questions, the system is required to have the ability of reasoning across multiple modalities. More importantly, the aforementioned three issues are not standalone but interweaved as conversation goes. Thus, MMCoQA is not the simple combi-

nation of multimodal QA and conversational QA but requires deep multimodal understanding and reasoning abilities across multi-turn conversations, which leaves ample room to study.

To advance the progress of building MMCoQA systems using data-driven approaches, we construct the MMConvQA dataset, the first dataset for MMCoQA (see Table 1). Each question is fully annotated with not only the natural language answer but also the related evidence. Besides, the valuable decontextualized self-contained questions are also annotated for all questions. Hence, MMConvQA can be used to develop individual system modules for multimodal conversational search, conversational question rewrite, and multimodal QA systems. Accordingly, we introduce an end-to-end baseline model and provide a set of bench-marking results, which may facilitate a lot of exciting ongoing researches in the area.

The contributions of this work are threefold:

- To the best of our knowledge, this is the first work towards the multimodal conversational question answering problem. We clearly define the research scope of this task and identify its potential research challenges.
- We construct the first dataset, MMConvQA, for the multimodal conversational QA task. MMConvQA contains multiple supervised labels, including related evidence, answers, and decontextualized questions, which facilitates the data-driven approaches in this community.
- We introduce an end-to-end model as the baseline and report a set of results. Experiment results indicate the significant room for future improvement. Besides, the data and codes of this work are released¹.

2 Related Work

Conversational QA is a relatively new topic in the QA community. Benefiting from the released dataset (Reddy et al., 2019; Choi et al., 2018), text-based conversational QA has been greatly developed. Researchers proposed to model and filter the conversation context via binary term classification (Voskarides et al., 2020) and question rewriting (Elgohary et al., 2019; Vakulenko et al., 2021; Yu et al., 2020). Recently, some efforts (Qu et al., 2020; Li et al., 2022; Anantha et al., 2021b)

¹<https://github.com/liyongqi67/MMCoQA>.

Table 1: Comparison of MMConvQA with datasets from related research tasks. *Conversational* denotes the questions are presented in a conversation, and *Retrieval* denotes the related evidence needs to be retrieved rather than directly given or given along with some negative ones. *DQ* means the dataset contains decontextualized questions.

Task	Dataset	Conversational	Modality	Retrieval	DQ
Conversational QA	QuAC (Choi et al., 2018)	✓	Text	✗	✗
	CoQA (Reddy et al., 2019)	✓	Text	✗	✗
	QReCC (Anantha et al., 2021a)	✓	Text	✓	✓
	CSQA (Saha et al., 2018b)	✓	KB	✓	✗
Multimodal QA	MMQA (Talmor et al., 2021)	✗	Multi	✗	-
	Manymodal QA (Hannan et al., 2020)	✗	Multi	✗	-
Conversational Search	CAsT (Dalton et al., 2020)	✓	Text	✓	✓
	SaaC (Ren et al., 2021)	✓	Text	✓	✗
Multimodal Conversational QA	MMConvQA (this work)	✓	Multi	✓	✓

expanded conversational QA to the open-domain setting, where the related passages must be retrieved rather than given directly. In addition to text-based conversational QA, knowledge-based conversational QA (Christmann et al., 2019) was also developed to answer conversational questions based on a knowledge base. Saha et al. (2018a) created a large-scale dataset and Shen et al. (2019) proposed a multi-task learning framework to resolve coreference in conversations and detect entities simultaneously. However, these existing methods only involved one knowledge source and the important visual cues were overlooked.

Our work is also closely related to multimodal QA. Essentially, the task of VQA (Jing et al., 2020; Shah et al., 2019) is multimodal and involves images and textual questions. However, in this work, we are more interested in question answering with multimodal knowledge sources. In fact, QA with multiple mediums has been studied for a long time. For example, in the year of 2011, Nie et al. proposed to enrich textual question answering with image and video data. Besides, Textbook QA (Kembhavi et al., 2017) and TVQA (Lei et al., 2018) were also explored under specific scenes. Recently, Hannan et al. introduced the ManymodalQA challenge, where the questions are ambiguous and the modality is not easily determined based solely upon the question. Talmor et al. innovatively introduced the complex question scenario to multimodal QA, where a complex question requires several modalities to answer. In this work, we believe that conversational QA is a natural scenario for combining multimodal knowledge sources, where different modalities are dynamically required as the conver-

sation moves on.

3 Dataset Construction

In fact, there are few websites or applications where we can directly obtain a huge amount of questions that are answered with multimodal knowledge sources, let alone in the conversational form. Fortunately, we notice that the MMQA (Talmor et al., 2021) dataset contains a number of complex questions answered with multiple modalities of knowledge. Considering that an important intention of developing conversational QA systems is to gradually satisfy users’ complex information needs via multi-turn conversations (Dalton et al., 2020), we intuitively propose to decompose these complex questions into conversational questions (Saha et al., 2018c). For example, as shown in Figure 2, the complex question “The player not wearing a helmet in 2019-20 Buffalo Sabres season free agents was on what team?” can be better presented in a conversation “Q1: Which player not wear a helmet in 2019-20 Buffalo Sabres season free agents” and “Q2: He was on what team in that season?”. However, if we obtain conversational questions only by decomposing complex questions, the number of questions in a conversation is rather limited since a complex question can only be decomposed into two or three questions. Therefore, we automatically generate potential conversations as references for annotators to refine.

Generate potential conversations. Observing that the follow-up questions in a conversation are usually related to the topics that have occurred in the previous questions or answers, we thus add the questions that contain the same entities into one po-

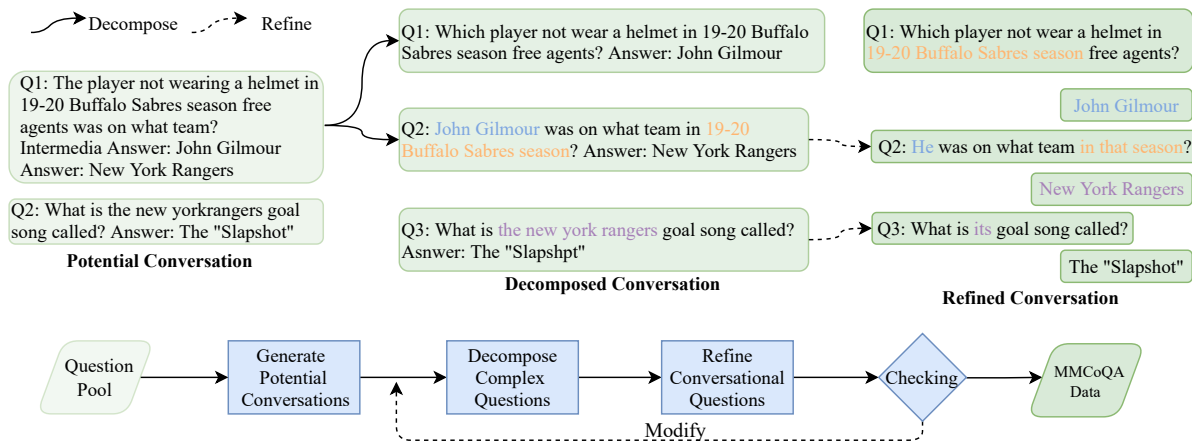


Figure 2: Overall pipeline for MMConvQA dataset creation and a real annotation example in the dataset.

tential conversation. Specifically, for all questions in the question pool, i.e., the MMQA dataset, we identify the entities of the question text and answer text. We then randomly select a question from the question pool as the seed of a conversation. We argue that a user may be interested in the entities that he/she have asked and the new entities occurred in the system’s responses. We hence randomly select one from the identified entities in previous questions and answers as the user’s next point of interest. Then we randomly select a question from the question pool that contains the selected entity as the follow-up question in the conversation. Once a question is selected to conduct a conversation, it will be removed from the question pool to keep the diversity of conducted conversations. Continually add follow-up questions until the conversation turn exceeds a certain number or there is no corresponding question in the question pool. Repeat the above process, and finally we obtain a number of artificial conversations.

The automatically generated conversations are unnatural: 1) there are a lot of complex questions requiring multi-hop logical reasoning, which are not common in daily conversations (Reddy et al., 2019). 2) The sequential questions in the potential conversations lack coherence of dialogues such as coreference and ellipsis. And 3) some questions in the conversation may be not consistent with the whole conversation. Therefore, annotators are involved to manually decompose complex questions and refine (including rewrite, delete and rearrange) the conversational questions towards the real conversation scenario.

Decompose complex questions. To facilitate the decomposition of complex questions, the types and intermediate answers of complex questions

provided in the MMQA dataset, are also shown to the annotators. The types of questions indicate the logic and the target number of decomposed questions for a complex question. For example, Q1 of the Potential Conversation in Figure 2 is a complex question and its type is “Compose(TableQ, ImageQ)”. “Compose(A,B)” means question A containing an entity is the answer of question B, while “TableQ” and “ImageQ” indicate that questions A and B can be answered with tables and images, respectively. Therefore, the annotator can easily decompose this complex question into two sequential questions according to its type. Notably, each annotator is required to decompose a complex question into self-contained questions that can be answered without the conversation context.

Refine conversational questions. After the decomposition, the same annotator refines conversational questions for an artificial conversation. Each annotator is showed some typical examples in the existing conversational QA datasets before the annotation. After fully understanding the linguistic phenomena in conversational QA, such as coreference and ellipsis, annotators write conversational questions for artificial conversations. It is worth mentioning that they have the right to delete and rearrange questions in a conversation to guarantee the smooth conversation flow. They can also report to delete a whole conversation if they think it is poor-quality.

Data quality. Four students that majored in computer science and have NLP research experience are invited to make annotations. To ensure the quality of collected conversation data, we apply the 5-step scheme of training, annotation, checking, modification, and re-checking. Before the collection of data, we carry out training for all participants to

Table 2: Statistics for MMConvQA dataset. TextQ, TableQ, and ImageQ refer to the questions related to text, tables, and images, respectively.

	items	values
	# Dialogs	1,179
	# Questions	5,753
QA pairs	# Avg.Q in Dialogs	4.88
	# Min.Q in Dialogs	3
	# Max.Q in Dialogs	10
Knowledge Collection	# Passages	218,285
	# Tables	10,042
	# Images	57,058
Modality Analysis	# TextQ	2,624 (45.6%)
	# TableQ	1,715 (29.8%)
	# ImageQ	1,414 (24.6%)

explain the annotation guidelines (see Appendix A) for about two hours. Each conversation will be checked by another annotator and the unqualified ones will be returned to modify. It is worth mentioning that since we only write conversational questions rather than give answers to questions, we do not need to calculate the annotation agreement of answers.

4 Dataset Analysis

MMConvQA contains 1,179 conversations and 5,753 QA pairs. There are 4.88 QA pairs on average for each conversation, as summarized in Table 2. The multimodal knowledge collection consists of 218,285 passages, 10,042 tables, and 57,058 images. Each question is annotated with the related evidence (a table, an image or a passage in the knowledge collection), and a natural language answer. Besides, each question is also accompanied with a corresponding self-contained question.

Question Analysis. Figure 3 shows sunburst plots of question types in MMConvQA. We can see that most of the first words are similar to those questions in other conversational QA datasets (Choi et al., 2018; Reddy et al., 2019; Saha et al., 2018b). “The” and “In” are frequently used because they usually relate to the coherence of conversations, such as “The actor”. There are also some special patterns in MMConvQA featuring multi-modalities. For example, “What Color” pattern is related to the visual modality and “How Many” may refer to the tables. On average, each question contains 14.4 words, while this number in the MMQA dataset is 19.2. This illustrates that we well decompose the

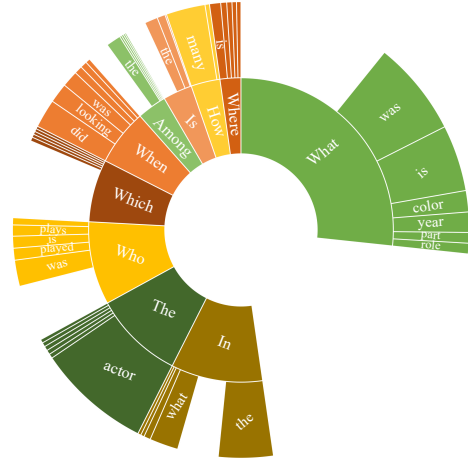


Figure 3: Distribution of the digram prefixes of questions in the MMConvQA dataset.

complex questions. The average number of words for gold questions is 15.5, which is slightly bigger than that of the conversational questions. This is because conversational questions embody the linguistic phenomena of dialogues, such as coreference and ellipsis, thus have less words than gold questions. It is worth mentioning that two different complex questions may produce the same single question. Therefore, there are some duplicated questions in a conversation.

Answer Analysis. The types of answers in MMConvQA are diverse. Most of answers are text spans of passages, cells of tables, and titles of images, whereas some answers do not exactly overlap with the evidence. For example, the answer to the question “The singer of ‘Take Me As I Am’ is shown wearing what item on her neck?” is “scarf”, which needs to be detected based on a related image rather than the title of the image. Apart from single answers, 9.9% questions require a list of answers. For example, the answer to the question “who is the owner of cape town knight riders?” is “Shah Rukh Khan and Juhi Chawla”. On average, each answer contains 2.11 words.

Modality Analysis. As summarized in Table 2, there are 45.6% questions can be answered with textual passages. Besides, 29.8% and 24.6% questions must be answered based on images and tables, respectively. Among the 1,179 conversations in this dataset, 57.7% conversations involve two different modalities of knowledge and 24.4% conversations involve three modalities. This indicates that as conversations proceed, questions dynamically require different modalities of knowledge to answer. To better illustrate the conversation flow,

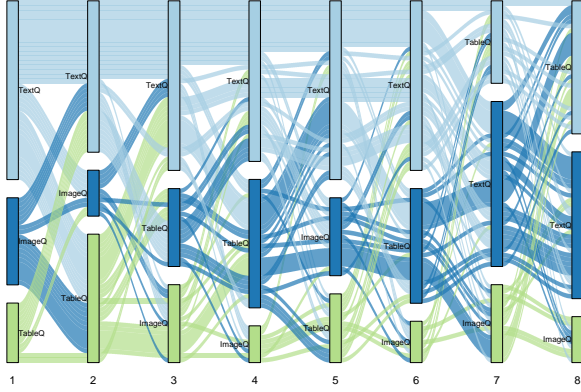


Figure 4: Transitions of modalities as conversation goes. The x-axis indicates the turn number and the y-axis indicates the modalities of questions. The height of one modality reflects the number of questions in each conversation turn of that modality, and the width of the bonds is proportional to the frequency of transition among modalities.

we visualize the transition of modalities as conversation progresses in Figure 4. It is observed that the transitions of modalities are frequent. For example, about 70% table questions at the first turn transform to the text and the image questions at the second turn. And as the turn number increases, the bonds become cluttered, which indicates that more conversations involve multiple modalities.

Linguistic Phenomena. To measure the quality of the conversational questions and analyze their linguistic phenomena, we sample 100 follow-up questions in the development set and annotate various phenomena. Our analysis shows that around 33% questions do not rely on coreference with the conversational history and are answerable on their own. Around 57% questions contain explicit coreference markers such as he, she, it. The remaining 10% do not have explicit coreference markers but refer to an entity or event implicitly. Another feature of open-retrieval conversational QA is the topic switch. Among the questions, 24% change the conversation topic (WikiEntity).

5 MAE Model

We introduce a Multimodal Conversational QA system with Adaptive Extractors, MAE for short, as a baseline model. As Figure 5 illustrates, MAE divides the MMCQA task into three steps: conversational question understanding, multimodal evidence retrieval, and adaptive answer extraction.

5.1 Problem Formulation

Assume that the current turn in a conversation is k and the current question is q_k . The conversation context for the current question q_k is denoted as $\mathcal{H}_k = \{q_1, a_1, \dots, q_{k-1}, a_{k-1}\}$. A multimodal knowledge collection that contains different modalities of items is given, denoted as $\mathcal{C} = \{\mathcal{C}_p \cup \mathcal{C}_t \cup \mathcal{C}_i\}$, where \mathcal{C}_p , \mathcal{C}_t , and \mathcal{C}_i are the sets of passages, tables, and images, respectively. The system is required to retrieve the related evidence from the knowledge collection \mathcal{C} and extract a natural language span \hat{a}_k to answer the question q_k .

5.2 Question and Multimodal Knowledge Encoder

To understand the current question with the conversation context \mathcal{H}_k , we apply the sliding window mechanism (Qu et al., 2020) to filter the previous questions. We feed the reformatted question q'_k into the BERT network (Devlin et al., 2019) to obtain the question representation, which is formulated as,

$$\mathbf{v}_q = \mathbf{W}_q \mathbf{F}_q(q'_k), \quad (1)$$

where \mathbf{F}_q is the BERT based question encoder, \mathbf{W}_q is the question projection matrix, and $\mathbf{v}_q \in \mathbb{R}^{d_q}$.

For different modalities of items in \mathcal{C} , we pass them to different knowledge encoders. For each passage p_j in \mathcal{C}_p , we obtain its representation \mathbf{v}_p^j as follows,

$$\mathbf{v}_p^j = \mathbf{W}_p \mathbf{F}_p(p_j), \quad (2)$$

where \mathbf{F}_p is the BERT based passage encoder, \mathbf{W}_p is the passage projection matrix, $\mathbf{v}_p^j \in \mathbb{R}^{d_p}$. Following the prior work (Herzig et al., 2020; Talmor et al., 2021), we linearize tables by rows as t'_j to obtain their representations. The table's representation is computed via,

$$\mathbf{v}_t^j = \mathbf{W}_t \mathbf{F}_t(t'_j), \quad (3)$$

where \mathbf{F}_t is the BERT based encoder and $\mathbf{v}_t^j \in \mathbb{R}^{d_t}$. For an image i_j in \mathcal{C}_i , its representation is obtained as,

$$\mathbf{v}_i^j = \mathbf{W}_i \mathbf{F}_i(i_j), \quad (4)$$

where \mathbf{F}_i is a pretrained ResNet (He et al., 2016) network on ImageNet, and $\mathbf{v}_i^j \in \mathbb{R}^{d_i}$. Noticed that d_q, d_p, d_t, d_i have the same dimension.

5.3 Evidence Retrieval

To facilitate large-scale retrieval, we apply the dense retriever mechanism inspired from open-domain QA (Karpukhin et al., 2020). Differently,

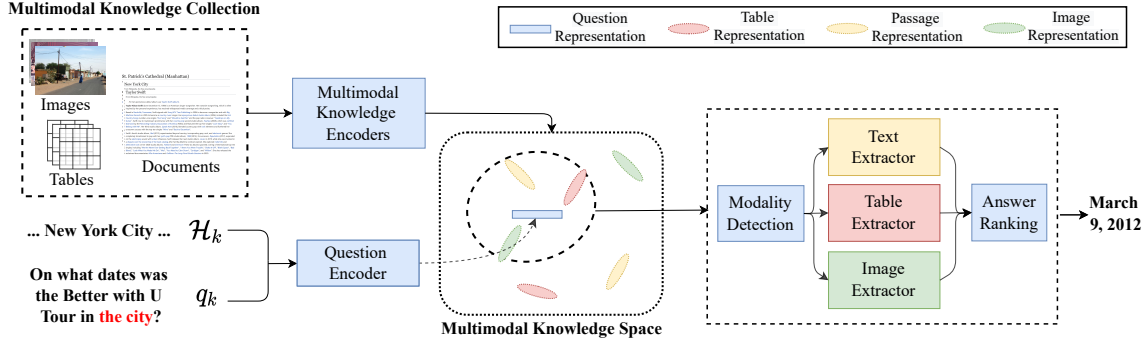


Figure 5: Illustration of the baseline model MAE.

we have three knowledge encoders \mathbf{F}_p , \mathbf{F}_t , \mathbf{F}_i , and they are independent from questions in order to enable strong precomputed multimodal encodings and execute the efficient maximum inner product search (Lee et al., 2019). We first pretrain the question encoder and the knowledge encoders and then input all items in \mathcal{C} into knowledge encoders to obtain their representations. The parameters of the knowledge encoders are frozen in the following training phases. Benefiting from this, we can efficiently calculate the similarity s_a between a given question embedding \mathbf{v}_q and all knowledge item embeddings via the inner product, and select the top- N_r items \mathcal{L}_r as evidence, where N_r is the number of the retrieved items.

5.4 Adaptive Answer Extraction

The retrieved evidence list \mathcal{L}_r contains items of different modalities, and different modalities need different answer extractors. We hence first detect the most appropriate modality for the question.

We regard the modality detection as a multi-class classification task where the network takes a question as input to predict the probabilities of three modalities. The classifier is formulated as,

$$\mathbf{s}_b = f(\mathbf{W}_c \mathbf{F}_c(q'_k)), \quad (5)$$

where $f(\cdot)$ denotes the softmax function, \mathbf{F}_c is the question encoder and $\mathbf{s}_b \in \mathbb{R}^3$.

TextExtractor. It is basically a machine reading comprehension model. Given the reformulated question and a passage in \mathcal{L}_r as input, TextExtractor predicts an answer span by computing two scores for each token in a passage in \mathcal{P}_r to be the start token and the end token, respectively.

TableExtractor. Following the previous work (Herzig et al., 2020), we concatenate the question text to the linearized table sequence, and encode them using BERT. Two linear classifiers are then followed to compute the probability of the

token being the start token and the end token of the answer span, respectively.

ImageExtractor. We collect the answers in the training set as the answer set for testing (Talmor et al., 2021). We extract the visual feature \mathbf{v}_i for an image with the ResNet, and append the question text with all the answers in the answer set as a text sequence. And then we input the text sequence into the BERT to obtain the representations for all tokens, which are then simply combined with the visual feature \mathbf{v}_i . Similarly, two linear classifiers are then followed to compute the probability of the token in the text sequence being the start token and the end token.

The answer extraction score s_c for a candidate answer predicted by the above three extractors is defined as the average of the probabilities of the start and the end token. For each candidate answer, we compute its final score as the sum of the retrieval score s_a , the modality score s_b , and the answer extraction score s_c . The training details are illustrated in Appendix B.

6 Experiments

6.1 Evaluation Protocols

We comprehensively evaluated the baseline models based on their performance in evidence retrieval and answer extraction. We adopted Recall and NDCG to evaluate the coverage and the rank position of the retrieval list. Following previous conversational QA tasks (Reddy et al., 2019), we reported macro-average F1 in the word level and Exact Match (EM) to estimate the performance of answer extraction.

6.2 Baseline Models

We evaluated the open-retrieval conversational QA system **ORConvQA** and a multimodalQA model **ManyModalQA** on our MMCoQA dataset. And

Table 3: Performance of various methods on the test set. ER denotes the related evidence needs to be retrieved, and EG means the related evidence is manually included. Recall and NDCG are computed for top-2000 retrieved items.

Methods	Dev				Test				
	Recall	NDCG	F1	EM	Recall	NDCG	F1	EM	
ORConvQA	14.11	1.91	3.02	1.20	19.05	2.34	1.87	1.06	
ManyModelQA	-	-	2.31	0.73	-	-	1.82	0.96	
ER	MAE	40.96	6.08	2.39	1.20	41.53	6.10	2.19	1.36
	w/o context	31.17	4.32	2.13	0.71	33.28	4.63	1.74	0.82
	Gold question	62.13	12.43	7.06	3.27	63.39	12.46	6.29	3.73
	Gold answer	39.93	5.94	3.49	2.24	42.54	6.54	3.58	2.88
EG	MAE	100	11.97	26.83	19.79	100	11.96	28.33	22.03
	w/o context	100	9.68	22.15	19.54	100	9.73	24.16	18.41
	Gold question	100	15.93	32.89	23.58	100	15.84	36.93	28.31
	Gold answer	100	11.20	30.18	21.51	100	11.78	32.29	24.92
MAE (QR)	45.34	8.37	4.88	2.31	46.32	8.78	4.91	2.92	
MAE (Pretrain)	42.17	7.21	4.59	2.07	42.71	7.66	3.59	2.88	

to better illustrate the characteristics of the dataset, we developed several variants of the MAE: including **w/o conversation context**, **gold question**, **gold answer**, **evidence given**, **QR(question rewrite)**, and **pretrain**. Please see Appendix C for the implement details.

6.3 Results Analysis

The results are summarized in Table 3. By analyzing the results, we gained the following insights. (1) The existing open-retrieval conversational QA and multimodal QA methods cannot handle the MM-CoQA problem well, since they are either single-modal or single-turn. (2) The results of the MAE variants partly evaluate the quality of the MMConvQA dataset. When the conversation context is removed, the performance drops, which verifies the dependency on the conversation context. Appending the previous gold answers or directly using the gold questions improve the performance, which is consistent with the dataset construction strategy. (3) Using extra data benefits the model, which illustrates that the size of the dataset is kind of small and pretraining can alleviate this problem. (4) When we manually complemented the relevant evidence into the retrieval list, it outperforms the normal MAE model a lot. It seems that the evidence retrieval is a bottleneck for the current model because the relation among multimodal knowledge is complex as claimed before.

Modality analysis. We summarized the performance of MAE-EG on the three different modal questions in Figure 6(a). It can be seen that the performance on ImageQ is the worst. It may be because that our ImageExtractor is a little coarse and

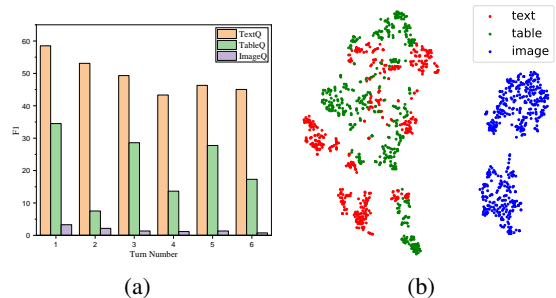


Figure 6: (a): Performance of MAE-EG on TextQ, TableQ, and ImageQ across conversation turns. (b): Visualization of the three different modal items.

more fine-grained interactions are expected. Besides, we selected some items that associated with same entities and visualized their embeddings in Figure 6(b). It is observed that the images’ embeddings are isolated, which illustrates that the visual and semantic meanings are not well-aligned. Some text’s and tables’ embeddings are partly syncretic, but it is still far away from an ideal common space where the embeddings of different modal items are evenly distributed according to their meanings. It seems that the successful dense retrieval scheme for document retrieval needs to be further modified for the multi-modal retrieval.

7 Conclusion

We define a novel and practical task, i.e., MM-CoQA, and identify its research challenges, including priority, consistency, and complementarity of multimodal knowledge. We construct the MMConvQA dataset, containing multiple supervised labels to facilitate related researches in this community. We also report a set of results and analyze the current bottleneck.

Acknowledgments

The work described in this paper was supported by Research Grants Council of Hong Kong (PolyU/5210919, PolyU/15207821), National Natural Science Foundation of China (62076212) and PolyU internal grants (ZVQ0).

References

- Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021a. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the International Conference of the North American Chapter of the Association for Computational Linguistics*, pages 520–534.
- Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021b. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the International Conference of the North American Chapter of the Association for Computational Linguistics*, pages 520–534.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.
- Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 729–738.
- Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. Cast-19: A dataset for conversational information seeking. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, pages 1985–1988.
- Yang Deng, Yuexiang Xie, Yaliang Li, Min Yang, Nan Du, Wei Fan, Kai Lei, and Ying Shen. 2019. Multi-task learning with multi-view attention for answer selection and knowledge base question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6318–6325.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the International Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.
- Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 5918–5924.
- Hanyin Fang, Fei Wu, Zhou Zhao, Xinyu Duan, Yueting Zhuang, and Martin Ester. 2016. Community-based question answering via heterogeneous social network learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1.
- Darryl Hannan, Akshay Jain, Mohit Bansal, li, and li. 2020. Mnymodalqa: Modality disambiguation and qa over diverse inputs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7879–7886.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the International conference on computer vision and pattern recognition*, pages 770–778.
- Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Mueller, Francesco Piccinno, and Julian Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333.
- Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. 2020. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11181–11188.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.
- 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 the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4999–5007.
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. Tvqa: Localized, compositional video question answering. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379.

- Yongqi Li, Wenjie Li, and Liqiang Nie. 2022. Dynamic graph reasoning for conversational open-domain question answering. *ACM Transactions on Information Systems*, 40(4):1–24.
- Liqiang Nie, Meng Wang, Zhengjun Zha, Guangda Li, and Tat-Seng Chua. 2011. Multimedia answering: enriching text qa with media information. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, pages 695–704.
- Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the International conference on research and development in Information Retrieval*, pages 539–548.
- Siva Reddy, Danqi Chen, Christopher D. Manning, li, and li. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten de Rijke. 2021. [Conversations with search engines: Serp-based conversational response generation](#).
- Amrita Saha, Vardaan Pahuja, Mitesh Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018a. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 705–713. AAAI.
- Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018b. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 705–713.
- Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018c. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 705–713.
- Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8876–8884.
- Tao Shen, Xiubo Geng, Tao Qin, Daya Guo, Duyu Tang, Nan Duan, Guodong Long, and Daxin Jiang. 2019. Multi-task learning for conversational question answering over a large-scale knowledge base. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 2442–2451.
- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hananeh Hajishirzi, and Jonathan Berant. 2021. Multimodal qa: complex question answering over text, tables and images. In *International Conference on Learning Representations*.
- Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the International Conference on Web Search and Data Mining*, pages 355–363.
- Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, pages 921–930.
- Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesaro, Bowen Zhou, and Jing Jiang. 2018. R 3: Reinforced ranker-reader for open-domain question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the International conference on research and development in Information Retrieval*, pages 1933–1936.


A Annotation Guidelines

A.1 Decompose complex questions

For a complex question, we give the question text and its question type to the annotators. We also provide the final and intermediate answers to a complex question. For example, as shown in Table 4, we give the question text of a complex question and its question type “Compose(TableQ,ImageQ)”. We have identified this complex question can be decomposed into two questions: the first one is a “ImageQ” and its answer is “John Gilmour”; The second one is a “TableQ” and its answer is “From New York Rangers”. It is noticed that the question type of the original question is very helpful, because it indicates the logical flow of the complex question. The question types and their meanings are listed as follows:

- TableQ: Return a question asked over tables.
- TextQ: Return a text corpus question.
- ImageQ: Return a question about a single image.

Table 4: An annotation example for constructing the MMConvQA dataset. ♣ denotes the provided information for annotators, and ♠ denotes that the column needs to be filled in by annotators.

Potential Conversation♣	Type♣	Answers♣	Decomposed Type♣	Decomposed Conversation♠	Refined Conversation♠												
The player not wearing a helmet in 2019-20 Buffalo Sabres season free agents was on what team?	Compose(TableQ, ImageQ)	John Gilmour	ImageQ														
		From New York Rangers	TableQ														
What is the new york rangers goal song called?	TextQ	The "Slapshot"	TextQ														
The corresponding evidence for the above three questions♣:		<table border="1" data-bbox="831 573 1029 705"> <thead> <tr> <th>Player</th> <th>Team</th> </tr> </thead> <tbody> <tr> <td>Jean-Sebastien Dea</td> <td>from Florida Panthers</td> </tr> <tr> <td>Andrew Hammond</td> <td>from Minnesota Wild</td> </tr> <tr> <td>Curtis Lazar</td> <td>from Calgary Flames</td> </tr> <tr> <td>Scott Wedgewood</td> <td>to Tampa Bay Lightning</td> </tr> <tr> <td>John Gilmour</td> <td>from New York Rangers</td> </tr> </tbody> </table>	Player	Team	Jean-Sebastien Dea	from Florida Panthers	Andrew Hammond	from Minnesota Wild	Curtis Lazar	from Calgary Flames	Scott Wedgewood	to Tampa Bay Lightning	John Gilmour	from New York Rangers		<div data-bbox="1118 573 1302 705" style="border: 1px solid black; border-radius: 10px; padding: 5px;">When the Rangers score a goal at Madison Square Garden the "Slapshot". (aka "The New York Rangers Goal Song") song is played following....</div>	
Player	Team																
Jean-Sebastien Dea	from Florida Panthers																
Andrew Hammond	from Minnesota Wild																
Curtis Lazar	from Calgary Flames																
Scott Wedgewood	to Tampa Bay Lightning																
John Gilmour	from New York Rangers																

- **Compose(*;*)**: Take a single question containing a single WikiEntity as the first argument, and a single question that produces that WikiEntity as the output answer as its second argument. For example, `Compose("Where was Barack Obama born?"; "Who was the 44th president of the USA?")`, the function replaces the WikiEntity in the first-argument single question with the second-argument single question and returns the resulting a complex question "Where was the 44th president of the USA born?".
- **Intersect(*;*)**: Take two single questions that return lists of more than one WikiEntity, and returns their intersection as the answer. E.g. "Who was born in Hawaii and is the parent of Sasha Obama?".
- **Compare(*;*)**: Take two single questions and each returns one WikiEntity that can be linked to one cell in a table.

When decomposing complex questions, annotators should follow these instructions:

- Decomposed questions keep close to the original questions as possible.
- Decomposed questions should keep consistent with the given answers.
- Decomposed questions should be independent and can be answered without any conversation context.

A.2 Refine conversational questions

After the complex question decomposition step, we have obtained a sequence of single questions. Now

we need to refine these questions into a natural conversation. Please follow these instructions:

- The refined questions should depend on the conversation context as possible and are hard to answer without conversational context.
- Annotators can use some pronouns to replace the entities that occurred in previous questions or answers. Annotators can also use some elliptical sentence like "When?", "How?". Some synonyms are also encouraged.
- Keep the whole conversation smooth as possible. You can rearrange questions and delete some low-quality questions. You can also report to delete the whole conversation.

B Training Details

Recall that we encode all the items \mathcal{C} offline for efficient retrieval. Specifically, we follow the previous work (Qu et al., 2020) to pretrain the three encoders so that it can provide reasonably good retrieval results to the subsequent components for further processing. After offline encoding, a set of item representations are obtained.

We define the loss for the evidence retrieval as,

$$L_{er} = - \sum_{j=1}^{N_r} (y \log(S_a^j) + (1 - y) \log(1 - (S_a^j))), \quad (6)$$

where S_a^j is the retrieval score of an item in \mathcal{L}_r and y denotes whether it is a positive item or not. The modality detection loss L_{md} is a typical cross-entropy loss used for training multi-class classification, while the answer extraction loss used for

training the three extractors is as follows.

$$L_{ae} = - \sum_{k=1}^{N_{to}} (y_1 \log(S_s^k) + (1 - y_1) \log(1 - (S_s^k))) - \sum_{k=1}^{N_{to}} (y_2 \log(S_e^k) + (1 - y_2) \log(1 - (S_e^k))), \quad (7)$$

where y_1 and y_2 indicate whether the token is the start token and the end token, respectively. The final loss is defined as the sum of the above losses.

C Implement Details

The data, code, and parameters are uploaded in the supplementary material. Specifically, the d_q , d_p , d_t , and d_i are set to 128, and N_r is set to 10. In the pretraining phase, we set the batch size to 4 and use Adam optimizer with learning rate 0.0001 to train the knowledge encoders for 12 epochs. In the following training phases, the parameters of knowledge encoders are frozen. We set the batch size to 1 and use Adam optimizer with learning rate 0.0001 to train the question encoder and extractors. We select the parameters that perform best in the development set and evaluate the model in the test set. The experiments are conducted on a server with a 3090 GPU card and Ubuntu operating system.

For ORConvQA and ManyModalQA, we did not use extra data, like VQA data, to pretrain the models for a fair comparison. And since ManyModalQA does not contain a evidence retrieval component, we apply our evidence retrieval component to it. **Without conversation context:** Remove the conversation context, and use the current question alone. **Gold question:** Replace the reformulated question q'_k with the gold question q_k^* . **QR:** Rewrite the current question based on the conversation context. **Pretrain:** Use the ORQuAC data to pretrain the evidence retrieval component. **Gold answer:** Append the previous gold answers to the reformulated question q'_k . **Evidence given:** Manually complement the evidence that supports the answer into the retrieved item list \mathcal{I}_r if it is not retrieved.