ChiQA: A Large Scale Image-based Real-World Question Answering Dataset for Multi-Modal Understanding

Bingning Wang  
Tencent Inc.  
Beijing, China

Feiyang Lv  
Tencent Inc.  
Beijing, China

Ting Yao  
Tencent Inc.  
Beijing, China

Jin Ma  
Tencent Inc.  
Beijing, China

Yu Luo  
Tencent Inc.  
Beijing, China

Haijin Liang  
Tencent Inc.  
Beijing, China

ABSTRACT
Visual question answering is an important task in both natural language and vision understanding. However, in most of the public visual question answering datasets, the questions are human-generated which are specific to the given image, such as ‘What color are her eyes?’. The human generated crowdsourcing questions are relatively simple and sometimes have the bias toward certain entities or attributes [1, 55].

In this paper, we introduce a new question answering dataset based on image-ChiQA. It contains the real-world queries issued by internet users, combined with several related open-domain images. The system should determine whether the image could answer the question or not. Different from previous VQA datasets, the questions are image-independent queries that are more various and unbiased. Compared with previous image-retrieval or image-caption datasets, ChiQA not only measures the relatedness but also measures the answerability, which demands more fine-grained vision and language reasoning.

ChiQA contains more than 40K questions and more than 200K question-images pairs. A three-level 2/1/0 label is assigned to each pair indicating perfect answer, partially answer and irrelevant. Data analysis shows ChiQA requires a deep understanding of both language and vision, including grounding, comparisons, and reading. We evaluate several state-of-the-art visual-language models, demonstrating that there is still a large room for improvements on ChiQA.

CCS CONCEPTS
• Information systems → Multimedia databases.

KEYWORDS
question answering, multimodal learning, datasets

1 INTRODUCTION
Question Answering (QA) is one of the long-standing problems in natural language understanding [20, 46]. Recently, with the introduction of large-scale QA datasets such as SQuAD [49] and Natural Questions [37], combined with the development of large-scale pre-training language models such as BERT [13] and GPT-3 [9], etc, some of the state-of-the-art machine reading comprehension models even surpass human performance1.

Most of the QA datasets are mono-modal, where the questions, the answers and the resources (unstructured text or structured knowledge base) are text-based [7, 37, 49]. However, from the perspective of cognitive science [36, 41] and real-world applications [25], the image sometimes satisfies the user’s demand more straightforwardly and conveniently. For example, to answer the question “what is the size of iPhone 13”, an image illustrating the shape of four different versions of iPhone13 is more clear and more intuitive than a detailed text description. Therefore, it becomes more and more important to incorporate the image in question answering.

A lot of models and datasets have been proposed to tackle the multi-modal visual question answering problem, such as VQA-1.0 and v2.0 [5, 18], CLEVR [32], and GQA [29], etc. In most of those datasets, a synthetic or real-world image is provided and human annotators are asked to generate questions with specific focuses, such as entities, relations, or both. However, there are several limitations of those previous VQA datasets. First of all, most of the previous VQA datasets are image-dependent, that is, the questions are asked after they see the image, such as a question starts with “Do you see a…” . A shortcoming of a context-dependent dataset is that the questions sometimes lack diversity, and models could easily learn the linguistic features merely based on the question and give the answer without understanding the context [18, 55]. Efforts to address this problem have mainly focused on making the datasets more balanced and diverse [18]. Nevertheless, benchmarks that involve crowdsourced question generation are bound to encode certain cognitive and social biases [1, 51]. In addition, the human-generated questions may merely be used for evaluating the model, but they are hard applied to a real-world problem.

1https://rajpurkar.github.io/SQuAD-explorer/
Q: what is the size of iPhone 13

The earth is made up of three different layers: the crust, the mantle and the core. The crust is the outside layer of the earth and is made of solid rock, mostly basalt and granite. There are two types of crust: oceanic and continental. Oceanic crust is denser and thinner and mainly composed of basalt. Continental crust is less dense, thicker, and mainly composed of granite.

Q: what does an alpaca look like

Alpacas look like small llamas or long-necked camels with no humps, especially when recently sheared. They have shaggy necks and camel-like faces with thick lips, pronounced noses, and long ears. Their large, expressive eyes seem to exhibit both wisdom and childlike curiosity.

Figure 1: Examples show that for some questions, image answer is more clear and satisfactory than text-based answers. The left part is the text answer while the right side is the image answer.

Secondly, most of the answers in previous VQA datasets are text-based entities, relations, or explanations, which are associated with local regions in the images. However, for a question like “what does an alpaca look like”, a single image illustrating an alpaca could perfectly serve as the answer. Further providing a complicated description of the alpaca’s appearance is difficult and unnecessary. In addition, short-form text answers sometimes restrict the annotator to focus on local structure but not the global information [15, 37, 63]. Therefore, a text answer may not be indispensable in the scenario of information seeking VQA.

Furthermore, most of the previous VQA datasets are developed in English. Efforts on multi-modal learning in other languages are mainly focused on cross-lingual representation [16, 27]. We argue that sometimes the linguistic information for one language is hard to transfer to other languages in VQA. For example, the pineapple has two references in Chinese: ‘菠萝’ and ‘凤梨’\(^2\), when we asked in Chinese about their difference ‘菠萝和凤梨的区别是什么’, little cross-lingual information could be borrowed from other languages. Therefore, developing a specific VQA dataset for a specific language is also important.

In this paper, we propose a new Chinese image question answering dataset called ChiQA. In ChiQA, given a real-world query and several related images, a three-level 2/1/0 label is assigned to each

\(^2\)https://zh.wikipedia.org/zh-hans/%E8%8F%A0%E8%98%BF, this two species of pineapple only grown in Southeast Asia, sometimes they are deemed as the same fruit but actually they have a minor difference.

Figure 2: Examples of the proposed ChiQA dataset. For each question, several related images are labelled with 2-1-0 scores indicating whether they are answer (✓), not the answer(✗) or the partially answer (✓✗) to the question.
The collection on ChiQA consists of a rigorous inspection step to maintain the quality. We collected more than 40,000 queries with more than 200,000 question-image pairs. Some examples of ChiQA are shown in Figure 2. Qualitative analysis of the data shows that ChiQA requires a deep understanding of both language and vision, including grounding, reading, logical inference, etc. We evaluate several state-of-the-art visual-language models such as ALBEF [38], which are far behind human performance, demonstrating that there is still large room for improvements on ChiQA.

2 RELATED WORK

2.1 Question Answering

Question Answering has been a long-term focus in NLP that can date back to 1960s, when Green Jr et al. [19] proposed a system to answer questions about baseball games. Since then, many works have been done to use diverse data resources to answer any type of question. Recently, due to the development of large-scale dataset SQuAD [49], and the pretraining language models such as BERT [13] and GPT-3 [9] in NLP, the performance on QA has been greatly improved and even surpass the human. However, since the questions in SQuAD are human-created which contains some unfavorable bias [55], recent efforts in QA has focused on open-domain real-world questions, such as Natural Questions [37], MS-COCO [40], etc. Among them, our proposed ChiQA is very related to WikiQA [67], it maintains more than 3,000 questions, of which about two-thirds are not answerable questions since they never have a positive sentence in the corresponding paragraphs. However, our proposed ChiQA is multi-media and large-scale.

2.2 Visual Question Answering

There have been years of effort in connecting the visual and textual information for joint multimodal learning [6, 33, 73]. Geman et al. [17] introduced a restricted visual Turing test to evaluate visual and language understanding. Malinowski and Fritz [42] proposed the DaQUAR, which is a small size VQA dataset built upon indoor scene RGB-D images. Currently, most of the VQA datasets are built based on MS-COCO dataset [40] which provides both images and captions. Antol et al. [5] introduce the task of free-form and open-ended VQA which takes the image and a free-form, open-ended, natural-language question about the image as input, and produces a natural-language answer as the output. VQA2.0 [18] and CLEVR [32] datasets are designed to address bias and reasoning limitations of VQA1.0, respectively. Visual Genome [35] is a very large-scale VQA dataset built with images from the Visual Genome project, it includes structured annotations of scene graphs describing the visual elements of the scenes, their attributes, and relationships between them. Visual7w [71] is a subset of the Visual Genome, specifically, all the objects mentioned in the questions are visually grounded with bounding boxes of their depictions in the images. Wang et al. [64] proposed KB-VQA which requires employing external knowledge rather than learning a mapping from image and question. Tapaswi et al. [61] introduce the MovieQA to answer the question about a movie, it requires multiple sources of information such as video clips, plots, subtitles, etc. Recently, Singh et al. [52] proposed text-VQA, where the task is to answer human-generated questions about the embedded text in the image.

Based on these datasets, a lot of cross-modality frameworks have been proposed to solve this problem, such as modular neural networks [4, 68], attention-based networks [28, 59]. Recently, inspired by the language model pretraining in NLP, some cross-modality vision and language pre-training methods have also been proposed [21, 38, 39, 53, 60]. Those pre-training methods utilize large-scale, probably noisy, image and caption pairs and achieve state-of-the-art results on various VQA problems.

2.3 Image Retrieval

Image retrieval is also related to our proposed ChiQA. In this task, we are given a text, and the system should retrieve related images with respect to the input text from the images pool, or verse visa. Two widely used image retrieval datasets are Flickr30K [69] and MS-COCO [40]. The efforts to improve the performance is focused on large-scale contrastive representation learning [38, 47, 48]. However, in image retrieval, the text is sometimes the descriptive caption of the corresponding image that can be viewed as a sort of translation. While in VQA the text and image are non-parallel that the text may only ask about a small part of the image. Therefore, ChiQA is a more fine-grained task compared with image retrieval.

2.4 Image Captioning and Visual Entailment

Image caption is a popular research area of visual and language understanding that deals with image understanding and a language description for that image. MS-COCO [40] is the most popular benchmark dataset of image captioning. Inspired by the sequence-to-sequence model for machine translation, most captioning models mainly adopt an encoder-decoder framework, where an encoder network encodes the input image into an embedding space and a decoder network takes the encoder feature as input and generates the output caption [3, 50, 62]. Some recent works also utilize large-scale pretraining which greatly improves the performance on image captioning [26, 30, 39, 70].

Visual Entailment is the task to predict whether there exists some relationship between the images and texts, which is inspired by the natural language inference task [8]. Xie et al. [66] design an inference task where the premise is image and the system should predict whether it could entail/contradict the hypothesis text. Suhr et al. [57] proposed the NLVR2 in which the premise consists of two images and the task is to determine whether a caption is true regard to the premise image pairs.

Our proposed ChiQA focuses on real-world multi-modal learning in which the questions are user queries issued to search engine. In addition, we did not utilize MS-COCO as our image source, since it is object-based [40] where the images are obtained by a closed set of object categories such as ‘dog’, ‘banana’, etc. It is not as diverse as the open domain web images. We summarize the key characteristics of the different VQA datasets in Table 1.

3 DATA COLLECTION

Each example in ChiQA contains a user query and five images. Each image is related to but not necessarily answerable to the question. We label each image with three-level labels, 2/1/0, to indicate their answerability. In addition, for each question-image pair, we also
Table 1: A summary of major VQA datasets and their main characteristics. The question in our proposed ChiQA is a real world query that is not specific to the target image. And the images are obtained from the open-domain web.

| Dataset       | Question Source | Image Source       | Answer Type                     | Size   | Domain       | Language   |
|---------------|-----------------|--------------------|---------------------------------|--------|--------------|------------|
| DAQUAR        | Crowdourcing     | NYU-DEPTH-V2       | Objects/Color/Number            | 12,468 | Close        | English    |
| CLEVR         | Synthesis        | Synthesis          | Objects/YesNo/Number            | 999,968| Close        | English    |
| VQA-1.0       | Crowdourcing     | MS-COCO & Abstract Scenes | Objects/YesNo/Number          | 614,163| Open        | English    |
| VQA-2.0       | Crowdourcing     | MS-COCO            | Objects/YesNo/Number            | 1,105,904| Open        | English    |
| Visual Genome | Crowdourcing     | MS-COCO            | Open-ended                      | 1,445,322| Open        | English    |
| Visual7W      | Crowdourcing     | MS-COCO            | Multiple-choice                 | 327,939| Open        | English    |
| Visual Madlibs| Crowdourcing     | MS-COCO            | Close                           | 360,000| Open        | English    |
| KB-VQA        | Crowdourcing     | MS-COCO            | Objects/YesNo/Number            | 2,402  | Close        | English    |
| FVQA          | Crowdourcing     | MS-COCO & ImageNet | Entity                          | 5,826  | Open        | English    |
| FM-IQA        | Crowdourcing     | MS-COCO            | Open-ended                      | 316,193| Open        | Chinese    |
| TextVQA       | Crowdourcing     | Open Images        | Open-ended                      | 45,336 | Open        | English    |
| MovieQA       | Crowdourcing     | Plot Synopses      | Multiple-choice                 | 14,944 | Movie        | English    |
| ChiQA         | User Queries    | Web Score          | Score                           | 210,681| Open        | Chinese    |

Figure 3: The labeling process of the ChiQA dataset. First of all, we use rules and models to select the question from the user queries issued to search engine. Then we adopt Google Image search API to obtain the image URL and then crawl the corresponding image. Then the question, images are labeled by the crowdsourcing workers to obtain labels. Next, the labeled samples are checked by an authoritative checker to maintain the quality of the data. Finally, an active learning process is applied to the labeled data and used to select the remaining data to be labeled.

3.1 Question Collection

We use the user queries issued to search engine as our question source. Since user’s query contains a large amount of samples that have no question intent, such as searching for a specific entity, homepages, etc. We use a binary classification model trained in a weakly supervised way to determine whether a query is a question or not. To build such a model, we collect a large but weakly supervised training data in four ways: 1) we select the query that has a high click ratio of the community question answering (CQA) site, such as Quota, Zhihu, etc, as the positive query; 2) the title of the CQA dataset as the positive data; 3) inspired by Natural Questions [37], we use rules to select the positive query from the log data: a) whether the question contains the interrogative pronoun such as ‘who’, ‘how’, etc. b) the title containing the informative words such as ‘the process of’, ‘the method of’, etc; 4) the query that has high click ratio of the non-QA site, such as video site, homepage, novel site, etc, is treated as the negative sample.

We use the weakly supervised data to train a simple bi-directional LSTM [24] model with binary cross-entropy loss. We found the model is simple yet the performance is satisfactory compared with a more heavyweight model like BERT. On the held-out human-labeled test set, our model achieves more than 90% accuracy with more than 80% recall. We use this model to select more than 75k queries to the next step.

3.2 Image Collection

After we obtain the question, we issued them to Google Image search api³. The Google Image API returns 100 most related images for the input query and every image URL is validated by default. However, we found in some cases the image may not be available, so we use the in-house spider to crawl the images and dismiss those images that are unavailable. Furthermore, we only keep images with both dimensions greater than 200 pixels, and the ratio of large-to-small dimensions is no more than three. In this way, we filter out images that are too small or are very tall or wide, which can be of low-resolution after image augmentations like up-sampling.

³https://serpapi.com/images-results
3.3 Image Labelling

After we collected the question and corresponding five images, we ask crowdsourcing workers to label the answerability. Annotation is performed using a custom annotation interface in Figure 4. The guidelines and interface divide the annotation task into three conceptual stages, where all three stages are completed by a single annotator in a single session.

**Question Identification**: Contributors of ChiQA determine whether the question could be answered by an image. A good image question is the one that itself could be answered by an image, such as ‘what is the difference ...’, the question that is ambiguous, incoherent, or containing baseless statement are regarded as the invalid question and not involves in the subsequent process. Note that 1) we regard the non-factoid opinion-seeking question as good since we believe for some questions like ‘red trousers go well with?’ an image could satisfy the question well compared with the text answer. 2) The question that can be answered by a single entity, like ‘Who is the director of The Shawshank Redemption?’, are regarded as bad question since it requires nothing but world knowledge and grounding. 3) For some queries that are falsely classified as a question by the model introduced in section 3.1, it should be given a label -1.

**Image Identification**: For those questions that are labeled as good (1), annotators then select whether the image is perfect, partially or unable to answer the question, with a three-level 2/1/0 label. For ChiQA, there is some intermediate zone where the answer image could be either right or wrong, so we add a level 1 to indicate the image could somewhat answer the question but the user who takes the image may conduct another query to confirm the answer. We did not adopt the quintile(1-5) labeling standard since we found in the preliminary process that this difference is too subtle to discriminate.

Some guideline examples for labeling the question and image are shown in Table 2.

3.4 Quality Inspection

We employ a very rigid data evaluation process to maintain the quality. First of all, all data are checked by authoritative experts who are well-informed of the rules and could make a fair judgment of the data. For each batch of annotated data, the experts sample one-fifth of the data and annotate it again. If the data had less than 90% accuracy, this batch of data is rejected and relabelled by the crowdsourcing worker.

Secondly, since we employ the outsourcing group as our workforce, the annotation standard and accuracy of different outsourcing groups may vary a lot. Therefore, instead of asking only one group of annotators, we ask three groups of annotators with the same expert checker. The three groups annotate independently and their annotated batched data are checked by the expert, after three rounds of annotation, we select the group that has better average accuracy.

![Figure 4: The annotation interface of ChiQA. The left panel is the label frame and the right is the image.](image-url)
judged by the expert, and the remaining data are annotated by this winner group.

### 3.5 Active Data Selection

One lesson we learned from the previous data collection process is that the dataset involves crowdsourced samples is bound to encode certain cognitive and social biases [1, 18, 51]. To make our data challenging enough and unbiased, we adopt an active-learning based data collection process consisting of two phases. In the first phase, we sample data randomly and annotate them. Then, we use the labeled data obtained from the first phase to train an ALBEF [38] model, a state-of-the-art model on vision and language learning. This ALBEF model is utilized as our selector to select the hard example in the next round: for a new sample, if the model’s prediction has a very high entropy which means the selector has low confidence, so it is hard for the current model and deserves to be labeled, otherwise, the sample may be so easy since the model could make a confident prediction, making this sample not very valuable to be labeled [58].

The active learning data selection process indeed makes the dataset more unbiased. Similar with some previous works [5], we found the labeled data in the first phase contains some undetectable fake patterns. For example, the question including the word ‘的技’ (the skill of) are labeled as the valid question but nearly all of the corresponding images are labeled as unanswerable. Consequently, the model trained on this dataset may disregard the image and predict a very low score only if the question contains that word. The active learning process makes this biased sample less likely to be annotated in the next round, making the data more balanced.

### 3.6 Test Set Collection

Since the test set demands a higher quality than the training set, we get the test set as follows: First of all, we randomly sample 2,500 samples from the training set and ask the crowdsource workers (probably different from the annotator that previously label this sample) to annotate them again. If the new labels are same as the old labels, they are passed to the test set; otherwise, we ask the authoritative experts to relabel them again. Finally, we obtain 2,362 test questions and more than 40,000 training questions, the statistics of ChiQA are shown in Table 3:

|       | #Questions | #Images | #2 | #1 | #0 |
|-------|------------|---------|----|----|----|
| Train | 40,635     | 198,440 | 35,771(18.03%) | 73,660(37.12%) | 89,009(44.85%) |
| Test  | 2,362      | 12,214  | 1,760(14.51%)  | 3,604(29.41%)  | 6,850(56.08%)  |

Table 3: The statistics of ChiQA. #2 means the number of images that are labeled with score 2 in the dataset.

### 4 DATA ANALYSIS

This section analysis our proposed ChiQA in three aspects: the question aspect, the image aspect, and the vision and language understanding aspect.

#### 4.1 Question Analysis

**Frequent words:** we plot the most frequent words of our proposed ChiQA dataset in Figure 5, we can see that ChiQA contains various types of words such as ‘的区别’ (the difference) and ‘的位置’, etc.

**Question Classification:** To analyze the domain distribution of the labeled data, we use our in-house question-classification model to classify the questions in ChiQA. The result is shown on the left of Figure 6. We can see that the question in ChiQA spans a wide range of domains, such as healthy, medical, education, etc. The domain distributions are even that none of them is dominant. The diversity of the domain makes the cross-modal learning on ChiQA more challenging since it should be more generalized.

**Interrogative words:** we also analyze the interrogative words of ChiQA. Different from other languages such as English, there are no apparent interrogative words such as ‘who’, ‘when’, ‘what’, etc. Therefore, we randomly annotated some questions to extract their keywords such as ‘如何’ (‘how to’), ‘什么’ (‘what’), etc. Then we map them to the interrogative words. For example, we map ‘的原因’ (‘the reason of’) to the interrogative words ‘为什么’ (‘why’). The question analysis of ChiQA is shown in Figure 6.

#### 4.2 Image Analysis

We use images from the open-domain web that is very various, to analyze the characteristics of the images, we adopt an off-the-shelf
where the corresponding questions are valid, the output is the DETR \[11\], a Transformer-based end-to-end object detection model that achieves state-of-the-art results on object detection, as our object extractor. The DETR is applied to the image to facilitate a better understanding of the image in ChiQA. Besides, we can see that a large number of examples require higher-level reasoning skills such as causal inference and logical reasoning, which are barely mentioned in previous VQA datasets. The prerequisite of high-level reasoning skills makes ChiQA challenging enough.

### 4.3 Reasoning Skills

Previous analysis of VQA corpus mostly focused on vision-based understanding, such as object relation \[18\], composition \[57\], etc. However, we believe knowing the reasoning skills requirement of VQA datasets is also important for the development of vision and language understanding systems. Inspired by the previous works of machine reading comprehension analysis \[54, 56\], we also identified 5 prerequisite reasoning skills for ChiQA\(^4\).

1. **Grounding**: the ability to locate and discriminate the target object mentioned in the question.
2. **Reading**: the ability to recognize the text in the images with arbitrary fonts and layout.
3. **Comparison**: is the knowledge to compare different objects with specific attributes, such as height, size, etc.
4. **Casual Inference**: \[56\] is the knowledge about causality between the cause and effect.
5. **Logical Knowledge**: understanding the predicate logic such as negation, conditionals, etc.

Based on these prerequisite skills, we randomly sample 200 samples from the test set of ChiQA and annotate them, the statistics and examples are shown in Figure 7. We can see that except for grounding, more than 80% of the examples require reading. This is in contrast with previous VQA datasets that mainly focus on non-textual objects. The reading skills and grounding skills are combined to facilitate a better understanding of the image in ChiQA. Besides, we can see that a large number of examples require higher-level reasoning skills such as causal inference and logical reasoning, which are barely mentioned in previous VQA datasets. The prerequisite of high-level reasoning skills makes ChiQA challenging enough.

### 5 EXPERIMENTS

#### 5.1 Evaluation Metrics

In ChiQA, the labels are three-level scores, so we adopt two types of metrics that are widely used:

- **NDCG@N**: When graded relevance judgments are available for a query, for instance in our ChiQA the three-point scale between

\[^4\] As indicated by Sugawara and Aizawa \[54\], this categorization is somewhat empirical and provisional without very solid theoretical background.

![Figure 7: Prerequisite reasoning skills for ChiQA. From top to the bottom are the question, the image, the reasoning skill, the explanation, and the proportion. Note that the reasoning skills are not mutually exclusive so some examples may require more than one reasoning skill.](image)

![Figure 8: Distribution of the objects in our corpus. The objects are extracted via DETR and mapped to MS-COCO object categories. We remove person object since it is too large for better illustration.](image)
zero to two. A popular incarnation of this metric is Normalized discounted cumulative gain. We report the NDCG scores at truncation levels 1, 3, and 5.

**Mean Average Precision (MAP@N)** is another widely used metric in ranking results, it incorporates the trade-off between precision and recall.

**Accuracy/Precision/Recall/F1**: Besides the ranking based criteria, we also incorporate the classification based metric, which measures the absolute score instead of the relative score. We consider two types of granularity: **Binary**, which takes 2 as positive and others (0,1) as negative. **Ternary**, a vanilla classification metric taking all three-level scores into account. Note that we use Macro-F1 in the ternary setting. For the binary result of precision and recall, we tune the threshold to obtain the best result.

### 5.2 Models

We evaluate a wide range of state-of-the-art vision and language understanding models on ChiQA. Inspired by previous works of image-text matching, we first represent the text and image by a text encoder and an image encoder, then their representations are merged to obtain a cross-modal representation. Finally, a prediction layer is applied on top of that representation to obtain the score. The encoders we consider are either pre-trained (✓) or not (○):

- **Random**: for each pair of test data, we simply choose a label from 2/1/0 randomly. The random baseline measures the lower bound of performance on ChiQA.
- **Majority**: every test sample is assigned with the score that appears mostly in training data. So every sample gets the same score and the rank-based metrics such as NDCG are meaningless.
- **LSTM-ResNet**: Before the prevailingness of large-scale pre-training, Long short term memory network [24] and residual networks [22] are de-facto backbone for most language understanding and vision understanding systems. We represent the question by a two-layer bi-directional LSTM and the image by a 152-layer Resnet. Then their representations are fed to a cross-modal spatial attention layer proposed by [71], and finally, the mean pooled representations are used for prediction.
- **BERT+ViT**: the text encoder and image encoder are initialized by Transformer-based BERT [13] and ViT [14], where both models are pre-trained on unlabelled data in a self-supervised way. In BERT+ViT, the presentation of text are initialized by a BERT\textsubscript{base}, and image are initialized by ViT-B/16 with weights pre-trained on ImageNet-21k. Then their representations are fed to the Transformer based self-attention layers. Finally, the representations of \textsubscript{CLS} token in BERT are used for prediction.
- **ALBEF**: is a large-scale vision and language representation learning has shown promising improvements on various vision-language tasks [38]. ALBEF also utilizes the ViT and BERT to encode the text and images, however, their representations are aligned by a contrastive learning objective before they are fused for cross-modal learning. In addition, the text encoder and multimodal encoder are both based on BERT model, making the architecture more efficient.
- **Pretrain**: In the raw model of BERT+ViT and ALBEF, only the text encoder and image encoder are pre-trained because the available cross-modal pre-training models are all based on English. For cross-modal pretraining on Chinese, we adopt Wukong [21], a large-scale Chinese pre-training dataset containing more than 100 million text-image pairs, and further pre-train BERT+ViT and ALBEF by a contrastive-learning method introduced in ALBEF [38].
- **Wenlan**: is a large-scale multimodal pre-training model that was trained on more than 30 million text-images pairs and compact nearly one billion parameters. It outperforms both UNITER [12] and OpenAI CLIP on various downstream tasks. Since the model are not publicly available so we adopt their public api3 to obtain the representations of the question and the image, and use the cosine similarity as the output.
- **Human**: In addition to these models, we also ask two volunteers to obtain the human performance of ChiQA. Since the test set is very large, we only sample 100 of them to the volunteers to get an estimated result.

### 5.3 Common Setup and Objective

The hidden size of the LSTM is set to 512. For all image encoders, we resize the images with resolution 224 × 224. The input question is truncated with the max length of 64. The batch size was set to 256. We use the Adam [34] optimizer with 400 warm-up steps and linearly decay the learning rate from 2 × 10^{-5}, \beta_1, \beta_2, \epsilon was set to 0.9, 0.999 and 10^{-8}, respectively. Since the public available pretrained ALBEF models are based in English, we translate the questions in ChiQA into English to obtain the off-the-shelf result. We randomly select 2000 samples from the training set for development. We implement BERT\textsubscript{base} and ViT-B/16 by Huggingface Transformers [65]. All experiments are conducted through 8 A100 GPUs.

**Objective**: For all model with fine-tuning, the output is a scalar value and we normalize the output with sigmoid function. We also normalize the label 2, 1, 0 to 1, 0.5, 0, respectively, the final objective for fine-tuning is a pointwise binary cross-entropy loss:

\[
\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot \log s_i + (1 - y_i) \cdot \log (1 - s_i) \quad (1)
\]

where \(y_i\) is the normalized labels and \(s_i\) is the normalized prediction. \(N\) is the size of data in a batch.

### 5.4 Main Results

The result is shown in Table 4. We can see that directly applying the previous state-of-the-art cross-modal methods yield a poor performance, only achieving a negligible improvement over the random baselines. This implies that ChiQA is very hard and the models only involve the large-scale weakly supervised contrastive learning (ALBEF, Wenlan) may not able to discriminate the fine-grained information required for visual question answering. In addition, the poor performance of these models demonstrates that ChiQA is very different from the weakly supervised image-caption pairs. Since the image caption only captures the relevance, the ChiQA focuses on answerability.

Furthermore, we observe that the additional cross-modal pretraining could bring some improvement but the margin is minor compared to the mono-modal pretraining. Therefore, the contrastive learning based cross-modal pretraining in ChiQA may not be as useful as they have achieved on other tasks, a better pretraining strategy should be explored. Finally, the model fine-tuned on

3https://github.com/chuhaojin/WenLan-api-document
ChiQA: A Large Scale Image-based Real-World Question Answering Dataset for Multi-Modal Understanding

| NDCG@1 | MAP@1 | T-ACC | B-ACC | B-F1 |
|--------|--------|--------|--------|------|
| 56.90  | 57.39  | 57.48  | 45.91  | 47.56 |

Table 4: Main result of different models on ChiQA. We report Macro-F1 in ternary classification.

| NDCG@1 | MAP@1 | T-ACC | T-F1 | B-ACC | B-F1 |
|--------|--------|--------|------|--------|------|
| 64.90  | 27.73  | 65.98  | 60.05 | 84.97  | 53.14 |

Table 6: Experimental results of employing different external knowledges. DETR is the state-of-the-art objective detection models, and OCR is the optical character recognition module to extract the text in the image.

5.5 Ablation Study

5.5.1 Listwise vs. Pointwise. All the models mentioned before are optimized by the pointwise objective, however, previous works on learning to rank demonstrate that the listwise ranking objective has superior performance [2]. Therefore, we replace the pointwise objective in equation 1 with ListNet, which is a listwise loss introduced in [10], to enable the cross-image interaction, we concatenate all image features to a single sequence features and then fed them to the cross-modal encoder. The result is shown in Table 5. We can see that the performance of listwise loss is not comparable with the pointwise loss, especially for the absolute metrics such as accuracy and F1. Our findings suggest that ChiQA is more than just an image ranking problem, the system should predict the invariant answerability score to achieve good performance.

5.5.2 External Knowledge. Some previous works on cross-modal learning indicate that incorporating the external knowledge, such as objective detection [39], could improve the downstream task a lot. We also consider two types of external knowledge for ChiQA:  
- **Object Dection**: most of the previous works on cross-modal learning resort to object detection to extract region-based image features and build the multimodal encoder thereon. We use DETR [11], a state-of-the-art object detector to predict boxes around both stuff and things classes on MS-COCO, then the region vectors extracted by DETR are treated as the input to the multimodal encoder.  
- **Optical Character Recognition (OCR)**: as we have shown in Sec. 4.3, more than 80% of the questions in ChiQA require reading text in the image. Therefore, we use our in-house Chinese OCR system to extract the embedded text in the images, and append the extracted text to the question with a special SEP token, the concatenated texts are fed to the text encoder, then their mixed representations interact with the images to get the final score.

The result of employing external knowledge is shown in Table 6, we can see that different types of the resources have different influences: When we only take the object representations to the model, the result even drops, therefore, the off-the-shelf object detection model may lose some information that is important for ChiQA. On the other hand, the enhancement of text information introduced by OCR could boost the performance a lot, so the reading ability plays a very important role in VQA, the attention to the textual content, in combination with other content in the images should be further explored.

6 Conclusion

This paper introduces the ChiQA, a large-scale image-independent visual question answering dataset. In ChiQA, given the real-world user questions and related images, the task is to predict whether the image could answer the question. We adopt a very rigorous data inspection process to guarantee quality. In addition, we adopt an active learning based data selection process to improve the diversity and balance of the data. The analysis demonstrates ChiQA is very challenging and requires a lot of high-level reasoning skills such as reading and comparison. Experimental results of many state-of-the-art language and vision understanding models demonstrate there is still a large gap between model and human performance, suggesting that there is still room for improvement in ChiQA.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don’t just assume; look and answer: Overcoming priors for visual question
