

WildQA: In-the-Wild Video Question Answering

Santiago Castro* Naihao Deng* Pingxuan Huang* Mihai Burzo Rada Mihalcea
University of Michigan – Ann Arbor, USA
{sacastro, dnaihao, pxuanh, mburzo, mihalcea}@umich.edu

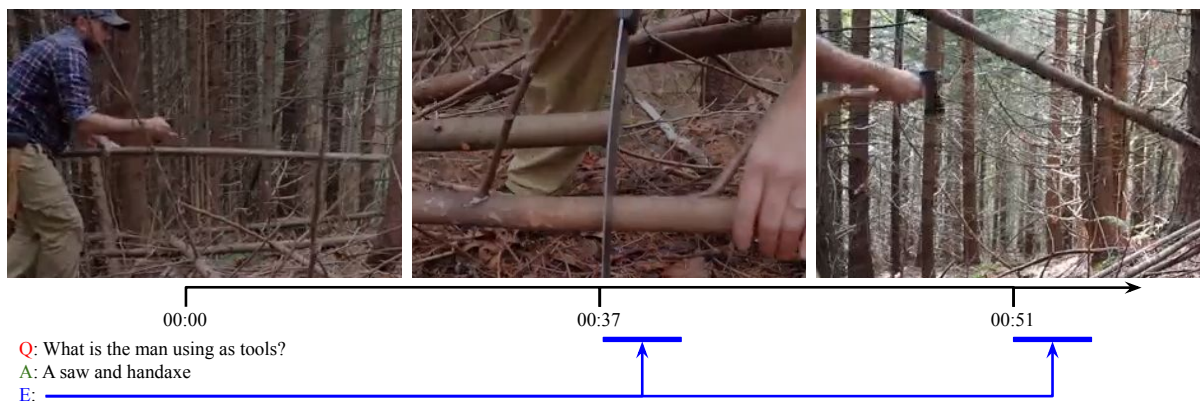


Figure 1: An example from our WildQA dataset, showing a question (Q), an answer (A), and evidence (E) that supports the answer. The corresponding part of the videos is provided as evidence for the question.

Abstract

Existing video understanding datasets mostly focus on human interactions, with little attention being paid to the "in the wild" settings, where the videos are recorded outdoors. We propose **WILDQA**, a video understanding dataset of videos recorded in outside settings. In addition to video question answering (Video QA), we also introduce the new task of identifying visual support for a given question and answer (Video Evidence Selection). Through evaluations using a wide range of baseline models, we show that WILDQA poses new challenges to the vision and language research communities. The dataset is available at <https://lit.eecs.umich.edu/wildqa/>.

1 Introduction

Video understanding plays an important role in the development of intelligent AI systems, as it enables the effective processing of different modalities of information (Li et al., 2021a). Various tasks have been proposed to examine the ability of models' to understand videos, including video question answering (Video QA), video captioning, and fill-in-the-blank tasks (Xu et al., 2017; Tran et al., 2016;

Castro et al., 2022). Recent years have witnessed significant progress in video understanding, including new benchmarks (Tapaswi et al., 2016; Grauman et al., 2021) as well as advanced sophisticated benchmarkmodels (Jin et al., 2019; Radford et al., 2021).

There are however several drawbacks associated with existing video understanding research. First, existing video understanding benchmarks focus on common human activities as typically appearing in cooking videos (Zhu et al., 2017) or in movies (Tapaswi et al., 2016), leading to a limited set of video domains. Second, most video understanding benchmarks adopt a multiple-choice format, where models select an answer from a set of candidates (Jang et al., 2017; Castro et al., 2020). Models trained under such a setting cannot be used in real-life applications because candidate answers are not provided (Castro et al., 2022). Third, videos included in existing benchmarks are typically short (Kim et al., 2016), and the performance of models on longer videos is not well studied.

We address these challenges in our dataset construction process. First, we propose the WILDQA dataset in which we collect "in the wild" videos that are recorded in the outside world, going beyond daily human activities. Figure 2 shows the differ-

*: Equal contribution

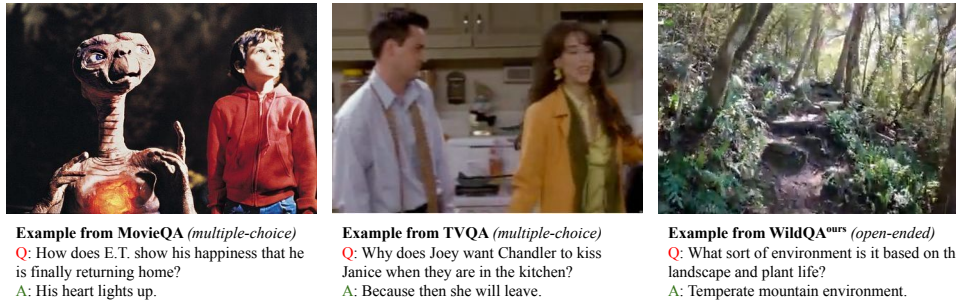


Figure 2: Examples from MovieQA (Tapaswi et al., 2016), TVQA (Lei et al., 2018), and our WildQA dataset. The previous datasets mostly focus on human interactions in a multiple-choice setting, while ours focus on scenes recorded in the outside world in an open-ended setting. We only list a single answer here for illustration purposes.

ence between the WILDQA dataset and previous question answering datasets. Second, we adopt the challenging answer generation approach, aiming to build a system that can answer questions with an open-ended answer, rather than selecting from a predefined set of candidate answers. Third, the average video length in our dataset is one minute, longer than the video clips in most of the existing datasets in Table 3, which presents a novel challenge for video understanding algorithms.

Using the WILDQA dataset, we address two main tasks. First, we address the task of video question answering (**Video QA**), aiming to generate open-ended answers. Second, we introduce the task of retrieving visual support for a given question and answer (**Video Evidence Selection**). Finding the relevant frames in a video for a given question-answer pair can help a system in its reasoning process, and is in line with ongoing efforts to build interpretable models (Jacovi and Goldberg, 2020). For each of these two tasks, we evaluate several baseline models, including multi-task models that combine the two tasks together. Figure 1 shows an example from our dataset, including an example of a question, answer, and supporting video evidence.

To summarize, the main contributions of this paper are:

1. We propose WILDQA, a multimodal video understanding dataset where video scenes are recorded in the outside world.
2. We propose two tasks for WILDQA: Video QA and Video Evidence Selection, aiming to build more interpretable systems.
3. We test several baseline models; experimental results show that our dataset poses new challenges to the vision and language research communities.

2 Related Work

Multimodal Question Answering. Two popular and representative tasks are Visual Question Answering (Visual QA) on images, and Video Question Answering (Video QA) on videos. Visual QA has attracted attention for a long time (Malinowski and Fritz, 2014; Zhang et al., 2016; Ren et al., 2015; Zhu et al., 2016). Recently, much progress has been made in Video QA. Researchers proposed various datasets such as TVQA that contain videos from movies or TV series (Tapaswi et al., 2016; Lei et al., 2018, 2020a) or videos from the Internet spanning from YouTube videos to Tumblr GIFs (Zeng et al., 2017; Ye et al., 2017; Jang et al., 2017; Yu et al., 2019b). Other datasets such as MSVD-QA (Xu et al., 2017) contain videos from the existing corpus (Chen and Dolan, 2011) or cartoon videos (Kim et al., 2016). Recent Video QA datasets have stronger focuses such as temporal relations (Mun et al., 2017), multi-step and non-factoid answers (Colas et al., 2020), natural interactions (Zadeh et al., 2019), characters in the video (Choi et al., 2021), question answering in real life (Castro et al., 2020), incorporating external knowledge (Garcia et al., 2020), and videos recorded from the egocentric view (Fan, 2019; Grauman et al., 2021). To the best of our knowledge, we are the first to collect videos from the outside world.

Researchers have also developed various methods to handle the Video QA task, including joint reasoning of the spatial and temporal structure of a video (Zhao et al., 2017; Gao et al., 2019; Huang et al., 2020; Jiang et al., 2020), integrating memory to keep track of past and future frames (Kim et al., 2017; Gao et al., 2018a; Zhao et al., 2018; Fan et al., 2019; Yu et al., 2020), various attention mechanisms (Zhu et al., 2017; Zhang et al., 2019;

Li et al., 2019; Yu et al., 2019a; Kim et al., 2018; Jin et al., 2019), and others. Recently, pre-trained models have proved to be useful in various visual and language tasks (Radford et al., 2021; Chen et al., 2020; Zellers et al., 2021). However, the pre-trained visual and language models are typically encoder-only and cannot generate an answer in natural language on their own. Thus, such pre-trained encoder-only models do not fit into the open-end video question answering setting in our task.

Additionally, previous work has also investigated various reasoning tasks in a multimodal setting (Gao et al., 2016; Yang et al., 2018; Gao et al., 2018b; Zellers et al., 2019). Although it is not our focus, some questions in our dataset require a certain level of reasoning ability. Moreover, since our dataset is created by domain experts, there is domain knowledge involved in the questions as well.

Moment Retrieval. Moment Retrieval is the task of retrieving a short moment from a large video corpus given a natural language query (Escorcia et al., 2019; Lei et al., 2020b). Researchers have proposed or adapted various datasets for this task (Krishna et al., 2017; Hendricks et al., 2017; Gao et al., 2017; Lei et al., 2020b). The task of retrieving relevant parts in the video given the question (Video Evidence Selection) in our proposed dataset is akin to Moment Retrieval. However, moment retrieval focuses on retrieving the part of videos that the question describes, while Video Evidence Selection is to find parts of videos that can support the answer to the questions as shown in Figure 1. Prior work such as Tutorial-VQA (Colas et al., 2020) also adopt the setting of providing parts of the videos as answers to the question, but they did not include any text answers in their dataset.

Few-shot Learning. Recently, there is a trend to evaluate neural models in a few-shot learning setting (Huang et al., 2018; Mukherjee and Awadallah, 2020; Sun et al., 2020; Li et al., 2021b; Lee et al., 2021; Pfeiffer et al., 2022), where the model is tuned with a small portion of the data and tested against the rest. We adopt the few-shot learning setting for our dataset for both Video QA and Video Evidence Selection.

3 WildQA Dataset

Video Selection and Processing. Following Zadeh et al. (2019); Castro et al. (2020), we

start by collecting videos from YouTube. First, we identify five domains that primarily consist of outdoor scenes and are representative for the outside world, namely, *Agriculture*, *Geography*, *Human Survival*, *Natural Disasters*, and *Military*. We then manually collected videos from relevant YouTube channels for each domain.

Because the raw videos can be as long as an hour, we split the raw videos into short clips using PySceneDetect,¹ and concatenate these short clips so that the output video is approximately one minute. We use the output videos for the annotation process described below. More details for the video selection and video processing steps are discussed in Appendix A.1.

Question, Answer, and Evidence Annotation.

There are two phases in our annotation process, as shown in Figure 3. In **Phase 1**, annotators watch the video clips and come up with a hypothetical motivation. They ask one or more **questions** and provide an **answer** to each of the questions they ask. Annotators are also instructed to provide all the relevant parts in videos as pieces of **evidence** to support the answer to their question. After this step in the data collection, three of the authors of this paper manually review all the question-answer pairs for quality purposes. Next, in **Phase 2**, we collect more **answers** and **evidences** for each question from Phase 1. Over the entire annotation process, annotators spent a total of **556.81 annotation hours**, split into 77.05 hours in Phase 1 and 479.76 in Phase 2. Appendices A.2, A.3, and A.5 present the annotation instructions, annotation interfaces, and reviewing process for question-answer pairs, respectively.

Because we want to collect questions that domain experts are interested, as opposed to arbitrary questions, domain experts carry out the Phase 1 annotations. To demonstrate the quality difference of questions collected from domain experts versus non-experts, we conduct a pilot study. Appendices A.4 and A.6 discuss the pilot study and the annotators’ expertise, respectively.

Dataset Statistics. Tables 1 and 2 present statistics of the videos and associated questions for each of the five domains, along with other relevant statistics. Figure 4 shows the distribution of question types. Appendix A.7 discusses more statistics.

¹PySceneDetect uses the OpenCV (Bradski, 2000) to find scene changes in video clips (py.scsencedetect.com).

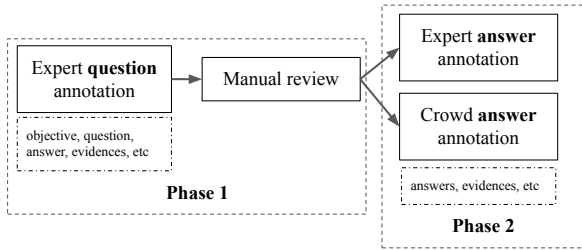


Figure 3: The two phases of data annotation.

Domain	video count	question count
<i>Agriculture</i>	85	109
<i>Human Survival</i>	95	309
<i>Natural Disaster</i>	70	187
<i>Geography</i>	46	110
<i>Military</i>	73	201
Total	369	916

Table 1: Video and question count for each domain.

Videos	369
Duration (in seconds)	71.22 ± 26.47
Questions	916
Question per video	2.48 ± 1.38
Question length (#tokens)	7.09 ± 2.60
Answer per question	2.22 ± 0.69
Answer length (#tokens)	9.08 ± 8.15
Evidence per answer	1.18 ± 0.80
Evidence length (s)	9.64 ± 10.96

Table 2: Dataset statistics for WildQA.

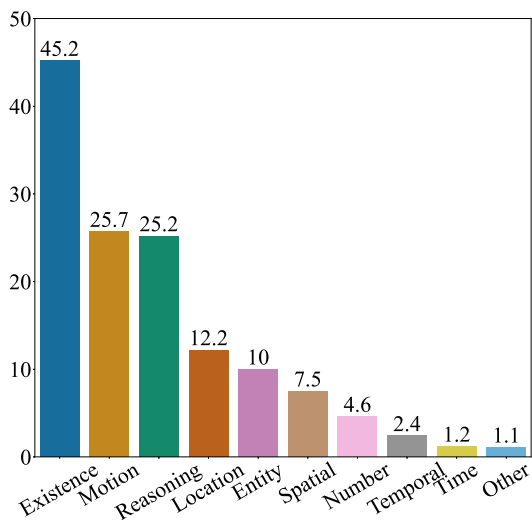


Figure 4: Percentage distribution of question types. Because one question might be classified into multiple categories, the scale summation is larger than 100%.



Q: What type of weather is happening?

A: Flooding and rain.

The weather is rain and flood.



Q: Where is the road at?

A: It is in a tundra environment

The road zig-zags across the landscape.

The road winds through a mountainous landscape.

The road is in an elevated area.

Figure 5: Examples of questions (Q) and answers (A) from WildQA. The first answer is collected during Phase 1 of the annotation process; all remaining answers are collected in Phase 2. More analyses in Appendix A.7.

Dataset Comparison. Table 3 shows the comparison between WILDQA and other existing datasets.

4 Video Question Answering

Following Xue et al. (2017), we adopt **free-form open-ended** video question answering for our video question answering (Video QA) task. Given a question q and a video v , the task is to generate an answer a in natural language.

We adopt a **few-shot learning** setting on our dataset, where models are fine-tuned on question-answer pairs corresponding to 30% of the videos for each domain. The tuned models are tested on data for the remaining 70% videos. The reason is that the time to annotate 30% of the data is around 23 hours, during which there are around 50 data points annotated for each domain, which is acceptable. We hypothesize that it is realistic to have such a setting because the potential end-users could spend around a day or two collecting data, and we can then quickly tune a model using it. Moreover, no repeated videos appear in different splits, following Lei et al. (2018). We end up having 264 question-answer pairs for 108 videos in our dev set and 652 pairs for 261 videos in the test set. We adopt BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) as the metrics to measure the quality of the generated answer. We run each model 3 times and report the scores of mean \pm

Dataset	Domain	VE?	#Videos	# Q	Avg dur. (s)	Annotation	QA Task
MovieQA (Tapaswi et al., 2016)	Movies	✓	6.7K	6.4K	203	Manual	MC
VideoQA (FiB) (Zhu et al., 2017)	Cooking, movies, web		109K	390K	33	Automatic	MC
MSRVTT-QA (Xu et al., 2017)	General life videos		10K	243K	15	Automatic	OE
MovieFiB (Maharaj et al., 2017)	Movies		128K	348K	5	Automatic	OE
TVQA (Lei et al., 2018)	TV shows	✓	21.8K	152K	76	Manual	MC
ActivityNet-QA (Yu et al., 2019b)	Human activity		5.8K	58K	180	Manual	OE
TVQA+ (Lei et al., 2020a)	TV shows	✓	4.2K	29.4K	60	Manual	MC, ES
KnowIT VQA (Garcia et al., 2020)	TV shows		12K	24K	20	Manual	MC
LifeQA (Castro et al., 2020)	Daily life		275	2.3K	74	Manual	MC
TutorialVQA (Colas et al., 2020)	Instructions	✓	76	6.2K	–	Manual	ES
NExT-QA (Xiao et al., 2021)	Daily life		5.4K	52K	44	Manual	MC, OE
FIBER (Castro et al., 2022)	Human actions		28K	2K	10	Manual	OE
WildQA	In-the-wild	✓	369	916	71.2	Manual	OE, ES

Table 3: Comparison between our WILDQA and other existing datasets. **VE?**: Whether the dataset provides “Video Evidences”?; **MC**: “Multiple Choice” question answering; **OE**: “Open Ended” question answering; **ES**: “Evidence Selection”. We adapt the comparison table from [Zhong et al. \(2022\)](#).

standard deviation in Table 4.

4.1 Baselines

Human Baselines. We report the average BLEU and ROUGE scores by leaving one annotator out in Table 4 (**Human**).

Text-only Models. We implement several baselines that only use the question-answer pairs in the dev set. **Random** randomly chooses answers from the dev set. **Common** always predicts the most common answer in the dev set; **Closest** employs embedding produced by a pretrained roberta-base model ([Liu et al., 2019](#)). In the inference, **Closest** retrieves the answers for the dev set question whose embedding has the highest cosine similarity to the test question. We also fine-tune T5 ([Raffel et al., 2020](#)) using question-answer pairs from the dev set (**T5 T**).

Text + Visual Models. Following [Castro et al. \(2022\)](#), we concatenate the text features with the visual features and input the concatenated features to the T5 model (**T5 T+V**). We extract I3D ([Carreira and Zisserman, 2017](#)) video features and take one feature per second.

Multi-task Learning. Multi-task learning has proved to be successful in various domains ([Collobert and Weston, 2008](#); [Deng et al., 2013](#); [Girshick, 2015](#)). Following [Caruana \(1993\)](#), we train **Multi_{T+V, SE}** which combines T5 T+V and T5 SE (the Video Evidence Selection model described in Section 5) with a shared T5 encoder between the tasks of Video Question Answering and Video Evidence Selection. We also train **Multi_{T+V, IO}**

Model name	ROUGE-1	ROUGE-2	ROUGE-L
Random	5.0 ± 0.2	0.5 ± 0.1	4.9 ± 0.2
Common	10.6 ± 0.0	0.0 ± 0.0	10.6 ± 0.0
Closest	19.5 ± 0.0	6.2 ± 0.0	18.7 ± 0.0
T5 T ^{0-shot}	0.8 ± 0.0	0.0 ± 0.0	0.8 ± 0.0
T5 T	33.8 ± 0.2	17.7 ± 0.1	32.4 ± 0.3
T5 T+V	33.1 ± 0.3	17.3 ± 0.4	31.9 ± 0.2
Multi _{T+V, IO}	34.0 ± 0.5	18.8 ± 0.7	32.8 ± 0.6
Multi _{T+V, SE}	33.8 ± 0.8	18.5 ± 0.7	32.5 ± 0.8
Human	40.8 ± 0.0	18.1 ± 0.0	36.3 ± 0.0

Table 4: ROUGE scores for the task of Video Question Answering. For comparison, we test the out-of-box T5 model under the zero-shot setting (T5 T^{0-shot}).

which combines T5 T+V and T5 IO (another Video Evidence Selection model described in Section 5) in a similar way. The loss function during the fine-tuning is:

$$L = \alpha L_1 + \beta L_2 \quad (1)$$

where L_1, L_2 are the losses for Video Question Answering and Video Evidence Selection, respectively; α, β are the weights for the two tasks. The selection process behind the values of α and β are presented in Appendix C.

4.2 Results

Table 4 reports F1 scores of ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) for our baseline models. For comparison, we also test the out-of-box T5 model on our test split under the zero-shot setting (**T5 Text^{0-shot}** in Table 5).

T5-based models significantly outperform the random baselines as well as the out-of-box T5

model, which suggests that the T5-based models acquire certain levels of question-answering ability in the tuning stage. However, adding visual features does not improve the model’s performance. This might be due to the *challenges of attending to the visual features at the corresponding parts in the video*, because both models under multi-task learning outperform the text-only baseline, suggesting that attending to the correct part of the video helps the answer generation process.

All baseline models underperform human baselines on ROUGE scores, especially on ROUGE-1 and ROUGE-L scores, suggesting that there is room for improvement. However, the ROUGE-2 score for human annotators is low because although human annotators tend to use the same word to describe the object that appears in the video, there are large variations in terms of expressing the ideas of their answers. More discussions on the diversity of the answers are in Appendix A.7.

5 Video Evidence Selection

Similar to Colas et al. (2020), given a video \mathbf{v} and a question \mathbf{q} , the video evidence selection task consists of predicting $\{(s_1, e_1), (s_2, e_2), \dots\}$, where (s_i, e_i) represents the time for start \mathbf{s} and end \mathbf{e} of a single span within the video \mathbf{v} . We also adopt the few-shot learning setting as described in Section 4 for the task of Video Evidence Selection. Similar to DeYoung et al. (2020), we design an Intersection-Over-Union (IOU) metric borrowed from Everingham et al. (2010). We define IOU as follows: given two time spans in the video, IOU is defined as the length of their intersection divided by the length of their union. The prediction is counted as a match if it overlaps with any of the ground truth spans by more than the threshold (0.5, following DeYoung et al., 2020). We use these partial matches to calculate an F1 score (IOU-F1 scores). As described in Section 4, we run each model three times and report the scores of mean \pm standard deviation in Table 5.

5.1 Baselines

As described in Section 4, we compute the average IOU-F1 score on the annotations from one annotator against the remaining annotators; we denote this metric as **Human**. The **Random** baseline consists of randomly choosing the start and end of a part within the original video as evidence. Similar to the structure Devlin et al. (2019) experiment

Model name	IOU-F1
Random	2.5 ± 0.3
T5 IO	1.1 ± 0.2
T5 SE	4.5 ± 0.8
Multi _{T+v, IO}	1.4 ± 0.3
Multi _{T+v, SE}	3.7 ± 2.4
Human	18.37 ± 0.0

Table 5: IOU-F1 scores for Video Evidence Selection.

on SQuAD (Rajpurkar et al., 2016), we build **T5 SE**; here, we feed the concatenated question embeddings and I3D visual features to the T5 encoder, and the T5 encoder outputs a sequence of the encoded states. We treat the subsequence corresponding to the visual features as the encoded hidden sequence $T_m \in R^H$ for the video frames (H denotes the dimension of the hidden sequence). We then multiply the sequence with two vectors $S, E \in R^H$. The T_i and T_j that maximize the likelihood are predicted as the **start and the end of the evidence**, respectively. During the training, we maximize their joint probability:

$$P_i P_j = \frac{e^{S \cdot T_i}}{\sum_m e^{S \cdot T_m}} \frac{e^{E \cdot T_j}}{\sum_m e^{E \cdot T_m}}$$

where P_i and P_j are the probability for the i being the start and j the end of the evidence, respectively.

Inspired by the Inside-Outside-Beginning (“IOB”) tagging scheme (Ramshaw and Marcus, 1995), we also formulate the evidence finding as a task of tagging whether a video frame is inside (“I”) the evidence, or outside (“O”) the evidence. We then build **T5 IO** by feeding the concatenated features to a T5 encoder. Similar to T5 Start End, we have an encoded sequence of $T_m \in R^H$ corresponding to the video frames. We then multiply the sequence with a vector $L \in R^H$ and apply a sigmoid function to the multiplication result. The model predicts the frame as “I” if the value at the corresponding position is greater than or equal to 0.5, otherwise it predicts “O”. We test Multi_{T+v, IO} and Multi_{T+v, SE} described in Section 4 on Video Evidence Selection as well.

5.2 Results

Table 5 shows the performance of the baseline models on the Video Evidence Selection task. All the baseline models perform significantly worse than the human annotators, and sometimes worse than the random baseline. This is understandable be-

Type	R1	IOU-F1
<i>Existence</i>	33.3 ± 0.3	5.3 ± 0.3
<i>Motion</i>	32.8 ± 0.6	3.1 ± 2.0
<i>Reasoning</i>	33.3 ± 0.4	3.1 ± 1.3
<i>Location</i>	26.2 ± 10.7	4.4 ± 1.4
<i>Entity</i>	33.2 ± 0.7	5.2 ± 0.7
<i>Spatial</i>	32.2 ± 0.6	2.4 ± 1.7
<i>Number</i>	33.8 ± 0.4	4.5 ± 0.7
<i>Temporal</i>	33.8 ± 0.6	3.8 ± 0.5
<i>Time</i>	33.1 ± 0.8	5.7 ± 1.0
<i>Other</i>	33.2 ± 0.6	5.3 ± 0.9

Table 6: $\text{Multi}_{T+V,SE}$ performance on different question types for Video QA (ROUGE-1) and for Video Evidence Selection (IOU-F1).

cause selecting evidence from a long video can be difficult. Additionally, multi-task learning makes the model’s performance worse. However, this could be due to the fact that the Video Evidence Selection itself is a hard task, and all the baseline models struggle with such a task. Although multi-task learning does not help Video Evidence Selection, as mentioned in Section 4, training with Video Evidence Selection does help Video QA. Thus, Video Evidence Selection is still an important task to improve a model’s ability to answer questions. We include more ablation studies in Appendix D.1.

6 Analysis and Discussion

Model Performance v.s. Question Types. Table 6 shows $\text{Multi}_{T+V,SE}$ ’s performance on different question types for Video QA and Video Evidence Selection respectively. Other ROUGE scores for Video QA follow similar trends as shown in Figure 14. According to Table 6, the model achieves good ROUGE-1 scores for Video QA when the model has a good IOU-F1 score for Video Evidence Selection such as its performance on *Existence*. The model has the highest ROUGE-1 variation on *Location* question types, with a relatively large variation for IOU-F1. The model’s ROUGE-1 score on *Spatial* questions is relatively low, with the lowest IOU-F1 score. $\text{Multi}_{T+V,SE}$ excels at question type *Entity* and *Existence* with relatively high IOU-F1 scores. One possible explanation could be that the average length of the answers generated for *Entity* and *Existence* are around eight tokens, which might be easier for the model to ground to the relevant part in the video.

Interestingly, even if the answers have similar lengths, the model struggles on *Motion* questions (with a relatively low IOU-F1 score). A possible

Model name	R1	R2	RL
T5 $T^{0\text{-shot}}$	0.8 ± 0.0	0.0 ± 0.0	0.8 ± 0.0
T5 $T_{TVQA}^{0\text{-shot}}$	9.1 ± 0.0	1.2 ± 0.0	8.8 ± 0.0
T5 $T_{TVQA,ours}$	32.4 ± 0.2	17.5 ± 0.2	31.6 ± 0.2
T5 T_{ours}	33.8 ± 0.2	17.7 ± 0.1	32.4 ± 0.3
T5 $T+V_{TVQA}^{0\text{-shot}}$	20.3 ± 0.0	8.1 ± 0.0	20.1 ± 0.0
T5 $T+V_{ours}$	33.1 ± 0.3	17.3 ± 0.4	31.9 ± 0.2
T5 $T+V_{TVQA,ours}$	33.7 ± 0.2	18.3 ± 0.1	32.6 ± 0.1

Table 7: ROUGE scores for the task of Video Question Answering for few-shot learning setting (the standard setting in our WildQA dataset introduced in Section 4) and zero-shot learning setting (“0-shot” in the superscript). Subscript “TVQA” means pre-training on the TVQA (Lei et al., 2018) dataset; subscript “TVQA,ours” means first pre-training the model on TVQA, then tuning the model on our WildQA dataset; subscript “ours” means tuning the model directly on our WildQA dataset.

reason could be that this type of questions provide a very abstract description of the action, which makes the model hard to attend to the relevant part of the video. For instance, an example of a *Motion* question is “Are there any structure or natural features being affected?”. To attend to the corresponding period in the video, the model needs to understand the word “affected” and the objects that are actually affected, which can be very difficult. The model also struggles to attend to the correct places in the video for the *Spatial* type of question. This might be because there is more than one entity in *Spatial* type of questions, and the model needs to locate all the objects appearing in various parts of the video, which is similarly complex. For instance, for the question “What effects did the weather have?”, the model needs to attend to “debris in the air”, “truck turnover” and “destruction of buildings”. For *Location* type of questions such as “What sorts of terrain is the vegetation present in?”, it might be difficult to attend to all the terrains of “forest”, “plateaus”, “mountainous”, “valleys”, and “arboreal” and to include them in the answer.

Domain Adaptation. Furthermore, we tune the $\text{Multi}_{T+V,SE}$ model on the dev set data from a single domain, and test it against data from other domains. Figures 6 and 7 show the model’s performance in different tuning and testing domains. Interestingly, the diagonal cells do not always have the darkest color, which indicates that inter-relations exist across domains. For instance, the model tuned on *Geography* performs relatively better for Video QA on *Human Survival* and *Agri-*

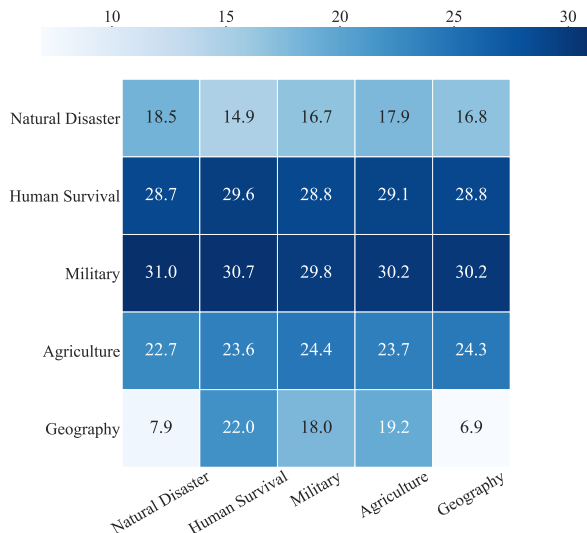


Figure 6: $\text{Multi}_{T+V, SE}$ performance (ROUGE-1) for Video QA when tuned on a single domain (y-axis) and tested against each domain (x-axis). The performances by the rest metrics for Video QA resemble the pattern here and are reported in Appendix D.

culture rather than itself. This suggests that the questions and videos from *Geography*, *Agriculture*, and *Human Survival* exhibit some similarity so that the model tuned on one domain can answer questions from the other domains relatively well. But answering questions from *Geography* can introduce the domain knowledge, an example of the answer is “Mountainous, temperate forest.”, where “temperate forest” is one of the terminologies specific to *Geography* domain. Training on these terminologies might confuse the model and hurt the performance. Thus, future research might be needed to study how to better incorporate domain knowledge into multi-modal question answering.

As for Video Evidence Selection, the patterns generally resemble the pattern in Figure 6, which means that in general, the model answers a question better if it can attend to the relevant part in the video. However, when tuned on *Human Survival* and tested on *Natural Disaster* the model performs relatively well on Video QA (with a 28.7 ROUGE-1 score) but less well on Video Evidence Selection (with a 0.7 IOU-F1 score). This might indicate that the model picks up some common patterns in the text rather than reasoning about the video and the question in an expected manner.

Pre-training on Other Datasets. We also pre-train the $T5_T$ and $T5_{T+V}$ using TVQA (Lei et al., 2018), a large-scale multimodal question an-

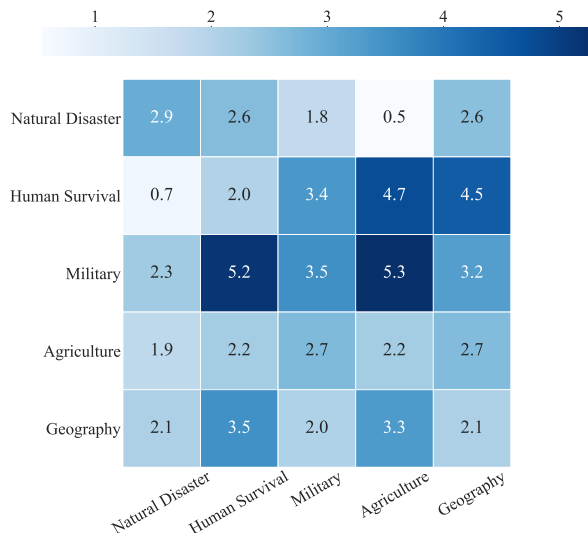


Figure 7: $\text{Multi}_{T+V, SE}$ performance (IOU-F1) for Video Evidence Selection when tuned on a single domain (y-axis) and tested against each domain (x-axis).

swering dataset with videos from TV series. We report the zero-shot learning performances as well as the few-shot learning performances for $T5_T$ and $T5_{T+V}$ in Table 7. We can see that pre-training on TVQA for text-only $T5_T$ does not help, which shows that the question styles in our dataset might be different from TVQA. For $T5_{T+V}$ which uses both text and visual features, pre-training on TVQA does help the model, which suggests that the pre-training helps the model take advantage of the visual features. $T5_{T+V}$ pre-trained on TVQA underperforms $T5_{T+V}$ trained together with $T5_{IO}$ (the $\text{Multi}_{T+V, SE}$ model) according to Table 4 and Table 7, suggesting that attending to the relevant part in the video helps the model better than training the model on more data. However, pre-training the model on the TVQA dataset reduces the variance of model performance, which suggests that training the model with more data helps the model perform consistently.

7 Conclusion

In this paper, we introduced a new and challenging benchmark, WILDQA, to promote domain diversity for video understanding. Specifically, we focused on five domains that involve long videos recorded in the outside world, which can be useful for applications in these domains. Instead of the traditional multiple-choice setting for Video Question Answering, we proposed to generate open-ended answers. We believe open-end answer generation can help construct systems that can answer end

users' questions in a more natural way. To help the model attend to the relevant parts in the videos, we also proposed the task of Video Evidence Selection. Through experiments, we showed the feasibility of these tasks, and also showed that jointly training for both Video Question Answering and Video Evidence Selection can improve the models' performance. In addition, we found it is easier to understand models' behavior by knowing which part of the video the model attends to when answering a question. We believe that this is an important step towards a trustworthy, explainable multimodal system. The dataset is available at <https://lit.eecs.umich.edu/wildqa/>.

Acknowledgement

We thank the anonymous reviewers for their constructive feedbacks. We thank Artem Abzaliev, Do June Min, and Oana Ignat for proofreading and suggestions. We thank William McNamee for the help with the video collection process, and all the annotators for their hard work on data annotation. We thank Yiqun Yao for the helpful discussions during the early stage of the project. This material is based in part upon work supported by the Automotive Research Center ("ARC"). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ARC or any other related entity.

Ethical Consideration

The videos we use are all publicly available on YouTube. The dataset includes a variety of domains, including videos in the *Military* domain, but we are ensuring that all the videos are only used for asking questions specific to the video content. Moreover, in our manual examination, we make sure that the question-answer pairs collected from our annotators are appropriate without any use of offensive language.

References

G. Bradski. 2000. *The OpenCV Library*. *Dr. Dobb's Journal of Software Tools*.

João Carreira and Andrew Zisserman. 2017. *Quo vadis, action recognition? A new model and the kinetics dataset*. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society.

Richard Caruana. 1993. *Multitask learning: A knowledge-based source of inductive bias*. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann.

Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. 2020. *LifeQA: A real-life dataset for video question answering*. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4352–4358, Marseille, France. European Language Resources Association.

Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan Stroud, and Rada Mihalcea. 2022. *FIBER: Fill-in-the-blanks as a challenging video understanding evaluation framework*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2925–2940, Dublin, Ireland. Association for Computational Linguistics.

David Chen and William Dolan. 2011. *Collecting highly parallel data for paraphrase evaluation*. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. *UNITER: Universal image-text representation learning*. In *European conference on computer vision*, pages 104–120. Springer.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. 2021. *DramaQA: Character-centered video story understanding with hierarchical qa*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1166–1174.

Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Zhe Wang, and Doo Soon Kim. 2020. *TutorialVQA: Question answering dataset for tutorial videos*. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5450–5455, Marseille, France. European Language Resources Association.

Ronan Collobert and Jason Weston. 2008. *A unified architecture for natural language processing: deep neural networks with multitask learning*. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM.

Li Deng, Geoffrey Hinton, and Brian Kingsbury. 2013. *New types of deep neural network learning for speech recognition and related applications: An overview*. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8599–8603. IEEE.

- 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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.
- Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. 2019. [Temporal localization of moments in video collections with natural language](#). *ArXiv preprint*, abs/1907.12763.
- Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. [The PASCAL visual object classes \(VOC\) challenge](#). *International journal of computer vision*, 88(2):303–338.
- Chenyou Fan. 2019. [EgoVQA – an egocentric video question answering benchmark dataset](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.
- Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. [Heterogeneous memory enhanced multimodal attention model for video question answering](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1999–2007. Computer Vision Foundation / IEEE.
- Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018a. [Motion-appearance co-memory networks for video question answering](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6576–6585. IEEE Computer Society.
- Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. [TALL: temporal activity localization via language query](#). In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5277–5285. IEEE Computer Society.
- Lianli Gao, Pengpeng Zeng, Jingkuan Song, Yuan-Fang Li, Wu Liu, Tao Mei, and Heng Tao Shen. 2019. [Structured two-stream attention network for video question answering](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6391–6398. AAAI Press.
- Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Yue Chai. 2016. [Physical causality of action verbs in grounded language understanding](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.
- Qiaozi Gao, Shaohua Yang, Joyce Yue Chai, and Lucy Vanderwende. 2018b. [What action causes this? towards naive physical action-effect prediction](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 934–945. Association for Computational Linguistics.
- Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. [KnowIT VQA: answering knowledge-based questions about videos](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10826–10834. AAAI Press.
- Ross B. Girshick. 2015. [Fast R-CNN](#). In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society.
- Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2021. [Ego4D: Around the world in 3,000 hours of egocentric video](#). *ArXiv preprint*, abs/2110.07058.
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. [Localizing moments in video with natural language](#). In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5804–5813. IEEE Computer Society.
- Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Minghui Tan, and Chuang Gan. 2020. [Location-aware graph convolutional networks for video question answering](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11021–11028. AAAI Press.
- Po-Sen Huang, Chenglong Wang, Rishabh Singh, Wentau Yih, and Xiaodong He. 2018. [Natural language to structured query generation via meta-learning](#). In *Proceedings of the 2018 Conference of the North*

- American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 732–738, New Orleans, Louisiana. Association for Computational Linguistics.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.
- Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. [TGIF-QA: toward spatio-temporal reasoning in visual question answering](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1359–1367. IEEE Computer Society.
- Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. 2020. [Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering](#). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11101–11108. AAAI Press.
- Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. 2019. [Multi-interaction network with object relation for video question answering](#). In *Proceedings of the 27th ACM International Conference on Multimedia, MM '19*, page 1193–1201, New York, NY, USA. Association for Computing Machinery.
- K Kim, C Nan, MO Heo, SH Choi, and BT Zhang. 2016. [PororoQA: Cartoon video series dataset for story understanding](#). In *Proceedings of NIPS 2016 Workshop on Large Scale Computer Vision System*, volume 15.
- Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. 2018. [Multimodal dual attention memory for video story question answering](#). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 673–688.
- Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. [DeepStory: Video story QA by deep embedded memory networks](#). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2016–2022. ijcai.org.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. [Dense-captioning events in videos](#). In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 706–715. IEEE Computer Society.
- Chia-Hsuan Lee, Oleksandr Polozov, and Matthew Richardson. 2021. [KaggleDBQA: Realistic evaluation of text-to-SQL parsers](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2261–2273, Online. Association for Computational Linguistics.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. [TVQA: Localized, compositional video question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.
- Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020a. [TVQA+: Spatio-temporal grounding for video question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online. Association for Computational Linguistics.
- Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020b. [TVR: A large-scale dataset for video-subtitle moment retrieval](#). In *European Conference on Computer Vision*, pages 447–463. Springer.
- Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021a. [UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics.
- Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. [Beyond rnns: Positional self-attention with co-attention for video question answering](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8658–8665. AAAI Press.
- Zhuang Li, Lizhen Qu, Shuo Huang, and Gholamreza Haffari. 2021b. [Few-shot semantic parsing for new predicates](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1281–1291, Online. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text summarization branches out*, pages 74–81.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

- RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv preprint*, abs/1907.11692.
- Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron C. Courville, and Christopher Joseph Pal. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7359–7368. IEEE Computer Society.
- Mateusz Malinowski and Mario Fritz. 2014. 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.
- Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. *Advances in Neural Information Processing Systems*, 33:21199–21212.
- Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. 2017. MarioQA: Answering questions by watching gameplay videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2886–2894. IEEE Computer Society.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.
- Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2953–2961.
- Yibo Sun, Duyu Tang, Nan Duan, Yeyun Gong, Xiaocheng Feng, Bing Qin, and Daxin Jiang. 2020. Neural semantic parsing in low-resource settings with back-translation and meta-learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8960–8967.
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding stories in movies through question-answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4631–4640. IEEE Computer Society.
- Du Tran, Maksim Bolonkin, Manohar Paluri, and Lorenzo Torresani. 2016. VideoMCC: a new benchmark for video comprehension. *ArXiv preprint*, abs/1606.07373.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE.
- Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia, MM '17*, page 1645–1653, New York, NY, USA. Association for Computing Machinery.
- Hongyang Xue, Zhou Zhao, and Deng Cai. 2017. Unifying the video and question attentions for open-ended video question answering. *IEEE Transactions on Image Processing*, 26(12):5656–5666.
- Shaohua Yang, Qiaozhi Gao, Sari Saba-Sadiya, and Joyce Yue Chai. 2018. Commonsense justification for action explanation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2627–2637. Association for Computational Linguistics.

- Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. 2017. [Video question answering via attribute-augmented attention network learning](#). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 829–832. ACM.
- Ting Yu, Jun Yu, Zhou Yu, Qingming Huang, and Qi Tian. 2020. [Long-term video question answering via multimodal hierarchical memory attentive networks](#). *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3):931–944.
- Ting Yu, Jun Yu, Zhou Yu, and Dacheng Tao. 2019a. [Compositional attention networks with two-stream fusion for video question answering](#). *IEEE Transactions on Image Processing*, pages 1204–1218.
- Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019b. [ActivityNet-QA: A dataset for understanding complex web videos via question answering](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9127–9134. AAAI Press.
- Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. [SocialIQ: A question answering benchmark for artificial social intelligence](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8807–8817. Computer Vision Foundation / IEEE.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. [MERLOT: Multimodal neural script knowledge models](#). In *Advances in Neural Information Processing Systems 34*.
- Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. [Leveraging video descriptions to learn video question answering](#). In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4334–4340. AAAI Press.
- Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. [Yin and yang: Balancing and answering binary visual questions](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5014–5022. IEEE Computer Society.
- Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, Fei Wu, and Yueting Zhuang. 2019. [Frame augmented alternating attention network for video question answering](#). *IEEE Transactions on Multimedia*, 22(4):1032–1041.
- Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. [Video question answering via hierarchical spatio-temporal attention networks](#). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3518–3524. ijcai.org.
- Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueting Zhuang. 2018. [Open-ended long-form video question answering via adaptive hierarchical reinforced networks](#). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 3683–3689. ijcai.org.
- Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. [Video question answering: Datasets, algorithms and challenges](#). *arXiv preprint arXiv:2203.01225*.
- Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2017. [Uncovering the temporal context for video question answering](#). *International Journal of Computer Vision*, 124(3):409–421.
- Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. [Visual7W: Grounded question answering in images](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4995–5004. IEEE Computer Society.

A Annotation Details

A.1 Video Selection and Processing

Video Selection. For the video selection part, as mentioned in Section 3, first, we identify 5 domains, *Agriculture*, *Geography*, *Human Survival*, *Natural Disasters*, and *Military*, to collect videos recorded in the outside world. We then identify eight (8) YouTube channels and crawl videos from those channels. During crawling, we manually substitute irrelevant videos such as advertisements with videos that contain scenes mostly recorded in the outside world from the same channel.

Video Processing. As mentioned in Section 3, we clip the raw videos into short clips by PySceneDetect because the raw videos can be as long as an hour. We then concatenate these short clips so that the output video will be around 1 minute. **The output videos are used for the following annotation process.** We want to include longer videos because the videos recorded in the outside world usually contain less information compared to the videos about human interactions. Besides, if the concatenated video is at the end of the original video, it is allowed to be shorter than 1 minute. We select the concatenated videos that only contain scenes recorded in the outside world. If none of the concatenated videos satisfies, we manually clip the original videos to get an output video.

A.2 Annotation Instructions

As mentioned in Section 3, we have 2 phases in our annotation process as shown in Figure 3. In Phase 1, annotators come up with a hypothetical motivation, ask questions, and provide the corresponding answers with relevant parts of the video as evidence. Phase 2 is to collect answers and evidence for questions we collect in Phase 1. The following are the instructions for these two phases.

Instructions for Phase 1

We need help for this Video QA task based on video content (including the audio).

In this task, we suppose you can hypothetically send a robot to a place that you want, for many hours, so as to collect information that you need. In this hypothetical scenario, you have an objective that you want the robot to learn about. This robot can chart territory and is able to answer questions based on recorded videos. Therefore,

after it comes back, you can ask questions to help you satisfy your objective, then this robot will provide you with answers, as well as video evidence clips to support the answers.

In this task, to simplify, the provided videos represent places where you could potentially have sent the robot and are much shorter (a few minutes). Given a recorded video, please help us provide one hypothetical objective that makes sense with it, along with questions, answers, and evidence. Specifically, you should pretend to be both the information-seeker and the robot, which means that as the robot, you could watch the recorded video, and you should provide answers and video evidence clips; as the information-seeker, you have an objective, **not** watch the whole video (because of practical reasons), and you can only ask questions and receive answers and video evidence clips as feedback.

1. Basic Instructions

- You will need to propose a hypothetical objective (or topic, intention, motivation), to motivate the questions, that makes sense for the given video.
- You will need to provide as many questions as you need (to satisfy your objective) with regard to the content in the videos and that seek to understand more about the proposed objective.
- You will first watch the video, but when you are providing the objective and questions, please pretend you **haven't** seen it before.
- You will need to provide at least one question for each video. **The more the better.**
- You will need to identify the source of your question (whether it is based on the visual scene or the audio) and classify your question accordingly.
- You will need to provide the correct answer to the question you asked, as supported by the content in the video.
- You will need to provide video evidence (video clip) to support your question and answer.
- If one video doesn't make sense at all, or there's no possible objective for this video that makes sense, please comment at the bottom of this annotation page (and fill in

the mandatory fields for the corresponding video with placeholder values).

2. How To Propose Hypothetical Objective

- For each video, you need to come up with a hypothetical objective (or intention, motivation, topic) that makes sense for this video, and briefly explain it.
- Your questions should all relate to this objective.
- Example 1:
 - Objective: I want to learn about the water in the territory.
 - Question 1: How big is the lake?
 - Question 2: Are there any boats in the lake?
 - Question 3: Where is the river?
 - ...
- Example 2:
 - Objective: people/life movement
 - Question 1: Is there any sign that wildlife has passed this area?
 - Question 2: How much traffic is there on the road?
 - ...

3. How To Ask Your Question

- Your question should relate to your proposed objective.
- For each video, after you finish one question, you could click the Add one more question for this video button to continue to provide another question for this video. On the contrary, if you want to delete one question, you could click the Delete this Question button.
- Ask one question at a time.
 - E.g., "*Are there any people? What are they doing?*" is not appropriate.
- When you provide multiple questions for the same video, make sure these questions are **independently** asked.
 - E.g., "What is growing on pine trees?" and "What is their color?" are not independent.
- The answer should be derived from the video (visual or audio).
 - E.g., "*Why do they run every morning?*" is not a good question.

- Ask from the 3rd person point of view.
 - E.g., "What do *we* have on this farm?"
-> "What do *They* have on this farm?"
- Try to balance the questions such that the answers are not too repetitive (E.g., too many 'yes' answers).
- Ask questions matter-of-factly (as **objectively** as possible). Stick to what you can see or hear from the video.
 - E.g., "*Does it make people feel good here?*" is somehow subjective.
- Don't ask questions about how's the video being recorded, the camera-person or the camera itself. Ask about the content itself. Ignore what the camera-person is doing.
 - E.g., "What's the *cameraman* doing?" / "How fast is the *camera* moving?" are not good questions.

4. How to identify the Question Category

We have some basic categories: **Motion, Spatial Relationship, Temporal Relationship, Reasoning, Number, Entity, Existence, Time, Location, Other.**

If your questions fall into **multiple categories**, please check all categories that apply.

Here are some example questions under each category:

- **Motion:** What is the group of soldiers doing?
- **Spatial Relationship:** What is driving beside the motorcycle?
- **Temporal Relationship:** What happens before the black smoke rises?
- **Reasoning:** What makes changing between targets possible for the missile?
- **Number:** How many fighters are flying?
- **Entity:** What is the target of the bullet?
- **Existence:** Is there a lake by the mountain?
- **Time:** How long can the missile fly?
- **Location:** Where is the tank?
- **Others**

5. How To Provide Correct Answer

- Your answer should be written as **full sentences** (at least one).
 - E.g., "*Left*" -> "*The landspout bends toward the left.*"

- The answer should be derived from the video (visual or audio).
 - E.g., "*These plants are green because they contain chlorophyll.*" is not a good answer.
- Provide answers matter-of-factly (as objectively as possible). Stick to what you can see or hear from the video.
 - E.g., "*beautiful*" is likely not a good word to use within an answer.
 - E.g., "*This takes some bravery to do.*" is somehow subjective.
- Don't answer about how's the video being recorded, the camera-person, or the camera itself. Answer about the content itself. Ignore what the camera-person is doing.
 - E.g., "There are two people, i.e. a running child, and the *cameraman*." is not a good answer.
- When you enter numbers, please enter digits instead of text.
 - "*Seventeen*" -> "*17*"

6. How to provide video evidence

- The video evidence consists of **all** the parts of the video that support the answer to your given question.
- You need to provide at least one video evidence clip (intervals within the video) for each question.
- You need to provide both the **start point and end point** for all the video evidence you identify in the video;
- You could use your mouse or ←/→ key to **click or drag the process bars** of start point and end point. When you click or drag the bar, the above video will change accordingly, so you could locate the points according to the video screen.
- For each video evidence clip, the end point should be **greater than zero**, and the end point should be greater or equal to the start point.
- The video evidence clips (the time gap between the start point and the end point) should be as short as possible.

Instructions for Phase 2

We need help for this Video Question Answering task based on video content (including the audio).

1. Basic Instructions

- You will first watch the video, then answer the questions, each question in turn.
- You will need to provide at least one answer for each question (ignoring differences such as upper/lower case, or the article). **The more answers the better**, but every answer should be correct.
- You will need to identify the source of your answer (whether it is based on the visual scene or the audio).
- For each answer, you will need to provide video evidence (video clip) to support your answers. See below for additional information.
- If one video or question is not available, please comment at the bottom of this annotation page (and fill the mandatory fields for this video/question with placeholder values).
- There are five questions, you need to finish all five questions according to the content in the video (including audio).

2. How To Answer

- Provide one or more answers for each question.
- Each answer should be written as full sentences (at least one).
 - E.g., "Left" -> "The landspout bends toward the left."
- The answer should be derived from the video (visual or audio).
 - E.g., "*These plants are green because they contain chlorophyll.*" is not a good answer.
- Respond matter-of-factly (as objectively as possible). Stick to what you can see or hear from the video.
 - E.g., "*beautiful*" is likely not a good word to use within an answer.
 - E.g., "*This takes some bravery to do.*" is somehow subjective.
- Answer in 3rd person point of view.
 - E.g., "We raise cattle on this farm." -> "They raise cattle on this farm."
- Don't answer about how's the video being recorded, the camera-person, or the camera itself. Answer about the content itself. Ignore what the camera-person is doing.

- E.g., "There are two people, i.e. a running child, and the cameraman." / "The camera is moving fast." are not good answers.
- When you enter numbers, please enter digits instead of text.
 - "Seventeen" -> "17"
- Use your best judgment.

3. How to provide video evidence

- The video evidence consists of all the frame intervals of the video that support the answer to your given question.
- You need to provide at least one video evidence clip (interval within the video) for each question.
- You need to provide both the start point and end point for all the video evidence you identify in the video;
- You can use your mouse or ←/→ key to click or drag the process bars of the start point and end point. When you click or drag the bar, the above video will change accordingly, so you could locate the points according to the video screen.
- For each video evidence clip, the end point should be greater than zero, and the end point should be greater or equal to the start point.
- The video evidence clips (the time gap between the start point and the end point) should only cover the actual evidence and not more (in other words, it should be as short as possible).

A.3 Annotation Interface

Figure 8 shows the annotation interface for Phase 1. Figure 9 shows the annotation interface for Phase 2.

A.4 Pilot Study Comparison between Annotations from Experts v.s. Non-Expert

Before the formal annotation, we compare the non-experts and experts' annotations for both phases. For Phase 1, we randomly selected 45 videos from each domain to be annotated by both the experts and crowdworkers. Following Castro et al. (2022), we set the AWS annotation qualification as HIT approve rate >92%, the number of HITs approved >1000, the location is either Canada or U.S., and the reward as \$6/HIT (around \$9/h).

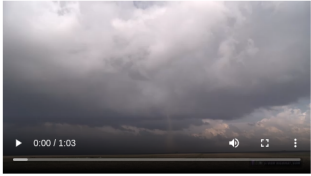
Video Question Answering

Figure 8: Interface for Phase 1 annotation. After watching the video, annotators provide a **motivation**, ask **questions** and provide corresponding **answers** by filling the blank. They provide parts of the videos as **evidence** to support each of the question-answer pairs by dragging the moving bar.

Video Question Answering

► Instructions

Video 1



Please carefully read the [instructions](#) before performing the task. Watch this video while you listen to its [audio](#) (you may need to adjust your device volume). When the video ends, do not click on any other video. Place your mouse on the video, control buttons will then appear.

Question-1:

Which direction does the landspot bend toward?

Question-1 Answer-1

Your Answer:

Please type your answer here

500/500 characters remaining

Answer in 3rd person point of view. Respond matter-of-factly (as objectively as possible). Stick to what you can see or hear from the video. Your answer should be written as full sentences (at least one). Answer about the content itself (not the video making process, the camera-person, or the camera itself)

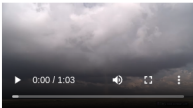
What is your answer based on: (Check all that apply):

Scene Audio

Please suggest how confident you are with your answer:

Very High Confidence High Confidence Moderate Confidence Low Confidence Very Low Confidence

Locate video evidences to support your answer:



The video evidence consists of all the frame intervals of the video that support the answer to your given question. You need to provide at least one video evidence clip (interval within the video) for one question. You need to provide both the **start point** and **end point** for all the video evidences you identify in the video; You can use your mouse or **- / -** key to **click** or **drag** the **process bars** of start point and end point. For each video evidence clip, the end point should be **greater than zero**, and the end point should be **greater or equal** to the start point. The video evidence clips (the time gap between the start point and the end point) should only cover the actual evidence and not more (in other words, it should be as short as possible).

Start point:

End point:

Click here to **Add one more video evidence**

Please finish above fill-in and classification tasks.

click here to **Add one more answer for this question**

Figure 9: Interface for Phase 2 annotation. After watching the video and given the question from Phase 1, annotators provide **answers** with the corresponding **evidence**.

	R	I	P	Overall
expert	2.7	2.5	2.1	2.4
crowd	0.8	0.7	0.5	0.7

Table 8: Average scores of the pilot study for Phase 1 (from 0 to 3). **R**: Relevance; **I**: Interestingness; **P**: Professionalism; **Overall**: Overall Score

After annotation, two authors of this paper who do not know the source of annotation evaluate and score in terms of Relevance, Interestingness, and Professionalism for each annotation from 0 to 3. We define Relevance, Interestingness, and Professionalism as follows:

- **Relevance**: how relevant a question and an answer are to the video. Good relevance indicates that the question is related to the video and focuses on the major events, objects, or people in the video. A relevant answer should address the question and can be derived from this video.
- **Interestingness**: whether the question interests you. In other words, whether you are interested in the question and answer, given a video.
- **Professionalism**: how detailed and precise the question and answer are. Good professionalism can be demonstrated by the precise usage of terminologies and numbers, and accurate description in the answer.
- **Overall Score**: the average score of the score for Relevance, Interestingness, and Professionalism.

For each category, the higher score indicates the better the annotation demonstrates that characteristic. Table 8 lists the scores and Table 9 presents some annotation examples. From both the empirical and numerical results, we could see there is a significant quality gap for the annotation from experts versus from crowdworkers. Therefore, we decide to employ domain experts for Phase 1.

For Phase 2, we randomly select 104 *Geography* videos and questions from the questions annotated in Phase 1 to be annotated by both experts and crowdworkers. Moreover, we set the reward as \$3/HIT (around \$9/h) and employ the AWS **Master**² as the crowdworkers. Table 10 lists the results of the pilot study for Phase 2. According to Table 10, crowdworkers perform similarly to experts

²https://www.mturk.com/worker/help/what_is_master_worker

	Objective	Question	Answer
E	Precipitation	What types of precipitation are occurring?	Rain and hail.
C	Very like	Nice	Nice
E	I want to learn about the people	What type of weapons are they carrying?	M4's
C	The soldiers are caught on the ship.	What they are doing in this video?	They caught the ship.
E	Storm	Where is the storm?	In a field.
C	Motivation	5	Very amazing

Table 9: Examples in pilot study for Phase 1. **E**: Expert; **C**: Crowd

Annotator	R1	R2	RL	IOU-F1
Expert	23.63	8.05	21.22	12.24
Crowd	20.03	3.24	17.69	8.50

Table 10: ROUGE and IOU-F1 scores for the pilot study in Phase 2. Note that the scores here are lower than the scores for the human baselines in Tables 4 and 5. This is because we only compare the collected answers to a single answer here, while in Tables 4 and 5 we calculate the average scores of one annotator against the remaining as described in Section 4.

in Phase 2. Considering the annotation efficiency, we decide to employ both experts and crowdworkers to annotate more diversified answers for each question in Phase 2. Note that the ROUGE scores in Table 10 are lower than the scores for the human baselines in Tables 4 and 5. This is because we only compare the collected answers to a single answer in Table 10, while in Tables 4 and 5, we calculate the average scores of one annotator against the remaining as described in Section 4.

A.5 Question and Answer Correction

After we collect annotation from Phase 1, the authors of this paper check the quality of the collected question and answers and modify the question and answers accordingly. Specifically, we:

- Delete the questions that can be answered without watching the video (e.g. Q: "If water can get through the hut's roof; can the wind go through the hut's roof?", A: "Yes the wind can go through the hut's roof.")
- Modify the question or the answer to 3rd person view (e.g. change Q: "Do we have aircraft that we can do a touch and go landing like a helicopter?" to Q: "Do they have

Annotator ID	Expertise	Assigned Domains (# Q)
0	Geography	Geography (94) ; Natural Disaster (187)
1	Geography	Geography (16) ; Human Survival (74)
2	Veteran	Military (26) ; Human Survival (146)
3	Veteran	Military (70) ; Human Survival (89)
4	Veteran	Military (12)
5	Veteran	Military (8)
6	Veteran	Military (85)
7	Biology	Agriculture (88)
8	Biology	Agriculture (21)

Table 11: Information for expert annotators who annotate the questions, together with their assigned domains and number of questions (# Q) in the parentheses.

aircraft that can do a touch and go landing like a helicopter.")

- Exclude the man holding the camera in the answer if it is a first-person view video.
- Modify questions that are not independently asked (e.g. "Where are they?", where "they" refers to the "paved and unpaved roads" in the previous question. Therefore, we change the question to "Where are the roads?")
- Split questions that include multiple sub-questions into several questions.

Some of the annotators from Phase 2 do not annotate any evidence (leaving the evidence from the start to the end of the video). Thus, we empirically filter out evidence longer than 1/4 of the video.

A.6 Annotator Information

Table 11 shows the expertise of each expert, together with their assigned domains of annotation and the number of questions they annotate in their assigned domains in Phase 1.

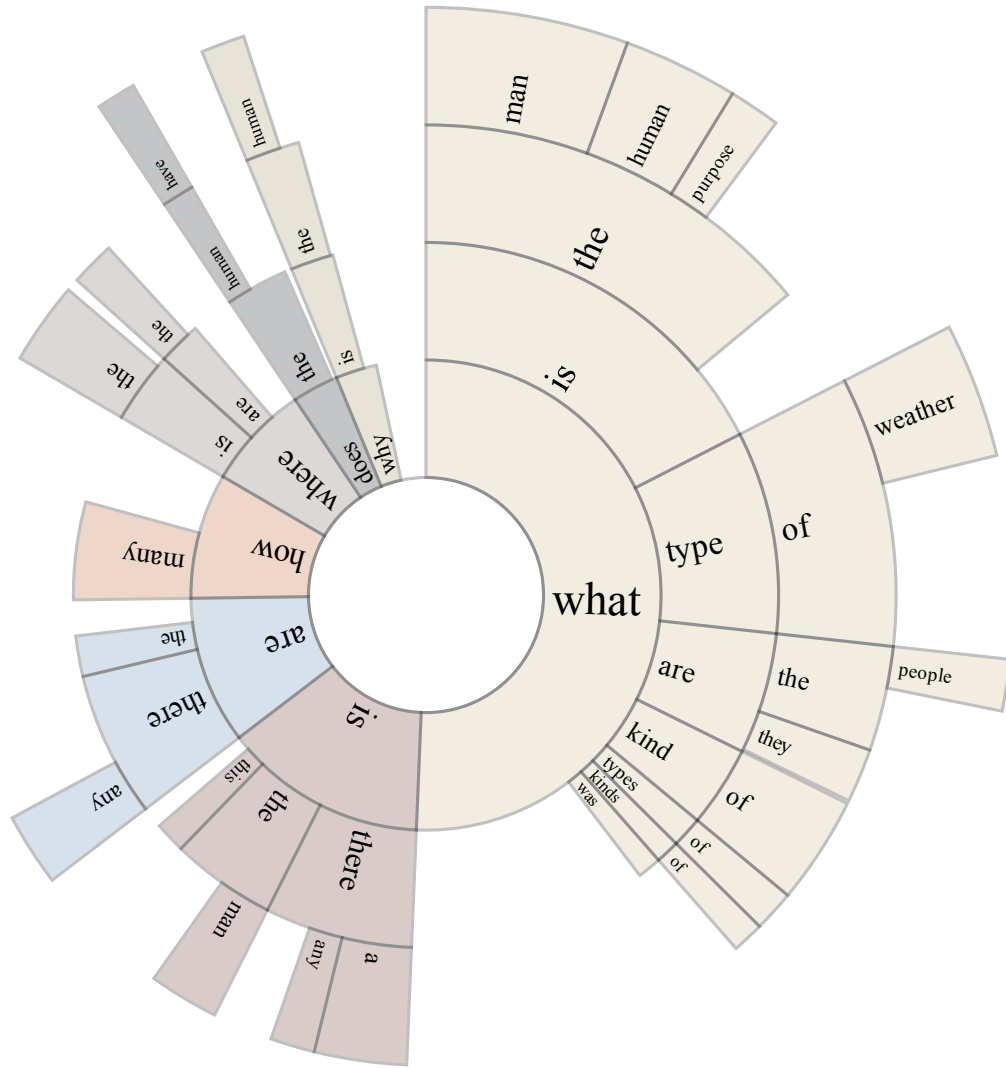


Figure 10: Distribution of questions by the first four tokens. The ordering of words starts from the center to outside.

A.7 Dataset Analysis

Figure 10 presents question distributions in terms of words.

Questions Types. Table 12 examines the frequent words for each domain, which demonstrates the characteristics of the domain. Take *Natural Disaster* as an example, the 3 most frequent words are used in 20.63% of sentences. Besides, Figure 4 in Section 3 lists the annotators’ self-reported question types. One thing we observe is that questions that start with “What” possess a large proportion of all the questions. Such questions might be hard to classify into certain question types (Castro et al., 2020), so we allow annotators

Domain	top1	top2	top3
<i>Agriculture</i>	farm	agricultural	understand
<i>Natural Disaster</i>	weather	people	flooding
<i>Human Survival</i>	man	determine	human
<i>Geography</i>	people	topography	water
<i>Military</i>	military	aircraft	determine

Table 12: Most common 3 words for each domain after removing stop-words.

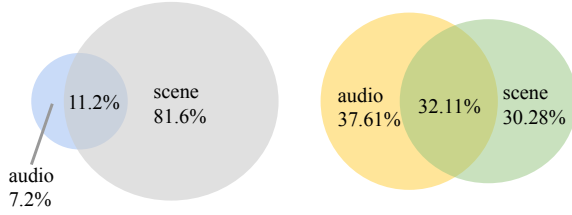


Figure 11: Venn diagrams showing whether the question depends on visual (scene) or audio from the original video. The left is for the entire dataset, while the right is for the *Agriculture* domain.

to choose multiple question types for a single question. Empirically speaking, questions that start with “is(are)”/“where”/“how many” are commonly relevant to “Existence”/“Location”/“Number” questions. In our dataset, their distribution trend (“is(are)”: 24.13% > “where”: 7.21% > “how many”: 4.48%) is akin to the trend of the distribution of the reported question types (“Existence”: 45.20% > “Location”: 12.23% > “Number”: 4.59%). Moreover, although we have “human”, “man” and “people” as the most frequent words in some domains, the most frequent words in domains such as *Military* are “military”, and “aircraft”, which demonstrates that our dataset does not only focus on human interactions as most of the existing datasets do.

Information Needed. As shown in the left Venn figure in Figure 11, generally, most questions are based on the visual (scene). Such a distribution is also justified by the distribution of the question types. The dominant question types we have in Figure 4 are *Motion*, *Spatial*, *Existence* and *Entity*, which typically focus on visual information. However, in *Agriculture* (the right Venn figure in Figure 11), the audio-based questions take more portion, because videos in *Agriculture* usually focus on farming tips, instructions for using tools, etc. In this paper, we do not experiment with models that use audio or transcripts from the video. Future research might look into letting models use audio and transcripts on our dataset.

Answer Similarity/Diversity. We have similar and diversified answers collected in our dataset. Figure 5 gives 2 examples: answers from the upper example are similar to each other; for the lower example, answers diverse a lot between Phase 1 and Phase 2 annotations or even within Phase 2. However, all of the answers are acceptable given the video. The similarity demonstrates the reliability

Videos	369
Duration (s)	71.22 ± 26.47
Questions	916
Question per video	2.48 ± 1.38
Question length (#tokens)	7.09 ± 2.60
Answer length (#tokens)	8.62 ± 8.90
Evidence per answer	1.53 ± 0.76
Evidence length (s)	9.09 ± 13.45

Table 13: Annotation statistics for Phase 1. “#tokens” represent the number of tokens.

Crowd annotated answers	932
Expert annotated answers	182
Total	1114
Answer per question	1.22 ± 0.69
Answer length (#tokens)	9.45 ± 7.46
Evidence per answer	0.89 ± 0.72
Evidence length (s)	10.43 ± 5.81

Table 14: Annotation statistics for Phase 2. “#okens” represents the number of tokens.

of the Phase 2 annotation. Meanwhile, the diversified answers help to better evaluate models.

B Annotation Statistics

Tables 13 and 14 list the statistics for annotation in Phase 1 and Phase 2, respectively.

C Details of Multi-task Learning

Tables 15 and 16 report the model performances under different sets of α, β for Equation (1). We highlight the rows we report in Table 4 in Section 4.2, Table 4 in Section 4.2, Table 5 in Section 5.2, and Table 5 in Section 5.2.

D Experiment Results

Figures 12 and 13 report *Multi-Task* model’s performance on Video QA by ROUGE-2, and ROUGE-L, respectively. Figure 14 demonstrates that ROUGE scores follow a similar trend as mentioned in Section 6.

D.1 Ablation Study on Video Evidence Selection

To investigate whether the vision part is indeed needed by the baseline models for the Video Evidence Selection task, we conduct an ablation study

β	R1	R2	RL	IOU-F1
0.5	33.8 ± 0.8	18.5 ± 0.7	32.5 ± 0.8	3.7 ± 2.4
1.0	32.2 ± 0.7	17.6 ± 0.5	31.0 ± 0.6	1.9 ± 1.7
1.5	33.8 ± 0.3	18.0 ± 0.9	32.5 ± 0.3	1.5 ± 0.1

Table 15: We set $\alpha = 1$ throughout all the experiments, and report the corresponding Multi-T+V,SE performances on Video QA (ROUGE scores) and Video Evidence Selection (IOU-F1 scores). We highlight the row we report in Table 4 in Section 4.2 and Table 4 in Section 4.2.

β	R1	R2	RL	IOU-F1
0.5	34.0 ± 0.5	18.8 ± 0.7	32.8 ± 0.6	1.2 ± 0.1
1.0	33.4 ± 0.6	18.4 ± 0.2	32.1 ± 0.6	1.4 ± 0.3
1.5	32.8 ± 0.3	18.3 ± 0.3	31.7 ± 0.2	1.0 ± 0.2

Table 16: We set $\alpha = 1$ throughout all the experiments, and report the corresponding Multi-T+V,IO performances on Video QA (ROUGE scores) and Video Evidence Selection (IOU-F1 scores). We highlight the row we report in Table 5 in Section 5.2 and Table 5 in Section 5.2.

Model name	IOU-F1
T5 IO _{random}	1.1 ± 0.3
T5 IO	1.1 ± 0.2
T5 SE _{random}	2.7 ± 1.9
T5 SE	4.5 ± 0.8

Table 17: Ablation study on Video Evidence Selection. We feed T5 IO_{random} and T5 SE_{random} the question concatenated with a random sequence, while we feed T5 IO and T5 SE the question with the actual video sequence.

using T5 IO and T5 SE (introduced in Section 5). We take a random sequence of the same length as the original video sequence and feed the random sequence instead of the original video sequence to the model. Table 17 shows the results of the comparison between these different settings. T5 IO performs roughly the same as T5 IO_{random}, which indicates that the model struggles to utilize visual information. T5 IO even underperforms the random baseline which can achieve an IOU-F1 score of 2.5 ± 0.3 (as shown in Table tab:few-shot-evidence-results-10-epochs). However, T5 SE outperforms T5 SE_{random}, suggesting that T5 SE uses visual features to locate the evidence of the question.

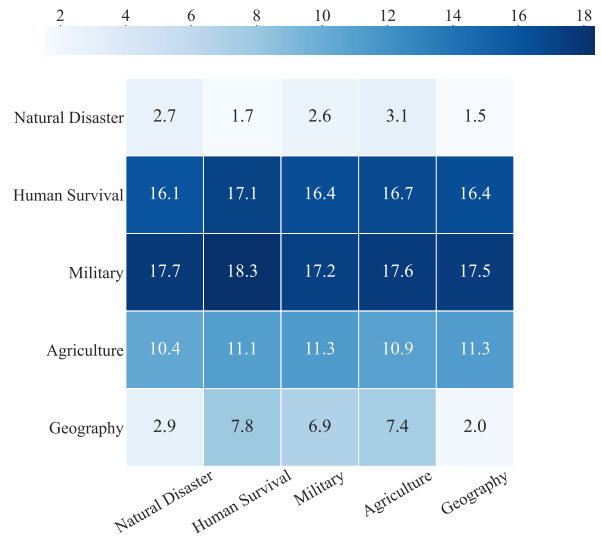


Figure 12: Multi-Task ROUGE-2 scores for Video QA when tuned on a single domain (y-axis) and tested against each domain (x-axis).

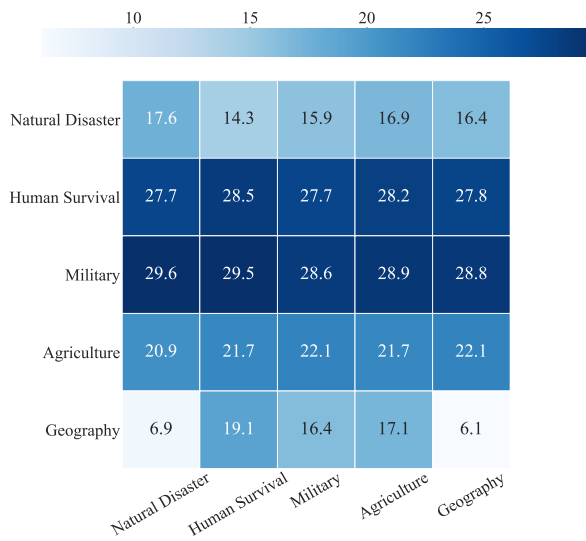


Figure 13: Multi-Task ROUGE-L scores for Video QA when tuned on a single domain (y-axis) and tested against each domain (x-axis).

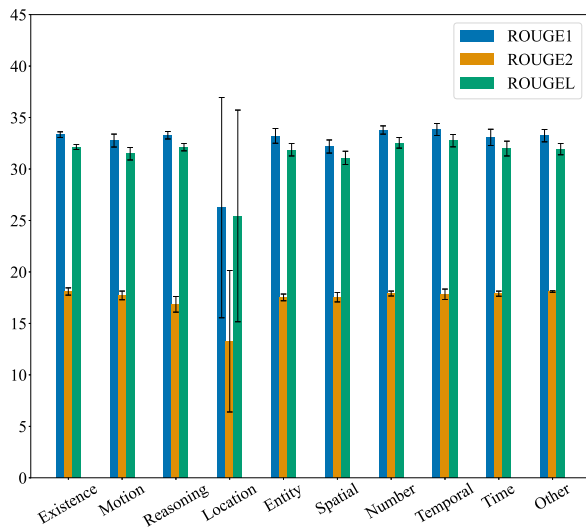


Figure 14: $\text{Multi}_{T+V, SE}$ performance on different question types for Video QA. For each question type, we report ROUGE-1, ROUGE-2, and ROUGE-L scores from left to right. We can see that different ROUGE scores follow similar trends, we only report ROUGE-1 in Table 6 in Section 6.