Towards Video Text Visual Question Answering: Benchmark and Baseline

Supplementary Materials

Minyi Zhao[1]*, **Bingjia Li**[1]*, **Jie Wang**[2], **Wanqing Li**[2], **Wenjing Zhou**[2], **Lan Zhang**[2]
Shijie Xuyang[1], **Zhihang Yu**[1], **Xinkun Yu**[1], **Guangze Li**[1], **Aobotao Dai**[1], **Shuigeng Zhou**[1]†
[1]Shanghai Key Lab of Intelligent Information Processing, and School of
Computer Science, Fudan University, Shanghai 200438, China
[2]ByteDance, China
{zhaomy20, bjli20, abdai20, sgzhou}@fudan.edu.cn
{wangjie.bernard, liwanqing.0415, zhouwenjing.233, zhanglan.11}@bytedance.com
{shijiexuyang, gzlifd}@gmail.com {zhyu21, xkyu21}@m.fudan.edu.cn

1 License and Copyright

As mentioned in our paper, the videos are collected from YouTube. During collection, workers were told to try their best to download only those videos that were available with a Creative Commmons CC-BY (v3.0) License. Since we do not own the copyright for these videos, we provide M4-ViteVQA for non-commercial research purposes only. The detailed license and responsibility agreement are given in the next section.

2 Responsibility Agreement

In order to ensure that researchers can reasonably use the data for research purposes only, all researchers must sign the following agreement when using M4-ViteVQA.

- The researcher shall use the M4-ViteVQA dataset only for non-commercial algorithm research and educational purposes. The researcher can not use the M4-ViteVQA dataset for any other purposes, including but not limited to distribution, commercial usage, etc...
- The researcher takes full responsibility for his or her use of the M4-ViteVQA dataset and shall defend and indemnify the dataset, including their affiliates, employees, trustees, officers and agents, against any and all claims arising from the researcher's use of the M4-ViteVQA dataset.
- The researcher agrees and confirms that authors reserve the right to terminate the researcher's access to the M4-ViteVQA dataset at any time.
- If the researcher is employed by a for-profit business entity, the researcher's employer shall also be bound by these terms and conditions, and the researcher hereby shall represent that he or she is fully authorized to enter into this agreement on behalf of such employer.

3 Accessibility

To access M4-ViteVQA, researchers must sign Sec. 2 to get the download links of M4-ViteVQA. During the review phase, in order to ensure that reviewers can access the dataset anonymously, we temporarily dispense with the signing step and directly provide a download link.

[*] This work was mainly done while the first two authors are interns in ByteDance with equal contribution.
[†] Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

4 Ethical Issues and Potential Negative Societal Impacts

All the workers in this paper are employees of ByteDance and are paid according to local standards. Therefore, there are no ethical issues for M4-ViteVQA.

The possible negative societal impact is the personal information in M4-ViteVQA. We have utilized the internal algorithm from ByteDance to preprocess all the videos to mask this information and pass the check.

5 Maintenance Plan

Zhao, as the first author, is a Ph.D. student focusing on video and OCR research topics at Fudan University since 2020 and will graduate at least after 2025. Zhao will maintain the benchmark [3] at least until 2025.

6 Annotation Instruction

In this section, we give the detailed annotation instructions of M4-ViteVQA. We use ByteDance's internal annotation platform to complete the labeling. Therefore, we cannot provide the annotation interface.

6.1 Annotation Step

In brief, given a video clip, we should write 3 to 7 question-answer (QA) pairs based on the texts and visual information from the video. And the answer must come from the video or 'Yes' or 'No'. If the answer is 'Yes' or 'No', it must be answered via reasoning texts in the video.

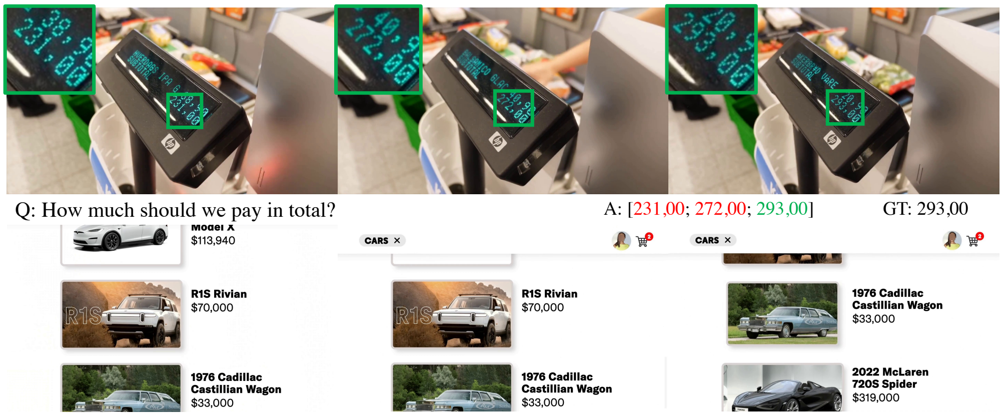
There are four items you should label for each QA pair.

- Question: An interrogative sentence. That is, a question should start with various types of words including What, How, Where, Is, Does, Do, etc... And ends with '?'. The type of question should be diverse. Nevertheless, it is hard to raise some types of questions for some video clips. Therefore, there should be at least two different question types (Where, Is, Does, etc...) in one video clip. We recommend raising some meaningful and practical questions, which can benefit its downstream applications. We require at least three QA pairs for each video clip. For some text-rich and meaningful video clips, we recommend you write down up to seven questions. We hope that on average there are five questions for each clip. The question should be a text-related question. That is, the question should be answered by reasoning texts in the video.
- A selection of '{Easy,Hard}: Easy: This question can be answered via one static frame. Hard: This question can only be answered via jointly understanding multiple frames.
- A selection of '{Text, Vision, Knowledge}'. Vision: This question should be answered via the texts and the visual information from the video. Text: This question can be answered by purely understanding the semantics of texts in the video. Knowledge: This question should be answered via some external knowledge from life.
- Answer: Standard answers to questions. There are only two sources of answers, and other sources are not allowed: Texts from video; 'Yes' or 'No'. If the answer is from the video, it may be diverse. In this case, please write down the most detailed and original form (Case sensitive). You can write down up to 2 answers split by ';' (semicolon). If the answer is 'Yes' or 'No', we hope the proportion is half-half. That is, the number of 'Yes' and 'No' should be similar.

6.2 Verification Step

In this phase, we should check the quality of the labeling. The key point of this phase is to check whether the question can be answered and whether the provided answer is correct. That is, for each

[3] https://github.com/bytedance/VTVQA



Q: How much should we pay in total? A: [231,00; 272,00; 293,00] GT: 293,00

Q: Which is the most expensive car in the website? A: [Model X; 1976 Cadillac Castillian Wagon; 2022 McLaren 720S Spider] GT: [2022 McLaren 720S Spider; 720S Spider]

Figure 1: Two examples from shopping category. 'A' indicates the answers returned by TextVQA models and wrong answers are colored in red.

labeled question, we should watch the video, write down your own answer and check whether it is the same as the original one. Besides, please also check whether this answer comes from the video or is 'Yes' or 'No'. If the QA pair is unqualified (the answer does not come from the video), please delete it.

7 Introduction of nine categories in M4-ViteVQA

As mentioned in our paper, M4-ViteVQA has 9 categories. These 9 categories cover scene texts recorded from daily life (*i.e.,* shopping and driving) and embedded texts in online media (*i.e.,* game and movie). It also contains various video themes and scenes. Here, we introduce each category to highlight the diversity of our dataset.

7.1 Shopping

This category focuses on understanding the events of shopping (both offline and online). The main questions include asking about prices, products, product features, etc... Both online shopping and offline shopping are included in this category. Besides, wide shopping venues are selected in M4-ViteVQA, including supermarket, shopping mall and etc... Two examples are given in Fig. 1.

7.2 Traveling

The traveling category records some street view videos and descriptions of natural scenes. Two representative cases are given in Fig. 2.

7.3 Driving

Driving category videos are mainly captured by vehicles. As shown in Fig. 3, this category requires the model to answer information about landmarks, road signs, etc... This category is suffering from motion blur and low resolution issues, which bring huge challenges to ViteVQA models. We hope this category will spur research in related fields such as auto-driving.

7.4 Vlog

This category includes filmed videos which recorded people's daily lives and street scenes. The photographer's personal private information has been processed. One example is given in Fig. 4.

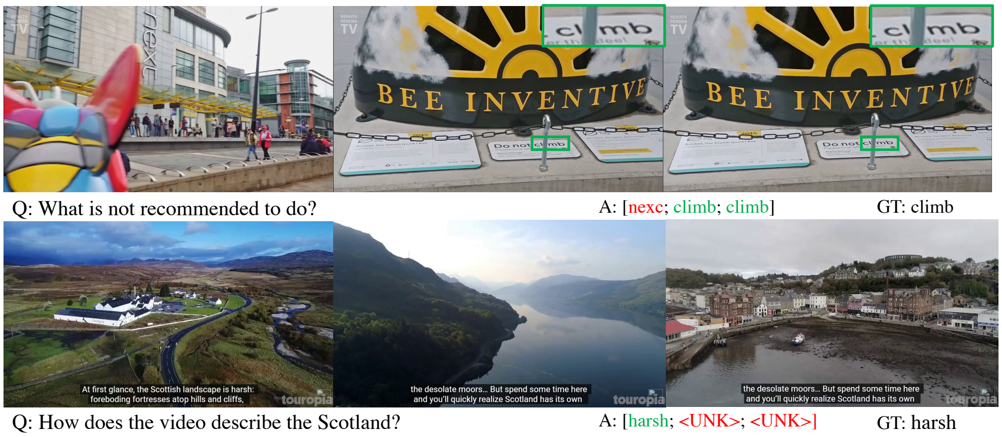

BEE INVENTIVE
climb
Do not climb
Q: What is not recommended to do?
A: [nexc; climb; climb]
GT: climb
At first glance, the Scottish landscape is harsh: foreboding fortresses atop hills and cliffs,
the desolate moors... But spend some time here and you'll quickly realize Scotland has its own
touropia
Q: How does the video describe the Scotland?
A: [harsh; <UNK>; <UNK>]
GT: harsh

Figure 2: Two examples from traveling category. ‘A’ indicates the answers returned by TextVQA models and wrong answers are colored in red.

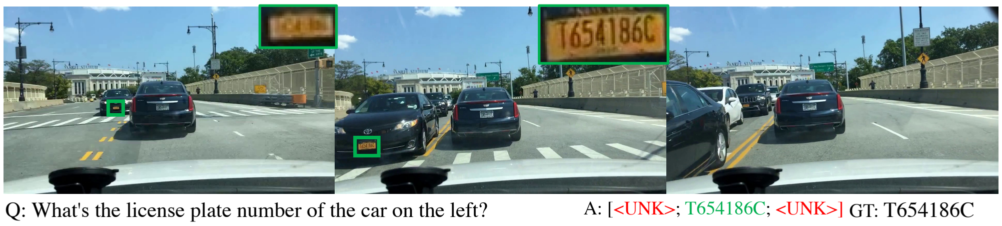

T654186C
Q: What's the license plate number of the car on the left?
A: [<UNK>; T654186C; <UNK>]
GT: T654186C

Figure 3: An example from driving category. ‘A’ indicates the answers returned by TextVQA models and wrong answers are colored in red.

7.5 Sport

This category includes sports videos including football, basketball, rugby and many other sports. Models need to understand temporal events to answer questions about scores, goals, etc... An example is given in Fig. 5.

7.6 Advertisement

As shown in Fig. 6, this category consists of advertisements collected from the network.

7.7 Movie

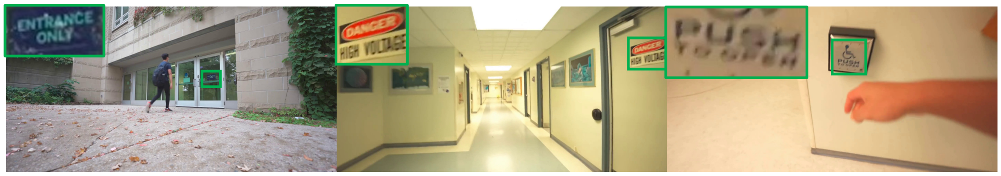
This category mainly consists of dialogues in movies and TV series. Unlike natural texts, the movie class requires the model to have the ability to read and understand contextual dialogue from the subtitle. One example is given in Fig. 7.

7.8 Game

This category consists of videos collected from several popular video games. This category requires the model to have an understanding of the events in the game. The layout of the texts in the game also poses a huge challenge for ViteVQA models. One example is given in Fig. 8.

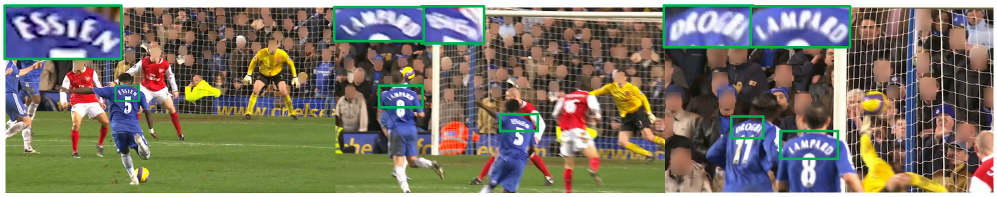
7.9 Talking

This category requires the model to understand the information in news and speeches. An example is shown in Fig. 9.



Q: What does the button he press at the end say? A: [<UNK>; <UNK>; PUSH TO OPEN] GT: PUSH TO OPEN

Figure 4: An example from vlog category. 'A' indicates the answers returned by TextVQA models and wrong answers are colored in red.



Q: Who is shooting the goal? A: [ESSIEN; LAMPARD; LAMPARD] GT: ESSIEN

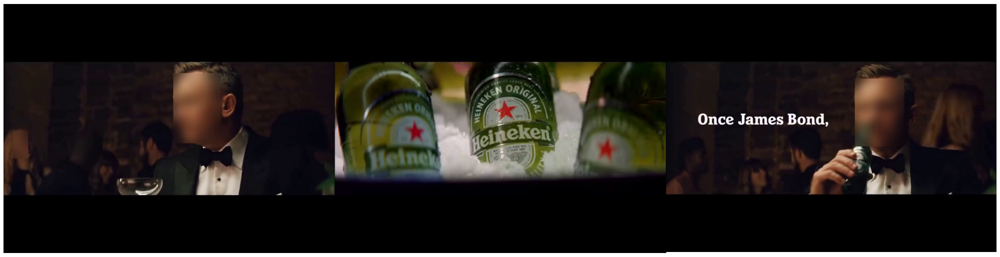
Figure 5: An example from sport category. 'A' indicates the answers returned by TextVQA models and wrong answers are colored in red.

8 Visualization

In this section, we visualize some cases in M4C [1] and T5-ViteVQA. The results are given in Fig. 10. As can be checked in Fig. 10, M4C can not solve some temporal action-related (1st case) and temporal layout-related (2nd case) questions. Although T5-ViteVQA has a better temporal reasoning ability, it still fails in some cases where texts change rapidly. For example, in the 3rd case of Fig. 10 the rapid change of the number of the items in the cart introduces huge challenges in reading and tracking these OCR tokens as well as understanding them. This indicates that our proposed T5-ViteVQA still has great space for improvement.

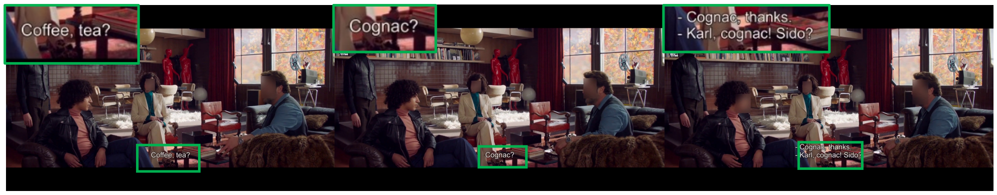
References

[1] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa," in *CVPR*, 2020, pp. 9992–10 002.



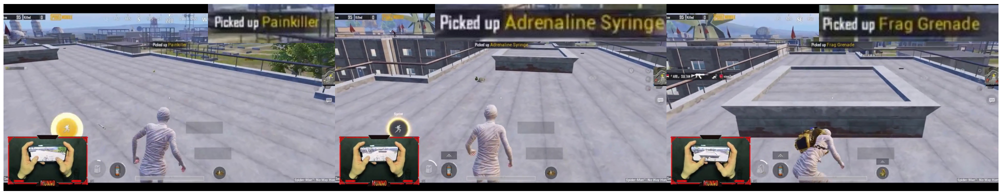
Q: What brand is the beer? A: [<UNK>; Heineken; <UNK>] GT: Heineken

Figure 6: An example from advertisement category. ‘A’ indicates the answers returned by TextVQA models and wrong answers are colored in red.



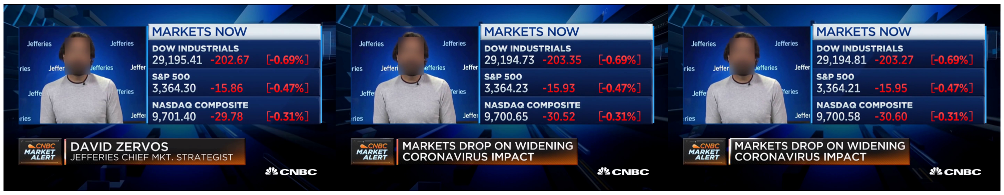
Q: What does he want for drink? A: [<UNK>; <UNK>; Cognac] GT: Cognac

Figure 7: An example from movie category. ‘A’ indicates the answers returned by TextVQA models and wrong answers are colored in red.



Q: What was the second thing he picked up? A: [PainKiller; Adrenaline Syringe; Cognac] GT: Adrenaline Syringe

Figure 8: An example from game category. ‘A’ indicates the answers returned by TextVQA models and wrong answers are colored in red.



Q: What is the S&P 500 index shown at last? A: [3,364.30; 3,364.23; 3,364.21] GT: 3,364.21

Figure 9: An example from talking category. ‘A’ indicates the answers returned by TextVQA models and wrong answers are colored in red.

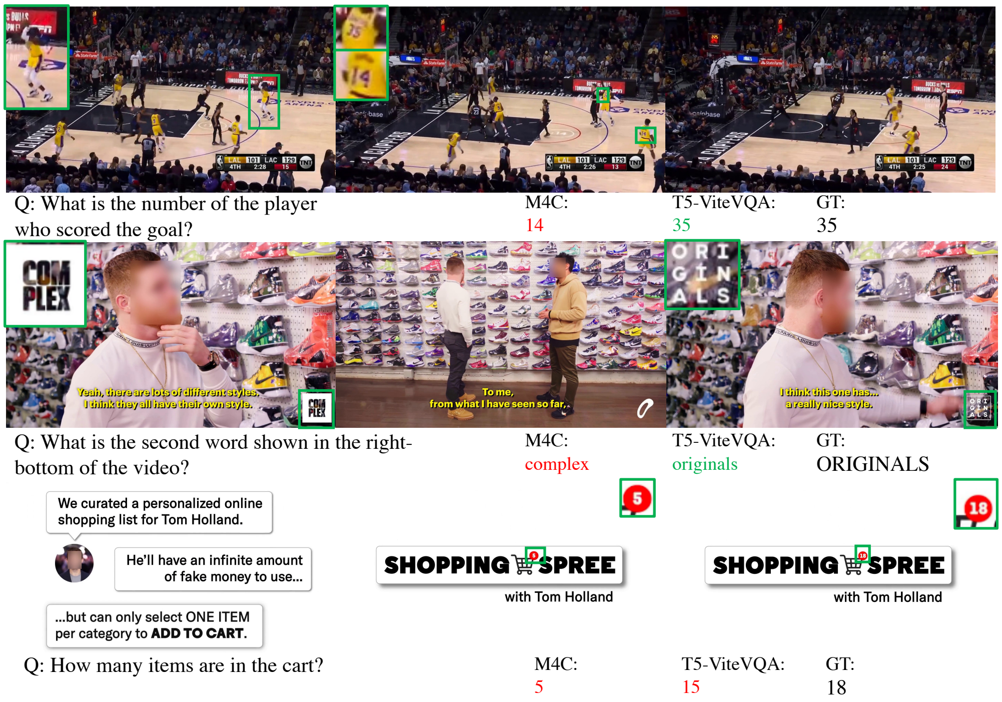

Q: What is the number of the player who scored the goal?
M4C: 14
T5-ViteVQA: 35
GT: 35
Yeah, there are lots of different styles, I think they all have their own style.
To me, from what I have seen so far,
I think this one has... a really nice style.
Q: What is the second word shown in the right-bottom of the video?
M4C: complex
T5-ViteVQA: originals
GT: ORIGINALS
We curated a personalized online shopping list for Tom Holland.
He'll have an infinite amount of fake money to use...
...but can only select ONE ITEM per category to ADD TO CART.
SHOPPING SPREE with Tom Holland
5
18
Q: How many items are in the cart?
M4C: 5
T5-ViteVQA: 15
GT: 18

Figure 10: Some qualitative examples on the M4-ViteVQA dataset. We compare our method with M4C [1] and find that our method performs better than M4C.