

Supplementary Material: Interactive Video Retrieval with Dialog

Sho Maeoki¹, Kohei Uehara¹, and Tatsuya Harada^{1,2}

¹The University of Tokyo, ²RIKEN

{maeoki, uehara, harada}@mi.t.u-tokyo.ac.jp

A. About the AVSD Dataset

According to [1], in the annotation process of the AVSD dataset, two workers on Amazon Mechanical Turk (AMT) were asked to help annotate a video, one of whom is a questioner and the other is an answerer. The questioner is presented with 3 static images (frames) from the video v^{true} i.e., beginning, middle and end of the video, which we write F_b, F_m, F_e respectively, then asks a question. The answerer, who has already watched the video and read a script D_0 about the video, responds. At the end of the dialog, the questioner has to provide a written summary of what happened in the video based on his/her prediction about the target video v^{pred} inspired by the dialog history. Therefore, the questioner needs to ask questions to obtain a good understanding of what is actually happening in the video. Additionally, two workers are implicitly encouraged to hold conversations with rich information unique to videos in this protocol. Based on his/her experience θ_{exp} , the questioner asks questions $\tilde{Q} = \{\tilde{q}_t\}_{t=1}^T$ such that

$$\tilde{q}_t = \text{human}(H_1, \dots, H_{t-1}, F_{b,m,e}, v_{t-1}^{pred}; \theta_{exp}), \quad (19)$$

which facilitates writing the summary of the video. Then, the answerer replies with answers $\tilde{A} = \{\tilde{a}_t\}_{t=1}^T$, where

$$\tilde{a}_t = \text{human}(\text{DH}_{t-1}, v^{true}). \quad (20)$$

The question generation process in Eq. 13 is based on the dialog history, candidate videos, and the videos in the database. On the other hand, Eq. 19 is based on questions and answers, three frames from the video, and the prediction about the video. These two processes are similar as the questioner is required to guess the true video given partial information including dialog, candidates/prediction about the video. More specifically, the AVSD questioner can take into account $F_{b,m,e}$ and v_{t-1}^{pred} in Eq. 19, which allow the AVSD questioner to consider possible candidate videos imagined from both of them. Hence, $F_{b,m,e}$ and v_{t-1}^{pred} play a similar role to $Cand_{t-1}$ in Eq. 13. Besides, the workers' experience θ_{exp} would also play a similar role to V in Eq. 13. Thus, we can assume $\tilde{Q} \approx \hat{Q}$.

B. Performance Analysis w.r.t λ_{feat} and λ_{ques}

While most of the hyperparameters mentioned in Sec 4.1. are determined based on [1] and [3], λ_{feat} and λ_{ques} are determined based on the results in Table 4.

Table 4. Comparison of retrieval performance depending on λ_{feat} . λ_{ques} is set to 2 for all the experiments in the table. λ_{feat} is set to 1000 in the experiments in the main paper as a result of this study.

	R@1	R@5	R@10	MeanR	MRR
$\lambda_{feat} = 1$	0.200	0.900	2.00	382	0.0126
$\lambda_{feat} = 10$	1.20	5.00	9.60	210	0.0442
$\lambda_{feat} = 100$	3.90	14.8	23.4	112	0.104
$\lambda_{feat} = 1000$	4.40	15.0	24.4	109	0.113
$\lambda_{feat} = 10000$	2.60	10.3	19.3	143	0.0777

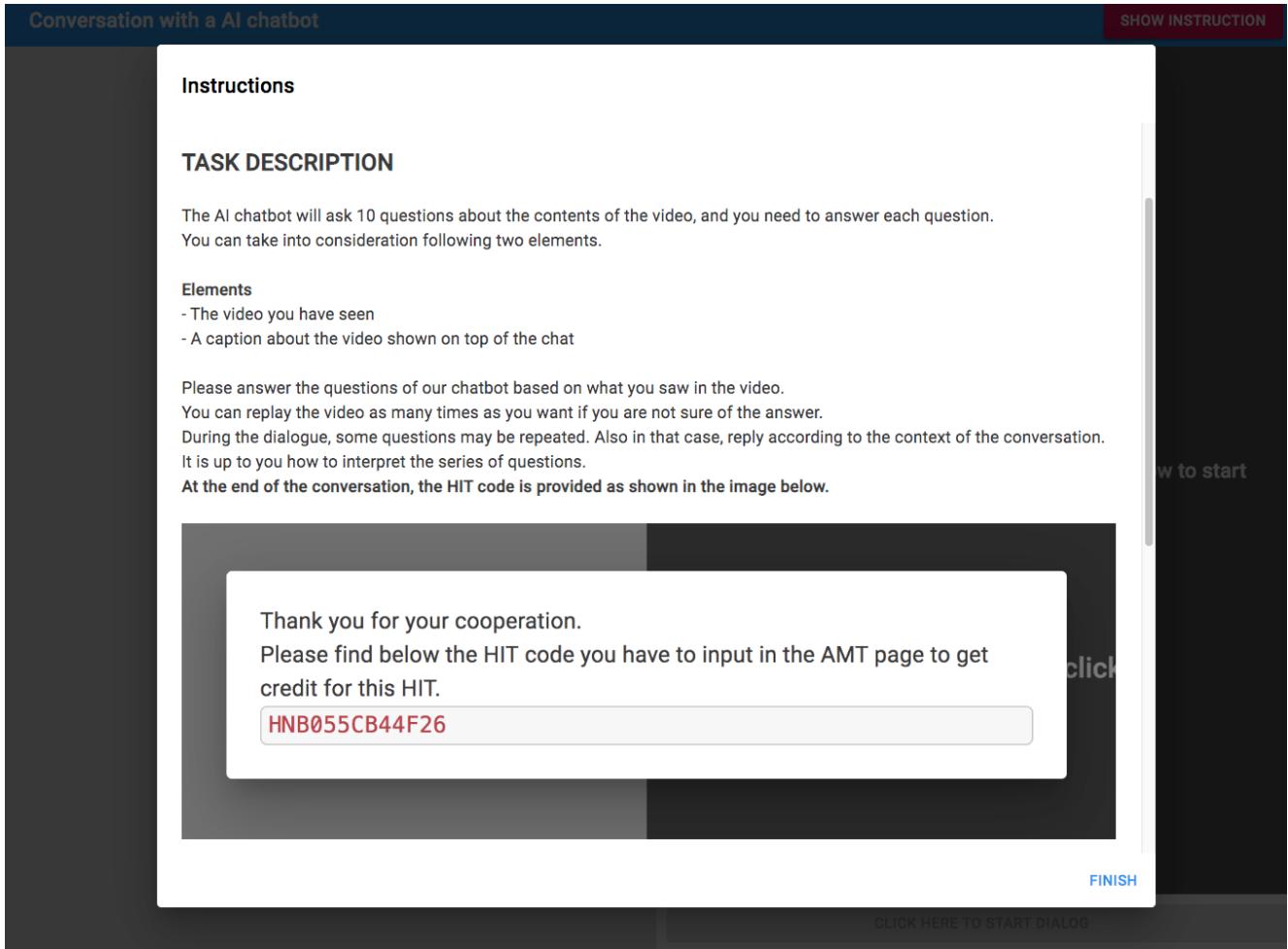


Figure 6. Instruction of the user study 1.

C. User Study Settings

Fig. 6 and Fig. 7 show the instructions in the user study and Fig. 8 is the appearance of the AMT interface in the user study. The instructions in Fig. 6 and Fig. 7, some of which are based on [2], are displayed to users before starting dialog. In the user study, the users were asked to respond to questions generated by the model while seeing the GT video. The main goal in this user study is to see if dialog contributes to improvement in video retrieval performance even when the model interacts with humans in reality. Therefore, GT captions for videos were first displayed to users to avoid a significant difference from the dataset in terms of dialog contents. The questions asked to the users were generated by the model. In our user study, at the end of each round, the model generates a question q_{t+1} conditioned on the dialog history including a video caption D_0 , previous questions q_1, \dots, q_t and user's answers a_1, \dots, a_t . Of course, users are also allowed to take into account the dialog history as well as caption and video. In the end, the GT video ranks in the database are calculated using the dialog data obtained.

D. Qualitative Results

Fig. 9, ..., Fig. 14 are examples of qualitative results as a consequence of the user study. Overall, dialog improves retrieval performance though in some cases dialog history is not properly utilized to improve the retrieval performance as in Fig. 12. Also, we can see there are quite a few generic questions and some of them are repeated. However, these examples include questions that could improve retrieval performance, *e.g.*, questions about temporal relationships, about scenes, about audio information. On the other hand, it is noteworthy in some examples questions targeting on specific topics are generated based on the input query. For instance, in Fig. 9 and Fig. 10, questions using specific words (*e.g.*, broom, towel) are generated.

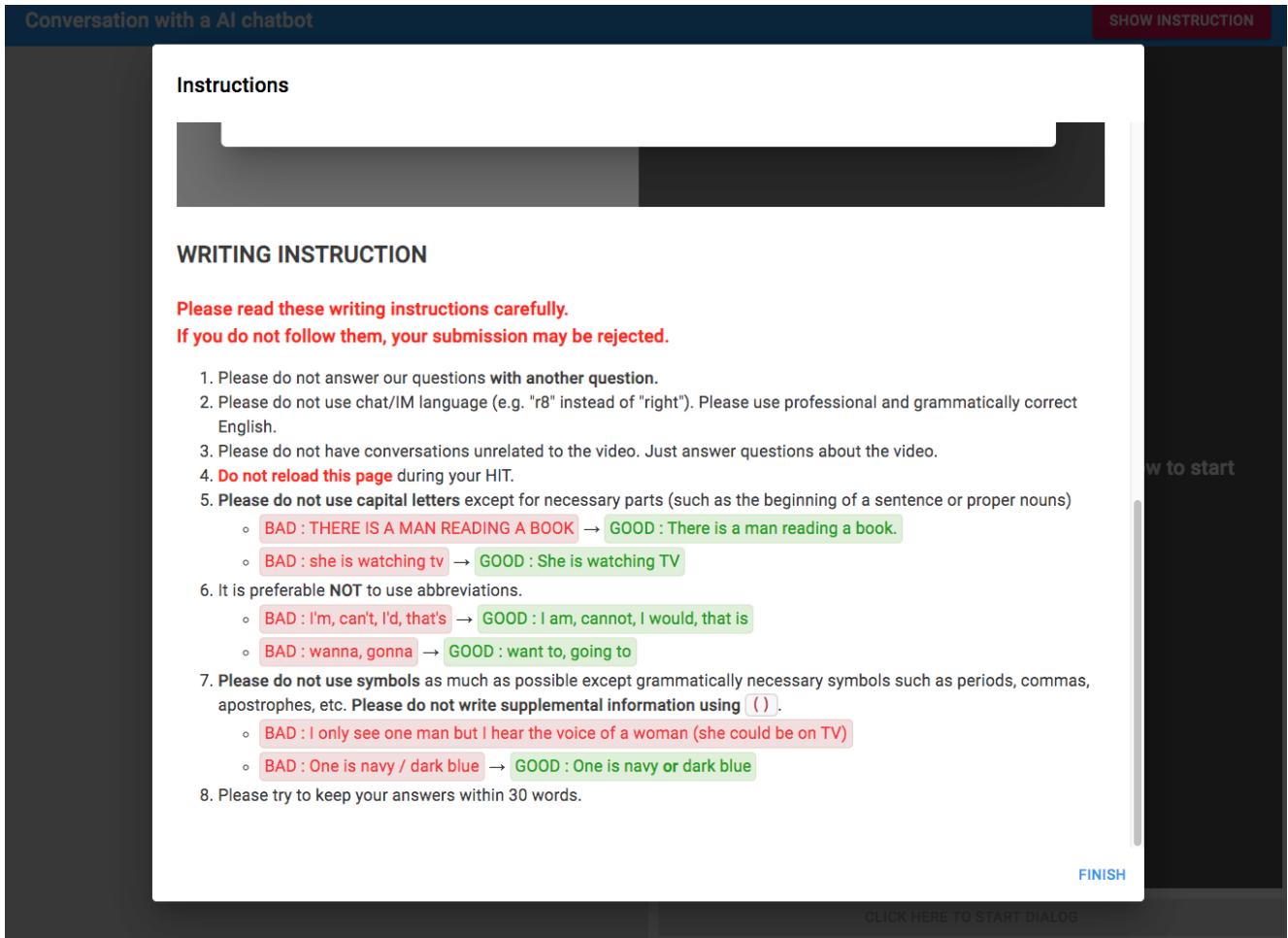


Figure 7. Instruction of the user study 2.

Additionally, we saw questions such as “*Can you tell what he is eating?*”, “*Can you tell what he is cooking?*” are generated for target videos involved with cooking, which is also based on specific topics. Note we show the qualitative results until the 4-th round in Fig. 9, ..., Fig. 12 as the highest R@10 is remarked in Fig. 4 (b).

References

- [1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. Audio-Visual Scene-Aware Dialog. In *CVPR*, 2019.
- [2] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *CVPR*, 2017.
- [3] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval. In *ICMR*, 2018.

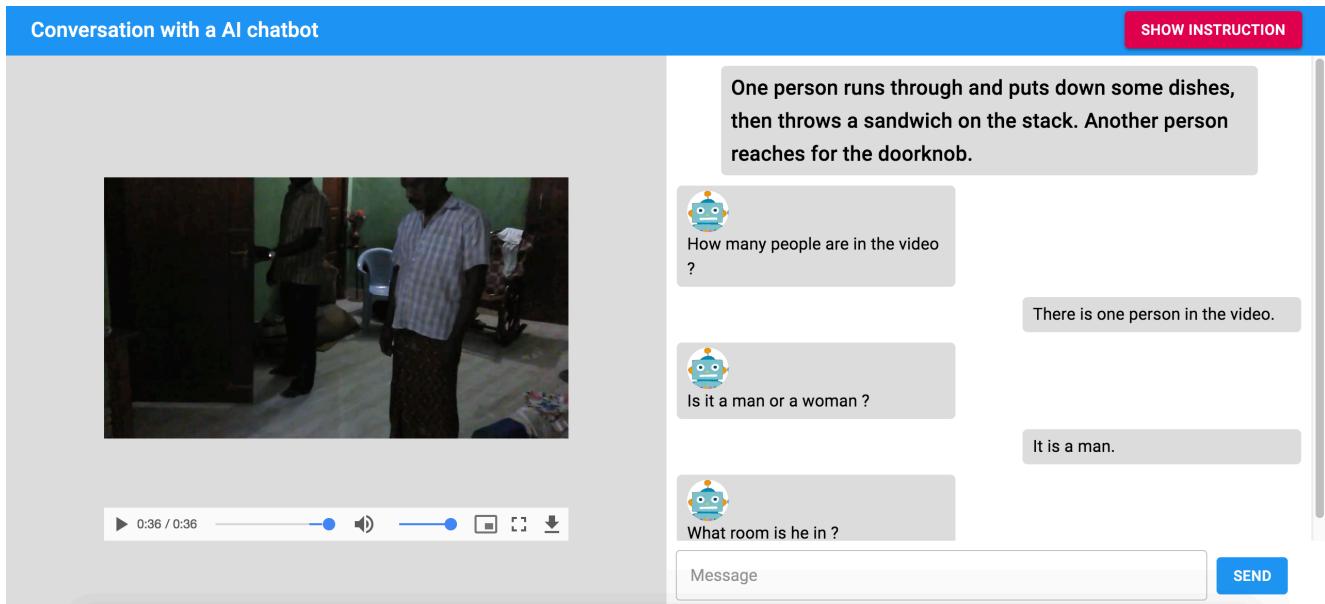


Figure 8. Appearance of the AMT interface in the user study.

caption	GT rank	Top Video
A person laughs as they walk down the stairs . The person hangs a towel from a broom stick , and leaves .	1	GT Video A
Q1:How many people are in the video ?	1	
Q2:Is the person a man or a woman ?	3	Video A.I
Q3:What is she doing at the beginning of the video ?	1	
Q4:What does she do with the broom ?	1	GT Video A



Figure 9. Qualitative result 1.

caption		GT rank	Top Video
A person is washing the clothes after undressing and talking a towel off the shelf .		86	Video B.I
Q1:How many people are in the video ?	A1:One.	107	
Q2:What is he doing at the beginning of the video ?	A2: Unzipping her jacket.	60	
Q3:What is the first thing that happens in the video ?	A3:The woman closes a cabinet.	61	Video B.II
Q4:What does she do with the towel ?	A4:Puts it in the washer.	27	



GT Video B



Video B.I



Video B.II

Figure 10. Qualitative result 2.

caption		GT rank	Top Video
A person takes dishes off a shelf then starts drinking some water .		58	Video C.I
Q1:How many people are in the video ?	A1:There are one Person in the video	127	
Q2:What is he doing at the beginning of the video ?	A2: She Opened the Kitchen Cabinet	8	Video C.II
Q3:What does she do next ?	A3: She took out some Bowls	1	GT Video C
Q4:What does she do after that ?	A4:She Opened the Tap	1	



GT Video C



Video C.I



Video C.II

Figure 11. Qualitative result 3.

caption		GT rank	Top Video
A person awakens in a kitchen holding a phone and sitting at a table .		558	Video D.I
Q1: How many people are in the video ?	A1: There is only one person in the video.	577	Video D.II
Q2: What is he doing at the beginning of the video ?	A2: He is resting by keeping his head on the table.	618	Video D.I
Q3: What does he do after that ?	A3: He wakes up after that and he check over his phone.	641	Video D.III
Q4: Does he ever leave the room ?	A4: No, he does not leave the room.	622	
			
GT Video D	Video D.I		
			
Video D.II	Video D.III		

Figure 12. Qualitative result 4.

caption		GT rank
A person opens the door to their closet and throws their blanket inside .		81
Q1: How many people are in the video ?	A1: One person.	70
Q2: What is he doing at the beginning of the video ?	A2: He was opening the door to his closet.	76
Q3: What does he do next ?	A3: He put on the light.	60
Q4: What is the first thing that happens in the video ?	A4: A man with a cloth in his hand opening a door.	36
Q5: What does he do after that ?	A5: He switched on the light.	30
Q6: What does he do after that ?	A6: He throws the cloth in his hand into the closet.	25
Q7: Does he do anything else ?	A7: No, he did nothing else.	24
Q8: Does he do anything else ?	A8: He did nothing else after he threw the cloth.	20
Q9: What does he do after that ?	A9: Nothing else.	20
Q10: Is there anything else I need to know about the video ?	A10: The guy has a blue shirt and a black pant on.	10

Figure 13. Qualitative result 5.

caption		GT rank
A person is holding a camera , taking photos of another person who is sitting on a sofa watching TV and holding a plate of food .		101
Q1: How many people are in the video ?	A1: There are two people in the video	57
Q2: Are they talking to each other ?	A2: No	55
Q3: What is the first thing that happens in the video ?	A3: A man is sitting in the sofa with a plate in one hand and Television remote in other hand	46
Q4: What is the first thing that happens ?	A4: The man is operating the television remote.	38
Q5: Is there anything else I need to know about the video ?	A5: A woman is taking picture of the man sitting in the sofa with her camera.	36
Q6: Is there anything else I need to know ?	A6: The man is watching the television	34
Q7: How does the video end ?	A7: The video end by showing the man again	34
Q8: Can you tell what time of day it is ?	A8: It may be noon time	41
Q9: Is there anything else I need to know about the video ?	A9: The guy has a blue shirt and a black pant on.	43
Q10: Can you tell what time of day it is ?	A10: I cannot tell	44

Figure 14. Qualitative result 6.