Supplementary Material
Describing Unseen Videos via Multi-Modal Cooperative Dialog Agents

Ye Zhu[1], Yu Wu[2,3], Yi Yang[2], and Yan Yan[1⋆]

[1] Texas State University, San Marcos, USA
{ye.zhu, tom_yan}@txstate.edu
[2] ReLER, University of Technology Sydney, Australia
yu.wu-3@student.uts.edu.au,yi.yang@uts.edu.au
[3] Baidu Research, Beijing, China

We mainly provide additional qualitative experimental results and analysis in this document.

Internal Dialog Generations. Selected examples of the generated dialog interactions between two agents are shown in Fig. 1 and Fig. 2. The corresponding round of the ground truth dialog is also given as comparisons. During the training process, the question and answer decoders are trained to imitate the ground truth dialog by providing the ground truth words as sequence input labels, but the entire model is optimized in the final video description task. As shown in the figures, the generated dialog is not always natural and precise in the sense of natural language understanding, although the generated questions and answers sometimes do contain keywords that are later used in the final descriptions (*e.g.*, kitchen, fridge, couch, dish). *Q-BOT* also tends to ask repeated questions in the internal dialog generation process, especially with small starting round number.

Video Descriptions. More qualitative results about the final video description task are presented in Fig. 2. We observe that *Q-Basic* tends to generate similar descriptions for different inputs. It is a common and existing problem in the related research fields that combine natural language processing and computer vision [2, 1]. Other methods (*i.e.*, *Q-Strong*, *A-BOT* and our *Q-BOT* with cooperative learning) alleviate the problem and include more specific details in the generated descriptions. However, the sentence structures are not as varied as the human (GT) performance, which leads to another diversity problem that researchers seek to solve [3, 5, 4]. The qualitative results show that our *Q-BOT* with cooperative learning is able to describe some details that are not presented in the initial inputs (*e.g.*, the refrigerator in the first example), which proves that *Q-BOT* obtains useful information from the dialog interactions with *A-BOT*.

Further Analysis. We further analyze different aspects of this work. Firstly, we believe this task setting is of great potential to provide a new perspective that brings more secure and reliable AI systems. The task setting with implicit information sources is not only limited to the proposed video description task, but also applicable in a wider context such as scene understanding and segmentation. Secondly, under the overall framework of multi-modal learning

⋆ Corresponding author.

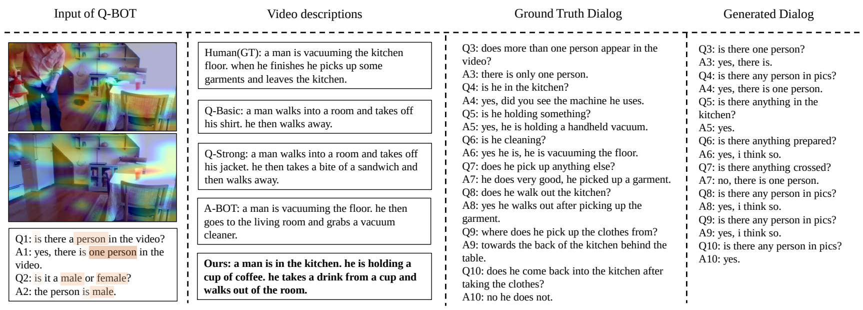

Fig. 1. Qualitative example with complete generated internal dialog.

with implicit information sources, there remain multiple future directions that worth being further exploited. For example, given the fact that *Q-BOT* tends to ask similar questions during the internal dialog process, we can consider incorporating internal supervision to motivate *Q-BOT* to ask more diverse and informative questions. In the meanwhile, we observe from the experiments that the performance does not necessarily improve when given more chances to freely ask/answer questions during the internal dialog interactions for a given video, indicating the fact that *Q-BOT* might have the potential to acquire the necessary information in less than 10 rounds of question-answer interactions. One possible future direction inspired by the observation is to encourage more efficient dialog interactions between the two agents.

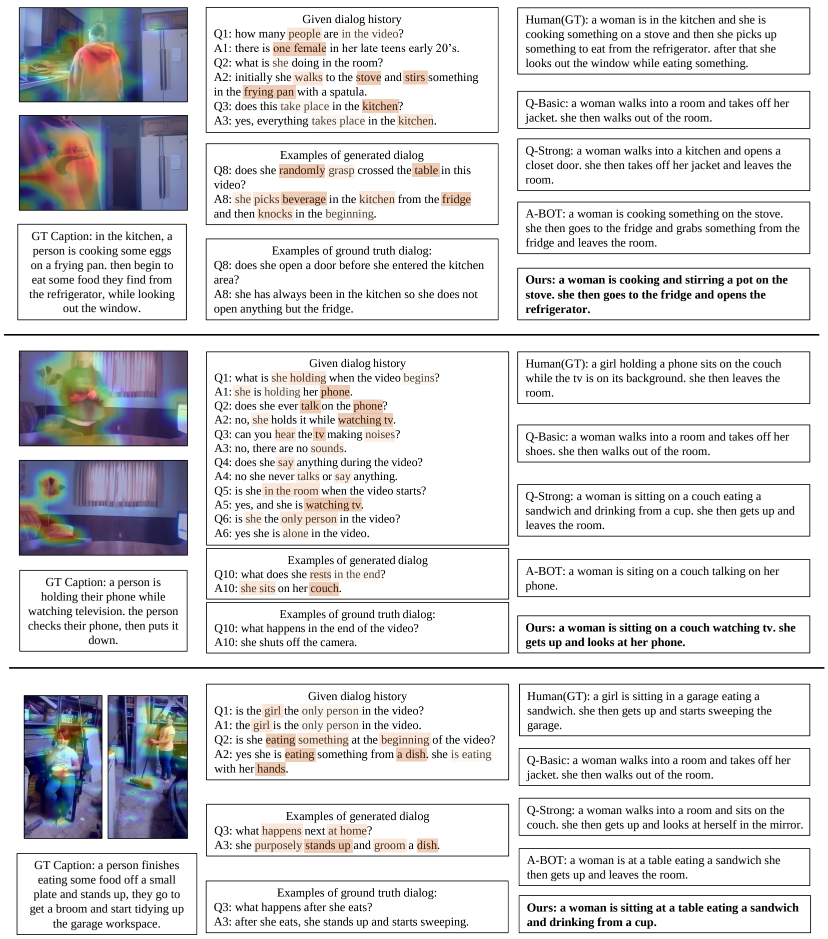

Fig. 2. More qualitative results for the unseen video description task. We provide the input static images to *Q-BOT* on the left. The given dialog history, examples of generated dialog and the ground truth dialog are presented in the middle. Comparisons of the final descriptions among different methods are shown on the right.

References

1. Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., et al.: Audio visual scene-aware dialog. In: CVPR (2019)
2. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: CVPR (2017)
3. Jain, U., Lazebnik, S., Schwing, A.G.: Two can play this game: visual dialog with discriminative question generation and answering. In: CVPR (2018)
4. Wang, L., Schwing, A., Lazebnik, S.: Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In: NeurIPS (2017)
5. Wu, Q., Wang, P., Shen, C., Reid, I., Van Den Hengel, A.: Are you talking to me? reasoned visual dialog generation through adversarial learning. In: CVPR (2018)