Supplementary material of "On the hidden treasure of dialog in video question answering"

Deniz Engin[1,2] François Schnitzler[2] Ngoc Q. K. Duong[2] Yannis Avrithis[1]
[1]Inria, Univ Rennes, CNRS, IRISA [2]InterDigital

A. Additional qualitative analysis

Dialog summarization In the example of Figure 4, Howard says, "I invited *her*." in scene B. Our dialog summarization interprets this sentence by assigning the correct character name: "Howard invited *Bernadette* in." Hence, we can answer the question of scene A, "Who did Howard invite to join him and Raj in Raj's lab?" correctly. Thanks to the episode dialog summary spanning all scenes and the use of character names instead of pronouns, our method can answer character-related questions correctly.

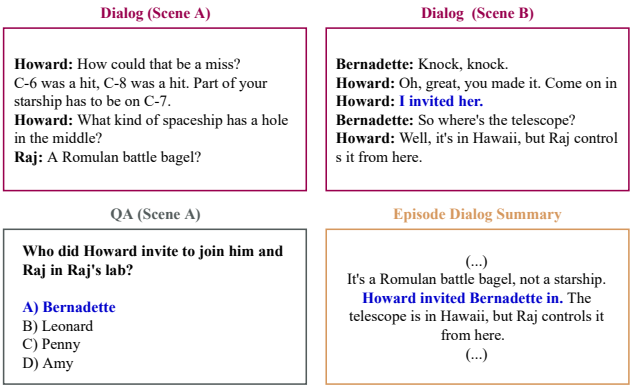

Figure 4: Dialog summarization converts pronouns in dialog to character names in episode dialog summary, supporting question answering. In particular, "I" is substituted by "Howard" and "her" by "Bernadette".

Plot *vs*. episode dialog summary A comparison of plot summary and episode dialog summary is given in Figure 5. There are three different topics in the story line, and each is highlighted with the same color in both summaries. The first topic, highlighted in purple, is "Sheldon's forgotten flash drive." The second, highlighted in yellow, is "Sheldon's grandmother." The third, highlighted in blue, is "Asking Summer out." The plot summary is topic-centered, while the episode dialog summary is following the narrative order. Hence, topics may be fragmented in the latter. The episode dialog summary has more detail than the plot. In particular, it contains enough information to answer the question *Why does Sheldon's grandmother call him Moon Pie?* That is, *because he's nummy-nummy*. This information is missing from the plot summary, which focuses on the main topics/events of an episode. Even though the episode dialog summary is noisy, it contains details that help in question answering.

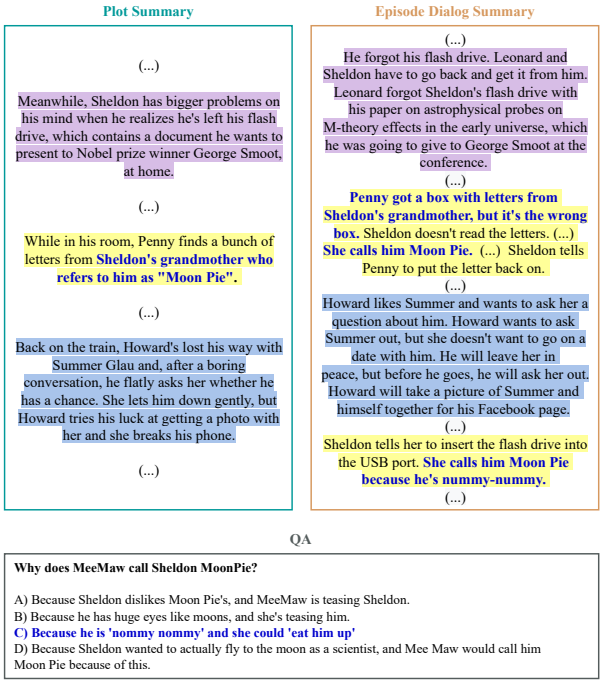

Figure 5: An example of plot summary and episode dialog summary, with each topic highlighted in the same color in both summaries. Phrases relevant to QA in blue. Only the episode dialog summary contains enough information to answer the question.

Failure cases Figure 6 shows examples of failed predictions of our model along with stream attention scores for different question types. The model receives three input sources, question/answers and attention scores over inputs.

Stream	Inputs	Soft Attn.	Vis.	Text.	Temp.	Know.	All
Single	P	-	0.656	0.594	0.628	0.712	0.683
	P	✓	0.666	0.623	0.593	0.735	0.702
	E	-	0.604	0.721	0.733	0.765	0.723
	E	✓	**0.676**	**0.750**	**0.779**	**0.785**	**0.756**
Multi	V + S + P	-	0.732	0.688	0.674	0.720	0.717
	V + S + P	✓	0.739	0.699	0.628	0.728	0.723
	V + S + E	-	0.707	0.772	0.721	0.700	0.711
	V + S + E	✓	**0.755**	**0.783**	**0.779**	**0.789**	**0.781**

Table 5: *Effect of temporal attention on single-stream QA* on KnowIT VQA. Soft Attn.: soft temporal attention on single-stream training. We use soft temporal attention for multi-stream QA, but this is still affected by the temporal attention used in single-stream training. V: video; S: scene dialog summary; P: plot; E: episode dialog summary.

Method	Soft Attn.	Vis.	Text.	Temp.	Know.	All
Product	-	0.728	0.645	0.744	0.756	0.736
	✓	0.743	0.659	0.756	0.751	0.739
Modality weighting [10]	-	0.716	**0.815**	0.791	0.776	0.768
	✓	0.708	0.786	0.767	0.787	0.769
Self-attention	-	0.753	0.804	**0.802**	0.766	0.769
	✓	**0.759**	0.764	0.767	0.777	0.771
Multi-steam attention	-	0.743	0.790	0.779	0.785	0.776
	✓	0.755	0.783	0.779	**0.789**	**0.781**
Multi-steam self attn.	-	0.749	0.797	0.791	0.768	0.768
	✓	0.755	0.768	0.756	0.777	0.770

Table 6: *Effect of temporal attention on multi-stream QA* on KnowIT VQA for fusion of video, scene dialog summary, and episode dialog summary input sources. Soft Attn.: soft temporal attention on multi-stream training. We use soft temporal attention for single-stream QA of episode dialog summary.

Figure 6(a) refers to a knowledge question, which requires recurrent knowledge of the whole TV show. In other words, the correct answer cannot be found in episode dialog summary. The question is answered as "a lasagna" found in episode dialog summary, even though it is wrong.

Figure 6(b) refers to a textual question, which should have been answered by scene dialog summary. However, scene dialog summary does not contain the correct answer. Our model gives most attention to episode dialog summary. The prediction is made according to the highlighted text, which is the same in both sources. However, this prediction refers to the wrong person.

Figure 6(c) refers to another knowledge question, which could be answered by the highlighted text in episode dialog summary. Even though episode dialog summary has the most attention, the prediction is incorrect.

Figure 6(d) refers to another textual question, which should have been answered by scene dialog summary. Although both scene dialog summary and episode dialog summary include the correct answer, and episode dialog summary has the most attention, the prediction indicates the wrong person.

Figure 6(e) refers to a temporal question. The scene dialog summary and episode dialog summary imply that *Raj* and *Howard* might be changing the tire. The video description is not helpful either. Hence, our model predicts *Raj*, while the correct answer is *Howard*.

Figure 6(f) is a visual question. However, the video description fails to convey relevant information to answer the question. The other inputs do not contain relevant information either. One of the character names appearing in episode dialog summary is predicted, which is incorrect.

B. Additional ablation studies

Hyperparameter validation *Modality weighting* [10] fusion method requires selection of hyperparameter β_ω. Figure 7 shows validation accuracy *vs.* β_ω for fusion of video, scene dialog summary and episode dialog summary. We choose $\beta_\omega = 0.7$ for both soft and hard temporal selection to report results in Table 3 and Table 6. The remaining weight of $1-\beta_\omega$ is evenly distributed over individual stream losses as 0.1 per stream.

Effect of temporal attention on single-stream QA We investigate the effect of our soft temporal attention (Subsection 6.2) on single-stream QA for episode input sources. We also evaluate the effect of single-stream training with soft or hard temporal attention on multi-stream attention, where we use soft temporal attention. According to Table 5, temporal attention improves the accuracy of plot and episode dialog summary by 1.9% and 3.3%, respectively. Accordingly, the accuracy of multi-stream QA on the same episode sources as well as video and scene dialog summary increases by 0.6% and 7.0%, respectively. The gain is higher when episode dialog summary is used, since the episode dialog summary is longer than plot.

Effect of temporal attention on multi-stream QA Table 6 shows the effect of *soft temporal attention* on multi-stream QA for fusion of video, scene dialog summary and episode dialog summary input sources. We use soft temporal attention for single-stream QA of episode dialog summary. In all fusion methods, the overall accuracy is improved by using soft temporal attention.

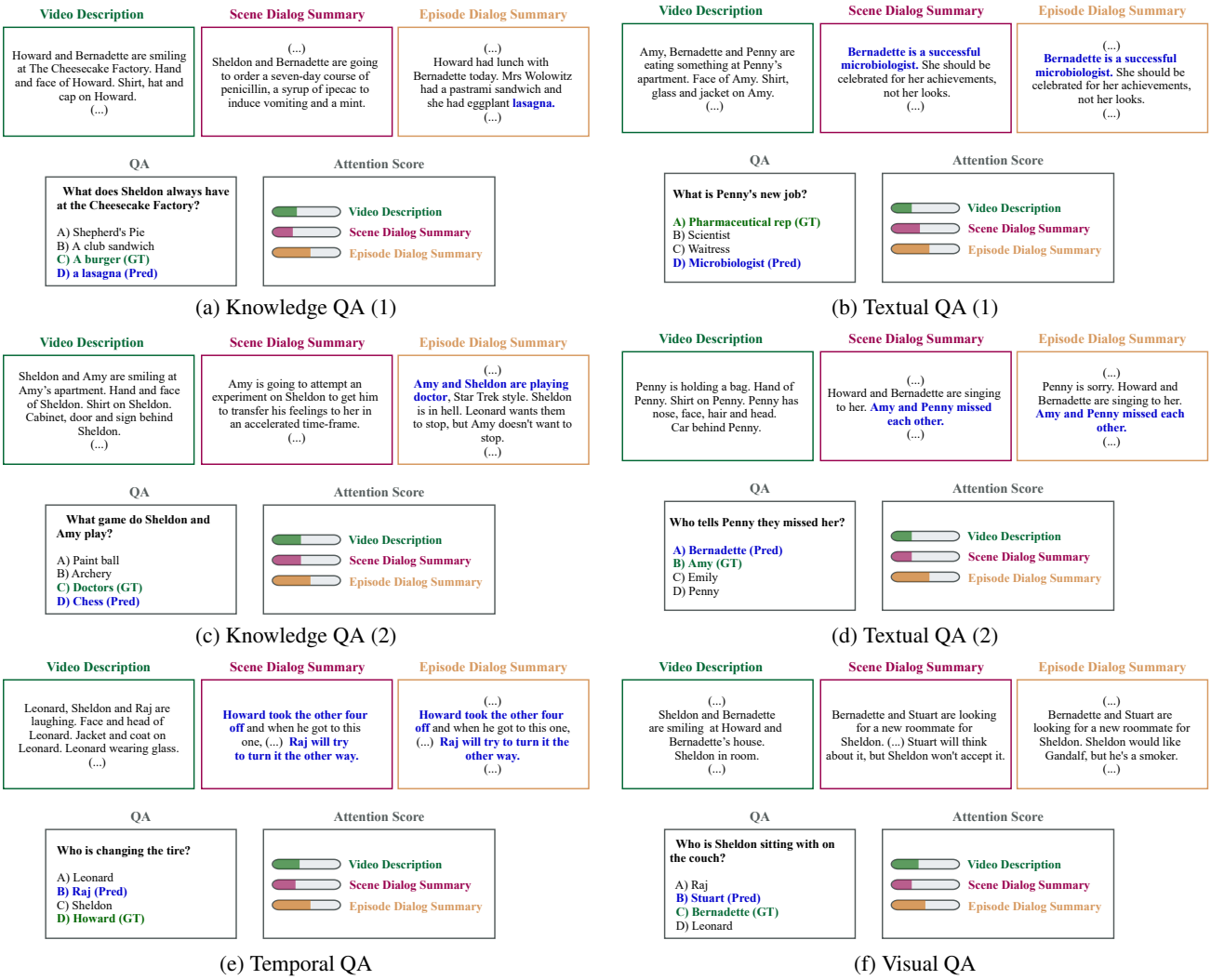

Figure 6: *Failed predictions of multi-stream attention*. We highlight in blue the part of the source text that might be relevant to answering the question. "Pred"/blue: model predictions. "GT"/green: ground truth.

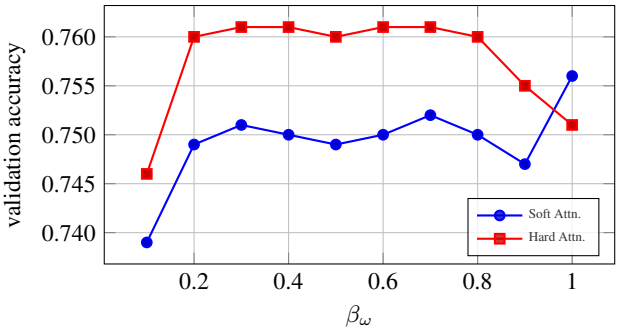

Figure 7: Accuracy *vs*. β_ω for fusion of video, scene dialog summary and episode dialog summary by modality weighting [10] on KnowIT VQA validation set.

Different input combinations Table 7 shows the accuracy of multi-stream QA for different input combinations, where the number of input streams varies in $\{2, 3, 4, 5\}$. Scene dialog summaries improves the accuracy compared with single-stream QA results in Table 2. Moreover, using the episode dialog summary always improves the overall accuracy by a large margin. The best overall accuracy of 0.781 is achieved by video, scene dialog summary, and episode dialog summary.

Question type $\leftrightarrow$ attention scores We perform significance testing for the dependence between the question type and attention scores. There are 2 independent variables in the scores of 3 streams, whose values we discretize into 10×10 bins. We form a $4 \times 10 \times 10$ joint histogram of question type (X) and scores (Y) and compute the mu-

Analysed Inputs	Inputs	Vis.	Text.	Temp.	Know.	All
D V P	D+V	0.693	0.768	0.593	0.554	0.611
	D+P	0.732	0.721	0.674	0.723	0.723
	D+V+P	0.734	0.725	0.663	0.724	0.724
D V P S	D+S	0.664	0.786	0.628	0.548	0.604
	V+S	0.689	0.721	0.581	0.549	0.601
	P+S	0.716	0.710	0.628	0.727	0.719
	D+V+P+S	0.734	0.732	0.663	0.725	0.726
D V P S E	D+E	0.743	0.812	0.779	0.779	0.775
	V+E	0.732	0.761	0.767	0.788	0.772
	P+E	0.716	0.743	0.721	**0.791**	0.766
	D+S+E	0.743	**0.822**	**0.802**	0.771	0.772
	V+S+E	**0.755**	0.783	0.779	0.789	**0.781**
	P+S+E	0.739	0.779	0.733	0.783	0.771
	D+V+P+S+E	0.751	0.797	0.744	0.781	0.775

Table 7: *Multi-stream QA accuracy* on KnowIT VQA: comparison of different input combinations for multi-stream attention. D: dialog; V: video; P: plot; S: scene dialog summary; E: episode dialog summary.

tual information $I(X;Y)$. We perform a G-test[1] with $G = 2N \cdot I(X;Y)$, where $N = 2361$ is the number of test questions. Finally, using a chi-square distribution of $3 \times 9 \times 9$ DoF, we find a p-value of 1.52×10^{-25} for the null hypothesis. This indicates that attention scores depend on question type.

Replacing attention scores with oracle scores determined by question type Assuming that we know the question type for the test set, we perform an *oracle* experiment where attention scores are based on question type rather than our fusion method. We only consider visual, textual, and knowledge types of question. In particular, we assign visual questions to video input, textual questions to scene dialog summary and knowledge questions to episode dialog summary. We exclude temporal questions since they can be answerable by scene dialog summary or video. Only 3.6% of questions are of temporal type in the test set. We find that our multi-stream attention method (0.781%) is 3.6% better than the oracle experiment (0.745%). This indicates that our fusion mechanism is more effective than a naïve oracle that assumes more knowledge.

[1] https://en.wikipedia.org/wiki/G-test