1 TECHNICAL APPENDIX

1.1 Validation of Hypothesis Mentioned in Auxiliary Head Section in Methodology

We claim that one primary source of hallucination is the inadequacy of the extracted visual features. Inadequate motion features lead to action hallucination, whereas inadequate object and appearance features lead to object hallucination. To validate the hypothesis, we added an increasing amount of white noise to reflect the increasing inadequacy scenario and calculated the action and object hallucination rate. As shown in Fig 1, if we increase the amount of inadequacy

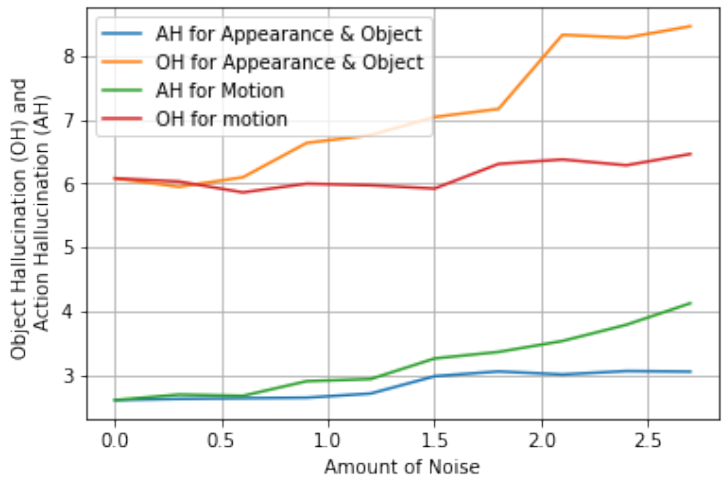

Fig. 1. The action and object hallucination plot for the increasing amount of added noise to the visual features. The experiments are done on MSVD data using the MARN (Pei et al. 2019) model. For AH and OH, lower is better.

for motion feature, the action hallucination increases rapidly than object hallucination. Similarly, for appearance and object features, object hallucination increases rapidly than action hallucination.

1.2 Validation of Hypothesis Mentioned in Context Gates Section in Methodology

According to our claim, another primary source of hallucination is the improper influence of features during intra-modal and multi-modal fusion. To validate and

reflect the scenario, we perturbed features' influence and checked their hallucination rate. As shown in Fig 2, during intra-modal fusion, the object hallucination increases significantly when we increase the influence of motion features over object and appearance features. Similarly, for dominant object and action features, action hallucination increases rapidly. Finally, we have shown the COAHA metric values for the increasing influence of language prior over source context during multi-modal fusion. We can see the hallucination increases with dominant language prior.

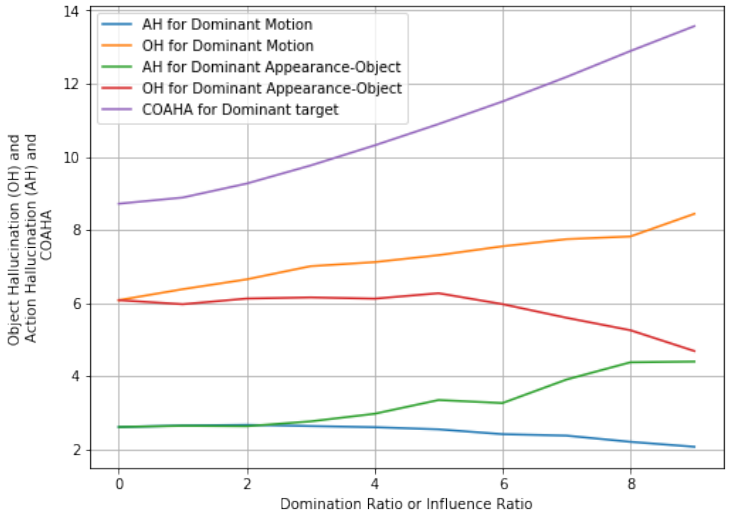

Fig. 2. Action hallucination (AH), object hallucination (OH), and COAHA values at the increasing amount of influence of one feature over others. The experiments are done on MSVD data using the MARN (Pei et al. 2019) model. For AH, OH, and COAHA, lower is better.

1.3 Study on Running Memories

Context gates are responsible for choosing the right amount of influence for features during the generation of each word. In order to predict the influence, the context gates rely on what is already generated so far. To that end, other methods (Tu et al. 2017) have used the LSTM hidden memory as one of the inputs to the context gate. In our case, a separate running visual and language memory performed better. The comparison of performance is shown in Table 1.

1.4 Performance Comparison of Different Decoder Architectures

For the language decoder, we have used the LSTM network. We have also examined GRU and Transformer models. As shown in Table 2, the performance of

Memory Type	B@4	M	R	C
Runing Memory	53.3	36.5	74.0	99.9
LSTM Memory	52.4	36.2	73.6	91.6
Runing + LSTM	52.2	36.4	73.4	95.2

Table 1. Performance comparison of different types of memories as input to the context gates.

LSTM and Transformer models are close to each other and better than GRU. For simplicity and to mainly focus on hallucination, we have not used any external language models.

Decoder Type	B@4	M	R	C
LSTM	53.3	36.5	74.0	99.9
GRU	50.1	35.9	72.5	93.5
Transformer	54.1	36.6	73.7	98.1

Table 2. Performance comparison of different types of decoder architectures.