CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Modality Shifting Attention Network for Multi-modal Video Question Answering

Junyeong Kim[1], Minuk Ma[1], Trung Pham[1], Kyungsu Kim[2], Chang D. Yoo[1]

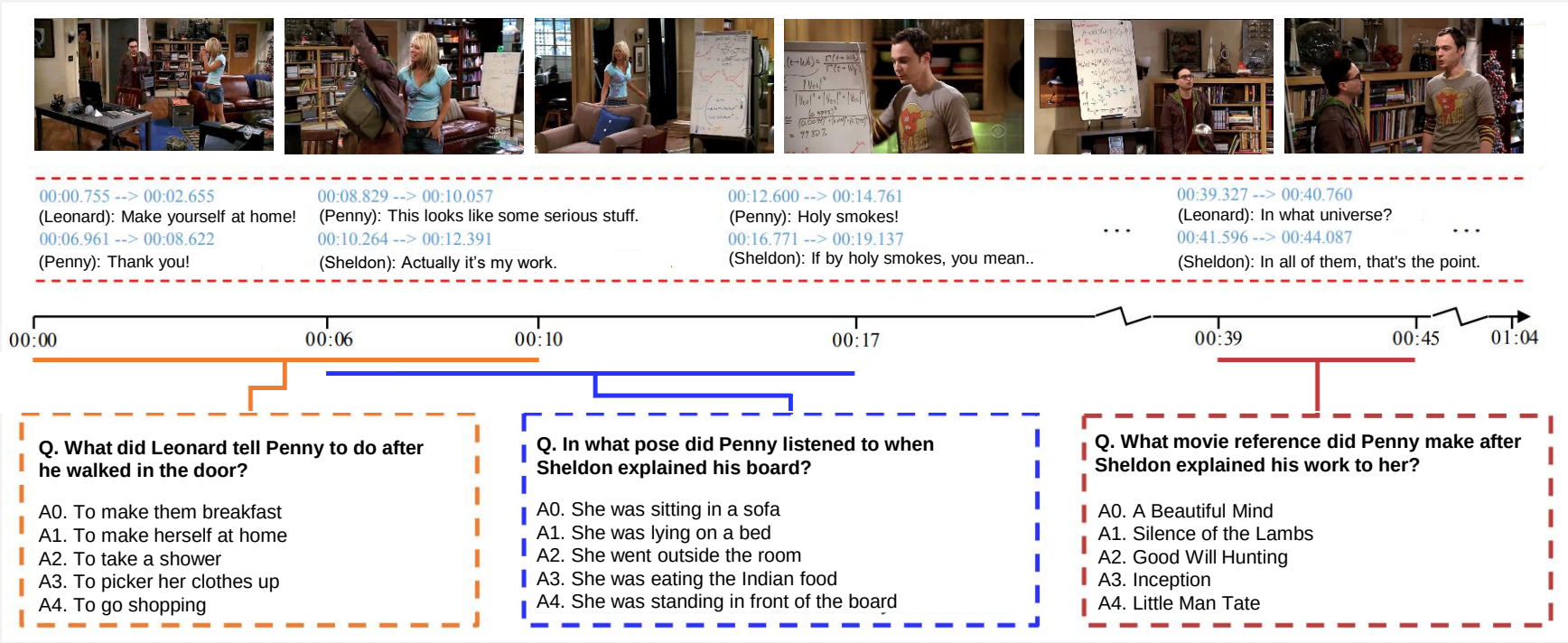
[1]Korea Advanced Institute of Science and Technology (KAIST)

[2]Samsung Electronics

KAIST

AIM Lab.
Artificial Intelligence & Machine Learning Lab.

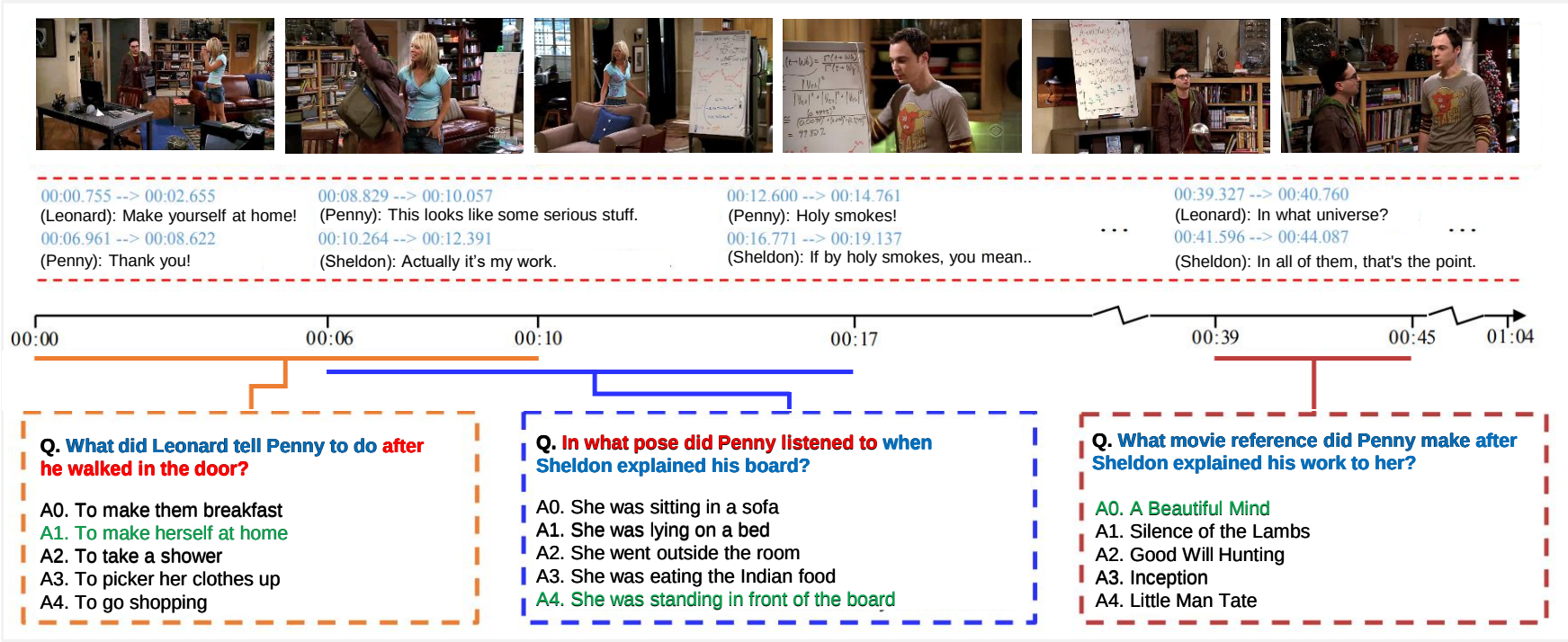
Multi-modal Video Question Answering (MVQA)



(Lie et al. TVQA: Localized, Compositional Video Question Answering, EMNLP 2018)

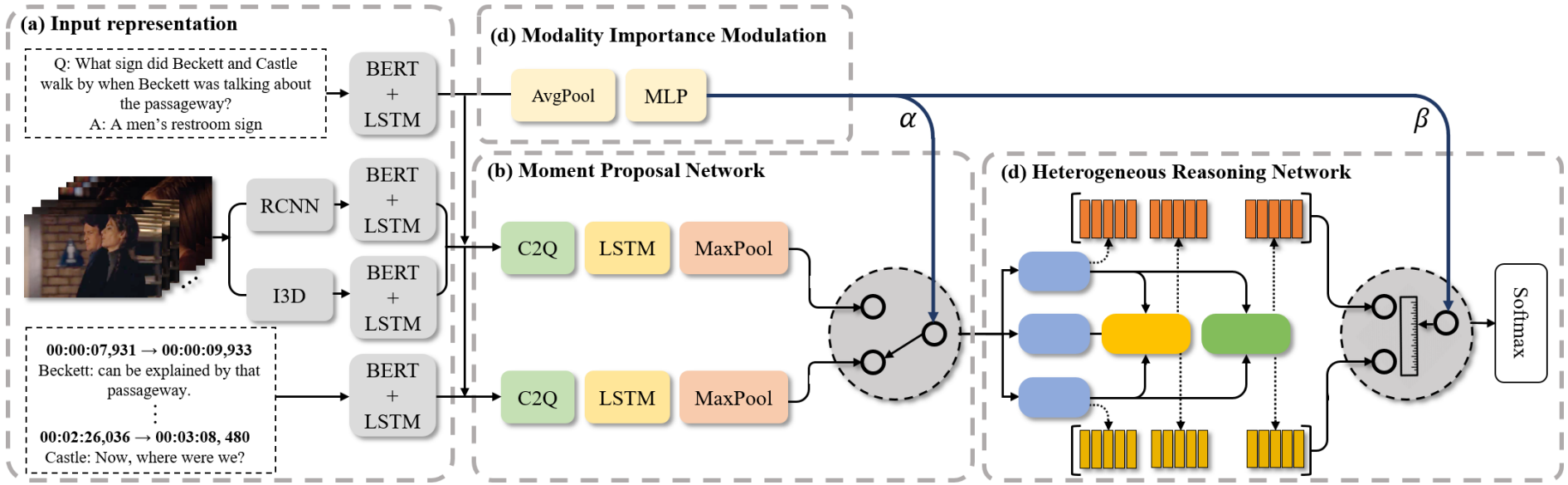
It aims to solve the multiple-choice question for a given multi-modal video.

Motivation



Our main motivation is that **MVQA can be decomposed** into (1) temporal localization, (2) answer prediction, and each sub-task might **require different modalities**.

Modality Shifting Attention Network (MSAN)



Modality Shifting Attention Network (MSAN) is proposed:

(1) Moment Proposal Network (MPN): localizes the key moment to answer the question

(2) Heterogeneous Reasoning network (HRN): applies heterogeneous attention to infer the answer

(3) Modality Importance Modulation (MIM): put weights on more important modality for each sub-task

Experiments

Table 3. Comparison with the state-of-the-art method on TVQA dataset. "img" is imagenet feature, "reg" is regional feature and "vcpt" is visual concept feature and "acpt"is action concept feature.

Methods	Text Feat.	Video Feat.	test Acc.
two-stream [19]	GloVe	img	63.44
		reg	63.06
		vcpt	66.46
PAMN [14]	GloVe	img	64.61
		vcpt	66.77
MTL [13]	GloVe	img	64.53
		vcpt	67.05
ZGF	-	-	68.90
STAGE [20]	BERT	reg	70.23
MSAN	GloVe	vcpt	68.18
		vcpt	70.92
MSAN	BERT	acpt	68.57
		vcpt+acpt	**71.13**

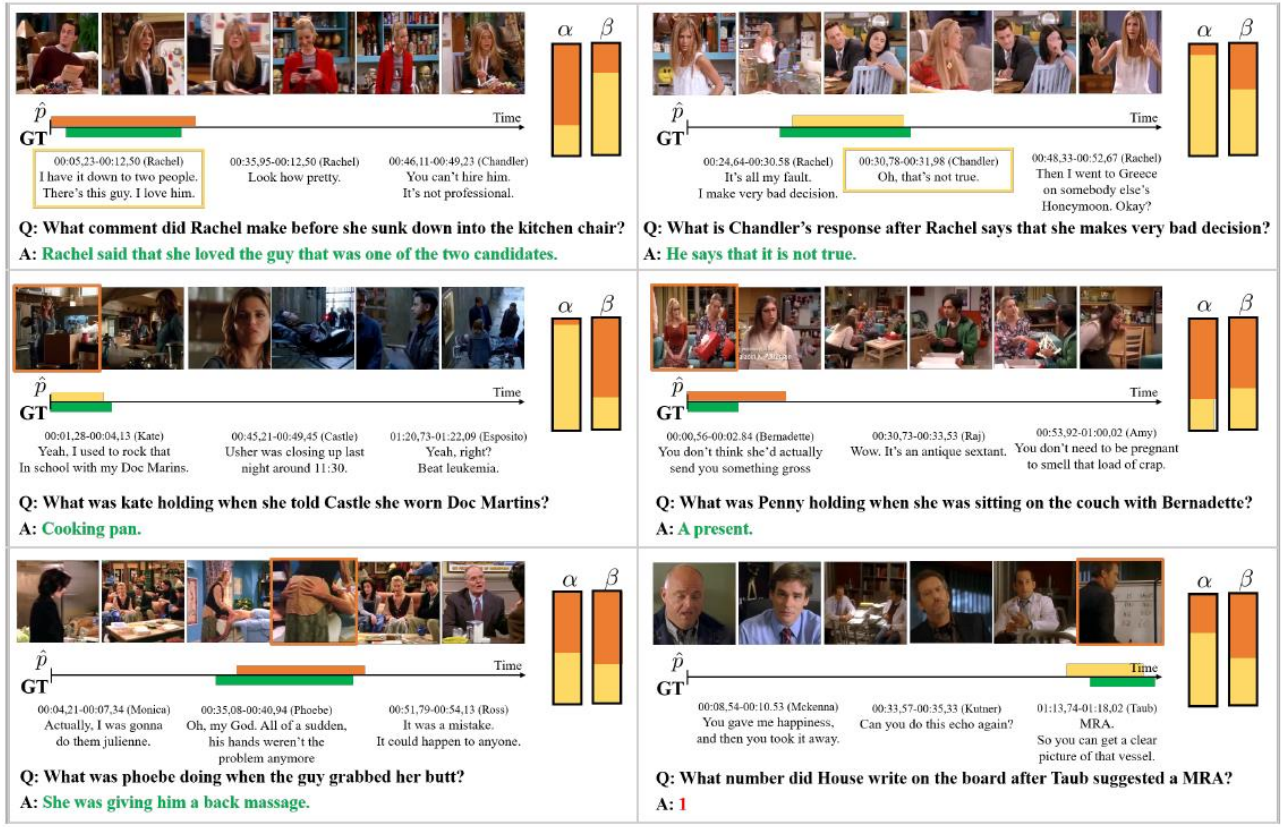


Figure 6. Visualization on the inference path of MSAN (the last example is a failure case). Each example provides MIM weights, the localized temporal moment $\hat{p}$ and ground-truth (GT) temporal moment. Video and subtitle modality are represented with orange and yellow color, respectively. The proposed MSAN dynamically modulates both modalities according to the input question.

Our proposed MSAN shows SOTA result on TVQA benchmark