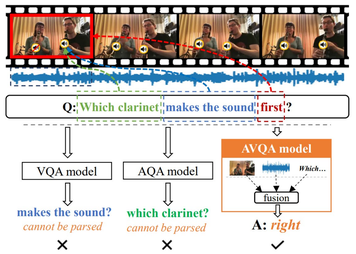
Learning to Answer Questions in Dynamic Audio-Visual Scenarios

Guangyao Li[1,†], Yake Wei[1,†], Yapeng Tian[2,†], Di Hu[1,*], Chenliang Xu[2], Ji-Rong Wen[1]

[1]Renmin University of China [2]University of Rochester

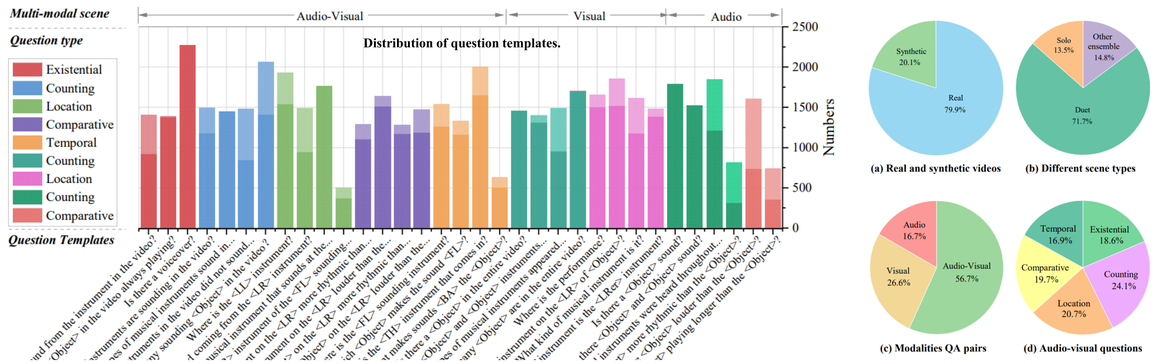
CVPR JUNE 19-24 2022

Motivation

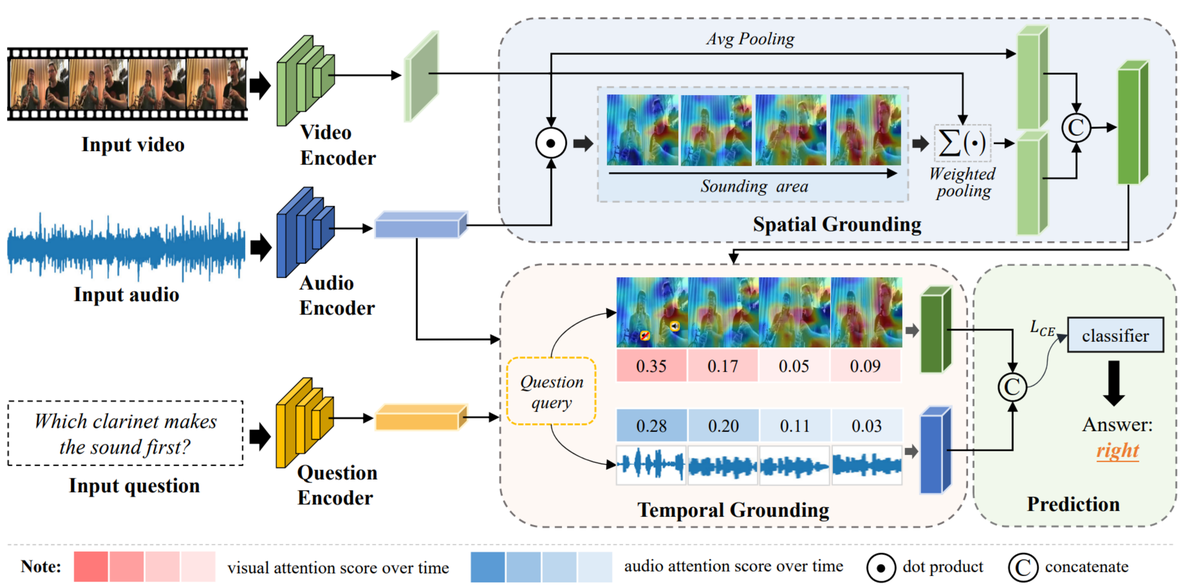


- How to make machines integrate multimodal information, especially the natural modality such as the audio and visual ones.
- Most methods remains limited ability for **cross-modal reasoning**, under complex audio-visual scenarios.
- The Audio-Visual Question Answering (AVQA) task, which aims to answer questions regarding different visual objects, sounds, and their associations in videos.

ST-AVQA Dataset



Framework



Evaluation Results

Method	A Question	V Question	A-V Question	All
Q	65.19	44.42	55.15	54.09
A+Q	67.78	62.75	63.86	64.26
V+Q	68.76	67.28	63.23	65.28
AV+Q	70.67	69.72	65.84	67.72
AV+Q+TG	73.01	73.18	68.02	70.27
AV+Q+TG+SG	74.06	74.00	69.54	71.52

* TG: Temporal Grounding; SG: Spatial Grounding.

Table1. Ablation study on input modalities and the proposed modules.

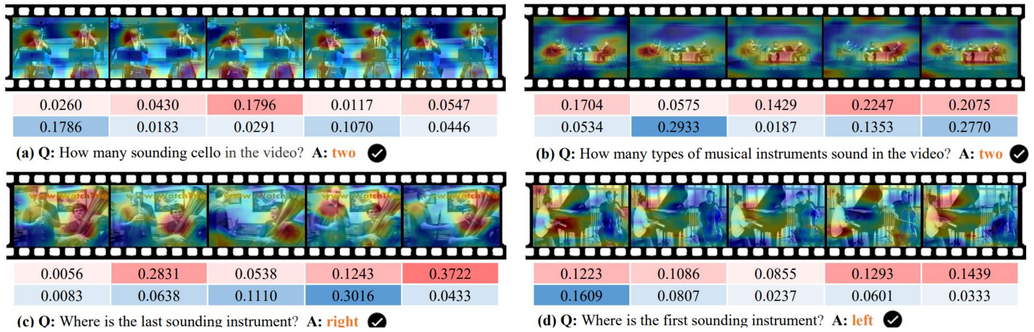
Method	Accuracy
Q-only*	43.50
Q+V*	41.70
Q+V+A*	41.45
Q+V†	42.01
Q+V+A†	41.95

Table2. Experiments on TVQA dataset

Task	Method	Audio Question			Visual Question			Audio-Visual Question						All
		Counting	Comparative	Avg.	Counting	Location	Avg.	Existential	Location	Counting	Comparative	Temporal	Avg.	Avg.
AudioQA	FCNLSTM [7]	70.45	66.22	68.88	63.89	46.74	55.21	82.01	46.28	59.34	62.15	47.33	60.06	60.34
	CONVLSTM [7]	74.07	**68.89**	72.15	67.47	54.56	60.94	**82.91**	50.81	63.03	60.27	51.58	62.24	63.65
VisualQA	GRU [3]	72.21	66.89	70.24	67.72	70.11	68.93	81.71	59.44	62.64	61.88	60.07	65.18	67.07
	BiLSTM Attn [53]	70.35	47.92	62.05	64.64	64.33	64.48	78.39	45.85	56.91	53.09	49.76	57.10	59.92
	HCAttn [26]	70.25	54.91	64.57	64.05	66.37	65.22	79.10	49.51	59.97	55.25	56.43	60.19	62.30
	MCAN [46]	77.50	55.24	69.25	71.56	70.93	71.24	80.40	54.48	64.91	57.22	47.57	61.58	65.49
VideoQA	PSAC [25]	75.64	66.06	72.09	68.64	69.79	69.22	77.59	55.02	63.42	61.17	59.47	63.52	66.54
	HME [6]	74.76	63.56	70.61	67.97	69.46	68.76	80.30	53.18	63.19	62..69	59.83	64.05	66.45
	HCRN [23]	68.59	50.92	62.05	64.39	61.81	63.08	54.47	41.53	53.38	52.11	47.69	50.26	55.73
AVQA	AVSD [30]	72.41	61.90	68.52	67.39	74.19	70.83	81.61	58.79	63.89	61.52	61.41	65.49	67.44
	Pano-AVQA [47]	74.36	64.56	70.73	69.39	75.65	72.56	81.21	59.33	64.91	64.22	63.23	66.64	68.93
	Our method	**78.18**	67.05	**74.06**	**71.56**	**76.38**	**74.00**	81.81	**64.51**	**70.80**	**66.01**	**63.23**	**69.54**	**71.52**

Table 3. Audio-visual video question answering results of different methods on the test set of ST-AVQA. The top-2 results are highlighted.

Visualization



The sounding area and key timestamps are accordingly highlighted in spatial and temporal perspectives

Conclusion

- We build the large-scale ST-AVQA dataset of musical performance, which contains more than 9K videos annotated by over 45K QA pairs, spanning over different modal scenes.
- A spatio-temporal grounding model is proposed to solve the fine-grained scene understanding and reasoning over audio and visual modalities.
- Extensive experiments show that AVQA benefits from multisensory perception and our model is superior to recent QA approaches especially on the questions that measures spatio-temporal reasoning ability of models.

二维码

中国人民大学高瓴人工智能学院 Gaoling School of Artificial Intelligence, Renmin University of China

UNIVERSITY of ROCHESTER

★ This poster is made by Guangyao Li

E-Mail: guangyaoli@ruc.edu.cn **Phone/WeChat:** +86 18810990726