Learning Trimodal Relation for Audio-Visual Question Answering with Missing Modality – *Supplementary Material* –

Kyu Ri Park[1], Hong Joo Lee[2,3†], and Jung Uk Kim[1†]

[1] Kyung Hee University, Yong-in, South Korea
{kyuri0924, ju.kim}@khu.ac.kr
[2] Technical University of Munich, Munich, Germany,
[3] Munich Center for Machine Learning (MCML), Munich, Germany
hongjoo.lee@tum.de

Table 1: Results on MUSIC-AVQA dataset under full modalities ($\mathcal{V}$: visual, $\mathcal{A}$: audio, $\mathcal{Q}$: question).

Method	Scenario	Audio Question			Visual Question			Audio-Visual Question						All
		Cnt.	Comp	Avg	Cnt.	Loc	Avg	Exist	Loc	Cnt.	Comp	Temp	Avg	Avg
AVSD [2]	Input: $\mathcal{A}+\mathcal{V}+\mathcal{Q}$	76.71	65.22	72.46	72.48	**76.71**	74.62	**81.31**	**71.11**	61.81	62.96	**64.81**	**68.63**	70.89
AVSD+Ours		**80.24**	**67.17**	**75.42**	**75.69**	74.69	**75.19**	80.26	70.51	**62.28**	**64.49**	61.68	68.19	**71.32**
Pano-AVQA [3]		77.79	64.89	73.01	73.48	73.62	73.55	**81.71**	71.90	60.41	63.14	**64.20**	68.59	70.68
Pano-AVQA+Ours		**81.91**	**65.49**	**75.85**	**77.78**	**74.61**	**76.18**	80.67	**72.65**	**62.39**	**64.94**	63.02	**69.13**	**72.19**
AVST [1]		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
AVST+Ours		**80.33**	**68.69**	**76.04**	**76.94**	75.35	**76.14**	80.97	**71.62**	63.48	64.03	62.17	68.80	**72.02**

1 Our Methods with Full Modalities

In the main paper, our primary focus was on evaluating the effectiveness of our method in addressing missing modality scenarios. However, considering the ultimate goal of our approach, which is to enhance feature extraction, it is equally essential to evaluate its performance in contexts where all inputs are available. To check this aspect, we investigated whether our method yields improved accuracy when applied to Audio-Visual Question Answering (AVQA) tasks with complete inputs.

Table 1 shows the comparison between our method and the baseline of AVQA models. The results show that our method consistently outperforms the baseline, indicating its ability to extract superior features even in scenarios where all inputs are complete. This observation emphasizes the effectiveness of our approach not only in handling missing modality scenarios but also in situations where all modalities are complete. Leveraging trimodal knowledge, our method demonstrates superior performance in diverse audio-visual scenes, highlighting its robustness and flexibility.

Consequently, our approach emerges as an effective AVQA method capable of delivering accurate answers consistently, both in missing data scenarios and non-missing situations.

† Corresponding author

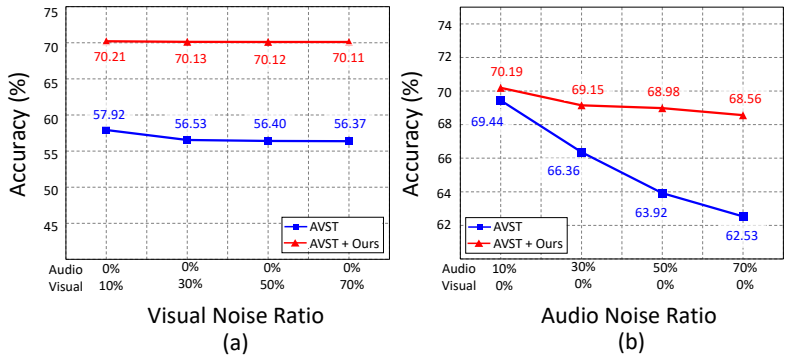

Fig. 1: AVQA results on the MUSIC-AVQA dataset vary based on the noise ratio of (a) visual and (b) audio modalities

2 Our Methods with Noisy Modalities

We also evaluated the effectiveness of our method in noisy environments that closely simulate real-world scenarios. While it is rarely possible to encounter noiseless conditions in real-world situations, the presence of noise is highly probabilistic, resulting in data with added noise. In visual contexts, noise can come from various sources such as cloudy weather, rain, snow, and fog, while in audio scenarios, ambient sounds such as wind and rain can obscure the desired information. In consideration of these issues, we attempted to validate the robustness of our method in dealing with noisy input data.

As shown in Fig. 1, we compared the results obtained by introducing noise into the input of both the original AVST model and AVST with our proposed method. Notably, AVST with our method exhibits resilience to noisy conditions, as evidenced by the comparison. Therefore, our results demonstrate the effectiveness of our approach even when confronted with noisy inputs that closely resemble real-world situations.

Furthermore, we conducted a comparative analysis between AVST and our method under conditions where both audio and visual modalities are noisy (see Fig. 2). Through this comparison, we observed that the presence of noise in the audio modality has a more pronounced effect than noise in the visual modality. We demonstrated the effectiveness of our proposed method regardless of which modality is more affected by noise. By conducting these various noise experiments, we provide strong evidence for the effectiveness of our method in real-world scenarios.

3 Qualitative Results

In this section, we provide a visual comparison of the AVQA results obtained by both AVST [1] and AVST with our method (*i.e.,* 'AVST+Ours'). For each piece of audio-visual data, four QA pairs exist, and we took only two of them

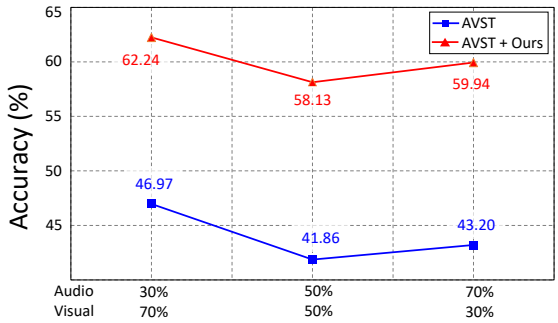

Fig. 2: AVQA results on the MUSIC-AVQA dataset based on the varying noise ratios for visual and audio modalities.

as examples, which are highly affected by the missing modal situation. Fig. 3 illustrates the examples of answers generated by AVST and 'AVST+Ours' on the MUSIC-AVQA dataset when the audio modality is missing. In addition, Fig. 4 shows the examples of answers generated by AVST and 'AVST+Ours' when the visual modality is missing. When a modality (audio or visual) is missing, as our method can utilize the trimodal knowledge to construct an augmented pseudo-modal feature that compensates the missing modal information. As a result, while AVST fails to provide accurate responses, our method adeptly provides the correct answers. These visualizations show the effectiveness of our method in handling missing modality scenarios by leveraging trimodal relations.

In conclusion, our method surpasses existing AVQA approaches across a spectrum of scenarios, including missing modalities, noisy inputs, and full modality contexts. This versatility demonstrates the robustness and effectiveness of our proposed method in addressing a wide range of real-world AVQA challenges.

References

1. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.R., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 19108–19118 (2022)
2. Schwartz, I., Schwing, A.G., Hazan, T.: A simple baseline for audio-visual scene-aware dialog. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 12548–12558 (2019)
3. Yun, H., Yu, Y., Yang, W., Lee, K., Kim, G.: Pano-avqa: Grounded audio-visual question answering on 360deg videos. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). pp. 2031–2041 (2021)

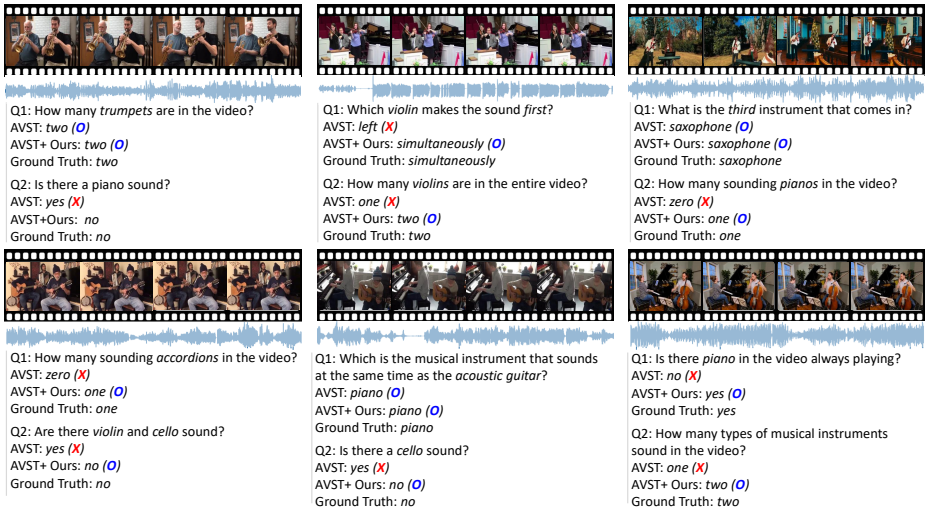

Fig. 3: Qualitative results comparison against AVST and ground-truth in audio missing situations on the MUSIC-AVQA dataset.

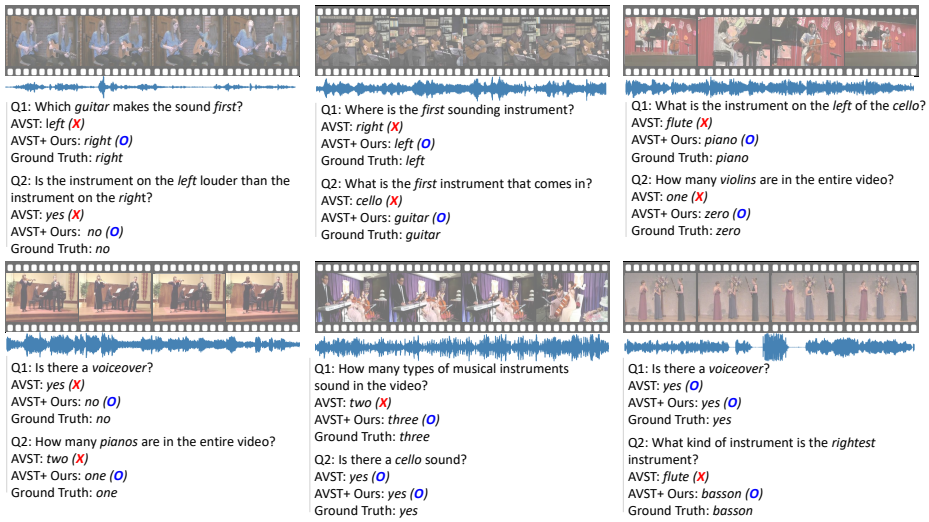

Fig. 4: Qualitative results comparison against AVST and ground-truth in visual missing situations on the MUSIC-AVQA dataset.