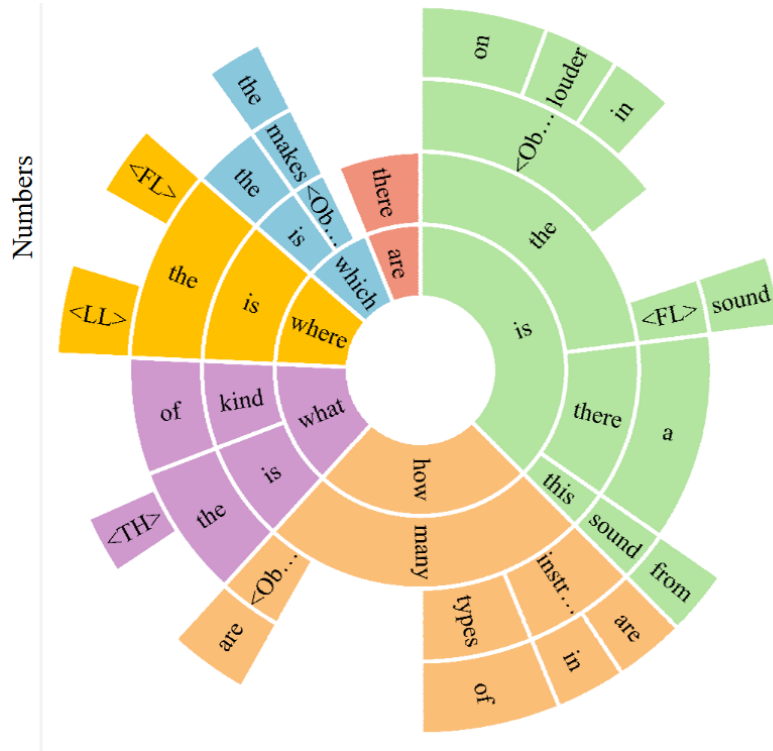


Method	MCCD	Audio QA			Visual QA			AVQA						All
		CNT	COMP	Avg.	CNT	LOC	Avg.	EXIST	LOC	CNT	COMP	TEMP	Avg.	Avg.
FCNLSTM	×	70.45	66.22	68.88	63.89	46.74	55.21	82.01	46.28	59.34	62.15	47.33	60.06	60.34
	✓	70.99	66.50	69.34	66.08	59.02	62.51	83.50	57.17	60.47	61.58	57.54	64.11	64.61
CONVLSTM	×	74.07	68.89	72.15	67.47	54.56	60.94	82.91	50.81	63.03	60.27	51.58	62.24	63.65
	✓	72.76	69.53	71.57	69.59	58.12	63.79	82.69	56.09	62.13	62.03	55.11	63.87	65.21
BiLSTM Attn	×	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
	✓	68.24	54.88	63.31	61.65	55.92	58.75	79.15	41.96	55.02	49.41	49.15	55.18	57.56
HCAtn	×	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
	✓	69.52	53.37	63.56	63.99	65.47	64.74	78.64	47.28	61.11	55.86	55.72	60.03	61.90
MCAN	×	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
	✓	78.27	56.57	70.27	71.93	71.18	71.55	81.48	54.24	65.77	55.86	46.84	61.54	65.74
GRU	×	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
	✓	73.35	66.16	70.70	67.25	71.43	69.36	81.98	60.11	63.08	62.76	61.19	65.84	67.63
HCRN	×	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
	✓	72.17	64.65	69.40	67.42	60.82	64.08	79.66	48.70	65.14	61.22	55.72	60.40	64.20
HME	×	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
	✓	72.17	64.65	69.40	67.42	60.82	64.08	79.66	48.70	65.14	61.22	55.72	60.40	64.20
PSAC	×	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
	✓	75.02	65.66	71.57	69.09	69.88	69.49	79.35	53.04	61.98	61.13	57.66	62.85	66.15
AVSD	×	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
	✓	72.07	63.97	69.09	67.42	74.53	71.02	81.17	59.13	63.08	62.49	63.50	65.82	67.77
LAViT	×	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
	✓	75.12	65.49	71.57	70.43	76.73	73.62	81.38	60.33	65.30	62.49	62.29	66.42	69.24

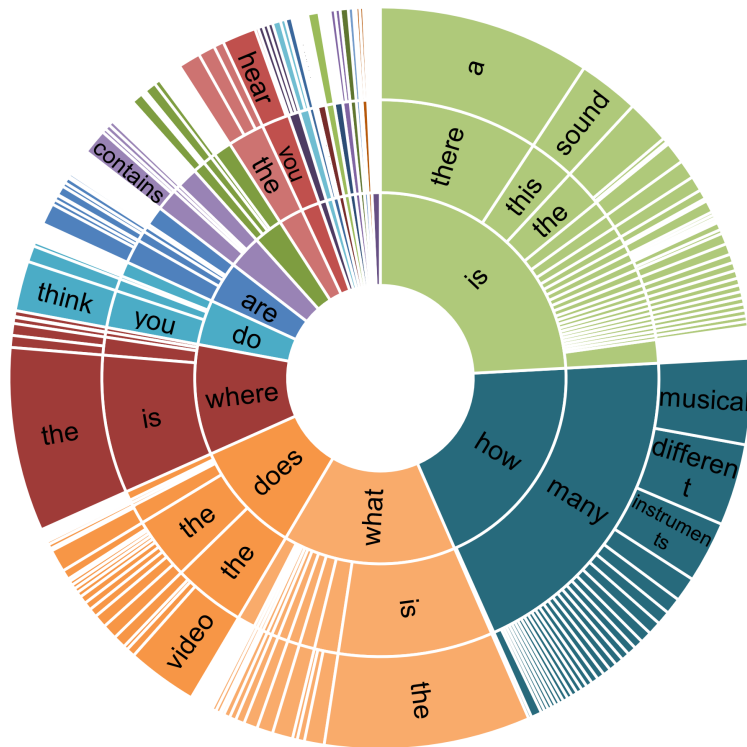
Table 1: Experimental results (%) on the MUSIC-AVQA test split. EXIST, LOC, CNT, COMP, and TEMP, which are question types, denote “Existential”, “Location”, “Counting”, “Comparative”, and “Temporal”, respectively. Avg. denotes the average accuracy.

Method	Audio QA				Visual QA				AVQA								All		
	CNT		COMP		CNT		LOC		EXIST		LOC		CNT		COMP		TEMP		Avg.
	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H	T			
FCNLSTM + MCCD	66.23	36.48	64.78	51.14	61.75	5.31	54.86	51.06	64.76	78.52	46.66	57.30	62.69	7.23	43.13	71.67	37.02	30.78	54.12
	62.51	34.44	61.19	51.26	61.11	5.66	57.73	50.36	62.48	82.40	45.49	60.09	62.07	7.16	44.55	69.46	36.55	30.74	54.55
CONVLSTM + MCCD	70.22	41.14	67.50	52.93	62.11	9.17	53.44	49.88	60.08	84.82	46.46	59.90	56.52	8.18	43.29	72.52	41.54	45.12	55.20
	68.38	41.58	68.39	52.10	61.46	9.56	54.17	50.33	59.61	83.11	55.29	56.52	59.13	7.82	45.31	72.70	41.26	45.40	55.74
BiLSTM Attn + MCCD	73.68	46.32	21.51	77.58	64.30	0.00	53.92	42.01	87.51	21.14	35.16	43.75	62.85	2.18	27.61	74.38	17.58	31.32	48.84
	73.30	45.16	20.71	77.48	64.41	0.00	56.08	42.54	87.47	21.04	34.47	43.51	63.33	2.18	26.01	75.48	17.92	32.67	49.55
HCAtn + MCCD	61.67	41.63	59.09	47.14	56.52	9.20	67.01	53.16	66.57	61.13	37.05	42.48	59.53	12.48	48.81	60.12	33.82	39.26	51.90
	62.50	41.43	58.89	47.42	56.65	8.85	67.31	52.92	66.82	59.87	38.25	42.53	59.38	12.42	57.39	52.01	32.84	39.55	52.29
GRU + MCCD	66.92	48.63	58.29	59.61	64.37	11.79	57.68	57.66	76.30	64.76	41.05	45.61	60.71	18.68	57.19	57.38	31.02	40.67	55.21
	69.94	48.09	56.31	63.77	66.24	13.36	63.55	57.59	83.04	54.16	43.36	43.36	57.89	18.36	53.93	59.65	30.82	38.23	55.70
MCAN + MCCD	75.02	60.16	58.89	50.09	64.58	26.69	66.48	62.25	51.29	67.29	46.11	61.61	64.76	25.28	50.57	52.40	34.64	58.05	57.27
	73.53	56.14	68.31	39.44	65.51	29.40	68.41	60.09	58.8	61.90	46.75	60.61	60.54	31.89	69.09	44.94	32.44	57.78	58.22
HCRN + MCCD	55.53	53.31	47.17	32.44	41.87	23.55	39.40	51.27	41.81	65.45	36.62	42.72	54.58	19.57	33.33	36.87	40.47	44.13	43.92
	51.96	49.21	43.42	36.78	41.13	20.71	37.79	50.99	44.38	58.40	35.05	46.33	54.39	20.90	34.50	33.14	40.13	44.00	42.87
HME + MCCD	62.60	53.95	54.97	58.29	50.95	16.46	73.25	58.60	65.74	66.49	33.79	46.03	63.18	17.18	53.20	60.57	33.95	41.57	53.66
	60.62	53.58	62.22	53.01	52.90	14.96	72.56	58.56	55.47	69.21	32.27	42.97	69.90	12.36	43.51	72.51	36.65	32.61	53.44
PSAC + MCCD	53.01	56.68	57.41	48.12	49.55	26.43	72.96	60.69	50.56	55.54	41.98	52.30	56.70	19.58	38.13	58.92	26.68	46.24	50.45
	55.14	52.26	64.70	44.45	52.34	22.15	72.06	60.70	58.97	52.35	41.18	49.78	53.28	18.85	42.60	64.53	25.81	45.68	51.64
AVSD + MCCD	54.00	47.84	60.61	47.79	60.34	10.07	74.78	61.43	66.28	61.98	33.00	40.35	46.21	8.06	51.98	66.00	40.14	41.52	52.33
	55.87	40.18	65.41	48.05	63.32	7.41	73.78	58.20	74.74	70.80	37.85	34.55	35.53	6.11	49.96	67.88	44.03	43.89	53.09
LAViT + MCCD	50.57	43.45	50.78	44.93	47.28	15.50	67.19	65.51	52.37	22.04	44.35	61.69	52.21	21.52	45.61	40.49	35.00	49.33	47.40
	45.05	45.09	57.33	41.26	48.62	17.00	69.91	65.90	60.61	29.57	43.17	57.57	53.92	22.09	54.46	35.35	33.99	49.40	48.91

Table 2: Experimental results (%) on the MUSIC-AVQA-R test split. The question types, such as CNT and COMP, are introduced in Table 1. H and T denote the head and tail accuracy.



(a) MUSIC-AVQA(Li et al., 2022)



(b) MUSIC-AVQA-R(Ours)

Figure 1: Distribution of questions based on the first few words.

References

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. [Learning to answer questions in dynamic audio-visual scenarios](#). In *CVPR*, pages 19108–19118.