



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



UNIVERSITÀ
DI PISA

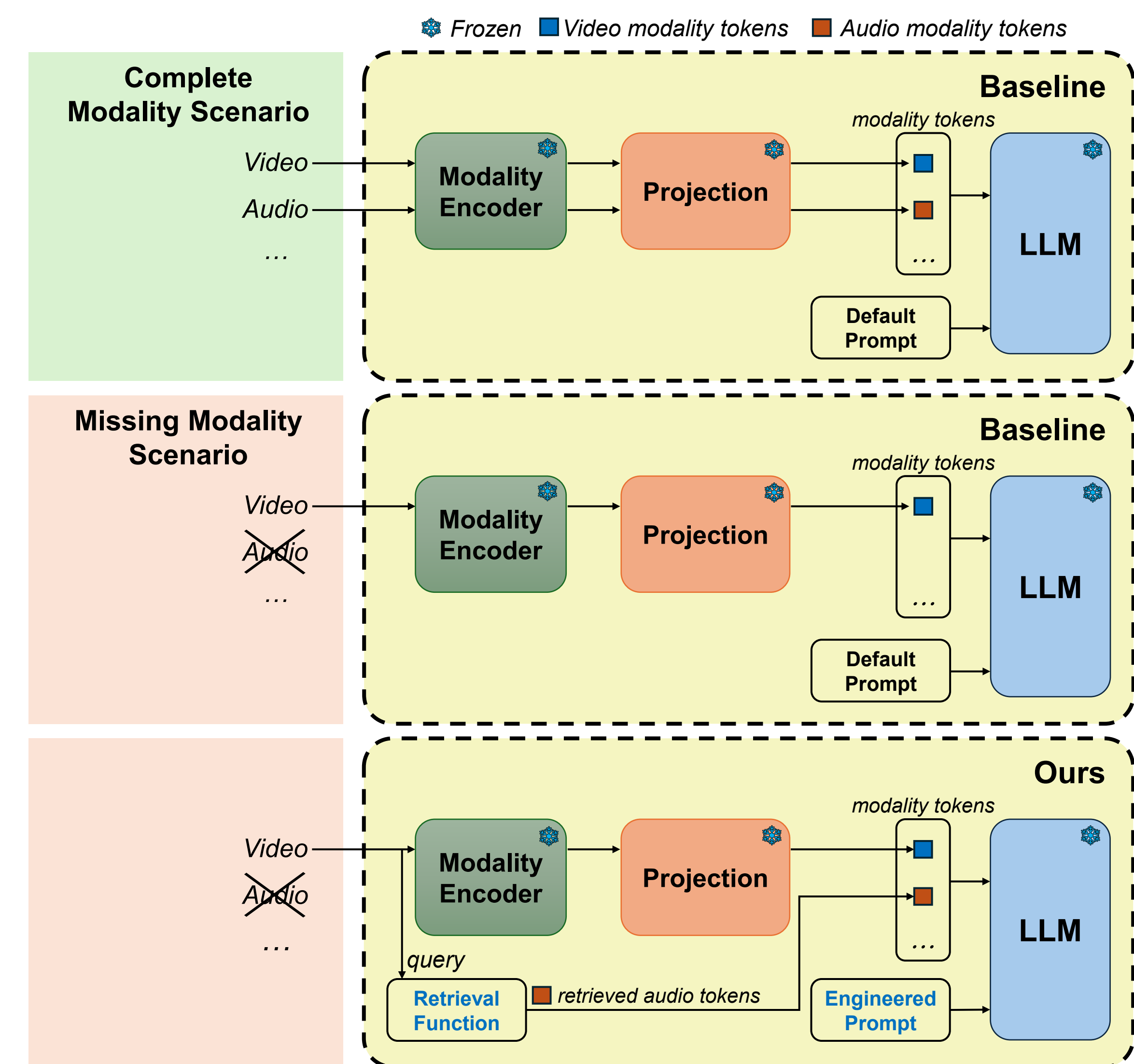


UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Missing Modality Problem with MLLMs

- In real-world scenarios, data samples often have **one or more absent or incomplete modalities**.

In this work, we are the first to assess the robustness of MLLMs under missing modality conditions.



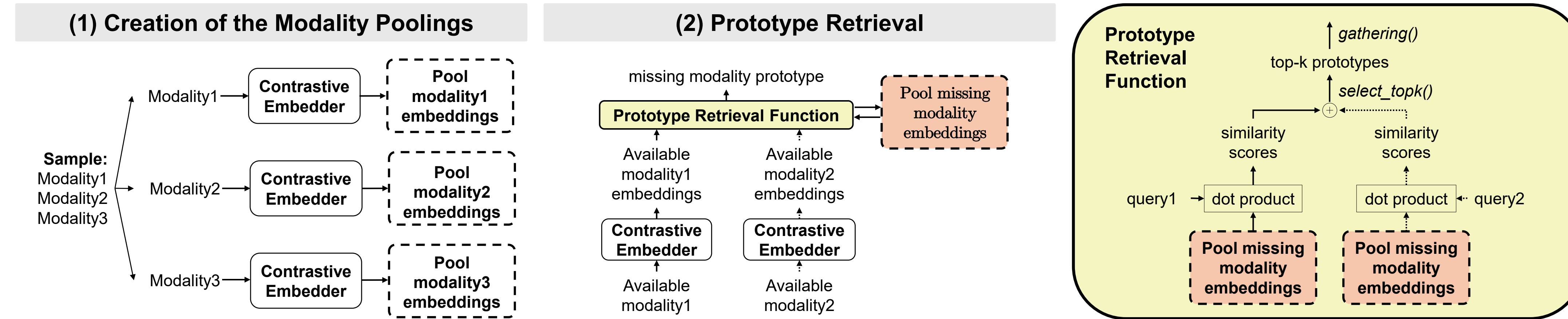
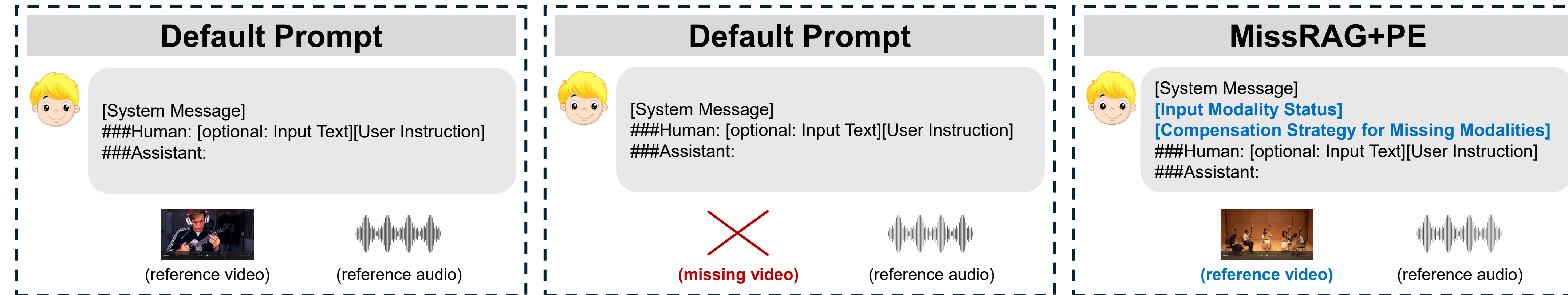
- We propose the first multimodal RAG framework to address the missing modality problem with MLLMs while keeping their parameters frozen!

MissRAG: Addressing the Missing Modality Challenge in Multimodal Large Language Models

Vittorio Pipoli^{*1,2}, Alessia Saporita^{*1,3}, Federico Bolelli¹, Marcella Cornia¹, Lorenzo Baraldi¹, Costantino Grana¹, Rita Cucchiara¹, Elisa Ficarra¹

¹University of Modena and Reggio Emilia, Italy; ²University of Pisa, Italy; ³University of Bologna, Italy
{name.surname}@unimore.it; ^{*}Equal Contribution

Our Proposal: MissRAG



Experimental Results

- MissRAG effectively mitigates the missing modality problem with MLLM across a wide range of tasks

Task	Audio-Video			Audio-Video						Audio-Video-Text													
	Question Answering			Captioning						Sentiment Analysis													
Dataset	MUSIC-AVQA			VALOR32K			CharadesEgo			MOSI							MOSEI						
Metric	Accuracy			CIDEr			CIDEr			Accuracy							Accuracy						
Method	MA	MV	C	MA	MV	C	MA	MV	C	MA	MV	MT	MAV	MAT	MVT	C	MA	MV	MT	MAV	MAT	MVT	C
ChatBridge	41.80	27.25	43.87	18.36	4.30	20.22	9.47	1.51	9.96	18.51	29.01	17.49	23.18	15.16	11.37	26.38	40.20	21.76	35.20	28.38	39.02	21.76	36.81
MissRAG+PE	45.14	43.18	46.10	19.09	10.20	20.66	10.85	4.37	10.25	24.78	30.09	18.22	23.76	16.18	17.93	28.12	41.21	40.95	37.05	40.46	36.19	37.33	38.94
Improvement	+3.34	+15.93	+2.23	+0.73	+5.90	+0.44	+1.38	+2.86	+0.29	+6.27	+1.08	+0.73	+0.58	+1.02	+6.56	+1.74	+1.01	+19.19	+1.85	+12.08	-2.83	+15.57	+2.13
OneLLM	48.89	39.48	49.86	20.41	3.34	26.40	2.98	1.20	3.76	48.54	73.91	40.23	75.88	38.34	41.69	56.41	30.82	57.70	40.22	58.75	38.14	46.10	40.09
MissRAG+PE	50.85	46.66	51.32	27.19	16.57	26.42	4.03	2.50	4.17	78.57	74.20	46.65	76.68	54.08	45.48	75.51	59.18	51.66	45.16	52.35	45.89	47.66	61.56
Improvement	+1.96	+7.18	+1.46	+6.78	+13.23	+0.02	+1.05	+1.30	+0.41	+30.03	+0.29	+6.42	+0.80	+15.74	+3.79	+19.10	+28.36	-6.04	+4.94	-6.40	+7.75	+1.56	+21.47
VideoLLaMA 2	77.20	59.57	79.94	19.36	7.02	22.40	10.17	0.69	12.20	46.06	43.59	10.50	74.34	10.20	4.37	25.80	41.36	44.24	28.55	53.17	28.07	22.00	36.62
MissRAG+PE	77.53	60.11	79.97	21.42	10.30	22.64	11.96	3.29	12.62	33.38	37.76	23.76	40.52	11.02	23.34	34.69	38.63	39.47	29.20	40.14	30.31	27.15	39.17
Improvement	+0.33	+0.54	+0.03	+2.06	+3.28	+0.24	+1.79	+2.60	+0.42	-12.68	-6.23	+13.26	-33.82	+082	+18.97	+8.89	-2.73	-4.77	+0.65	-13.03	+2.24	+5.15	+2.55



Ablation Studies

- Ablation study on the multimodal RAG system

Task	Audio-Video Question Answering			Audio-Video Captioning						Audio-Video-Text Sentiment Analysis													
	MUSIC-AVQA Accuracy			VALOR32K CIDEr			CharadesEgo CIDEr			MOSI Accuracy						MOSEI Accuracy							
Method	MA	MV	C	MA	MV	C	MA	MV	C	MA	MV	MT	MAV	MAT	MVT	C	MA	MV	MT	MAV	MAT	MVT	C
OneLLM	48.89	39.48	49.86	20.41	3.34	26.40	2.98	1.20	3.76	48.54	73.91	40.23	75.88	38.34	41.69	56.41	30.82	57.70	40.22	58.75	38.14	46.10	40.09
PE	51.72	41.81	51.32	23.79	3.77	26.42	3.25	1.22	4.17	68.66	79.74	44.90	76.11	39.48	42.72	75.51	48.12	64.58	49.43	47.69	31.79	41.73	61.56
MissRAG (k = 1)	49.61	45.55	-	25.00	15.83	-	3.93	2.2	-	65.60	63.41	39.50	56.71	49.97	38.78	-	42.20	36.34	28.89	31.53	26.23	22.04	-
MissRAG (k = 3)	49.67	45.69	-	22.87	17.13	-	3.99	2.17	-	63.70	63.85	46.79	53.35	40.96	44.46	-	41.43	31.70	26.89	26.74	31.36	22.84	-
MissRAG (k = 5)	49.50	46.27	-	21.01	17.16	-	4.22	2.14	-	62.68	60.50	47.52	51.60	49.56	48.98	-	41.34	30.78	24.55	25.80	31.96	21.59	-
MissRAG+PE (k = 1)	50.85	46.66	51.32	27.19	16.57	26.42	4.03	2.50	4.17	76.09	74.49	47.81	76.68	54.08	45.48	75.51	59.18	51.66	46.16	52.35	45.89	47.66	61.56
MissRAG+PE (k = 3)	50.86	46.80	51.32	24.86	17.34	26.42	4.00	2.31	4.17	76.38	74.05	48.69	72.01	49.71	51.02	75.51	57.67	49.04	41.21	43.79	44.47	37.48	61.56
MissRAG+PE (k = 5)	50.83	47.29	51.32	23.50	17.49	26.42	4.25	2.60	4.17	76.53	72.74	49.98	70.85	49.42	50.73	75.51	57.07	47.91	37.33	42.91	41.83	31.90	61.56
MissRAG+PE (k=†)	50.71	46.57	51.32	26.73	16.37	26.42	3.97	2.46	4.17	78.43	73.85	48.25	72.68	54.96	45.34	75.51	55.42	49.33	48.89	52.16	45.78	46.51	61.56

Ablation on the impact of the number k of retrieved prototypes in MissRAG conducted on OneLLM. The symbol † denotes that the prototype pool is constructed by merging the prototype pools from all datasets.

- Ablation studies on the prompt engineering strategy

Task	Audio-Video Question Answering			Audio-Video Captioning			Task	Audio-Video Question Answering			Audio-Video Captioning		
	Dataset	MUSIC-AVQA			VALOR32K	Dataset		MUSIC-AVQA			VALOR32K		
		Metric	Accuracy					CIDEr	Metric	Accuracy		CIDEr	
Method	MA	MV	C	MA	MV	C	Method	MA	MV	C	MA	MV	C
OneLLM	48.89	39.48	49.86	20.41	3.34	26.40	OneLLM	48.89	39.48	49.86	20.41	3.34	26.40
System Message	51.72	41.81	51.32	23.79	3.77	26.42	PE	51.72	41.81	51.32	23.79	3.77	26.42
User Instruction	48.68	41.02	49.27	23.00	3.58	27.30	PE w/o CSMM	49.68	40.79	51.01	23.06	3.69	26.22

Ablation on the system message vs. user instruction.

Ablation on the prompt engineering compensation strategy.

Qualitative Examples

