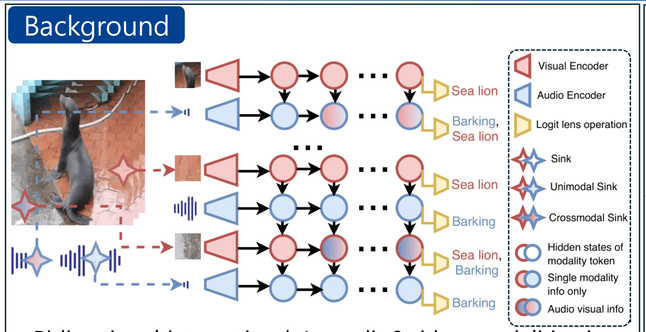
Probing Cross-modal Information Hubs in Audio-Visual LLM

Jihoo Jung [1], Chaeyoung Jung [1], Ji-Hoon Kim [2], Joon Son Chung[1]
[1]KAIST, [2]Chung-Ang University

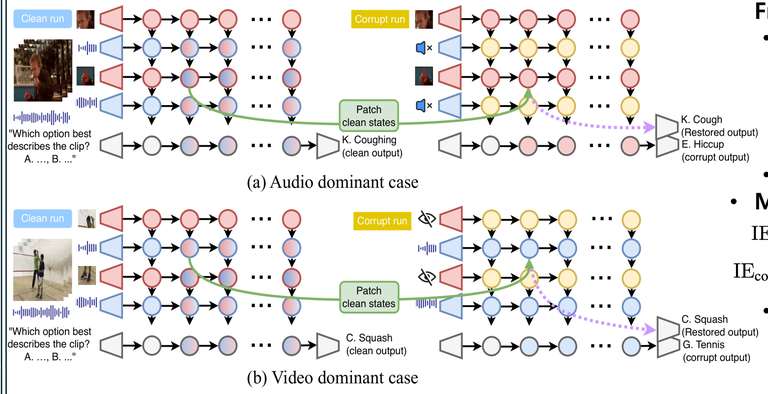
ICML International Conference On Machine Learning

Background



- Bidirectional interaction b/w audio&video modalities in AVLLMs.
- Where is information derived from one modality (audio or visual) is stored within the token of the other modality?

Unimodal Dominance Framework



- **Unimodal Dominance Framework**
  - A single modality governs the model's output by providing decisive cues, while its counterpart remains ambiguous.
  - Audio/Video Dominant case
- **Metrics for Causal Tracing**

$$\mathrm{IE}_{\mathrm{clean}}(S) = P_{h_S^{\mathrm{clean}}}[o_{\mathrm{clean}}] - P[o_{\mathrm{clean}}]$$

$$\mathrm{IE}_{\mathrm{corrupt}}(S) = P[o_{\mathrm{corrupt}}] - P_{h_S^{\mathrm{clean}}}[o_{\mathrm{corrupt}}]$$

- High -> token S acts as a critical repository of cross-modal information.

Where Is Cross-modal Information Located?

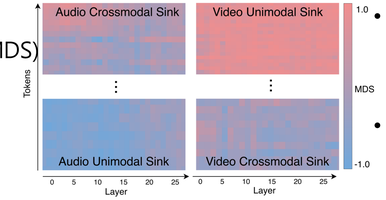
Modality	Ablation	Qwen2.5-Omni(7B)			Qwen2.5-Omni(3B)			video-SALMONN-o1(7B)			video-SALMONN2+(7B)			video-SALMONN2+(3B)		
		IE_{clean} ↑	IE_{corr} ↑	#Tokens	IE_{clean} ↑	IE_{corr} ↑	#Tokens	IE_{clean} ↑	IE_{corr} ↑	#Tokens	IE_{clean} ↑	IE_{corr} ↑	#Tokens	IE_{clean} ↑	IE_{corr} ↑	#Tokens
Audio Dominant	All	9.61	5.28	1440	7.83	3.48	1440	35.55	33.18	1820	6.45	5.27	1210	1.92	2.15	1210
	Object	5.04	2.44	613	3.53	1.12	580	16.22	15.06	852	3.78	3.93	500	0.72	1.16	447
	Sink (N=2)	**6.24**	**2.94**	603	**6.99**	**2.70**	605	**25.33**	**22.73**	818	**4.79**	**4.20**	565	**1.33**	**1.38**	506
	Sink (N=3)	4.31	1.94	362	6.36	2.08	354	21.42	19.67	514	3.73	3.49	360	0.93	0.94	297
	Sink (N=4)	3.26	1.23	256	5.50	1.64	243	19.10	17.79	364	3.23	3.33	256	0.69	0.65	195
	Random (N=2)	4.24	2.37	603	4.05	1.20	605	20.43	18.11	818	4.21	4.01	565	1.09	0.95	506
	Random (N=3)	2.97	1.55	362	2.71	0.72	354	12.65	13.54	514	3.12	3.51	360	0.67	0.74	297
	Random (N=4)	1.93	0.87	256	1.87	0.65	243	8.70	8.77	364	3.02	3.33	256	0.66	0.53	195
Video Dominant	All	8.21	13.63	249	2.43	8.85	249	3.63	4.08	153	0.46	1.86	60	-0.05	-0.04	60
	Object	4.97	8.44	149	1.59	6.41	149	2.07	0.40	78	0.22	1.71	7	-0.01	-0.06	7
	Sink (N=2)	**5.47**	**8.54**	144	**2.07**	**6.87**	147	**3.57**	**3.87**	117	**0.28**	**2.24**	9	-0.02	0.00	29
	Sink (N=3)	4.40	7.12	86	1.62	5.88	109	3.45	3.66	76	0.08	1.70	3	-0.03	-0.06	16
	Sink (N=4)	3.10	6.28	60	1.10	4.78	85	3.30	3.28	52	0.06	1.29	2	-0.01	**0.07**	13
	Random (N=2)	4.56	6.83	144	1.22	5.29	147	2.86	2.22	117	0.21	1.77	9	-0.02	0.01	29
	Random (N=3)	2.70	3.62	86	0.94	4.34	109	1.69	-0.44	76	0.12	1.28	3	0.00	0.02	16
	Random (N=4)	1.86	2.30	60	0.58	3.46	85	0.9	-2.52	52	0.21	1.28	2	-0.01	0.02	13

Finding 1 : Cross-modal information is primarily encoded in sink tokens.

Are Sink Tokens Homogeneous Cross-modal Information Holder?

- **Dissecting sink tokens**
  - Modality Dominance Score (MDS)

$$\mathrm{MDS}_i^l = \frac{\bar{A}^l_{\mathrm{video},i} - \bar{A}^l_{\mathrm{audio},i}}{\bar{A}^l_{\mathrm{video},i} + \bar{A}^l_{\mathrm{audio},i}}.$$



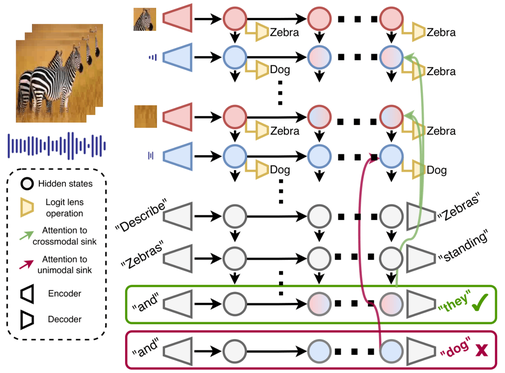
- **Unimodal sink tokens** receive attention mainly from their own modality.
- **Cross-modal sink tokens** receive attention mainly from the other modality.

Modality	Ablation	Qwen2.5-Omni(7B)			Qwen2.5-Omni(3B)			video-SALMONN-o1(7B)			video-SALMONN2+(7B)			video-SALMONN2+(3B)		
		IE_{clean} ↑	IE_{corr} ↑	#Tokens	IE_{clean} ↑	IE_{corr} ↑	#Tokens	IE_{clean} ↑	IE_{corr} ↑	#Tokens	IE_{clean} ↑	IE_{corr} ↑	#Tokens	IE_{clean} ↑	IE_{corr} ↑	#Tokens
Audio Dominant	Sink (N=2)	**6.24**	**2.94**	603	**6.99**	**2.70**	605	**25.33**	**22.73**	818	**4.79**	**4.20**	565	**1.33**	**1.38**	506
	Unimodal (N=2)	0.65	0.23	301	0.89	0.31	302	7.25	6.82	409	2.32	3.03	282	0.21	0.45	252
	Crossmodal (N=2)	5.58	2.95	301	6.57	2.33	302	21.30	19.93	409	4.16	3.69	282	1.27	1.14	252
	Sink (N=3)	**4.31**	**1.94**	362	**6.36**	**2.08**	354	**21.42**	**19.67**	514	**3.73**	**3.49**	360	**0.93**	**0.94**	297
	Unimodal (N=3)	0.92	0.39	181	1.02	0.18	177	7.02	7.03	257	2.06	2.87	180	0.19	0.42	148
	Crossmodal (N=3)	3.54	1.52	181	5.73	1.85	177	16.81	15.78	257	3.35	3.20	180	0.77	0.70	148
	Sink (N=4)	**3.26**	**1.23**	256	**5.50**	**1.64**	243	**19.10**	**17.79**	364	**3.23**	**3.33**	256	0.69	**0.65**	195
	Unimodal (N=4)	0.71	0.36	128	1.07	0.32	121	6.19	6.10	182	2.11	2.78	128	0.21	0.36	97
	Crossmodal (N=4)	2.70	0.99	128	4.90	1.28	121	14.24	13.95	182	2.82	3.05	128	**0.76**	0.64	97
Video Dominant	Sink (N=2)	**5.47**	**8.54**	144	**2.07**	**6.87**	147	**3.57**	**3.87**	117	**0.28**	**2.24**	9	-0.02	0.00	29
	Unimodal (N=2)	1.93	3.54	72	0.35	3.43	73	-0.01	-5.00	58	0.18	1.31	4	**0.00**	**0.03**	14
	Crossmodal (N=2)	3.03	4.53	72	1.25	4.48	73	3.53	3.72	58	0.26	2.19	4	-0.02	0.01	14
	Sink (N=3)	**4.40**	**7.12**	86	**1.62**	**5.88**	109	**3.45**	**3.66**	76	0.08	**1.70**	3	-0.03	-0.06	16
	Unimodal (N=3)	1.72	3.19	43	0.31	3.15	54	0.13	-4.57	38	0.08	1.44	1	0.00	**0.06**	8
	Crossmodal (N=3)	2.15	3.70	43	1.01	4.11	54	3.30	3.15	38	**0.20**	1.60	1	**0.01**	**0.06**	8
	Sink (N=4)	**3.10**	**6.28**	60	**1.10**	**4.78**	85	**3.30**	**3.28**	52	0.06	1.29	2	-0.01	**0.07**	13
	Unimodal (N=4)	1.27	2.80	30	0.24	2.77	42	0.18	-4.46	26	**0.08**	**1.33**	1	-0.02	-0.02	6
	Crossmodal (N=4)	1.45	3.02	30	0.63	3.57	42	3.00	2.56	26	0.07	1.25	1	**0.02**	0.04	6

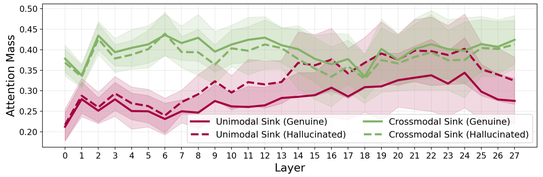
Finding 2 : Cross-modal sink tokens serve as the primary carriers of cross-modal information.

Application: Mitigating Object Hallucination in AVLLMs

- **Object hallucination in AVLLMs**
  - Misinterpreted object cues from one modality may persist across layers.
  - Incomplete suppression of erroneous cues leads to captions mixing correct and hallucinated objects.



- Prompt : Describe what you see and hear.
  - Ground truth label: Zebra braying
  - Genuine caption : **Zebras** are standing and they are braying.
  - Hallucinated caption : **Zebras** are standing and a dog is barking in the background.



- **Genuine objects:** attention to cross-modal sink tokens remains dominant across layers.
- **Hallucinated objects:** attention shifts toward unimodal sink tokens, sometimes exceeding cross-modal attention.

Dataset	Method	Qwen2.5-Omni(7B)				video-SALMONN-o1 (7B)			
		ALOHa ↑	Cs ↓	Ci ↓	F1 ↑	ALOHa ↑	Cs ↓	Ci ↓	F1 ↑
VGGSound-Animal	Vanilla	40.71	48.21	37.13	55.24	36.21	37.74	32.09	53.68
	PAI	39.52	51.24	38.11	55.11	36.99	35.26	31.18	53.16
	VCD	40.27	51.52	41.28	52.43	36.40	39.39	33.40	53.37
	ASD	**42.77**	**36.91**	**34.15**	52.44	**43.29**	**25.07**	**25.71**	50.89
VGGSound-All	Vanilla	35.02	30.70	**20.67**	58.69	32.74	30.63	22.39	53.40
	PAI	34.68	32.21	21.52	58.47	32.44	29.29	22.01	53.15
	VCD	34.60	32.63	22.36	57.09	30.28	30.76	24.31	50.02
	ASD	**38.89**	**29.65**	21.74	55.81	**36.63**	**21.11**	**18.42**	50.10
Audioset	Vanilla	38.24	8.92	10.93	69.73	36.81	11.39	14.91	67.27
	PAI	36.94	11.84	13.09	73.22	36.05	10.95	14.54	67.64
	VCD	36.98	12.28	14.88	71.12	32.50	9.34	12.52	67.74
	ASD	**38.32**	**8.54**	**10.20**	72.98	**39.64**	**6.57**	**9.50**	67.29

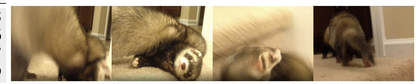
- **Adaptive Sink-Guided Decoding (ASD)**
  - Dynamically reweights attention between cross-modal and unimodal sink tokens during generation.

$$\tilde{A}_{t,j} \leftarrow A_{t,j} + \alpha |A_{t,j}|, \quad j \in \mathcal{S}_{\mathrm{cross}},$$

$$\tilde{A}_{t,j} \leftarrow A_{t,j} - \alpha |A_{t,j}|, \quad j \in \mathcal{S}_{\mathrm{uni}},$$

$$\log \tilde{P}(y_t|\mathbf{x}, y_{<t}) = \gamma_t \log P_{\mathrm{cali}}(y_t|\mathbf{x}, y_{<t}) + (1-\gamma_t) \log P_{\mathrm{orig}}(y_t|\mathbf{x}, y_{<t}).$$

$$\gamma_t^{\mathrm{base}} = \frac{\bar{A}_{t,\mathrm{uni}}}{\bar{A}_{t,\mathrm{uni}} + \bar{A}_{t,\mathrm{cross}}}$$



Prompt : Describe what you see and hear in a single sentence.

Base
A ferret is playing with a toy on the carpet while a dog is whimpering and a cat is purring.

Ours
A ferret is playing with a toy on the floor, and it's making a lot of noise.