AVHBench: A Cross-Modal Hallucination Benchmark for Audio-visual Large Language Models

Kim Sung-Bin*[1] Oh Hyun-Bin*[1] Lee Jung-Mok[1]
Arda Senocak[2] Joon Son Chung[2] Tae-Hyun Oh[3]

[1]Deptartment of Electrical Engineering, POSTECH
[2]School of Electrical Engineering, KAIST
[3]School of Computing, KAIST

*Equal contribution

Email:
sungbin@postech.ac.kr
hyunbinoh@postech.ac.kr

Motivation

- Recent advancements have expanded capabilities toward human-like video understanding through audio-visual Large Language Models (AV-LLMs)

Motivation

- Recent advancements have expanded capabilities toward human-like video understanding through audio-visual Large Language Models (AV-LLMs)
- However, AV-LLMs are prone to cross-modal hallucinations:
  - Hearing imaginary sounds from visual cues
  - Perceiving fake visual events from audio cues

Motivation

- Recent advancements have expanded capabilities toward human-like video understanding through audio-visual Large Language Models (AV-LLMs)
- However, AV-LLMs are prone to cross-modal hallucinations:
  - Hearing imaginary sounds from visual cues
  - Perceiving fake visual events from audio cues
- Currently, no suitable benchmark exists to validate such hallucinations

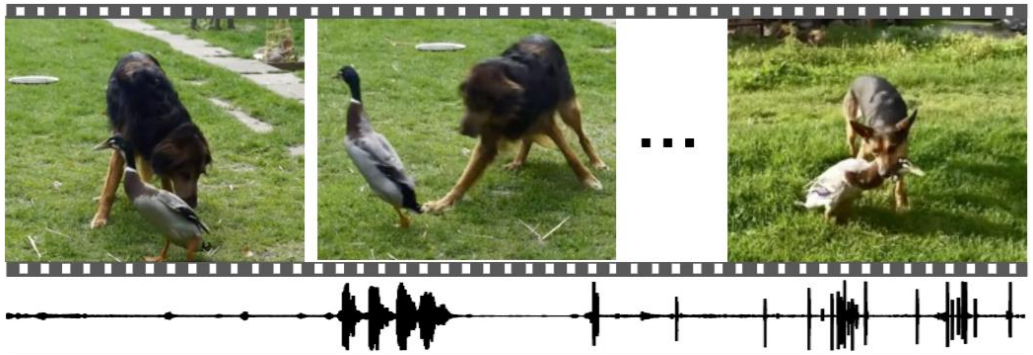
Motivation

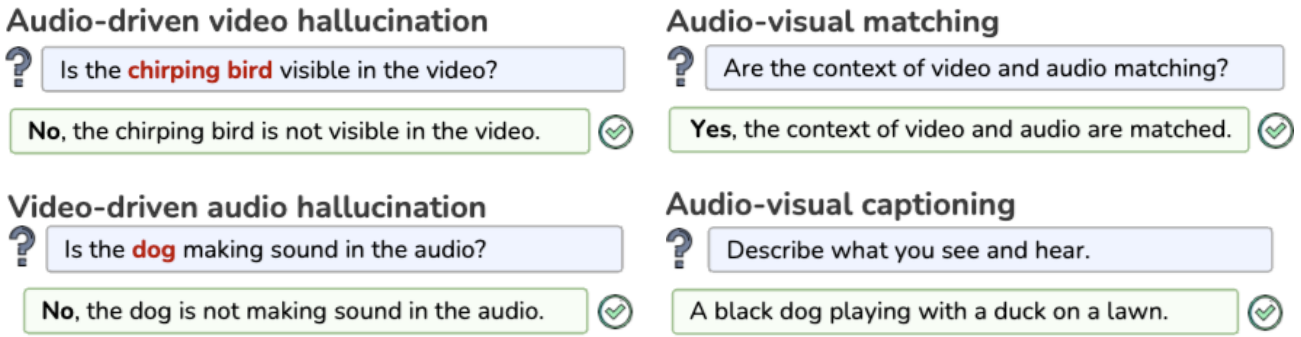
- Recent advancements have expanded capabilities toward human-like video understanding through audio-visual Large Language Models (AV-LLMs)
- However, AV-LLMs are prone to cross-modal hallucinations:
  - Hearing imaginary sounds from visual cues
  - Perceiving fake visual events from audio cues
- Currently, no suitable benchmark exists to validate such hallucinations

A benchmark is needed to evaluate the perception and comprehension capabilities of AV-LLMs regarding hallucinations

Audio-visual Hallucination Benchmark

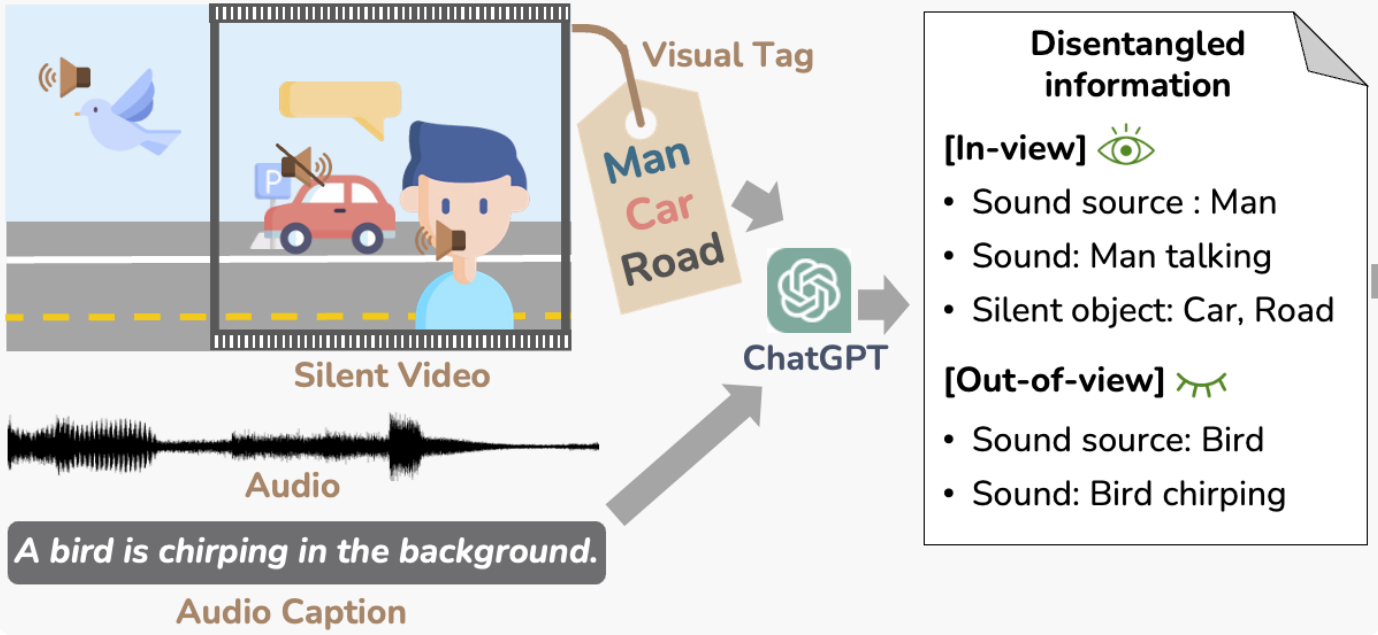
- Propose **A**udio-**V**isual **H**allucination **Bench**mark (AVHBench)
- Tasks in the benchmark (J: judgment / D: description)
  - [J] Audio-driven video hallucination
  - [J] Video-driven audio hallucination
  - [J] Audio-visual matching
  - [D] Audio-visual captioning





Semi-automatic dataset construction pipeline

Stage 1: Disentangle audio-visual information



Disentangled information

[In-view]

- Sound source : Man
- Sound: Man talking
- Silent object: Car, Road

[Out-of-view]

- Sound source: Bird
- Sound: Bird chirping

Stage 2: QA generation

Rule-based algorithm

Judgement Question

"Is there a {car} in the video?"

"Is the {car} making sound in the audio?"

"Is there a {bird chirping} sound in the audio?"

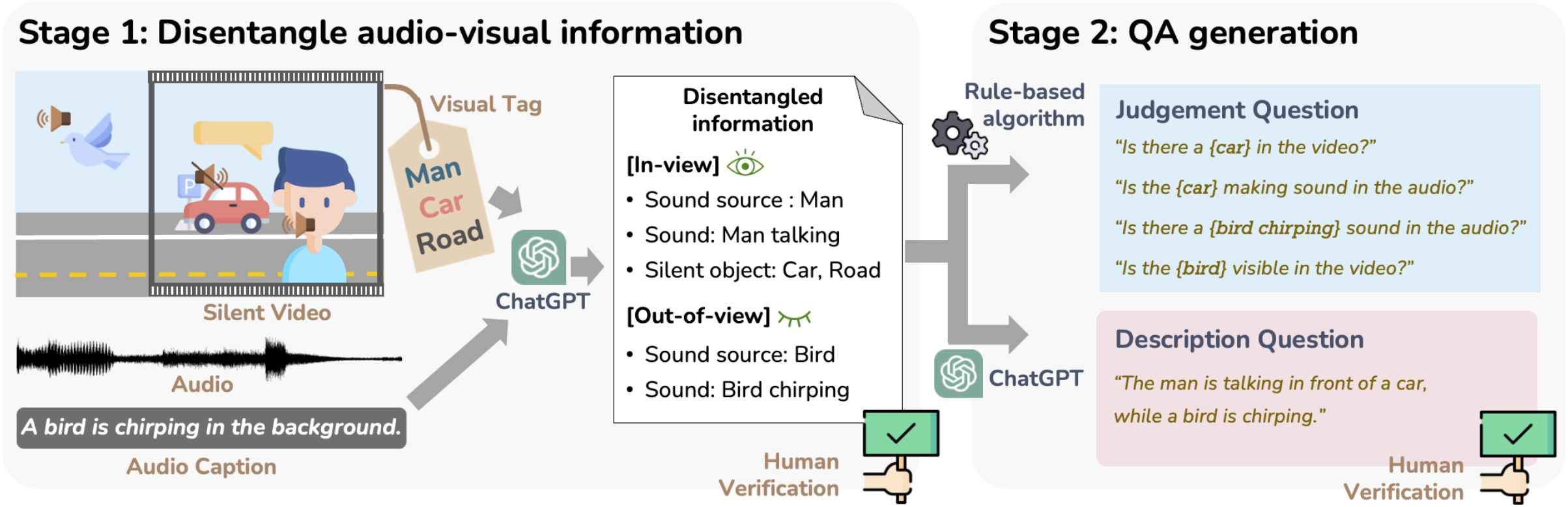
"Is the {bird} visible in the video?"

ChatGPT

Description Question

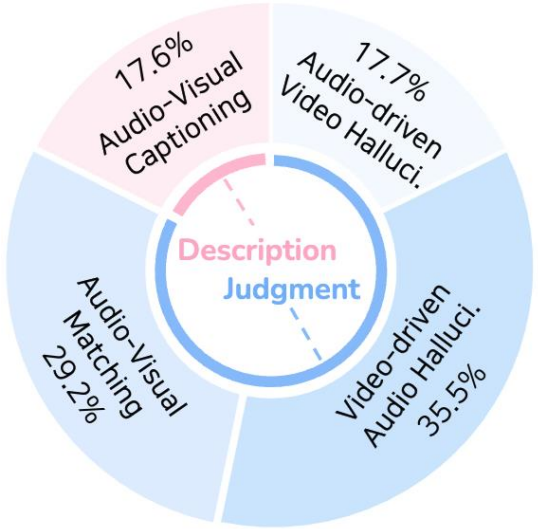
"The man is talking in front of a car, while a bird is chirping."

Semi-automatic dataset construction pipeline



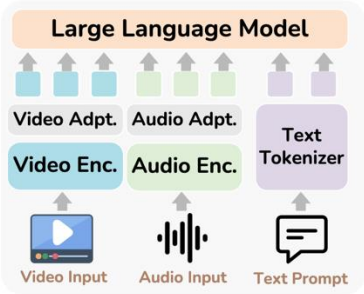
Overall dataset statistics

- Judgment tasks
  - Total 2,136 videos with 5,816 QnA pairs
  - Equally distributed yes/no answers
- Description task
  - Total 1,238 videos with corresponding captions



Task Type	Task	# QnA pairs	# Yes	# No	# captions
Judgment	A→V	1,250	625	625	-
	V→A	2,508	1,254	1,254	-
	A-V Mat.	2,058	1,029	1,029	-
Description	A-V Cap.	-	-	-	1,238
	Total	5,816	2,908	2,908	1,238

Results and analysis



(a) Multi-modal Inputs

(b) Uni-modal Inputs

(c) Caption Inputs

Model	Audio-driven Video Hallucination					Video-driven Audio Hallucination				
	Acc. (↑)	Precision (↑)	Recall (↑)	F1 (↑)	Yes (%)	Acc. (↑)	Precision (↑)	Recall (↑)	F1 (↑)	Yes (%)
X-InstructBLIP	18.1	16.0	15.0	15.5	46.9	16.3	14.5	38.5	21.1	77.0
ImageBind-LLM	50.3	50.2	87.1	63.7	86.7	50.0	50.0	99.3	66.5	99.3
Video-LLaMA	50.1	50.1	**100**	66.7	99.9	50.2	50.2	**100**	66.9	100
ChatBridge	52.9	**70.9**	52.9	48.9	77.6	32.8	**60.0**	32.8	39.8	14.8
PandaGPT	**58.5**	55.3	91.1	**68.8**	82.3	**61.3**	57.4	86.6	**69.1**	75.5
OneLLM	53.7	58.6	64.8	49.8	63.1	44.3	50.2	39.4	49.8	55
Random Choice	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0

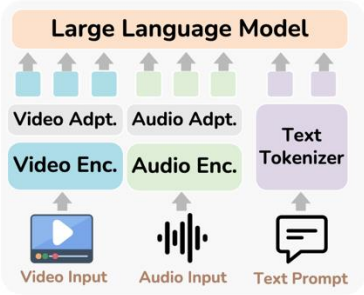
Model	Audio-visual Matching					Audio-visual Captioning		
	Acc. (↑)	Precision (↑)	Recall (↑)	F1 (↑)	Yes (%)	METEOR (↑)	CIDEr (↑)	GAVIE-A (↑)
X-InstructBLIP	15.1	18.6	18.9	18.8	52.6	6.10	3.40	2.83
ImageBind-LLM	50.0	50.0	**100**	**66.7**	100	11.5	16.0	3.35
Video-LLaMA	50.0	50.0	**100**	**66.7**	100	**14.0**	9.5	2.29
ChatBridge	29.9	48.3	29.9	33.9	13.0	13.7	**33.1**	**4.69**
PandaGPT	51.2	53.6	18.1	27.0	16.8	11.7	14.1	2.70
OneLLM	**60.1**	**67.7**	61.9	64.6	53.9	5.41	28.8	1.47
Random Choice	50.0	50.0	50.0	50.0	50.0	-	-	-

Q1. Are AV-LLMs robust against cross-modal hallucinations?

→ No, they are vulnerable to cross-modal hallucinations

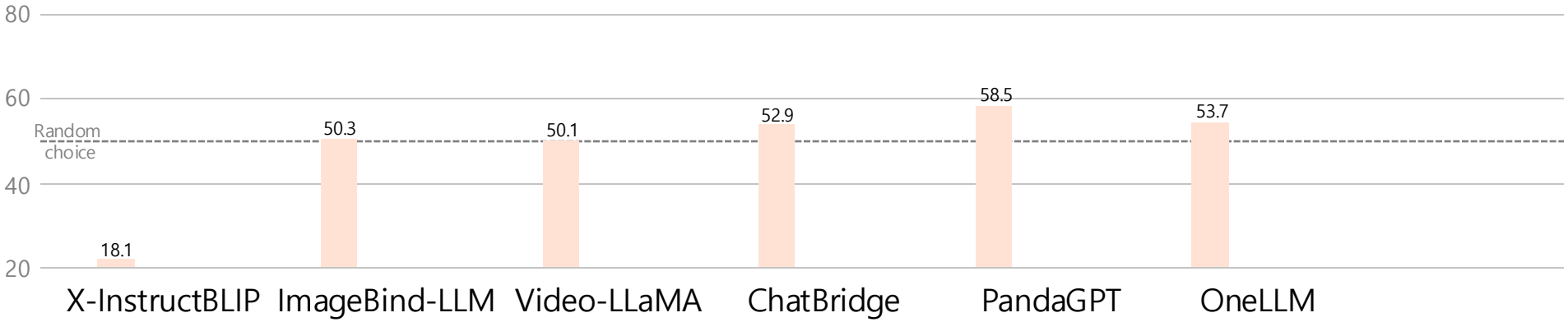
Results and analysis

: (a) Multi-modal Inputs

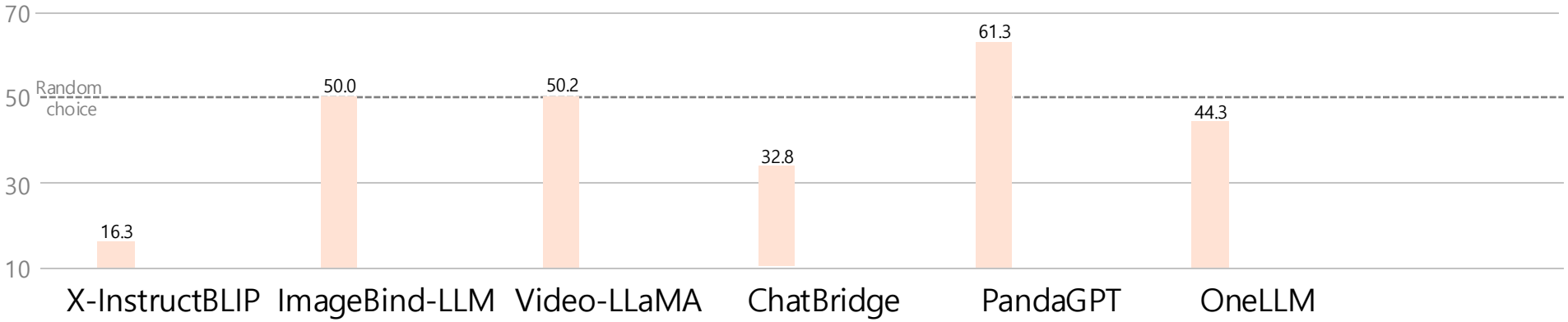


(a) Multi-modal Inputs (b) Uni-modal Inputs (c) Caption Inputs

Audio-driven Video Hallucination (Accuracy (%))



Video-driven Audio Hallucination (Accuracy (%))

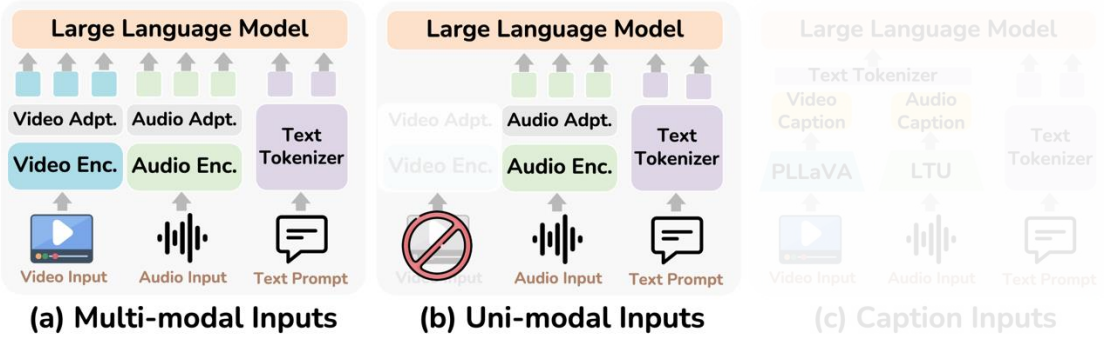


Q1. Are AV-LLMs robust against cross-modal hallucinations?

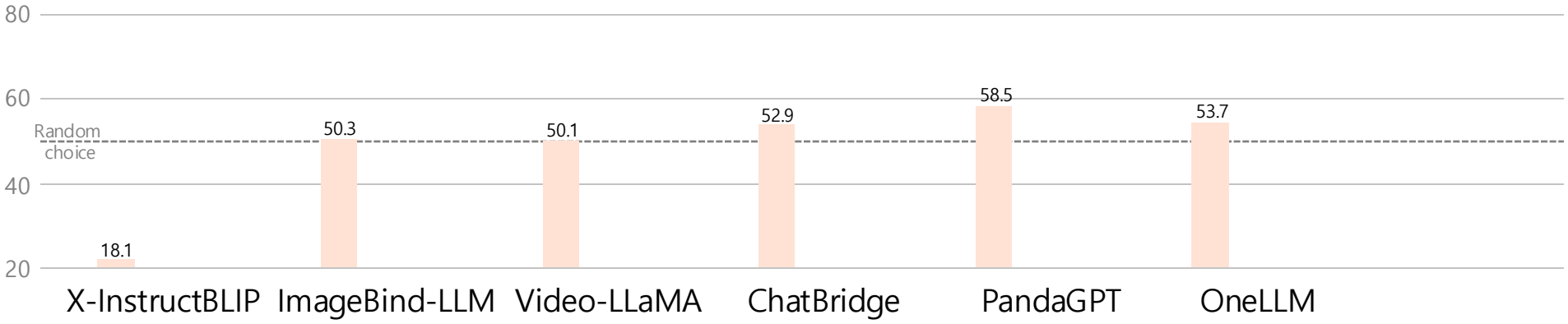
→ No, they are vulnerable to cross-modal hallucinations

Results and analysis

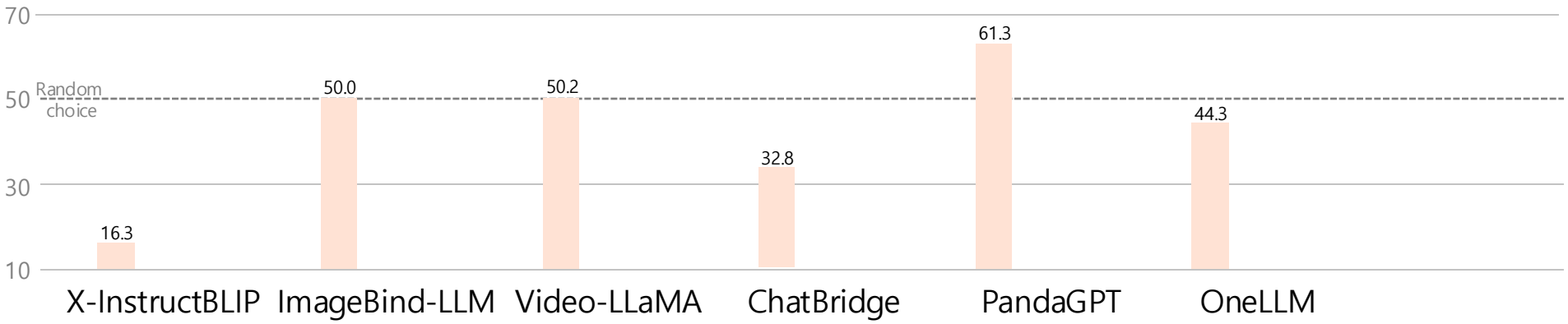
: (a) Multi-modal Inputs
: (b) Uni-modal Inputs



Audio-driven Video Hallucination (Accuracy (%))



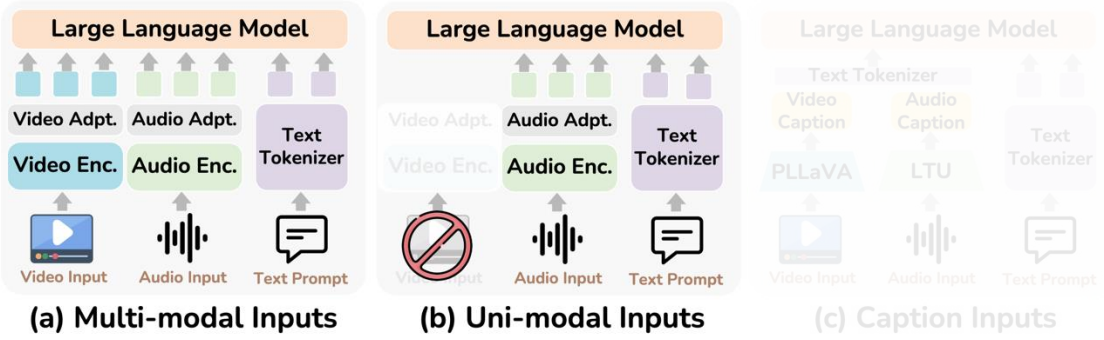
Video-driven Audio Hallucination (Accuracy (%))



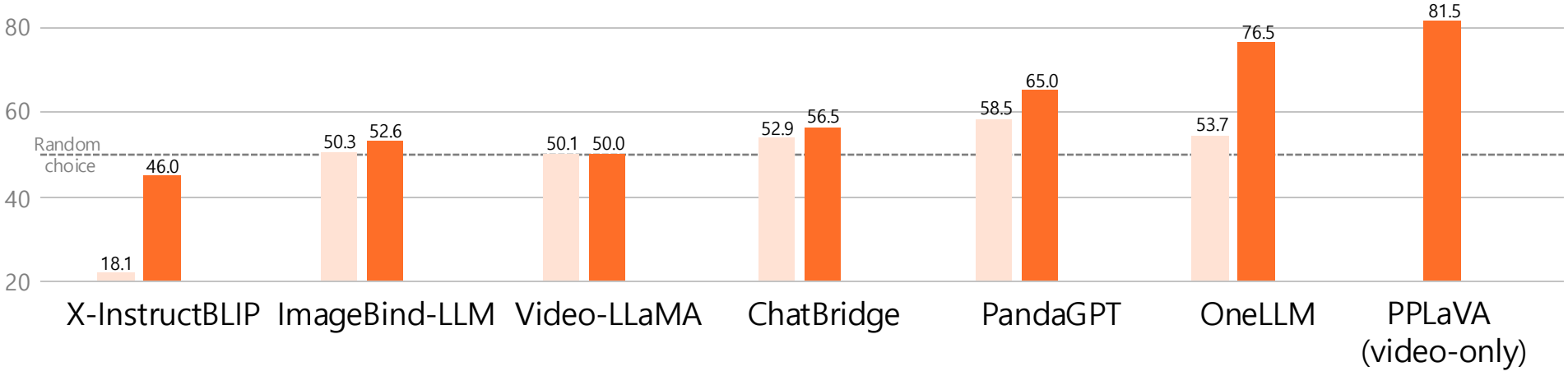
Q2. Are multi-modal inputs really helpful?

Results and analysis

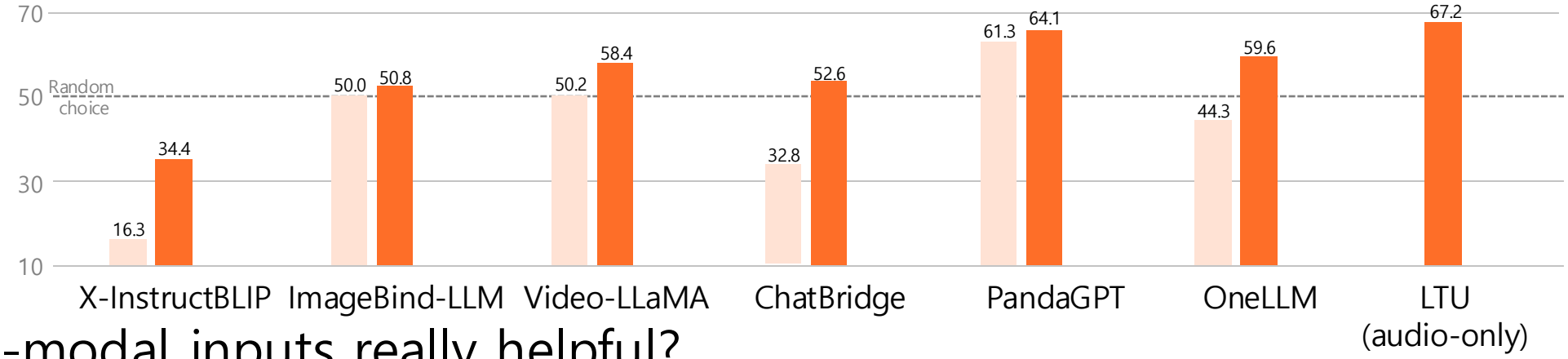
: (a) Multi-modal Inputs
: (b) Uni-modal Inputs



Audio-driven Video Hallucination (Accuracy (%))



Video-driven Audio Hallucination (Accuracy (%))

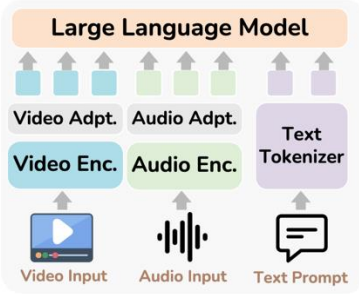


Q2. Are multi-modal inputs really helpful?

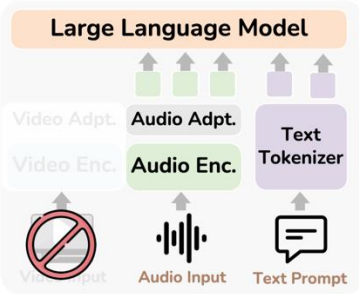
→ No, multi-modal signals tend to confuse the models' perception

Results and analysis

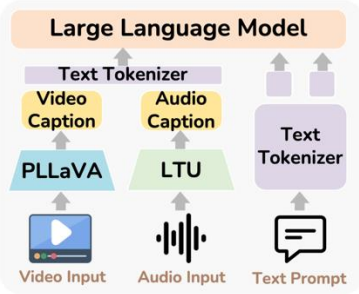
: (a) Multi-modal Inputs
: (b) Uni-modal Inputs
: (c) Caption-modal Inputs



(a) Multi-modal Inputs

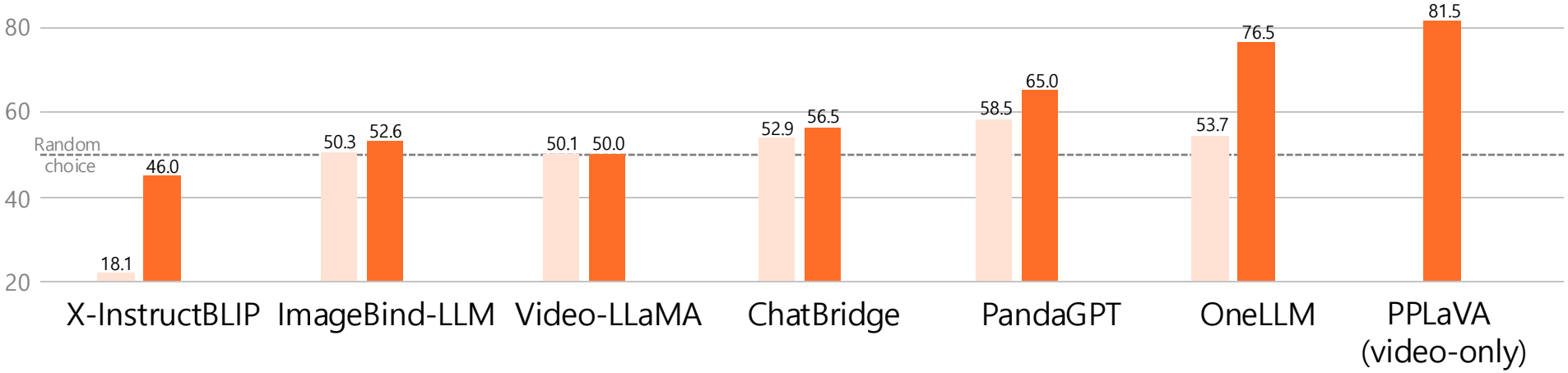


(b) Uni-modal Inputs

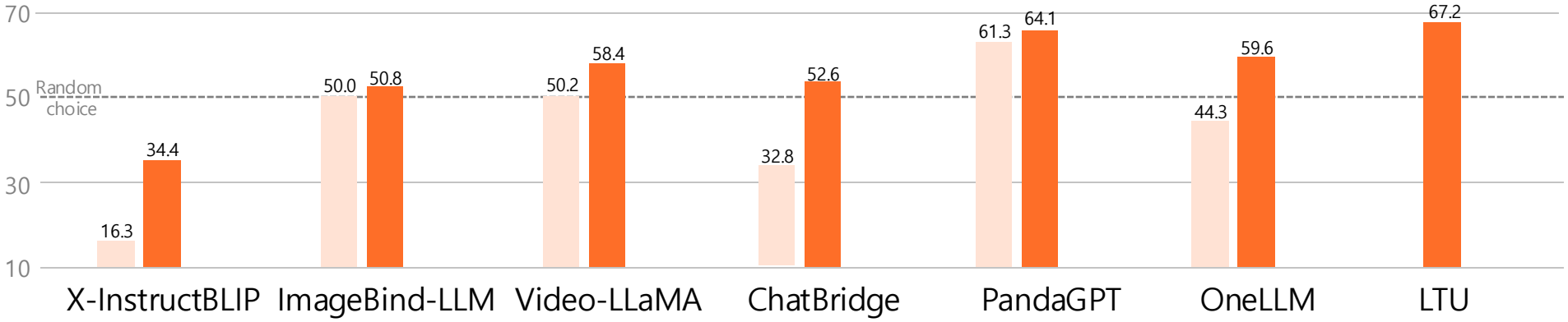


(c) Caption Inputs

Audio-driven Video Hallucination (Accuracy (%))



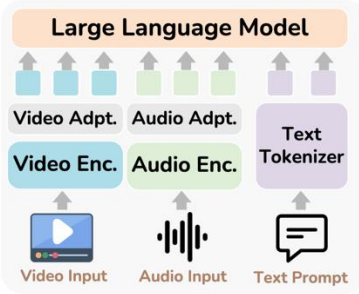
Video-driven Audio Hallucination (Accuracy (%))



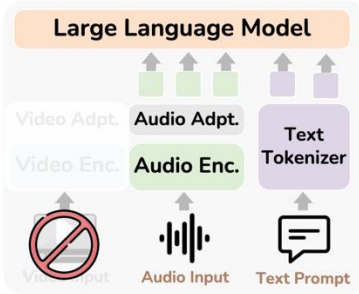
Q3. What could be the potential reason behind these hallucinations?

Results and analysis

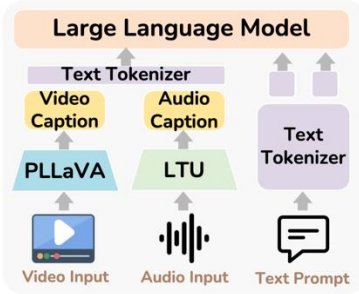
: (a) Multi-modal Inputs
: (b) Uni-modal Inputs
: (c) Caption-modal Inputs



(a) Multi-modal Inputs

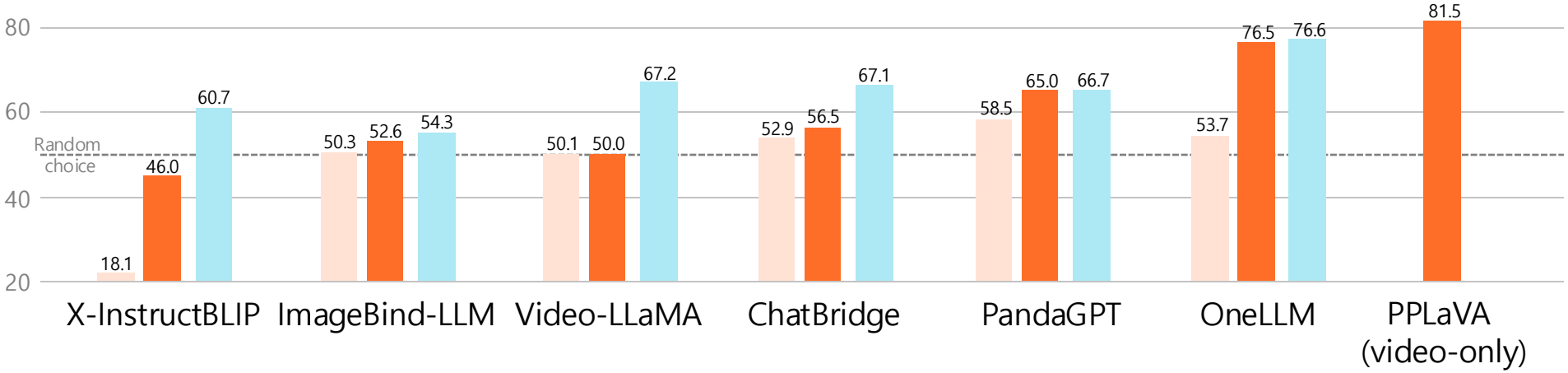


(b) Uni-modal Inputs

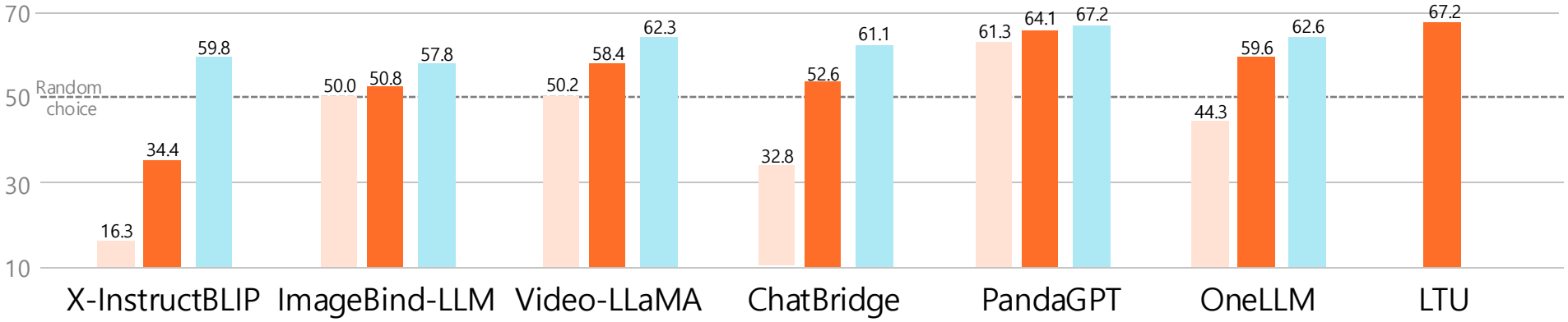


(c) Caption Inputs

Audio-driven Video Hallucination (Accuracy (%))



Video-driven Audio Hallucination (Accuracy (%))



Q3. What could be the potential reason behind these hallucinations?

→ The LLM's limited capacity to handle complex multi-modal signals

Results and analysis

Align.	FT	A→V (multi.)	A→V (uni.)	V→A (multi.)	V→A (uni.)	A-V Mat.	Audio-visual Captioning		
							METEOR (↑)	CIDEr (↑)	GAVIE-A (↑)
-	-	50.1	50.0	50.2	55.7	50.0	**14.0**	9.5	2.29
✓	-	52.8	50.4	58.1	63.5	51.3	9.5	18.9	3.49
-	✓	79.1	84.0	76.6	80.6	50.8	11.9	33.1	3.54
✓	✓	**83.9**	**85.2**	**77.3**	**81.1**	**55.6**	12.2	**35.6**	**3.82**

Enhancing robustness against cross-modal hallucinations

Q4. Can AV-LLMs be improved against cross-modal hallucinations?

→ Yes, by (1) enhancing the feature alignment, and (2) fine-tuning with LoRA

Results and analysis

Align.	FT	A→V (multi.)	A→V (uni.)	V→A (multi.)	V→A (uni.)	A-V Mat.	Audio-visual Captioning		
							METEOR (↑)	CIDEr (↑)	GAVIE-A (↑)
-	-	50.1	50.0	50.2	55.7	50.0	**14.0**	9.5	2.29
✓	-	52.8	50.4	58.1	63.5	51.3	9.5	18.9	3.49
-	✓	79.1	84.0	76.6	80.6	50.8	11.9	33.1	3.54
✓	✓	**83.9**	**85.2**	**77.3**	**81.1**	**55.6**	12.2	**35.6**	**3.82**

Enhancing robustness against cross-modal hallucinations

Align.	FT	VAST			AVinstruct				
		METEOR (↑)	CIDEr (↑)	GAVIE-A (↑)	METEOR (↑)	CIDEr (↑)	ROUGE-L (↑)	BLEU-4 (↑)	Acc. (%)
-	-	18.2	0.2	4.04	45.9	14.5	35.3	12.8	43.6
✓	-	19.2	20.7	3.68	42.2	27.1	41.5	14.9	52.6
-	✓	18.7	13.4	2.58	53.5	76.4	52.3	25.1	44.2
✓	✓	**22.1**	**47.6**	**5.09**	**58.1**	**102.0**	**55.8**	**28.5**	**57.8**

Generalization performance on other audio-visual reasoning dataset

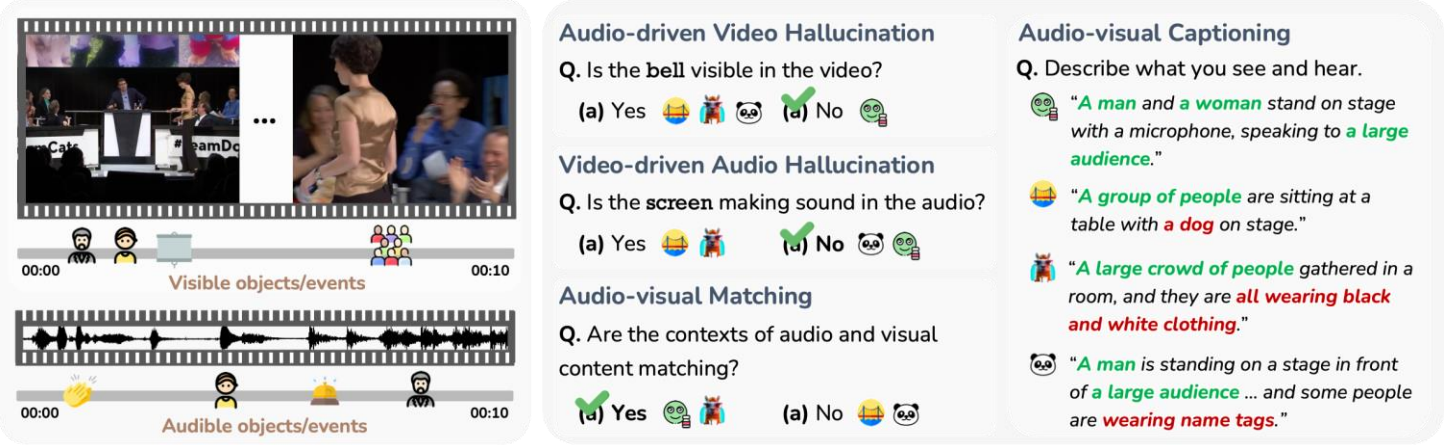
Q4. Can AV-LLMs be improved against cross-modal hallucinations?

→ Yes, by (1) enhancing the feature alignment, and (2) fine-tuning with LoRA

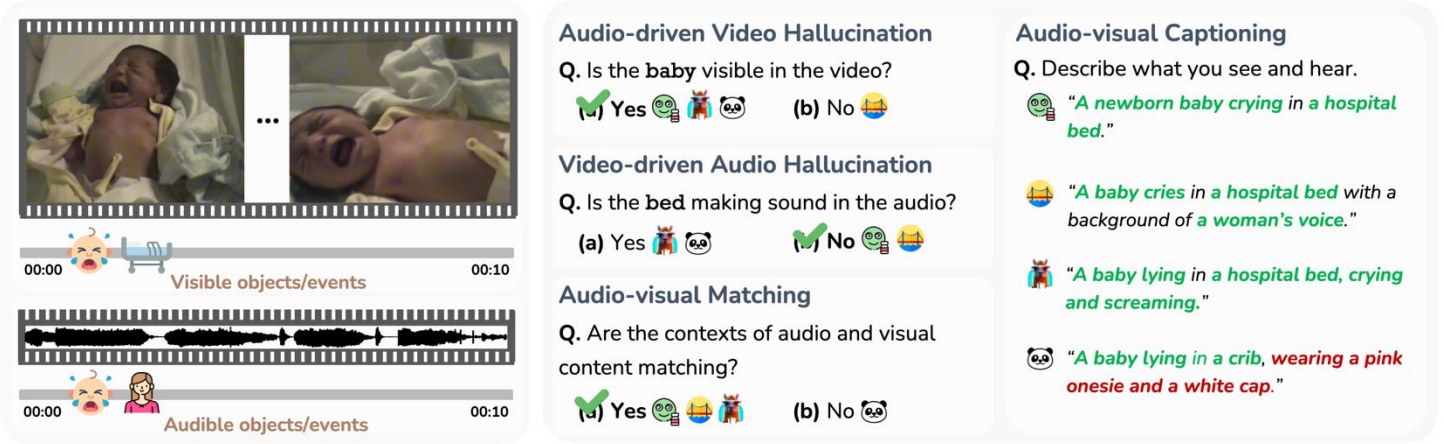
Qualitative results of AV-LLMs on AVHBench

AVHModel-Align-FT (ours) PandaGPT Video-LLaMA ChatBridge

Green: correct answer
Red: Incorrect answer



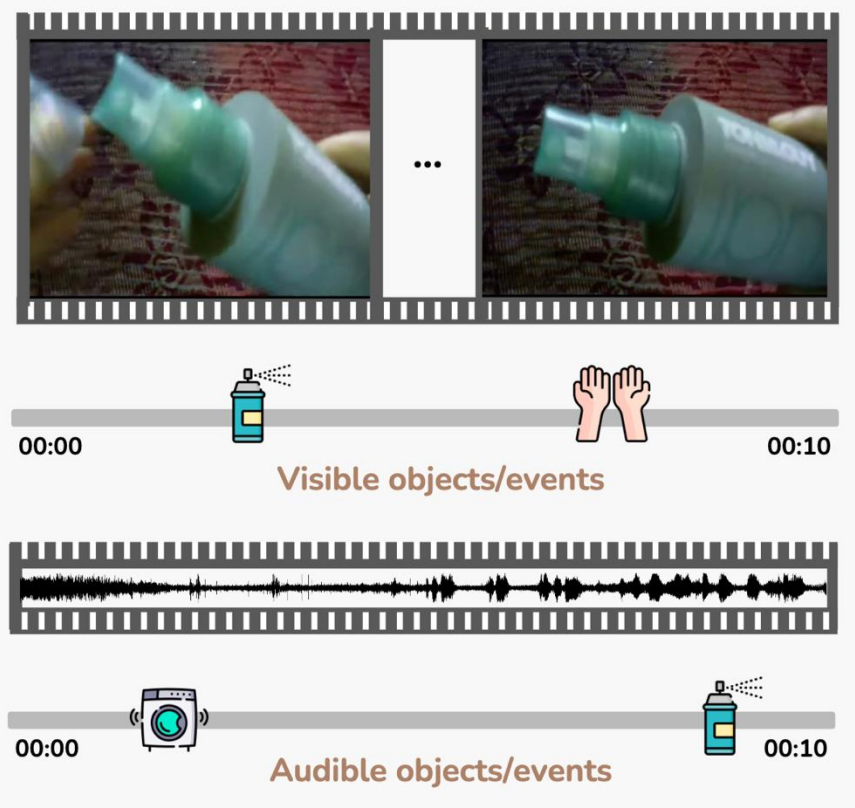
Quiz Show



Baby Crying

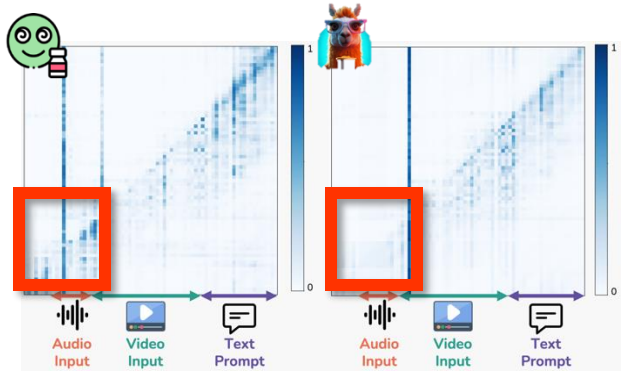
Which modalities do AV-LLMs attend to answer?

- We visualize the attention maps of the LLM layers in the AV-LLMs



Input Video & Audio

Q. Is the **bottle** making sound in the audio?



Video-driven Audio Hallucination

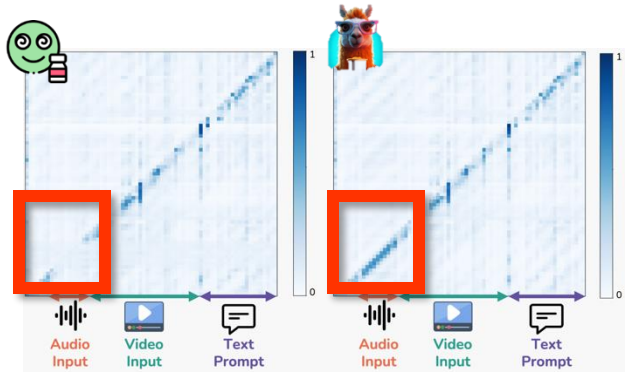
Higher attention on **audio** modality

Lower attention on **audio** modality

→ Ours leverages the audio information

→ Possibly driven by the audio feature alignment

Q. Is the **mechanical device** visible in the video?



Audio-driven Video Hallucination

Lower attention on **audio** modality

Higher attention on **audio** modality

→ Ours does not confused by the audio

→ Video-LLaMA is vulnerable to imaginary sound

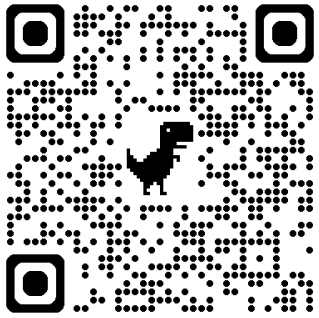
Conclusion

- Introducing AVHBench, a cross-modal hallucination benchmark for evaluating recent AV-LLMs

Email:
sungbin@postech.ac.kr
hyunbinoh@postech.ac.kr

Conclusion

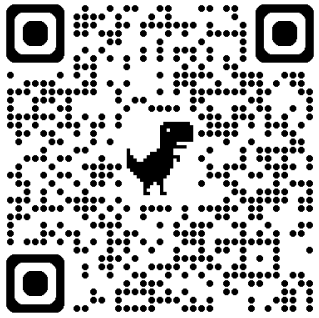
- Introducing AVHBench, a cross-modal hallucination benchmark for evaluating recent AV-LLMs
- Analyzing the phenomena of AV-LLMs using AVHBench
  - Susceptibility to cross-modal hallucinations when provided with multi-modal inputs
  - Tendency to perform better with uni-modal or text-only inputs compared to multi-modal



Email:
sungbin@postech.ac.kr
hyunbinoh@postech.ac.kr

Conclusion

- Introducing AVHBench, a cross-modal hallucination benchmark for evaluating recent AV-LLMs
- Analyzing the phenomena of AV-LLMs using AVHBench
  - Susceptibility to cross-modal hallucinations when provided with multi-modal inputs
  - Tendency to perform better with uni-modal or text-only inputs compared to multi-modal
- Enhancing feature alignment and the capacity to handle multi-modal signal could improve AV-LLMs' robustness against hallucinations



Email:
sungbin@postech.ac.kr
hyunbinoh@postech.ac.kr