Assessing Modality Bias in Video Question Answering Benchmarks with Multimodal Large Language Models

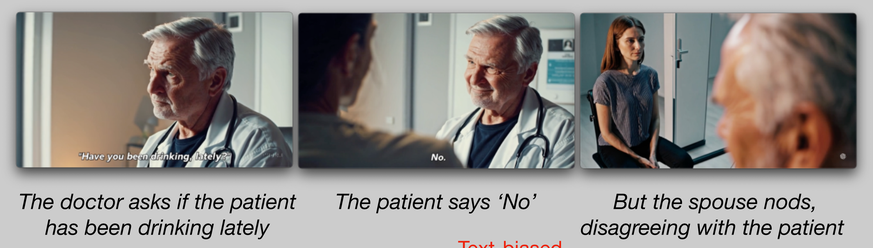
Jean Park[1], Kuk Jin Jang[1], Basam Alasaly[2], Sriharsha Mopidevi[2], Andrew Zolensky[1], Eric Eaton[1], Insup Lee[1], and Kevin B. Johnson[12]

Department of Computer and Information Science[1], Perelman School of Medicine[2]

University of Pennsylvania

Motivation

- Recent multimodal models have shown significant progress in tackling complex tasks (e.g. VidQA) that require integration of various modalities
- However, successfully integrating different modalities still remains as a siginificant challenge



The doctor asks if the patient has been drinking lately | *The patient says 'No'* | *But the spouse nods, disagreeing with the patient*

Input Question
Has the patient been drinking lately?

Biased AI -> "No"

Text-biased multimodal model would miss crucial visual cue!

True Multimodal AI -> "Yes"

- Current VidQA datasets are biased, leading multimodels trained on these dataset biased as well

Contributions

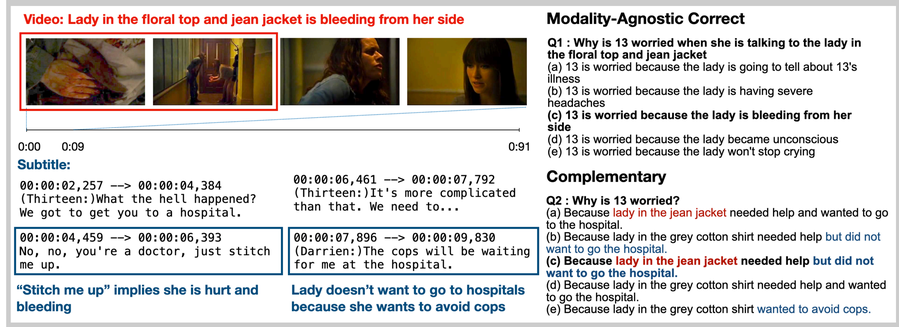
- We propose **Modality Importance Score (MIS)**
  - Measures **how much each modality contributes to answering a question** using Multimodal Large Language Models (MLLMs)

$$MIS^{i}_{m_j} = perf(q_i \mid M_j^{+}) - perf(q_i \mid M_j^{-})$$

  - Serves as an **effective proxy for human judgement**
    - Scalable and practical than manual annotation
- We **quantitatively assess modality bias in VidQA dataset** using MIS
- We reveal that **multimodal models do not optimally combine information from different sources** using MIS

Modality Bias - Examples

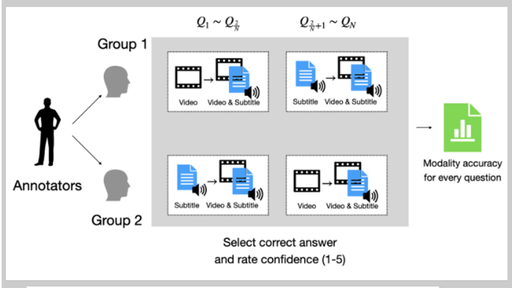
MIS reveals relative contribution of each modality compared to other



1. **Unimodal-biased:** Only answerable by single modality
   - Subtitle-biased (SB): *Why did the lady refuse to go to the hospital?*
   - Video-biased (VB): *What was the lady wearing?*
2. **Modality Agnostic Correct/Incorrect** (MA_C/MA_{IC})**:** Always correct/incorrect regardless of which subset of modaliteis used
3. **Complementary** (C)**:** Only answerable using a specific set of modalities

MLLM vs Human MIS Analysis

Goal: How well does MLLM-derived MIS align with human perception?

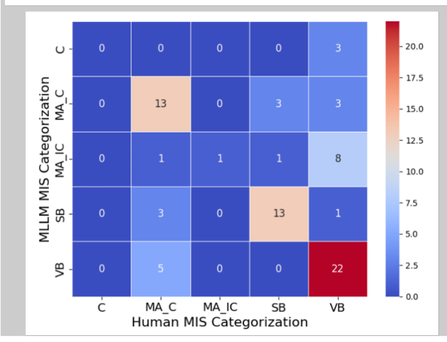


Human study

- Total 4 participants divided into 2 groups
  - Avg 7 hours per person
- Evaluated 197 questions from TVQA
  - Analysis focused on 77 unanimously agreed questions

MLLM Classification

- Zero-shot prompting by GPT-4 Turbo given different set of modalities
- We compute each question's MIS based on response accuracy and categorize them



MLLM vs Human MIS Analysis

- **No complementary questions** identified by humans
- MLLM often misclassified VB as MA_{IC}
- **Fair correlation** between human and MLLM based MIS
  - Potential for improved MIS accuracy as models advance
- MLLM-based MIS is much more **scalable and practical** than human evaluation
  - 1-2 hours w/ $20 compute vs 7 hrs per person

MLLM-derived MIS

Coverage of Questions in Multimodal Datasets

- TVQA
  - Source: TV shows
  - # of QAs: 1,019/15,253
- LifeQA
  - Source: YouTube videos
  - # of QAs: 372/372
- AVQA
  - Source: YouTube videos (VGG-Sound dataset)
  - # of QAs: 796/6,728

* **Modalities:** Video + Caption, **QA type:** Multiple Choice Answers

MLLM-derived MIS distribution

Goal: Analyze modality bias in datasets

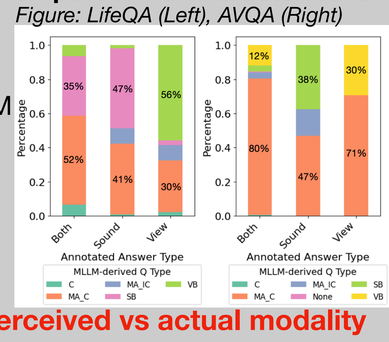
- Input: Subset of modalities (video, subtitle, both), prompt, question, answer candidates
- Method: Used GPT-4 Turbo
- **Distribution:** MA_C, SB, $VB \uparrow$ $C \downarrow$

Question Type	TVQA	LifeQA	AVQA
SB	224 (22.0%)	74 (19.9%)	39 (4.9%)
VB	345 (33.9%)	125 (33.6%)	93 (11.7%)
C	21 (2.1%)	9 (2.4%)	5 (0.6%)
MA_C	357 (35.1%)	135 (36.3%)	625 (78.5%)
MA_{IC}	71 (7.0%)	29 (7.8%)	32 (4.0%)
None	1 (0.1%)	0 (0.0%)	4 (0.5%)
Total	1,019	372	796

MLLM-derived MIS validation (LifeQA, AVQA)

Goal: Compare MLLM-derived MIS against dataset's question annotations based on perceived characteristics

Figure: LifeQA (Left), AVQA (Right)



- > 50% of questions annotated as "Both" classified as MA_C by MLLM
- Many "Sound" or "View" type are MA_C
- 2nd most common MLLM-derived type aligns with annotation

- **Reveals gap between perceived vs actual modality dependencies**
- **Truly multimodal questions are scarce**

Evaluation

Goal: Analyze how well models focus on information relevant to each question type

	Subtitle-biased			Video-biased		
	Orig.	SP (Δ)	VP (Δ)	Orig.	SP (Δ)	VP (Δ)
Merlot R*	91.5 ± 0.0	32.2 ± 3.8 (-59.3)	87.4 ± 1.9 (-4.1)	71.9 ± 0.0	72.0 ± 1.5 (+0.1)	43.2 ± 5.0 (-28.7)
FrozenBiLM	95.5 ± 0.0	31.3 ± 4.3 (-64.2)	96.3 ± 0.3 (+0.8)	75.4 ± 0.0	73.4 ± 2.7 (-1.9)	41.5 ± 4.4 (-33.9)
Llama-VQA	95.1 ± 0.0	37.3 ± 1.8 (-57.8)	94.3 ± 0.0 (-0.8)	56.9 ± 0.0	56.1 ± 0.3 (-0.8)	47.5 ± 1.5 (-9.4)
MiniGPT4*	61.4 ± 0.2	35.9 ± 3.6 (-25.5)	58.7 ± 3.5 (-2.8)	42.4 ± 0.8	40.9 ± 2.0 (-1.5)	38.6 ± 3.2 (-3.9)
Average	85.9 ± 0.0	34.2 ± 3.4 (-51.7)	84.2 ± 1.5 (-1.7)	61.6 ± 0.2	60.6 ± 1.6 (-1.0)	42.7 ± 3.0 (-19.0)

Table: Accuracy (%) comparision after feature permutation with five random seeds

*Orig: Original, SP: Subtitle Permutation, VP: Video Permutation, Δ: Difference between original and permutation

- **Method: Feature permutation**
  Replace each video's features (subtitle/image) with random features (different subtitle/image)
- **Observation**
  - Model accuracy drops significantly more after **permuting important feature** (e.g. subtitle for SB)
  - Accuracy drops slightly when permuting **less important modality feature** (e.g. video for SB)
- **Takeaway**
  - MLLM-derived MIS effectively identifies unimodal-biased questions
  - Model perform better with subtitles, likely due to prevalence of SB and MA_C questions
  - Models do not optimally combine information from different modalities

Conclusion

- We introduce new metric, Modality Importance Score (MIS), which effectively measures modality contributions for each question in multimodal dataset. MIS aligns with human assessment while being more efficient than manual annotation. Moreover, experiments revealed current open-source multimodal models struggle to properly reason on multimodal information due to modality bias in VidQA datasets.
- MIS has promise for **mitigating bias** associated with developing questions and answers in multimodal datasets.

Acknowledgement

- National Institutes of Health (NIH) #DP1-LM014558, #UL1TR001878
- National Science Foundation (NSF) #NSF-1915398
- Army Research Office MURI (ARO-MURI) #W911NF-20-1-0080
- University of Pennsylvania, Collaborative Research in Trustworthy AI for Medicine grant by ASSET

Contact: hlpark@seas.upenn.edu, GitHub: