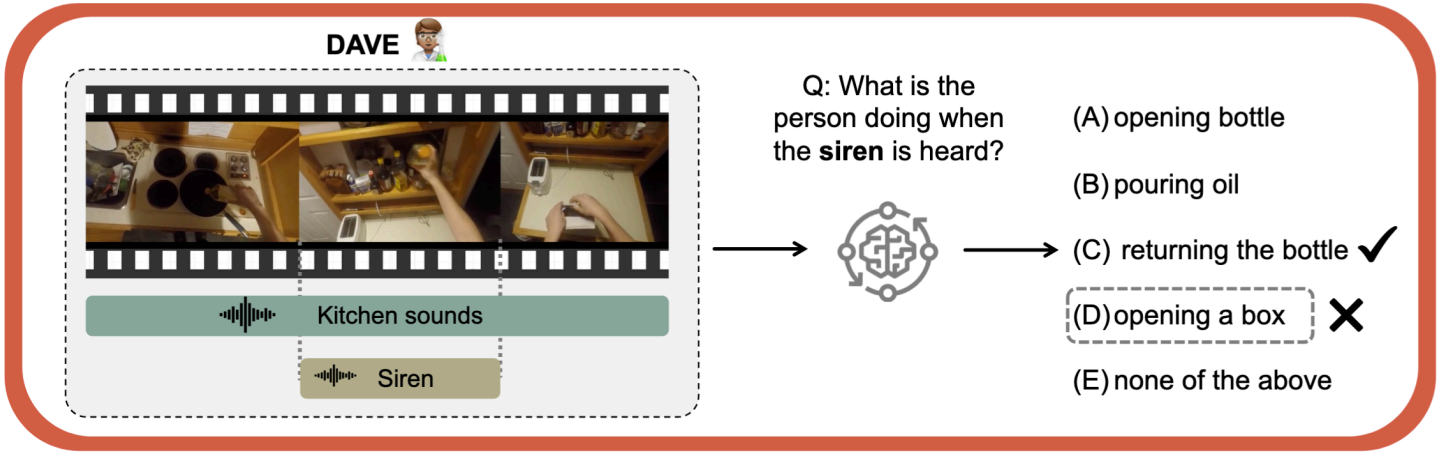
KU LEUVEN

How can we perform fine-grained analysis of multi-modal LLMs?: Performance evaluation and perspectives for future improvements

Matthew Blaschko
(thanks to Gorjan Radevski, Teodora Gagaleska, Tinne Tuytelaars, & Jiameng Li)

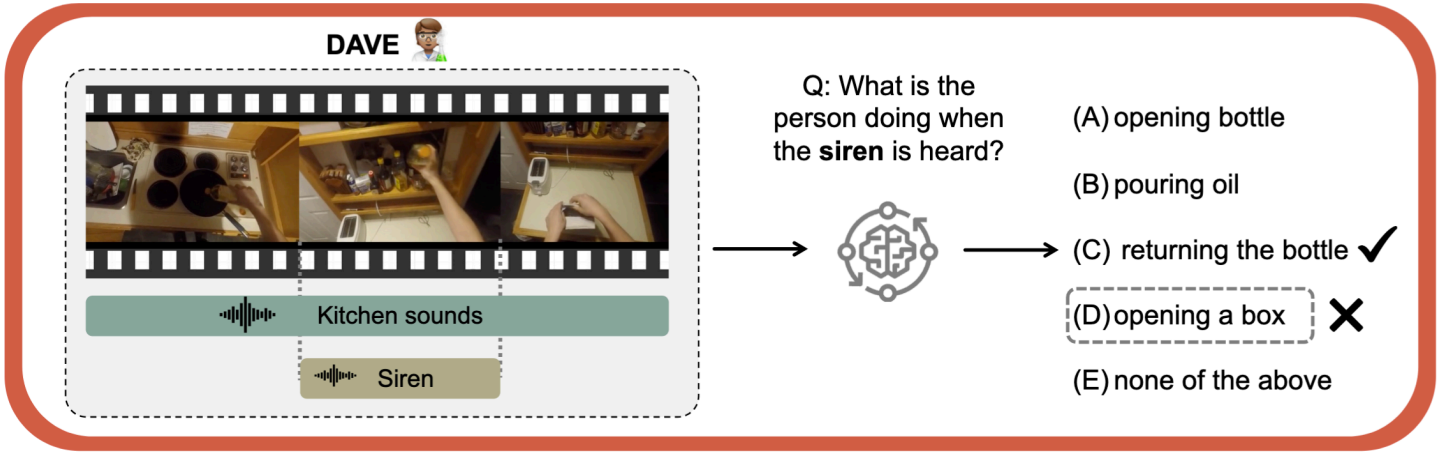
Multimodal reasoning in LLMs

- Multimodal reasoning depends on both modalities simultaneously
- Temporal dependence
- Dominance of one modality
- Inconsistencies between modalities



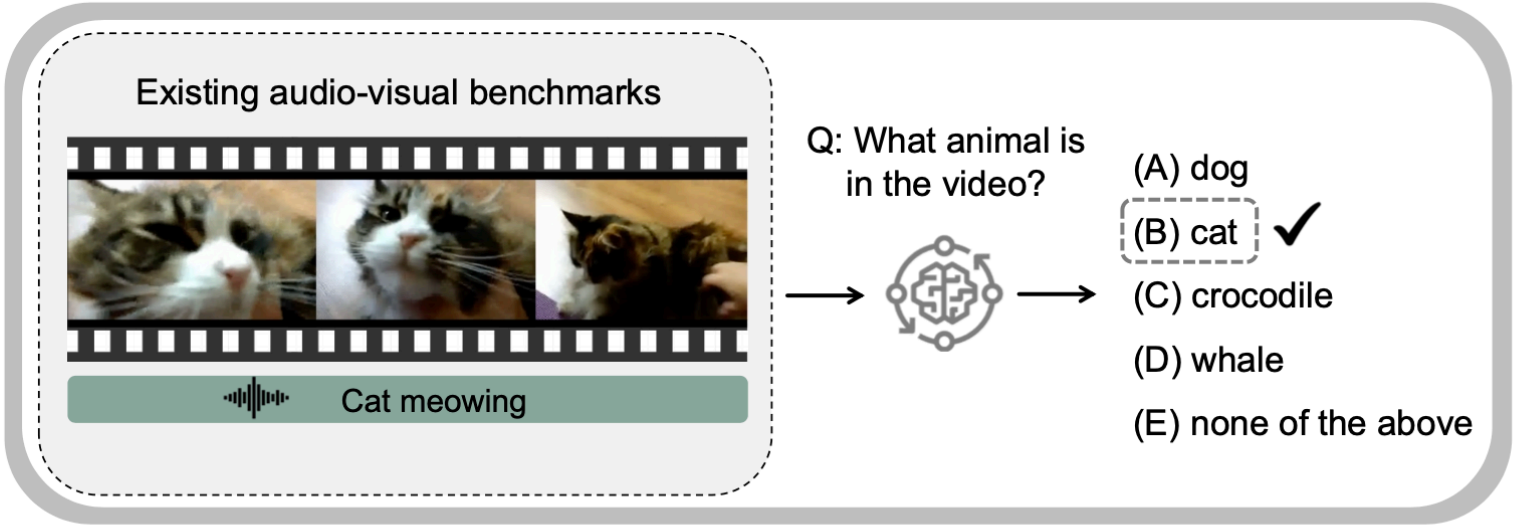
Measurement & refinement

- Benchmarks allow us to measure multi-modal LLM abilities
- Check appropriateness for use in specific applications
- Question answering most common
- Fine tuning on data enables improved performance



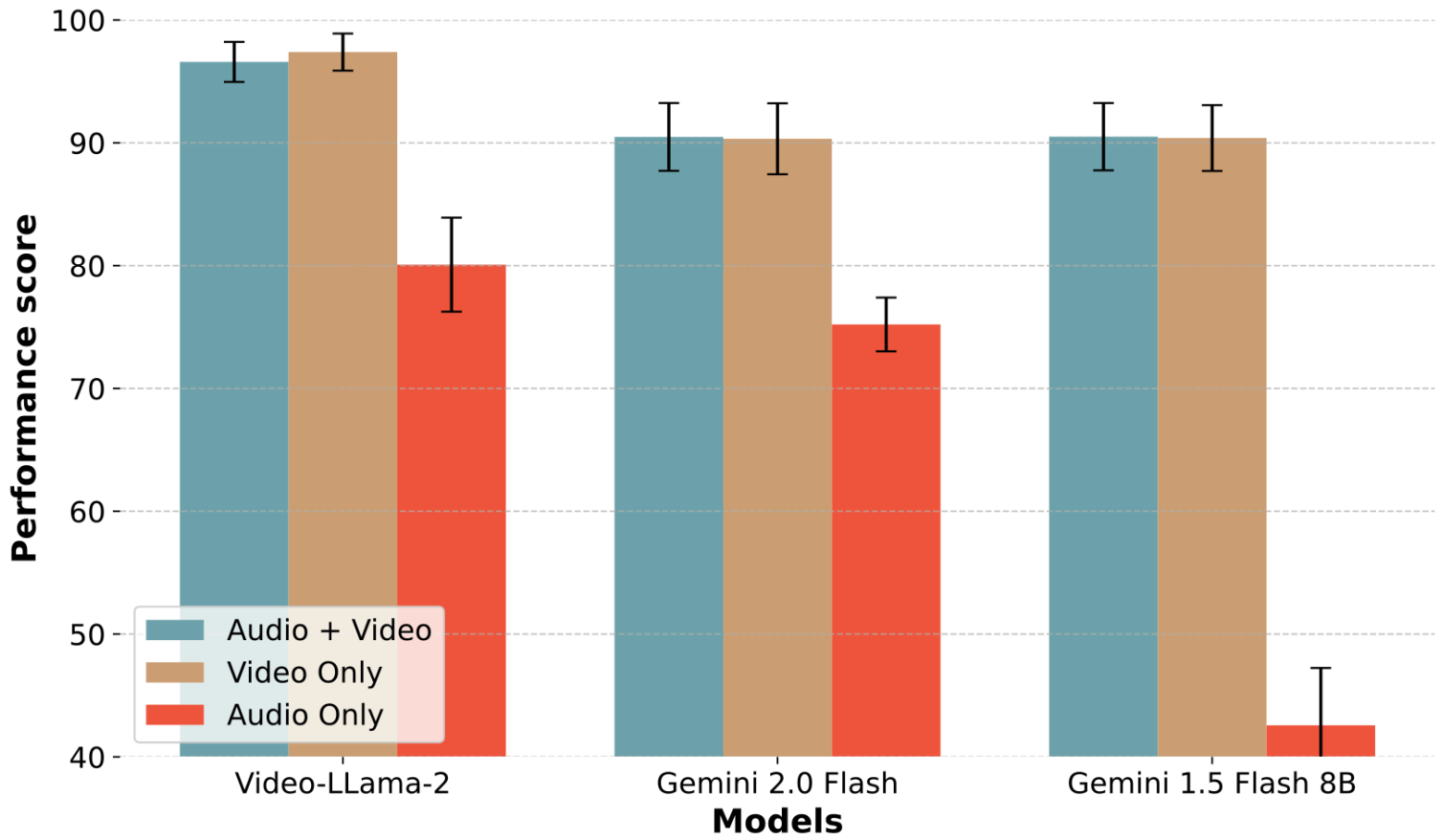
The problem with audio-visual evaluation

- Existing benchmarks are broken: **Strong visual bias**



- Are we testing **multimodal integration** or just vision understanding?

AVQA dataset: Video alone is sufficient



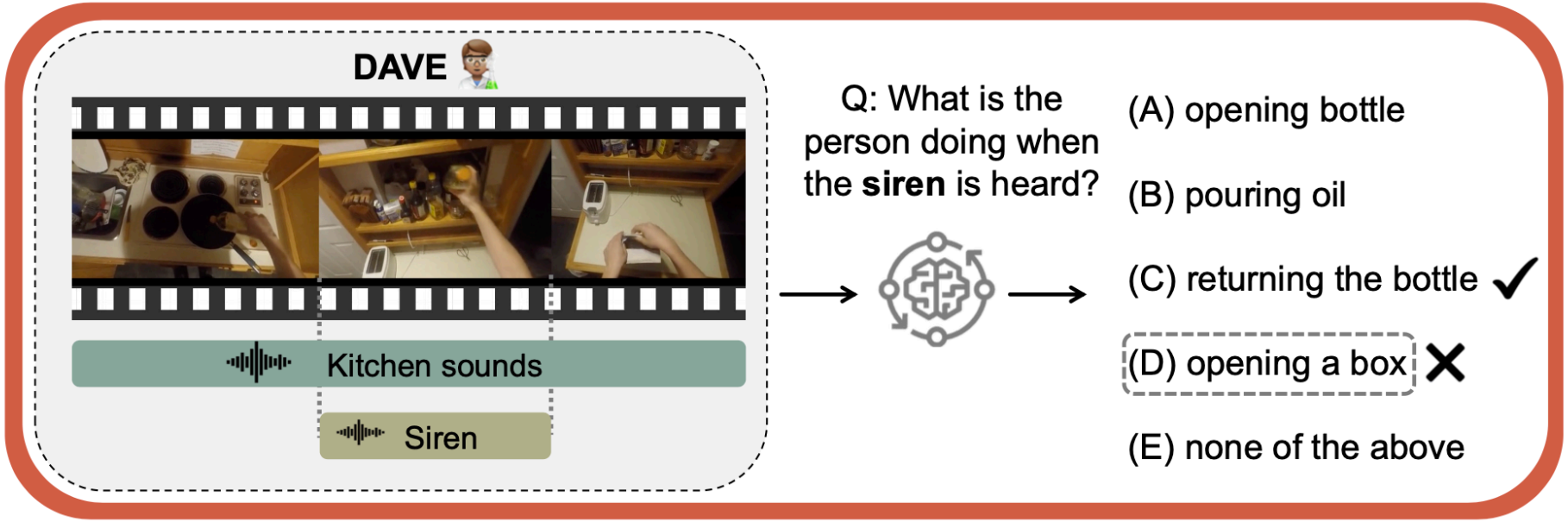
Yang et al., 2022. AVQA: A Dataset for Audio-Visual Question Answering on Videos

What is missing?

- **Temporal alignment:** Does the model know when sounds occur relative to actions?
- **True multimodal necessity:** Both modalities required to answer correctly
- **Diagnostic evaluation:** Which specific capability is failing?

DAVE 👨🏾‍🔬: Diagnostic Audio-Visual Evaluation

- Questions requiring **both audio and video**, decomposed into **testable subcategories**

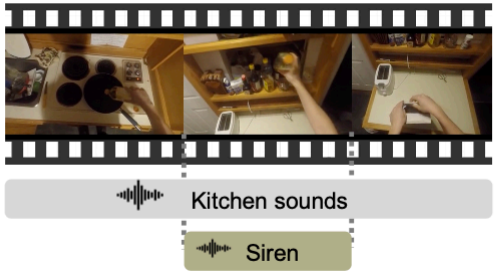


DAVE 👨🏾‍🔬: Diagnostic Audio-Visual Evaluation

```
What is the person doing when the laughing sound is heard in the background? Note
that the sound might not be present in the video, in which case the correct answer
would be 'None of the above'.
(A) first time: get meat from pan
(B) first time: place meat on dough
(C) second time: get meat from pan
(D) second time: place meat on dough
(E) none of the above
Answer only with the letter corresponding to your choice in parenthesis: (A), (B),
(C), (D) or (E). Do not include any other text.
**************************************************
Correct answer:  (D)
```

DAVE: Three Diagnostic Question Types

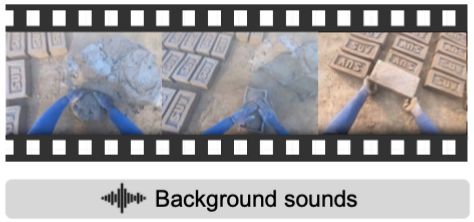
Multimodal synchronisation



Q: What is the person doing when the **siren** is heard?

(A) opening bottle
(B) pouring oil
(C) returning the bottle
(D) opening a box
(E) none of the above

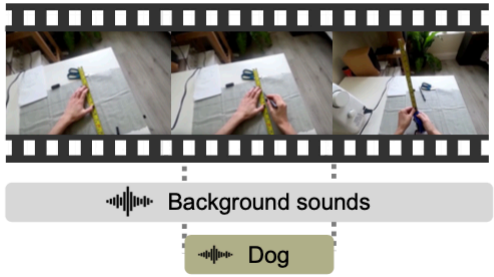
Sound absence detection



Q: What is the person doing when the **car horn** is heard?

(A) put the clay in the brick mould
(B) wipe the brick mould with sand
(C) take the clay
(D) place the brick on the ground
(E) none of the above

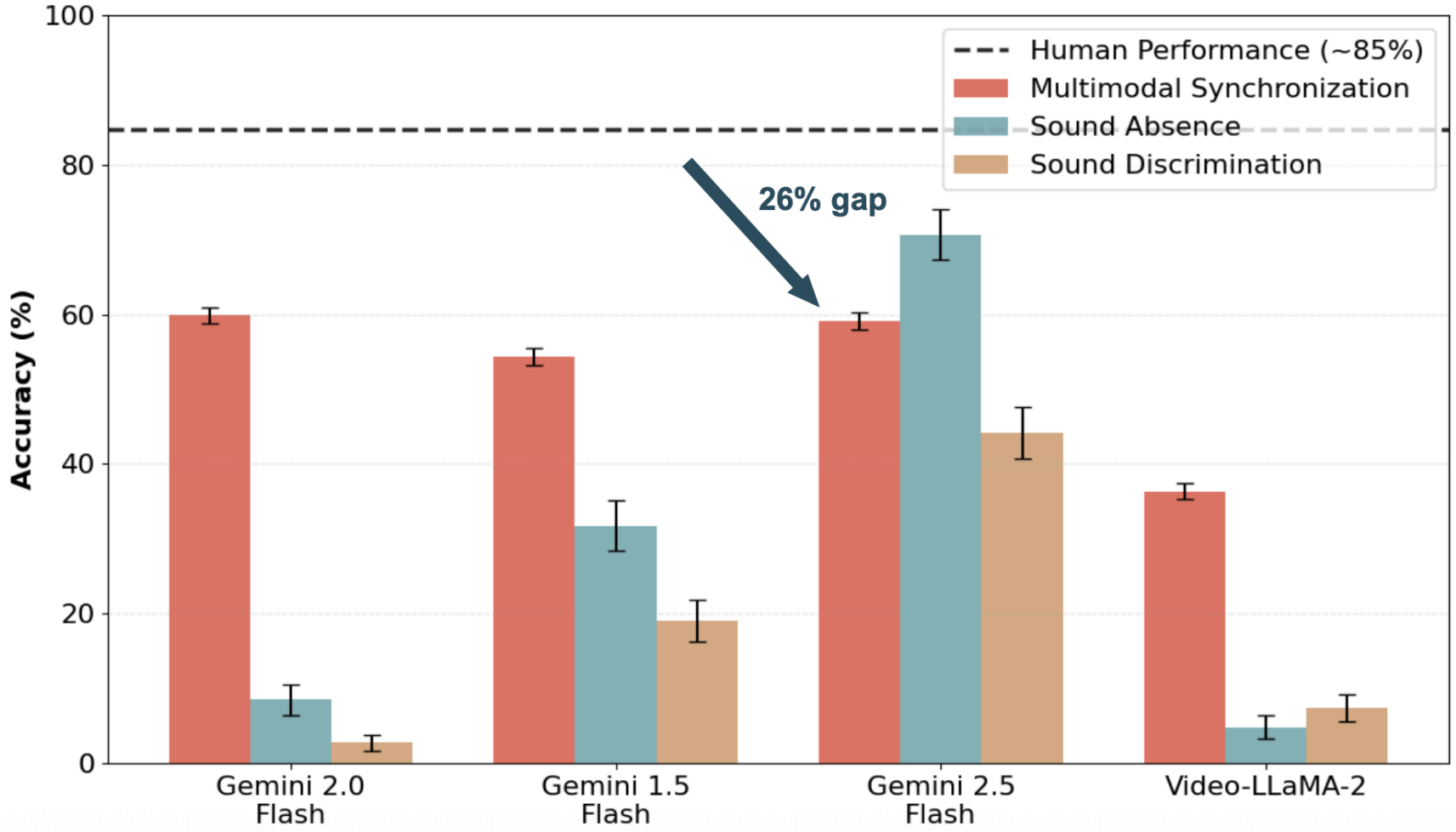
Sound discrimination



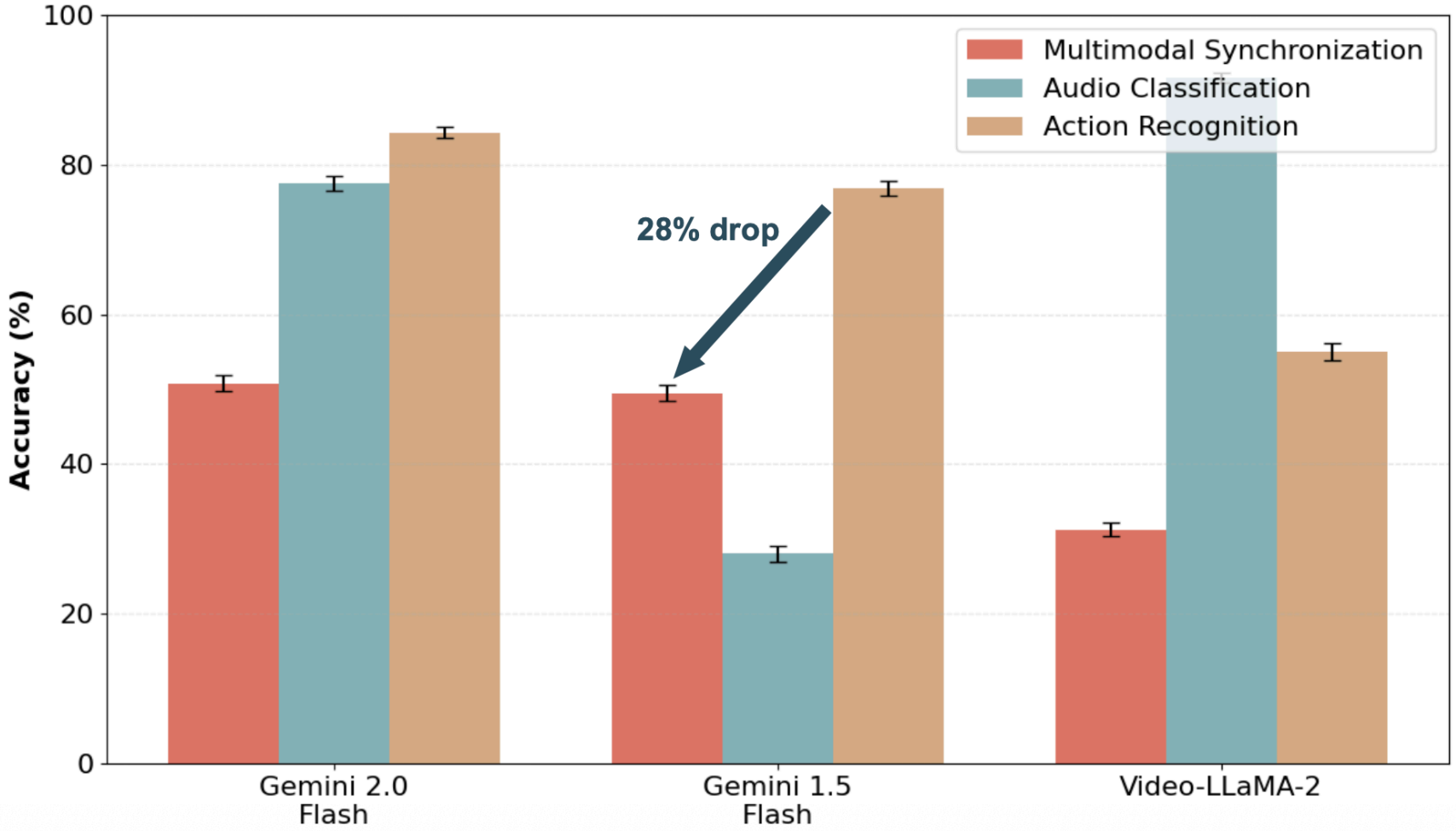
Q: What is the person doing when **coughing** is heard?

(A) mark a piece of cloth
(B) fold the measuring tape
(C) put the pen on the table
(D) measure a piece of cloth
(E) none of the above

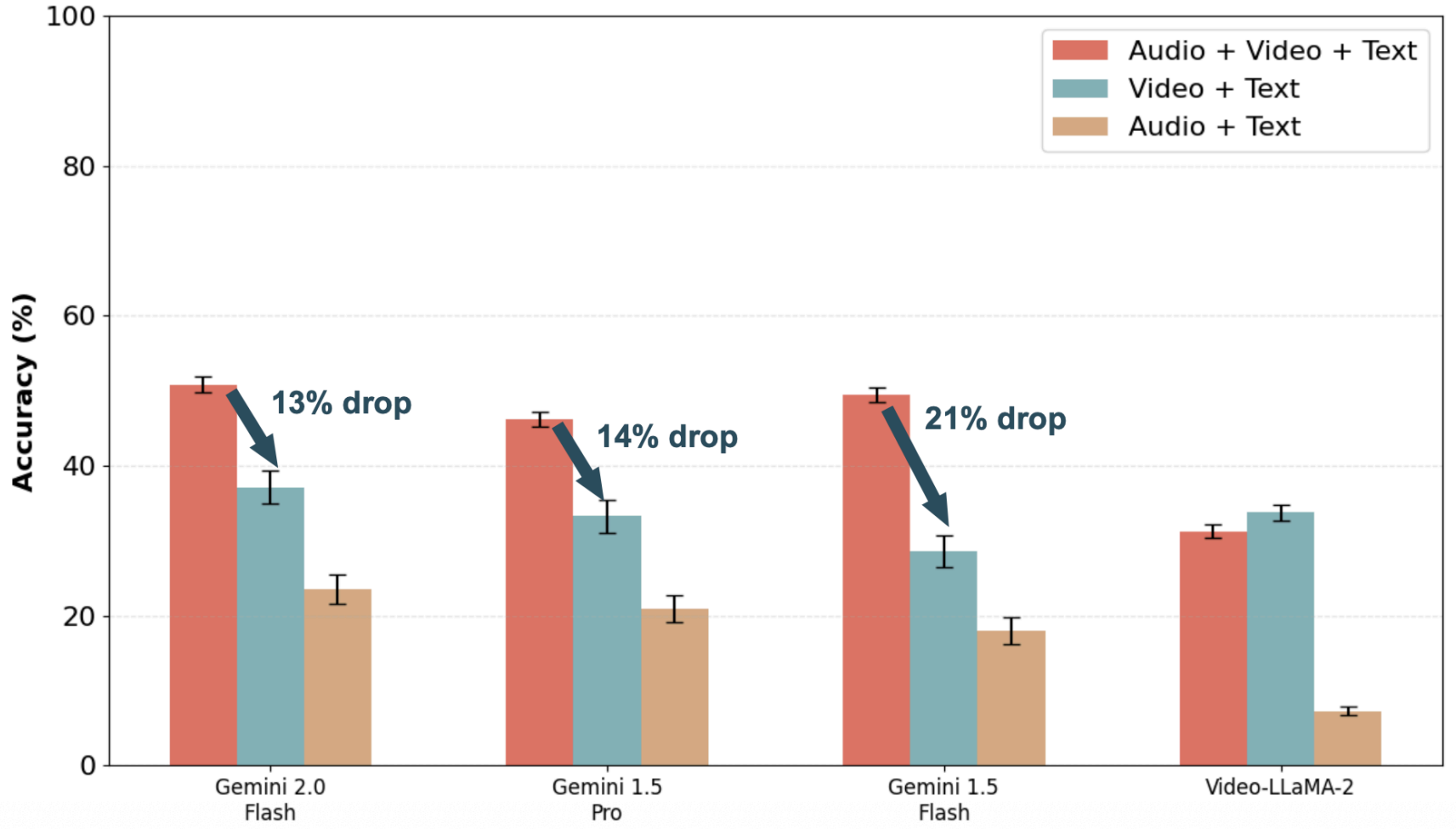
Models fail at audio-visual integration



The root cause: Integration, not perception



DAVE requires true multimodal reasoning



DAVE: Three critical insights

- Audio-visual models **trail behind humans** by 26% points
- Models **hallucinate** audio-visual connections
- **Bottleneck** is **temporal alignment**, not perception

🤗 **Data:** hf.co/datasets/gorjanradevski/dave
💻 **Code:** github.com/gorjanradevski/dave

Conclusions

- Evaluating multi-modal capabilities can be automated
  - Labeled datasets can be leveraged to generate known audio-visual alignment or image-text conflicts
  - Question answering gives largely automated evaluation
- Benchmarks show that there are shortcomings in current models
- Some commercial models substantially better than current open models
- Fine-tuning using these benchmarks can potential improve performance
  - Standard LoRA strategies can be applied
  - Need to evaluate generalization to other tasks

Selected References

- Gorjan Radevski, Teodora Popordanoska, Matthew B. Blaschko, Tinne Tuytelaars: DAVE: Diagnostic benchmark for Audio Visual Evaluation. Neural Information Processing Systems (NeurIPS), 2025.
- Konstantinos Kontras, Thomas Strypsteen, Christos Chatzichristos, Paul P. Liang, Matthew Blaschko, Maarten De Vos: Balancing Multimodal Training Through Game-Theoretic Regularization. Neural Information Processing Systems (NeurIPS), 2025.
- **Dongli Xu, Aleksei Tiulpin, Matthew B. Blaschko:** SoftCFG: Uncertainty-guided Stable Guidance for Visual Autoregressive Model. **arXiv:2510.00996, 2025.**
- Kontras, K., C. Chatzichristos, M. B. Blaschko, and M. De Vos: Improving Multimodal Learning with Multi-Loss Gradient Modulation. British Machine Vision Conference (BMVC), 2024.

Thank you!

Questions?