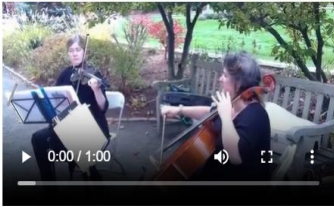


1. Aim - Measuring change in causality of vision language models using audio
2. Datasets –
 - a. Music VQA
 - i.

More video examples

Some video examples with QA pairs in the MUSIC-AVQA dataset. Through these examples, we can have a better understanding of the dataset, and can more intuitively feel the QA tasks in dynamic and complex audio-visual scenes



Question: How many instruments are sounding in the video?
Answer: two
 To answer the question, an AVQA model needs to first identify objects and sound sources in the video, and then count all sounding objects. Although there are three different sound sources in the audio modality, only two of them are visible. Rather than simply counting all audio and visual instances, exploiting audio-visual association is important for AVQA.

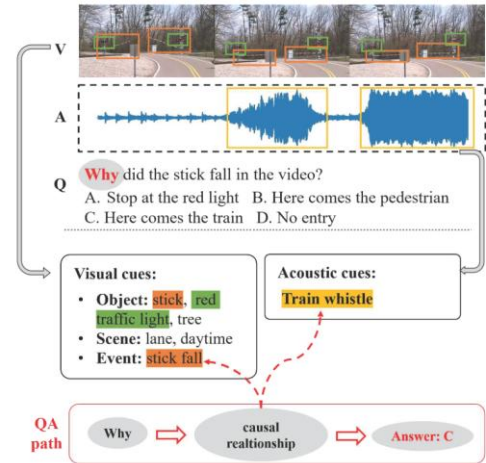
- b. Audiocaps

[Audio Classification] rumble | vehicle | speech | car | outside

[Video Captioning] A bus passing by with some people walking by in the afternoon.

[Audio Captioning] A muffled rumble with man and woman talking in the background while a siren blares in the distance.

- c. AVQA
 - i.



3. Structure of results –

Model	Baseline model	Proposed model 1	Proposed model 2	average
Dataset 1	accuracy			
Dataset 2				
Dataset 3				

- a. Audio Present
- b. Audio Masked

- c. Audio Swapped (Wrong) – here the causal link should break so a low accuracy will be there for a good model.
- 4. Models
 - a. [Image bind](#) – meta learning models by meta.
 - b. Base VLM – <https://arxiv.org/pdf/2312.17432> lot of model in this any we can use (fig 5)
 - c. Audio encoder – [Encodec](#), SAM Audio encoder, Wav2Vec 2.0(2020 old), AST: Audio Spectrogram Transformer (2021 old)
 - d. Fusion mechanism:
 - i. Early fusion (audio + visual embeddings)
 - ii. Late fusion (separate encoders + cross-attention)
 - iii. Multimodal transformer
- 5. Evaluation –
 - a. Total Effect (TE): The overall change in the output when an input (e.g., audio) changes.
 - b. Natural Direct Effect (NDE): How much the output changes based on the audio alone, keeping the video "frozen" at its original state.
 - c. Natural Indirect Effect (NIE): How much the audio influences the output through its effect on the visual perception.
 - d. Modality Sensitivity (The "Mute" Test)
 - i. **Baseline:** Run the full model (Audio + Video + Text).
 - ii. **Intervention:** Set the Audio embedding to zero (or white noise).
 - iii. **Analysis:** If the model's prediction remains the same with 99% confidence, the audio has **zero causal influence**, even if the model has high accuracy.
 - e. Counterfactual Robustness (The "Switch" Test)
 - i. This is the "gold standard" for causal VLM evaluation.
 - ii. **Setup:** Find two videos with different audio-driven labels (e.g., Video A: "Bird Chirping", Video B: "Chainsaw").
 - iii. **Execution:** Swap the audio. Pair Video A with Audio B.
 - iv. **Success Metric:** A causally sound model should change its prediction to "Chainsaw." If it still says "Bird," it is over-relying on visual cues (Visual Bias).
 - f. [CausalVLBench](#) – by changing the variable in the causal discovery graph.
- 6. Visualisations –
 - a. Audio-Visual Heatmaps: Use Grad-CAM or Integrated Gradients to show which audio frequencies (on a spectrogram) triggered which bounding boxes in the video.
 - b. Causal Discovery Graphs: Use algorithms like PC or IC* to automatically generate a graph from your model's hidden states to show the flow of information.

Reference –

1. For types of video language models - <https://link.springer.com/article/10.1007/s11263-025-02385-8>