Pano-AVQA: Grounded Audio-Visual Question Answering on 360° Videos

Heeseung Yun[1], Youngjae Yu[2], Wonsuk Yang[3], Kangil Lee[4], Gunhee Kim[1]

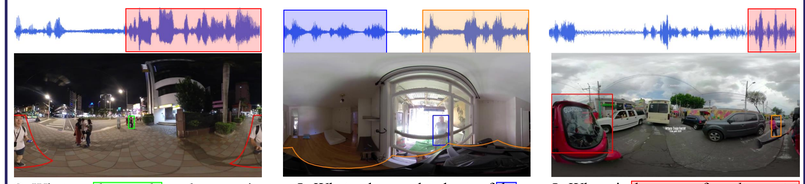
[1] Seoul National University, [2] Allen Institute for AI, [3] University of Oxford, [4] Hyundai Motor Company

2021 ICCV OCTOBER 11-17 VIRTUAL

https://github.com/hs-yn/PanoAVQA

Overview

- ✓ 360° videos convey holistic view for the surroundings
- ✓ Applications: autonomous vehicles, AR/VR, Automatic cinematography [1], saliency detection [2], audio spatialization [3], etc.



Q. What are the people on the opposite side of the *talking* man doing? A. Walking.

Q. What color are the shorts of the bald man *speaking in a grave voice*? A. White.

Q. Where is the cause of *rustling noise* in relation to a woman in pink clothes? A. Right.

- ✓ First large-scale audio-visual QA dataset on 360° videos
- ✓ Spherical spatial reasoning and audio-visual reasoning
- ✓ 51.7K QA pairs with answer grounding
- ✓ Possible extension from audio-visual scene graphs to embodied AI

Challenges

Spherical Spatial Reasoning

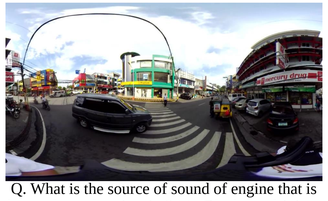
- ✓ Non-Euclidean spatial relations
- ✓ Distortion, discontinuity on a sphere
- ✓ Relative spatial relation without fixed orientation

Audio-Visual Reasoning

- ✓ Audio-visual cues beyond normal FoV
- ✓ Fine-grained audio-visual relationships from the videos in the wild

Q. Where is the cause of speech of a female *in relation to* a man with sunglasses? / A. **Right**.

Q. What color is the shirt **of the man who speaks first**? / A. Black.

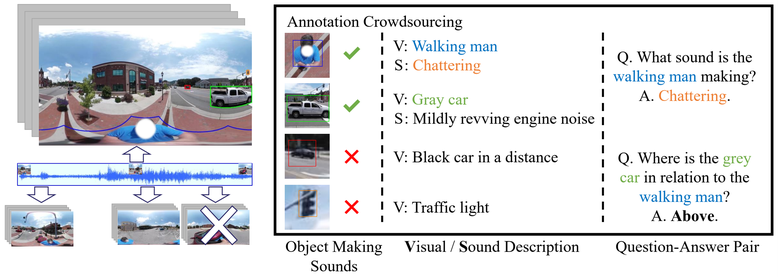


Q. What is the source of sound of engine that is **beneath** an electric wire? / A. Black car driving.

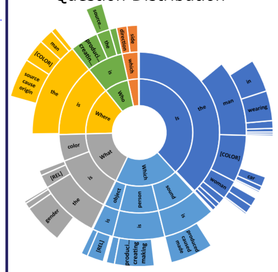
Q. Is a man wearing a mask **telling** something? / A. *No*.

Pano-AVQA Dataset

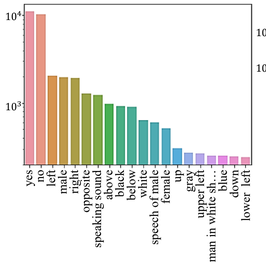
- ✓ 360° video clips (+multichannel audio) with audiovisual saliency
- ✓ 3-stage pipeline to facilitate annotation



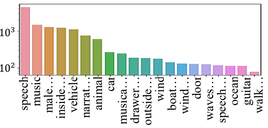
Question Distribution



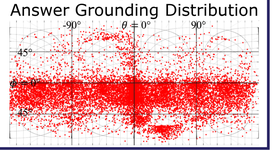
Answer Distribution



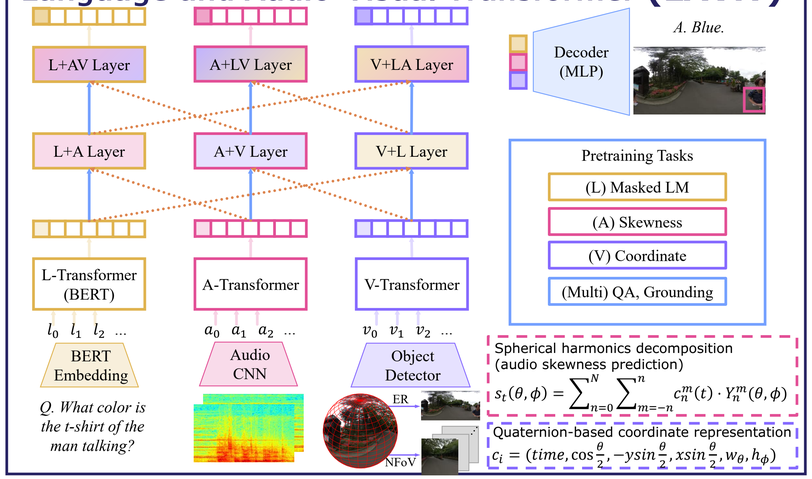
Audio Tag Distribution



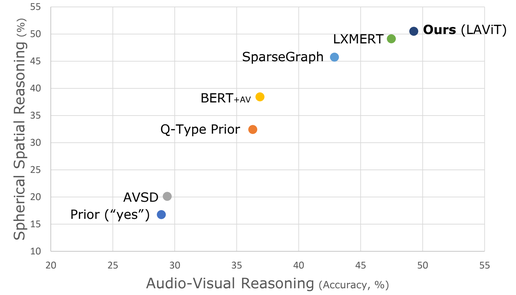
Answer Grounding Distribution



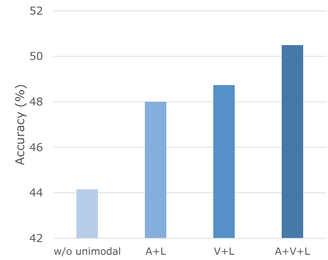
Language and Audio-Visual Transformer (LAViT)



Experiments

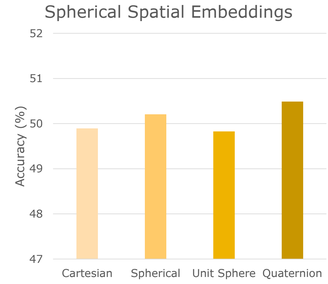


Influence of Modality



Influence of FoV

Spherical Spatial Embeddings



Q. What is below a man with striped shirt?
Answer: Selfie stick / Prediction: Camera

Q. Is a conversation sound created by a man wearing black mask?
Answer: Yes / Prediction: Yes

GT
Cartesian
Angular
Spherical
Quaternion

References

[1] Su et al. Pano2Vid: Automatic Cinematography for Watching 360 Videos. In ACCV, 2016.
[2] Cheng et al. Cube Padding for Weakly-Supervised Saliency Prediction in 360 Videos. In CVPR, 2018.
[3] Morgado et al. Self-Supervised Generation of Spatial Audio for 360 Video. In NIPS, 2018.