2021 ICCV OCTOBER 11-17 VIRTUAL

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

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, Gunhee Kim

SEOUL NATIONAL UNIV.
VISION & LEARNING

UNIVERSITY OF OXFORD

HYUNDAI

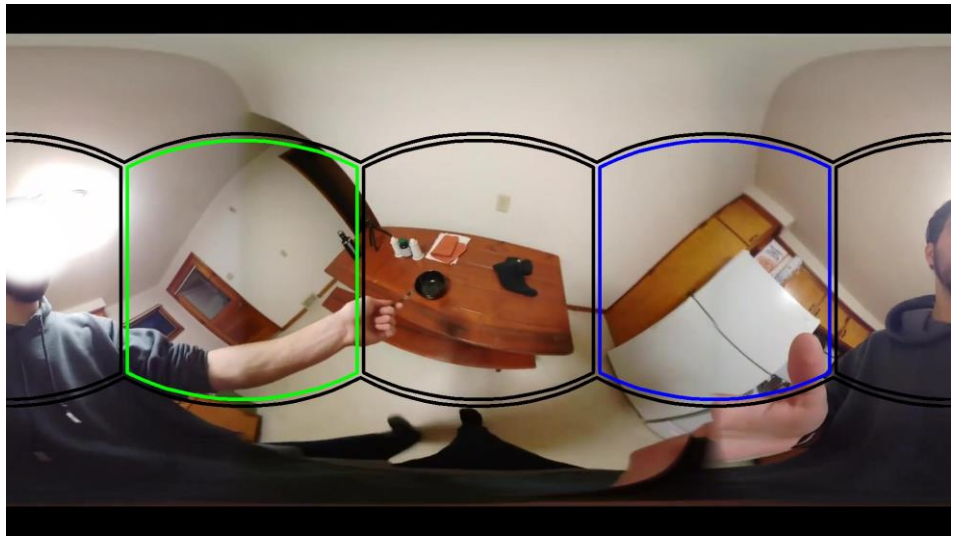
Motivation

- 360° Videos
  - Convey holistic views for the surroundings
  - Applications
    - Autonomous vehicles, robotics, AR/VR
    - Automatic cinematography[1], saliency detection[2], audio spatialization[3]

[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.

Motivation

- Challenges
  - Spatial relations on a sphere (distortion, discontinuity, etc.)
  - Audio-visual cues beyond normal field of views (NFoV)



Where is the door in relation to the fridge?
→ The door is on the opposite side of the fridge.

NFoV

Who is talking?
→ A man in green jacket is talking, not the man with a horse.

Motivation

- **Pano-AVQA**
  - First large-scale audio-visual QA dataset on 360° videos
  - Spherical spatial reasoning and audio-visual reasoning
  - 51.7K QA pairs with bounding box grounding

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

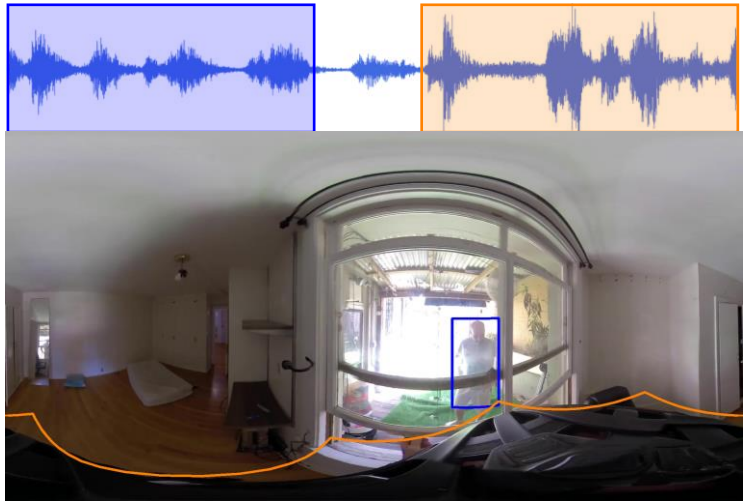
Q. What color are the shorts of the bald man *speaking in a grave voice*?

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

Motivation

- **Pano-AVQA**
  - First large-scale audio-visual QA dataset on 360° videos
  - Spherical spatial reasoning and audio-visual reasoning
  - 51.7K QA pairs with bounding box grounding

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.

Pano-AVQA: Definition

- Spherical spatial reasoning
  - *Relative* spatial relation without fixed orientation
    - *i.e.,* left/right to, opposite of, above/below, next to

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

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

Pano-AVQA: Definition

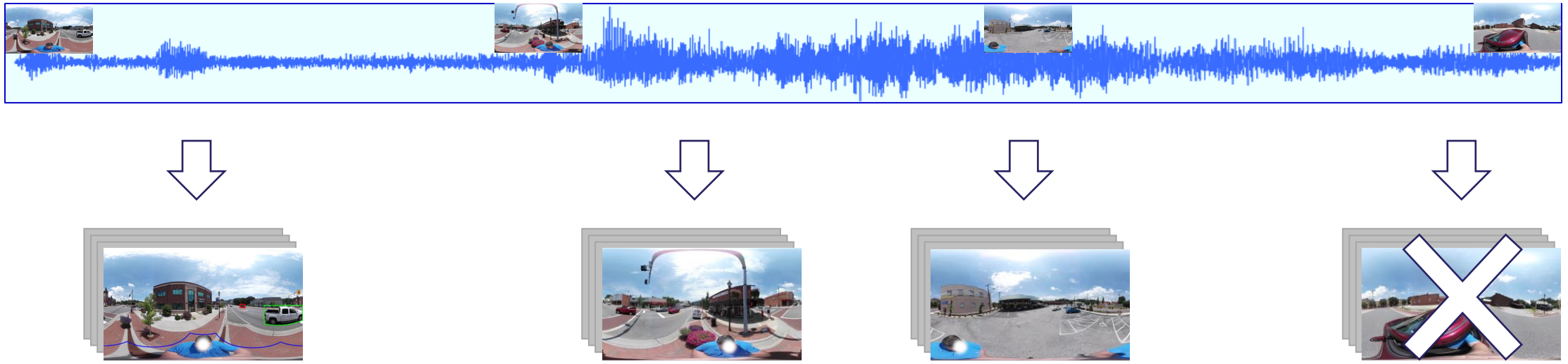
- Audio-visual reasoning
  - Identify the object from sound and vice versa
  - Fine-grained audio-visual relationships from the videos in the wild

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

Q. Is a man wearing red shirt and black mask **talking**?
A. *No*.

Pano-AVQA: Collection

- 360° videos harvested with diverse queries
  - Equirectangular format with multichannel audio
  - Extract & filter 5s clips with audio-visual saliency



Pano-AVQA: Annotation

- 3-stage pipeline
  - Less cognitive burdens for annotators
  - Bounding boxes from off-the-shelf object detector
    - Equirectangular + NFoV

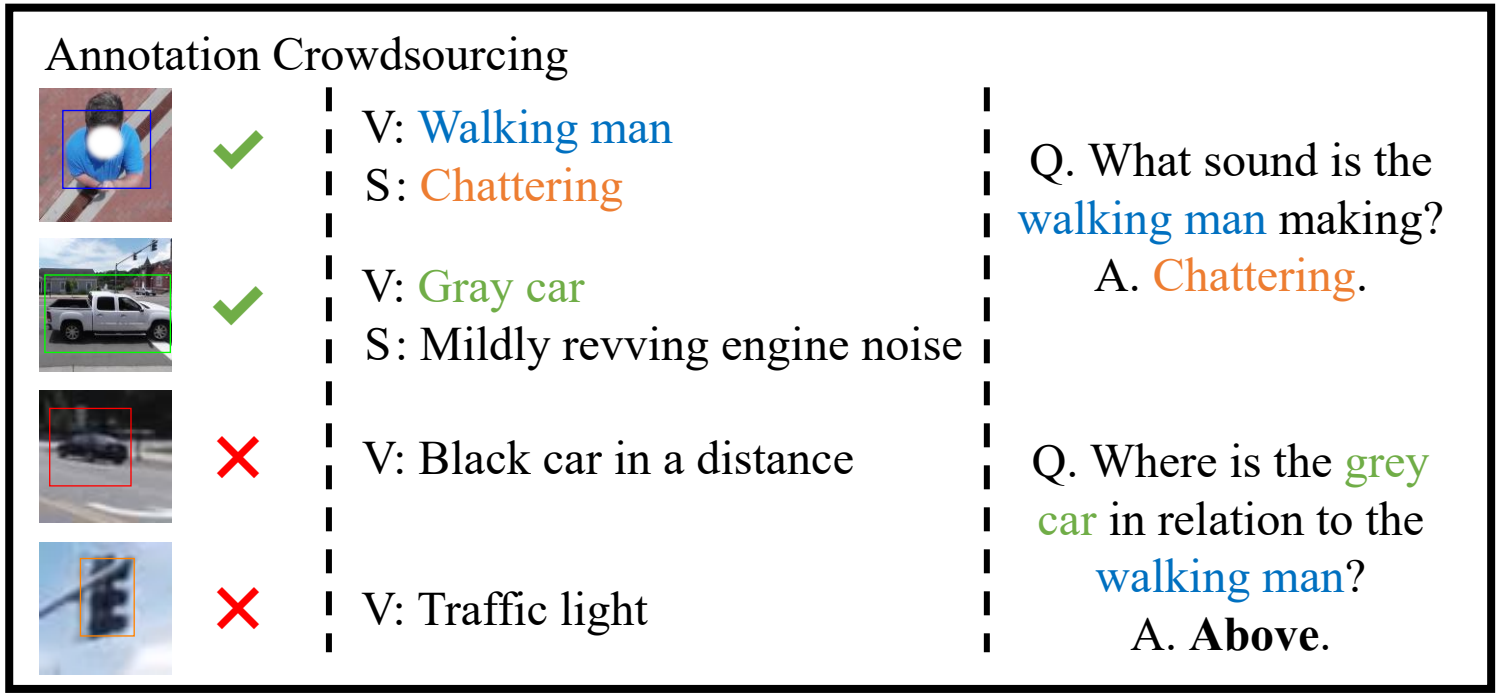
$$for\ (x,y) \in [-1,1]^2, (\theta,\phi) \in (-\pi,\pi) \times \left(-\frac{\pi}{2},\frac{\pi}{2}\right),$$

$$f(x,y) = \frac{M(\theta,\phi)\cdot(1,x,y)^t}{||M(\theta,\phi)\cdot(1,x,y)^t||},$$

$$M(\theta,\phi) = \begin{pmatrix} cos\theta cos\phi & -sin\theta & -cos\theta sin\phi \\ sin\theta cos\phi & cos\phi & -sin\theta sin\phi \\ sin\phi & 0 & cos\phi \end{pmatrix}.$$

Pano-AVQA: Annotation

- 3-stage pipeline
  - Less cognitive burdens for annotators
  - Question-answer pairs
  - Validation & postprocessing

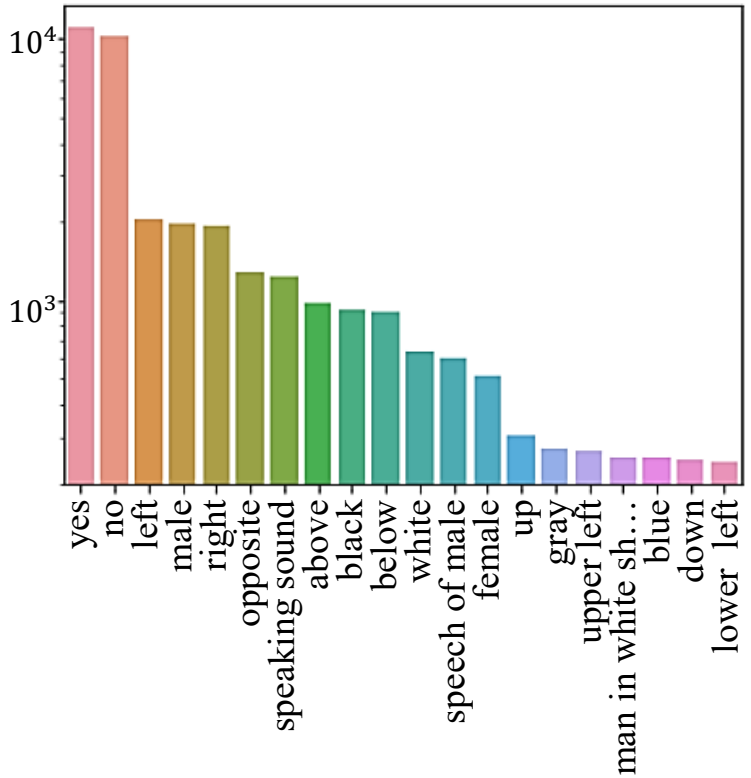


Pano-AVQA: Analysis

- Statistics
  - 51.7K QA pairs (39% SSR, 61% AVR)
  - Average length: question (12.1 words), answer (3.7 words)

Question Distribution

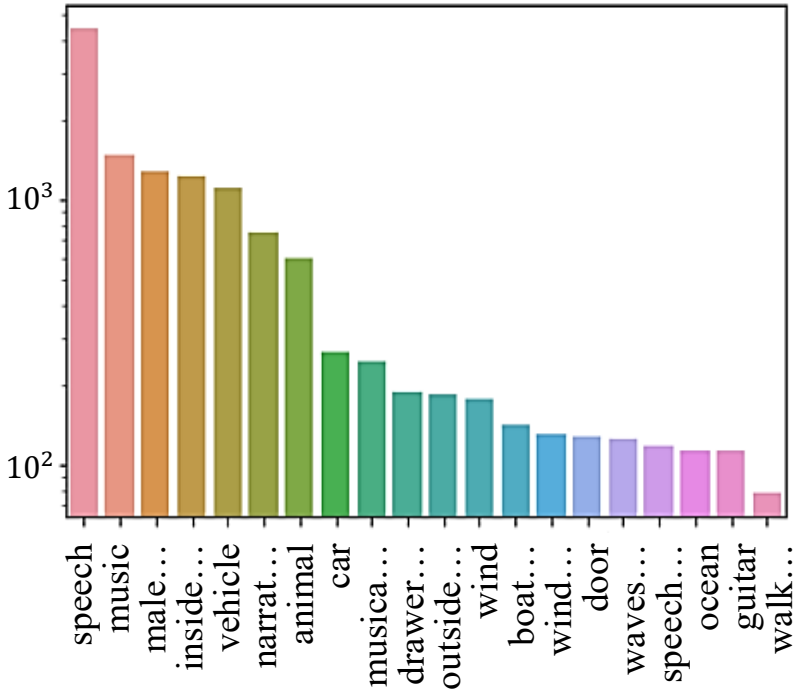
Answer Distribution



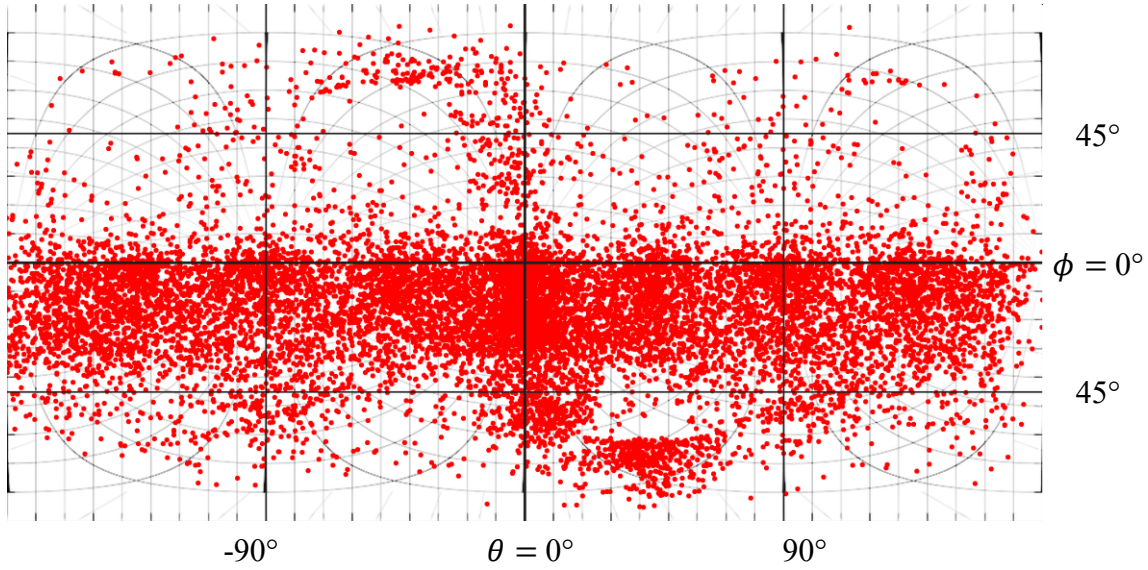
Pano-AVQA: Analysis

- Statistics
  - Various audio sources (top-3 Audioset taggings[1])
  - Diverse answer grounding distribution on a sphere

Audio Tag Distribution



Answer Grounding Distribution

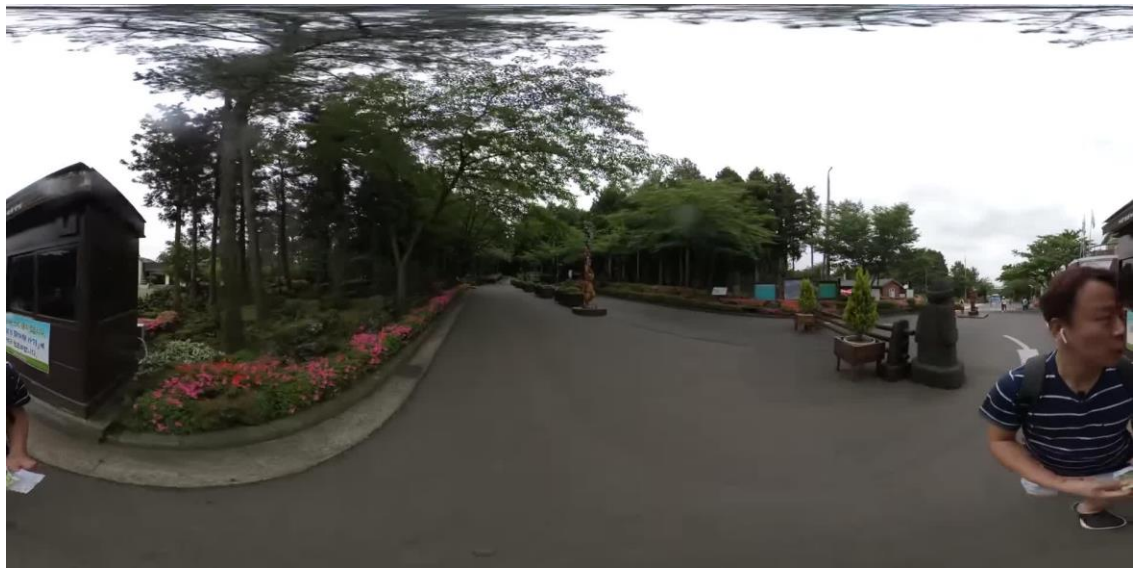


[1] Gemmeke et al. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In ICASSP, 2017.

LAViT

- **L**anguage and **A**udio-**Vi**sual **T**ransformer

Q. What color is the t-shirt of the man talking?



LAViT

- Input Representation

Language

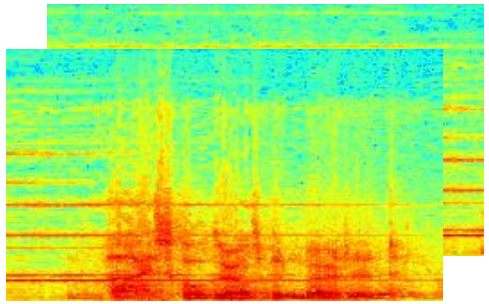
Q. What color is the t-shirt of the man talking?

BERT Embedding

[CLS], What, color, ...

l_0 l_1 l_2 ...

Audio

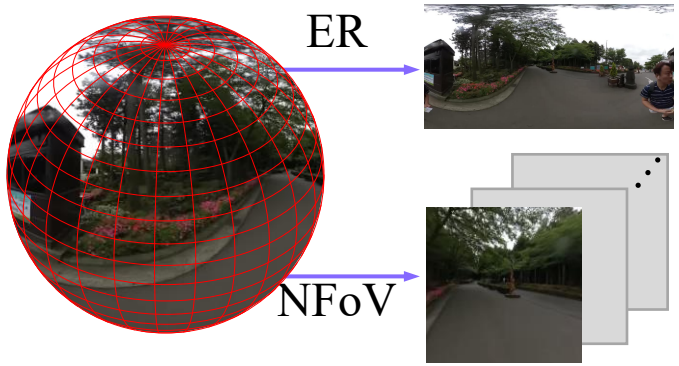


Audio CNN

[CLS],  ,  , ...

a_0 a_1 a_2 ...

Video



Object Detector

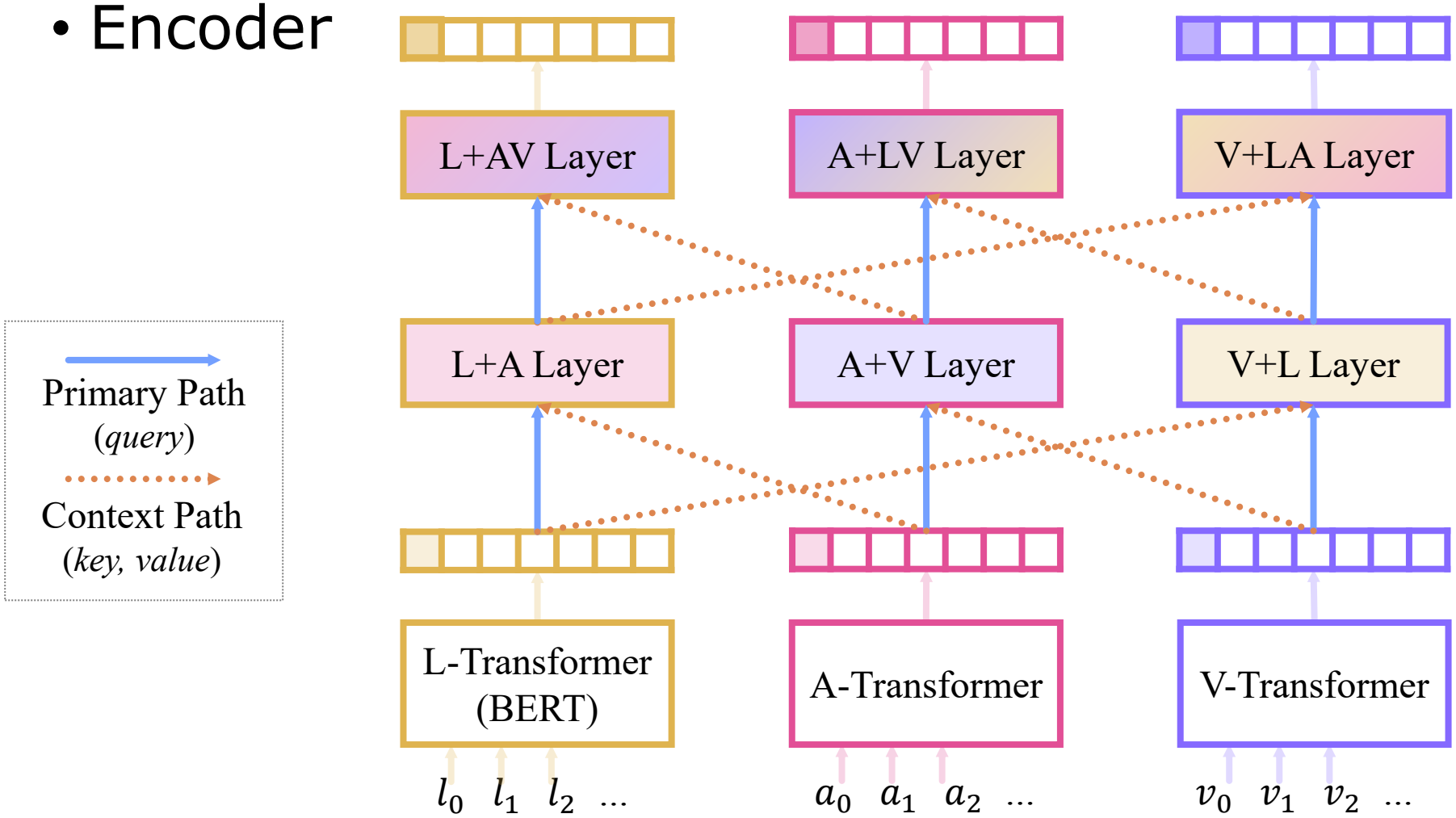
[CLS],  ,  ...

v_0 v_1 v_2 ...

$$c_i = (time, \cos\frac{\theta}{2}, -ysin\frac{\theta}{2}, xsin\frac{\theta}{2}, w_\theta, h_\phi)$$

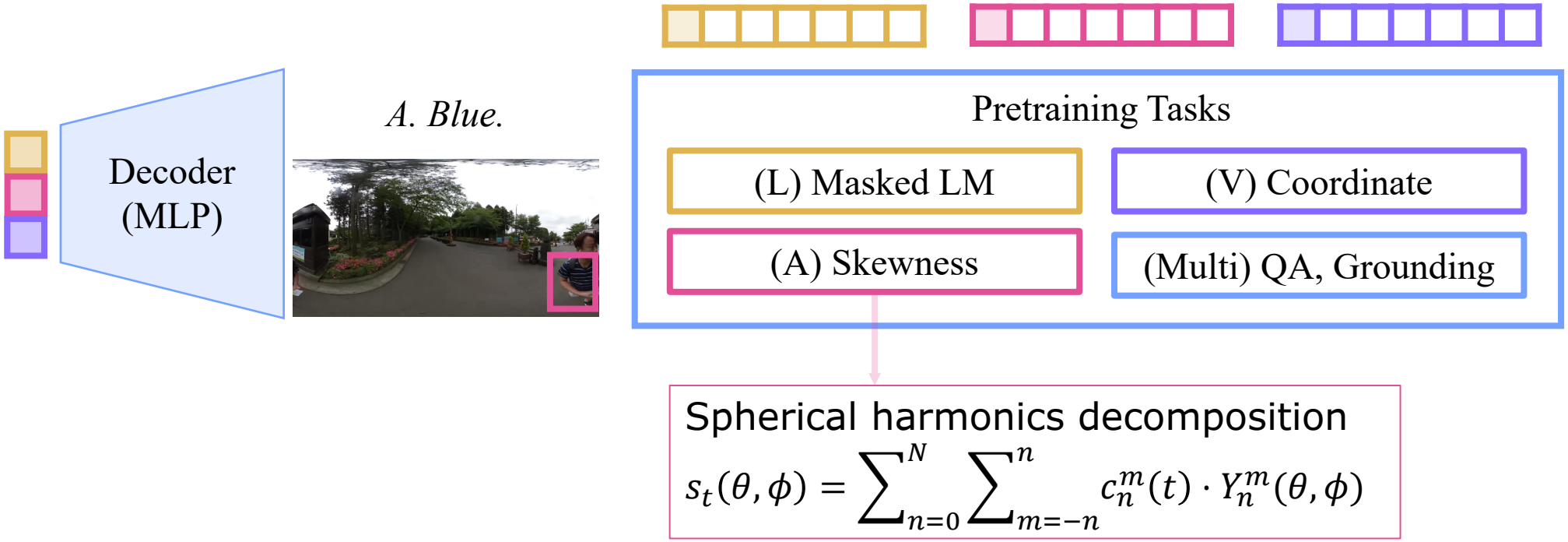
LAViT

- Encoder



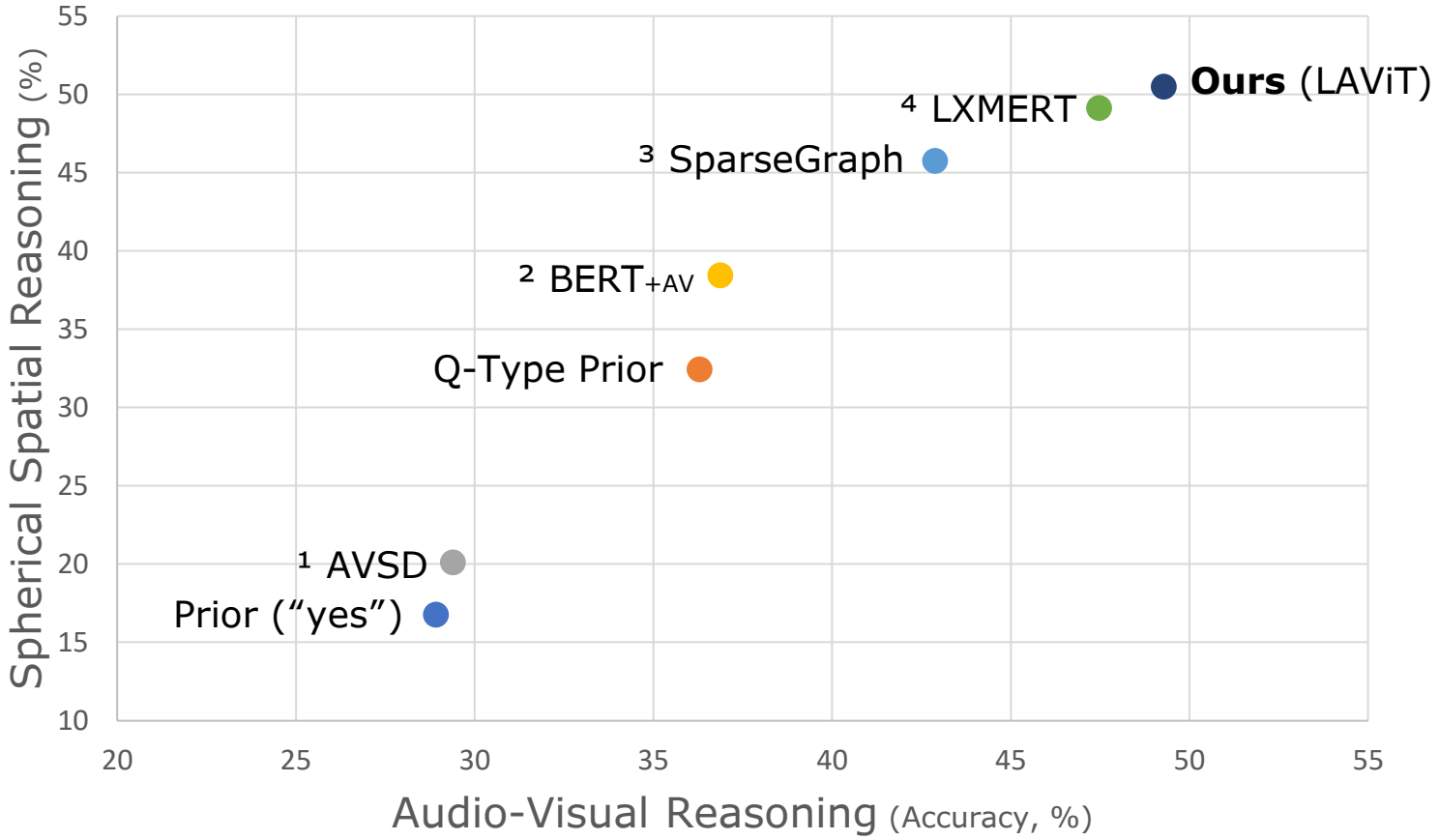
LAViT

- Decoder & Training
  - 2020-D answer classification
  - 5 Pretraining tasks + 2 fine-tuning tasks



Experiments

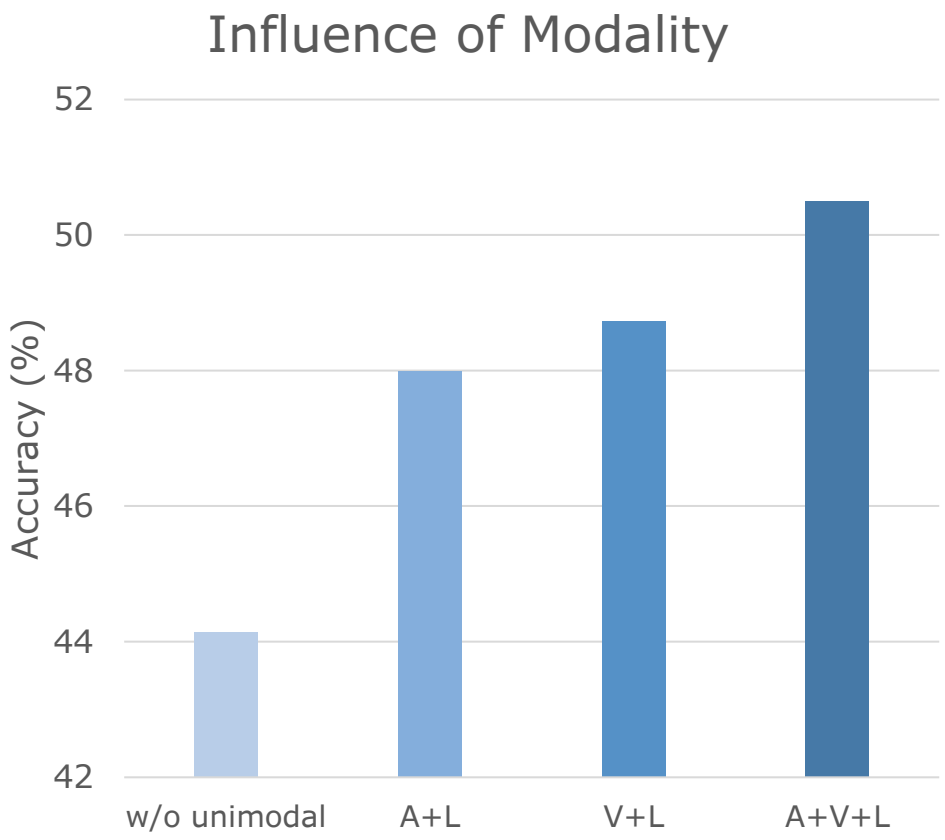
- Comparison with existing VQA models

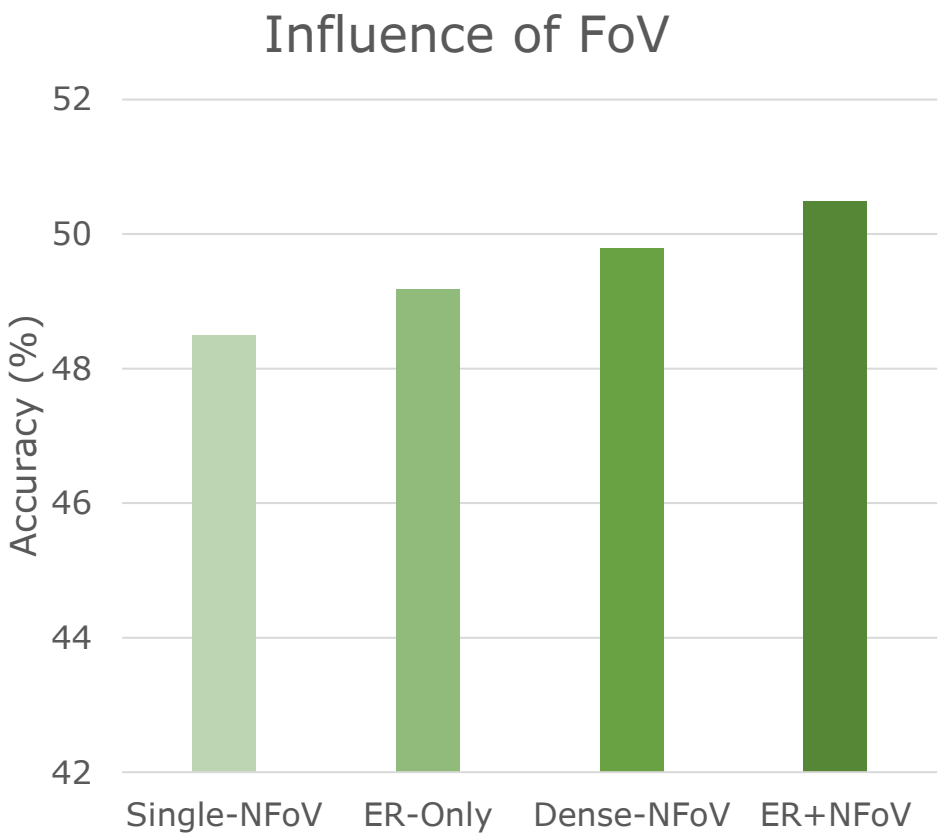


[1] Alamri et al. Audio Visual Scene-Aware Dialog. In CVPR, 2019.
[2] Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL, 2019.
[3] Norcliffe-Brown et al. Learning Conditioned Graph Structures for Interpretable VQA. In NeurIPS, 2018.
[4] Tan and Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In EMNLP, 2019.

Experiments

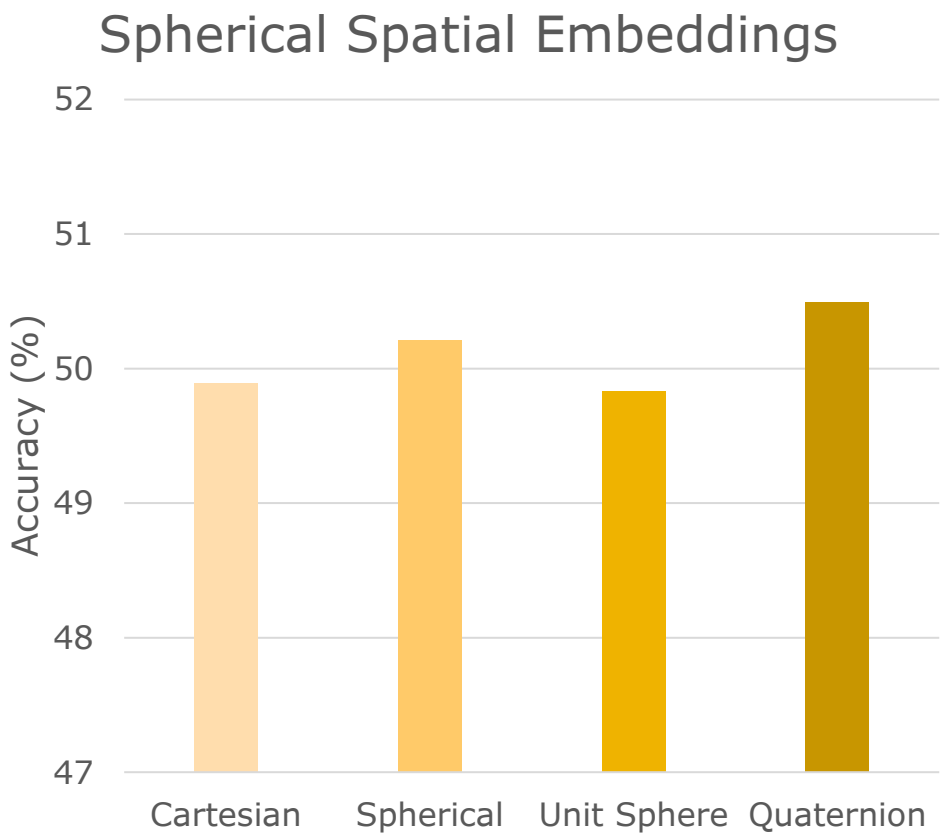
- Influence of modality / visual input



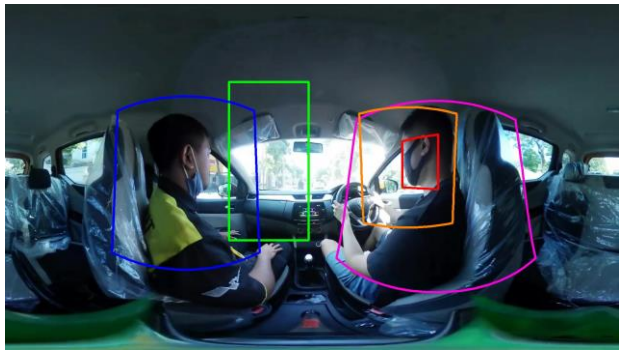


Experiments

- Influence of 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

Concluding Remarks

- Novel, large-scale audio-visual question answering dataset on 360° videos
- LAViT model for 360° video question answering
- Extension
  - audio-visual scene graphs
  - embodied AI (QA, navigation, etc.)

Thank You

Data, code available at 👇
https://github.com/hs-yn/**PanoAVQA**