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

In Appendix, we provide the details of the Pano-AVQA dataset construction, implementation details and experiments that are not fully described in the main paper.

A. Details on Dataset Construction

We provide details of clip filtering, annotation quality control, spherical spatial relation, and question-answer generation.

A.1. Filtering Clips

From raw 360° videos crawled online, we extract audio-visual clips as follows. We first ensure that each peak of audio amplitude is at least 5 seconds apart from another. To compare clips with different length, we compute dynamic time warping (DTW) with l_2-distance and compare the first feature vectors of Mel-frequency coefficients. As a result, we can reduce the chance of obtaining clips with similar sounds from a single video.

Second, computer-generated frames barely provide any visual context. To filter out such clips, we extract frames at a sampling rate of 1fps and discard any frame whose color histogram maximum is more than 0.5 (*e.g*., more than 50% of pixels are exactly the same in color). Any clips containing such a frame are flagged as synthetic and therefore discarded.

Clips may have salient audio peaks but simultaneously can be visually uninteresting. In particular, a non-negligible number of clips have few visual transitions, *i.e*., almost static. We thereby compute 64bit DCT image hash using pHash of each frame. Clips with less than three hash values are deemed as static and are neglected.

A.2. Annotation Quality Control

In total, 436 workers have participated in our data collection. On each sub-task, the workers have to pass the qualification test before the main task so that potentially underperforming workers are filtered out. To evaluate the workers' understanding of each sub-task in qualification tests, we create multiple-choice question from correct annotations collected from a small set of video clips to. Along with the qualification tests, we consistently monitor the submitted answers, approving eligible submissions while prohibiting abusive or dishonest workers from participating in our tasks. Fig. 7 is the user interface for collecting short descriptions in the Pano-AVQA dataset.

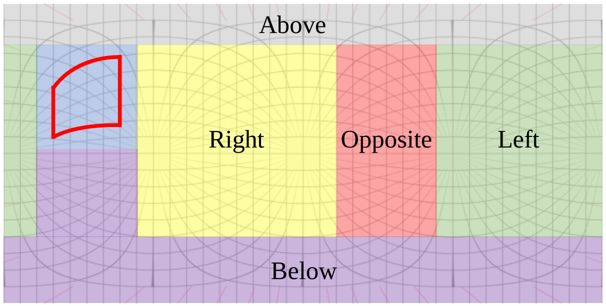


Figure 6. An illustration of spatial relations on a sphere. Since the principal orientation is not fixed in 360° videos, spatial relations change with respect to the reference object's coordinates (denoted by □.)

A.3. Spherical Spatial Relations

Fig. 6 visualizes our spherical spatial relation generation process. As discussed in Sec.3.1, 360° videos lack any principal orientation. Instead of assuming a fixed orientation for every 360° video, we first select a reference object and obtain the most appropriate spherical spatial relation between the two objects. We discard any objects overlapping more than $\frac{\pi}{8}$ or objects with ambiguity (*e.g*., objects belonging to more than one spatial category).

A.4. Question-Answer Generation

Table 4 enumerates question-answer templates used in our dataset, where permutations of word ordering or minor paraphrasing like active/passive voice are omitted for simplicity. Tokens are defined as follows:

- [SUB]: short description about the subject
- [OBJ]: short description about the object
- [SND]: short description about sound
- [REL_q]: relational expressions used in questions (*e.g*., left/right to, to the left/right of, etc.)

Task	Question	Answer
SS	Is [SUB] [REL_q] [OBJ]?	Yes/No
	Where is [SUB] in relation to (with respect to) [OBJ]?	[REL_a]
	On which side(direction) of [SUB] does [OBJ] exist?	[REL_a]
	Who(What) is [REL_q] [OBJ]?	[SUB]
	What color is [SUB] that is [REL_q] [OBJ]?	[color]
	What is [SUB] [REL_q] [OBJ] wearing?	[clothing]
	What is [SUB] [REL_q] [OBJ] doing?	[action]
AV	Is [SUB] [SPEECH]?	Yes/No
	Is [SUB] [SOUND] [SND]?	Yes/No
	Is [SND] [SOUND] by [SUB]?	Yes/No
	What is the gender of the person [SPEECH]?	Male/Female
	Who(What) is [SOUND] [SND]?	[SUB]
	Which sound is [SUB] [SOUND]?	[SND]
	Is [SND] [SOUND] by [SUB]?	Yes/No
	Where is the source/origin/cause of [SND] in relation to [OBJ]?	[REL]

Table 4. Question-answer templates used for constructing the Pano-AVQA dataset.

- [REL_a]: relational expressions used in answers (*e.g.*, above, below, etc.)
- [color], [clothing], [action]: keywords related to colors, cloths, actions, respectively
- [SPEECH]: speech-related verbs (*e.g.*, talking, speaking, chattering, etc.)
- [SOUND]: sound-related verbs (*e.g.*, making, causing, etc.)

Question templates with binary answers are augmented with unrelated visual or sound descriptions for balancing the answer distribution as well as reducing bias discussed in Sec.3.3. Fig. 8 displays generated question-answer examples of the Pano-AVQA dataset.

B. Implementation Details

B.1. Input Representation.

Spherical Non-Max Suppression. We first extract region proposals from both equirectangular (ER) and 18 NFoV projections, where NFoV point-of-views are from the equator (*i.e.*, $\phi = 0$, $\theta \in \{-\frac{3}{4}\pi, -\frac{\pi}{2}, -\frac{\pi}{4}, 0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3}{4}\pi, \pi\}$), southern hemisphere (*i.e.*, $\phi \in \{-\frac{\pi}{4}, -\frac{\pi}{2}\}$, $\theta \in \{-\frac{\pi}{2}, 0, \frac{\pi}{2}, \pi\}$), and northern hemisphere (*i.e.*, $\phi \in \{\frac{\pi}{4}, \frac{\pi}{2}\}$, $\theta \in \{-\frac{\pi}{2}, 0, \frac{\pi}{2}, \pi\}$). Region proposals extracted from each NFoV are then converted from local (*i.e.*, NFoV) to global (*i.e.*, ER) coordinates. We then apply non-max suppression using region proposal polygons with the threshold of 0.65.

Harmonics Decomposition. As discussed in Sec.4.3, we regard the stereo audio channel as 3D audio with two silent channels and apply spherical harmonics decomposition. For spherical harmonics Y_n^m, we utilize truncated spherical harmonics decomposition of an audio $s_t(\theta, \phi) = \sum_{n=0}^{N} \sum_{m=-n}^{n} c_n^m(t) \cdot Y_n^m(\theta, \phi)$ to extract the coefficient c_n^m. To compute spatial skewness τ, we take the sum of the coefficients for cases when $\theta < 0$ (= left) and $\theta > 0$ (= right) respectively, and take their difference. As a result of decomposition, we obtain skewness $\tau \in \mathbb{R}_{[-20,20]}$, where -20 indicates that the audio is coming from the rightmost point of the sphere and vice versa. Finally, we map τ from $\mathbb{R}_{[-20,20]}$ to $\mathbb{R}_{[-1,1]}$.

B.2. Computation Environment

We summarize some information about computing infrastructure for our experiments.

- GPU: NVIDIA TITAN RTX
- CPU: Intel(R) Xeon(R) Gold 6130 CPU
- OS: Ubuntu 16.04 LTS
- RAM: SAMSUNG DDR4 8G
- Relevant software libraries: Anaconda distribution of python ($\approx$ 3.8) and PyTorch ($\approx$ 1.7)

Please refer to our source code for more details.

C. Additional Experiments

In this section, we experiment on the effectiveness of pretraining tasks and audio representation. We also present qualitative examples generated by our LAViT from the Pano-AVQA *validation* split in Fig. 9.

C.1. Experiment on Pretraining Tasks

To explore the influence of pretraining tasks, we train a few ablation variants of our model: (i) LAViT$_{w/o\,A}$ pretrained without answers, (ii) LAViT$_{w/o\,A\,\&\,G}$ pretrained without both answers and grounding tasks and (iii) LAViT$_{w/\,11\,tasks}$ pretrained with 11 possible tasks, which

	MSE	Accuracy (%)		
Model	Ground	SS	AV	All
LAViT$_{w/\ 11\ tasks}$	0.627	48.51	51.22	50.17
LAViT$_{w/o\ A\ \&\ G}$	0.623	47.92	51.07	49.85
LAViT$_{w/o\ A}$	**0.586**	**49.54**	50.98	50.42
LAViT (ours)	0.629	49.29	**51.25**	**50.49**

Table 5. Results of various VQA models on the Pano-AVQA *test* split. SS denotes the spherical spatial reasoning task and AV denotes the audio-visual reasoning task.

		Accuracy (%)		
	Embeddings	SS	AV	All
A	Mono	42.40	50.70	47.50
	Stereo	48.02	50.57	49.58
	Stereo+C	48.51	51.44	50.30
	Stereo+C+F	49.29	51.25	50.49

Table 6. Results of different audio embeddings. C denotes coordinates, and F denotes utilization of different fully-connected layer for different audio channels.

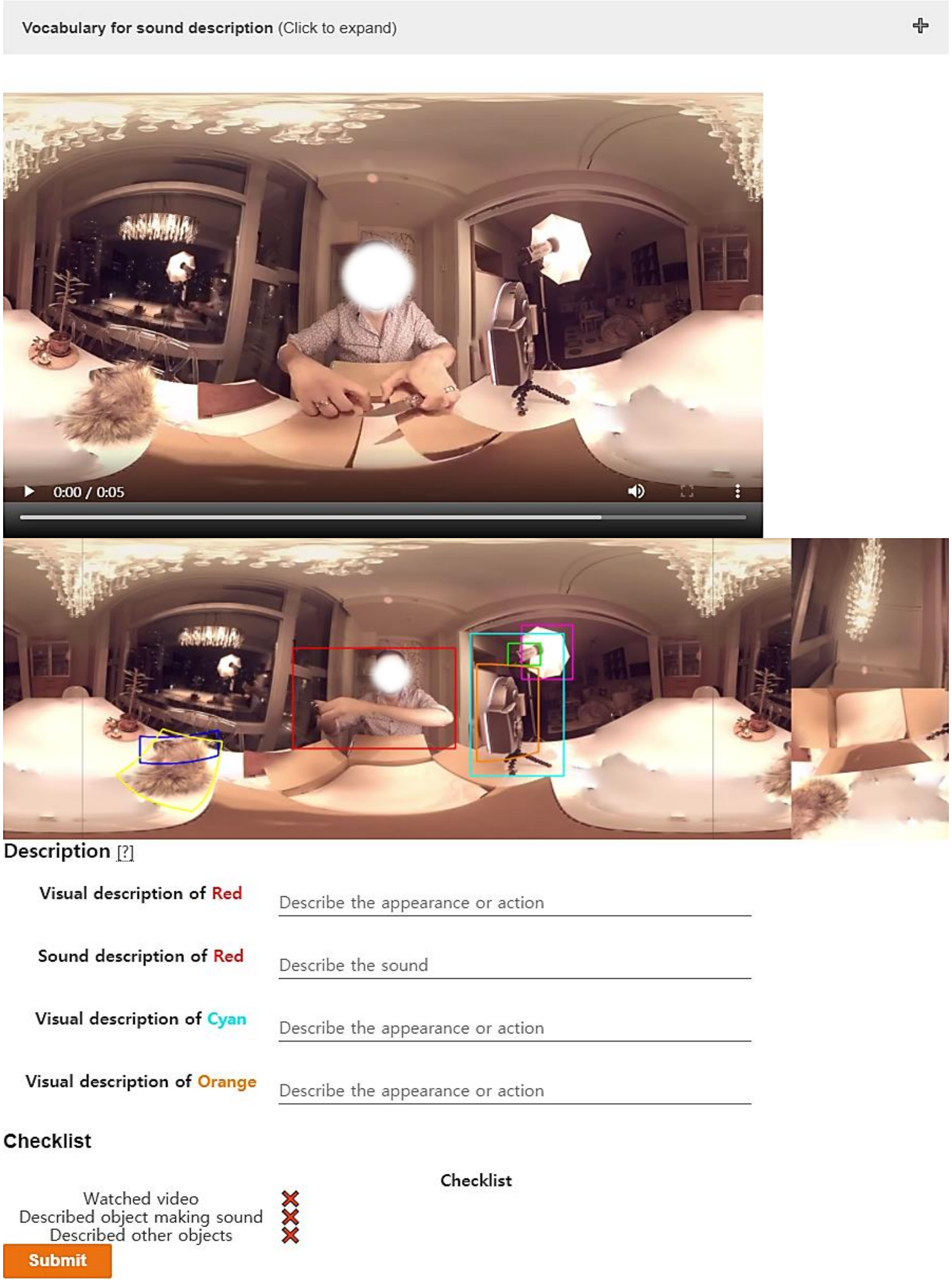
consist of five tasks in Fig.4-(c) as well as four four feature/pseudolabel regression and two multimodal matching tasks that are often used in transformer-based VQA models.

For LAViT$_{w/o\ A}$ that uses only the grounding prediction task results in a slight performance drop of audio-visual reasoning task (AV) accuracy, but the grounding prediction task can be beneficial for pretraining in terms of better answer grounding and spherical spatial reasoning. Excluding both multimodal tasks from the pretraining stage (LAViT$_{w/o\ A\ \&\ G}$) drops performance by 0.64%. Compared to LAViT$_{w/\ 11\ tasks}$, which is trained on twice as many pretraining tasks as our original model, our model still performs marginally better (*i.e.,* 0.32%). This suggests that simply adding more pretraining tasks is not necessarily beneficial to solve the problems of Pano-AVQA.

C.2. Analysis on Audio Representation

To explore the effectiveness of our audio representation, we experiment with a few other possible audio representations: mono audio-only, stereo audio-only, stereo audio with coordinate information, and stereo audio with both coordinate information and different fully-connected layers for each audio channel.

Table 6 outlines the results. Compared to mono audio-only, using stereo audio shows a performance gain of 2%. Using audio with coordinate information as input marginally adds 0.72%, where the influence of using different fully-connected layers for each audio channel is not significant.



Vocabulary for sound description (Click to expand)

Description [?]

Visual description of Red Describe the appearance or action

Sound description of Red Describe the sound

Visual description of Cyan Describe the appearance or action

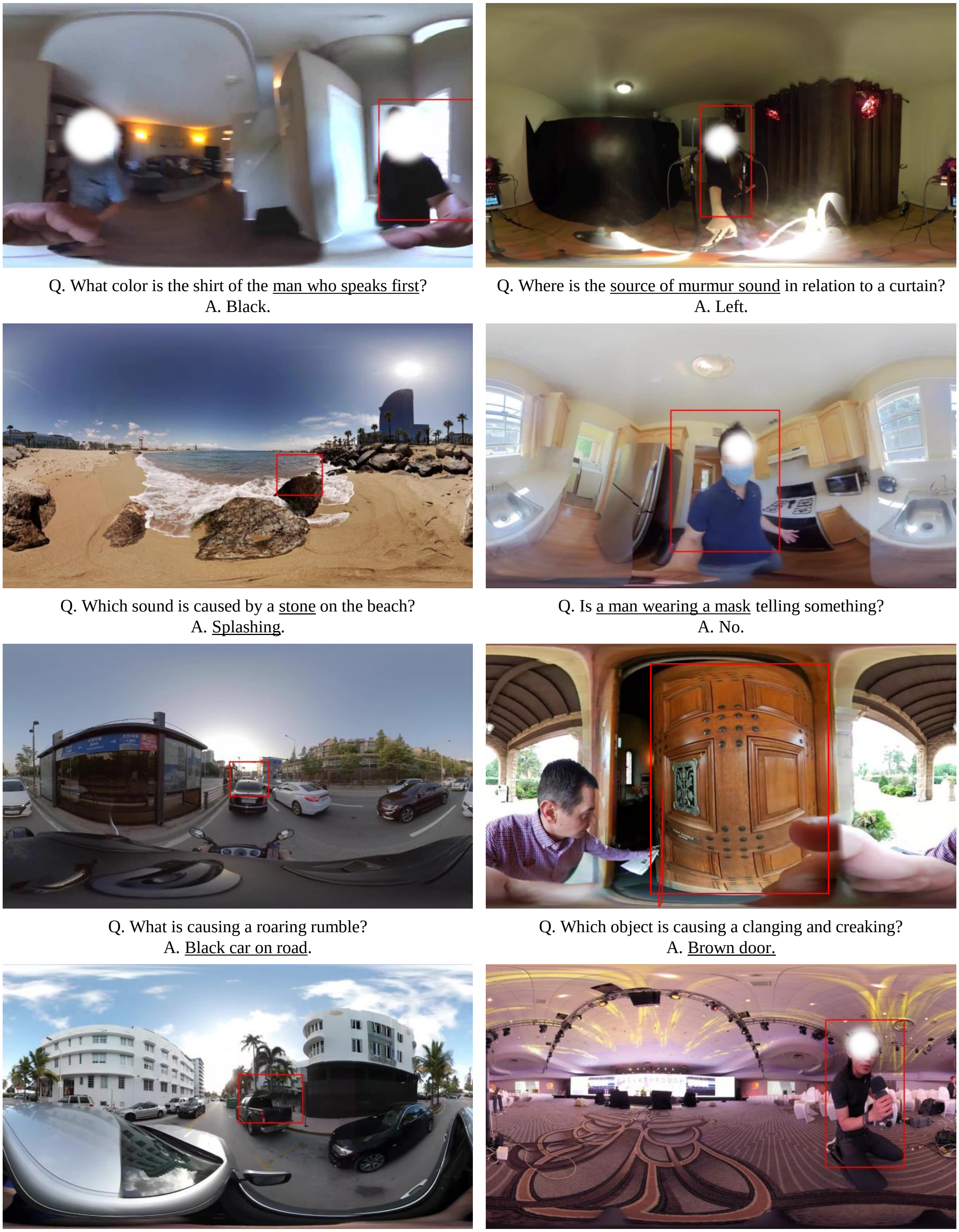
Visual description of Orange Describe the appearance or action

Checklist

Checklist

Watched video ✗
Described object making sound ✗
Described other objects ✗

Figure 7. The AMT user interface for visual and sound description.



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

Q. Where is the source of murmur sound in relation to a curtain?
A. Left.

Q. Which sound is caused by a stone on the beach?
A. Splashing.

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

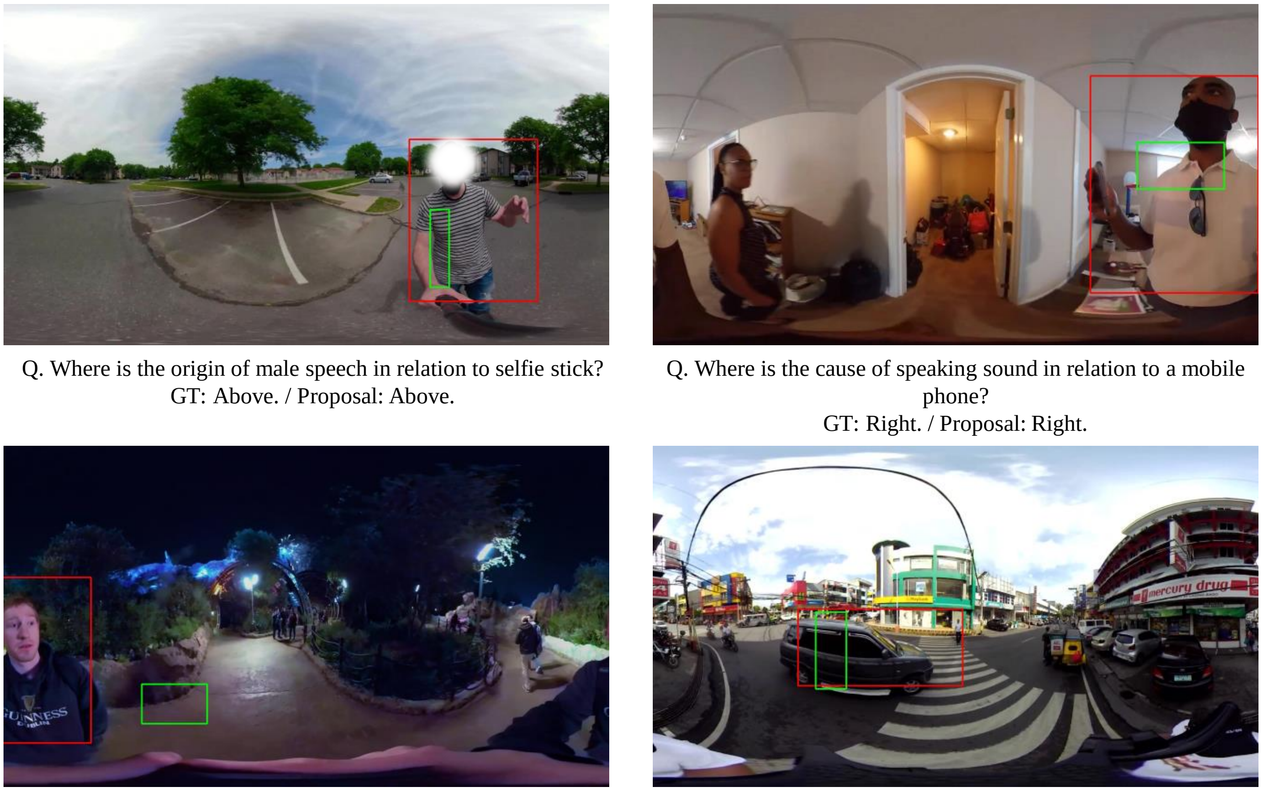
Q. What is causing a roaring rumble?
A. Black car on road.

Q. Which object is causing a clanging and creaking?
A. Brown door.

Q. Where is a black pickup truck in relation to a windshield of a car?
A. Upper right.

Q. Where is a black speaker on the floor in relation to the cause of speech of a male?
A. Opposite.

Figure 8. Question-answer pairs from Pano-AVQA *train* split.



Q. Where is the origin of male speech in relation to selfie stick?
GT: Above. / Proposal: Above.

Q. Where is the cause of speaking sound in relation to a mobile phone?
GT: Right. / Proposal: Right.

Q. What color is a hoodie that is to the opposite of a post lamp?
GT: Black. / Proposal: Black.

Q. What is the source of the sound of engine that is beneath an electric wire?
GT: Black car driving. / Proposal: Car in white color.

Figure 9. Answer and grounding proposals generated by LAViT from question-answer pairs in Pano-AVQA *validation* split. Red denotes ground-truth answer grounding and Green denotes grounding proposal generated by LAViT.