# 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
{heeseung.yun, yj.yu}@vision.snu.ac.kr, {wonsuk1001, smddls77}@gmail.com, gunhee@snu.ac.kr
https://github.com/hs-yn/PanoAVQA

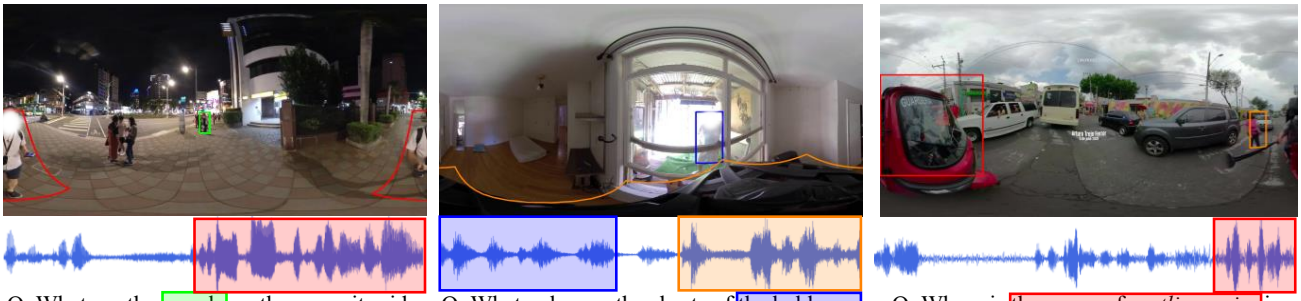



Figure 1. **Pano-AVQA** is a panoramic video question answering dataset for evaluating spherical spatial reasoning and audio-visual reasoning that goes beyond a normal field-of-view with limited context. **Pano-AVQA** introduces diverse new sets of questions from real-life surroundings, considering spherical spatial relations and audio-visual matching.

## Abstract

*360° videos convey holistic views for the surroundings of a scene. It provides audio-visual cues beyond predetermined normal field of views and displays distinctive spatial relations on a sphere. However, previous benchmark tasks for panoramic videos are still limited to evaluate the semantic understanding of audio-visual relationships or spherical spatial property in surroundings. We propose a novel benchmark named Pano-AVQA as a large-scale grounded audio-visual question answering dataset on panoramic videos. Using 5.4K 360° video clips harvested online, we collect two types of novel question-answer pairs with bounding-box grounding: spherical spatial relation QAs and audio-visual relation QAs. We train several transformer-based models from Pano-AVQA, where the results suggest that our proposed spherical spatial embeddings and multimodal training objectives fairly contribute to a better semantic understanding of the panoramic surroundings on the dataset.*

## 1. Introduction

Due to their capacity to capture entire surroundings without a restriction in the field of view, 360° videos have been gaining increasing popularity as a novel medium to record real-life scenery. As illustrated in Fig. 1, unlike conventional normal field-of-view (NFoV) videos, 360° videos allow users to attend to any regions of interest from the original real-life surroundings. As publicly available 360° videos surge from video-sharing platforms (*e.g.*, YouTube) and their applications of omnidirectional perception quickly spread from autonomous vehicles [1, 2], robotics [3, 4] to virtual & augmented reality [5, 6], visual understanding in 360° videos has warranted serious attentions in computer vision research.

The wide field of view of 360° videos brings forth new challenges in visual understanding that are underemphasized in the NFoV video understanding, including spherical spatial reasoning and audio-visual reasoning. Since 360° videos are encoded in a spherical ambient space, spatial reasoning in 360° video, namely spherical spatial reasoning, requires a novel approach to recognizing various relations between the objects all around. Moreover, 360° videos contain more diverse visual sources of sounds than conventional videos, which allows richer contextual audio-visual correspondences. Given that spatial attention for visual and auditory stimuli is inherent and even aligned in human [7], capturing the link among visual and auditory signals from panoramic videos can be highly beneficial to real-life scene understanding.

These two of the main cornerstones of 360° video understanding, namely spherical spatial reasoning and audio-visual reasoning, have been actively addressed by previous works, including automatic cinematography [8], panoramic saliency detection [9, 10], and self-supervised spatial audio generation [11]. Nonetheless, no known task incorporates linguistic queries to tackle the tasks in 360° video domain. To this end, we propose spatial and audio-visual question answering on 360° videos as a novel benchmark task for 360° video understanding.

In this work, we introduce the **Pano-AVQA** dataset as a new 360° video question answering dataset that necessitates fine-grained incorporation of visual, audio, and language modality on panoramic videos. We collect openly available 360° videos from online and annotate them with (audio, video, relationship) description pairs; as a result, we contribute 20K spatial and 31.7K audio-visual question-answer pairs with bounding box grounding from 5.4K panoramic video clips.

Upon this dataset, we propose a transformer [12]-based spatial and audio-visual question answering framework. By attending to the context provided by other modalities throughout training, our model learns to fuse holistic information from the panoramic surroundings. For this, we suggest quaternion-based coordinate representation for accurate spatial representation and an auxiliary task of audio skewness prediction that are broadly applicable to multi-channel audio inputs.

We summarize our main contributions as follows.

1. We propose novel benchmark tasks on spatial and audio-visual question answering on 360° videos towards a holistic semantic understanding of omnidirectional surroundings.

2. Since there is no existing dataset for this objective to the best of our knowledge, we contribute **Pano-AVQA** as the first large-scale spatial and audio-visual question answering dataset on 360° videos, consisting of 51.7K question-answer pairs with bounding box grounding.

3. We design an audio-visual question answering model for 360° videos that effectively fuses multimodal cues from the panoramic sphere. We incorporate this model with several baseline systems and evaluate them on the Pano-AVQA dataset.

## 2. Related Works

**Understanding of Panoramic Videos.** A large body of literature regarding 360° video extends visual understanding of panoramic videos to many pragmatic applications, such as automatic cinematography [8], highlight detection [13], summarization [14], tracking [15] and visual saliency detection [9, 10].

However, most of the prior works concentrate on diverse visual cues present in the panoramic videos. Some of the recent works like narrative description for grounding viewpoints [16], spatial audios for audio augmentation [11] or object removal [17] focus on exploiting modalities other than visual cues. Unlike the prior works, we exploit language queries to evaluate the understanding of audio-visual signals in panoramic videos. We provide a large-scale annotated dataset about panoramic videos, which can be potentially beneficial for audio-visual grounding or scene graph generation in panoramic videos.

**Multimodal Question Answering.** Stemming from image VQA [18], video VQA has been extensively studied [19, 20, 21, 22, 23, 24, 25] towards understanding of visual-linguistic relations in various contexts such as movies [19], TV shows [23, 25], web GIFs [20] and animation clips [22]. Recently, there have been emerging works on answering questions grounded on sound modality, including Diagnostic Audio Question Answering [26], Open-Domain Spoken Question Answering [27] and Audio-Visual Scene-aware Dialog [28].

Closest to our work is AVSD [28], which utilizes both audio and video information in the clip to answer the questions sequentially. Although AVSD evaluates the conversation capability of models, when it comes to audio-visual relationships, AVSD mainly focuses on the existence of sound (*ex. Do you hear any noise in the background?*). On the other hand, Pano-AVQA deals with fine-grained audio-visual relationships like grounding or spatial reasoning in 360° videos (*ex. What is on the opposite side of a loud honking?*). Specifically, we deal with a variety of spatial relations in the panoramic sphere, which sheds new light upon spatial reasoning in videos.

**Audio-Visual Scene Understanding.** Leveraging both audio and video for scene understanding has been broadly researched in the signal processing domain. Early works on multimodal audio-visual learning focused on improving audio-visual speech recognition [29, 30]. Owing to paired videos with audio prevalent in various platforms, recent approaches utilize representation learning of unlabeled videos [31, 32, 33, 34, 35, 36], which is beneficial for various downstream tasks like sound localization [4], audio spatialization [33], audio-visual source separation and co-segmentation [34, 37, 32, 31].

While these approaches showed some successes in audio-visual scene understanding, they assume that the viewpoint is already attended to a salient context. Some of the previous researches focus on audio-visual scene understanding on panoramic videos [35, 4], but they regard panoramic frames as normal ones, ignoring non-negligible distortion present in the panoramic video. Contrarily, we tackle the alignment of audio and video without predetermined context, *i.e.*, normal field of view, thereby con-

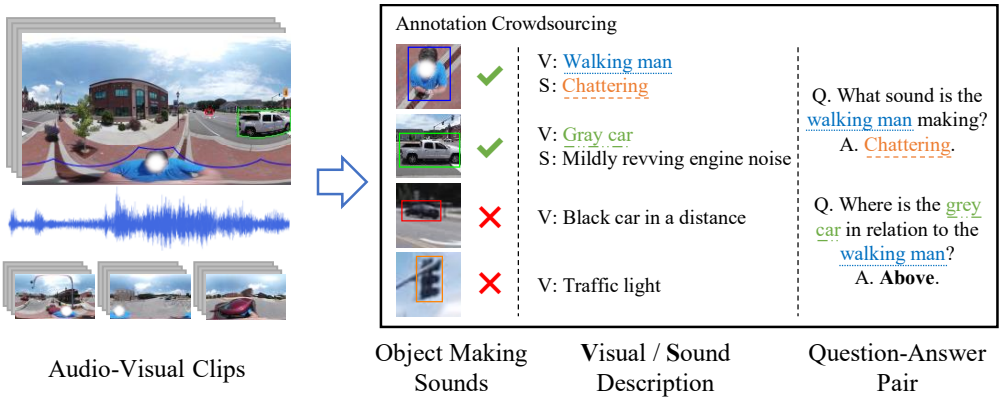



Figure 2. The data collection pipeline of the Pano-AVQA dataset discussed in Sec. 3.

sidering more context in surroundings.

## 3. Pano-AVQA Dataset

The objective of Pano-AVQA dataset is to provide a benchmark for fine-grained spatio-temporal and audio-visual question answering (QA) on panoramic videos. To achieve this goal, each question-answer pair should encapsulate audio signal as well as visual objects in the clip. Since no existing dataset can be used for this objective, we collect the data from scratch.

Fig. 2 illustrates our dataset collection pipeline. From the 360° videos collected online, we extract clips of about 5 seconds, from which we collect three types of annotations from human workers: (a) bounding boxes and sound grounding, (b) visual and sound descriptions, and (c) question-answer pairs. Please refer to the Appendix for the full description of dataset construction.

### 3.1. Task Definition

We introduce two new types of panoramic question answering tasks essential for panoramic scene understanding: (i) spherical spatial reasoning and (ii) audio-visual reasoning, where we design both tasks as open-ended questions. Please refer to Fig. 1 and Appendix for QA pair examples.

**Spherical spatial reasoning** tackles QAs that require recognizing spatial relations between objects in 360° videos. Since 360° videos lack any principal orientation, we only question *relative* spatial relation. That is, we consider the spatial relation of a target object to a reference object. Each answer can be a name or an attribute (*e.g.*, color, action, *etc.*) of the object, or one of the following spatial relations: *left/right to, opposite of, above/below, or next to.* One exemplar template of this task include *Where is [object1] in relation to [object2]? / [relation].*

**Audio-visual reasoning** covers queries about identifying the object from sound and vice versa for a specific visual object and the sound the object is making. Possible answers include the object or sound themselves or their attributes like color or loudness. Two example templates of this task are *Who/what is making [sound]? / [object].* or *Which sound is [object] making? / [sound].*

### 3.2. Data Collection

We collect 360° videos from YouTube using 58 keywords (*e.g.*, sports, tour, indoor, cooking) to foster diversity in context. For consistency, we convert every video into an equirectangular format and discard videos with mono channel audio. For valid audio-visual QA pairs, the video must contain clear, discernible audio signals. Since raw video is often too long and contains uneventful contents, we extract clips of interest spanning five seconds on average. We implement an automated extractor that reads raw audio source and video frames and slices around *audio peaks* whose root mean square amplitudes are greater than those of surrounding segments by at least the standard deviation of the root mean square amplitude of the entire audio.

During extraction, we apply the following filters to ensure the quality of clips. First, we reduce the chance of including similarly sounding clips using the $\ell_2$ distance between Mel-frequency coefficients of each candidate clip. Second, we discard clips containing synthetic or computer-generated frames by inspecting skewness in color histograms. Third, we filter out static clips; we compute the 64bit DCT image hash using pHash of each frame and neglect any clips with less than three hash values. Finally, using off-the-shelf object detector [38], we remove clips with less than three salient objects. In addition to automatic filtering, we inspect any remaining invalidity, including occlusion, post-dubbing, and the existence of background music.

### 3.3. Data Annotations

It can be too cognitively burdensome even for humans to directly create a question-answer pair involving visual and audio features from 360° videos. Therefore, we decompose the entire annotation pipeline into three subtasks

| Dataset | Task | C | # Clips | Length (hr) | Additional information |
|---|---|---|---|---|---|
| Pano2Vid [8] | NFoV cinematography | H | 86 | 7.3 | NFoV videos |
| Deep360Pilot [15] | Object tracking | H | 91 | 1.71 | Object track |
| Yu et al. [13] | Highlight detection | H | 115 | 72 | NFoV videos |
| Lee et al. [14] | Summarization | H | 285 | 92.23 | Photostream |
| Narrated360 [16] | NFoV grounding | H | 864 | 3.98 | - |
| YT-ALL [11] | Audio spatialization | H | 1146 | 113.1 | - |
| REC-STREET [11] | Audio spatialization | R | 43 | 3.5 | - |
| OAP [35] | Object prediction | R | 165 | 15 | - |
| **Pano-AVQA** | Question answering | H | 5.4k | 7.69 | QA with grounding |

Table 1. Comparison of Pano-AVQA with existing 360° video datasets. Column **C** denotes collection procedure, where **H** indicates the dataset harvested online and **R** is the dataset recorded with a custom apparatus.

to reduce complexity while obtaining fine-grained annotations: bounding box collection, visual / sound description, and question answer generation. The results of each subtask are validated before proceeding onto the next subtask.

**Bounding Box Collection.** First, we provide workers with a set of candidate bounding boxes and ask them to choose those that enclose objects that are making a sound. These objects should be either clearly identified as a sound source or humanly inferrable despite occlusion (*e.g.*, man in a mask talking). To obtain candidate bounding boxes, we run Detectron 2 [38] pretrained on ImageNet detection dataset [39] to the central frame of the clip. We pretrain the model from scratch using the ImageNet detection dataset, which includes many sound-making objects such as guitar and drum. To capture objects of different sizes with minimum distortion, we extract bounding boxes from both equirectangular and multiple NFoV projections. We then calibrate the coordinates of the bounding boxes from the perspective projections to the spherical coordinates. Given coordinate $(x, y) \in [-1, 1]^2$ and perspective $(\theta, \phi) \in (-\pi, \pi) \times (-\pi/2, \pi/2)$, we use straightforward yet effective strategy to obtain the spherical coordinate $f(x, y)$:

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

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

**Visual and Sound Description.** Workers are asked to briefly describe (1) the appearances or actions of the annotated objects and (2) the sound they are making (if any). Writing a sound description is not as straightforward as writing a visual description. To assist workers with creating more graphic descriptions, we provide them with sound-describing words (*e.g.*, shout, strum, bang, *etc.*) extracted from audio classification and captioning datasets [40, 41, 42]. We also refrain workers from describing the sound via visual keywords (*e.g.*, *shout in male voice* instead of *man yelling next to a table*) or content of the speech (*e.g.*, *woman explaining the history of the museum*).

**Question Answer Pairs.** Given short descriptions of objects and sounds, we finally create the spherical spatial and audio-visual QA pairs. Following collection practices in existing video QA datasets [20, 43, 44], we combine manual and automated QA generation.

From the collection of object and sound descriptions for each video, we follow the templates discussed in Sec. 3.1 to generate QA pairs. To obtain spatial relations for the spherical spatial reasoning task, we use bounding box coordinates to manually designate the relations between the objects into one of the following categories: *next to, opposite of, left/right to, and above/below*.

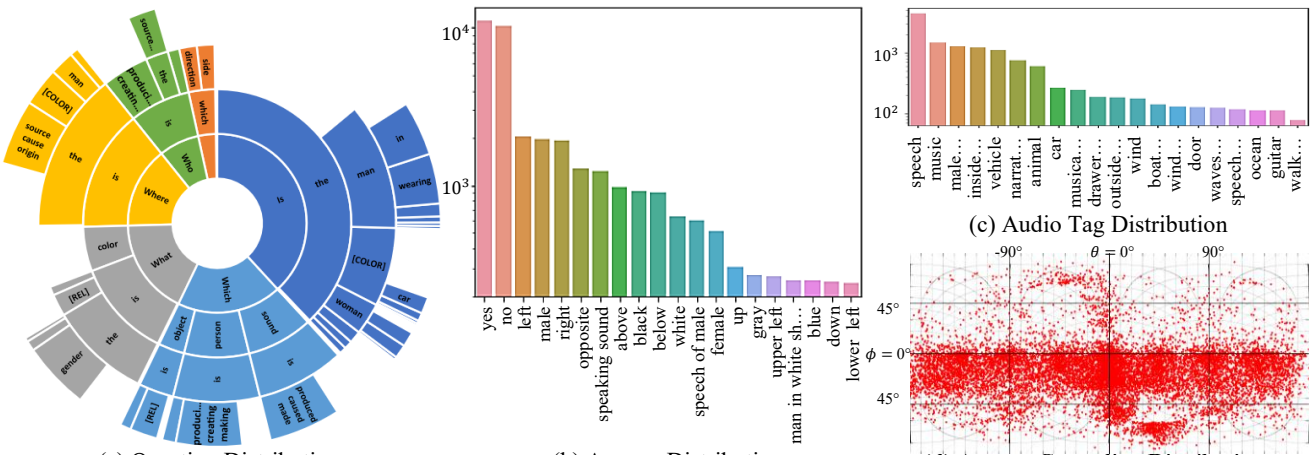
One limitation of a template-based generation is that the answer distribution may have a strong statistical bias with some words in the question template, leaving the question answerable without taking the context into account. For example, the abundance of *man/woman* annotated with *utterance-related sound descriptions* might bring in a misconception that all visible people in the scene are speaking. To alleviate this problem, we generate additional QA pairs by replacing original descriptions with unrelated audio and visual descriptions or throwing identical questions on *counterexample* clips like with non-speaking persons in this case.

**Postprocessing.** To ensure the grammatical correctness of collected QA pairs, we use *LanguageTool*[1] for proofreading. We also manually validate whether the question is answerable from the video, bounding boxes are correct, and sound description is included in the QA pair in any form for audio-visual QAs.

### 3.4. Data Analysis

Pano-AVQA consists of 51.7K QA pairs (42.8K training, 3.7K validation, 5.3K testing) from 5.4K clips extracted from 2.9K videos. There are in total 5.8K unique answers, with an average length of 3.7 words. The average question length is 12.1 words. Compared to other datasets on 360° videos in Table 1, Pano-AVQA contributes a large-scale and

[1] https://github.com/languagetool-org




(a) Question Distribution (b) Answer Distribution (c) Audio Tag Distribution (d) Answer Grounding Distribution

Figure 3. Illustrations of Pano-AVQA dataset statistics. (a) Distribution of first n-grams in questions. (b) Distribution of top-20 frequent answers. (c) Distribution of top-3 AudioSet [40] taggings. (d) Distribution of center points of bounding box groundings for answers.

diverse dataset on 360° videos along with additional annotations, *i.e.*, QA with groundings, relevant to video clips.

Among the QA pairs, 20K pairs belong to spherical spatial reasoning, and 31.7K pairs belong to audio-visual reasoning. We can easily notice the prevalence of questions with spatial relations (the words "next to, opposite of, left/right to, and above/below" are aggregated to `[REL]` token for visibility) or words relevant to audio-visual reasoning like *source, origin, causing* and *producing* from the sunburst diagram in Fig. 3(a).

Containing audio signals from diverse sources is crucial for audio-visual reasoning in real-life. Fig. 3(c) shows the distribution of top-3 Audioset [40] taggings obtained by running pretrained audio neural networks [45]. Although the human sound (*e.g.*, speech, narration, *etc.*) tag is the most frequent due to the prevalence of vlog in the video set, our dataset still contains a sizable number of other tags like vehicles, animal, and musical instruments. Moreover, human speech depends on factors like vocal tone, pace, and style *etc*. Our dataset reflects these different patterns by generating QA pairs from detailed descriptions of human speeches like *narration in loud tone, murmuring, etc.*

Along with QA pairs, our dataset contains 51.7K objects annotated with bounding boxes that are the most relevant to answering the question, *i.e.*, answer grounding. Fig. 3(d) illustrates the distribution of the center points of the bounding boxes. While the majority of the points are located near the equator (*i.e.*, $\phi = 0°$), considerable amounts of boxes are well spread away from the equator and even positioned near the poles. This distribution demonstrates that our dataset reflects various spherical spatial properties of 360° videos from a wide, holistic perspective.

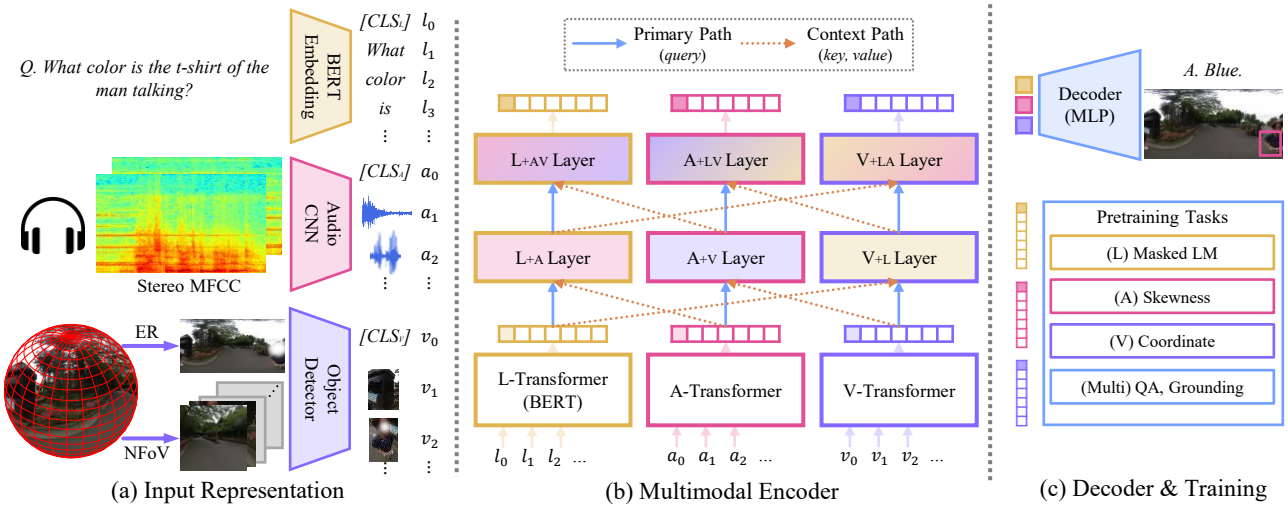
# 4. Approach

To address the new problems of audio-visual question answering on panoramic videos, we present a model named LAViT (**L**anguage **A**udio-**Vi**sual **T**ransformer), as illustrated in Fig. 4. It focuses on resolving two challenges of modeling (i) the feature representation of the video, audio, and language and (ii) the encoder-decoder structure that reconciles three different modalities. In summary, we tackle these issues by (i) extracting spherical spatial embedding from a set of visual objects and audio events, and (ii) utilize transformer-based architecture as a multimodal encoder, inspired by its recent success in VQA research [46, 47, 48, 49, 50, 51].

## 4.1. Input Representations

**Visual Representation.** We first uniformly sample the video at 1 fps (*i.e.*, about five panoramic frames) to reduce computational complexity while maintaining the temporal context in the video. As explained in Sec. 3.3, we use faster R-CNN [52] trained with ImageNet Detection [39] to extract and represent region proposals. We apply it to both equirectangular and NFoV projections, which are complementary since the former format shows key objects more continuous and larger, and the latter format displays objects with less distortion. We apply non-maximum suppression using spherical coordinates $(\theta, \phi, w_\theta, h_\phi)$ to filter out overlapping proposals from the two different projections with an IoU threshold $\tau = 0.65$. If there are too many objects detected, we only keep top-35 proposals with higher confidence. Finally, we obtain object embeddings $\{b_i\}_{i=1}^{N}$ per 360° video, where $N = 35$ is the number of proposals.

Next, we convert the Cartesian coordinates of the region proposals into the rotation quaternion based spatial repre-




(a) Input Representation (b) Multimodal Encoder (c) Decoder & Training

Figure 4. Overview of the proposed architecture named LAViT (**L**anguage **A**udio-**Vi**sual **T**ransformer).

sentation $\{c_i\}_{i=1}^{N}$ to reflect the spherical geometry:

$$c_i = (t, \cos\frac{\theta}{2}, -y\sin\frac{\theta}{2}, x\sin\frac{\theta}{2}, w, h), \tag{2}$$

where $t$ denotes the time step in seconds, $\theta$ is a rotation angle from the bottom of the sphere $(0, 0, -1)$ to the center of the object, the unit vector $(x, y, z)$ is the position of object center, and $(w, h)$ is the width and height. For the uniqueness of the axis of rotation, we only select the axis on a horizontal plane, *i.e.*, XY-plane, thereby omitting the z-axis from the rotation quaternion.

Finally, we obtain **visual representations** $\{v_i\}_{i=0}^{N}$, where $v_i = f_b(b_i) + f_c(c_i)$ for $i \geq 1$ using linear FC layers $f_b, f_c$. We obtain $v_0$ by average-pooling $\{v_i\}_{i=1}^{N}$, and use it as a special visual symbol $\texttt{[CLS}_v\texttt{]}$ similar to $\texttt{[CLS]}$ symbol in [53] or $\texttt{<IMG>}$ token in [47].

**Audio Representation.** We use stereo audio to reflect the spatial information of the surroundings [54]. As a feature extractor, we adopt a VGG-like CNN [45] trained with AudioSet [40]. We run the extractor to audio signals on the left and right channels separately. Since segmenting the audio into equal lengths may result in mixing different events, we need a reasonable way to recognize when the audio event changes. Motivated by CTC [55], we regard audio segments with the same top-k classes as a single event. Therefore, we split the audio stream into multiple segments whose top-k labels ($k = 3$) are identical. For each audio event, we max-pool the corresponding audio features, thereby obtaining left channel audio embeddings $\{a_i^0\}_{i=1}^{M}$ and right channel audio embeddings $\{a_i^1\}_{i=1}^{M}$, where $M$ is the number of events.

We finally obtain **audio representation** $\{a_i\}_{i=0}^{N}$, where $a_i = f_a^0(a_i^0) + f_a^1(a_i^1)$ for $i \geq 1$ using linear FC layers $f_a^0, f_a^1$. $a_0$ corresponds to a special audio symbol $\texttt{[CLS}_a\texttt{]}$, where we average pool the rest of the audio representations.

**Language Representation.** We use the WordPiece tokenizer [56] to split the questions into tokens and use pre-trained $\text{BERT}_{\text{base-uncased}}$ [53] to extract **language representations** $\{l_i\}_{i=0}^{K}$, where $l_0$ is a special language symbol $\texttt{[CLS}_l\texttt{]}$.

## 4.2. Encoder

The encoder of our model consists of three unimodal encoders and one multimodal encoder as shown in Fig. 4(b).

**Unimodal Encoder.** To each of the language, audio, and visual input representations $\{l_i\}_{i=0}^{K}, \{a_i\}_{i=0}^{M}, \{v_i\}_{i=0}^{N}$, we first apply layer normalization [57] and feed them into the corresponding unimodal encoder, for which we use the encoder module of Transformer [12]. We stack nine encoding layers for language and five layers for audio and visual modality, as in [46]. The number of layers can be adjusted in the context of computing resources or performance.

**Multimodal Encoder.** We utilize the encoding layers of Transformer for multimodal encoding as well, but with different attention input. To be specific, we use the primary modality as an attentional query (*i.e.*, primary path) and another modality as an attentional key-value (*i.e.*, context path) so that two different modalities can be fused in one encoding layer. We stack two encoding layers per modality to perform this with the other two modalities. For unimodal encoder output $\{l_i'\}_{i=0}^{K}, \{a_i'\}_{i=0}^{M}, \{v_i'\}_{i=0}^{N}$ and Transformer encoding layer $T(\text{primary, context})$, we obtain multimodal encoder output $\{\hat{l}_i\}_{i=0}^{K}, \{\hat{a}_i'\}_{i=0}^{M}, \{\hat{v}_i'\}_{i=0}^{N}$:

$$l_{ai}' = T(l_i', \{a_j'\}),\ a_{vi}' = T(a_i', \{v_j'\}),\ v_{li}' = T(v_i', \{l_j'\}),$$
$$\hat{l}_i = T(l_{ai}', \{a_{vi}'\}), \hat{a}_i = T(a_{vi}', \{v_{li}'\}),\ \hat{v}_i = T(v_{li}', \{l_{ai}'\}).$$

**Decoder.** We obtain the average-pooled representations $\hat{v}_0$, $\hat{a}_0$, $\hat{l}_0$ from multimodal encoder output

$\{\hat{l}_i\}_{i=0}^{K}, \{\hat{a}'_i\}_{i=0}^{M}, \{\hat{v}'_i\}_{i=0}^{N}$, which are used as the special symbols `[CLS`$_v$`]`, `[CLS`$_a$`]`, `[CLS`$_l$`]`, respectively. We finally concatenate all three aggregated representations $\hat{v}_0, \hat{a}_0, \hat{l}_0$ and feed them into two three-layered MLPs, one for predicting answer label, for which we take argmax onto the output and the other for answer grounding.

### 4.3. Training

Following the training practice of transformer-based architectures, we utilize pretraining and finetuning objectives to train the model. For pretraining, we randomly mask visual, audio, and language input representations $\{v_i\}_{i=1}^{N}, \{a_i\}_{i=1}^{M}, \{l_i\}_{i=1}^{K}$ with a probability of 0.15 and train the model with the following pretext tasks.

**Language Pretraining Task.** We use masked token prediction with cross-entropy loss as suggested in [53], by predicting the masked part of the language input.

**Visual Pretraining Tasks.** Instead of predicting the representation itself or its classification label, we add an MLP that predicts spherical spatial embedding from the masked visual representation with a smooth L1 loss.

**Audio Pretraining Tasks.** Designing a pretext for audio representation is less straightforward than visual ones. We thus propose *spatial skewness prediction* of the masked audio representations. Compared to phoneme classification or speaker classification generally adapted in audio transformers [58, 59], which may be limited in the utterance domain, our spatial skewness prediction can generally be applied to any media with multichannel audio and without any teacher model. We regard the stereo audio channel as a 3D audio with two silent channels and apply spherical harmonics decomposition to measure spatial skewness from given audio, *i.e.*, from which direction the audio is coming. That is, from the truncated spherical harmonics decomposition of an audio $s_t(\theta, \phi) = \sum_{n=0}^{N} \sum_{m=-n}^{n} c_n^m(t) \cdot Y_n^m(\theta, \phi)$, where $Y_n^m$ is the spherical harmonics, we extract the coefficient $c_n^m$, which reflects how much sound is originated from position $(\theta, \phi)$. We map the obtained skewness from $\mathbb{R}_{[-20,20]}$ to $\mathbb{R}_{[-1,1]}$ and train an MLP with a smooth L1 loss to predict the masked audio representation's skewness along with the timestamp (*i.e.*, start time and duration).

**QAs with Grounding.** We use the Question-Answer pairs with grounding as a multimodal task for both pretraining and fine-tuning. We formulate the question-answering task as a classification problem where the model selects the best answer candidate over the 2020-D answer table, which covers approximately 93% of the questions. Specifically, we provide aggregated representations from multimodal encoder $(\hat{v}_0, (\hat{a}_0, \hat{l}_0)$ as input to an MLP to predict the answer and coordinate grounding, respectively. We train answer prediction with a cross-entropy loss and coordinate grounding with a smooth L1 loss.

**Implementation Details.** Except for input feature extraction, we train our model end-to-end with a batch size of 32, gradient accumulation of 4, and dropout with a rate of 0.1. We optimize with AdamW [60] with an initial learning rate of 1e-4 for three epochs as pretraining and fine-tune the model for another seven epochs with a learning rate of 5e-5. In both stages, we aggregate all losses from the tasks with equal weights but the grounding task, which is set to 0.2 to balance its influence against the question-answer task. We use a linear scheduler with a warmup rate of 0.1.

## 5. Experiments

### 5.1. Experimental Setup

**Baselines.** To evaluate the proposed encoding strategies of different modalities, we compare with AVSD [28], BERT [53], SparseGraph [61] and LXMERT [46]. AVSD suggests a late fusion-based approach for audio-visual dialog, for which the pretrained BERT can be a better language backbone. SparseGraph and LXMERT are chosen as the representative models for image question answering. For a fair comparison, we use the same tokenizer and multimodal encoder (including audio) as in LAViT.

**Different Spherical Spatial Representations.** As claimed in [61], providing appropriate spatial embedding is paramount for good performance in visual question answering. To explore the effectiveness of using quaternion representation for spatial embedding in the spherical panorama, we experiment with a few other possible spatial representations: Cartesian coordinates $(x, y, w, h)$, spherical coordinates $(\theta, \phi, w_\theta, h_\theta)$, and normal 3D coordinates $(x, y, z, w_\theta, h_\theta)$.

**Evaluation Metrics.** We measure the accuracy on the Pano-AVQA test split as the percentage of correctly answered questions. As mentioned in Sec. 4.3, the VQA task is formulated as a classification problem; selecting the best word over the dictionary vocabularies. For the answer grounding task that predicts bounding box coordinates, we use the mean squared error.

### 5.2. Results and Analyses

**Comparison with VQA Models.** Effective multimodal fusion is one of the paramount issues to correctly address the questions in the Pano-AVQA dataset. In Table 2, the sharp performance drop of AVSD and BERT$_{+AV}$ compared to our model suggests that late fusion-based approaches are less adept at incorporating different modalities. Compared to SparseGraph [61] and LXMERT [46] that can effectively fuse visual and language modalities, our model performs 5.85% and 2% better, respectively.

Good performance of prior-based models may imply that the answer distribution is skewed toward a few popular answers. Accuracies of prior-based models in our dataset are 21.47 and 32.49, which are lower than those in VQA [18]

| Model | MSE | Accuracy (%) | | |
|---|---|---|---|---|
| | Ground | SS | AV | All |
| Prior ("yes") | - | 28.92 | 16.75 | 21.47 |
| Q-Type Prior | - | 36.30 | 32.42 | 32.49 |
| AVSD [28] | - | 29.40 | 20.10 | 24.60 |
| BERT$_{+AV}$ [53] | - | 36.88 | 38.43 | 37.83 |
| SparseGraph [61] | - | 42.89 | 45.74 | 44.64 |
| LXMERT [46] | - | 47.48 | 49.12 | 48.48 |
| LAViT$_{w/o\ \text{unimodal}}$ | - | 39.42 | 47.14 | 44.14 |
| LAViT$_{A+L}$ | - | 46.90 | 48.68 | 47.99 |
| LAViT$_{V+L}$ | 0.556 | 48.75 | 48.71 | 48.73 |
| LAViT$_{\text{Single-NFoV}}$ | - | 47.14 | 49.37 | 48.50 |
| LAViT$_{\text{ER-Only}}$ | 0.605 | 47.63 | 50.17 | 49.18 |
| LAViT$_{\text{Dense-NFoV}}$ | 0.593 | 47.68 | 51.13 | 49.79 |
| **LAViT (ours)** | 0.629 | **49.29** | **51.25** | **50.49** |

Table 2. Results on Pano-AVQA *test* split. SS denotes spherical spatial reasoning task and AV denotes audio-visual reasoning task.

| | Embeddings | MSE | Accuracy (%) | | |
|---|---|---|---|---|---|
| | | Ground | SS | AV | All |
| V | Cartesian | *0.166 | 47.48 | 51.41 | 49.89 |
| | Spherical | *3.496 | 48.95 | 51.01 | 50.21 |
| | Unit sphere | *1.378 | 49.49 | 50.05 | 49.83 |
| | **Quaternion** | *0.629 | 49.29 | 51.25 | 50.49 |

Table 3. Experimental results of different spherical spatial embeddings. *The grounding errors of different representations are not comparable as they have different error scales.

(*i.e.*, 29.66 and 37.54, respectively).

**Ablation.** Our model without the unimodal encoders (LAViT$_{w/o\ \text{unimodal}}$) attains 6.35% performance drop, which indicates the importance of loading pretrained language model as well as maintaining the context of unimodal input. Opting out either visual or audio input decreases performance by 2.5% and 1.76%, implying the importance of utilizing both modalities.

**Influence of FoV Selection.** The model trained with single NFoV in videos, which corresponds to a video captured with a conventional camera, is 2% lower than our model, denoting the importance of a wider field of view. Meanwhile, the performance of the ER-only model is lower than the model trained with dense NFoV, which is presumably due to overlooking smaller objects. Still, utilizing both ER and NFoVs as in Fig. 4(a) shows the best performance.

**Spherical Spatial Representations.** Table 3 shows that the unit sphere and quaternion-based spatial embeddings perform better in the spherical spatial reasoning task, while the Cartesian coordinates works the worst. Although the Cartesian-based model has the lowest grounding error, it is mainly due to the error scale of Cartesian coordinates. Thus, the ground errors beween the spatial embeddings are not directly comparable. Fig. 5 displays different answer grounding proposals per geometry. In general, embeddings with spherical spatial information performs better than Cartesian-based proposals. Still, our quaternion-based approach displays notable localization ability compared to other proposals, especially in the examples from the second column. Please refer to the Appendix for more inference examples and visualization.

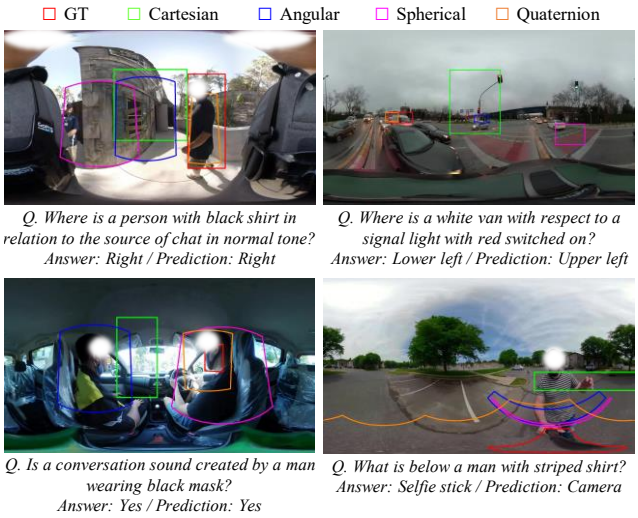



Figure 5. Qualitative examples of answer grounding from Table 3.

## 6. Conclusion

Our work extended existing works on panoramic video understanding by proposing video question answering as a novel task to evaluate spherical spatial and audio-visual reasoning capacity of models in 360° surrounding. To evaluate this, we introduced a large-scale Pano-AVQA dataset consisting of 51.7K QA pairs with bounding boxes from 5.4K panoramic videos. Also, we designed LAViT as a new audio-visual QA transformer framework that extends cross-modal attention to leverage three modalities.

Moving forward, for better reasoning in 360° videos, it can incorporate audio-visual scene graphs as an additional annotation. Another promising direction to use our 360° datasets to address Embodied Question Answering (EQA) [62, 63, 64, 65] and language-guided embodied navigation [66, 67] in a simulated 3D interactive environment.

**Acknowledgement**. We thank the anonymous reviewers for their thoughtful suggestions on this work. This work was supported by AIRS Company in Hyundai Motor Company & Kia Corporation through HKMC-SNU AI Consortium Fund, Brain Research Program by National Research Foundation of Korea (NRF) (2017M3C7A1047860) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01082, SW StarLab). Gunhee Kim is the corresponding author.

## References


[1] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. WoodScape: A Multi-task, Multi-camera Fisheye Dataset for Autonomous Driving. In *ICCV*, 2019. 1

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *CVPR*, 2020. 1

[3] Yasamin Heshmat, Brennan Jones, Xiaoxuan Xiong, Carman Neustaedter, Anthony Tang, Bernhard E Riecke, and Lillian Yang. Geocaching with a Beam: Shared Outdoor Activities through a Telepresence Robot with 360 Degree Viewing. In *CHI*, 2018. 1

[4] Yoshiki Masuyama, Yoshiaki Bando, Kohei Yatabe, Yoko Sasaki, Masaki Onishi, and Yasuhiro Oikawa. Self-Supervised Neural Audio-Visual Sound Source Localization via Probabilistic Spatial Modeling. In *IROS*, 2020. 1, 2

[5] Wen-Chih Lo, Ching-Ling Fan, Jean Lee, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 360 Video Viewing Dataset in Head-mounted Virtual Reality. In *ACM MMSys*, 2017. 1

[6] Maximilian Speicher, Jingchen Cao, Ao Yu, Haihua Zhang, and Michael Nebeling. 360Anywhere: Mobile Ad-hoc Collaboration in any Environment using 360 Video and Augmented Reality. *ACM HCI*, 2018. 1

[7] David V Smith, Ben Davis, Kathy Niu, Eric W Healy, Leonardo Bonilha, Julius Fridriksson, Paul S Morgan, and Chris Rorden. Spatial Attention Evokes Similar Activation Patterns for Visual and Auditory Stimuli. *Journal of Cognitive Neuroscience*, 2010. 1

[8] Yu-Chuan Su, Dinesh Jayaraman, and Kristen Grauman. Pano2Vid: Automatic Cinematography for Watching 360° Videos. In *ACCV*, 2016. 2, 4

[9] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency Detection in 360° Videos. In *ECCV*, 2018. 2

[10] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. Cube Padding for Weakly-Supervised Saliency Prediction in 360 Videos. In *CVPR*, 2018. 2

[11] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-Supervised Generation of Spatial Audio for 360 Video. In *NIPS*, 2018. 2, 4

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *NeurIPS*, 2017. 2, 6

[13] Youngjae Yu, Sangho Lee, Joonil Na, Jaeyun Kang, and Gunhee Kim. A Deep Ranking Model for Spatio-Temporal Highlight Detection from a 360° Video. In *AAAI*, 2018. 2, 4

[14] Sangho Lee, Jinyoung Sung, Youngjae Yu, and Gunhee Kim. A Memory Network Approach for Story-Based Temporal Summarization of 360 Videos. In *CVPR*, 2018. 2, 4

[15] Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun. Deep 360 Pilot: Learning a Deep Agent for Piloting through 360° Sports Video. In *CVPR*, 2017. 2, 4

[16] Shih-Han Chou, Yi-Chun Chen, Kuo-Hao Zeng, Hou-Ning Hu, Jianlong Fu, and Min Sun. Self-View Grounding Given a Narrated 360 Video. In *AAAI*, 2018. 2, 4

[17] Ryo Shimamura, Qi Feng, Yuki Koyama, Takayuki Nakatsuka, Satoru Fukayama, Masahiro Hamasaki, Masataka Goto, and Shigeo Morishima. Audio–Visual Object Removal in 360-Degree Videos. *The Visual Computer*, 2020. 2

[18] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 2, 7

[19] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *CVPR*, 2016. 2

[20] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *CVPR*, 2017. 2, 4

[21] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. Video Question Answering via Attribute-Augmented Attention Network Learning. In *SIGIR*, 2017. 2

[22] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. DeepStory: Video Story QA by Deep Embedded Memory Networks. In *IJCAI*, 2019. 2

[23] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*, 2018. 2

[24] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. In *AAAI*, 2019. 2

[25] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. KnowIT VQA: Answering Knowledge-Based Questions about Videos. In *AAAI*, 2020. 2

[26] Haytham M Fayek and Justin Johnson. Temporal Reasoning via Audio Question Answering. *arXiv:1911.09655*, 2019. 2

[27] Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. ODSQA: Open-Domain Spoken Question Answering Dataset. In *IEEE SLT*, 2018. 2

[28] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio Visual Scene-Aware Dialog. In *CVPR*, 2019. 2, 7, 8

[29] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal Deep Learning. In *ICML*, 2011. 2

[30] Nitish Srivastava and Russ R Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. In *NeurIPS*, 2012. 2

[31] Andrew Owens and Alexei A Efros. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In *ECCV*, 2018. 2

[32] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 2

[33] Ruohan Gao and Kristen Grauman. 2.5D Visual Sound. In *CVPR*, 2019. 2

[34] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The Sound of Pixels. In *ECCV*, 2018. 2

[35] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic Object Prediction and Spatial Sound Super-Resolution with Binaural Sounds. In *ECCV*, 2020. 2, 4

[36] Di Hu, Feiping Nie, and Xuelong Li. Deep Multimodal Clustering for Unsupervised Audiovisual Learning. In *CVPR*, 2019. 2

[37] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The Sound of Motions. In *ICCV*, 2019. 2

[38] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3, 4

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 4, 5

[40] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *ICASSP*, 2017. 4, 5, 6

[41] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an Audio Captioning Dataset. In *ICASSP*, 2020. 4

[42] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in the Wild. In *NAACL*, 2019. 4

[43] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging Video Descriptions to Learn Video Question Answering. In *ECCV*, 2016. 4

[44] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *ACM MM*, 2017. 4

[45] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM TASLP*, 2020. 5, 6

[46] Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*, 2019. 5, 6, 7, 8

[47] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, 2019. 5, 6

[48] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv:1908.03557*, 2019. 5

[49] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*, 2020. 5

[50] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*, 2020. 5

[51] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*, 2020. 5

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, 2015. 5

[53] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019. 6, 7, 8

[54] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-Supervised Moving Vehicle Tracking with Stereo Sound. In *ICCV*, 2019. 6

[55] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *ICML*, 2006. 6

[56] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144*, 2016. 6

[57] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. In *NIPS Deep Learning Symposium*, 2016. 6

[58] Yung-Sung Chuang, Chi-Liang Liu, Hung-yi Lee, and Lin-shan Lee. SpeechBERT: An Audio-and-text Jointly Learned Language Model for End-to-end Spoken Question Answering. In *INTERSPEECH*, 2020. 7

[59] Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Shang-Wen Li, and Hung-yi Lee. Audio ALBERT: A Lite BERT for Self-supervised Learning of Audio Representation. In *IEEE SLT*, 2021. 7

[60] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2018. 7

[61] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning Conditioned Graph Structures for Interpretable Visual Question Answering. In *NeurIPS*, 2018. 7, 8

[62] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *CVPR*, 2018. 8

[63] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *CVPR*, 2018. 8

[64] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019. 8

[65] Piotr Mirowski, Matthew Koichi Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. Learning to navigate in cities without a map. In *NeurIPS*, 2018. 8

[66] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *CVPR*, 2019. 8

[67] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv:1801.02209*, 2018. 8