

A Technical Appendices and Supplementary Material

A.1 Experimental Details

A.1.1 Baseline Experiments

We evaluated several baseline models to assess spatial reasoning capabilities of audio-visual methods. These include ISSL [39], ACL-SSL [31], and VideoLLaMA2 [9]. The models were reproduced using publicly available codebases or adapted from official checkpoints. All models were tested on our proposed Hear You Are QA dataset.

ISSL. This model is a ResNet-based sound source localization method that originally uses 224×224 input images, unlike ours, which uses 880×224 panoramic inputs. Unlike transformer-based models that rely on positional embeddings, ISSL does not require them, allowing it to operate directly on equirectangular panoramic images without spatial tokenization or interpolation. We used the raw 360° panoramic input as-is, without any resizing or slicing. Following the original article, we sort the heatmap values of each image and retain the top T pixels. In this experiment, we use the top 0.5% as the threshold. Afterward, we sum the values along the vertical axis and divide them into angle bins corresponding to 30° . The bin with the largest sum is selected as the localization answer. For sound classification, we perform audio retrieval by retrieving the most similar audio feature from the test set for each test audio. If the retrieved audio belongs to the same category, it is considered a correct answer.

ACL-SSL. The ACL-SSL model is also trained on 224×224 input images, but unlike ISSL, it is based on a transformer architecture and relies on positional embeddings. To apply the model to 360° panoramic inputs (880×224), we first sliced each equirectangular image into four vertical segments of 220×224 , and then resized each slice to 224×224 to match the model’s expected input format. We ran the model on each slice independently and then concatenated the resulting heatmaps to construct a panoramic heatmap. This step is not part of the original design, but we adopt it to enable panoramic localization. Since ACL-SSL focuses on *semantic alignment* rather than spatial reasoning, this slicing and recombination process introduces minimal distortion, and spatial continuity is not critical for performance. To obtain the final localization answer, we apply a fixed threshold of 0.5 to the heatmap and consider only pixels with values above the threshold. We then sum the values along the vertical axis, divide them into 30° angle bins, and select the bin with the highest sum. For sound classification, we perform audio retrieval in the same manner as ISSL. The most similar audio feature from the test set is retrieved, and if it belongs to the same category, it is considered correct.

VideoLLaMA2. We adapted the VideoLLaMA2 framework to our multimodal setting by using the same model architecture and encoders as our full method. The only difference lies in the audio input, as this baseline receives *monaural* audio instead of binaural signals. We trained the model with the R+M+Q configuration, which uses panoramic RGB, monaural audio, and text question input. This corresponds to the ablation setting in Table 3 and serves as a strong LLM-based baseline for multi-modal reasoning without spatial modeling.

Qualitative Comparison with Baselines

Figure 4 and Figure 5 present qualitative comparisons between the ACL-SSL baseline and our proposed model. These visualizations illustrate the grounding performance of each method on representative Q1 (non-matching) and Q8-type questions, which require both semantic recognition and spatial localization of sounding objects.

As shown in Figure 4, ACL-SSL generates heatmaps that highlight regions semantically aligned with the audio but lacks the spatial precision to distinguish between multiple matching candidates. In the first column, the ACL-SSL model merely segments both chickens and electric blenders. In contrast, our model can identify which specific object is making the sound by leveraging spatial understanding. In the second column, the cell phone ringing sound originates from the *pitcher* and the *hamper*. Since these objects are not semantically related to the sound, the ACL-SSL model fails to localize the correct region. However, as shown in Figure 5, our model recognizes the spatial



Figure 4: Qualitative results from the ACL-SSL baseline. The model highlights semantically matching regions but fails to distinguish the actual sound source due to the lack of spatial reasoning.

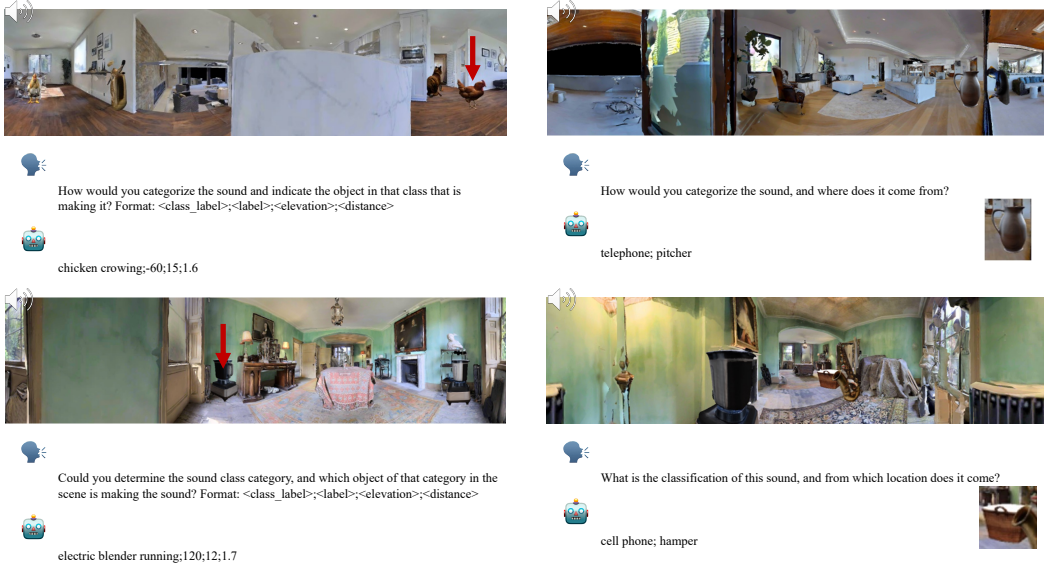


Figure 5: Qualitative results from our model. By leveraging spatial audio cues, the model accurately localizes the sound source and identifies the correct visual object at that location.

829 audio cues and localizes the sound source, enabling it to infer what visual object is present at that
830 location and produce the correct answer.

831 These results underscore the importance of spatial reasoning in audio-visual understanding. While
832 semantic-only models like ACL-SSL may succeed in object detection, they fall short in tasks requiring
833 disambiguation. By explicitly modeling the spatial alignment between binaural audio and panoramic
834 vision, our model can resolve such ambiguities and make accurate spatial predictions.

835 A.1.2 Encoder Warm Start QA Generation

836 To pre-train each encoder on spatially grounded audio and visual representations respectively, we
837 constructed a simple uni-modal QA dataset derived from simulation metadata. Each data sample
838 contains a 360° image, spatial audio, and object positions with annotations.

839 We synthesized QA pairs in two modalities:

- 840 • **Audio-based QA:** Given a binaural waveform, questions ask for either the *class label* or the
841 *spatial position* (azimuth, elevation, and distance) of the sound source.
- 842 • **Visual-based QA:** Given a binaural waveform, questions ask for either the *class label* or the
843 *spatial position* (azimuth, elevation, and distance) of the visual object.

844 Answer Format Examples:

- 845 • <wav> What are the predicted azimuth and elevation angles,
846 and the distance to the sound source?
847 Answer: (90, -10), 2.3 meters
- 848 • <wav> What sound did you detect?
849 Answer: typewriter
- 850 • <rgb> What are the predicted azimuth and elevation angles,
851 and the distance to the typewriter?
852 Answer: (90, -10), 2.3 meters
- 853 • <rgb> What visual objects did you detect at (60, 0), 1.7
854 meters?
855 Answer: typewriter

856 As we use Spatial-AST [52] as the audio encoder, which is already pre-trained to capture spatial
857 cues, we only need to train the audio projector to align with the LLM backbone. This makes the
858 adaptation process relatively simple. In contrast, the image encoder [47] is not initially designed
859 for 360° panoramic inputs and suffers from geometric distortion. To address this, we first LoRA
860 fine-tune the image encoder to recognize object class labels under panoramic distortion. Once it
861 learns to handle such geometric transformations, we further train it to answer questions requiring
862 spatial position prediction, such as azimuth, elevation, and distance.

863 A.1.3 Reproducibility.

864 We will release the full codebase, panoramic image dataset, reverb files, model checkpoints, and de-
865 tailed instructions for reproducing all experiments upon acceptance. Please refer to the VGGSound [8]
866 for the audio files used in this study.

867 A.2 Dataset Details

868 A.2.1 Explanation on the Visual Scene

869 The Hear You Are QA dataset contains 360° panoramic images of realistic indoor environments.
870 These scenes are populated with both sound-emitting and silent visual objects, distributed across
871 diverse azimuth angles. The height of each object is randomly sampled within 0.5 meters from the
872 floor to provide visually plausible augmentation without introducing unrealistic placements. Although
873 elevation is included in both the training and evaluation stages, it is largely negligible in practice and
874 is therefore excluded from performance metrics, except for Q3-type questions where elevation is
875 explicitly required.

876 A.2.2 Generated 3D Objects

877 We use Stable Diffusion 3 [35] and InstantMesh [49] to synthesize new 3D audio-visual objects,
878 enabling the diversification of spatial grounding scenarios. The size of each object category is
879 manually determined based on the common sense judgments of three annotators. We classify objects
880 into four size levels: smallest, small, medium, and large. For each size level, we define a representative
881 base size and apply a random variation of $\pm 20\%$ to introduce natural variation.

882 A.2.3 Explanation on Azimuth and Elevation

883 Figure 6 consists of two visualizations. The top image is a 2D equirectangular projection of a 360°
884 indoor scene. The bottom image shows a circular representation of the same scene, in which the

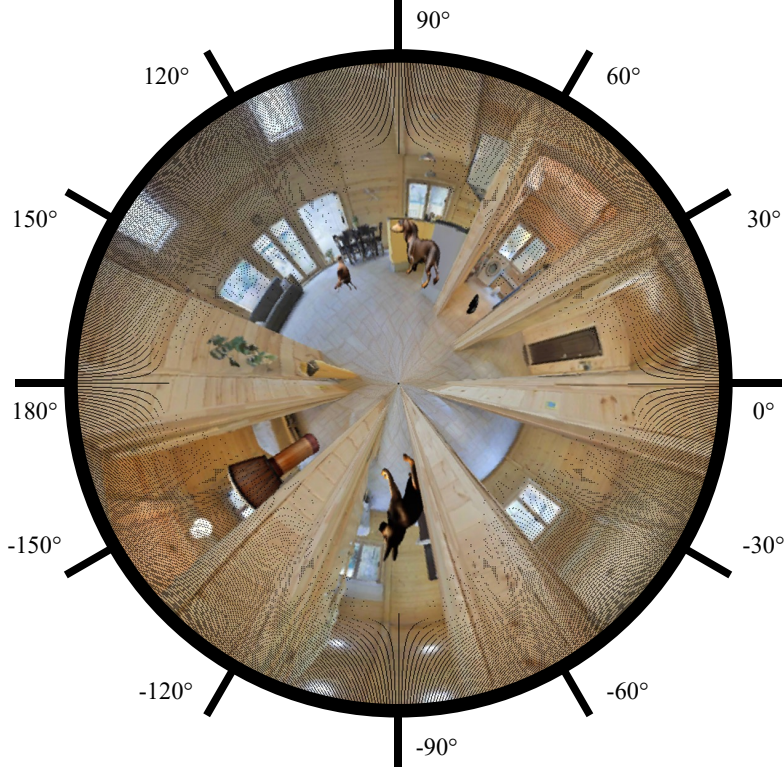


Figure 6: Equirectangular (top) and circular views of a 360° scene with azimuth annotations (bottom).

panoramic view is reprojected into a top-down format. Azimuth angles are annotated around the circle, ranging from -180° to 180° , with 90° indicating the agent's front-facing direction. This visualization helps provide an intuitive understanding of how spatial directions are represented in the panoramic setting.

Figure 7 illustrates how azimuth and elevation angles are defined on a spherical coordinate system. The *azimuth* (θ) represents the horizontal angle around the vertical axis, and the *elevation* (ϕ) indicates the vertical angle above or below the horizontal plane. In our setup, the agent is facing $\theta = 90^\circ$, which serves as the reference front-facing direction. The full range of these angles is defined as:

$$\theta \in [-180^\circ, 180^\circ], \quad \phi \in [-90^\circ, 90^\circ]$$

This spherical representation is used to define the 3D positions of sound sources and visual objects relative to the agent. It allows for a consistent spatial grounding of audio-visual inputs across different environments.

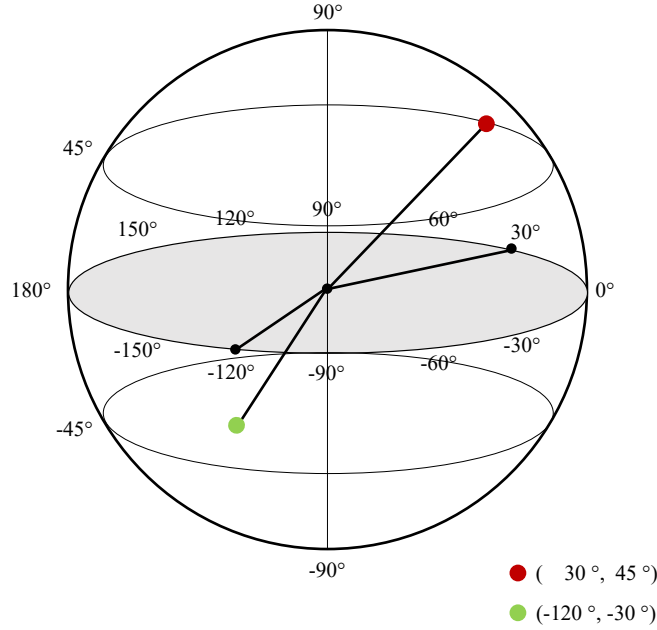


Figure 7: Spherical visualization of azimuth (θ) and elevation (ϕ) angles in a 360° panoramic setting. The agent is positioned at the center, facing 90°, with azimuth ranging from -180° to 180° and elevation from -90° to 90° .

896 A.2.4 Question Types

897 To help readers understand the design and purpose of each question type in our dataset, we provide ex-
 898 planations along with qualitative examples. Each example highlights a representative 360° panoramic
 899 scene, the associated question, and the correct answer.



Figure 8: Q1 example: Identify the sound class and locate its matching visual object.

900 **Q1: Spatial Correspondence** The scene includes a backpack and a dog, along with a cell
 901 phone sound that has no corresponding visual object. The question is: “What is the sound class
 902 category? Where is the sound coming from?” The correct answer is cell phone; backpack.
 903 Although the phone itself is not visible, the sound is localized at the position of the backpack. The
 904 model is expected to recognize the audio as a cell phone ringtone and associate it with the
 905 backpack, which occupies the same location.

906 **Q2–Q4 and Invisible Audio Settings** To provide a clearer understanding of the invisible audio
 907 settings in Q2–Q4, we present both a panoramic view (Figure 9) and a bird’s-eye view (Figure 10).
 908 For the bird’s-eye view, we include two settings: one with a visible audio-emitting object on the



Figure 9: Panoramic view used for illustrating Q2–Q4 scenarios.

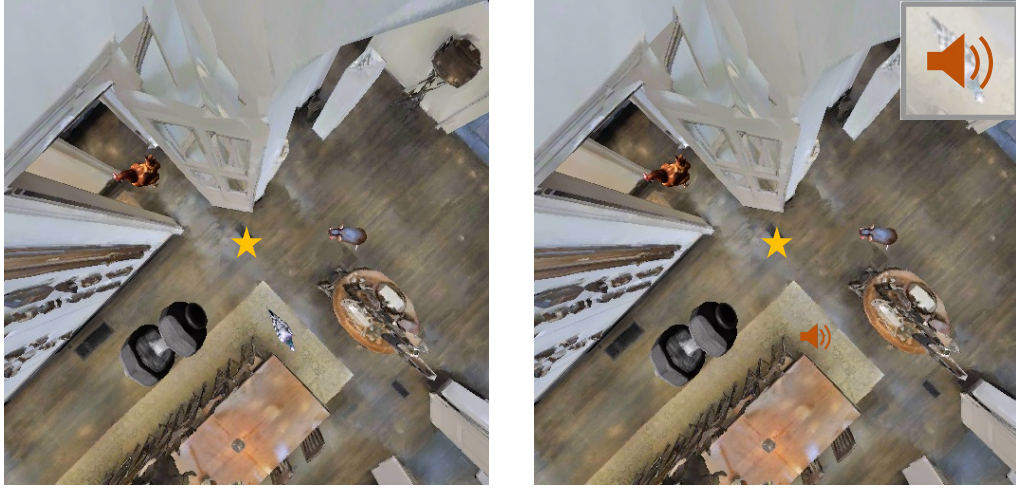


Figure 10: Bird's-eye view of audio source settings. The left side shows a visible sound source; the right side shows an invisible one.

909 left, and another with an invisible one on the right. In the latter case, the sound is still assigned to a
 910 specific location, even though no corresponding visual object is present. This invisible audio setting
 911 corresponds to the condition analyzed in the ablation study shown in Table 3



Figure 11: Left: object-to-agent distances (Q2). Right: angular and spatial relationship between two objects (Q4).

912 **Q2, Q4: Relative Location** Figure 11 illustrates the spatial setups involved in Q2 and Q4. The left
 913 part shows the relative distance between each object and the agent (yellow star), corresponding to
 914 Q2. Two sets of concentric circles are drawn: blue for the pigeon and green for the dumbbell.
 915 Since the blue circles are smaller, the pigeon is closer to the agent. The right part corresponds to
 916 Q4 and visualizes the spatial and angular relationship between the two objects. The blue and green

lines indicate the directions from the agent to the dumbbell and the pigeon, respectively, and the angle between them represents their azimuthal separation. The red line connects the two objects and indicates their Euclidean distance.

Q3: Relative Location Figure 12 illustrates the object coordinates and the question format used in Q3. The left part shows a bird’s-eye view with an XZ coordinate system, where the agent is placed at the origin (yellow star). Each object is plotted with its relative position, and larger x- or z-values indicate positions farther to the left or behind, respectively. The right part presents examples of Q3-style questions, where the model is asked to estimate the relative location of one object from another. Labels such as “Left, Behind” or “Right, Front” are derived from their spatial relationship on the coordinate grid.

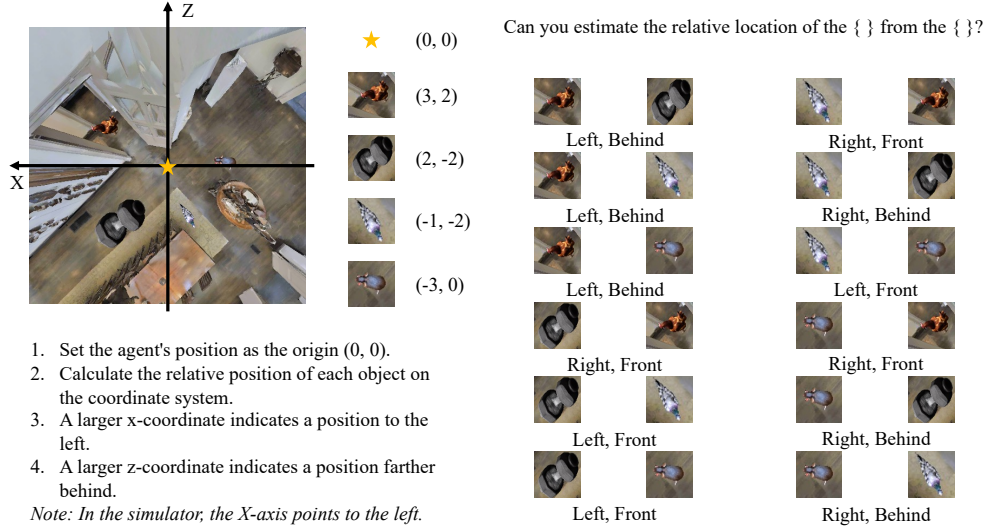


Figure 12: Bird’s-eye view of object locations (left) and Q3-style relative location question examples (right).



Figure 13: Q5 example: Estimate sound position and identify the emitting object class.

Q5: Spatial & Semantic Correspondence (One visual object semantically matches the audio)
The scene includes a dog, a double bass, a cup, and a mandolin. The sound is that of a mandolin, and it is spatially localized at $(-69, -17)$, 2.2 meters. The question is: “What is the object in the scene located at $(-69, -17)$, 2.2 meters? Is it making a sound?” The correct answer is mandolin; Yes. To answer correctly, the model must identify the object located at the specified coordinates and determine whether the sound is coming from that location.

Q6: Spatial & Semantic Correspondence (Multiple visual objects semantically match the audio)
The scene includes two alarm clocks, a harp, and an electric guitar. The sound is coming from the direction of one of the alarm clocks, around -100 in azimuth. The question is: “What



Figure 14: Q6 example: Determine if a visible object is emitting sound.

936 *is the object in the scene located at (43, -24), 1.3 meters? Is it making a sound?*", focuses on the
 937 other alarm clock, located at (43, -24), 1.3 meters, and asks whether it is emitting sound. The correct
 938 answer is `alarm clock`; No. To answer correctly, the model must determine whether the sound
 939 is coming from the location specified in the question.

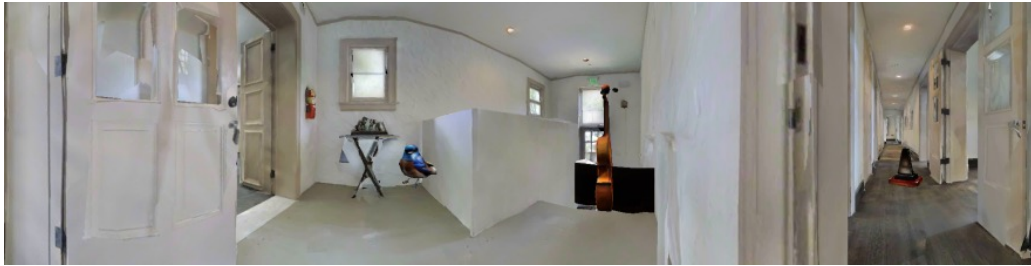


Figure 15: Q7 example: Select the correct sound-emitting object among candidates.

940 **Q7: Spatial & Semantic Correspondence (One visual object semantically matches the audio)**
 941 The scene includes a barn swallow calling, a metronome, and a double bass. The
 942 sound is that of the barn swallow calling. The question is: *"Given multiple visual objects,*
 943 *which one is making a sound, and where is it located?"* The correct answer is `barn swallow`
 944 `calling`; 123; -11; 2.2. The model must classify the sound, match it to the correct visual
 945 object among similar distractors, and provide its spatial location in azimuth, elevation, and distance.



Figure 16: Q8 example: Find the spatial position of a known audio category.

946 **Q8: Spatial & Semantic Correspondence (Multiple visual objects semantically match the audio)**
 947 The scene includes two dogs, a djembe, a bird, and a set of castanets. The sound is that
 948 of dog barking, and it is coming from the dog positioned at (-86, -9), 1.9 meters. The question
 949 is: *"Could you determine the sound class category, and which object of that category in the scene is*
 950 *making the sound?"* The correct answer is `dog barking`; -86; -9; 1.9. The model must
 951 classify the sound, match it to the correct visual object among similar candidates, and predict its
 952 spatial location in azimuth, elevation, and distance.



Figure 17: Q9 example: Identify the sound class and confirm source visibility.

Q9: Semantic Co-occurrence The scene includes an acoustic guitar, a baby, a bassoon, and a banjo. The sound is that of an acoustic guitar. The question is: “What is the sound class category? Is the sound source visible in the scene?” The correct answer is acoustic guitar; Yes. To answer correctly, the model must classify the sound and verify whether a visual object of the same class appears at the location where the sound is coming from.

A.2.5 Paraphrase Prompt Templates

To increase linguistic diversity and reduce model overfitting to rigid question structures, we employed GPT-4o to generate paraphrased templates for each of the 9 question types (Q1–Q9). Approximately 20 paraphrases were generated per type, and we present three representative prompts per type below. The full set will be released with the dataset and codebase.

- **Q1: Spatial Correspondence**
 - What is the sound class category? Where is the sound coming from?
 - Can you identify the sound category and its source?
 - What kind of sound is it, and what is its source location?
- **Q2: Relative Location (closer)**
 - Is the sound source of the {A} closer to the agent than it is to the {B}?
 - Does the sound of the {A} come from a closer position to the agent than the visual object {B}?
 - Is the {A}’s sound coming from a point nearer to the agent than the visual object {B}?
- **Q2-far: Relative Location (farther)**
 - Is the sound source of the {A} farther to the agent than it is to the {B}?
 - Is the agent farther from the sound of the {A} than to that of the {B}?
 - Is the acoustic origin of the {A} more distant from the agent than the visual object {B}?
- **Q3: Relative Location**
 - What is the distance between the {A} sound and the visual object {B}, and how is {A} positioned relative to {B}?
 - Can you assess the distance between the {A} sound and the visual object {B}, and determine the relative position of {A} with respect to {B}?
 - How would you describe the relative placement of {A} to {B} based on the sound?
- **Q4: Relative Location**
 - What is the distance between the {A} sound and the visual object {B}, and what is the angle formed by the agent’s gaze toward both?
 - Can you estimate the distance from the {A} sound to the {B}, and the angle between the agent’s gaze direction toward the {A} and the {B}?
 - How would you assess the angle between the agent’s gaze toward {A} and {B}, and their relative distance?
- **Q5: Spatial & Semantic Correspondence (One visual object semantically matches the audio)**

- 992 – What is the object in the scene located at {azimuth}, {distance} meters? Is it making a
- 993 sound?
- 994 – Which object is found at {azimuth}, {distance} meters, and is it currently making a
- 995 sound?
- 996 – Can you identify the object positioned at {azimuth}, {distance} meters, and is it
- 997 emitting any sound?
- 998 • **Q6: Spatial & Semantic Correspondence (Multiple visual objects semantically match**
- 999 **the audio)**
- 1000 – What is the object located at {azimuth}, {distance} meters in the scene, and is it
- 1001 producing a sound?
- 1002 – Can you determine the object located at {azimuth}, {distance} meters and confirm if it
- 1003 is producing sound?
- 1004 – What is the object at {azimuth}, {distance} meters, and is it the source of the audible
- 1005 signal?
- 1006 • **Q7: Spatial & Semantic Correspondence (One visual object semantically matches the**
- 1007 **audio)**
- 1008 – Given multiple visual objects, which one is making a sound, and where is it located?
- 1009 – Which object among the visual objects is producing a sound, and where is it placed?
- 1010 – From the visual objects in the scene, which one is producing a sound, and where is it
- 1011 positioned?
- 1012 • **Q8: Spatial & Semantic Correspondence (Multiple visual objects semantically match**
- 1013 **the audio)**
- 1014 – What is the sound class, and which object of that type in the scene is the source of the
- 1015 sound?
- 1016 – Can you determine the category of the sound and identify the object within that category
- 1017 that is generating it?
- 1018 – Which object of the sound class is producing the audio in the scene?
- 1019 • **Q9: Semantic Co-occurrence**
- 1020 – What is the sound class category? Is the sound source visible in the scene?
- 1021 – Can you specify the sound type and indicate whether its source can be seen in the
- 1022 scene?
- 1023 – What is the classification of the sound, and is the sound-emitting object in view?

1024 A.2.6 Prompt Formatting Strategy

1025 To ensure stable and structured responses from the language model, we applied prompt formatting
 1026 with question-specific instructions and examples. The prompt suffix varied depending on the question
 1027 type (q1–q9), and typically contained the following elements:

- 1028 • **A directive phrase**, e.g., Please provide the answer in the following
- 1029 format: ...
- 1030 • **A format schema**, e.g., <label>; <distance>
- 1031 • **A concrete example**, e.g., (e.g., J; 3.2)
- 1032 The examples were dynamically generated per instance. For example:
- 1033 – Azimuth labels (A–L) were randomly selected for each question from a uniform pool.
- 1034 – Distances were sampled from [0.5, 4.0] meters.
- 1035 – Elevations from [−20, 20] degrees.
- 1036 – Class names (e.g., blender, accordion) from the dataset’s audio/visual categories.
- 1037 • **A label explanation block** (for azimuth questions), e.g., A: 180°, B: −150°, ...

1038 The mapping between question type and appended instruction is as follows:

- 1039 • **Q1**

- 1040 - Suffix:
 1041 Please provide the answer in the following format:
 1042 class_category; sound_source (e.g., xylophone;eagle)
 1043 - Purpose: to ensure joint prediction of the sound category and its visual counterpart.
- 1044 • **Q2**
- 1045 - Suffix:
 1046 Please provide the answer Yes or No
 1047 - Purpose: to elicit binary (Yes/No) responses based on relative spatial reasoning.
- 1048 • **Q3**
- 1049 - Suffix:
 1050 Please provide the answer in the following format:
 1051 <left_right>;<up_down>;<front_behind>;<distance> (e.g.,
 1052 left;up;behind;2.3)
 1053 - Purpose: to guide spatial reasoning using direction and distance.
- 1054 • **Q4**
- 1055 - Suffix:
 1056 Please answer using labels A-L, where: A: 180°, B:
 1057 -150°, ..., L: 150°. Format: <label>;<distance> (e.g.,
 1058 J;3.2)
 1059 - Purpose: to map coarse directions to discrete azimuth bins with distance.
- 1060 • **Q5, Q6, and Q9**
- 1061 - Suffix: Please provide the answer in the following format:
 1062 <class_label>;<yes_no> (e.g., vacuum;Yes)
 1063 - Purpose: to encourage semantic grounding with binary decision making.
- 1064 • **Q7 and Q8**
- 1065 - Suffix:
 1066 Please answer using labels A-L, where: A: 180°, ..., L:
 1067 150°. Format: <class_label>;<label>;<elevation>;<distance>
 1068 (e.g., accordion;H;5.0;1.8)
 1069 - Purpose: to map coarse directions to discrete azimuth bins with distance, and to jointly
 1070 classify the sound type and localize the object.

1071 **Why Discrete Labels (A–L)?** We discretized the azimuth direction into 12 evenly spaced bins
 1072 (A–L), each representing a 30° increment in the clockwise direction starting from A (180°), with J
 1073 corresponding to the front (90°). This approach offers several advantages. Notably, it helps avoid
 1074 biased numeric outputs during training. When the model was prompted to directly generate azimuth
 1075 values as raw numbers, we found that it frequently produced certain patterns such as 123, likely
 1076 influenced by pretraining exposure to common number sequences. These patterns disrupted training
 1077 stability, making discrete labels a more robust and interpretable alternative. In addition, it facilitates
 1078 simple evaluation in direction-of-arrival and localization tasks. The specific azimuth bin definitions
 1079 are shown in Table 4.

Table 4: Azimuth label mapping used in directional prompts.

Label	A	B	C	D	E	F	G	H	I	J	K	L
Degree	180°	-150°	-120°	-90°	-60°	-30°	0°	30°	60°	90°	120°	150°

1080 A.2.7 Spatial Audio Experience via Rotating Agent

1081 To help readers directly experience how spatial sound changes with orientation, we prepared a simple
 1082 interactive demo using a panoramic image. The scene is rotated in 30° increments, resulting in 12
 1083 viewpoints that cover a full 360° turn. Each viewpoint is accompanied by spatial audio corresponding
 1084 to the listener’s orientation.

1085 To explore this experience, please unzip the provided `rotate.zip` file and open `index.html` in
1086 your browser. The webpage allows you to perceive how the spatial characteristics of sound evolve as
1087 the agent rotates around the scene.

1088 Figure 18 provides a schematic top-down view to help readers intuitively understand the relative
1089 angle of each rotation step. Figure 19 shows a screenshot of the actual webpage, where the panoramic
1090 image and corresponding audio player are displayed.

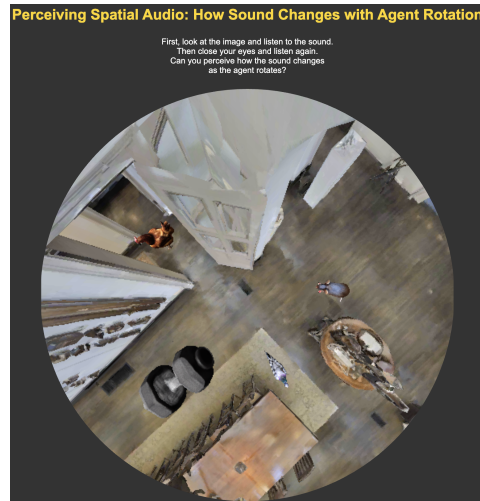


Figure 18: A schematic top-down view illustrating the agent’s 360° rotation. Used for angle reference only.



Figure 19: Screenshot of the interactive demo page showing the panoramic image and spatial audio player.

1091 A.2.8 Test Sample Viewer

1092 We provide a viewer for test samples. This interface displays a series of 360° panoramic images from
1093 the test set, each paired with corresponding spatial audio.

1094 To use the viewer, unzip the provided `test_samples.zip` file and open `test_samples.html`
1095 in your browser. The demo page allows users to visually inspect the scene while listening to the
1096 associated audio, which was used as input during model inference.

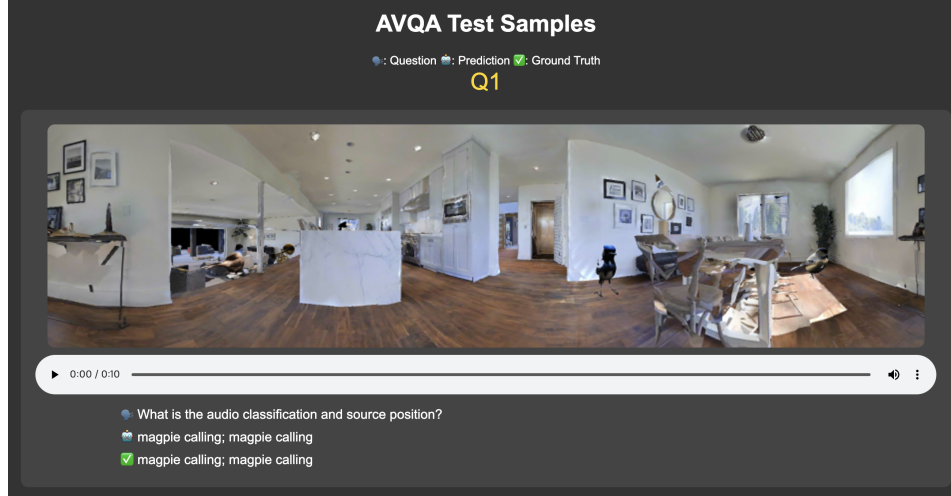


Figure 20: Screenshot of the test sample viewer. Each scene is presented with panoramic image and spatial audio.

A.2.9 New Assets

We introduce the *Hear You Are QA* dataset, a new benchmark for evaluating spatial audio-visual reasoning. It consists of 360° panoramic images, spatialized audio rendered using room impulse responses, and corresponding question–answer pairs across diverse spatial understanding tasks. The dataset is entirely synthetic and does not include any real-world human likenesses or sensitive content; therefore, no consent from individuals was required. The dataset will be released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

A.2.10 Dataset Statistics

Table 5 summarizes the number of samples per question type across the train, validation, and test splits. The dataset was designed to broadly cover key aspects of audio-visual spatial reasoning. Q1 and Q9 include a larger number of samples, based on the intuition that effectively disentangling semantic alignment and spatial localization during training can benefit the learning of other tasks as well. The remaining question types (Q2–Q8) are uniformly distributed to ensure balanced coverage of diverse spatial reasoning scenarios.

Table 5: Number of samples per question type in each split.

Question Type	Train	Val	Test
Q1	172,794	3,600	3,600
Q2	86,373	1,800	1,799
Q3	86,373	1,800	1,799
Q4	86,373	1,800	1,799
Q5	86,373	1,800	1,799
Q6	86,389	1,799	1,800
Q7	86,373	1,800	1,799
Q8	86,389	1,799	1,800
Q9	172,794	3,600	3,600

Visual and Audio Category Distributions

Figures 23–28 show distributions of the most frequent visual and audio categories.



Figure 22: 3D mesh examples from InstantMesh applied on diffusion outputs.

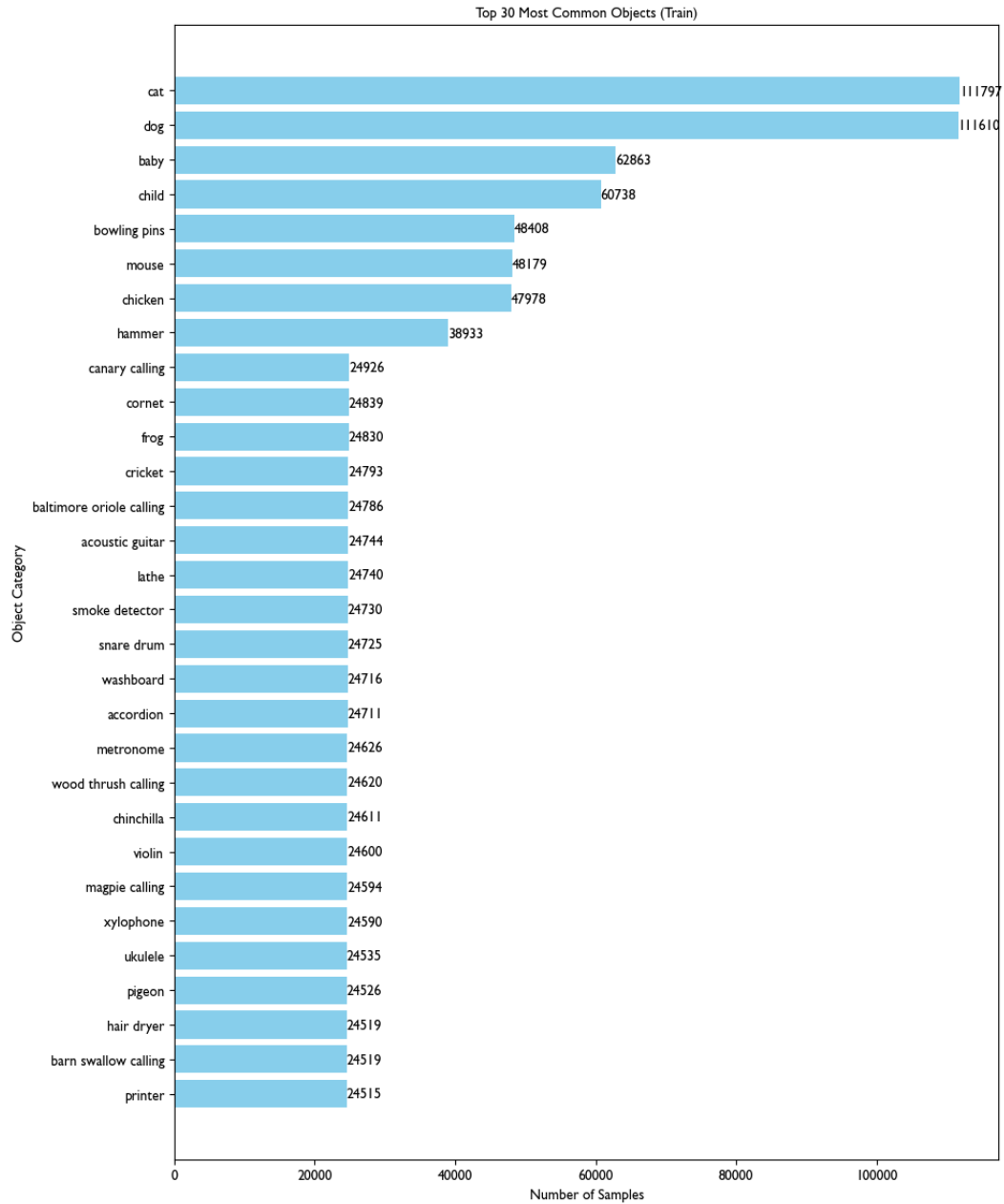


Figure 23: Top 30 most frequent visual object categories in the training set.

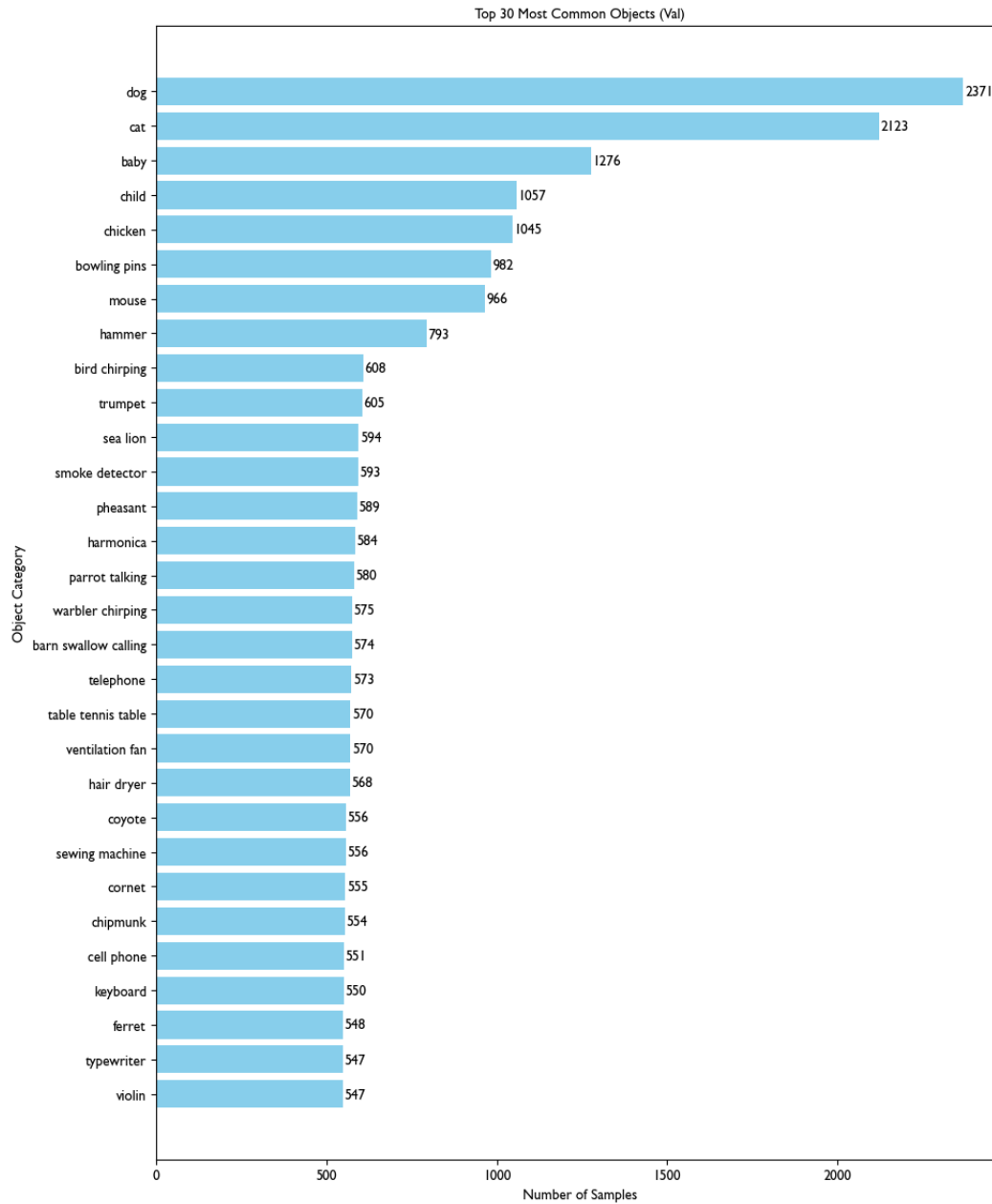


Figure 24: Top 30 most frequent visual object categories in the validation set.

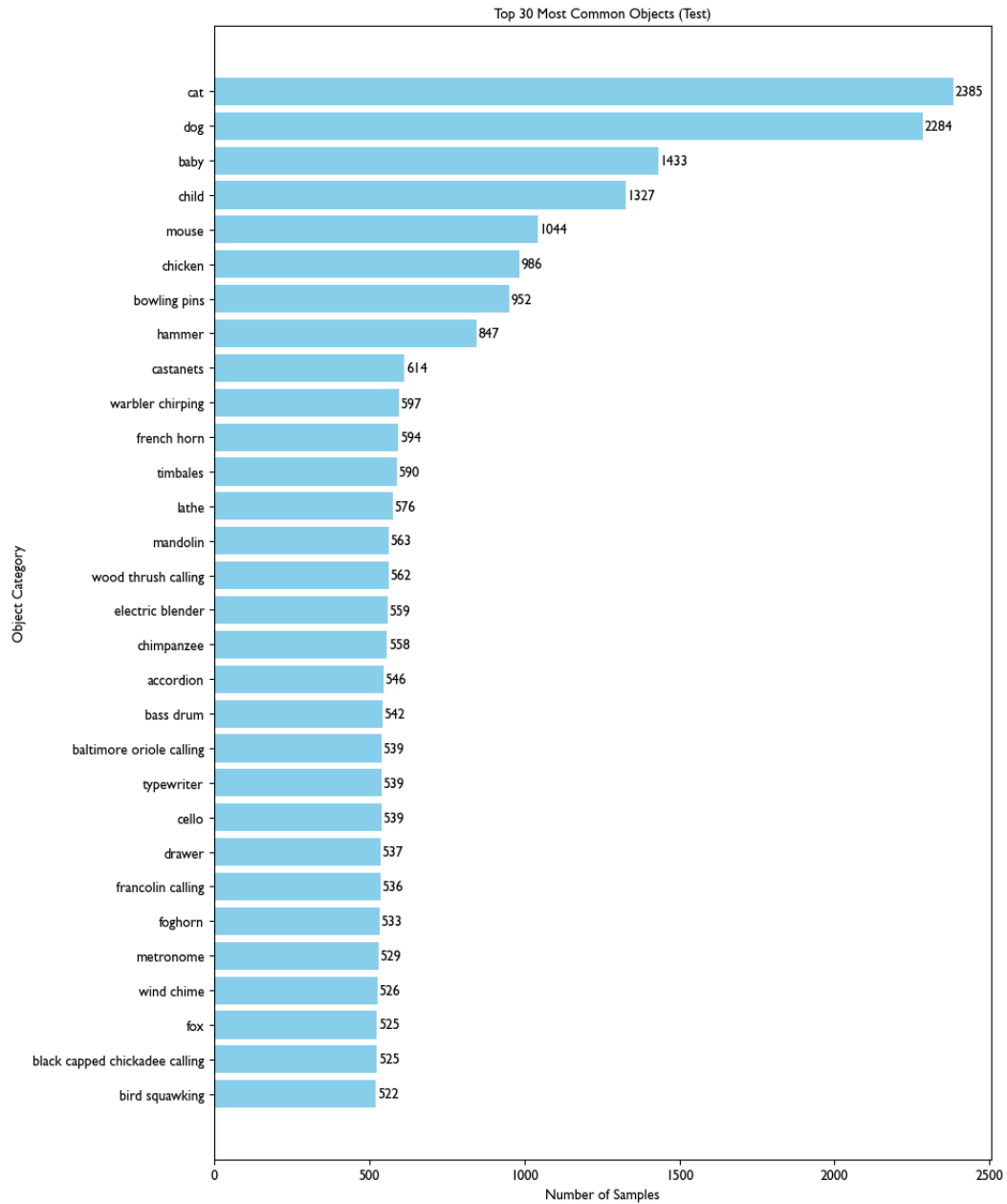


Figure 25: Top 30 most frequent visual object categories in the test set.

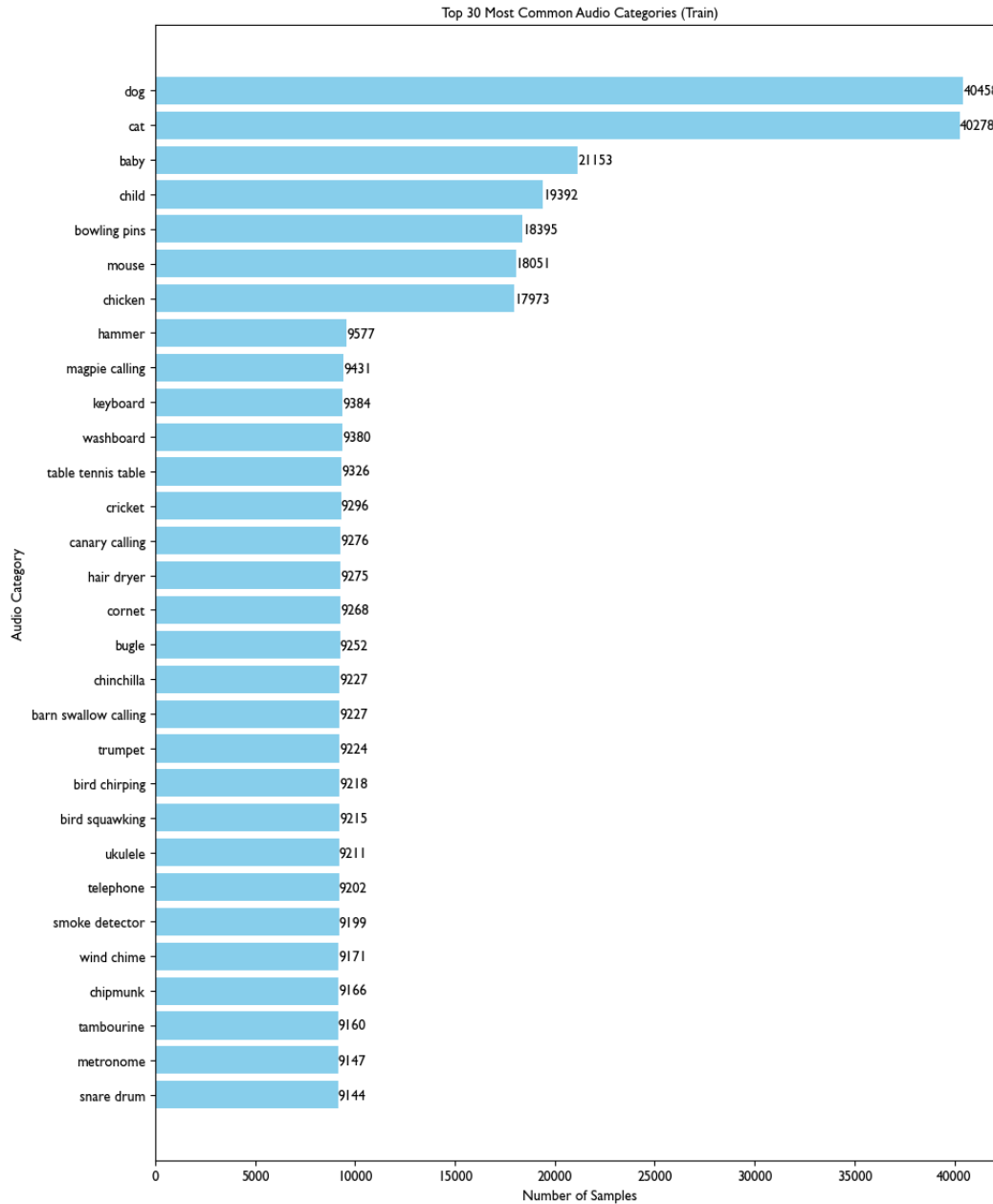


Figure 26: Top 30 most frequent audio object categories in the training set.

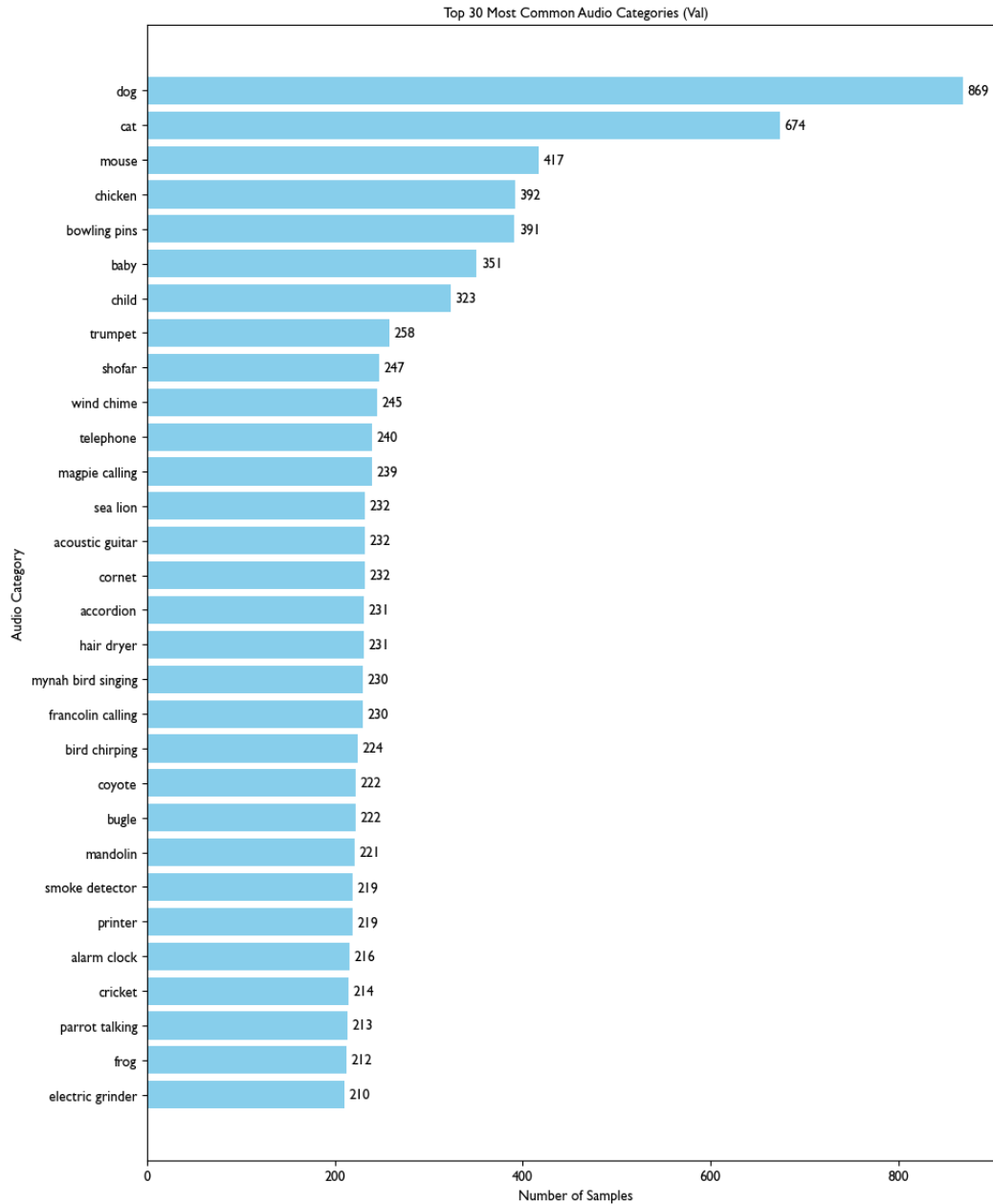


Figure 27: Top 30 most frequent audio object categories in the validation set.

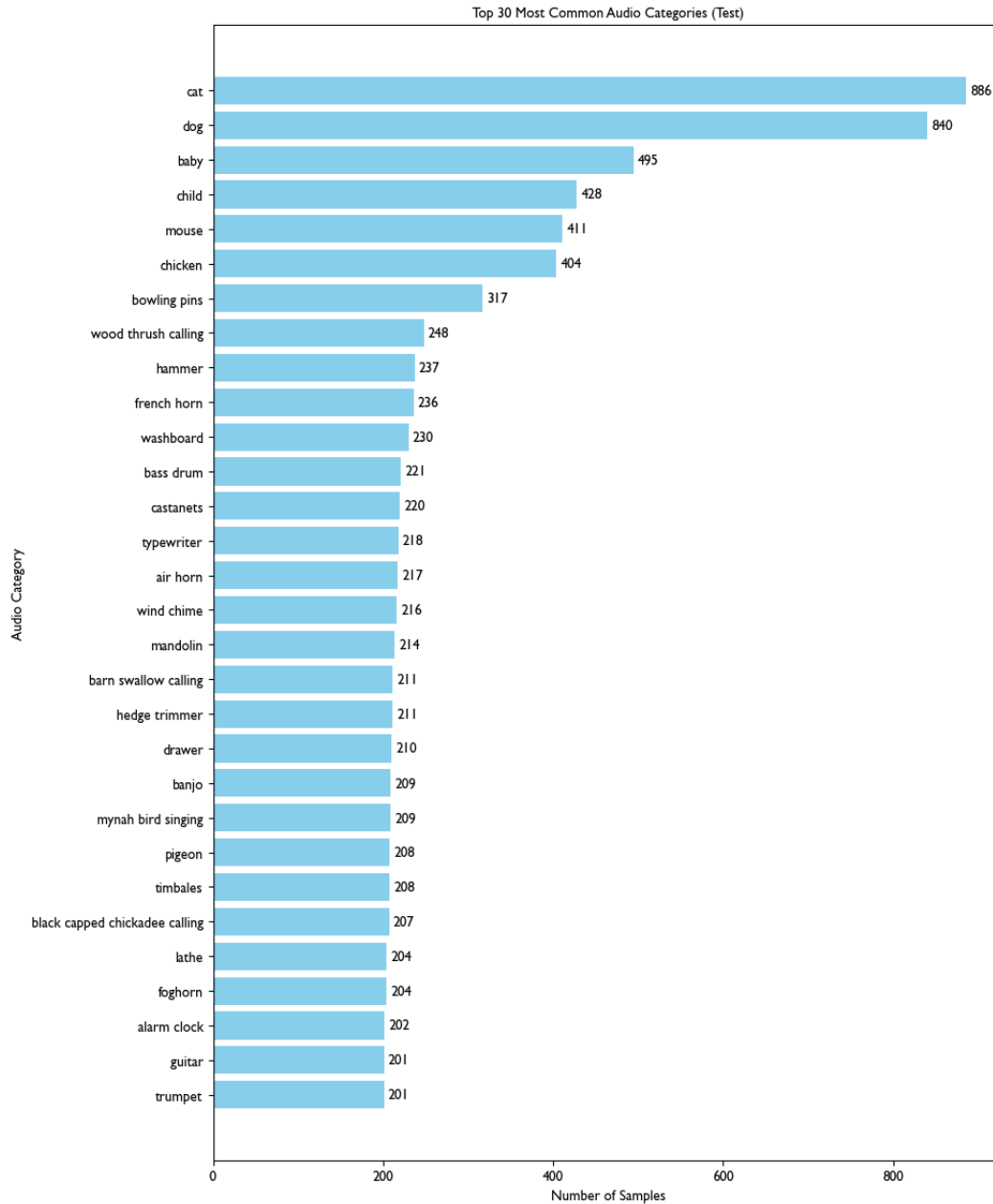


Figure 28: Top 30 most frequent audio object categories in the test set.

1113 A.3 Broader Impact

1114 Our work contributes to the development of multimodal agents capable of spatial reasoning in
1115 realistic 3D environments. By combining panoramic vision and binaural audio, we offer new
1116 benchmarks that encourage spatially grounded audio-visual understanding. Potential applications
1117 include assistive robotics and embodied AI systems. However, care must be taken to prevent misuse,
1118 such as surveillance.

1119 A.4 License of Existing Assets

1120 We use audio assets and 3D scene data derived from publicly available datasets, and generate 3D
1121 objects using publicly available code and pre-trained weights, all under permissible licenses:

- 1122 • **VGGSound** [8]: Available for commercial and research use under the Creative Commons
1123 Attribution 4.0 International (CC BY 4.0) license. Copyright remains with the original video
1124 owners. (<https://github.com/hche11/VGGSound>)
- 1125 • **Matterport3D** [2]: Distributed under the Matterport3D Terms of Use for academic research,
1126 and subject to a CC BY-NC-SA 3.0 US license. (<https://github.com/niessner/Matterport>)
- 1127 • **SoundSpaces 2.0** [4]: Licensed under CC BY 4.0. Task datasets and models derived from
1128 Matterport3D are covered under Matterport3D’s terms.
1129 (<https://github.com/facebookresearch/sound-spaces>)
- 1130 • **InstantMesh** [49]: Code is licensed under the Apache License 2.0.
1131 (<https://github.com/TencentARC/InstantMesh>)
- 1132 • **Stable Diffusion 3** [35]: Code is licensed under the MIT License. Model
1133 weights are released by StabilityAI under the CreativeML Open RAIL++-M License.
1134 (<https://huggingface.co/stabilityai/stable-diffusion-3-medium>)

1135 All generated assets are synthetic and not intended to resemble real individuals or sensitive environ-
1136 ments. Proper credit has been given to the original authors, and we have fully complied with the
1137 terms of each license.

1138 A.5 LLM Usage

1139 Our method leverages a large language model (VideoLLaMA2) as the final reasoning module. Visual
1140 and audio features are first projected and aligned through a multi-query transformer, then passed
1141 to the LLM via a prompt-based format. For question generation and template population, we used
1142 ChatGPT-4o to ensure linguistic diversity and format consistency. All prompts were manually
1143 inspected to avoid hallucinated or biased patterns.

1144 In addition, we used ChatGPT-4o to assist with grammar and style corrections during paper writing.
1145 No part of the scientific content or experimental contribution was generated by the model.