Hear you are: Teaching LLMs Spatial Reasoning with Vision and Spatial Sound

Supplementary Material

8. Technical Appendices and Supplementary Material

8.1. Experimental Details

8.1.1. Baseline Experiments

We evaluated several baseline models to assess spatial reasoning capabilities of audio-visual methods. These include ISSL [43], ACL-SSL [32], and VideoLLaMA2 [10]. 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 4 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 audio cues and localizes the sound source, enabling it to infer what visual object is present at that location and produce the correct answer.

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

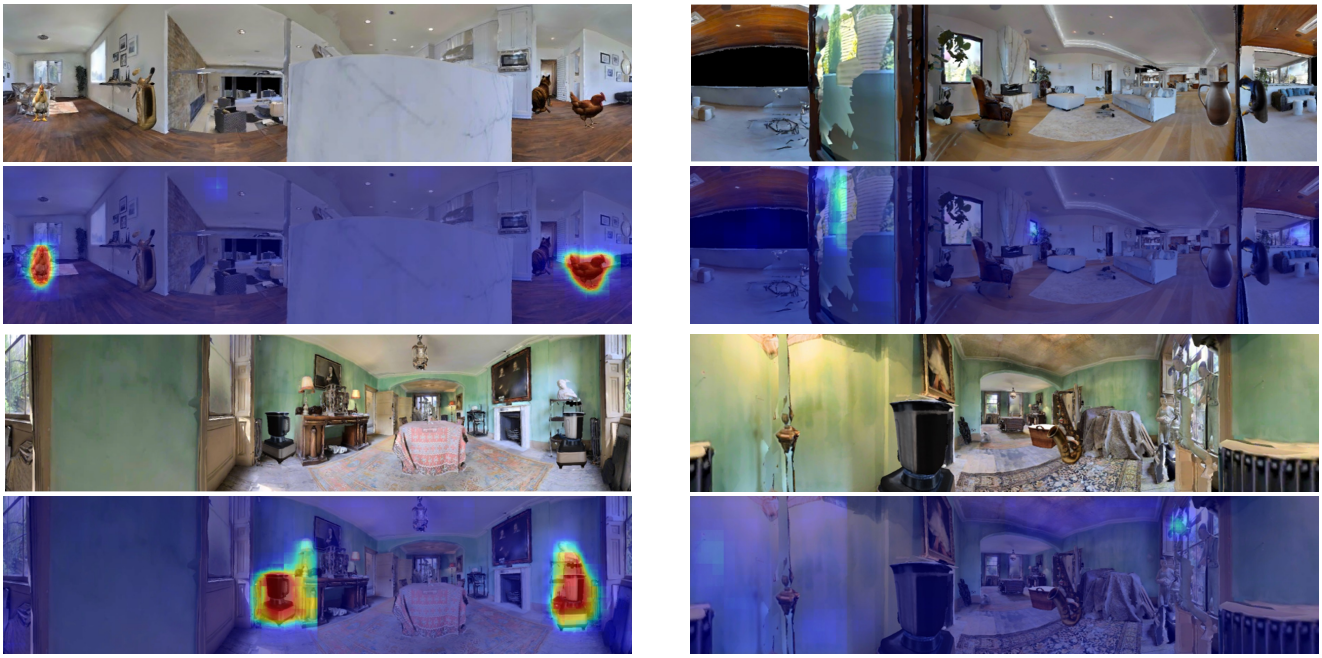

Figure 4. Qualitative results from the ACL-SSL baseline. The model highlights semantically matching regions but fails to distinguish the actual sound source due to the lack of spatial reasoning.

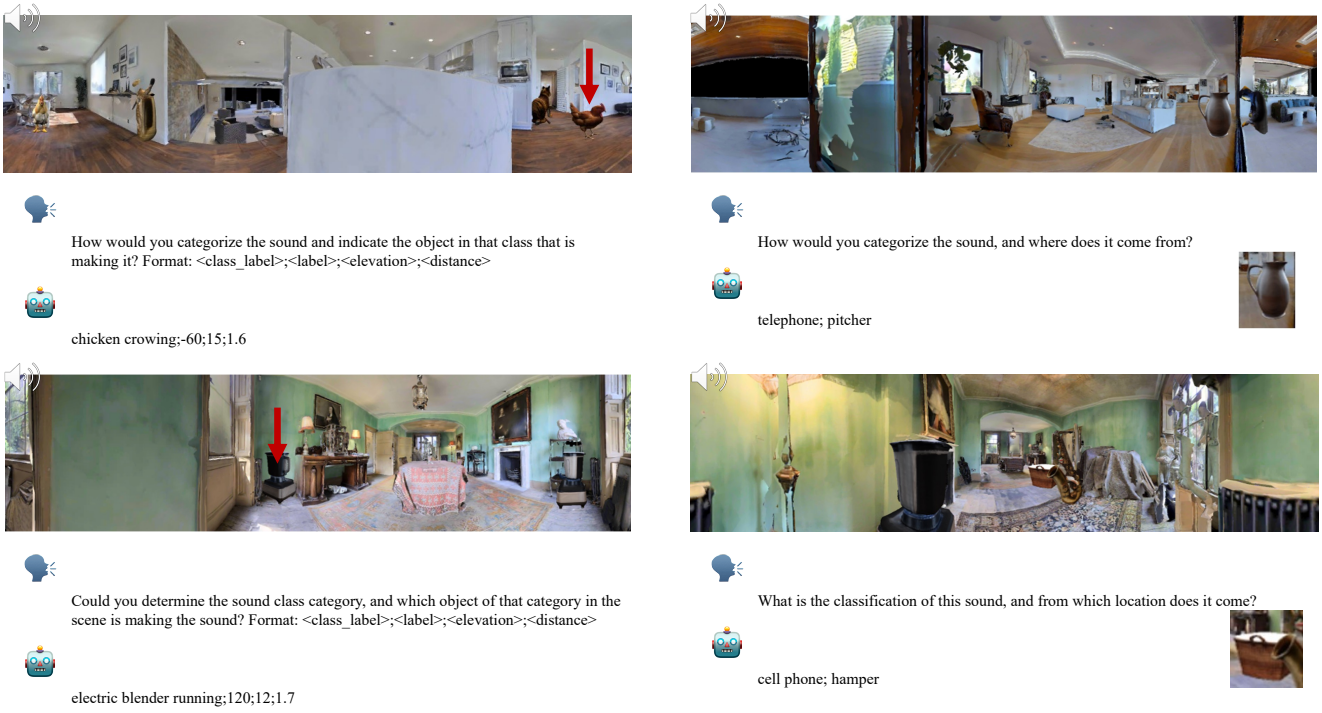

Figure 5. Qualitative results from our model. By leveraging spatial audio cues, the model accurately localizes the sound source and identifies the correct visual object at that location.

8.1.2. Encoder Warm Start QA Generation

To pre-train each encoder on spatially grounded audio and visual representations respectively, we constructed a simple

uni-modal QA dataset derived from simulation metadata. Each data sample contains a 360° image, spatial audio, and object positions with annotations.

We synthesized QA pairs in two modalities:

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

Answer Format Examples:

- `<wav> What are the predicted azimuth and elevation angles, and the distance to the sound source?`
  Answer: `(90, -10), 2.3 meters`
- `<wav> What sound did you detect?`
  Answer: `typewriter`
- `<rgb> What are the predicted azimuth and elevation angles, and the distance to the typewriter?`
  Answer: `(90, -10), 2.3 meters`
- `<rgb> What visual objects did you detect at (60, 0), 1.7 meters?`
  Answer: `typewriter`

As we use Spatial-AST [56] as the audio encoder, which is already pre-trained to capture spatial cues, we only need to train the audio projector to align with the LLM backbone. This makes the adaptation process relatively simple. In contrast, the image encoder [51] is not initially designed for 360° panoramic inputs and suffers from geometric distortion. To address this, we first LoRA fine-tune the image encoder to recognize object class labels under panoramic distortion. Once it learns to handle such geometric transformations, we further train it to answer questions requiring spatial position prediction, such as azimuth, elevation, and distance.

8.1.3. Reproducibility.

We will release the full codebase, panoramic image dataset, reverb files, model checkpoints, and detailed instructions for reproducing all experiments upon acceptance. Please refer to the VGGSound [6] for the audio files used in this study.

8.2. Dataset Details

8.2.1. Explanation on the Visual Scene

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

8.2.2. Generated 3D Objects

We use Stable Diffusion 3 [36] and InstantMesh [53] to synthesize new 3D audio-visual objects, enabling the diversification of spatial grounding scenarios. The size of each object category is manually determined based on the common sense judgments of three annotators. We classify objects into four size levels: smallest, small, medium, and large. For each size level, we define a representative base size and apply a random variation of ±20% to introduce natural variation.

8.2.3. Explanation on Azimuth and Elevation

Figure 6 consists of two visualizations. The top image is a 2D equirectangular projection of a 360° indoor scene. The bottom image shows a circular representation of the same scene, in which the 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 $\theta = 90°$, which serves as the reference front-facing direction. The full range of these angles is defined as:

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

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.

8.2.4. Question Types

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

Q1: Spatial Correspondence The scene includes a `backpack` and a `dog`, along with a `cell phone` sound that has no corresponding visual object. The question is: *"What is the sound class category? Where is the sound coming from?"* The correct answer is `cell phone; backpack`. Although the phone itself is not visible,

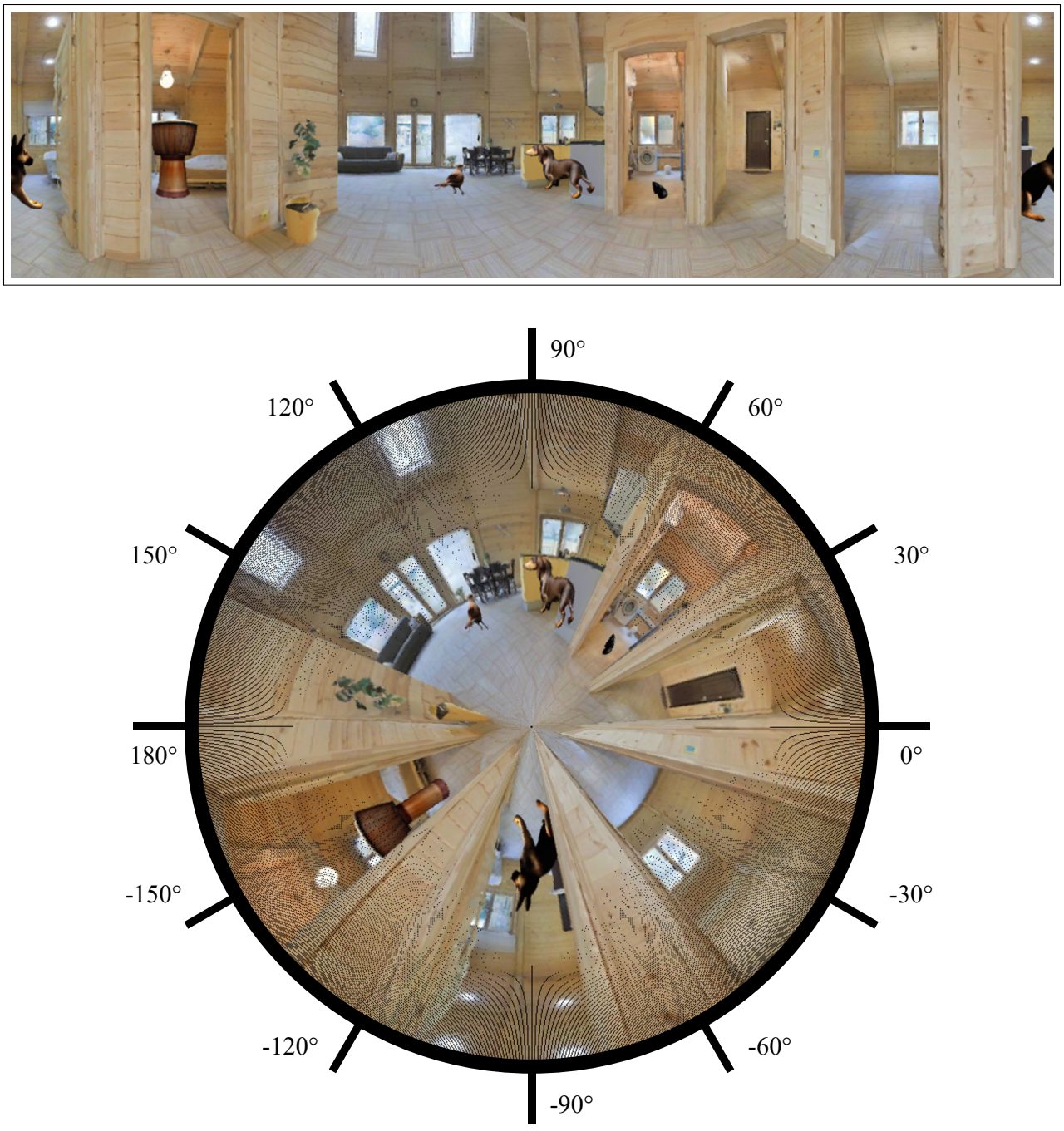

Figure 6. Equirectangular (top) and circular views of a 360° scene with azimuth annotations (bottom).

the sound is localized at the position of the backpack. The model is expected to recognize the audio as a `cell phone` ringtone and associate it with the `backpack`, which occupies the same location.

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

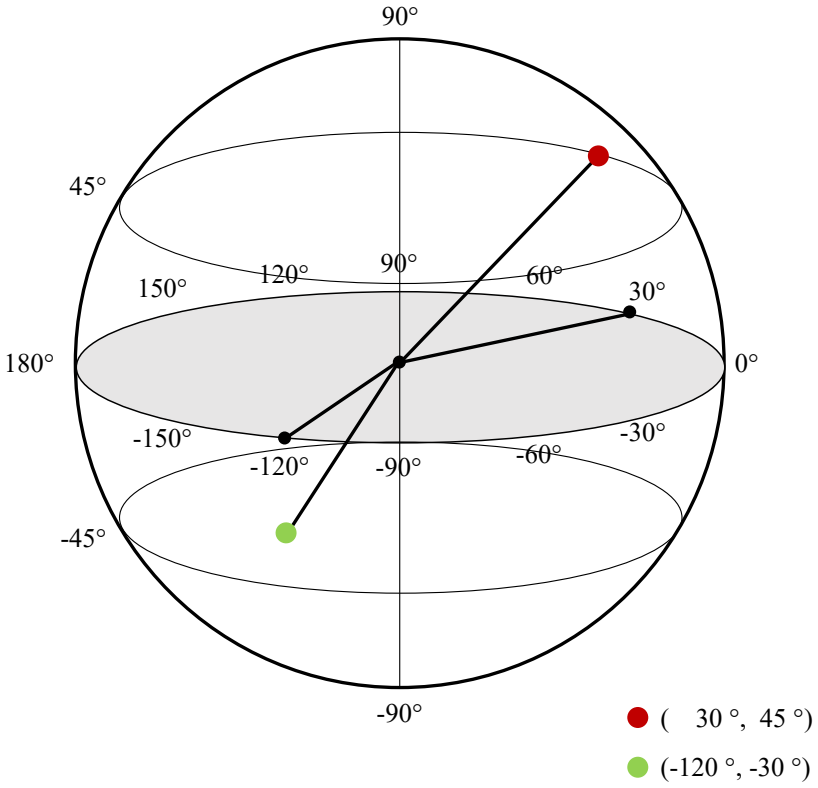

Figure 7. Spherical visualization of azimuth (θ) and elevation (ϕ) angles in a 360° panoramic setting. The agent is positioned at the center, facing 90°, with azimuth ranging from -180° to 180° and elevation from -90° to 90°.

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

right. In the latter case, the sound is still assigned to a specific location, even though no corresponding visual object is present. This invisible audio setting corresponds to the condition analyzed in the ablation study shown in Table 4.

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

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.

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.

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.

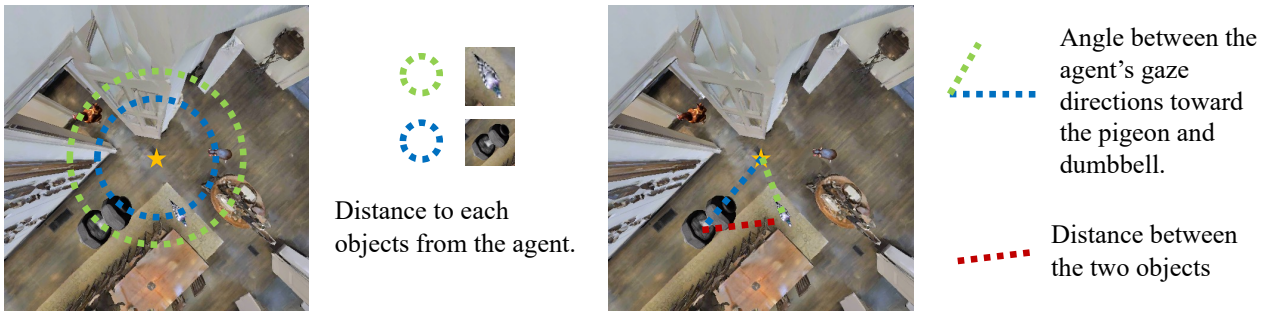

Figure 11. Left: object-to-agent distances (Q2). Right: angular and spatial relationship between two objects (Q4).

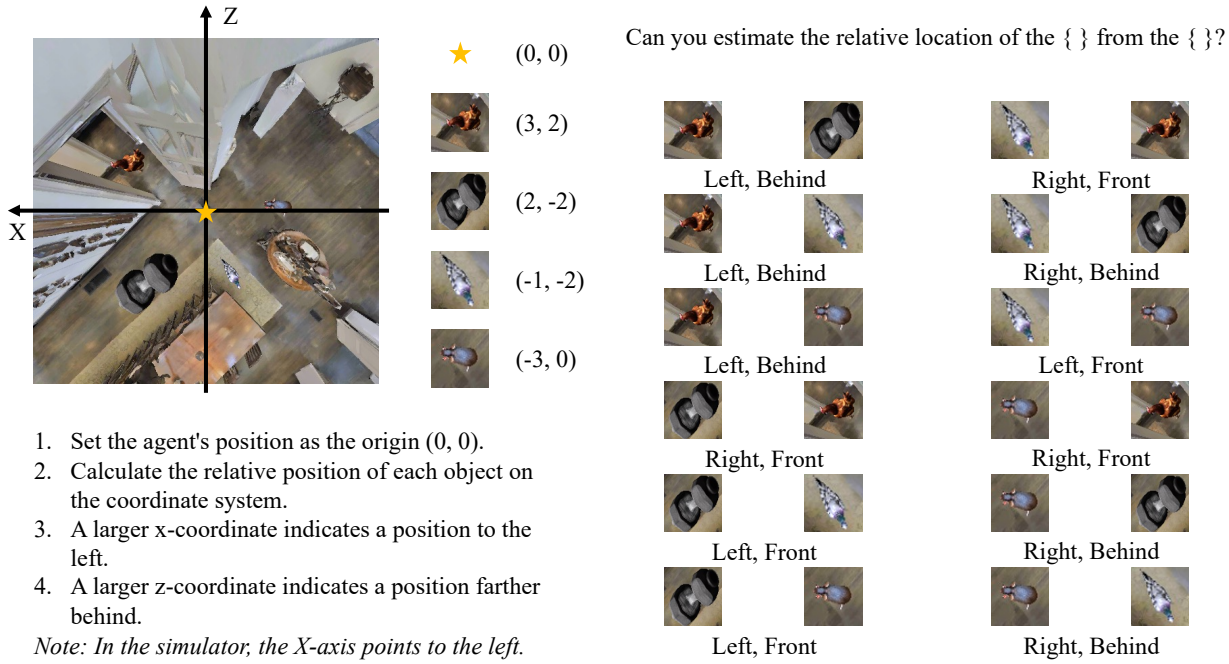

Figure 12. Bird's-eye view of object locations (left) and Q3-style relative location question examples (right).

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

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

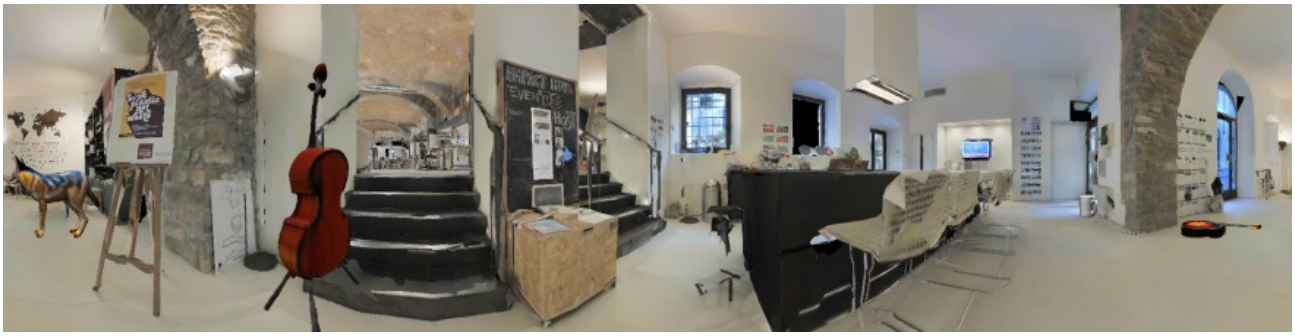

Figure 13. Q5 example: Estimate sound position and identify the emitting object class.

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

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

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.

8.2.5. 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**

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

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

  - 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)**
  - What is the object in the scene located at {azimuth}, {distance} meters? Is it making a sound?
  - Which object is found at {azimuth}, {distance} meters, and is it currently making a sound?
  - Can you identify the object positioned at {azimuth}, {distance} meters, and is it emitting any sound?
- **Q6: Spatial & Semantic Correspondence (Multiple visual objects semantically match the audio)**
  - What is the object located at {azimuth}, {distance} meters in the scene, and is it producing a sound?
  - Can you determine the object located at {azimuth}, {distance} meters and confirm if it is producing sound?
  - What is the object at {azimuth}, {distance} meters, and is it the source of the audible signal?
- **Q7: Spatial & Semantic Correspondence (One visual object semantically matches the audio)**
  - Given multiple visual objects, which one is making a sound, and where is it located?
  - Which object among the visual objects is producing a sound, and where is it placed?
  - From the visual objects in the scene, which one is producing a sound, and where is it positioned?
- **Q8: Spatial & Semantic Correspondence (Multiple visual objects semantically match the audio)**
  - What is the sound class, and which object of that type in the scene is the source of the sound?
  - Can you determine the category of the sound and identify the object within that category that is generating it?
  - Which object of the sound class is producing the audio in the scene?
- **Q9: Semantic Co-occurrence**
  - What is the sound class category? Is the sound source

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

visible in the scene?
  - Can you specify the sound type and indicate whether its source can be seen in the scene?
  - What is the classification of the sound, and is the sound-emitting object in view?

8.2.6. Prompt Formatting Strategy

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

- **A directive phrase**, e.g., `Please provide the answer in the following format: ...`
- **A format schema**, e.g., `<label>;<distance>`
- **A concrete example**, e.g., `(e.g., J;3.2)`

  The examples were dynamically generated per instance. For example:
  - Azimuth labels (A–L) were randomly selected for each question from a uniform pool.
  - Distances were sampled from $[0.5, 4.0]$ meters.
  - Elevations from $[-20, 20]$ degrees.
  - Class names (e.g., `blender`, `accordion`) from the dataset's audio/visual categories.
- **A label explanation block** (for azimuth questions), e.g., `A: 180°, B: -150°, ...`

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

- **Q1**
  - Suffix:
    `Please provide the answer in the following format: class_category; sound_source (e.g., xylophone;eagle)`
  - Purpose: to ensure joint prediction of the sound category and its visual counterpart.
- **Q2**
  - Suffix:
    `Please provide the answer Yes or No`
  - Purpose: to elicit binary (Yes/No) responses based on relative spatial reasoning.
- **Q3**
  - Suffix:
    `Please provide the answer in the following format: <left_right>;<up_down>;<front_behind>; <distance> (e.g., left;up;behind;2.3)`
  - Purpose: to guide spatial reasoning using direction and distance.
- **Q4**
  - Suffix:
    `Please answer using labels A{L, where: A: 180°, B: -150°, ..., L: 150°. Format: <label>;<distance> (e.g., J;3.2)`
  - Purpose: to map coarse directions to discrete azimuth bins with distance.
- **Q5, Q6, and Q9**
  - Suffix:
    `Please provide the answer in the following format: <class_label>;<yes_no> (e.g., vacuum;Yes)`
  - Purpose: to encourage semantic grounding with binary decision making.
- **Q7 and Q8**
  - Suffix:
    `Please answer using labels A{L, where: A: 180°, ..., L: 150°. Format: <class_label>;<label>;<elevation>; <distance> (e.g., accordion;H;5.0;1.8)`
  - Purpose: to map coarse directions to discrete azimuth bins with distance, and to jointly classify the sound type and localize the object.

Why Discrete Labels (A–L)? We discretized the azimuth direction into 12 evenly spaced bins (A–L), each represent-

Table 5. **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°

ing a 30° increment in the clockwise direction starting from A (180°), with J corresponding to the front (90°). This approach offers several advantages. Notably, it helps avoid biased numeric outputs during training. When the model was prompted to directly generate azimuth values as raw numbers, we found that it frequently produced certain patterns such as `123`, likely influenced by pretraining exposure to common number sequences. These patterns disrupted training stability, making discrete labels a more robust and interpretable alternative. In addition, it facilitates simple evaluation in direction-of-arrival and localization tasks. The specific azimuth bin definitions are shown in Table 5.

8.2.7. Spatial Audio Experience via Rotating Agent

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

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

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

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

Table 6. **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

8.2.8. Test Sample Viewer

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

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

8.2.9. Dataset Statistics

Table 6 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.

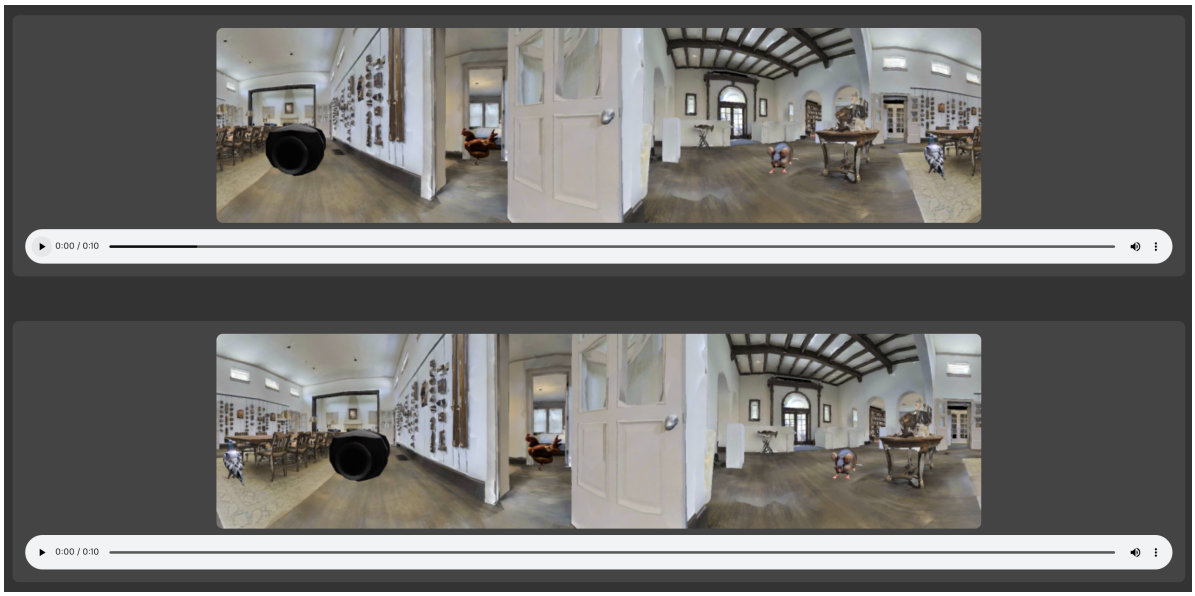

Figure 19. Screenshot of the interactive demo page showing the panoramic image and spatial audio player.

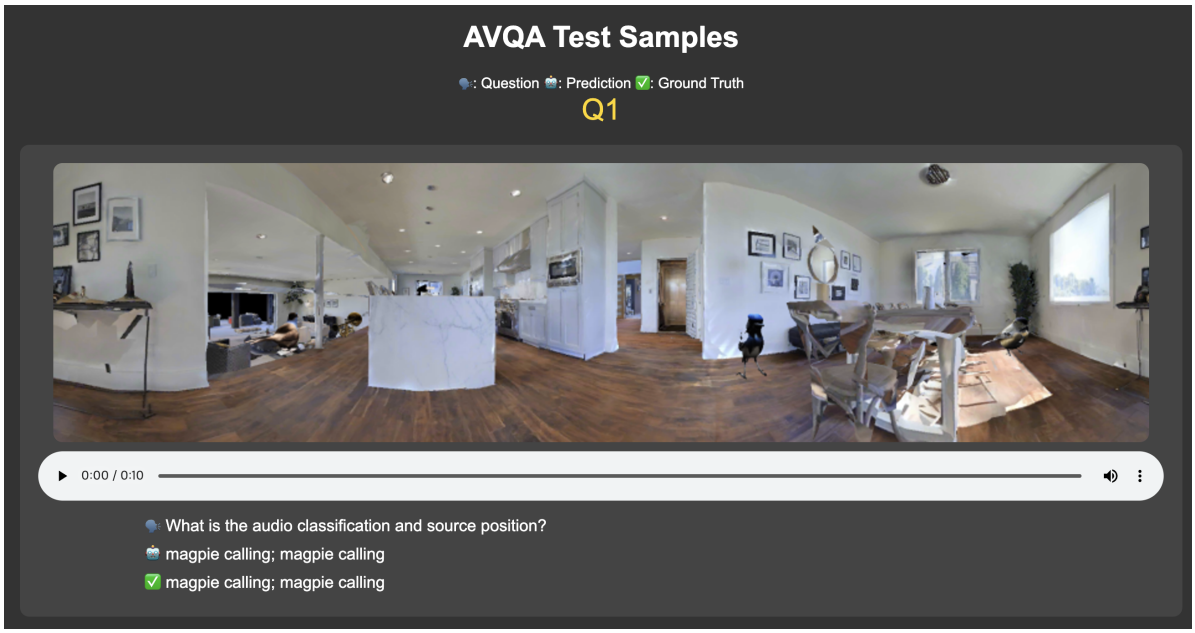

Figure 20. Screenshot of the test sample viewer. Each scene is presented with panoramic image and spatial audio.

Visual and Audio Category Distributions

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

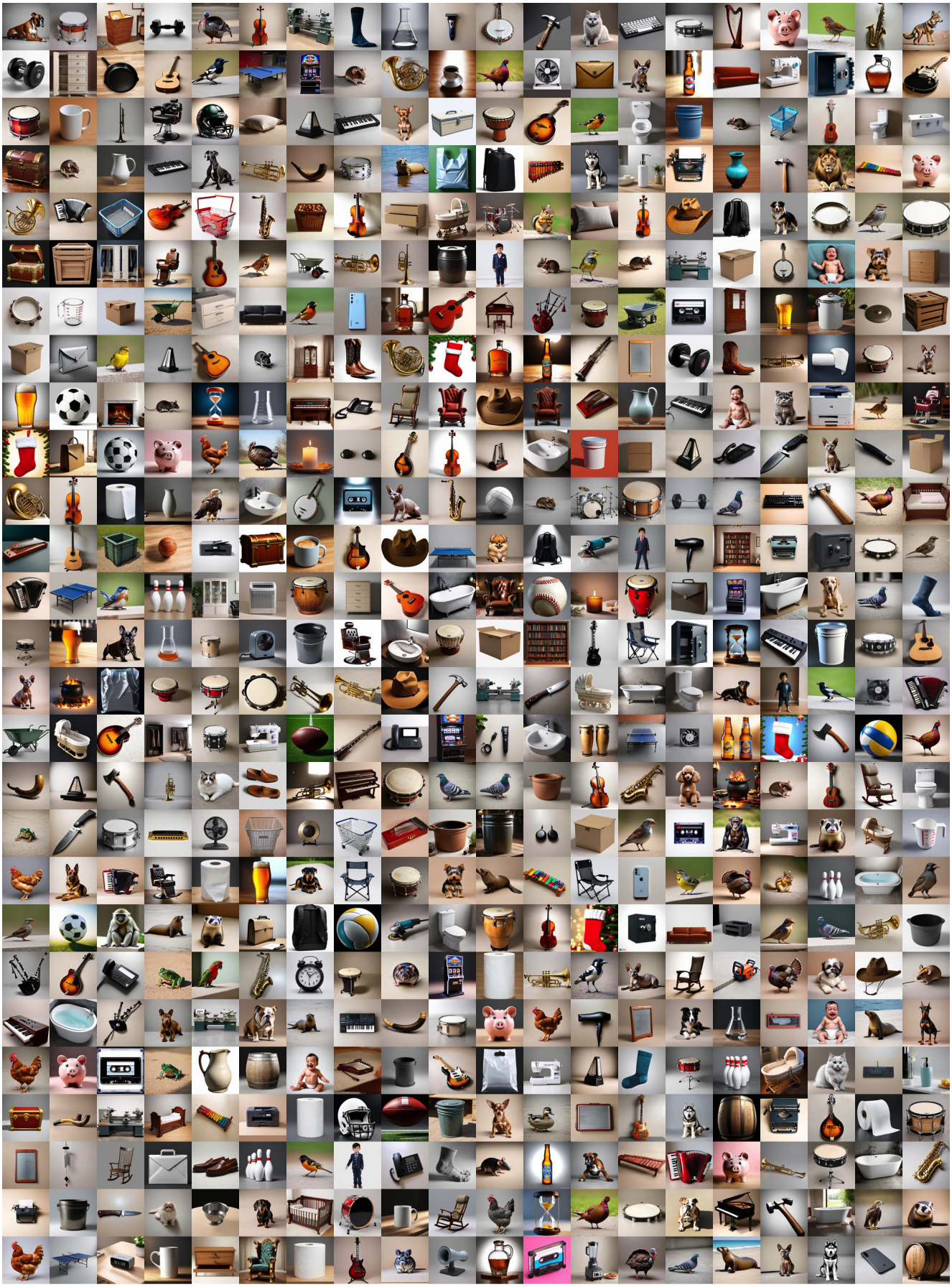

Figure 21. Generated visual samples using Stable Diffusion3.

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

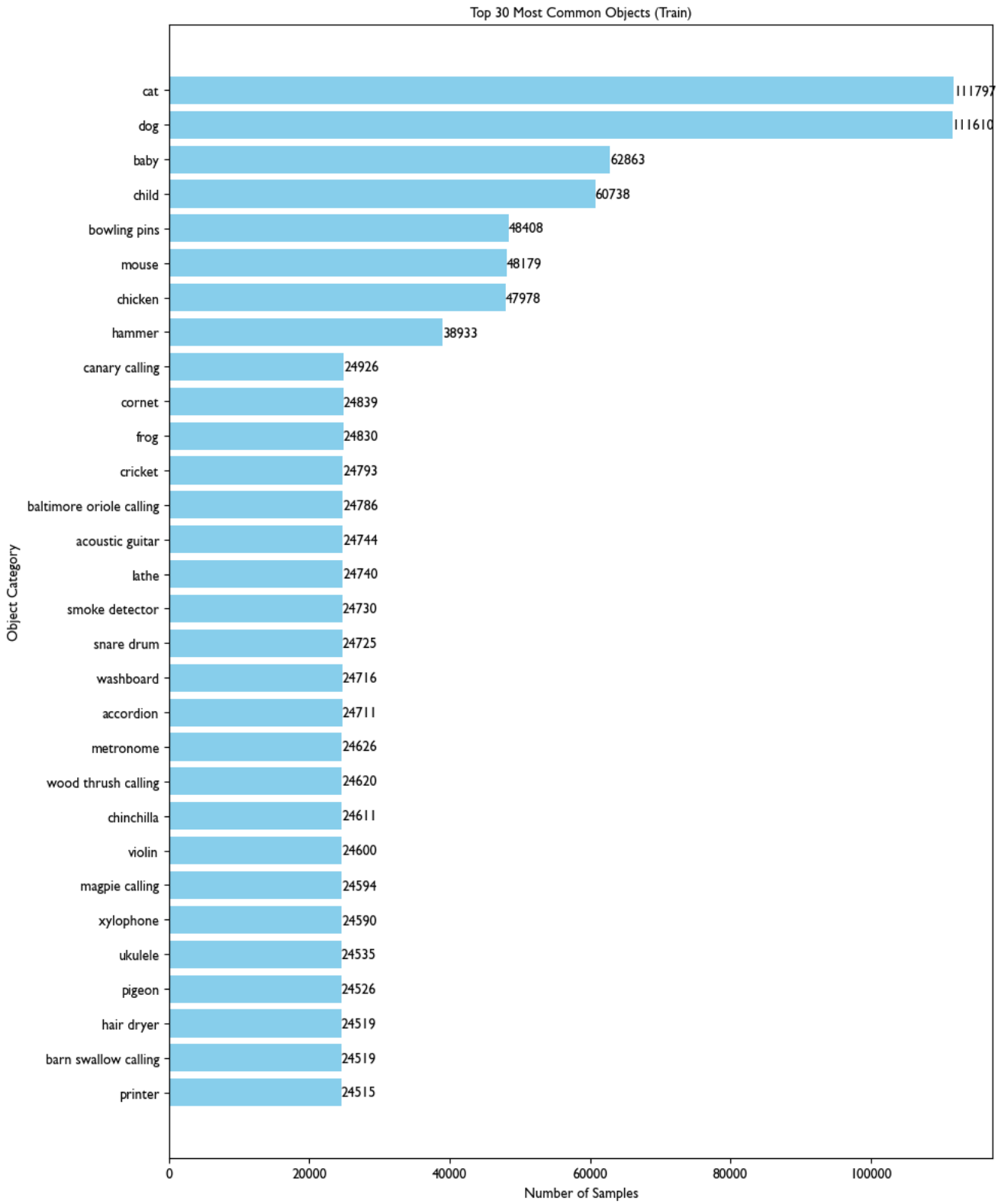

Figure 23. Top 30 most frequent visual object categories in the training set.

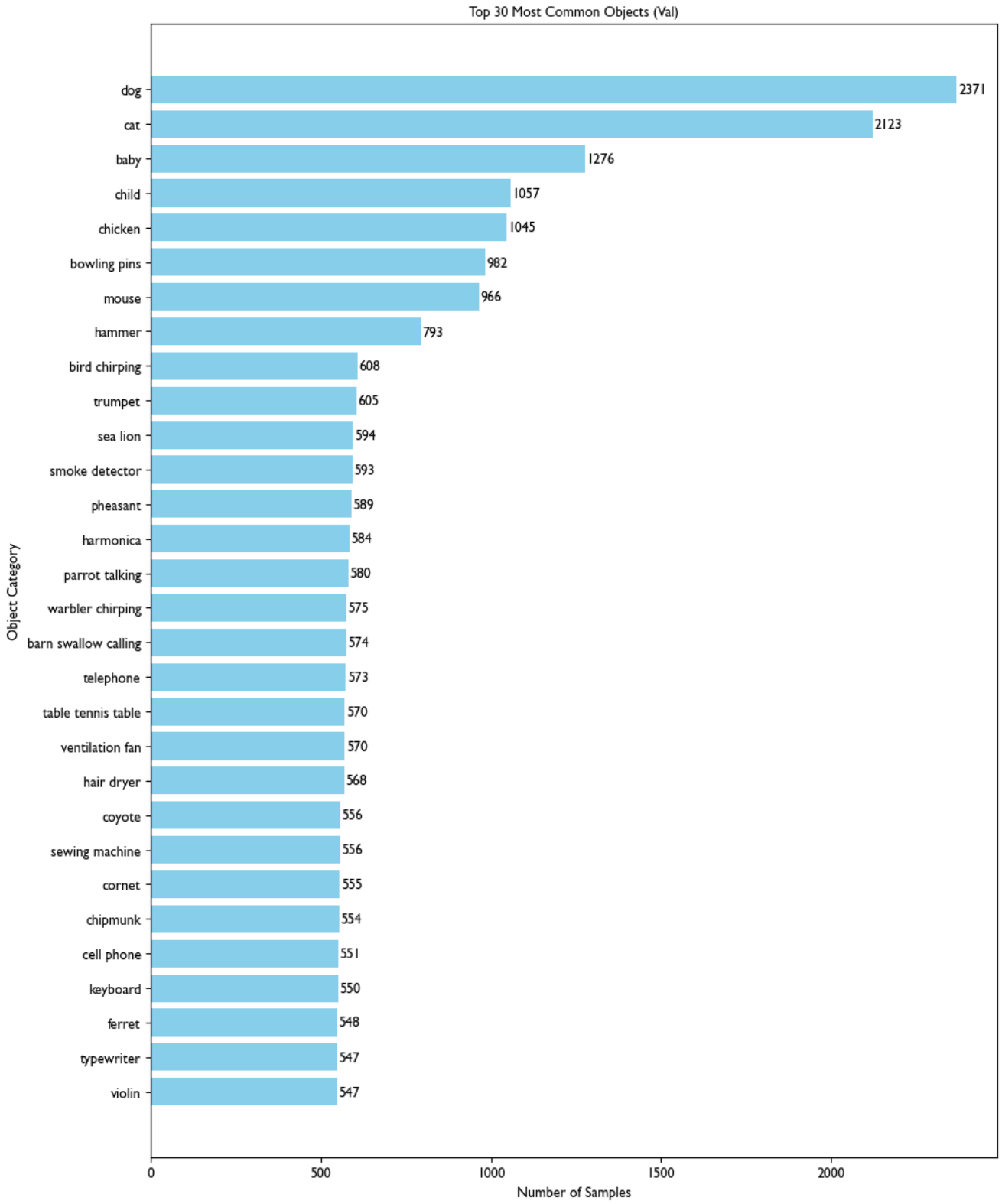

Figure 24. Top 30 most frequent visual object categories in the validation set.

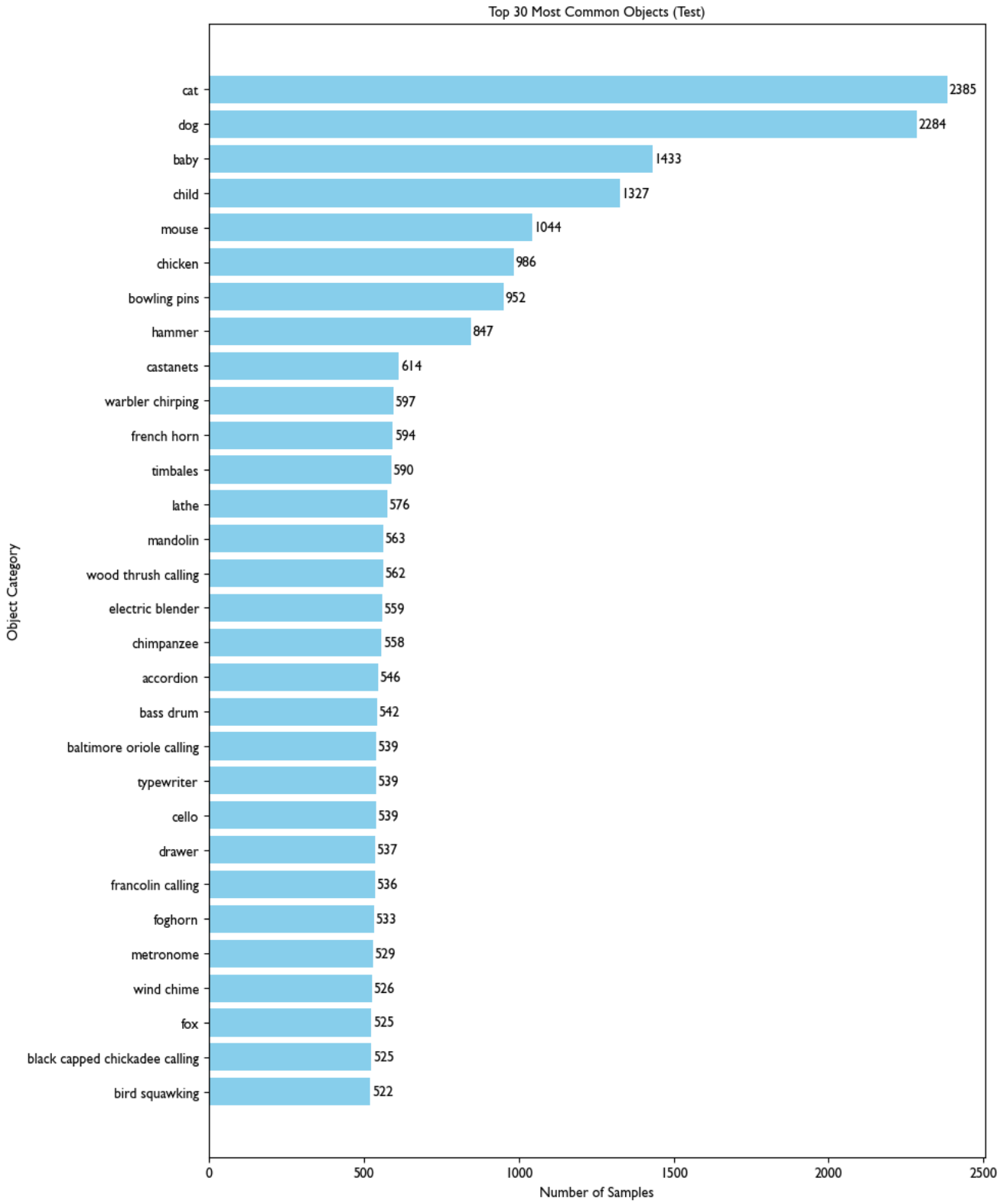

Figure 25. Top 30 most frequent visual object categories in the test set.

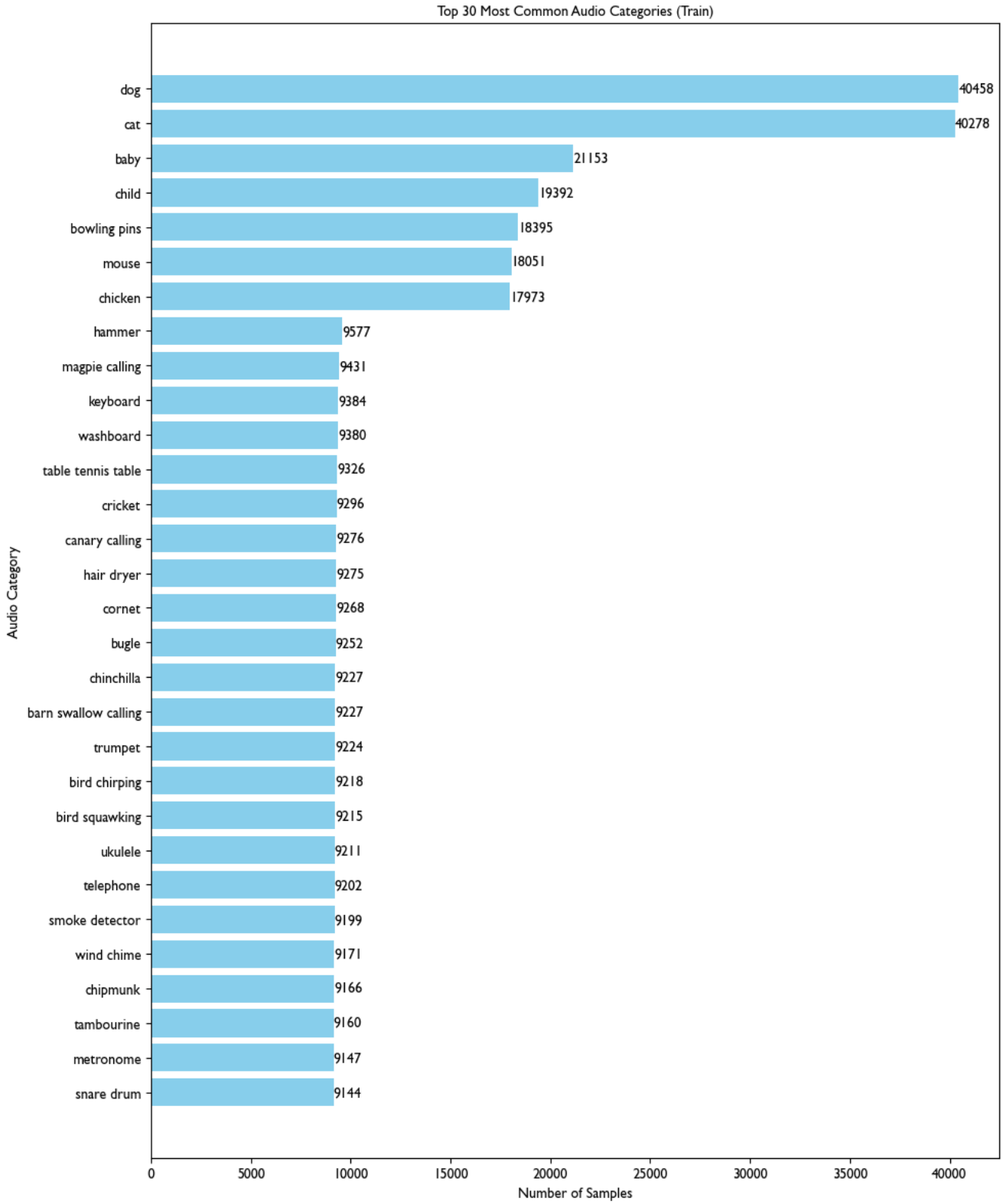

Figure 26. Top 30 most frequent audio object categories in the training set.

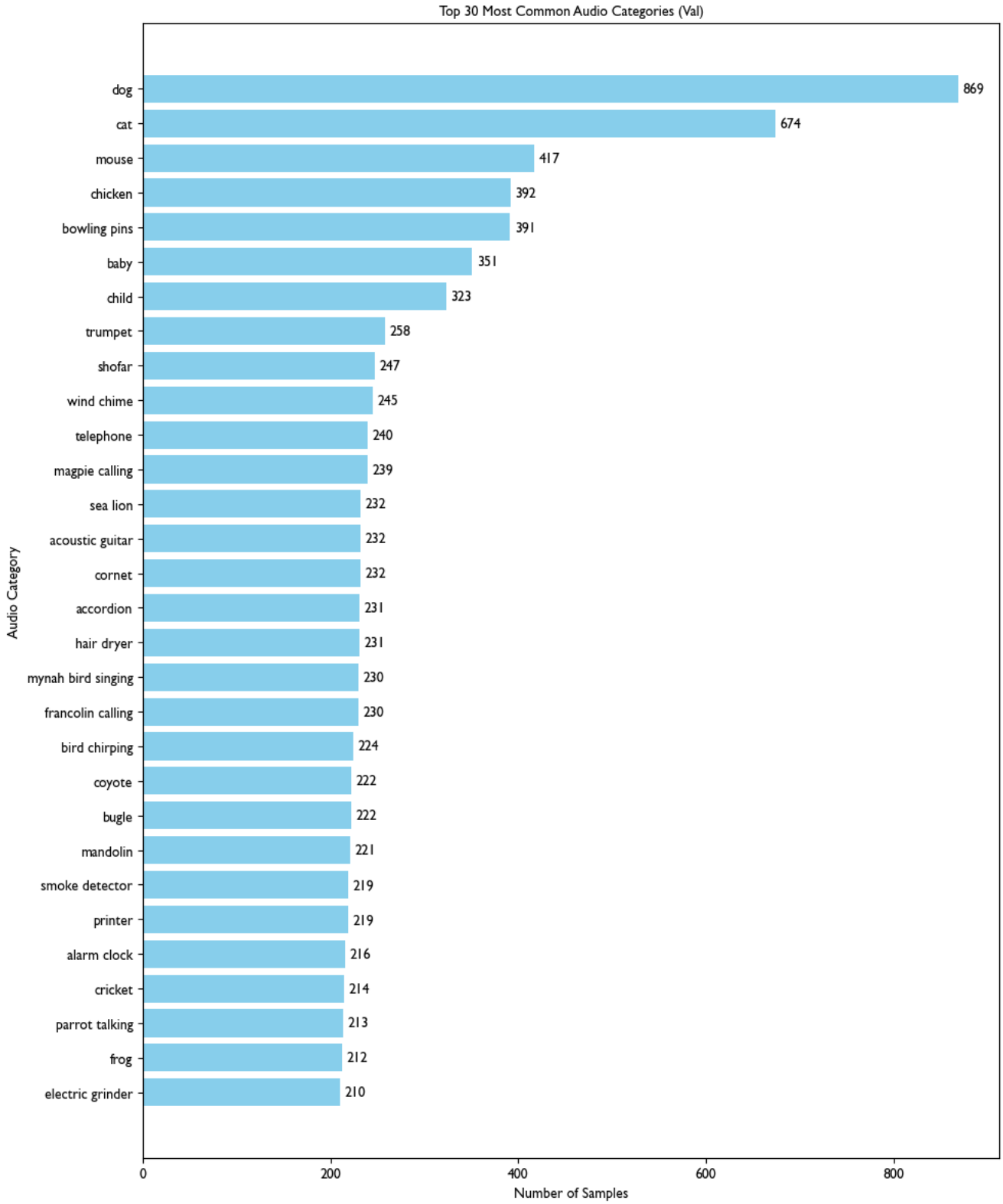

Figure 27. Top 30 most frequent audio object categories in the validation set.

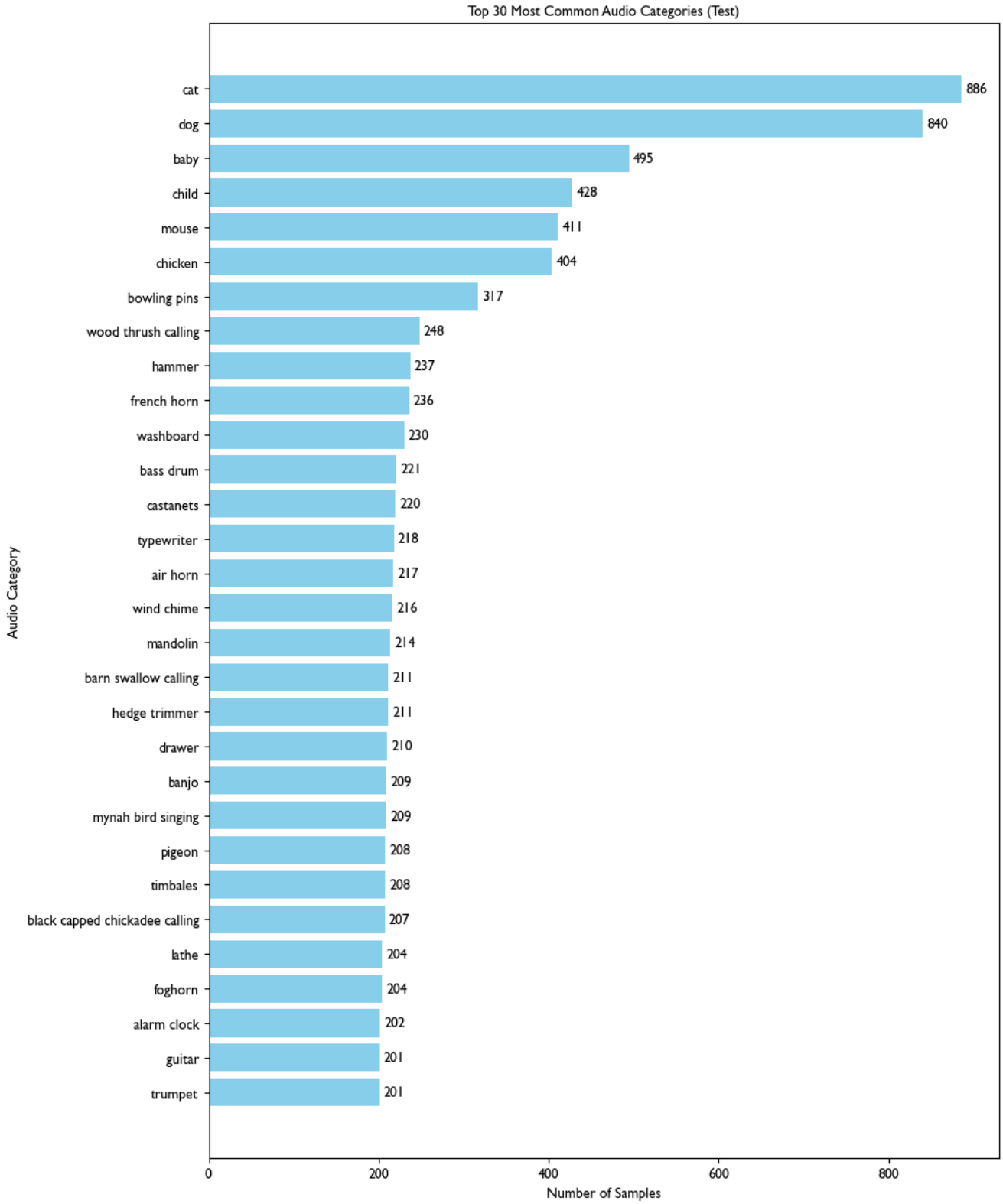

Figure 28. Top 30 most frequent audio object categories in the test set.