

# Acoustic Field Video for Multimodal Scene Understanding

Daehwa Kim  
Carnegie Mellon University  
Pittsburgh, Pennsylvania, USA  
daehwak@cs.cmu.edu

Chris Harrison  
Carnegie Mellon University  
Pittsburgh, Pennsylvania, USA  
chris.harrison@cs.cmu.edu










	Conventional Video	Acoustic Field Video		
 Smart Glasses			<b>Prompt</b>   "Is the dishwasher done running?"	
			<b>Gemini w/ Conventional Video</b>	Yes, the dishwasher is done running. The blue indicator light is illuminated, and there are no audible sounds of it being in operation. ✗
			<b>Gemini w/ Conventional + Acoustic Field Video</b>	No, the dishwasher is still running. ✓
 Smart Speaker			<b>Prompt</b>   "What stage of coffee making are we in?"	
			<b>Gemini w/ Conventional Video</b>	The coffee is in the brewing stage, as the espresso machine is actively extracting a shot into the cup. ✗
			<b>Gemini w/ Conventional + Acoustic Field Video</b>	Water is being boiled in the electric kettle. ✓
 Mobile Robot			<b>Prompt</b>   "Which room is the user calling you from?"	
			<b>Gemini w/ Conventional Video</b>	The user's speech is cut off before they state which room they are in, so the specific room cannot be identified. ✗
			<b>Gemini w/ Conventional + Acoustic Field Video</b>	The user is calling from the room through the open doorway in the center. ✓

Figure 1. Many everyday scene-understanding tasks remain unconstrained when relying solely on conventional video (with audio) input. We show that adding acoustic field video, a modality that visualizes the spatial distribution of sound, significantly improves multimodal reasoning. Moreover, the necessary hardware can be practically integrated into many platforms, from smart glasses to robots.

## Abstract

We introduce and explore a new multimodal input representation for vision–language models: acoustic field video. Unlike conventional video (RGB with stereo/mono audio), our video stream provides a spatially grounded visualization of sound intensity across a scene, offering a new and powerful dimension of perceptual understanding. Our real-time pipeline uses low-cost beamforming microphone arrays — already common in smart speakers and increasingly present in robotics and XR headsets — yet this sensing capability remains unutilized for scene understanding. To assess the value of spatial acoustic information, we constructed an evaluation set of 402 question–answer scenes, comparing a SOTA VLM given conventional video with and without paired acoustic field video. Results show a clear and consistent improvement when incorporating spa-

tial acoustic data; the VLM we test jumps from 38.3% correct to 67.4%. Our findings highlight that many everyday scene understanding tasks remain unconstrained when relying solely on visual and audio input, and that acoustic field data provides a promising and practical direction for multimodal reasoning. A video demo is available at <https://daehwakim.com/seeingsound>

## 1. Introduction

Vision–language models (VLMs) demonstrate remarkable performance across a wide range of multimodal reasoning tasks, from visual question answering to embodied dialogue. However, despite their growing sophistication, most VLMs remain grounded in just two sensory modalities: visual imagery (RGB video) and non-spatial audio (typically stereo streams). This leaves a critical gap in perceptual un-

derstanding: the ability to reason about where, and thus from what objects, sounds are being generated within a scene.

Sounds in natural environments are often byproducts of physical processes and human activity: engines hum, electronics beep, fans blow, doors creak, and kitchen appliances whir. These distributed auditory cues accompany specific object states, motions, or interactions, providing valuable implicit information about the world. For example, a boiling kettle indicates temperature and phase change; a whirring fan reveals mechanical activity; muffled speech implies distance or occlusion. Sound, therefore, is not merely an auxiliary sensory channel, it is a diagnostic signal of a scene’s underlying dynamics. Localizing these acoustic signatures provides a powerful complementary signal to visual data, particularly for reasoning about state, activity, and causality.

For this reason, spatial hearing plays a central role in human perception, enabling us to locate sound sources, resolve ambiguities in complex environments, and integrate auditory cues with vision to infer scene structure. While recent multimodal research has incorporated audio spectrograms or raw waveforms, these representations encode frequency and time but lack explicit spatial grounding. As a result, current models struggle with questions requiring the linkage between sound and location. This is especially true for generic or repeated sounds, such as motors humming, fans blowing, or electronics beeping, which could originate from many objects within a scene. In such cases, audio alone cannot resolve the source (see Figures 1, 2 and 4).

To address this gap, we introduce a new input modality for multimodal understanding: acoustic field video. Each frame encodes the spatial distribution of acoustic energy across a scene, effectively visualizing where sounds occur. Our real-time pipeline generates this stream using a microphone array and well-established beamforming algorithms, producing acoustic maps that are spatially and temporally aligned with RGB video frames and conventional audio. This alignment yields a rich, multimodal representation that unifies visual appearance, auditory content, and the spatial structure of scene sounds. While acoustic fields are not new (often referred to as sound pressure level (SPL) maps or related terms in acoustics), this is the first work to employ them as direct input to VLMs and to evaluate their utility for scene understanding and reasoning.

Our implementation requires low-cost microphone arrays, hardware already common in smart speakers and increasingly in domains such as robotics and XR headsets. Compared to other extended sensory modalities explored for use with VLMs — most notably thermal imagery [9, 13] — microphone arrays built from commodity MEMS components are inexpensive, small, and power efficient (e.g., the ICS-41350 MEMS microphone costs under \$0.75 in

volume, measures 3.5×2.65×0.98 mm, and consumes just 185  $\mu$ A in its always-on mode). This allows for microphone arrays to be integrated into even highly-constrained worn devices, such as Meta’s new Ray-Ban Display Glasses, which contain a six-microphone array.

To demonstrate the potential of this approach, we constructed a diagnostic benchmark of 402 question–answer (QA) pairs that span domestic, commercial, and industrial scenes. We open source both this dataset and our software pipeline. For evaluation, we use Gemini 2.5 Pro, a representative state-of-the-art VLM, to compare QA performance using traditional RGB+sound inputs against the same data augmented with acoustic field video. The addition of this modality yields substantive improvements (from 38.3% to 67.4% accuracy), particularly for questions involving localization, spatial attribution, and multi-source reasoning—tasks often unsolvable from RGB and stereo audio alone.

## 2. Related Work

### 2.1. Multimodal and Vision-Language Models

Recent years have seen rapid progress in large-scale vision–language models (VLMs) that jointly process vision and text. Models such as Flamingo [2], Kosmos-2 [31], LLaVA [20], and Gemini learn cross-modal representations that support image captioning, visual QA, and grounded dialogue without task-specific supervision. Video-centric VLMs further extend this to temporal reasoning and audio for open-ended video QA [7, 29], and AVQA systems that incorporate mono-channel audio [15, 19, 37] further highlight the value of combining sound and vision for scene understanding.

While these models achieve impressive performance, they remain primarily bimodal, relying on static RGB images or video streams and text. Audio, when included, is typically represented as 1D waveform embeddings or spectrograms that encode temporal–frequency structure but lack explicit spatial grounding. Consequently, even advanced VLMs are limited in their ability to infer where in a scene a sound originates, a key factor in human spatial understanding and embodied perception. Our work complements this line of research by introducing a spatially grounded acoustic representation that can be seamlessly integrated into VLM pipelines, extending their perceptual range beyond visual and textual modalities.

### 2.2. Modalities Beyond RGB Video & Audio

A parallel line of research explores how additional modalities—such as mono/stereo audio, depth, thermal, and motion—can complement visual and linguistic cues. AudioSet [8] and its derivatives provide large-scale audio–visual corpora for cross-modal pretraining, while mod-

els such as SoundSpaces [5], ImageBind [9], LanguageBind [40], and AVE [36] explore joint embeddings between text/vision to other modalities for localization, retrieval, and semantic understanding. These efforts have shown that sound enriches scene understanding by providing temporal continuity, semantic cues, and motion-related context not easily captured by vision alone.

However, most existing audio–visual–language models treat sound as a non-spatial signal—typically as mel-spectrograms or learned audio embeddings aggregated over time. This approach captures what is heard, but not where it occurs. As a result, models struggle with questions or reasoning tasks that depend on spatial localization, multi-source attribution, or occlusion relationships. In contrast, our proposed acoustic field (AF) video introduces an explicitly spatial acoustic modality, derived from beamformed microphone arrays, that visualizes sound energy across a scene. This enables multimodal models to reason about both the identity and spatial distribution of sound sources, bridging the gap between existing audio–visual fusion approaches and fully spatialized multimodal perception.

Methods such as Progressive Spatio-Temporal Perception for Audio-Visual Question Answering [19] model temporal dynamics and cross-modal correlations to improve AVQA performance. Bridging Audio and Vision [15], explores self-supervised methods to associate sounds with their visual sources in video. While these approaches successfully link objects and their auditory signatures, they treat audio as a non-spatial signal, limiting the ability to reason about precise sound locations or objects. As a result, if there are multiple objects in the scene of the same type, there can be confusion. Likewise, if a sound is generic, such as a motor hum, it may not be possible to guess which device is producing the sound in a complex scene such as a kitchen or workshop. The main example offered in [15] is two clarinet players, which start at different times, and thus can rely on visuo-temporal data for disambiguation. But this crucial visual information is not readily available in a wide range of scenes, especially non-human scenes. When actions are not visible, such as a computer fan running at max speed in a server rack, or a leaking faucet or toilet, there will likely be confusion.

In contrast, our approach leverages real-world live sensor data to derive spatial acoustic maps that explicitly encode the spatial distribution of acoustic energy, enabling multimodal models to disambiguate. To our knowledge, no prior work has explored using acoustic field video for audio-visual question answering.

## 2.3. Audio Scene Understanding

Beyond multimodal VLMs, a large body of work studies audio scene understanding without video. Audio-only approaches tackle acoustic scene classification, event detec-

tion, and audio-driven video understanding, including predicting visual motion or learning visual features from sound in a self-supervised way [1, 3, 4, 10, 17, 39]. Audio QA benchmarks such as AVQA [18] further require models to answer semantic questions about sound events and environments from spectrograms or waveforms. These methods show that sound alone carries rich information about activities, materials, and object states, but they operate in non-spatial feature spaces and are not designed to interface directly with general-purpose VLMs.

When video is available, most work emphasizes sound source localization and separation using vision. Early audio-visual correspondence methods such as “Look, Listen and Learn” [3] and “Sound Source Localization in the Wild” [22] localize sounding objects from unconstrained video, and later approaches refine this with contrastive learning and self-supervision for robust localization [6, 16, 21, 21, 27, 28, 30, 34, 35, 35, 38]. These pipelines typically output a heatmap or mask used for detection, retrieval, or sound isolation, rather than as a primary input to a reasoning system. In contrast, we derive real-time acoustic field (AF) video from beamformed microphone arrays and feed this spatially explicit sound-intensity map directly into an off-the-shelf VLM alongside RGB video, allowing us to study how explicitly spatialized acoustics improve zero-shot scene understanding and question answering rather than localization or separation alone.

## 3. Implementation

We now describe the main components of our system, including both hardware and software. Our pipeline is made open source at <https://www.github.com/anonymized-for-review>.

### 3.1. Hardware

We note that many contemporary devices already contain microphone arrays. For instance, the Apple HomePod includes a six-microphone array [14]), the Azure Kinect DK camera contains a seven-microphone array [25], and Meta’s recently announced Ray-Ban Display XR glasses have a six-microphone array [24]. Of course, many research systems have experimented with microphone arrays, from robots [23] to smartwatches [12]. Likewise, we envision our approach being used on multiple different platforms, including mobile robots, smart environment infrastructure, and worn devices with ego-centric views (see examples in Figure 1 and Section 4).

As a proof of concept, we use an off-the-shelf 16-channel MiniDSP UMA-16 v2 microphone array [26], seen in our Video Figure. The array measures  $132 \times 202 \times 18$  mm and enumerates as a multichannel microphone under the standard USB Audio Class (UAC) protocol. To this hardware, we add a USB webcam at the center with a  $72^\circ$  diagonal

field of view. Both the microphone array and webcam connect to a 2024 MacBook Air M3, on which all processing (other than the VLM) occurs in real time. In Section 3.6, we breakdown sources of latency.

### 3.2. Conventional RGB Video and Audio Streams

For RGB video, we use frames streamed from the USB webcam, downscaled to 640×360. For audio, we create a stereo audio stream (44.1 kHz) using the upper-left and upper-right microphones on our array (12.6 cm apart).

### 3.3. Acoustic Field Video Software Pipeline

Our software pipeline is agnostic to the microphone element count and geometry of the array. The latter is defined in an XML file read by our software at runtime. The sixteen-channel microphone array we used implemented the UAC protocol, allowing for plug-and-play operation. Audio was sampled at 44.1 kHz, with a chunk size of 2048, and thus our pipeline runs at ~22 FPS (we discuss latency in Section 3.6).

For acoustic beamforming, we utilize the Multiple Signal Classification (MUSIC) algorithm [33] as implemented in *Acoular* Python [32] (1,024-point FFT, Hann window, 50% overlap). As a frequency-domain method, MUSIC requires a discrete set of frequencies for analysis. We therefore selected four center frequencies (2, 4, 6, and 8 kHz) that span a broad portion of the acoustic spectrum relevant to human activities and everyday environments. To compensate for frequency-dependent gain differences and suppress background noise, we first subtract different noise floor thresholds (18, 20, 23, and 27 dB) and then clip each SPL map to a narrow top dynamic range (0.2, 0.2, 0.5, and 0.5 dB below the maximum observed value) for our 2, 4, 6, and 8 kHz SPL maps, respectively. We then average the four resulting maps to produce a single composite acoustic field. To stabilize temporal fluctuations, we apply an eight-frame median filter (approx. 370 ms). This resulting map is rendered with a jet colormap, alpha-blended onto a gray-scaled copy of the video frame. The latter is streamed as our acoustic field video for downstream vision-language modeling, with a resolution of 640×360, matching our RGB video stream.

### 3.4. Zero-shot Scene Understanding

We use Gemini-2.5-Pro as a representative state-of-the-art multimodal understanding model. Beyond a language prompt, we use two input formats. First is RGB video + stereo audio, which we refer to as *Conventional Video*. Second, we have the RGB video + acoustic field video, arranged as a stacked pair, plus stereo audio, which we refer to as *Conventional + Acoustic Field Video*.

We interface with Gemini using the google-generativeai Python API [11]. We use the following prompt:

Be definitive in your answers. Avoid hedging words like "potentially", "possibly", or "probably", and other speculative language. Also, answer concisely; one or a few sentences at most. I am giving you a video clip with audio of a scene.

The video clip has two synchronized visualizations of the same camera view. The top is the standard video of the scene. Bottom is the same video, but overlaid with a sound pressure map (jet color scheme, e.g., blue is no or low sound, and oranges and reds are louder sound sources). The sound pressure map shows where sounds are coming from. Your answer should not explicitly mention the video or the sound pressure map.

Using this information, I want you to answer the following question:

Only the black portions of the prompt are used when inputting *Conventional Video* (for baseline evaluation purposes, discussed in Section 5), and the complete prompt (black and blue portions) is used when inputting *Conventional + Acoustic Field Video*.

### 3.5. Live Mode

We also take advantage of Gemini 2.5 Pro’s “Talk to Gemini Live” mode to ingest a live audio-video stream, which could come from a robot, AI glasses, etc. In our case, we pass Gemini our *Conventional + Acoustic Field Video*. We use the same prompt as the prior section, sans the very last sentence. In this way, it is open-ended for the user to define a desired assistive use. They could say, for instance, “*let me know if there are any hazards while I am working*”, or “*keep me informed about the status of my 3D print job*”. We offer some illustrative examples in our Video Figure.

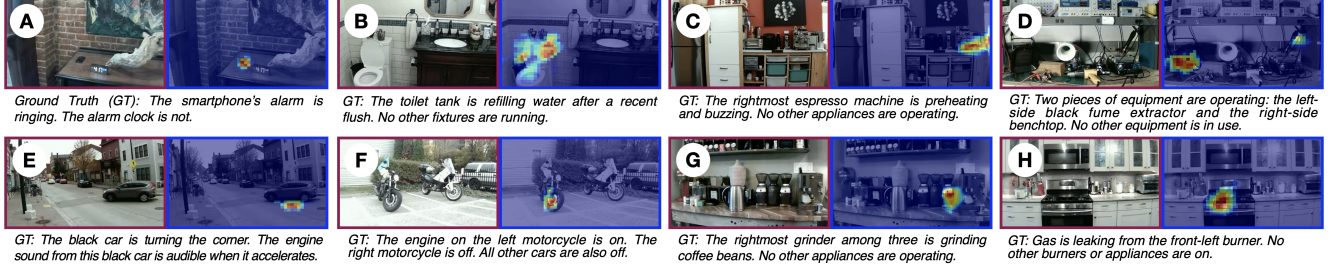
### 3.6. Latency

Our approach is comparatively lightweight with modest latency; VLM token generation dominates any interaction. There are two main sources of latency in our system: sensor latency and beamforming computation. Starting first with sensors: the USB Audio Class (UAC) protocol is highly optimized on modern operating systems, with just a few milliseconds of latency, while USB Video Class (UVC) devices have on the order of tens of milliseconds of latency. Our beamforming stage (with four runs of the MUSIC algorithm) takes 58 ms on our 2024 MacBook Air M3. These two main processes, plus various other smaller latencies (e.g., moving video/audio frames into Python, rendering the acoustic field onto live video for debugging) combine for a total latency of around 90 ms. We note our Python-based pipeline is not heavily optimized, and this reported latency should be considered an upper bound.

## 4. Potential Applications

We envision AI agents augmented with acoustic field sensing being valuable across a wide range of scenarios. Below we outline several representative use cases that illustrate the potential of this modality to enhance perception, interaction, and situational awareness. Please also see our Video





Prompt	Gemini with <b>Conventional Video</b>	Gemini with <b>Conventional + Acoustic Field Video</b>
<b>A</b> You are a mobile robot. When the user asks "please stop that" or "please check that", describe what action you would take.	Based on the audio and visual information, I will approach the table, identify the source of the beeping alarm, and press the appropriate button to silence it.	✗ I would go to the table and silence the alarm coming from the smartphone. ✓
<b>B</b> What is happening in the scene?	A bathroom exhaust fan is running.	✗ The toilet tank is filling with water. ✓
<b>C</b> What is the noise?	The noise is the humming sound made by the refrigerator's motor.	✗ The noise is coming from the espresso machine. ✓
<b>D</b> Describe this scene and the prominent objects in it, and try to assign actions or states to objects you believe are active.	This is an electronics workbench featuring several pieces of test equipment on a shelf, including multiple power supplies and a multimeter. On the bench itself are a soldering station, a yellow heat gun, a fume extractor, and various electronic components. A loud, continuous blowing or fan noise is audible throughout the clip, indicating that a device like the fume extractor or heat gun is actively running.	✗ In an electronics workshop, a small black fan on the left is running, and a soldering iron on the right is also active and producing a faint sound. ✓
<b>E</b> What is the noise?	The noise is the sound of cars driving on a city street.	✗ The noise is generated by the engine and tires of the dark SUV as it turns right at the intersection. ✓
<b>F</b> Which vehicle has been started?	Based on the audio, one of the motorcycles has been started.	✗ The black motorcycle on the left has been started. ✓
<b>G</b> Which grinder is making noise?	None of the three grinders are making noise. The sound is coming from the espresso machine on the far right, which is brewing coffee into the cup.	✗ The coffee grinder on the far right is making the noise. ✓
<b>H</b> What is the noise?	The noise is the electronic beeping of a kitchen appliance, such as a microwave or oven.	✗ A gas stove burner is hissing. ✓

Figure 2. Top: Example scenes (conventional RGB and acoustic field still frames shown) drawn from our test set. Bottom table: Example prompts, along with output from Gemini given either *Conventional Video* or *Conventional + Acoustic Field Video* as input.

Figure for additional examples using our live-mode implementation.

#### 4.1. Worn Devices (Ego-Centric View)

Worn devices such as AI pins and XR glasses have a unique ego-centric view onto the world. With acoustic field video, such systems could infer where sounds originate relative to the user, allowing them to better understand and monitor activities, interactions, and context. For example, AI glasses could proactively let users know “the kettle is boiling” (Figure 1 middle), “you left the stove on” (Figure 2 H), and “you should turn on the fume extractor while soldering” (Figure 4 I).

#### 4.2. Mobile Robots

For autonomous mobile robots, spatialized acoustic understanding enhances both perception and interaction. Robots equipped with microphone arrays can identify and localize sounds that indicate human presence, mechanical operation, or environmental changes. For instance, a household robot could infer that a laundry machine has finished its cycle, or that someone called its name from another room (Figure 1 bottom). In workplace settings, acoustic cues such as machine vibrations, dripping taps, alarms, can reveal operational states that are difficult to capture visually (Figure 1 & 2).

### 4.3. Smart Speakers & Smart Environments

Although fixed in position, smart speakers, security systems, and other ambient IoT infrastructure occupy ideal vantage points for continuous monitoring of activity in homes and workplaces. When augmented with vision and acoustic field sensing, these systems could move beyond simple sound detection (e.g., breaking glass alert) to spatial reasoning (e.g., user accidentally dropped a glass). For example, a smart speaker could localize a crying infant, detect if a stove burner was left on (Figure 2 H), track the usage of appliances (Figure 1 middle), monitor water consumption (Figure 2 B), or identify noise from a kitchen appliance (Figure 4 K). In office or factory environments, acoustic fields could integrate with building systems such as HVAC or lighting to adapt to occupancy and activity levels. Moreover, these maps could inform nearby robots or embodied agents, allowing a cooperative ecosystem where devices share a common, spatially grounded understanding of sound events in their surroundings.

## 5. Evaluation

### 5.1. QA Scene Data Collection

There is no public data set that contains paired conventional video and acoustic field video. It may be possible to partially simulate such data using audiovisual segmenta-

tion [34, 35], however results can be inconsistent if there are multiple objects of the same class in the scene or if sounds are generic (hums, vibrations, whirs, electronic beeping, etc.) and could be attributed to more than one present object. For this reason, we do not believe there is a suitable substitute for real-world data at this time, and as such, we created our own dataset for evaluation, which we also make freely available for replication and advancement.

All recordings were captured using the hardware and software described in Section 3. One capture instance consists of a synchronized, five-second triplet of data: conventional (RGB) video, acoustic field video, and stereo audio. Data collection was conducted in ten diverse environments: bathroom, bedroom, kitchen, office, office kitchenette, utility room, electronics room, fabrication workshop, parking lot, and a road. We did not control for background noise, and so most of our data contains, e.g., HVAC noise and background chatter.

We endeavored to curate a diverse range of questions that would be applicable to different use cases (see Section 4). For all instances captured, we asked three common questions: First, simulating a mobile robot use case, we asked: “When the user says ‘please stop that’ or ‘please check that’, what action would you take?”. Targeting more generic scene understanding, we also asked: “What is happening in the scene?”, and “What is the noise?”. In addition to these three common questions, we also included at least one (max five) custom scene-specific queries, such as in the office kitchenette: “What stage of coffee making are we in?”. In total, our evaluation set contains 402 QA instances.

## 5.2. Procedure

To test whether adding acoustic field video improves zero-shot scene understanding, we compare two input modalities: *Conventional Video* (i.e., RGB video + stereo audio; baseline) vs. *Conventional + Acoustic Field Video* (ours). The task is to answer a question for a given five-second question-scene instance. We use the same model, Gemini 2.5 Pro, for both conditions (see prompt in Section 3.4). To prevent context carryover and standardize quality, we initialize a new inference session for each instance. As we have 402 instances, tested with two input modalities, our procedure generated 804 QA instance pairs. Including uploading data to Gemini, this automated process took around 3 hours.

## 5.3. Human Raters

To evaluate the quality of VLM output, we recruited three human raters (mean age = 20.3; two identified as women). We developed a basic web interface to facilitating process all of the data. For each QA instance, raters (wearing headphones) watched the five second conventional (RGB) video

clip including stereo audio, along with a short ground-truth text description and question about the scene. Raters could re-play the video as they saw fit to understand each scene and question being asked.

Once satisfied they understood the scene, the raters proceeded to evaluate the correctness of a VLM-generated answer. Two buttons were offered: “correct” and “not exactly”. The raters were told during orientation that an answer “should only be marked ‘correct’ if it matches the ground truth without contradictions, omissions, or vague language,” while answers with “with wrong or vague responses should be marked as ‘not exactly’.” Upon selecting one of these options, they were shown a second VLM-generated answer, and again told to assess correctness. The presentation order of the two input modality conditions was randomized and counterbalanced. Lastly, for the same QA instance, raters were shown both answers side-by-side and asked: “Which do you believe is the better answer?”. They could select the first answer, second answer, or equal preference. For final labels, we took the raters’ majority vote; i.e., for an answer to be deemed correct, incorrect, or preferred, two of three or three of three raters had to rate it as such. There is a potential corner case with this methodology, which is a three-way tie between preference for answer A, answer B, and equal preference, but this never occurred in our data.

The raters completed this task independently, which took around 4 hours to complete. In total, our raters provided 3618 responses (402 QA instances  $\times$  3 ratings  $\times$  raters). We see substantial agreement on correctness for both input modalities: Fleiss’  $\kappa = 0.72$  for *Conventional Video* and  $\kappa = 0.65$  for *Conventional + Acoustic Field Video*. When our raters had a preference between the two answers, we see a similarly high inter-rater reliability, with a Krippendorff’s  $\alpha$  of 0.78.

## 6. Results & Discussion

We now present our main findings, also summarized in Figure 3, before moving to the discussion.

### 6.1. QA Scene Accuracy

Across all 402 QA instances, the baseline *Conventional Video* input condition was rated as 38.3% correct, while our *Conventional + Acoustic Field Video* input condition was rated as 67.4% correct. The magnitude of this increase strongly suggests that acoustic field data materially helps the VLM disambiguate sound sources and relate audio evidence to the visual scene.

Breaking down this data a different way, we see that in 24.4% of QA instances, the raters believed both VLM input conditions were incorrect, and 30.1% of the time raters believed both input conditions were correct. More interesting is when answers derived from *Conventional Video*

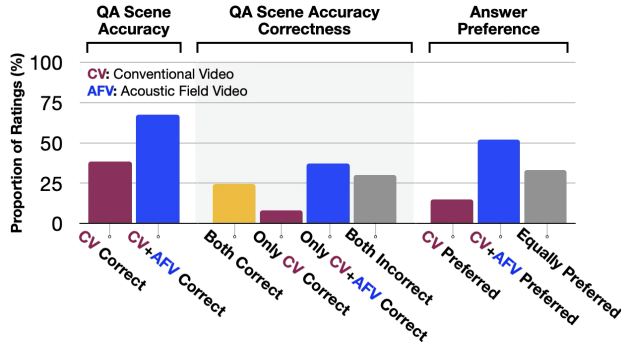


Figure 3. QA scene understanding with and without acoustic field video. Left: Overall answer accuracy when the VLM sees only *Conventional Video* (CV) versus *Conventional + Acoustic Field Video* (CV+AFV); Middle: Breakdown of correctness. Right: Human raters’ answer preferences.

were rated as correct, while answers from the *Conventional + Acoustic Field Video* were rated as incorrect (8.4% of QA instances). However, a much larger proportion shows the opposite behavior, where answers from the VLM given *Conventional + Acoustic Field Video* were rated as correct while the *Conventional Video* input was wrong (37.3%).

## 6.2. Answer Preference

When comparing answers from the two VLM input conditions side-by-side, answers derived from *Conventional Video* input were preferred in 14.9% of the time, while answers from *Conventional + Acoustic Field Video* input were preferred 52.0% of the time (with the remaining 33.1% being judged as equally valid).

## 6.3. Improved Attention

We observed that when an acoustic field video is provided, the VLMs attention improved, leading to both more correct and more succinct answers (see examples in Figure 4, I-M). The VLM also tended to elevate the active object to the beginning of its reply. Lastly, we also saw the absence of sound in the acoustic field being utilized. For example, as can be seen in Figure 4 I, where a user is soldering without an exhaust fan running, the VLM is able to reason that the fan is off using the acoustic field.

## 6.4. Failure Cases

We also noted some interesting failure cases, a few of which we highlight in Figure 4, N-P. As reported in Section 6.1, only in 8.2% of the QA instances did the VLM get the correct answer using *Conventional Video*, but got the wrong answer using *Conventional + Acoustic Field Video*. We analyzed these 33 failure instances for high-level themes. We found occasional misclassifications of sounds — for instance, the humming of a running microwave was described

as “beeping” and “finishing”. We also observed instances of misattributions of sound — for example, at the beginning of one clip, a parked car honks its horn which is clearly visible in the acoustic field video; however, another car drives through the scene in the second half of the clip (which is very salient) and the honk is incorrectly attributed to the moving car. Finally, we saw in some clips with quiet but audible background noise (especially from HVAC) that the model sometimes attributed activity to inactive objects in the scene, even without a cue present in the acoustic field (e.g., an inactive 3D printer sitting on a table). It may be that prompt engineering could resolve some of these issues.





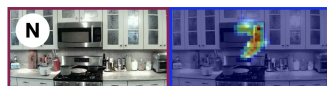


## 7. Limitations & Future Work

While our results demonstrate clear benefits of incorporating acoustic field video into multimodal reasoning, there are several limitations of note. First, our evaluation dataset, although diverse in scene type and acoustic configuration, remains modest in size and limited to real-world recordings from a single array geometry. Broader datasets spanning varied microphone arrangements, room impulse characteristics, more outdoor environments, and cluttered multi-source scenes will be important to generalize our initial findings.

We also note that our implementation uses frequency-domain MUSIC beamforming with a small set of discrete analysis frequencies. Although this approach offers high angular resolution and robustness, it imposes computational overhead and a fixed spatial grid that may limit responsiveness in dynamic scenes. Moreover, MUSIC relies on narrowband assumptions and is known to degrade in low-SNR settings, highly reverberant environments, or when sound sources exhibit broad spectral content not well captured by our chosen frequency bands. Future work could explore alternative wideband beamforming techniques, neural beamformers, learned spatial acoustic encoders, or hybrid classical–neural architectures that jointly optimize spatial and semantic representations. Nonetheless, MUSIC was sufficient for our proof-of-concept implementation (i.e., the goal of this work was not to make advances in beamforming, but rather explore the potential utility of the signal). More advanced methods would likely unlock further accuracy gains against conventional video inputs.

Additionally, while acoustic field video provides explicit spatial grounding, VLMs ingest it only through conventional visual pathways. As a result, the model performs no audio–acoustic fusion at the level of raw spatial features; instead, spatial sound structure must be inferred from a colorized overlay. Training VLMs (or dedicated multimodal encoders) to directly process acoustic fields in a native tensor format may unlock significant performance gains and reduce reliance on handcrafted visualization choices (e.g., colormaps, clipping thresholds). Similarly, joint training



			
<p>GT: The white fume extractor beside the person soldering is off. Also, no other equipment is on.</p>	<p>GT: The smartphone's alarm is ringing. The alarm clock is not.</p>	<p>GT: The fridge door isn't fully closed, so its alarm is sounding. No other appliances are operating or sounding an alarm.</p>	<p>GT: The printer is currently printing.</p>
			
<p>GT: The rightmost espresso machine is preheating and buzzing. No other appliances are operating. The only active machine is the espresso machine.</p>	<p>GT: The microwave is running. No other appliances are on.</p>	<p>GT: The orange car, second from the right between the red and black cars, is honking.</p>	<p>GT: The printer is off. The AC is on with noise from the left-side ceiling.</p>

Prompt	Gemini with <b>Conventional Video</b>	Gemini with <b>Conventional + Acoustic Field Video</b>
<b>I</b> Is the white fume extractor on?	Yes, the sound of its fan running can be heard in the audio. ❌	The white fume extractor is not on. ✅
<b>J</b> Describe this scene and the prominent objects in it, and try to assign actions or states to objects you believe are active.	In a room with an exposed brick wall, a digital alarm clock on a wooden table is actively sounding its alarm. The clock displays a temperature of -4°F. Also on the table are a smartphone and a large piece of driftwood. An abstract painting hangs on the wall above. ❌	A smartphone and a digital clock are on a wooden table in front of a brick wall. The smartphone's alarm is ringing. ✅
<b>K</b> Describe this scene and the prominent objects in it, and try to assign actions or states to objects you believe are active.	This video shows a modern kitchen with white cabinets, granite countertops, and stainless steel appliances. The prominent objects are a French door refrigerator, a microwave, and a gas stove. An appliance is actively beeping, creating a repetitive alarm sound. ❌	In a kitchen with white cabinets and stainless steel appliances, a refrigerator is beeping as a door-ajar alarm. ✅
<b>L</b> Describe this scene and the prominent objects in it, and try to assign actions or states to objects you believe are active.	This video shows the interior of an office or workshop. Prominent objects include a large multifunction printer on the right, a blue recycling bin, a box of paper, and a shelving unit with storage bins. The printer is active, making a mechanical whirring sound as it operates. ✅	In an office or workshop, a large multifunction printer on the right is operating. ✅
<b>M</b> You are a mobile robot. When the user asks "please stop that" or "please check that", describe what action you would take.	I would investigate the source of the humming noise coming from the kitchen appliances. ❌	I will go to the espresso machine and turn it off. ✅
<b>N</b> Has the microwave finished?	No, the microwave has not made a sound to indicate that it has finished. ✅	Yes, the microwave is beeping, indicating it has finished. ❌
<b>O</b> You are a mobile robot. When the user asks "please stop that" or "please check that", describe what action you would take.	Based on the video, I would identify the orange SUV as the source of the horn honking. ✅	I would investigate the black SUV parked on the right, as its horn is honking. ❌
<b>P</b> Is the printer printing?	No, the printer is not printing. There are no sounds of a printing mechanism operating, nor is any paper moving through the machine. ✅	No, the printer is not printing. The sound is coming from the equipment on the left. ❌

Figure 4. Examples of good attention and failure cases (separated by black horizontal rule). Top: Example scenes (*Conventional + Acoustic Field Video*) drawn from our test set. Bottom table: Example prompts, along with output from Gemini given either *Conventional Video* or *Conventional + Acoustic Field Video* as input.

on synchronized RGB, audio, and acoustic field data could yield richer cross-modal bindings and more consistent attributions of sounds to objects.

Finally, although we evaluate zero-shot performance, in-the-wild deployment introduces additional challenges: user motion with wearable devices, array self-noise, user-motion generated noise, and device heterogeneity within and across manufacturers. Understanding the sensitivity of acoustic-field-driven reasoning to these factors, and developing calibration-free or self-adaptive techniques, remains an open question for deployment.

## 8. Conclusion

We presented acoustic field video as a new spatially grounded acoustic modality for multimodal vision-language reasoning. By visualizing where sound originates and aligning this information with RGB video and audio, acoustic field video enables VLMs to connect auditory evidence to specific objects and locations in a scene. Using a real-time, low-cost beamforming pipeline and a 402 QA-instance benchmark, we show that augmenting conventional RGB+audio inputs with acoustic field video yields

substantial gains in zero-shot scene understanding and is consistently preferred by human raters. As microphone arrays continue to proliferate across smart speakers, robots, and wearable devices, our results highlight acoustic field video as a practical and powerful path toward more perceptually grounded multimodal intelligence.

## References

- [1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision – ECCV 2020*, pages 208–224, Cham, 2020. Springer International Publishing. 3
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736. Curran Associates, Inc., 2022. 2
- [3] Relja Arandjelovic and Andrew Zisserman. Look, listen and



- learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3
- [4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 3
- [5] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vincenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision – ECCV 2020*, pages 17–36, Cham, 2020. Springer International Publishing. 3
- [6] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16867–16876, 2021. 3
- [7] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-lms, 2024. 2
- [8] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 2
- [9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 2, 3
- [10] Laurent Girin, Sharon Gannot, and Xiaofei Li. Chapter 3 - audio source separation into the wild. In *Multimodal Behavior Analysis in the Wild*, pages 53–78. Academic Press, 2019. 3
- [11] Google. google-generativeai. <https://pypi.org/project/google-generativeai/>, 2025. Python package, version <x.y.z>. Accessed: 2025-11-11. 4
- [12] Teng Han, Khalad Hasan, Keisuke Nakamura, Randy Gomez, and Pourang Irani. Soundcraft: Enabling spatial interactions on smartwatches using hand generated acoustics. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, page 579–591, New York, NY, USA, 2017. Association for Computing Machinery. 3
- [13] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
- [14] iFixit. Homepod teardown, 2018. Accessed: 2025-11-14. 3
- [15] Seung jae Lee and Paul Hongsuck Seo. Bridging audio and vision: Zero-shot audiovisual segmentation by connecting pretrained models, 2025. 2, 3
- [16] Yanhao Jia, Ji Xie, S Jivaganesh, Hao Li, Xu Wu, and Mengmi Zhang. Seeing sound, hearing sight: Uncovering modality bias and conflict of ai models in sound localization, 2025. 3
- [17] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, page 213–224, New York, NY, USA, 2018. Association for Computing Machinery. 3
- [18] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19108–19118, 2022. 3
- [19] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 7808–7816, New York, NY, USA, 2023. Association for Computing Machinery. 2, 3
- [20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 2
- [21] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 3742–3753, New York, NY, USA, 2022. Association for Computing Machinery. 3
- [22] Xian Liu, Rui Qian, Hang Zhou, Di Hu, Weiyao Lin, Ziwei Liu, Bolei Zhou, and Xiaowei Zhou. Visual sound localization in the wild by cross-modal interference erasing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):1801–1809, 2022. 3
- [23] Alejandro Luzanto, Nicolás Bohmer, Rodrigo Mahu, Eduardo Alvarado, Richard M. Stern, and Néstor Becerra Yoma. Effective acoustic model-based beamforming training for static and dynamic hri applications. *Sensors*, 24(20), 2024. 3
- [24] Meta Platforms, Inc. Meta ray-ban display glasses and meta neural band, 2025. Accessed: 2025-11-14. 3
- [25] Microsoft Corporation. Azure kinect dk, 2020. Accessed: 2025-11-14. 3
- [26] miniDSP. Uma-16 v2 usb mic array, 2018. Accessed: 2025-11-14. 3
- [27] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Advances in Neural Information Processing Systems*, pages 37524–37536. Curran Associates, Inc., 2022. 3
- [28] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Computer Vision – ECCV 2022*, pages 218–234, Cham, 2022. Springer Nature Switzerland. 3
- [29] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models, 2023. 2

- [30] Sooyoung Park, Arda Senocak, and Joon Son Chung. Marginnc: Robust sound localization with a negative margin. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. [3](#)
- [31] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. [2](#)
- [32] E Sarraj, G Herold, A Kujawski, S Jekosch, AJR Pelling, M Czuchaj, T Gensch, S Oertwig, J Tschavoll, J Ostendorf, et al. Acoular–acoustic testing and source mapping software. *Version v25*, 3, 2024. [4](#)
- [33] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986. [4](#)
- [34] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Aligning sight and sound: Advanced sound source localization through audio-visual alignment, 2024. [3](#), [6](#)
- [35] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6420–6429, 2023. [3](#), [6](#)
- [36] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. [3](#)
- [37] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *Computer Vision – ECCV 2024*, pages 146–164, Cham, 2025. Springer Nature Switzerland. [2](#)
- [38] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. [3](#)
- [39] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [3](#)
- [40] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2024. [3](#)