Bat: Learning to Reason about Spatial Sounds with Large Language Models

Zhisheng Zheng [1 2] Puyuan Peng [1] Ziyang Ma [2] Xie Chen [2] Eunsol Choi [1] David Harwath [1]

Abstract

Spatial sound reasoning is a fundamental human skill, enabling us to navigate and interpret our surroundings based on sound. In this paper we present Bat, which combines the spatial sound perception ability of a binaural acoustic scene analysis model with the natural language reasoning capabilities of a large language model (LLM) to replicate this innate ability. To address the lack of existing datasets of in-the-wild spatial sounds, we synthesized a binaural audio dataset using AudioSet and SoundSpaces 2.0. Next, we developed SpatialSoundQA, a spatial sound-based question-answering dataset, offering a range of QA tasks that train Bat in various aspects of spatial sound perception and reasoning. The acoustic front end encoder of Bat is a novel spatial audio encoder named Spatial Audio Spectrogram Transformer, or Spatial-AST, which by itself achieves strong performance across sound event detection, spatial localization, and distance estimation. By integrating Spatial-AST with LLaMA-2 7B model, Bat transcends standard Sound Event Localization and Detection (SELD) tasks, enabling the model to reason about the relationships between the sounds in its environment. Our experiments demonstrate Bat's superior performance on both spatial sound perception and reasoning, showcasing the immense potential of LLMs in navigating and interpreting complex spatial audio environments. Our demo, dataset, code and model weights are available at: https://zhishengzheng.com/BAT.

1. Introduction

The evolution of Large Language Models (LLMs) has enabled their application beyond textual data, giving rise to multimodal language models capable of performing image understanding tasks such as visual question answering and image captioning (Li et al., 2023; Liu et al., 2023; OpenAI, 2023; Peng et al., 2024). Other work has integrated sound perception ability into LLMs, enabling them to perform tasks such as audio question answering, speech recognition, speech translation, and speech synthesis (Deshmukh et al., 2023; Wang et al., 2023a; Chu et al., 2023; Borsos et al., 2023; Gong et al., 2024). However, despite these LLM-based models' ability to handle numerous visual and audio tasks with impressive performance, so far none of them can handle in-the-wild spatial audio input. Humans are equipped with binaural hearing ability, allowing us to not only identify the type of sound we are hearing, but also infer how far away the sound is, what direction it is coming from, and whether multiple sound sources at different locations are present at once. Consider scenarios such as locating someone calling from upstairs, hearing your phone ringing from behind a sofa, or sensing footsteps approaching from behind; these situations all demand spatial audio perception and reasoning, a capability not yet present in the aforementioned models.

To bridge this gap, we introduce Bat, the first spatial audio-based LLM designed to reason about sounds in a 3-D environment. Recognizing the current lack of large-scale datasets for in-the-wild spatial audio and spatial audio-based question answering, we synthesized a large-scale binaural audio dataset using Audioset (Gemmeke et al., 2017) clips as sound sources and Soundspaces 2.0 (Chen et al., 2022) for simulating diverse, 3-D, reverberant acoustic environments. In conjunction with this dataset, we developed SpatialSoundQA, a diverse collection of question answering tasks designed to train and evaluate spatial sound understanding across varying levels of complexity.

The successful training and evaluation of Bat hinge on the ability to encode rich spatial audio information accurately. However, existing encoders often specialize in either monaural event detection (Baade et al., 2022; Huang et al., 2022; Chen et al., 2023; 2024b) or are limited to first-order ambisonics (Shimada et al., 2021; Wang et al., 2023b; Kim

[1] Department of Computer Science, University of Texas at Austin, USA [2] Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. Correspondence to: Zhisheng Zheng <zszheng@utexas.edu>, David Harwath <harwath@utexas.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

et al., 2023), focusing on a limited range of sound events. Other approaches are solely dedicated to direction-of-arrival (DoA) estimation in specific multi-channel formats (Yang et al., 2022a; Wang et al., 2023c), lacking the breadth needed for comprehensive spatial audio understanding tasks. To overcome these limitations, we developed SPATIAL-AST, a novel multi-task spatial encoder that is not only capable of sound event detection and spatial localization but also excels in distance perception, thereby providing a comprehensive solution for spatial audio tasks.

In our experiments, we found that for SPATIAL-AST is both effective and parameter-efficient. Using multi-step training curriculum, the SPATIAL-AST encoder by itself can achieve strong performance across multiple tasks: a mean Average Precision (mAP) of 50.03% for audio event classification, a Mean Angular Error (MAE) of 17.94° for direction of arrival estimation, and a Distance Error Rate (DER) within 0.5 meters of the actual location at 32.54% for distance estimation. By fusing the SPATIAL-AST encoder to the LLaMA-2 (Touvron et al., 2023b) and adopting an appropriate training curriculum, our BAT not only excels in tasks involving perception but also demonstrates spatial reasoning abilities in scenarios with mixed sound sources (e.g. "Is the stereo system to the left of the barking dog?"), achieving a 76.89% accuracy in answering spatial sound reasoning questions.

Our key contributions are summarized as follows:

- We present the first spatial audio-based question answering dataset SPATIALSOUNDQA, offering a range of 3-D audio understanding tasks from perception to reasoning, which provides a platform for training models to perform spatial sound understanding tasks.
- We propose SPATIAL-AST, a binaural spatial audio encoder architecture that jointly performs sound event detection, spatial localization, and distance estimation that achieves strong performance across all three tasks.
- We introduce BAT, which integrates SPATIAL-AST with the LLaMA-2 LLM, resulting in a model capable of answering complex reasoning questions about multiple sound sources situated within a 3-D environment.

2. Related Work

2.1. Spatial Audio

Spatial audio is a family of techniques used to create the illusion of sound sources existing within a 3-D space around the listener. Encapsulating traditional stereo and surround sound systems to more recent methods such as ambisonic audio, spatial audio is used in applications ranging from virtual reality (Mystakidis, 2022) to advanced theater systems. In the realm of artificial intelligence and machine learning, spatial audio presents unique challenges for intelligent agents situated in 3-D physical spaces in accurately localizing and interpreting sound sources (Evers et al., 2020; Guizzo et al., 2022). To address these challenges, researchers have developed acoustic simulation techniques (Scheibler et al., 2018; Chen et al., 2022) and algorithms that leverage spatial audio (Yang et al., 2022b; Wang et al., 2023c). Existing spatial audio datasets, such as YouTube-360 (Morgado et al., 2020), YouTube-ASMR (Yang et al., 2020), Pano-AVQA (Yun et al., 2021), and STARSS23 (Shimada et al., 2023), offer varying levels of spatial audio information. However, they often suffer from inconsistent quality, lack crucial ground truth labels like sound source distance and direction, and are limited in scope, hindering the development of spatial audio-based machine learning models.

2.2. Sound Event Localization and Detection

The Sound Event Localization and Detection (SELD) task (Adavanne et al., 2018) merges sound source localization with sound event detection (SED), a task which the DCASE community[1] later incorporated into their challenges as Task 3. Subsequent works (Wang et al., 2020; Zhang et al., 2021; Wang et al., 2023b) adapted architectures like GRU (Cho et al., 2014), TDNNF (Povey et al., 2018), and Conformer (Gulati et al., 2020) for SELD tasks, achieving strong performance. However, these models have primarily focused on shallow spatial audio perception, and the potential of leveraging large language models to enable spatial reasoning remains unexplored in prior work, which is the main focus of our work.

2.3. Multimodal Large Language Models

Recent models (OpenAI, 2023; Li et al., 2023; Liu et al., 2023; Peng et al., 2024) have equipped LLMs with the ability to reason about and create images and videos. For the audio modality, AudioGPT (Huang et al., 2023) integrates ChatGPT as a versatile interface for a wide array of audio and speech applications. Pengi (Deshmukh et al., 2023) proposes the use of transfer learning to frame all audio tasks as text-generation challenges, utilizing both audio and text inputs to produce results without additional fine-tuning. LTU (Gong et al., 2024) proposes a monaural audio QA dataset, and combines an Audio Spectrogram Transformer (AST) (Gong et al., 2021) feature extractor with a LoRA-adapted (Hu et al., 2022) LLaMA (Touvron et al., 2023a) LLM to train the model to reason and answer questions about the sounds in a clip. Crucially, LTU differs from our work in that LTU is trained only on monaural sound data, whereas we consider multi-channel spatial sound rendered by a reverberant, 3-D environment. SALMONN (Tang et al.,

[1] https://dcase.community/community_info

Table 1. SPATIALSOUNDQA Overview: The first four types focus on perception, while the last emphasizes reasoning. DP: Distance Prediction; DoA: Direction-of-Arrival. Numbers (e.g., 139K, 15.9%) indicate the QA sample count and their percentages in the dataset.

Type	#Sources	Example
A: Detection (139K, 15.9%)	1	**Q:** Identify the sound events in the audio clip. / **A:** baby laughter; laughter; speech **Q:** What are the distinct sounds present in this audio clip? / **A:** heart sounds, heartbeat
B: DoA & DP (139K, 15.9%)	1	**Q:** How would you describe the location of this audio clip? / **A:** right, front, below; 2.5m **Q:** At what distance and in which direction, is the music's sound originating? / **A:** left, behind, below; 5m
C: Detection (118K, 13.5%)	2	**Q:** Identify the sound events in the audio clip coming from the right, front, below, approximately 3 meters away. / **A:** slosh; speech **Q:** What sound events can you detect in the audio recording emanating from the left, behind, above, roughly 0.5 meters away? / **A:** music; musical instrument; steelpan
D: DoA & DP (118K, 13.5%)	2	**Q:** In which direction and how far away is the source of the heart sounds, heartbeat's sound? **A:** left, behind, below; 1m **Q:** Where is the sound of the music coming from? / **A:** left, behind, below; 3m
E: Reasoning (358K, 41.2%)	2	**Q:** When measuring the direct line distance, is the sound produced by wheeze closer to you than the sound from bird flight, flapping wings? / **A:** No **Q:** Is the source of both explosion and speech's sounds from your left side? / **A:** Yes **Q:** Is the sound from the music on the front and the sound from the rattle on the behind? / **A:** No **Q:** Does the sound of electric shaver, electric razor appear on the behind of the sound of waterfall? / **A:** Yes **Q:** Can you estimate the distance from the sound of the speech to the sound of the dog? / **A:** 1.5m **Q:** What is the sound on the above side of the sound of the vibration? / **A:** croak; frog **Q:** Could you determine whether the singing's sound is to the left or right of the steam's sound? / **A:** left

2024) uses OpenAI's Whisper model (Radford et al., 2023) and BEATs (Chen et al., 2023) as dual auditory encoders to extract speech and audio representations, and then concatenates them to LLMs to enable the generation of responses. Qwen-audio (Chu et al., 2023) focuses solely on training the Whisper-based audio encoder using large-scale datasets, achieving proficiency in universal audio understanding across a diverse range of tasks and audio types. However, despite their impressive performance in the audio domain, none of these models have the capability to perceive and reasoning about spatial audio that is situated in diverse, reverberant, and complex 3-D environments.

3. SPATIALSOUNDQA

The training of BAT requires a spatial audio-based question answering dataset. Current datasets like SpatialVLM (Chen et al., 2024a) and Pano-AVQA (Yun et al., 2021) are focused on vision and lack spatial audio information (or audio altogether). To fill this void, we introduce SPATIALSOUNDQA, the first spatial audio-based question answering dataset, which is designed to emphasize the complexities of the spatial audio, free from visual influences, and caters to the unique demands of spatial audio perception and reasoning.

3.1. Spatial Audio Generation

Collecting and annotating spatial audio in the real world is a time consuming task that is made more complex by environmental acoustic variability and limitations in recording equipment. In order to efficiently create a dataset with a large amount of variability across environments and types of sound sources that is also rich in ground-truth metadata about these sources, we have adopted a simulation-based approach for generating spatial audio data for the dataset.

Spatial Audio Simulator. We use the state-of-the-art audio simulator, SoundSpaces 2.0 (Chen et al., 2022). This platform performs on-the-fly geometry-based sound rendering, enabling realistic acoustic reverberation with arbitrary source-receiver locations. Many simulation parameters can be easily modified, such as the material properties of a room's walls and the objects within the room, as well as the geometry of the receiver's microphone array. This enables us to curate a diverse and highly realistic dataset that also offers easy access to the ground-truth sound generation parameters, such as the spatial location and orientation of each source within the environment. Specifically, we leverage Matterport3D (Chang et al., 2017) for our environmental meshes, which includes highly detailed mesh renderings of 90 complete buildings, each featuring an average of 24.5 rooms across 2.61 floors with an average floorspace of 517.34 m^2. For a given arbitrary source location s, a monaural sound source A^s, receiver location r, and the receiver's heading direction θ in a specific mesh environment, the audio signal A^r received by a microphone is given by the convolution of the room impulse response with the sound source, as shown in the following equation:

$$A^r_m(t) = R_m(t, s, r, \theta) * A^s(t) \tag{1}$$

where $t \in [1, T]$ represents the time sample index, $m \in [1, 2, ..., n]$ represent the microphone index and

$R_m(t, s, r, \theta)$ is the impulse response between the source and receiver.

To minimize perceptual dissonance in auditory localization when visual cues are absent, we ensure that both the sound source and the receiver are located within the same room. We position the receiver at coordinates that allow for an upright stance. As for the sound sources, they are placed at random locations throughout the room. In total, 21,131 reverberations are generated, averaging approximately 9.58 reverberations per room.

Sound Sources. Previous spatial audio datasets (Yang et al., 2020; Shimada et al., 2023) often encompass a limited range of audio events, typically restricted to categories like music, speech, and domestic sounds. To broaden our scope and better represent real-world acoustic scenarios, we sample from AudioSet (Gemmeke et al., 2017) to specify our monaural sound source A^s. AudioSet consists of roughly 2 million 10-second in-the-wild YouTube clips used for audio classification, with each clip weakly annotated with 527 types of audio events, often featuring multiple overlapping events within a single clip. However, certain categories among these 527 types are discernible only through visual cues (e.g., "Single-lens reflex camera" vs. "Camera"). To enhance the dataset's reliability, we exclude labels that require visual information by manual inspection. Additionally, to mitigate the impact of poor-quality training samples, we remove labels with less than 50% quality.[2] Ultimately, given that the input audio undergoes convolution with reverberation, we further exclude most noise-related labels, such as "Noise", "Echo" and "Outside, urban or manmade". This curation results in a set of 355 audio event labels that are appropriate for our setting, identifiable solely through audio cues. After filtering by these categories, we are left with 1,861,750 clips in the AudioSet-2M split, while our filtered AudioSet-20K split contains 18,373 clips, and our evaluation set consists of 17,148 clips. Lastly, to counteract any volume variability in the source audio, we applied loudness normalization across all clips.

3.2. Question-Answer Pair Generation

We curated SPATIALSOUNDQA to include diverse question-answer pairs, all centered around the challenges of spatial audio perception and reasoning. Similar to LTU (Gong et al., 2024), each sample in the dataset is structured as a tuple of (audio, question, answer), where the audio and question serve as inputs to the model, and the answer acts as the target label. Table 1 illustrates the array of tasks in SPATIALSOUNDQA, ranging from basic perception to complex spatial reasoning. To enhance variety, the questions in these tasks are paraphrased using GPT-4, ensuring a broad spectrum of queries. The answers, on the other hand, are generated uniformly through a systematic rule-based approach, maintaining consistency across the dataset. We enumerate each question type in the following paragraphs.

Sound Event Detection (Type A & C). For this task, the question is "What sound events can you detect in the audio recording?" and its paraphrases. The answer may consist of a sorted list of sound classes present in the audio clip, arranged in alphabetical order. Question type A utilizes audio with 1 sound source, while C uses 2.

Direction and Distance Estimation (Type B & D). This task aims to identify the direction and distance of a sound source. For establishing ground truth, the three-dimensional space is segmented into eight regions. Each region is defined by a tuple representing the directional axis (left/right, front/behind, above/below) with respect to the receiver. Additionally, distance is quantified in increments of 0.5 meters, spanning a range from 0 to 10 meters. A typical question might be, "Where is the sound of the music coming from?" accompanied by GPT-generated paraphrased versions. The format for answers is constrained to expressions like "left, front, above; 2.5m". Question type B utilizes audio with 1 sound source, while D uses 2.

Spatial Reasoning (Type E). This task differs from the previous perception-focused tasks by introducing scenarios where two sound events occur concurrently, originating from distinct distances and directions. When generating these scenarios, we ensure that each mix contains two sound sources with entirely different sound events and orientations. The challenge here is to discern the spatial differences between these overlapping sounds. For example, a question might be, "What is the sound on the `left` side of the sound of the *dog barking*?" This task demands both perception and complex reasoning. The model must implicitly separate the sound sources based on their unique classes, spatially localize each source, and then analyze the relationship between the sources in the context of the question. However, the encoder SPATIAL-AST was not been exposed to such mixed data types during its pre-training stage, implying a lack of audio separation ability. This task is thus designed to test the model's in-context learning capabilities and effectively utilize the latent representations from the encoder to address this complex multi-source audio reasoning challenge. To make evaluation straightforward, we have chosen a binary (Yes/No) response format for these questions. By adopting binary answers, we can more effectively gauge the model's reasoning performance and maintain consistent assessment standards across different question types.

4. Method

Figure 1 illustrates our overall architecture for the BAT model. We will first introduce SPATIAL-AST, our

[2] For example, the estimated quality of the "clatter" label is 20%. Source: https://research.google.com/audioset/dataset/clatter.html.

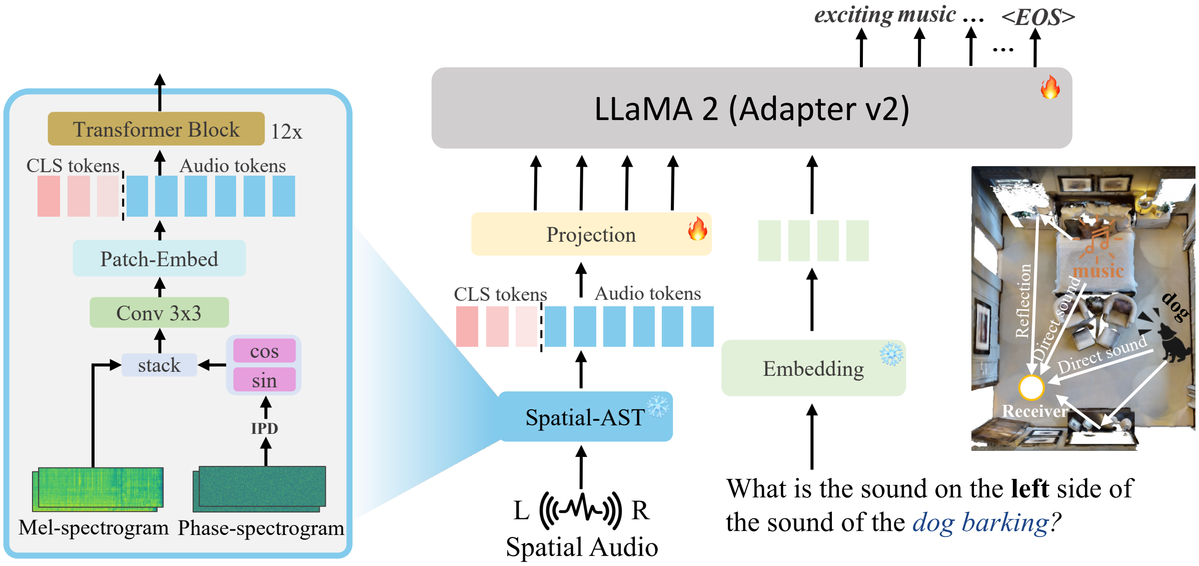

Figure 1. Illustration of the BAT model structure. The rightmost part shows the binaural audio generation process.

novel spatial audio encoder that acts as a front end for BAT. We then discuss how it is integrated with the LLaMA-2 LLM in order to build spatial reasoning ability on top of SPATIAL-AST's auditory perception ability.

4.1. Spatial Audio Encoder: SPATIAL-AST

We propose a novel architecture SPATIAL-AST to capture spatial audio information. The model is shown on the left-hand side of Figure 1.

Front-End. Our encoder integrates both Mel-Spectrogram and Interaural Phase Difference (IPD). As detailed in the SPATIAL-AST section of Figure 1, we initially transform the time-domain signal $x(n)$ into the frequency domain $X(t, f)$ using the short-time Fourier transform (STFT), as represented by Equation 2:

$$X_i(t,f) = \sum_{n=0}^{N-1} x_i(n) w(n-t) e^{-j\frac{2\pi f n}{N}}, \; i \in \{1,2\} \quad (2)$$

Here $w(n)$ denotes an N-point window function, and i indicates the left or right channel in the binaural recording.

Subsequently, we compute both the Mel-Spectrogram and the Interaural Phase Difference (IPD) based on these frequency domain signals $X_i(t, f)$. The Mel-Spectrogram for each channel is calculated as per Equation 3, and the IPD is derived from the phase spectrograms of the left and right channels, as indicated in Equation 4:

$$S_i(t,m) = log\left(|X_i(t,f)|^2 \times \text{melW}\right) \quad (3)$$

where melW is a M-bin Mel filter bank.

$$IPD(t,f) = \angle \frac{X_2(t,f)}{X_1(t,f)} \quad (4)$$

To address the numerical instabilities due to phase wraparound in IPD, we apply cosine and sine transformations, which are then weighted by melW to align with the dimension of the Mel-spectrogram. The final front-end output, $\mathcal{Z}$, is a concatenation of these processed components. It includes both the Mel-Spectrograms of the left and right channels, as well as the cosine and sine transformations of the IPD, all combined as outlined in Equation 5:

$$\begin{aligned} \mathcal{Z} = [&S_1; S_2; \cos(\text{IPD}) \times \text{melW}; \\ &\sin(\text{IPD}) \times \text{melW}], \quad \mathcal{Z} \in \mathbb{R}^{4\times T\times M} \end{aligned} \quad (5)$$

Backbone. The initial input $\mathcal{Z}$ is first passed to a 3x3 2D convolution followed by a batch normalization (Ioffe & Szegedy, 2015) layer and a GELU (Hendrycks & Gimpel, 2016) layer to fuse the inter-channel information. Subsequently, we employ a Patch-Embed CNN with a kernel size of (16×16) and corresponding strides in both time and frequency dimensions, ensuring non-overlapping segmentation of the features into distinct patch tokens. To extract spatial cues, we concatenate three [CLS] tokens at the beginning of the audio tokens, each specifically designated for extracting information about the audio's category, distance, and direction respectively. All tokens are then fed into 12-layer Transformer encoder blocks. Finally, we process the three [CLS] tokens using three separate linear layers to get predictions for their corresponding tasks.

Table 2. The perception-to-reasoning curriculum.

Stage	Question Type	# Tr. Samples	Percentages
I	A, B	278K	31.8%
II	A, B, C, D	514K	58.8%
III	A, B, C, D, E	872K	100%

Pre-Training Objective. The spatial encoder is pre-trained to handle three tasks, namely sound event detection, distance prediction, and directional prediction, with the latter being further divided into azimuth and elevation angle predictions. We employ cross-entropy loss for all three tasks. To discretize prediction targets for distance and directional prediction, we categorize distance into intervals of 0.5 meters and angle into intervals of 1 degree for both azimuth and elevation.

The final pre-training objective of the spatial encoder is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{dis} + \lambda_3 \mathcal{L}_{doa} \tag{6}$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{dis}$, $\mathcal{L}_{doa}$ represent the losses for detection, distance, and direction respectively, while λ_1, λ_2, λ_3 are hyperparameters.

4.2. BAT: A Model for Reasoning about Spatial Sounds

On its own, SPATIAL-AST is capable of predicting the type, direction, and distance for an input spatial sound. To extend this ability to encompass spatial reasoning about multiple sound sources present within an environment, we fuse SPATIAL-AST to the LLaMA-2 7B LLM (Touvron et al., 2023b). Input spatial audio is first processed by the SPATIAL-AST encoder, and then a projection module is used to map its set of output tokens into LLaMA-2's input text embedding space. For efficient fine-tuning, we use the LLaMA-adapter v2 (Gao et al., 2023).

The fine-tuning objective of BAT is to predict the answer text conditioned on its paired question text and the corresponding audio input. The model seeks to maximize the probability of predicting the next answer token, applying cross-entropy loss for each token position $1 \leq \tau \leq T$ in the sequence, which can be mathematically represented as:

$$\mathcal{P}_\theta \left(y_\tau \mid y_{\tau-1}, \cdots, y_1; a_l, a_{l-1}, \cdots, a_1 \right) \tag{7}$$

where a_l is the l-th token of audio embedding, and y_τ denotes the τ-th token of text output.

For the training of BAT, we devised a perception-to-reasoning curriculum, as detailed in Table 2. The process begins with a "warm-up" phase focusing on single-source perception tasks (question types A and B), enabling the model to adapt to audio modality inputs and infer basic properties of a single audio source including its class, distance, and direction. In the second stage, we introduce multi-source scenarios (question types C and D), where the model is taught to implicitly perform sound source separation by answering questions about specific sources within mixed audio. The final stage unlocks the full capabilities of BAT, integrating reasoning tasks (question type E), such as interpreting spatial relationships and distances between multiple sound sources. The importance of this curriculum is evidenced in Table 10 in Appendix H.

5. Experiments

5.1. Implementation Details

Input Audio Processing. The 10-second audio sources are first loudness normalized by scaling them so that each clip has the same total sound energy. After convolution with the room impulse response from SoundSpaces, we trim or pad these clips to 10 seconds to align with AudioMAE (Huang et al., 2022). The resulting waveforms are binaural with 2 channels at a 32kHz sampling rate. We use a window size of 1024, a hop size of 320, and 128 mel-bins to compute the Short-Time Fourier Transforms (STFTs) and mel-spectrograms. As a result, for a 10-second recording from AudioSet, the concatenated Mel-spectrogram and IPD feature dimension is (4, 1024, 128).

Spatial Audio Encoder. We implement a patch masking ratio of 0.25 in both time and frequency during training, and the unmasked tokens are then concatenated with three learnable [CLS] tokens. We initialize the weights of the transformer blocks using the official pretrained AudioMAE (Huang et al., 2022) checkpoint. The encoder is trained on 8 RTX 3090 GPUs, with each epoch taking approximately 10 minutes. Recognizing the greater challenge posed by sound event detection compared to distance and directional prediction, we train our encoder in two stages. In the first stage, we focus exclusively on detection loss, setting the weights λ_2 and λ_3 to zero as per Equation 6. In the second stage, we reintroduce the losses for distance and direction, adjusting the loss weights to $\lambda_1 = 1250$, $\lambda_2 = 1$, and $\lambda_3 = 2$ to broaden the capabilities of the model. This two-stage pre-training approach is crucial for achieving a balanced performance across different tasks. The effects of varying these weights on model performance for various tasks are detailed in Appendix E.

LLM. Diverging from the original two-stage learnable parameters design of LLaMA-adapter v2, We concurrently train parameters including zero-initialized attention, projection, norm, bias, and scale to simplify the training process. The training is completed on 8 V100 GPUs. For generation, we employ a greedy decoding and set the temperature to 0.1 and nucleus sampling (Holtzman et al., 2020) top p to 0.75.

Table 3. This table presents a comparative analysis of various models using both monaural and binaural data, focusing on key performance metrics: mean Average Precision (mAP, ↑), Error Rate at 20° ($ER_{20°}$, ↓), Mean Angular Error (MAE, ↓), and Distance Error Rate (DER, ↓). Results for AudioMAE models were derived through inference using their official checkpoints, without further training. Models that utilize anechoic audio data (no reverberation) are grayed-out for clarity.

		Data		Performance Metrics			
		Audio Type	Input Features	mAP (↑)	$ER_{20°}$ (↓)	MAE (↓)	DER (↓)
AudioMAE (Huang et al., 2022)		Anechoic	Mel-spectrograms	53.56	-	-	-
AudioMAE (Huang et al., 2022)		Monaural	Mel-spectrograms	47.18	-	-	-
SPATIAL-AST (ours)	Joint-Training			51.39	95.85	88.52	**26.87**
	Stage-1			52.26	-	-	-
	Stage-2			50.69	95.35	89.42	28.25
SELDnet (Adavanne et al., 2018)		Binaural	Mel-spectrograms, IPD	42.66	25.19	19.21	38.46
SPATIAL-AST (ours)	Joint-Training		Mel-spectrograms	49.94	25.50	18.57	34.85
SPATIAL-AST (ours)			Mel-spectrograms, IPD	50.57	24.81	18.16	35.29
SPATIAL-AST (ours)	Stage-1	Binaural	Mel-spectrograms	51.70	-	-	-
	Stage-2			50.86	30.94	21.66	28.60
SPATIAL-AST (ours)	Stage-1		Mel-spectrograms, IPD	**52.26**	-	-	-
	Stage-2			50.03	**23.89**	**17.94**	32.54

Baseline. For our encoder baseline comparisons, we select established models such as AudioMAE (Huang et al., 2022) and SELDnet (Adavanne et al., 2018). To ensure a fair comparison, SELDnet's feature extraction network is augmented with a 12-layer transformer block, aligning the model parameters of the compared architectures to approximately 90M. Additionally, for BAT, we use a version with monaural input for comparison, which has an architecture similar to LTU (Gong et al., 2024), serving as a baseline.

Evaluation Metrics. In the evaluation of our spatial audio encoder, outlined in Table 3, we use the mean Average Precision (mAP) for sound event detection; Error Rate at 20° ($ER_{20°}$), as referenced in (Mesaros et al., 2019), and Mean Angular Error (MAE) for Direction of Arrival (DoA) accuracy, evaluating whether predictions deviate more than 20° from the reference point and quantifying the average angular deviation between predictions and ground truth; and Distance Error Rate (DER) for measuring the accuracy of distance predictions within 0.5 meters of the actual location. For the evaluation of BAT, as shown in Table 4, we apply similar metrics, including mAP and DER. For the DoA task, where the three-dimensional space is segmented into eight distinct sections, we use accuracy (Acc) as the performance metric. Additionally, for reasoning tasks in question type E, Binary Accuracy (BA) serves as the evaluation metric.

5.2. Performance of SPATIAL-AST on SELD Tasks

Table 3 presents the experimental results for our proposed SPATIAL-AST audio encoder. We compare SPATIAL-AST against current state-of-the-art models such as AudioMAE (Huang et al., 2022) and SELDnet (Adavanne et al., 2018). We also compare spatial cue extraction between monaural and binaural data. Additionally, we explore the impact of two pre-training methodologies, including joint-training and two-stage approaches, and highlight the importance of IPD.

Monaural vs. Binaural. To support monaural input, we modified the 3x3 convolution module, setting its input channel to 1, and only feed a monaural mel spectrogram as input. Despite binaural data typically offering richer spatial information, our encoder demonstrates robust performance with monaural inputs, especially in sound event detection (SED) and distance prediction (DP) tasks. The encoder achieves a mAP of 51.39 for monaural input, surpassing both its binaural counterpart, which scores 49.94, and AudioMAE, which achieves 47.18 in mAP. Additionally, our encoder with monaural input even attains the best Distance Error Rate (DER) compared to models using binaural input. However, as monaural audio cannot capture spatial information, it shows low performance in the direction-of-arrival (DoA) task, with an $ER_{20°}$ of 95.82 and an MAE of 88.52. The performance of monaural data suggests it is sufficiently effective for SED and DP tasks. The effectiveness of monaural input can be attributed to the fact that binaural audio does not necessarily provide additional benefits for sound category perception. The results from the single-task setting detailed in Appendix D corroborate this conclusion.

Joint-training vs. Two-stage training. Joint-training emerges as the more effective approach for monaural data, outperforming two-stage training in both detection and dis-

Table 4. Performance evaluation on SPATIALSOUNDQA. The metrics include mAP for detection accuracy, represented by two columns corresponding to question types A and C, respectively; Accuracy (Acc) for assessing the accuracy of identifying directions and Distance Error Rate (DER) for distance prediction, with two columns each for question types B and D; and Binary Accuracy (BA) for the accuracy of binary (Yes/No) responses in reasoning (question type E). Input types include monaural audio (M), binaural audio (B), prompt (P), or ground truth sound source parameters (GT). The "Random" and "Oracle" rows serve as baseline and ideal performance benchmarks.

Model	Input	Perception (Type ACBD)†			Reasoning (Type E)		
		Detection (mAP ↑)*	DoA (Acc ↑)	DP (DER ↓)	Direction	Distance	Avg.
Random	P	0.61 | 0.59	12.57 | 12.41	67.33 | 67.46	50.00	50.00	50.00
Monaural BAT (ours)[1]	M + P	24.15 | 6.42	14.31 | 11.93	34.17 | 56.26	57.69	51.36	54.53
One-stage BAT[1]	B + P	26.23 | 9.06	**76.49** | 37.09	29.28 | 51.62	65.12	81.98	73.55
One-stage BAT[1, 2]		25.50 | 5.67	72.74 | 35.63	31.51 | 73.68	59.59	49.76	54.68
BAT	P	0.69 | 0.65	12.88 | 13.14	51.65 | 71.73	58.28	50.67	54.48
	B + P	**26.34** | **9.89**	75.54 | **37.65**	**29.16** | **47.90**	**69.77**	**84.01**	**76.89**
Oracle[1]	GT + P	94.16 | 94.26	99.20 | 99.81	0.00 | 0.00	98.21	100.00	99.10

† We grey out the columns corresponding to question types A and B, where the model receives a single sound source.

* Like LTU (Gong et al., 2024), we use a text encoder (text-embedding-3-small) to encode BAT's output and the labels from the evaluation dataset, then compute the cosine similarity between these text embeddings for evaluation. mAP potentially underestimates BAT's performance.

[1] Only train stage III. [2] We removed question types C and D from SPATIALSOUNDQA.

tance prediction tasks, achieving a mAP of 51.39 and DER of 26.87. However, for binaural data, two-stage training consistently demonstrates improvements in its second stage, regardless of whether IPD is used. The enhancement is particularly significant with the usage of IPD, achieving the lowest MAE of 23.89 and DER of 17.94, demonstrating the importance of including multi-microphone phase information for spatial audio perception. This also underscores the importance of employing a two-stage training approach for enhanced performance in binaural settings.

Model Architecture. When compared to AudioMAE and SELDnet, our encoder exhibits enhanced performance in reverberant environments. As AudioMAE is tailored for monaural input, we focus our comparison on monaural scenarios. In these contexts, our proposed model surpasses both joint-training and two-stage training methods. This achievement is especially significant given that our model has been finetuned for much fewer training hours. For simplicity, we compare our model with SELDnet in joint-training, where it surpasses SELDnet on all metrics.

5.3. Performance of BAT on SPATIALSOUNDQA

Table 4 shows the performance of various configurations of our BAT model on SPATIALSOUNDQA, broken down across the different question types. BAT performs strongly across all question types, outperforming the random baseline by a very large margin. We observe that the one-stage BAT generally lags behind the three-stage BAT in performance across various metrics. Intriguingly, we observed a significant drop in the model's reasoning ability when question types C and D, related to sound source separation, were removed from the training set. This is evident from the Binary Accuracy (BA) results, which approximate random chance in the reasoning tasks. Analysis of the one-stage BAT's outputs reveals that the model struggles to differentiate between sound sources and their respective positions when confronted with reasoning questions. This finding highlights the importance of sound source separation tasks in training BAT to effectively handle complex reasoning, underscoring the necessity of a multi-stage curriculum that gradually introduces the model to increasingly sophisticated spatial audio perception and reasoning tasks.

The "Monaural BAT" row shows the performance of BAT trained with monaural audio. The architecture of monaural BAT is similar to that of LTU (Gong et al., 2024), so we use it as a baseline for comparison. The results reveal that monaural BAT faces significant challenges in reasoning tasks, achieving a Binary Accuracy (BA) of only 54.53%. This suggests that monaural audio input alone cannot provide sufficient spatial information for complex reasoning challenges, highlighting the usefulness of SPATIALSOUNDQA for learning spatial audio reasoning capabilities. More ablations can be found in Appendix H.

6. Limitations and Future Work

We are confident that BAT will significantly contribute to the development of spatial audio perception and reasoning as well as multimodal LLM. However, as with any research tool, it is crucial to acknowledge certain limitations and assumptions inherent in our approach.

One of the primary challenges we encounter is extracting as much information as possible from spatial audio, with a specific focus on how to better utilize phase information. Currently, our model handles a maximum of two sound sources, with an emphasis primarily on audio. Looking ahead, there is potential to expand into multi-source scenarios and to integrate both audio and speech processing for a more holistic approach. Additionally, while our current framework is limited to binaural audio, exploring ambisonics could provide a more immersive and realistic spatial audio experience. Moreover, expanding SPATIALSOUNDQA to be more open-ended would better align with human usage patterns and preferences, allowing for a wider range of queries and responses.

Another important consideration is the integration of additional modalities, such as visual information, to complement auditory cues. This multimodal approach might enhance the model's understanding and interpretation of complex environments. Lastly, the sim2real gap remains an aspect that requires further investigation. Observing how our model performs in real-world scenarios, as opposed to simulated environments, will be crucial in assessing its practical applicability and effectiveness.

7. Conclusion

In this work, we have presented BAT, the first LLM capable of processing spatial audio input. To train and evaluate BAT, we introduced SPATIALSOUNDQA, the first extensive spatial audio-based question answering dataset. We also proposed SPATIAL-AST, a novel spatial audio encoder capable of efficiently handling sound event detection, spatial localization, and distance estimation. BAT and SPATIALSOUNDQA showcase the immense potential of LLMs for reasoning about spatial sound. SPATIALSOUNDQA also enables rich future work, such as reasoning about the environment itself (materials, shape, layout), modeling moving sounds to for tracking problems, or incorporating the visual modality which SoundSpaces2.0 natively supports.

Acknowledgements

We thank Changan Chen, Linfeng Feng, and Jin Sob Kim for their constructive help. This work was supported in part by various grants and institutions. Zhisheng Zheng was supported by Zhiyuan College at Shanghai Jiao Tong University. Puyuan Peng and David Harwath's work was supported by the National Science Foundation under Grant No. 2238605. Xie Chen was partially supported by the National Natural Science Foundation of China (No. U23B2018, No. 62206171), the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102, and the International Cooperation Project of PCL.

Impact Statement

Our research on BAT, the first spatial audio-based large language model, alongside the SPATIALSOUNDQA dataset, holds significant potential for broad impact across various fields:

Advancements in Spatial Audio: Our work addresses a key gap in spatial audio perception and reasoning. By developing an LLM specifically for spatial audio, we open up new possibilities for applications where spatial sound cues play a crucial role, such as virtual reality, gaming, and audio engineering. This can lead to more immersive and realistic experiences in these domains.

Enhancements in Multimodal Large Language Models: The integration of spatial audio into LLMs represents a significant step towards truly multimodal AI systems. Our model not only demonstrates the capability of LLMs in processing complex spatial audio information but also paves the way for future research in combining auditory and other sensory modalities, such as visual or tactile, to create even more sophisticated AI models.

Empowering Embodied AI: The ability to interpret and reason about spatial sounds can significantly enhance embodied AI systems, like robots or autonomous vehicles. These systems can utilize spatial audio cues for better navigation and interaction with their environment, leading to more effective and safer applications in real-world scenarios.

References

Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *Proc. JSTSP*, 2018.

Baade, A., Peng, P., and Harwath, D. F. Mae-ast: Masked autoencoding audio spectrogram transformer. *Proc. Interspeech*, 2022.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., et al. AudioLM: A language modeling approach to audio generation. *Proc. TASLP*, 2023.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., et al. Matterport3D: Learning from RGB-D data in indoor environments. *Proc. 3DV*, 2017.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Driess, D., et al. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024a.

Chen, C., Schissler, C., Garg, S., Kobernik, P., Clegg, A., et al. SoundSpaces 2.0: A simulation platform for visual-acoustic learning. *Proc. NeurIPS*, 2022.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., and Wei, F. BEATs: Audio pre-training with acoustic tokenizers. *Proc. ICML*, 2023.

Chen, W., Liang, Y., Ma, Z., Zheng, Z., and Chen, X. EAT: Self-supervised pre-training with efficient audio transformer. *arXiv preprint arXiv:2401.03497*, 2024b.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *Proc. SSST*, 2014.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., et al. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Deshmukh, S., Elizalde, B., Singh, R., and Wang, H. Pengi: An audio language model for audio tasks. *Proc. NeurIPS*, 2023.

Evers, C., Löllmann, H. W., Mellmann, H., Schmidt, A., Barfuss, H., et al. The LOCATA challenge: Acoustic source localization and tracking. *Proc. TASLP*, 2020.

Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., et al. LLaMA-Adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., et al. AudioSet: An ontology and human-labeled dataset for audio events. *Proc. ICASSP*, 2017.

Gong, Y., Chung, Y.-A., and Glass, J. AST: Audio spectrogram transformer. *Proc. Interspeech*, 2021.

Gong, Y., Luo, H., Liu, A. H., Karlinsky, L., and Glass, J. Listen, think, and understand. *Proc. ICLR*, 2024.

Guizzo, E., Marinoni, C., Pennese, M., Ren, X., Zheng, X., et al. L3DAS22 challenge: Learning 3d audio sources in a real office environment. *Proc. ICASSP*, 2022.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., et al. Conformer: Convolution-augmented transformer for speech recognition. *Proc. Interspeech*, 2020.

Hak, C. C., Wenmaekers, R. H., and Van Luxemburg, L. Measuring room impulse responses: Impact of the decay range on derived room acoustic parameters. *Proc. Acta Acustica*, 2012.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. *Proc. ICLR*, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., et al. LoRA: Low-rank adaptation of large language models. *Proc. ICLR*, 2022.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., et al. Masked autoencoders that listen. *Proc. NeurIPS*, 2022.

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., et al. AudioGPT: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*, 2023.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proc. ICML*, 2015.

Kim, J. S., Park, H. J., Shin, W., and Han, S. W. AD-YOLO: You look only once in training multiple sound event localization and detection. *Proc. ICASSP*, 2023.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proc. ICML*, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Proc. NeurIPS*, 2023.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *Proc. ICLR*, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *Proc. ICLR*, 2019.

Mesaros, A., Adavanne, S., Politis, A., Heittola, T., and Virtanen, T. Joint measurement of localization and detection of sound events. *Proc. WASPAA*, 2019.

Morgado, P., Li, Y., and Nvasconcelos, N. Learning representations from audio-visual spatial alignment. *Proc. NeurIPS*, 2020.

Mystakidis, S. Metaverse. *Encyclopedia*, 2022.

OpenAI. GPT-4V(ision) system card. 2023.

Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., et al. Grounding multimodal large language models to the world. *Proc. ICLR*, 2024.

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., et al. Semi-orthogonal low-rank matrix factorization for deep neural networks. *Proc. Interspeech*, 2018.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. *Proc. ICML*, 2023.

Scheibler, R., Bezzam, E., and Dokmanic, I. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. *Proc. ICASSP*, 2018.

Shimada, K., Koyama, Y., Takahashi, N., Takahashi, S., and Mitsufuji, Y. ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection. *Proc. ICASSP*, 2021.

Shimada, K., Politis, A., Sudarsanam, P., Krause, D., Uchida, K., et al. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *arXiv preprint arXiv:2306.09126*, 2023.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., et al. SALMONN: Towards generic hearing abilities for large language models. *Proc. ICLR*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Wang, J., Du, Z., Chen, Q., Chu, Y., et al. LauraGPT: Listen, attend, understand, and regenerate audio with GPT. *arXiv preprint arXiv:2310.04673*, 2023a.

Wang, Q., Wu, H., Jing, Z., Ma, F., Fang, Y., Wang, Y., Chen, T., Pan, J., Du, J., and Lee, C.-H. The USTC-Iflytek system for sound event localization and detection of dcase2020 challenge. *Proc. DCASE 2020*, 2020.

Wang, Q., Jiang, Y., Cheng, S., Hu, M., Nian, Z., et al. The nerc-slip system for sound event localization and detection of dcase2023 challenge. *Proc. DCASE 2023*, 2023b.

Wang, Y., Yang, B., and Li, X. FN-SSL: Full-band and narrow-band fusion for sound source localization. *Proc. Interspeech*, 2023c.

Yang, B., Ding, R., Ban, Y., Li, X., and Liu, H. Enhancing direct-path relative transfer function using deep neural network for robust sound source localization. *Proc. TIT*, 2022a.

Yang, B., Liu, H., and Li, X. SRP-DNN: Learning direct-path phase difference for multiple moving sound source localization. *Proc. ICASSP*, 2022b.

Yang, K., Russell, B., and Salamon, J. Telling left from right: Learning spatial correspondence of sight and sound. *Proc. CVPR*, 2020.

Yun, H., Yu, Y., Yang, W., Lee, K., and Kim, G. Pano-AVQA: Grounded audio-visual question answering on 360deg videos. *Proc. ICCV*, 2021.

Zhang, Y., Wang, S., Li, Z., Guo, K., Chen, S., and Pang, Y. Data augmentation and class-based ensembled cnn-conformer networks for sound event localization and detection. *Proc. DCASE 2021*, 2021.

A. Why Do We Use Large Language Models?

The use of Large Language Models (LLMs) in our spatial audio reasoning framework is necessary due to several critical reasons. First, LLMs enable complex questions to be posed using natural language, allowing for a more flexible and intuitive query format, such as "Is the dog further away from you than the stereo?" or "Is the dog to the left or to the right of the stereo?" This capability eliminates the need for a cascade of separate processing steps like source separation, classification, DoA, distance classifiers, and rule-based templates, which would significantly limit the types of questions that could be addressed. Additionally, LLMs facilitate the joint training of the spatial audio encoder and reasoning module in an end-to-end manner, reducing the risk of error propagation inherent in segmented processing approaches. Moreover, LLMs offer the ability to generalize across different questions with overlapping expressions, enhancing the model's adaptability and versatility in responding to a wide range of spatial audio-related queries. Last but not least, this task also intends to further expand the application boundaries of LLMs, demonstrating their versatility and effectiveness in domains beyond traditional text and image processing.

B. Hyperparameters

We present the specific training hyperparameter configurations for SPATIAL-AST & BAT in Table 5.

Table 5. Training hyperparameters of SPATIAL-AST & BAT

Configuration	SPATIAL-AST		BAT		
	Stage-1	Stage-2	Stage-I	Stage-II	Stage-III
Sound Source	AudioSet-2M		AudioSet-20K		
Audio Normalization	✓		✓		
Augmentation	✓		✗		
Weighted sampling	✓		✗		
Optimizer	AdamW (Loshchilov & Hutter, 2019)				
Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.95$				
Weight decay	0.0001		0.05		
Base learning rate	0.001		0.001		
Learning rate schedule	half-cycle cosine decay (Loshchilov & Hutter, 2017)				
Warm-up epochs	10	5	2		
Epoch partitioning factor	10		10		
Epochs	60	40	2	2	3
Batch size	64		2		
GPUs	8 RTX 3090		8 Tesla V100		

C. Room Acoustics Characteristics

Table 6 presents the average RT60 values across some common room types in the Matterport3D (Chang et al., 2017) training settings. RT60 is a standard acoustic measurement that is defined as the time it takes for the sound pressure level to reduce by 60 dB (Hak et al., 2012). This metric varies across different rooms due to factors such as room size, object arrangement, and construction materials, each influencing the room's acoustic characteristics.

Table 6. Average reverberation time (RT60) for different rooms.

	Living Room	Bedroom	Closet	Bathroom	Toilet	Garage	Lounge	Kitchen	Dining Room
RT60	0.2076	0.1917	0.1950	0.1906	0.1706	0.2190	0.2224	0.2061	0.2201

D. Performance of SPATIAL-AST in a Single-Task Setting

We evaluate the performance of our encoder in single-task scenarios, specifically focusing on classification, distance prediction, and direction-of-arrival (DoA). As shown in Table 7, we analyze the impact of including or omitting Interaural Phase Difference (IPD), as well as comparing monaural versus binaural formats, on the performance of each task. Our findings reveal that IPD generally enhances performance across all tasks, indicating the value of spatial information. Moreover, we observe that monaural data provides sufficient information for tasks like sound event detection and distance prediction, achieving the best performance in Distance Prediction and comparable results to binaural data in Sound Event Detection. However, binaural data significantly improves performance in tasks involving direction perception, underscoring the encoder's potential to handle multiple tasks effectively and the importance of audio format in spatial audio perception.

Table 7. Performance of the encoder in single-task scenarios: classification, distance prediction, and Direction-of-Arrival (DoA). We compare the impact of including or excluding Interaural Phase Difference (IPD) on the performance of each task. Key evaluation metrics include mean Average Precision (mAP, ↑), Error Rate at 20° ($ER_{20°}$, ↓), Mean Angular Error (MAE, ↓), and Distance Error Rate (DER, ↓).

Task	Data		Performance Metrics			
	Audio Format	Input Features	mAP (%) ↑	$ER_{20°}$ (%) ↓	MAE (°) ↓	DER (%) ↓
Detection	Monaural	Mel-spectrogram	52.13	-	-	-
Direction-of-Arrival			-	96.80	85.47	-
Distance Prediction			-	-	-	**21.44**
Detection	Binaural	Mel-spectrogram	51.70	-	-	-
		Mel-spectrogram, IPD	**52.26**	-	-	-
Direction-of-Arrival		Mel-spectrogram	-	21.42	15.98	-
		Mel-spectrogram, IPD	-	**20.11**	**15.36**	-
Distance Prediction		Mel-spectrogram	-	-	-	25.55
		Mel-spectrogram, IPD	-	-	-	22.60
Generalist	Binaural	Mel-spectrogram, IPD	50.03	23.89	17.94	32.54

E. Ablations of the Spatial Audio Encoder SPATIAL-AST

Loss Weights. As detailed in Equation 6, we fix $\lambda_2 = 1$ and $\lambda_3 = 2$, and by adjusting the value of λ_1, we monitor changes in the model's performance. As depicted in Figure 2, with the increase of λ_1, the mAP consistently improves. However, for the MAE metric in direction-of-arrival (DoA), the performance gradually deteriorates. This presents a trade-off between the two metrics. Prioritizing classification, we opt for settings that yield the highest mAP. In subsequent experiments, we fix the hyperparameters at $\lambda_1 = 1250, \lambda_2 = 1, \lambda_3 = 2$.

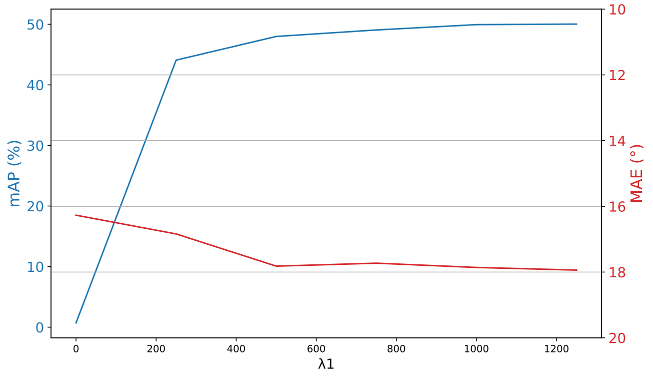

Figure 2. Classification (mAP) and Direction-of-Arrival (MAE) for different λ_1 values.

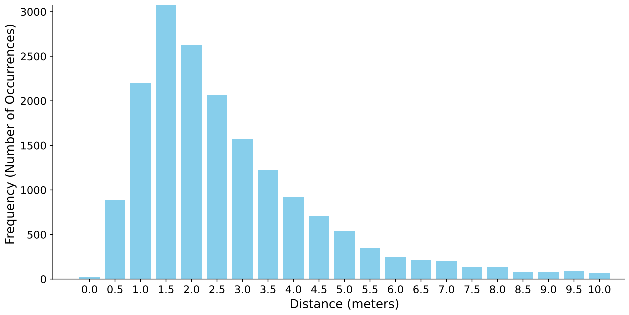

Figure 3. Frequency distribution of training reverberations' distances ranging from 0 to 10 meters.

Table 8. Ablation study on monaural audio input: We employ our proposed architecture to train and infer under four distinct scenarios, with and without audio normalization and reverberation.

Normalization	Reverberation	mAP (%) ↑	DER (%) ↓
✗	✗	52.64	49.49
✓	✗	52.87	49.51
✗	✓	51.13	39.57
✓	✓	51.39	**26.87**

Audio Normalization and Reverberation. In this ablation study, we assess the effects of audio normalization and reverberation on the encoder's performance, using monaural and fixing $\lambda_1, \lambda_2, \lambda_3 = 1250, 1, 2$ for simplicity, as outlined in Table 8. The addition of reverberation indeed impacts the model's ability to detect sound events, as it makes the task more challenging without adding extra event information. A comparison between rows 1 and 2 with rows 3 and 4 indicates that removing reverberation significantly enhances mean Average Precision (mAP), increasing from 51.13% to 52.64%. However, it's noteworthy that audio without reverberation lacks spatial information, leading to a Distance Error Rate (DER) for distance perception that is nearly equivalent to random guessing. Based on the distance distribution described in Figure 3, random sampling yields approximately 67% DER, and when sampling is limited to specific distances like 1.0, 1.5, and 2.0 meters, it results in around 53% DER. The implementation of audio normalization further improves the encoder's performance in both classification and distance tasks. Comparing rows 3 and 4, the mAP increases slightly from 51.13% to 51.39%, while DER significantly drops from 39.57% to 26.87%, underscoring the benefits of audio normalization in our model.

F. SPATIALSOUNDQA Generation

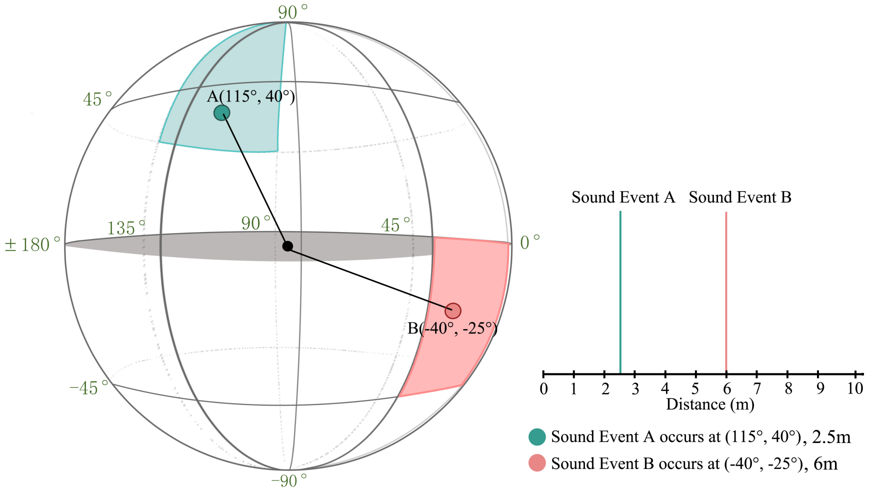

Figure 4. Sound events A and B occur in different directions and distances simultaneously.

In the process of generating SPATIALSOUNDQA, we simulate spatial audio scenarios where multiple sound events occur simultaneously, each at distinct directions and distances from the receiver. For example, as depicted in Figure 4, consider a situation where two sound sources, A and B, are active concurrently. Sound event A occurs at an azimuth of 115 degrees and an elevation of 40 degrees, situated 2.5 meters away from the receiver. In contrast, sound source B occurs at coordinates of -40 degrees azimuth and -25 degrees elevation, positioned 6 meters distant. It's important to note that in the actual dataset construction, the direction and distance parameters are often not as distinct as in this example. With these ground truth spatial parameters, we can then construct a diverse array of questions and answers concerning the spatial relationships between the two sound sources, as shown in Table 1.

In supplement to Table 1, we provide details for each question type, including the initial prompt used and the corresponding answer template.

Table 9. Prompts and answer templates for SPATIALSOUNDQA

Type	Prompt	Answer template
A: Detection	Identify the sound events in the audio clip.	sound event1; sound event2
B: DP & DoA	Where is the audio clip coming from? How would you describe the location of the {sound event}'s sound?	{d1}, {d2}, {d3}; {distance}
C: Detection	Identify the sound events in the audio clip coming from the {d1}, {d2}, {d3}, approximately {distance} away.	sound event1; sound event2
D: DP & DoA	Where is the sound of the {sound event} coming from?	{d1}, {d2}, {d3}; {distance}
E: Reasoning	Do the sound of {sound event1} and the sound of {sound event2} both come from your {d} side? Can you hear the {sound event1}'s sound from the {d1} and the {sound event2}'s from the {d2}? Would you find the sound of {sound event1} on the {d} side of the sound of {sound event2}? In terms of straight-line distance, does the sound of {sound event1} reach you from a closer point compared to the sound of {sound event2}?	Yes/No

G. Prompt Details

For our spatial audio-based large language model BAT, we designed a specific prompt to guide its responses to various spatial audio tasks. The prompt structure is as follows:

```
"Based on the audio you've heard, refer to the instruction and provide a response.\n\n
### Instruction:\n{instruction}\n\n### Response:"
```

To form the input, we replace "{instruction}" with our specific question, and then concatenate a fixed length of learnable query tokens at the beginning.

H. Ablations of BAT

Table 10. The influence of different training curriculum on BAT's reasoning ability.

Model	#Sources	Stage			Reasoning (BA ↑)		
		I	II	III	Direction	Distance	Avg.
Random	-	-	-	-	50	50	50
BAT	2	✓	✗	✗	0	0	0
		✓	✓	✗	0	0	0
		✗	✗	✓	65.12	81.98	73.55
		✗	✗	✓*	59.59	49.76	54.68
		✗	✓	✓*	64.69	79.16	71.93
		✓	✓	✓*	69.56	83.51	76.54
		✓	✓	✓	**69.77**	**84.01**	**76.89**

* We removed question types C and D from SPATIALSOUNDQA.

In Table 10, we thoroughly examine the influence of different training curriculum on BAT's reasoning ability. Initially, the model encounters difficulties with reasoning tasks, particularly when facing unfamiliar reasoning-style questions, as evidenced by its inability to produce accurate binary (Yes/No) responses. A one-stage version of BAT surprisingly achieves a respectable overall Binary Accuracy (BA) of 73.55% in reasoning tasks. However, when we remove question types C and D, which are related to sound source separation, from the one-stage training, the model's reasoning ability drops significantly, exhibiting near-random performance with a BA of only 54.68%. Intriguingly, when BAT is first trained on sound source separation tasks before proceeding to the final reasoning stage, it retains its ability to reason effectively, even in the absence of separation-related questions during the final stage. The last row of Table 10 showcases the optimal results achieved through our three-stage curriculum, demonstrating the importance of a comprehensive training approach in equipping BAT with robust reasoning skills.

Table 11. Impact of different tokens from SPATIAL-AST on One-stage BAT's reasoning ability.

CLS tokens	Audio Tokens	Reasoning (BA ↑)		
		Direction	Distance	Avg.
✓	✗	60.32	72.92	66.62
✗	✓	61.70	78.95	70.33
✓	✓	**65.12**	**81.98**	**73.55**

Table 12. Impact of learnable query size on reasoning accuracy.

Learnable Query	Reasoning (BA ↑)		
	Direction	Distance	Avg.
32	61.42	83.49	72.45
64	**69.77**	**84.01**	**76.89**

Table 11 demonstrates that the combination of both CLS and audio tokens yields the best performance in reasoning tasks. However, it is noteworthy that using only three CLS tokens as a representation for spatial audio also results in impressive performance. As per Table 12, a larger learnable query size in the projection module correlates with improved reasoning efficacy.