THE WORLD IS NOT MONO: ENABLING SPATIAL UNDERSTANDING IN LARGE AUDIO-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing large audio-language models perceive the world as ”mono”—a single stream of audio that ignores the critical spatial dimension (”where”) required for universal acoustic scene analysis. To break down this fundamental limitation, we introduce a framework that enables models like Qwen2-Audio to understand and reason about the complex, three-dimensional acoustic world. Our framework achieves this through three core contributions: First, we build a large-scale, synthesized binaural audio dataset to provide the rich spatial cues. Second, we design a novel Mixture-of-Experts (MoE) architecture, where a learnable router directs outputs from parallel semantic and spatial encoders to specialized expert pathways. Finally, we employ a progressive training curriculum, advancing from supervised fine-tuning (SFT) to reinforcement learning via Group Relative Policy Optimization (GRPO), to evolve the model's capabilities from basic perception to advanced reasoning. On our comprehensive benchmark, the model demonstrates a strong capability for spatial understanding. By enabling this spatial perception, our work provides a clear pathway for leveraging the powerful reasoning abilities of large models towards holistic acoustic scene analysis, advancing from one-dimensional semantic recognition to three-dimensional spatial intelligence.

1 INTRODUCTION

The expansion of Large Language Models (LLMs) from text-centric processing to multimodal intelligence has brought renewed attention to the auditory domain. Recent Large Audio-Language Models (LALMs) have demonstrated remarkable capabilities in understanding and following instructions related to the semantic content of audio—the 'what' in speech, general sounds, and music (Chu et al., 2024; Yang et al., 2024). However, prevailing LALM frameworks predominantly treat audio as a monophonic time series, lacking explicit modeling of spatial attributes—the 'where'. Technical reports and public benchmarks for these models rarely address reasoning about direction, distance, or spatial relationships, a gap that constitutes the primary motivation for our work.

This disregard for spatial dimensions starkly contrasts with the fundamental mechanisms of biological hearing. As established in classical auditory scene analysis, the brain leverages binaural cues, such as Interaural Time Differences (ITD) and Interaural Level Differences (ILD), to localize sound sources and segregate auditory streams. This process provides a ”spatial release from masking” and underpins the 'cocktail party effect,' enabling focused listening in complex acoustic environments. A model incapable of processing these binaural cues is, in a perceptual sense, ”spatially deaf,” responding only to a one-dimensional projection of a three-dimensional sound field.

This limitation imposes significant constraints on real-world applications in robotics, augmented reality, and assistive hearing, all demanding a joint understanding of audio's content and spatial arrangement. Recent explorations have begun extending the ”encoder-to-LLM” paradigm to spatial audio, typically by processing binaural signals or Ambisonics recordings (You et al., 2025). For instance, BAT(Zheng et al., 2025) introduced spatial question answering, while subsequent methods improved reasoning by decoupling spatial features or using contrastive learning for localization (Devnani et al., 2024). Despite these promising first steps, the research landscape remains fractured, with most large-scale models focused on semantics (Tang et al., 2024; Huang et al., 2023) while specialized systems handle spatial physics (Shimada et al., 2021). We propose to bridge this di-

vide, using the reasoning core of an LLM to unify these capabilities and process open-ended spatial queries with physically consistent representations.

However, the primary bottleneck to integrating spatial awareness into general auditory intelligence is not the lack of algorithmic starting points, but the scarcity of appropriate data and representations. Large-scale audio datasets like FSD50K (Fonseca et al., 2022), Clotho (Drossos et al., 2019), and MusicCaps (Agostinelli et al., 2023b) are rich in semantic labels but lack the binaural or 3D spatial metadata required for joint semantic-spatial alignment. Real-world spatial data, while valuable, remains limited in scale and diversity. Consequently, physically-consistent procedural simulation emerges as the most viable path forward. Tools such as Pyroomacoustics (Scheibler et al., 2017) and high-resolution Head-Related Transfer Function (HRTF) databases (Audio Engineering Society, 2015) enable the systematic generation of large-scale training corpora with complete "position-semantic-environment" annotations.

In this context, we introduce **The World is Not Mono (TWNM)**, a framework that extends LALMs from semantic recognition to spatial reasoning by integrating synthetic binaural data, specialized representation learning, and reinforcement learning alignment.

Contributions. The main contributions of this work are:

- **Synthetic binaural data pipeline**: construction of a scalable simulation framework using physically realistic Binaural Room Impulse Responses (BRIRs) / Head-Related Transfer Functions (HRTFs) to generate large-scale, spatially annotated audio–language QA pairs across diverse environments.
- **Task-aware MoE architecture**: design of a MoE framework with supervised routing to explicitly decouple semantic and spatial processing, while retaining a shared expert for generalization.
- **Training and alignment**: a progressive curriculum that first optimizes experts and the router with the LLM frozen, followed by joint fine-tuning. To further align spatial reasoning with natural language, GRPO (Shao et al., 2024) is employed, using rule-based rewards and relative advantages to improve output accuracy in multiple-choice spatial QA tasks.
- **Comprehensive evaluation**: demonstration of robust spatial understanding and relational reasoning on our promoted benchmarks, highlighting a viable path for spatially aware LALMs.

This work positions spatial perception not as an incremental feature, but as a fundamental dimension essential for achieving genuine auditory intelligence. The remainder of the paper is organized as follows: Section 2 reviews prior efforts on LALMs, spatial audio modeling, and data synthesis. Section 3 introduces our proposed framework. Section 4 presents evaluation results on our proposed benchmark. Finally, Section 5 concludes with discussions and future directions.

2 Related Work

Large Audio-Language Models Foundation models such as Qwen2-Audio (Chu et al., 2024), SALMONN (Tang et al., 2024), and AudioGPT (Huang et al., 2023) unify speech, sound, and music under a language interface, but remain focused on semantic tasks. Most benchmarks neglect localization and spatial reasoning, leaving models unable to process direction or distance.

Spatial Audio and LLMs Recent efforts extend the encoder-to-LLM paradigm to spatial audio. BAT(Zheng et al., 2025) evaluated binaural reasoning on the SpatialSoundQA benchmark, and ELSA aligned spatial audio and text via contrastive learning (Devnani et al., 2024). These approaches show potential but leave representation design and training strategies unresolved.

Computational Auditory Scene Analysis (CASA) SELD methods using FOA input, e.g., ACCDOA (Shimada et al., 2021) and STARSS23 (Shimada et al., 2023), jointly predict sound class and direction. While effective for structured outputs, they mismatch the open-vocabulary and conversational style of LALMs, and early QA adaptations remain limited.

Data Synthesis Because large-scale spatially annotated corpora are scarce, benchmarks often rely on simulation. SpatialSoundQA, for example, renders binaural or FOA mixtures by convolving

clean sources with room responses. Public content sets such as FSD50K (Fonseca et al., 2022), Clotho (Drossos et al., 2019), and MusicCaps (Agostinelli et al., 2023b), combined with tools such as Pyroomacoustics (Scheibler et al., 2017) and standardized HRTFs (Audio Engineering Society, 2015), provide reproducible pipelines, though still narrower in scope than large semantic datasets.

Training Paradigms Scaling and alignment methods from general-purpose LLMs also inform spatial modeling. MoE architectures (Fedus et al., 2022; Lepikhin et al., 2020) allow efficient specialization, while reinforcement learning techniques such as GRPO (Shao et al., 2024) refine reasoning and output formats. However, their application to spatial audio remains at an early stage.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND DESIGN PRINCIPLES

We address the task of conditional text generation from binaural audio input. Given a two-channel audio waveform $x \in \mathbb{R}^{B\times 2\times T}$ and a natural language prompt p, the model's objective is to generate a textual response y. To enable the model to comprehend not only the semantic content ("what") but also the spatial arrangement ("where") of the acoustic scene, we explicitly decompose the audio representation into two distinct components: a semantic embedding $z_{\text{sem}}(x)$ and a spatial embedding $z_{\text{sp}}(x)$. These representations are conditionally fused before being processed by the language decoder:

$$y \sim p_\theta\big(\ \cdot\ \big|\ \text{fuse}(z_{\text{sem}}(x),\ z_{\text{sp}}(x)),\ p\big).$$

Our framework is built upon three core design principles:

- **Decouple-and-Fuse.** Semantic and spatial attributes are modeled by independent, robust encoders. Their outputs are then adaptively combined via a supervised, conditional routing mechanism. This prevents representational entanglement, where the learning of one attribute might corrupt the other.
- **Minimally Invasive LLM Interfacing.** We avoid complex prompt templates and modifications to the LLM's native tokenization or positional encoding schemes. The final audio representation is simply appended to the prompt embeddings as a suffix, ensuring maximum compatibility with standard autoregressive frameworks.
- **Progressive Curricular Optimization.** We employ a multi-stage training curriculum that proceeds from representation learning to alignment and finally to policy optimization. This disentangles conflicting learning objectives and stabilizes the training of the composite model.

3.2 MODEL ARCHITECTURE

Our architecture, depicted in Figure 1, implements our "Decouple-and-Fuse" principle. The process begins as the input binaural audio is fed into two parallel backbones: a Semantic Encoder to extract a content representation (W) and a Spatial Encoder for locational cues (S'). These decoupled representations are then dynamically fused by our conditional MoE module. The MoE's router is a key component, conditioned on both a global audio context and the user's prompt intent, which allows it to intelligently route features to specialized experts. This process yields a unified audio embedding (H_{enc}) that is appended to the prompt embeddings and passed to the LLM decoder for text generation.

Semantic and Spatial Encoders. For semantic feature extraction, we use a frozen Whisper encoder. To specifically isolate content, the binaural input is downmixed to mono by averaging the left and right channels. This process yields a semantic sequence $W \in \mathbb{R}^{B\times L\times 768}$.

For spatial features, we adapt the dedicated encoder from (Wu et al., 2025). It processes the complex STFT of the binaural signal to preserve phase information crucial for localization. The model alternates between frequency-band modeling and temporal modeling using self-attention. Critically, an attractor-based aggregation mechanism allows it to handle a variable number of sound sources. The final output is projected and resampled into a spatial feature sequence $S' \in \mathbb{R}^{B\times L\times 768}$, aligned with the semantic features.

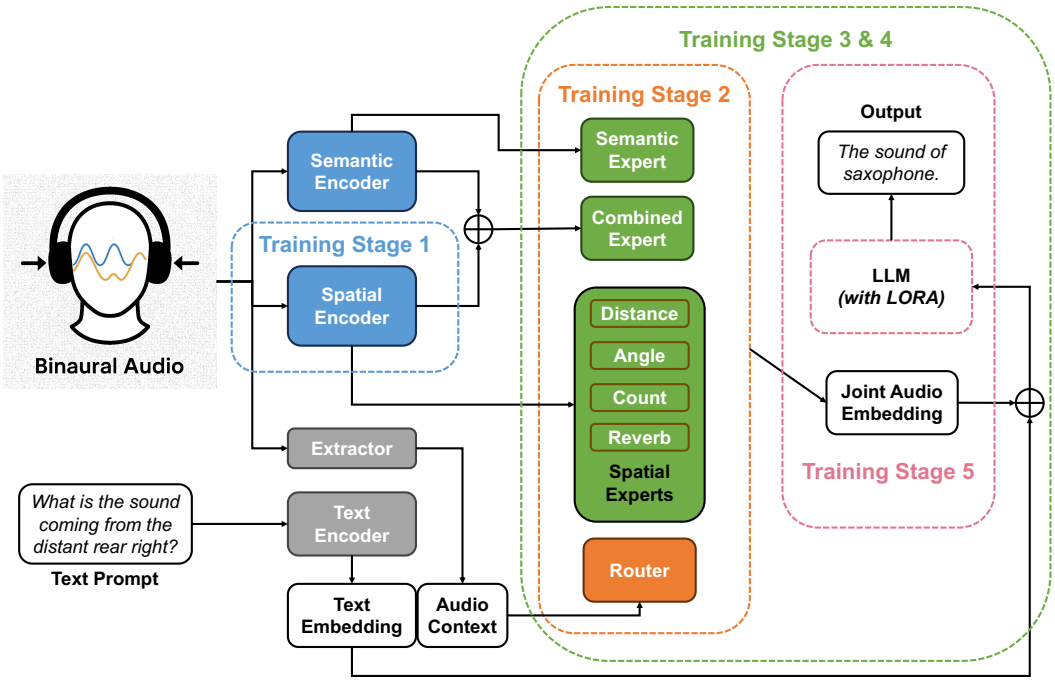

Figure 1: Model Architecture

Conditional MoE. The MoE module serves as the dynamic bridge between the decoupled representations and the decoder. Its design features two key innovations: expert specialization and hybrid conditional routing.

1. Expert Specialization. The module comprises six experts with distinct roles:

- **One Semantic Expert** (E_{wh}): Exclusively processes the semantic representation W.
- **Four Parallel Spatial Experts** ($\{E_{\text{sp}}^{(k)}\}_{k=1}^{4}$): Exclusively process the spatial representation S'. These experts are pre-oriented during their training towards distinct spatial attributes, namely direction, distance, room acoustics (reverberation), and source count/separation. This specialization enhances model interpretability and facilitates targeted supervision.
- **One Combined Expert** (E_{comb}): This expert is always active and does not participate in the dynamic routing. It receives the element-wise sum of the semantic and spatial representations, $C = W + S'$. We term C a Combined Steady Reference, as it provides the model with a stable, unweighted snapshot of the entire audio scene, which helps mitigate statistical drift and stabilize the learning process, especially in early training stages.

2. Hybrid Conditional Routing. The routing mechanism is conditioned on both the text prompt and the audio input. The conditioning signals are:

- **Audio Context** $c(x) \in \mathbb{R}^{256}$: A compact vector extracted directly from the raw waveform via a lightweight 1D CNN, global pooling, and an MLP. It captures global acoustic properties like loudness, reverberation, and source sparsity.
- **Prompt Intent** $\bar{e}(p) \in \mathbb{R}^{d_{\text{dec}}}$: A vector derived by mean-pooling the word embeddings of the input prompt p. It encodes the user's task intention (e.g., a query about "left" vs. "what sound").

These two vectors are concatenated and passed through an MLP to produce a 5-dimensional gating vector α:

$$\alpha = \sigma(\text{MLP}([\, c(x)\,;\, \bar{e}(p)\,])) \in (0,1)^5.$$

The five dimensions correspond to the one semantic and four spatial experts. We use a per-expert Sigmoid activation rather than a Softmax, enabling the simultaneous activation of multiple experts

for complex queries that require both semantic and spatial reasoning. The outputs of the gated experts, M, are then fused with the output of the always-active combined expert via averaging to produce the final hidden state H_{enc}:

$$M = \alpha_{\text{wh}} E_{\text{wh}}(W) + \sum_{k=1}^{4} \alpha_k \, E_{\text{sp}}^{(k)}(S'), \quad H_{\text{enc}} = \tfrac{1}{2}(M + E_{\text{comb}}(C)) \, .$$

LLM Interface. We employ the Qwen2-Audio 7B model as the language decoder. Its core parameters are kept frozen, and we use Low-Rank Adaptation (LoRA)(Hu et al., 2022) for parameter-efficient fine-tuning, applied only to the attention projection matrices (e.g., `q_proj`, `v_proj`). Following our principle of minimal intrusion, the audio hidden state H_{enc} is directly appended to the prompt embeddings $E(p)$, forming the input sequence $[\, E(p) \,;\, H_{\text{enc}} \,] \in \mathbb{R}^{B \times (L_p + L) \times d_{\text{dec}}}$. This simple and robust protocol eliminates the need for prompt engineering.

3.3 Learning: A Progressive Curriculum from Representation to Policy

To disentangle competing learning objectives, we adopt a five-stage curriculum. The input protocol and tokenization remain consistent across all stages to prevent training-inference mismatch.

Stage 1: Encoder Pretraining. In this preparatory stage, only the spatial encoder is trained. The objective is to learn a robust and general-purpose representation of spatial acoustics from a large-scale (potentially unlabeled) binaural audio dataset. This endows the encoder with a strong inductive bias for physical acoustic properties, decoupling this representation learning from the more complex cross-modal alignment task.

Stage 2: Router & Experts Training (Alignment I). The audio encoders and the LLM are frozen, and only the MoE module (experts and router) is trained. The loss function comprises two terms: the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ on the decoder output, and a router supervision loss $\mathcal{L}_{\text{router}}$. The supervision signal for the router is derived from task metadata (e.g., classifying a query as spatial, semantic, etc.). We employ teacher-forcing for the router weights α and add a light sparsity regularizer (e.g., ℓ_1 norm) to prevent expert averaging. The goal is for the model to learn when to rely on semantic, spatial, or combined cues.

Stage 3: SFT 1.0 (Alignment II). The encoders remain frozen, while the MoE module and the decoder's LoRA weights are trained jointly. The objective is to minimize a combined loss, $\mathcal{L}_{\text{CE}} + \lambda_{\text{r}} \mathcal{L}_{\text{router}}$. This stage focuses on bridging the modality gap between the fused audio representations and the LLM's latent space, refining the cross-modal interface.

Stage 4: SFT 2.0 (Formatting). In this stage, the encoders remain frozen, while the MoE module and the decoder's LoRA weights continue to be trained jointly. The key change is the removal of the router supervision loss $\mathcal{L}_{\text{router}}$. The optimization is driven solely by the cross-entropy loss $\mathcal{L}_{\text{CE}}$. The purpose is to transition the MoE from explicit supervision to end-to-end learning. The router's policy is now shaped exclusively by which expert combination ultimately leads to better text generation, fostering a co-evolution of the routing strategy and the desired output format (e.g., `<think>...</think><answer>...</answer>`).

Stage 5: GRPO Preference Optimization. Finally, only the LoRA weights are updated using GRPO. For each sample, we generate a group of m candidate responses $\{y_i\}_{i=1}^{m}$ from the current policy π_θ. Each response is scored by a reward function:

$$R(y) = R_{\text{fmt}}(y) + R_{\text{ans}}(y).$$

Here, R_{fmt} assesses structural compliance (e.g., proper tag closure), while R_{ans} measures correctness by parsing the content within the `<answer>` tag. GRPO uses the in-batch group mean as a baseline to compute the advantage:

$$A_i = \frac{R(y_i) - \frac{1}{m} \sum_{j=1}^{m} R(y_j)}{\text{std}_{j=1}^{m}(R(y_j))}$$

The policy is then updated by maximizing the following objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{(x,p),\{y_i\}\sim\pi_\theta}\left[\sum_{i=1}^{m} \log \pi_\theta(y_i|x,p) \cdot \text{stop\_gradient}(A_i)\right].$$

Keeping the encoders and MoE frozen during this stage confines the optimization to the small set of LoRA parameters, ensuring stability and efficiency.

3.4 Inference and Practical Considerations

- **Inference Protocol**: The inference process strictly follows the training protocol, using standard autoregressive decoding on the concatenated input embedding $[\, E(p)\,;\, H_{\text{enc}}\,]$.
- **Router Control**: The Sigmoid router allows for flexible control at inference time. A temperature parameter τ can be used to soften the expert weights, enabling a more robust "soft routing" mode. Alternatively, a threshold ϵ can be applied to enforce "hard routing" by activating only the top experts, reducing computational cost on resource-constrained platforms.
- **Training Stability**: To prevent mode collapse in the router, we incorporate two regularization techniques during training: a slow-varying regularization to penalize drastic changes in α between consecutive steps, and a load-balancing loss to encourage all experts to be utilized over the course of training.
- **Computational Efficiency**: The MoE design is efficient, as the experts can be computed in parallel along the channel dimension of the tensors. The primary latency stems from the STFT front-end and attention layers in the spatial encoder, both of which can be optimized for streaming applications using chunk-wise processing and state caching.
- **Output Interpretability**: The structured output provides significant benefits. The content of the `<answer>` tag can be directly parsed for automated evaluation. More importantly, the `<think>` tag offers a trajectory of the model's spatial reasoning, providing valuable insights for error analysis and future alignment with human preferences.

In summary, our methodology combines a decoupled representation with a progressive training curriculum to efficiently instill spatial reasoning into a pre-trained LALM.

4 Experiments

This section details the experimental setup designed to validate our proposed framework. We first describe our data generation pipeline and the construction of our comprehensive benchmark. We then present a detailed analysis of the main results, highlighting the effectiveness of our progressive training curriculum and discussing key findings from our evaluation.

4.1 Datasets and Simulation Pipeline

Lacking suitable public datasets for spatial audio reasoning, we developed a scalable simulation pipeline to generate data for our SFT and GRPO stages. This pipeline combines generation of physically-principled BRIRs with synthesis of complex acoustic scenes.

Data for SFT. The SFT phase utilizes two data formats. The initial SFT stage employs open-ended question-answer pairs, where each sample includes an audio path, a textual instruction, a ground-truth answer, and a `router_label` to supervise the MoE routing mechanism. A subsequent stage, SFT2.0, uses data formatted as multiple-choice questions. The answers in this stage are structured with a chain-of-thought rationale enclosed in `|<think>|...|</think>|` tags, followed by the final choice in `|<answer>|...|</answer>|` tags. This prepares the model for the structured output format required during the final RL phase. For all SFT data, audio is resampled to 44.1 kHz and standardized to a fixed length of 5 seconds by padding or truncation.

Data for GRPO. The GRPO phase uses a benchmark of multiple-choice questions designed to probe a wide range of spatial audio understanding abilities. Each data sample contains an audio path, a question, a set of options (e.g., A, B, C, D), and the correct answer key. During training, the question and options are concatenated into a single prompt.

Spatial Audio Simulation Pipeline.

- **BRIR Generation.** We create a large and diverse library of BRIRs using the `pyroomacoustics` simulator (Scheibler et al., 2017). We model shoebox-shaped rooms with varying dimensions and wall absorption coefficients, corresponding to low, medium, and high reverberation levels. For each simulated room, we randomly place a receiver and 30 candidate source positions. The receiver's head-related impulse responses (HRIRs) are sourced from the empirical FABIAN dataset (Brinkmann et al., 2017), ensuring perceptual realism. The resulting BRIR for each source-receiver pair is calculated and stored, along with comprehensive metadata describing the room acoustics and spatial geometry.
- **Scene Synthesis.** During training, we dynamically generate binaural audio samples. Dry audio clips are sampled from a collection of large-scale public datasets, including FSD50K for general sound events (Fonseca et al., 2022), Clotho v2 for environmental diversity (Drossos et al., 2019), MusicCaps for musical scenarios (Agostinelli et al., 2023a), and Emilia for multilingual speech (He et al., 2024). These dry sounds are convolved with the pre-computed BRIRs from a chosen scene, scaled by a random gain, and summed to create a multi-source, spatially coherent binaural mixture. This approach allows us to generate a virtually infinite amount of training data with precise ground-truth labels for source locations, event classes, and room characteristics.

Further details on data and simulation parameters are provided in Appendix A.

4.2 Training and Optimization Setup

Our training regimen follows the progressive curriculum outlined in Section 3, beginning with supervised fine-tuning and culminating in reinforcement learning.

Model Configuration. We use LoRA for parameter-efficient fine-tuning (PEFT), with a rank of $r = 8$ and $\alpha = 32$. To further reduce memory, we employ 4-bit NormalFloat (NF4) quantization via QLoRA(Dettmers et al., 2023).

Optimization and Hyperparameters. We use the AdamW optimizer across all stages. The learning rate is set to 5×10^{-5} for SFT stages and 1×10^{-5} for the GRPO stage. We use a batch size of 1 per device with gradient accumulation steps of 2 for SFT and GRPO, and 4 for SFT2.0. The training is conducted for 100, 50, and 5 epochs for SFT, SFT2.0, and GRPO, respectively. All training stages utilize a warmup ratio of 0.1 and gradient clipping at a norm of 1.0.

Reinforcement Learning with GRPO. In the final stage, we refine the model's policy using GRPO, implemented with the TRL library. For each prompt, we generate three candidate responses from the current policy. The reward function is a composite of a format reward, which encourages adherence to the `|<think>|...|</think>|...|<answer>|...|</answer>|` structure, and a result reward, which grants a bonus if the extracted answer matches the ground truth. We do not use a separate reference model during GRPO training (`ref_model=None`).

System and Distributed Training. All experiments were conducted with 8 NVIDIA RTX 5090 GPUs (32 GB VRAM each). We leverage `torchrun` for distributed training, utilizing the DeepSpeed ZeRO stage 2 strategy (Rajbhandari et al., 2020) with CPU-offloading for the optimizer to enhance memory efficiency. Our entire framework is built upon the Hugging Face Transformers library (Wolf et al., 2020).

4.3 Benchmark Construction

To rigorously evaluate spatial reasoning, we constructed a 1,000-question multiple-choice benchmark using an LLM-powered, semi-automated pipeline to ensure quality and diversity. The process involves several quality-controlled steps:

- **Scene Generation:** For each simulated acoustic scene, we first generate a detailed, text-based description of the scene's contents and spatial layout.

- **Question Generation:** Using task-specific prompts and the generated scene descriptions, the LLM creates questions that target three core competency areas: Perception (e.g., source counting, content identification), Integration (e.g., binding sounds to locations), and Reasoning (e.g., inferring spatial relationships between sources).
- **Quality Control:** We employ several checks, such as ensuring that questions about spatial relationships are only generated for scenes with two or more sources. After generation, the order of options is randomly shuffled to prevent positional bias, with the answer key updated accordingly.

This benchmark not only serves as the evaluation set but also as the foundation for our SFT2.0 "teacher" data. By prompting the LLM to provide a step-by-step reasoning process (`<think>`) and a final answer (`<answer>`) for each benchmark question, we automatically generate high-quality data for teaching the model the desired output format. Prompt details are provided in Appendix B.

Our generation pipeline successfully produced a benchmark that aligns with our target distribution across the three core competency areas. The final 1,000-question set consists of 32.1% Perception tasks (e.g., identifying sound content, counting sources, and recognizing environmental acoustics), 28.5% Integration tasks (e.g., binding sounds to specific locations and associating acoustic attributes), and 39.4% Reasoning tasks (e.g., inferring spatial relationships between sources, counterfactual reasoning, and summarizing the scene). This balanced distribution ensures a comprehensive evaluation of both the model's foundational perception abilities and its advanced reasoning capabilities.

4.4 Main Results

Our curriculum constitutes a staged ablation where each phase contributes a distinct capability. Moving from base alignment (SFT 1.0) to full instruction tuning (SFT 2.0) and finally to preference optimization (GRPO), overall accuracy rises from 25.10% to 50.10% and reaches 61.10% on our $1{,}000$-sample benchmark (95% CIs: 22.5–27.9%, 47.0–53.2%, 58.0–64.1%). We group tasks into three competency areas—Perception (3 tasks, $n = 321$), Integration (3 tasks, $n = 285$), and Reasoning (4 tasks, $n = 394$). Micro-averaged accuracies improve monotonically across all areas, with consistent trends under macro averages. Effect-size analysis shows a small-to-moderate gain from SFT 2.0 to GRPO (Cohen's $h = 0.22$) and a large jump from SFT 1.0 to SFT 2.0 ($h = 0.52$).

Table 1: Ablation study of our progressive training curriculum. Accuracy (%) is reported for each stage across three core competency areas and overall. Each stage builds upon the previous one, showing significant gains, particularly in Reasoning after GRPO.

Competency Area	SFT 1.0	SFT 2.0	GRPO
Perception	34.89	48.60	**59.50**
Integration	18.60	32.28	**46.67**
Reasoning	21.83	64.21	**72.84**
Overall Accuracy	25.10	50.10	**61.10**

Our curriculum constitutes a staged ablation where each phase contributes a distinct capability. Moving from base alignment (SFT 1.0) to full instruction tuning (SFT 2.0) and finally to preference optimization (GRPO), overall accuracy rises from 25.10% to 50.10% and reaches 61.10% on our $1{,}000$-sample benchmark (95% CIs: 22.5–27.9%, 47.0–53.2%, 58.0–64.1%). We group tasks into three competency areas—Perception (3 tasks, $n = 321$), Integration (3 tasks, $n = 285$), and Reasoning (4 tasks, $n = 394$). Micro-averaged accuracies improve monotonically across all areas, with consistent trends under macro averages. Effect-size analysis shows a small-to-moderate gain from SFT 2.0 to GRPO (Cohen's $h = 0.22$) and a large jump from SFT 1.0 to SFT 2.0 ($h = 0.52$).

Preference optimization yields the largest gains on complex reasoning. The scene summarization reasoning task reaches $98/98 = 100.0\%$ under GRPO, with a 95% confidence interval of 96.2–100.0%, and causal intent reasoning attains 84.38%. A plausible explanation is that preference optimization explicitly rewards coherent multi-step analyses that are only weakly elicited during supervised fine-tuning.

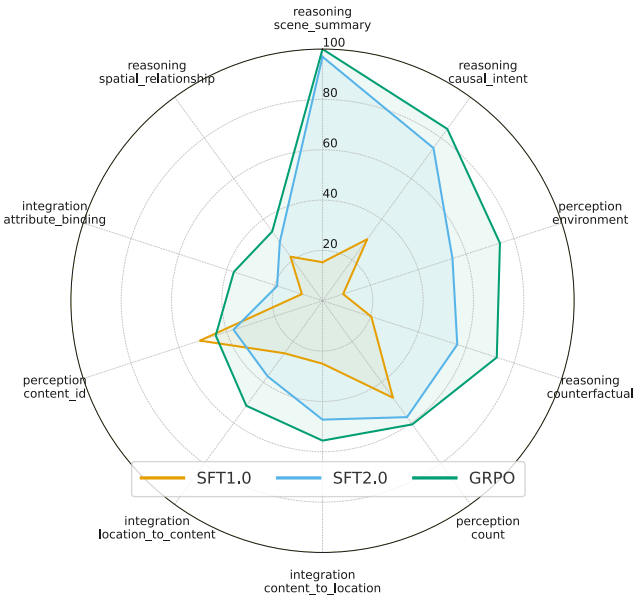

Figure 2: Task-wise accuracies on the ten benchmark categories for SFT1.0, SFT2.0, and GRPO.

Despite the overall gains, two categories remain challenging under GRPO: spatial-relationship reasoning at 34.02% and attribute binding integration at 37.07%. Manual inspection indicates a frequent pattern in which the model perceives individual events and locations correctly but fails during the relational aggregation step; for instance, it may infer several distances correctly yet choose the wrong object as the farthest. Targeted training signals that impose relative order constraints, along with contrastive objectives that tighten the association between objects and their attributes, are promising directions.

We also observe a decrease on the content identification task, where SFT 1.0 at 51.24% surpasses SFT 2.0 at 37.19% and GRPO at 44.63%. We interpret this as an alignment tax: emphasizing long-form reasoning can sometimes harm direct identification for simple inputs. To mitigate this trade-off, one can incorporate routing that sends simple cases to a direct-answer head while reserving chain-of-thought style decoding for complex cases, or apply minimum-reasoning regularization and early-exit mechanisms that discourage unnecessary deliberation on easy instances. Finally, because the scene summarization reasoning task exhibits a ceiling effect under GRPO, we plan to include harder variants with longer contexts to better probe headroom and to rule out unintended shortcuts. Additional case studies are provided in Appendix C.

5 Conclusion

In this work, we addressed the critical limitation of existing Large Audio-Language Models—their inability to perceive spatial audio—by introducing **The World is Not Mono (TWNM)**, a comprehensive framework for spatial audio understanding. We demonstrated that a combination of large-scale synthetic binaural data, a task-aware MoE architecture, and a progressive SFT-to-GRPO curriculum can successfully endow LALMs with robust spatial reasoning. Our ablation studies confirmed the value of each training stage, with preference optimization proving especially effective for complex reasoning tasks. By moving beyond the mono paradigm, our work establishes a clear pathway for three-dimensional auditory intelligence. Future work will focus on enhancing the model's generalization to diverse in-the-wild recordings and extending our framework to support multi-channel formats like Ambisonics, further broadening its applications in embodied AI, augmented reality, and assistive technologies.

6 ETHICS STATEMENT

All authors of this paper have read and adhered to the ICLR Code of Ethics. This research does not involve human subjects, personally identifiable information, or sensitive data. The datasets utilized are publicly available and have been handled in accordance with their specified licenses. While we have strived to ensure the fairness and robustness of our methods, we acknowledge that potential biases may exist in the underlying data, which could be reflected in the model's performance. We believe the potential societal benefits of this work in advancing the understanding of audio scene analysis outweigh the foreseeable risks. We declare no conflicts of interest.

7 REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our research. All datasets used in our experiments are publicly available. For the review period, we include anonymized supplementary materials containing core implementation modules and configuration files to facilitate inspection of our method. A fully runnable codebase with scripts and instructions will be released upon acceptance.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalan Borsos, Jesse Engel, Mauro Verzetti, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023a.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023b. URL https://arxiv.org/abs/2301.11325.

Audio Engineering Society. AES69-2015: Aes standard for file exchange—spatial acoustic data file format. AES Standard, 2015. URL https://www.aes.org/publications/standards/search.cfm?docID=96.

Fabian Brinkmann, Alexander Lindau, Stefan Weinzierl, Gunnar Geissler, Steven van de Par, Markus M"uller-Trapet, and Rogier Opdam. The fabian head-related transfer function data base. DepositOnce, Technische Universit"at Berlin, 2017. URL https://depositonce.tu-berlin.de/items/3b423df7-a764-4ce1-9065-4e6034bba759.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. URL https://arxiv.org/abs/2407.10759.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. Learning spatially-aware language and audio embeddings, 2024. URL https://arxiv.org/abs/2409.11369.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset, 2019. URL https://arxiv.org/abs/1910.09387.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022. URL https://arxiv.org/abs/2101.03961.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: An open dataset of human-labeled sound events, 2022. URL https://arxiv.org/abs/2010.00475.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. *arXiv preprint arXiv:2407.05361*, 2024. doi: 10.48550/arXiv.2407.05361.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2022.

Rongjie Huang et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*, 2023. URL https://arxiv.org/abs/2304.12995.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding, 2020. URL https://arxiv.org/abs/2006.16668.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. *arXiv preprint arXiv:1910.02054*, 2020.

Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulations and array processing algorithms. *arXiv preprint arXiv:1710.04196*, 2017. URL https://arxiv.org/abs/1710.04196.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and Yuki Mitsufuji. Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection, 2021. URL https://arxiv.org/abs/2010.15306.

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events, 2023. URL https://arxiv.org/abs/2306.09126.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024. URL https://arxiv.org/abs/2310.13289.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2020.

Donghang Wu, Jiaqi Du, Tianshu Qu, Qingbo Huang, and Dejun Zhang. Moving sound source localization and tracking based on envelope estimation for unknown number of sources. 05 2025.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of ACL*, 2024. URL https://aclanthology.org/2024.acl-long.109.pdf.

Yuhuan You, Yufan Qian, Tianshu Qu, Bin Wang, and Xueyang Lv. Spherical harmonic beamforming based ambisonics encoding and upscaling method for smartphone microphone array. In *Audio Engineering Society Convention 158*. Audio Engineering Society, 2025.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. Bat: Learning to reason about spatial sounds with large language models, 2025. URL https://arxiv.org/abs/2402.01591.

APPENDIX

A DATA SIMULATION AND BENCHMARK PARAMETERS

Our data generation pipeline is governed by a set of parameters designed to create diverse and realistic acoustic environments. The target scale for our generated data includes 10,000 unique acoustic scenes and a benchmark of 1,000 distinct questions. Key parameters are summarized below.

Room Dimensions: Rooms are sampled from three size categories with uniform probability within each range:

- Small: $x \in [3, 5]$ m, $y \in [4, 6]$ m, $z \in [2.5, 3.5]$ m
- Medium: $x \in [8, 12]$ m, $y \in [10, 15]$ m, $z \in [3, 5]$ m
- Large: $x \in [20, 30]$ m, $y \in [25, 35]$ m, $z \in [10, 15]$ m

Reverberation / wall absorption. We use absorption coefficient $\alpha \in [0, 1]$ where $\alpha = 0$ is perfectly reflective and $\alpha = 1$ is perfectly absorptive. To simulate different RT60 regimes we sample:

- High reverberation (reflective walls): $\alpha \in [0.05, 0.25]$,
- Medium reverberation: $\alpha \in [0.25, 0.5]$,
- Low reverberation (absorptive walls): $\alpha \in [0.5, 0.95]$.

Benchmark Task Distribution: The question generation process targets the following distribution across competency areas: Perception (30%), Integration (30%), and Reasoning (40%).

B PROMPT FOR BENCHMARK GENERATION

Prompt (for generating open-ended spatial QA)

You are a top-tier AI course designer preparing graduation exam questions for an advanced spatial audio model. This model has already mastered the basics of recognizing sound content (semantic), localizing sources (localization), perceiving the environment (acoustics), and counting (count).
Your task: given a `<scene_description>`, design 1–2 **open-ended**, complex questions that require integrating multiple abilities, and provide detailed, fluent, high-quality answers.
[Design Principles]
- **Reject simplicity**: do not ask "Where is the dog?"; instead ask "What is the main activity in the scene, and where does it occur relative to me in space?"
- **Encourage reasoning**: propose questions that require inference from multiple cues. For example: "Based on the room's reverberation and the sounds inside, what is the most plausible type of place?"
- **Simulate dialogue**: both questions and answers should read like a natural conversation between humans.
[Example] `<scene_description>`
Indoors with slight reverberation, two sounds are audible. One is keyboard typing from straight ahead at a close distance. The other is birdsong coming from outside the window, located up-right and farther away.
`</scene_description>`
`<generated_data>`
[
{
"instruction": "Based on what you hear, can you infer where I might be and what I'm doing?",
"answer": "It sounds like you're likely in a room—perhaps an office or at home. The clear, nearby keyboard typing suggests you're working or studying on a computer. Meanwhile, the birdsong from outside implies it's likely daytime and that you're close to a window."
}
]
`</generated_data>`
[Formal Task] `<scene_description>`
{{ insert RTSD here }}
`</scene_description>`
`<generated_data>`
`</generated_data>`

C CASE STUDY

To further illustrate the reasoning ability of our model, we present a representative case study drawn from the evaluation set. The task requires the model to infer the most plausible explanation of an acoustic scene based on the perceived spatialized audio.

Ground Truth (Reference)

Scene Description:
– **Scene type:** Mixed, single sound source
– **Sound event:** Respiratory sound (breathing)
– **Azimuth:** 317°, **Elevation:** -13°, **Distance:** 17.4 m
– **Room acoustics:** Anechoic, absorption coefficient ≈ 0.99, RT60 = 0.0 s

Correct Answer:
A. The listener may be using headphones to hear a specially recorded or processed audio, in which a breathing sound is precisely placed at a specific position in the sound field.

Question

Task: Based on the audio you hear, which of the following explanations is the most reasonable for this scene?

- A. The listener may be using headphones to hear a specially recorded or processed audio, in which a breathing sound is precisely placed at a specific position in the sound field.
- B. In a wide-open space, someone is breathing about 17 meters away from the listener, and the openness of the space makes the sound unusually clear.
- C. The listener is in a completely silent room, and due to over-concentration, hallucinates the distant sound of breathing.
- D. This is a live recording in a large concert hall, where a breathing sound from an audience member was mistakenly amplified during post-production.

Model Output (verbatim)

|<think>| The core of the question is to infer the most reasonable scene based on the heard sound.
First, I analyze the key sound elements in the audio. I heard a clear breathing sound, coming from my front-left, about 17 meters away. The sound is very clear, almost without echo, which indicates I am in an open environment, such as a large space. The sound position is very precise, at azimuth 313 degrees and elevation 1 degree, which shows that the sound source is fixed rather than moving randomly.
Based on these observations, I evaluate each option:
– Option A states that the breathing sound was placed at a specific position. This matches exactly with the sound location I heard (azimuth 313 degrees). A processed audio, where a sound is placed at a specific position in the recording, is entirely possible.
– Option B describes a wide-open space with someone breathing. This contradicts the clear, echo-free sound characteristics. A wide-open space would produce noticeable reverberation.
– Option C suggests that the sound is a hallucination from silence. This is clearly wrong because I distinctly heard a breathing sound.
– Option D suggests a concert hall live recording where the breathing was over-amplified. However, a concert hall should have noticeable reverberation, which is absent here, so D is not reasonable.
In summary, Option A best fits the sound characteristics: a clear, precisely located breathing sound placed in a large space.|</think>|
According to the audio analysis, the sound comes from a precise location in a nearly echo-free large space, indicating that it is likely a processed audio rather than a real-world event.
|<answer>| A |</answer>|

D STATEMENT ON THE USE OF LARGE LANGUAGE MODELS (LLMS)

In the preparation of this manuscript, LLMs were utilized for two distinct purposes. First, as a general-purpose tool for tasks such as grammar correction and language polishing of the manuscript text. Second, as a core component of our methodology, LLMs were employed in a semi-automated pipeline under human supervision to generate the questions and chain-of-thought rationales for our benchmark dataset, as detailed in Section 4.3. The core research ideation, experimental design, and the final analysis were conceived and executed by the authors. We confirm that the LLMs' role, while integral to the methodology, was that of a tool to implement our research design. The authors take full responsibility for all content presented in this paper, including the validity of the generated benchmark and any potential inaccuracies.