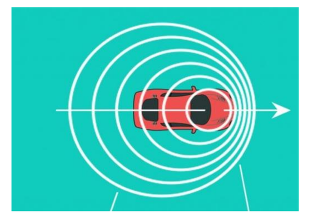
Teaching Physical Awareness to LLMs through Sounds

Weiguo Wang (NIO), Andy Nie (NIO, Peking University), Wenrui Zhou (NIO), Yi Kai (NIO), Chengchen Hu (NIO)

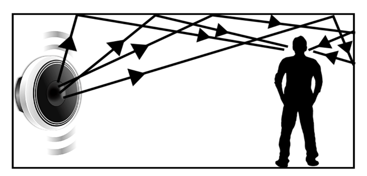
NIO · PEKING UNIVERSITY · ICML International Conference On Machine Learning

1. How Human Ears Understand Physical World ?



Doppler Effect

Identifies whether a car is approaching



Multipath Effect

Distinguishes indoor from outdoor environments

Binaural Hearing

Enables localization of sound sources

Sounds inherently carries rich physical information

2. Can Audio LLM Hears like Human Ears?

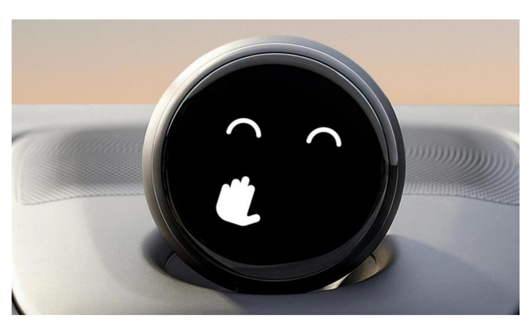
✕ No

While Audio LLMs perform well on speech content, they lack physical understanding

3. Why We Need Physical Understanding?

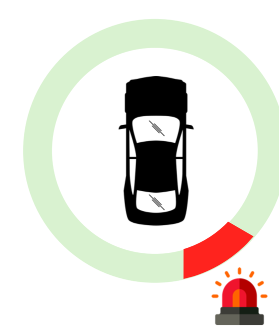
Voice-controlled Vehicle

Blocks unauthorized voice commands from outside the vehicle



Embodied AI Systems

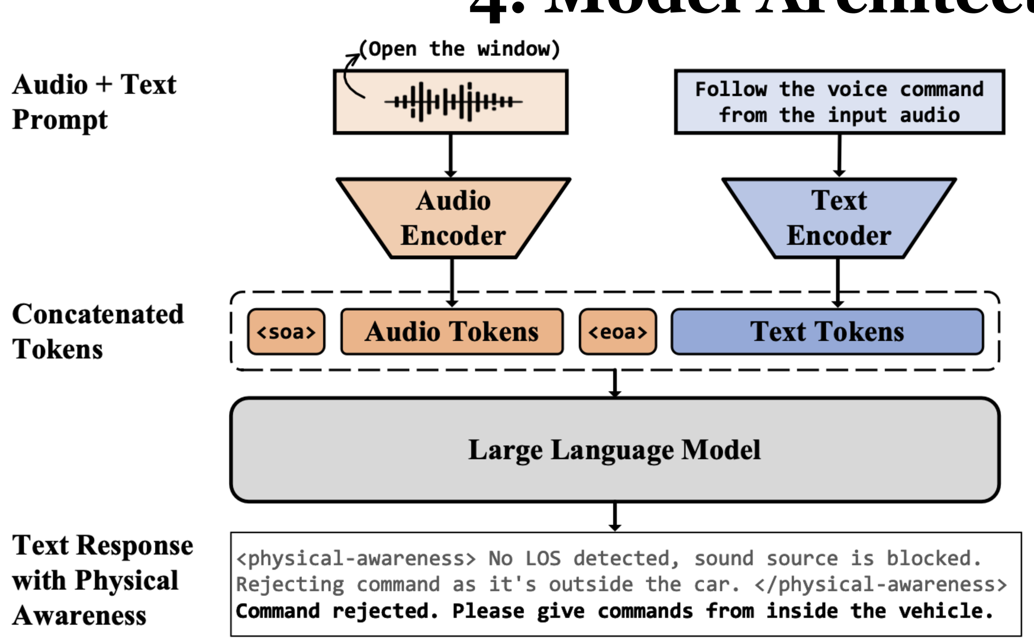
Uses sound localization to make systems more human-like



Siren Detection and Localization

Prevents "deaf driver" behavior, enhancing safety and awareness

4. Model Architecture



1. **Audio Encoder:** Converts raw audio into tokens
2. **Text Encoder:** Converts text input into tokens
3. **LLM:** Generates responses based on combined input

Following common practices, we adopt a common end-to-end architecture

5. Challenge I: Dataset Construction

How to collect and annotate a large-scale dataset ?

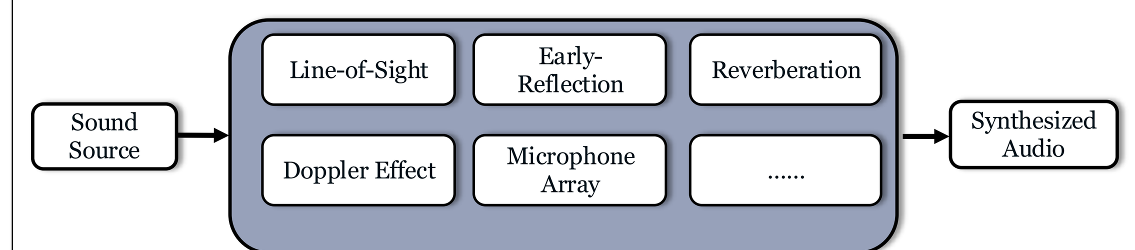
- **Data Collection ?** This requires extensive deployment of recording devices across various environments and conditions, which is expensive and not scalable
- **Data Annotation ?** Unlike text or images where humans can directly annotate content, audio physical cues cannot be labeled easily by humans.

Key Insight: The sound that we hear or microphones capture can be decomposed into two independent components:

$$y = h * x$$

- x the sound source
- h the physical channel through which it travels

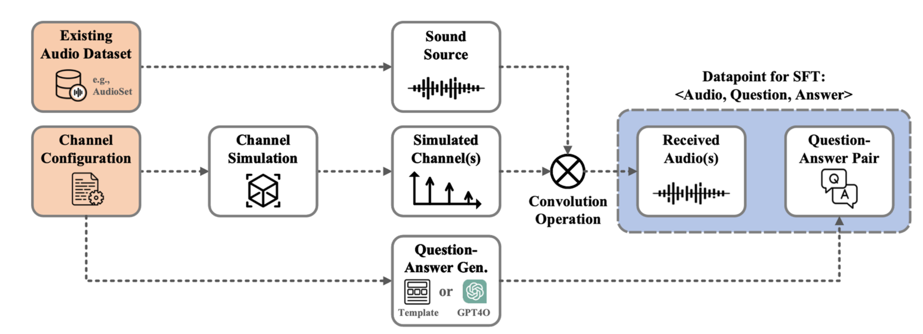
Solution: Synthesize audios by convolving real sounds with simulated channels



Acoustic Channel Simulator

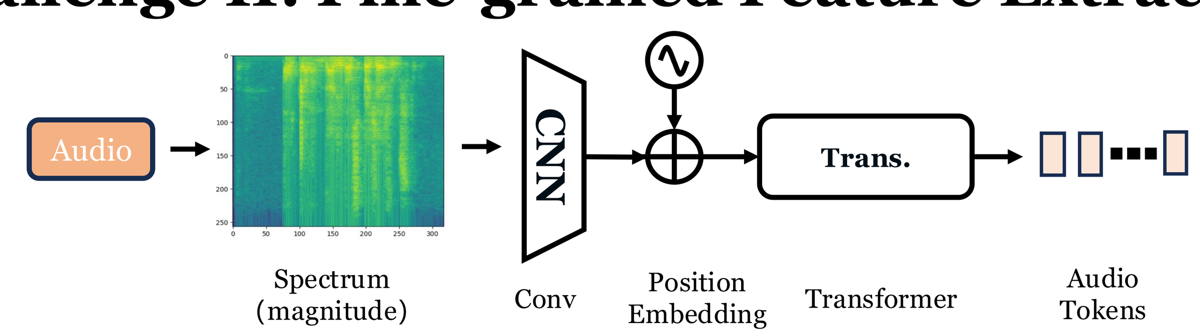
We develop a acoustic channel simulator, which models the full life cycle of sound — from emission to reception.

- We generate SFT data by convolving existing sounds with simulated channels.
- Using known simulation parameters, we automatically create aligned QA pairs.
- Each datapoint is a <Audio, Question, Answer> triplet, enabling grounded physical understanding.



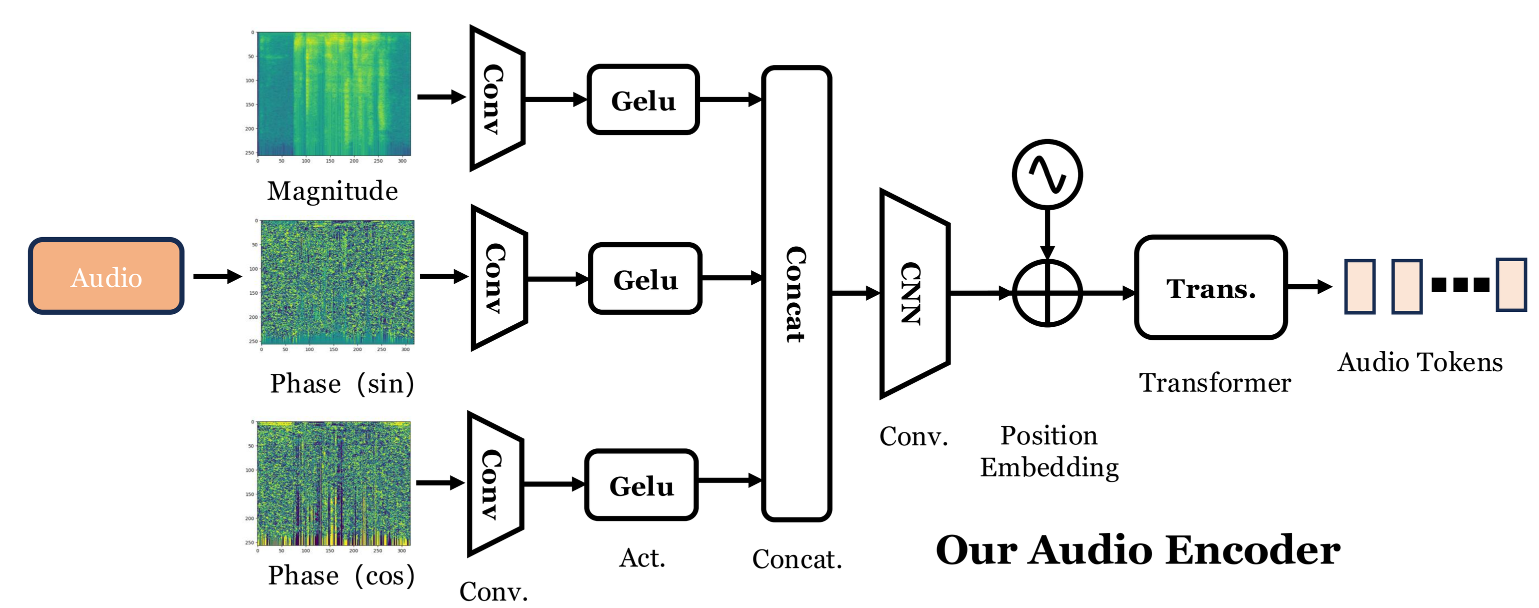
Synthesis Workflow of <Audio, Question, Answer>

6. Challenge II: Fine-grained Feature Extraction



Audio Encoder (OpenAI Whisper)

Problem: Audio encoders like Whisper fall short for physical understanding. Whisper mainly captures **magnitude features**, which work well for speech recognition—but lack the fine-grained phase information needed for physical cues



Our Audio Encoder

Solution: Our encoder incorporates both magnitude and phase (sin, cos) to retain physical characteristics of sound.

7. Main Results

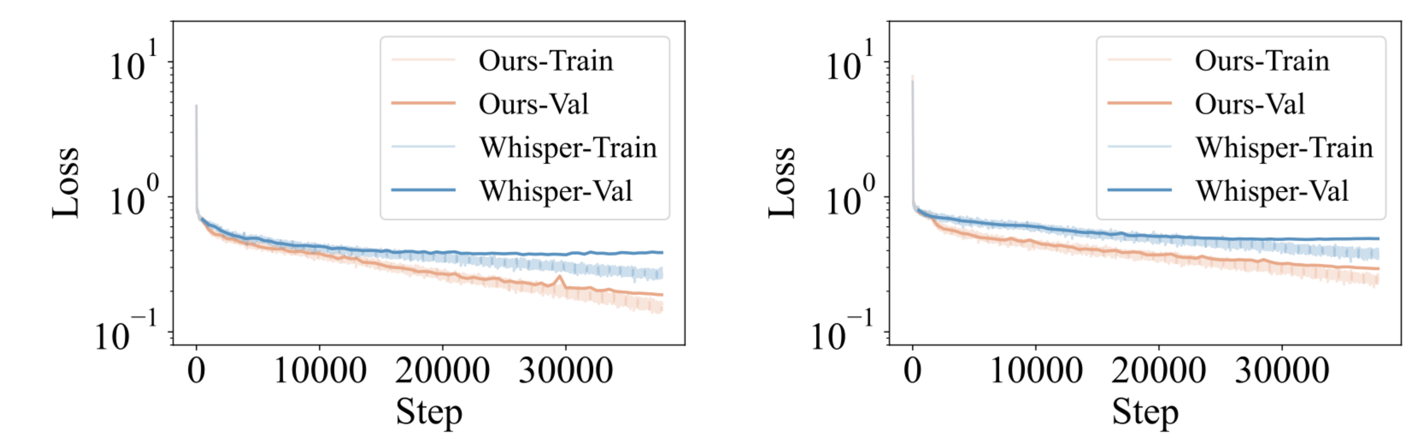
Table 2. Overall Performance. Values are presented as (Merged | Sole) where "Merged" indicates models trained on combined dataset and "Sole" indicates models trained separately for each task. By default, we focus on Merged results, with Sole results provided for reference.

Model Architecture		Task Performances (Merged | Sole)				
Audio Encoder	**LLM**	**LOS Detection** BCA (↑)	**Doppler Estimation** MAE_f (↓)	**DoA Estimation** MAE_t (↓)	**Multipath Analysis** TCA (↑)	**Range Estimation** REP (↓)
Whisper	Llama3.1-8B	0.867 | 0.906	1.213 | 3.147	5.585 | 5.601	0.845 | 0.889	12.572 | 17.182
	Qwen2-7B	0.881 | 0.910	1.042 | 0.575	2.716 | 6.873	0.848 | 0.897	10.609 | 12.901
ACORN	Llama3.1-8B	0.920 | **0.965**	0.791 | 0.557	1.423 | 1.349	0.890 | **0.945**	1.764 | **1.446**
	Qwen2-7B	**0.924** | 0.962	**0.181** | **0.263**	**0.907** | **1.167**	**0.903** | 0.944	**1.599** | 1.751
Performance on Open QA (Our Encoder + Qwen2-7B)		0.898 | 0.953	0.487 | 0.398	2.314 | 2.043*	0.906 | 0.908	2.852 | 1.900*
*Random Baseline***		0.50	10.00	66.67	0.33	33.33

We compare two audio encoders: OpenAI's Whisper and our encoder proposed. We pair each encoder with two different large language models (LLMs): Llama3- 8B and Qwen2 with 7B

Key Findings:

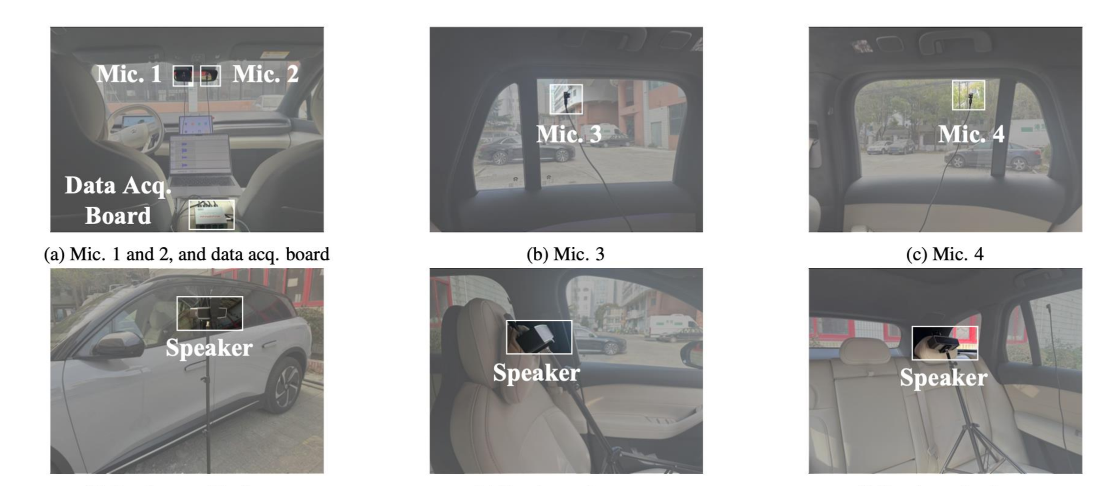
1. the feasibility of teaching LLMs to understand physical phenomena through sound
2. the superiority of our audio encoder over Whisper
3. the model-agnostic nature of our approach, evidenced by similar performance of different LLM architectures



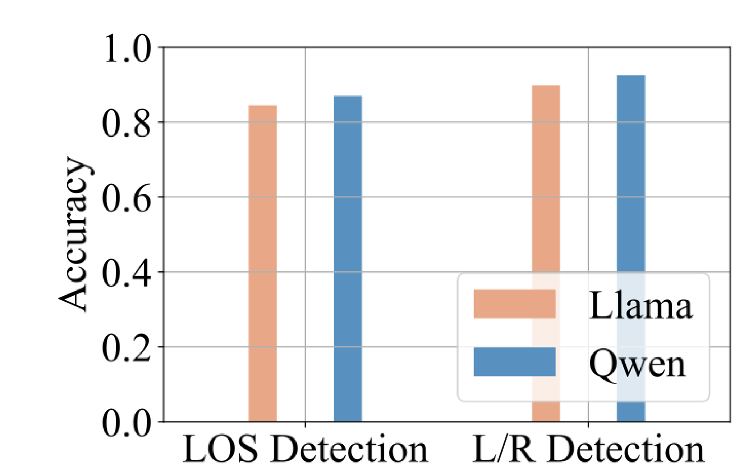
(a) Qwen2-7B (b) Llama3.1-8B

Loss History

- Our approach achieves faster convergence and lower final loss values during training across both Llama and Qwen architectures



Real-World Deployment



Results

- The results show the practical viability of our approach in the real world.

Contact Me:

wangwg.wwg@gmail.com

Homepage WeChat