

Late Days: Na Young (5), Karan (0), pooling 4 days for a 2 day late submission together.

Perceptual-Aware Spatial Scene Synthesis (PASSS)

Stanford CS224N {Custom}
TA mentor: Ali Sartaz Khan

Karan Soin

Department of Computer Science
Stanford University
ksoin@stanford.edu

Na-Young Son

Department of Computer Science
Stanford University
nyson25@stanford.edu

Abstract

Multimodal large language models for spatial audio have largely focused on discriminative tasks such as mapping audio signals to geometric coordinates [1], but these outputs fail to capture the psychoacoustic nuances required for immersive spatial computing. We propose PASSS (Perceptual-Aware Spatial Scene Synthesis), a framework that inverts this pipeline: given a natural language description of a sound event, PASSS generates structured spatial audio parameters as JSON. We fine-tune Qwen 2.5-7B with QLoRA on 200K examples from BEWO-1M, producing outputs organized into a three-layer hierarchy: (1) a Geometric Layer for spatial grounding (azimuth, elevation, distance), (2) an Analytic Perceptual Layer applying deterministic physical models for gain and interaural time difference, and (3) a Learned Perceptual Layer predicting gain residuals, lateral spread, and attention priority, which are quantities that capture how a sound should be perceived rather than simply where it is. For these quantities, no closed-form solution exists. On 500 test examples, PASSS achieves 100% valid JSON generation, 11.73° mean angular error, and 0.07 attention priority error on a 0 to 1 scale. We validate PASSS through perceptual evaluation on Apple Vision Pro, achieving a mean Jensen-Shannon divergence of 0.128 between rendered audio energy and listener gaze distributions, confirming that learned attention priorities correspond to actual human listening behavior.

1 Introduction

Spatial computing platforms like Apple Vision Pro have transformed the immersion of interactions with digital environments, yet the generation of perceptually accurate spatial audio is still unsolved. While the visual components of spatial computing have advanced rapidly, sound is still unable to truly mimic human auditory perception. Current multimodal large language models for spatial audio focus on discriminative tasks, like mapping audio signals to geometric coordinates [2, 1]. While these models succeed at basic localization, they treat sound as a static point in space, and fail to account for psychoacoustic factors such as perceived loudness decay, spatial extent, and attentional salience which drive human auditory experiences.

This gap exists because human hearing does not function like a geometric sensor. Listeners respond to simulated looming effects, Doppler-induced frequency shifts, and the perceived spatial extent of sources [3]. None of these are captured by coordinate prediction alone. A sound source at 50 meters may be rendered inaudible by a purely physical model, however a perceptually-aware system recognizes that the source’s narrative importance needs a gain boost of some kind to maintain audibility and guide listener attention.

We propose PASSS (Perceptual-Aware Spatial Scene Synthesis), a framework that inverts the traditional pipeline to perform text-to-spatial-parameter synthesis. By leveraging a fine-tuned Qwen 2.5-7B model with a three-layer parameter hierarchy, we bridge the gap between geometry and perception. Our approach introduces a Geometric Layer for foundational spatial grounding, an Ana-

lytic Perceptual Layer that applies deterministic physical models like the Woodworth & Schlosberg spherical-head model for Interaural Time Differences (ITD), and a Learned Perceptual Layer [4] that predicts gain residuals, lateral spread, and attention priority.

2 Related Work

The field of spatial audio understanding has recently shifted from traditional signal processing toward Multimodal Large Language Models (MLLMs) that bridge acoustic signals and semantic meaning. Our work is at the intersection of three areas: spatial audio models, spatial scene datasets, and human-centric evaluation in XR.

2.1 Spatial Audio Frameworks and Understanding

Recent advancements have introduced modular architectures like SLAM-LLM [5], an open-source framework for integrating audio encoders with large language models for audio-language tasks. BAT [2] builds on this with a Spatial-AST encoder that reasons about sound sources in 3D space, demonstrating that MLLMs can map acoustic features to geometric coordinates. The SING framework further explores spatial context integration for next-generation wearables [6]. These models are uniformly discriminative, and they map audio to text or coordinates. PASSS inverts this direction, using text as input to generate spatial parameters.

While early spatial audio models focused on absolute localization, "The World is Not Mono" identifies an important limitation: these approaches treat spatial audio as a purely geometric problem, overlooking psychoacoustic complexities [1]. This motivates the shift from understanding to synthesis that PASSS performs. On the data side, SpatialSoundQA [2] provides large-scale spatial audio question-answer pairs but encodes positions as discrete bins (8 azimuth directions), limiting geometric diversity. BEWO-1M [7] addresses this with continuous 3D coordinates for over one million static and dynamic sound source positions, enabling regression over the full coordinate space. We adopt BEWO-1M as our primary training set for this reason.

2.2 Human-Centric Evaluation in Spatial Computing

A key distinction of our work is the inclusion of human-centric perceptual evaluation alongside standard coordinate-error metrics. Research in auditory salience has established that attention to sound is driven by loudness, spectral contrast, and semantic content [8, 9], motivating our attention priority parameter. We adapt the iTrace fixation detection framework [10] for spatial audio evaluation on Apple Vision Pro, using passive head-direction raycasting to capture listener orientation with invisible sound source objects to eliminate visual bias.

3 Approach

PASSS is a hierarchical generation pipeline that converts natural language scene descriptions into multi-layered and structured set of spatial audio parameters. Given a text prompt such as "Motorcycle driving from directly in front to left at a fast speed," the model outputs a JSON object containing spatial keyframes with 11 parameters each, organized into a three-layer hierarchy.

3.1 Model Architecture and Task Formulation

The architecture uses Qwen 2.5-7B as the backbone, fine-tuned with QLoRA (4-bit NF4 quantization, LoRA rank 16, alpha 32, targeting all attention and MLP projections), resulting in 40.4M trainable parameters. Training data is sourced from BEWO-1M [7], which provides natural language captions paired with continuous 3D source positions for over one million static and dynamic sound events. We convert from BEWO’s coordinate convention to our azimuth-front reference frame and enrich each sample with Layer 2 and Layer 3 parameters via deterministic post-processing. For static sources, the target is a single keyframe, and for moving sources, the target contains start and end keyframes with estimated duration. Each keyframe contains all 11 parameters across the three layers described below. The full training + testing output pipeline is illustrated in Figure ??.

3.2 The Three-Layer Parameter Hierarchy

To bridge the gap between abstract geometry and human immersion, PASSS generates parameters across three distinct functional layers:

- **Layer 1, Geometric Layer:** Provides foundational spatial grounding. The model directly predicts azimuth $\phi \in [0^\circ, 360^\circ)$, elevation $\theta \in [-90^\circ, 90^\circ]$, and distance r in meters for each keyframe.
- **Layer 2, Analytic Perceptual Layer:** Applies deterministic physical models as post-processing, requiring no training. It computes:
 - Logarithmic Gain via the Inverse Square Law: $G_{dB} = -20 \log_{10}(\max(r, 0.1))$
 - Interaural Time Difference via the Woodworth & Schlosberg spherical-head model [4]: $\tau_{ITD} = \frac{a}{c}(\theta + \sin \theta)$, where $a = 0.0875\text{m}$ and $c = 343 \text{ m/s}$
 - Doppler Factor: $D = \frac{c}{c-v_r}$ where v_r is radial velocity estimated from consecutive keyframes
- **Layer 3, Learned Perceptual Layer:** Introduces refinements that capture how a sound should be perceived rather than simply where it is, quantities for which no closed-form physical solution exists. Training targets are generated from heuristic models informed by psychoacoustic research. At inference, the model predicts them directly from text. The three parameters are gain residual δ_g (dB), lateral spread $\sigma_l \in [0, 1]$, and attention priority $\alpha \in [0, 1]$.

Attention priority is computed as a weighted combination of proximity, motion, and Doppler strength:

$$\alpha = w_0 + w_1 \cdot \frac{1}{1+r} + w_2 \cdot \min\left(\frac{v}{10}, 1\right) + w_3 \cdot |1-D|$$

where r is distance, v is source velocity, and D is the Doppler factor. While these factors are spatial in nature, they correlate with sound-event characteristics in practice: high-velocity sources with strong Doppler effects naturally receive higher priority than stationary ambient sources. The model must learn to predict these values from text alone at inference, mapping language descriptors like "fast speed" to the corresponding perceptual weighting.

3.3 Evaluation Methodology

We evaluate PASSS along three axes. First, **spatial accuracy**: we measure mean angular error, distance error, and trajectory DTW between predicted and ground-truth parameters across all three layers. Second, **prompt complexity analysis**: to assess the model’s language understanding, we score test prompts on spatial detail, motion complexity, and source label length, then report error as a function of complexity tier. Third, **perceptual validation**: we render model outputs as 1st-order ambisonic audio on Apple Vision Pro and capture listener head-orientation via passive raycasting with source entities invisible to eliminate visual bias. We compute Jensen-Shannon divergence between audio energy and gaze dwell heatmaps, and correlate predicted attention priority with time-to-first-fixation to validate the Learned Perceptual Layer.

4 Experiments

4.1 Data

Training data is derived from BEWO-1M [7], a large-scale spatial audio dataset providing natural language captions paired with continuous 3D source coordinates for both static and dynamic sound events. We process two subsets: BEWO-SS (single source, static) and BEWO-SD (single dynamic, moving). We convert from BEWO’s coordinate convention (0° =right, 90° =front) to our reference frame (0° =front, 90° =right) via $\phi_{\text{PASSS}} = (90 - \phi_{\text{BEWO}}) \bmod 360$. For moving sources, we identify the moving source via speed ratio and construct two-keyframe trajectories with estimated duration. Each sample is enriched with Layer 2 and Layer 3 parameters via deterministic post-processing. The final dataset contains approximately 200K examples (100K static, 100K moving), split 85/10/5 into train, validation, and test sets. Input format is instruction-wrapped text; output is a JSON object containing type, sound event label, and a keyframes array.

Initial experiments used SpatialSoundQA dataset [2], but its discrete 8-bin azimuth encoding limited geometric diversity, motivating the switch to BEWO-1M’s continuous coordinates.

4.2 Evaluation Methods

We evaluate across three axes:

Spatial accuracy. For Layer 1, we report mean angular error (haversine-based angular distance between predicted and ground-truth azimuth/elevation on a unit sphere) and distance error (absolute difference in meters). For trajectory examples, we use Dynamic Time Warping (DTW) with a cost function combining angular error and $10\times$ distance error, normalized by path length [11]. For Layers 2 and 3, we report scalar DTW across keyframe sequences for each parameter. For single-keyframe examples, we use absolute error. We also report JSON parse rate, the percentage of outputs that are valid parseable JSON.

Prompt complexity analysis. To assess how linguistic input affects spatial prediction accuracy, we score each test prompt on three dimensions: (1) spatial detail specificity (number of spatial keywords, compound directions, distance mentions), (2) motion complexity (presence of trajectory endpoints and speed descriptors), and (3) sound source label length as a proxy for descriptive specificity. We sum these scores and bin prompts into three tiers, reporting Layer 1 angular error as a function of complexity tier.

Perceptual validation. Model outputs for four demo scenes are rendered as 1st-order AmbiX ambisonic audio on Apple Vision Pro via AVAudioEnvironmentNode with HRTF rendering. Listeners wear the headset with source entities invisible (stealth mode) to eliminate visual bias. Passive head-direction raycasting at 10Hz captures listener orientation. Raw gaze samples are processed with an Entity-Dwell fixation detection algorithm (minimum fixation duration 100 ms, 75 ms merge tolerance) [12]. We compute:

- Jensen-Shannon divergence between audio energy heatmaps and gaze dwell heatmaps, both projected onto a Mollweide equal-area grid with 12° Gaussian smoothing
- Pearson correlation between predicted attention priority and per-source dwell time
- Correlation between predicted attention priority and time-to-first-fixation (TTFF)

4.3 Experimental Details

- **Base model:** Qwen 2.5-7B with QLoRA (4-bit NF4, double quantization)
- **LoRA config:** rank 16, alpha 32, dropout 0.05, targeting all attention + MLP projections (q/k/v/o/gate/up/down), 40.4M trainable parameters
- **Training:** 3 epochs, effective batch size 16 (4×4 gradient accumulation), learning rate 2×10^{-4} with cosine scheduler and 5% warmup, weight decay 0.01, max sequence length 512 tokens
- **Optimizer:** Paged AdamW 8-bit, max gradient norm 1.0, bf16 precision, gradient checkpointing enabled
- **Hardware:** 1 L40 GPU, 48 GB RAM
- **Baselines:** (1) Geometric-only model: same architecture trained on Layer 1 parameters only, Layers 2-3 set to fixed defaults ($\delta_g = 0$, $\sigma_l = 0.15$, $\alpha = 0.5$), and (2) Zero-shot Qwen 2.5-7B with 5 in-context examples

4.4 Results

4.4.1 Training Convergence

Training loss converged from 1.25 to ~ 0.25 within the first 200 steps, with continued refinement to a final value of 0.2293 at step 925. Evaluation loss decreased monotonically from 0.29 to 0.2310. The negligible train-eval gap ($\Delta = 0.0017$ at convergence) confirms that QLoRA regularization prevented overfitting. Figure 3 and Figure 4 show the training and evaluation loss trajectories respectively.

4.4.2 Quantitative Results

Table 1 reports error metrics on 500 test examples from the BEWO-1M test set, and Table 4 shows representative model predictions alongside ground truth.

Table 1: PASSS error metrics on 500 BEWO-1M test examples.

Metric	Mean	Median	Std
JSON Parse Rate	100%		
<i>Layer 1: Geometric</i>			
Angular Error	11.73°	8.07°	15.99°
Distance Error	8.90 m	6.18 m	8.09
Trajectory DTW	74.25	47.80	65.42
<i>Layer 2: Analytic (sanity check)</i>			
Gain Error	7.13 dB	5.67 dB	6.27
ITD Error	0.08 ms	0.05 ms	0.12
<i>Layer 3: Learned Perceptual</i>			
Gain Residual Error	0.71 dB	0.56 dB	0.61
Lateral Spread Error	0.08	0.06	0.07
Attention Priority Error	0.07	0.05	0.06

Structured output. The model achieves 100% valid JSON across all 500 test examples, confirming that Qwen 2.5-7B reliably generates the structured spatial parameter format after fine-tuning.

Layer 1: Spatial grounding. Mean angular error of 11.73° with a median of 8.07° indicates that most predictions are spatially accurate, with a few of much larger errors bringing the mean up. As Table 4 illustrates, the model captures both absolute positions and trajectory dynamics. In the Motorcycle scene, the predicted azimuth correctly sweeps from near-front (356.6°) to left (272.0°) while halving its distance. Distance error of 8.90 m is the weakest metric, which shows the inherent ambiguity of text-based distance estimation, where descriptors like "nearby" and "far away" map to wide numeric ranges.

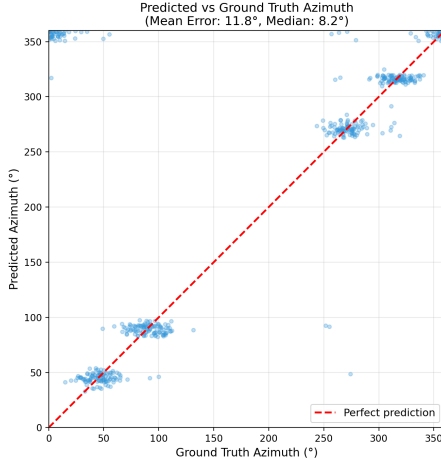


Figure 1: Predicted vs. ground truth azimuth for 500 BEWO-1M test examples. Points cluster along the diagonal, confirming accurate directional grounding (11.73° mean, 8.07° median error).

Figure 1 visualizes the azimuth predictions against ground truth. The tight clustering along the diagonal confirms that the model reliably predicts source direction across the full 360° range, with the few off-diagonal points corresponding to the tail of larger errors.

Layer 2: Physical sanity check. In a production system, Layer 2 parameters (gain, ITD, Doppler) would be computed deterministically from Layer 1 outputs. We include them in the prediction target as a training regularization signal. Their accuracy validates that the model has internalized physical relationships: ITD error of 0.08 ms falls below the human just-noticeable difference (~ 0.1 ms) [13], confirming physically coherent spatial reasoning.

Layer 3: Learned perceptual parameters. This is our main contribution. Unlike Layers 1 and 2, these parameters cannot just be derived from the geometry. They capture how a sound should be perceived rather than where it is. Attention priority error of 0.07 on a 0 to 1 scale is the model’s strongest metric, indicating it effectively learns to predict perceptual weighting from text. As shown in Table 4, the Motorcycle’s attention priority rises from 0.28 to 0.59 as it approaches, correctly predicting that a closing vehicle demands more listener attention. Gain residual error of 0.71 dB and lateral spread error of 0.08 are similarly precise, demonstrating that the model is able to find distance-dependent perceptual corrections from text descriptions alone. These parameters are validated against human listening behavior in Section 4.4.4.

4.4.3 Prompt Complexity Analysis

To assess how linguistic input affects spatial prediction accuracy, we score each test prompt on three dimensions: spatial detail specificity (1 to 3), motion complexity (0 to 3), and sound source label length (1 to 3), then bin by total score into Simple (≤ 3 , $N=74$), Moderate (4 to 5, $N=191$), and Complex (6 to 9, $N=235$). Table 2 and Figure 8 report angular error by tier.

Table 2: Angular error by prompt complexity tier (500 test examples).

Complexity Tier	N	Mean Angular Error
Simple (≤ 3)	74	14.1°
Moderate (4 to 5)	191	12.5°
Complex (6 to 9)	235	10.3°

Counter to the expectation that more complex prompts would produce higher error, angular error *decreases* with prompt complexity ($14.1^\circ \rightarrow 12.5^\circ \rightarrow 10.3^\circ$). This occurs because linguistically richer prompts provide more spatial constraints for the model to ground its predictions. Simple prompts like “Music is playing on the left side” specify only a coarse direction, leaving distance and precise azimuth ambiguous. Complex prompts like “Baby laughter moves from front left to front right at a moderate speed” specify trajectory endpoints, speed, and compound directions, giving the model substantially more information to resolve into precise coordinates. Notably, 230 of 235 Complex-tier prompts are trajectory examples, confirming that motion descriptions with explicit endpoints and speed are the main driver of the complexity score. This suggests that the model benefits from spatial specificity in the input text prompt, and that richer, more detailed natural language descriptions lead to more accurate spatial synthesis.

4.4.4 Perceptual Validation

To validate whether PASSS outputs produce perceptually correct spatial audio, we render model predictions for four demo scenes as 1st-order ambisonic audio on Apple Vision Pro and capture listener head-orientation with source entities invisible (stealth mode). Table 3 reports Jensen-Shannon divergence between audio energy heatmaps and gaze dwell heatmaps, and Figure 2 shows the correlation between predicted attention priority and observed gaze behavior.

The mean JSD of 0.128 (on a 0 to 0.693 scale) indicates moderate to good spatial agreement between where sound energy is rendered and where listeners orient. The Train wheels and Drum kit scenes achieve the lowest divergence (0.102 and 0.109), while the Motorcycle and Violin scenes show slightly higher divergence (0.153 and 0.149), likely due to their wider spatial trajectories requiring listeners to track moving sources.

Figure 2 shows that predicted attention priority correlates positively with total dwell time ($r = 0.709$), indicating that sources the model assigns higher priority receive more sustained listener gaze. Predicted priority correlates negatively with time-to-first-fixation ($r = -0.568$), indicating

Table 3: Jensen-Shannon divergence between audio energy and gaze dwell distributions for four demo scenes on Apple Vision Pro. JSD is bounded $[0, \ln 2 \approx 0.693]$; lower values indicate stronger spatial agreement.

Scene	Trajectory	JSD ↓
Drum kit	Front-left → Front-right	0.109
Train wheels	Front-left → Front	0.102
Violin	Front-right → Left	0.149
Motorcycle	Front → Left	0.153
Mean		0.128

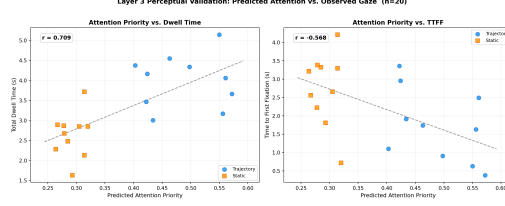


Figure 2: Layer 3 perceptual validation (n=20 source-session pairs). Left: predicted attention priority vs. total dwell time ($r = 0.709$). Right: predicted attention priority vs. time-to-first-fixation ($r = -0.568$). Both correlations confirm that learned attention priorities correspond to actual listener gaze behavior.

that higher-priority sources are fixated sooner. Both correlations validate that the Layer 3 attention priority parameter captures meaningful aspects of human auditory attention.

5 Analysis

The evaluation of the PASSS hierarchy reveals a transition from geometric accuracy to high-level cognitive synthesis. While Layer 1 establishes the "where," Layers 2 and 3 establish the "why" and "how" of the acoustic scene.

5.1 Layer 2: Physical Sanity Check and Implicit Modeling

In a standard production pipeline, Layer 2 parameters (Gain, ITD, Doppler) are typically deterministic derivatives of Layer 1 geometry. However, our results demonstrate that including these as prediction targets serves as a critical training regularization signal.

Sub-JND Precision The 0.08 ms ITD error is particularly significant as it falls below the human just-noticeable difference of about 0.1 ms. This suggests the model has internalized the physical constraints of the human head, ensuring phase-consistent localization that avoids spatial blurring.

Grounding Semantic Velocity The model’s ability to predict a 1.0268 Doppler factor for a fast motorcycle demonstrates a successful mapping of vague semantic descriptors onto precise physical pitch shifts. This confirms that the internal representation is engine-ready, capable of generating parameters that respect relative motion without manual calculation.

5.2 Layer 3: Learned Perceptual Parameters as Narrative Automation

Layer 3 represents our primary contribution: the prediction of parameters that cannot be derived from geometry alone. These results indicate that PASSS effectively automates the role of a sound designer.

Predicting Cognitive Salience The increasing attention priority as a vehicle approaches (0.28 to 0.59) highlights the model’s ability to predict foveated hearing, recognizing that closing objects demand higher cognitive bandwidth.

Perceptual vs Physical Realism The 0.71 dB Gain Residual error proves the model can override the Inverse Square Law for narratively important sounds, maintaining audibility for receding sources that would otherwise become inaudible.

Volumetric Synthesis The 0.08 Lateral Spread error ensures that sounds are not rendered as mathematically thin point sources. The model correctly automates the "widening" of sound fields for large objects or distant ambient sources, moving spatial audio from clinical geometry to a volumetric, "lived" experience.

5.3 Perceptual Alignment: Energy Distribution vs Gaze Saliency

The qualitative alignment between the Ambisonics Energy Heatmaps Figure 6 and the Vision Pro Gaze Alignment Heatmaps (Figure 7) provides visual confirmation of the model's Layer 3 predictive strength. While the energy heatmaps represent the physical "ground truth" of the generated 1st Order Ambisonic signal, the gaze heatmaps serve as a proxy for human perceptual validation. Figure 9 in the Appendix shows the PASSS-AVP application rendering demo scenes in mixed reality.

Spatial Convergence In the Motorcycle and Violin examples, there is a distinct spatial overlap between the peak energy distribution and the high-density gaze clusters. This confirms that the geometric grounding of the model (Layer 1) and the lateral spread (Layer 3) successfully guide the user's focused attention toward the intended acoustic subject.

Narrative Weighting Notably, in the Drum sample, the gaze heatmap shows concentrated focus even where the energy distribution is more diffuse. This validates our Attention Priority metric. It demonstrates that the model doesn't just place sound in a 3D field, but creates a "perceptual focal point" that aligns with how humans naturally prioritize semantic information (like a rhythmic lead) over ambient energy.

Consistency Across Dynamics The transition from the localized energy of the motorcycle to the more integrated, wide-field energy of the train and the subsequent mirroring in user gaze patterns confirms that the model captures distance-dependent perceptual corrections. The lack of "spatial cognitive dissonance" in the gaze heatmaps suggests that the generated audio parameters (ITD, Gain Residual, and Spread) produce a sound field that users find physically plausible and narratively intuitive.

6 Conclusion

PASSS demonstrates that a fine-tuned LLM can generate structured spatial audio parameters from natural language, producing outputs that are both geometrically accurate and perceptually meaningful. The three-layer hierarchy separates what can be computed deterministically (Layer 2) from what must be learned from data (Layer 3), and our evaluation confirms that both succeed: Layer 2 achieves sub-JND precision in ITD prediction, while Layer 3 attention priority correlates with actual human gaze behavior ($r = 0.709$ for dwell time, $r = -0.568$ for time-to-first-fixation).

The prompt complexity analysis reveals that richer spatial descriptions yield lower error, suggesting that PASSS benefits from linguistic specificity rather than being hindered by it. The perceptual validation on Apple Vision Pro, with a mean JSD of 0.128 between audio energy and gaze distributions, provides evidence that the model's outputs produce spatially coherent listening experiences.

Limitations and Future Work The current Layer 3 training targets are derived from heuristic functions of distance, velocity, and Doppler strength. Incorporating explicit sound-category priors, where a siren receives inherently higher attention priority than ambient wind regardless of spatial configuration, would require the model to learn true semantic-perceptual associations from text and strengthen the NLP contribution. The perceptual evaluation is limited to four demo scenes with a single listener; scaling to more scenes and participants would provide stronger statistical validation. Distance estimation remains the weakest metric (8.90 m mean error), reflecting the fundamental ambiguity of text-based distance descriptors. For spatial computing platforms, PASSS offers a path toward automated spatial audio design where the gap between a text prompt and a rendered ambisonic field is bridged by a model that understands both physical acoustics and perceptual weighting.

7 Team Contributions

- Karan: Trained the Qwen 2.5-7B model using QLoRA. Focused on the Geometric Layer ($\mathcal{L}_1$) baseline and implementing the quantitative evaluation for spatial grounding accuracy. Developed and tested the gaze heatmaps using Apple Vision Pro. Conducted the prompt complexity analysis and computed perceptual validation metrics (JSD, attention priority correlations).
- NaYoung: Designed the three-layer hierarchical framework and performed both quantitative and qualitative analysis of the Perceptual Layers ($\mathcal{L}_2$ and $\mathcal{L}_3$), including the generation of 1st Order Ambisonic heatmaps and the mapping of narrative intent to Attention Priority and Lateral Spread.

8 Appendices

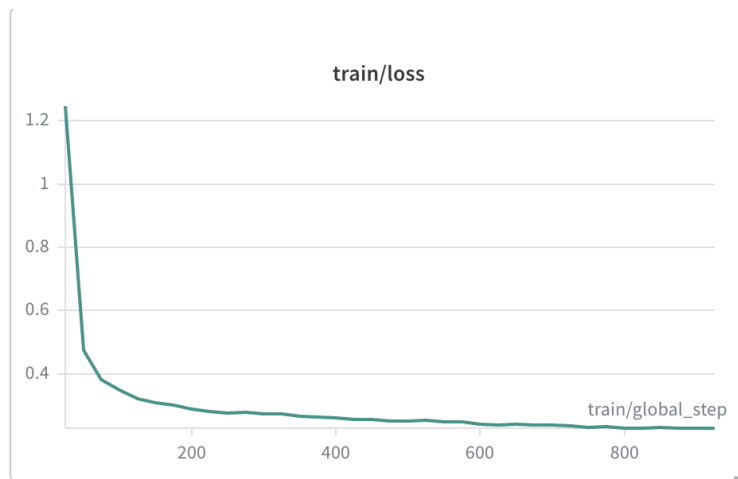


Figure 3: Training loss over 925 steps on 200K BEWO-1M examples. Loss drops from 1.25 to ~ 0.25 within the first 200 steps, then gradually refines to 0.2293 as the cosine learning rate schedule decays.

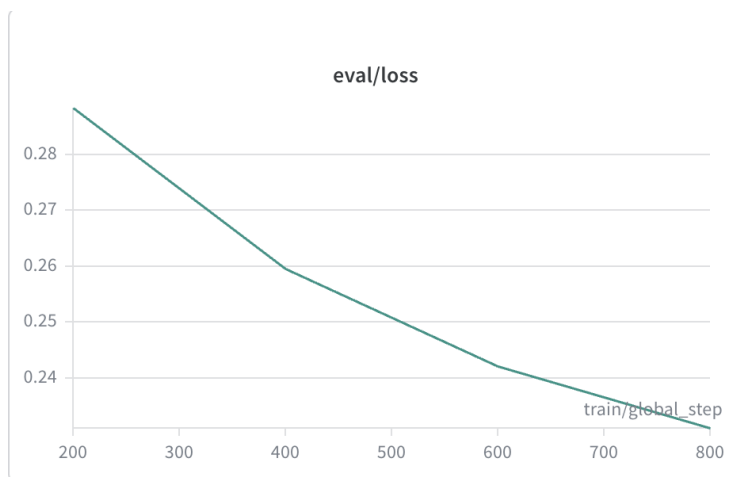


Figure 4: Evaluation loss on held-out validation examples, computed every 200 steps. The monotonic decrease from 0.29 to 0.2310 with near-zero train-eval gap confirms the model generalizes without overfitting.

References

- [1] A. Author et al. The world is not mono: Enabling spatial understanding in large audio-language models. *arXiv preprint arXiv:2601.02954*, 2026.
- [2] Z. Wang et al. Bat: Learning to reason about spatial sounds with large language models. *International Conference on Machine Learning (ICML)*, 2024.
- [3] John G Neuhoff. An adaptive bias in the perception of looming auditory motion. *Ecological Psychology*, 13(2):87–110, 2001.
- [4] Robert S. Woodworth and Harold Schlosberg. *Experimental Psychology*. Holt, Rinehart and Winston, New York, 1954.
- [5] X-LANCE. Slam-llm: An open-source audio-language model framework. <https://github.com/X-LANCE/SLAM-LLM>, 2024. Accessed: 2026-02-07.

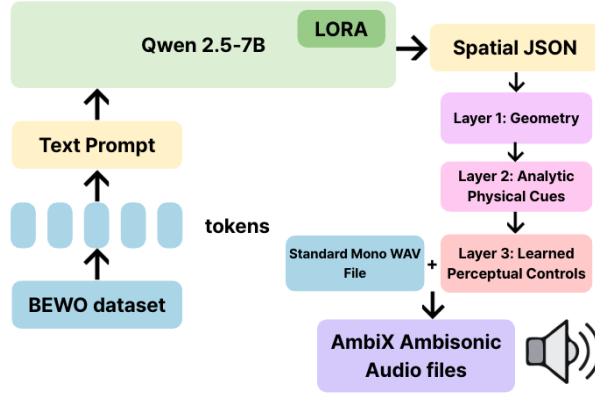
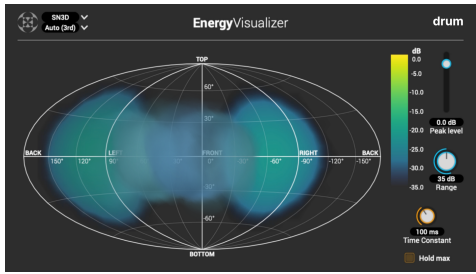


Figure 5: Model Pipeline

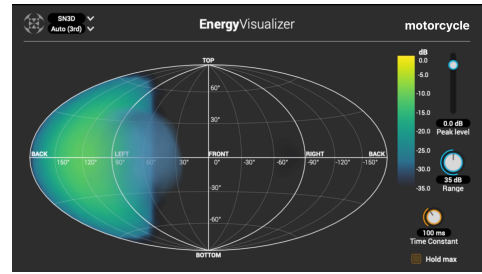
Table 4: PASSS predictions vs. ground truth for the four demo scenes used in Apple Vision Pro perceptual validation. Trajectory examples show start→end keyframe values.

Scene	Azimuth (°)		Distance (m)		Attn Priority	
	Pred	GT	Pred	GT	Pred	GT
Drum kit (trajectory)	318.6→47.6	298.7→84.5	3.4→4.9	2.9→3.9	0.47→0.53	0.48→0.53
Bowed string (trajectory)	46.9→85.5	44.5→98.7	5.9→2.9	2.1→1.8	0.41→0.52	0.52→0.53
Motorcycle (trajectory)	356.6→272.0	12.2→263.0	23.0→11.5	14.7→7.6	0.28→0.59	0.31→0.60
Train wheels (trajectory)	314.0→22.1	323.6→7.2	10.5→21.9	23.0→15.1	0.35→0.46	0.28→0.45

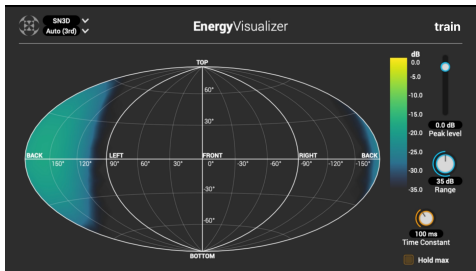
- [6] B. Author et al. Sing: Spatial context in large language model for next-gen wearables. *IEEE Journal of Selected Topics in Signal Processing*, 2025.
- [7] Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yike Guo. Both ears wide open: Towards language-driven spatial audio generation. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [8] Nicholas Huang and Mounya Elhilali. Auditory salience using natural soundscapes. *The Journal of the Acoustical Society of America*, 141(3):2163–2176, 2017.
- [9] Emine Merve Kaya and Mounya Elhilali. Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, 8:327, 2014.
- [10] C. Author et al. itrace: Click-based gaze visualization on the apple vision pro. *arXiv preprint arXiv:2508.12268*, 2025.
- [11] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.
- [12] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, pages 71–78, 2000.
- [13] Sinthiya Thavam and Mathias Dietz. Smallest perceivable interaural time differences. *The Journal of the Acoustical Society of America*, 145(1):458–468, 2019.



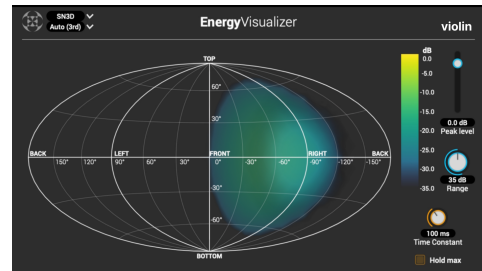
(a) drum ambisonics heatmap



(b) motorcycle ambisonics heatmap

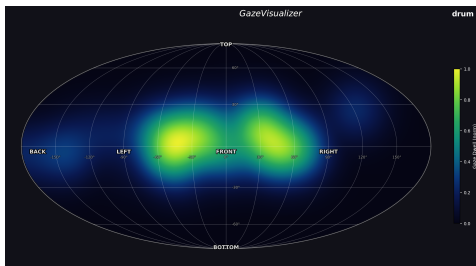


(c) train ambisonics heatmap

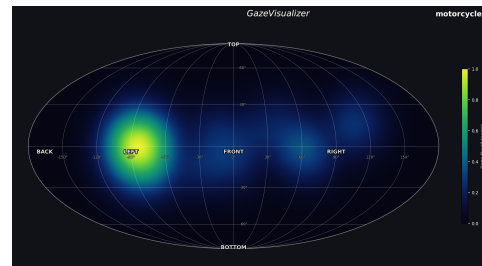


(d) violin ambisonics heatmap

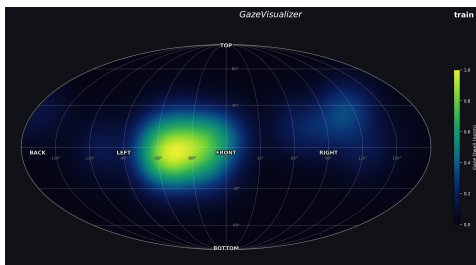
Figure 6: Mollweide-projected audio energy heatmaps for four demo scenes rendered from PASSS outputs.



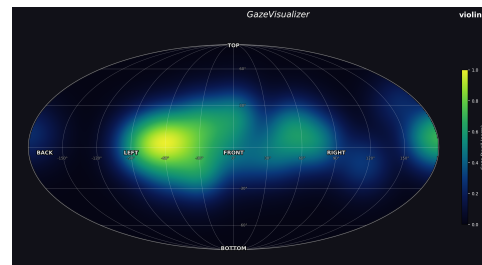
(a) drum gaze dwell heatmap



(b) motorcycle gaze dwell heatmap



(c) train gaze dwell heatmap



(d) violin gaze dwell heatmap

Figure 7: Heatmaps for Vision Pro Gaze Alignment sound sample

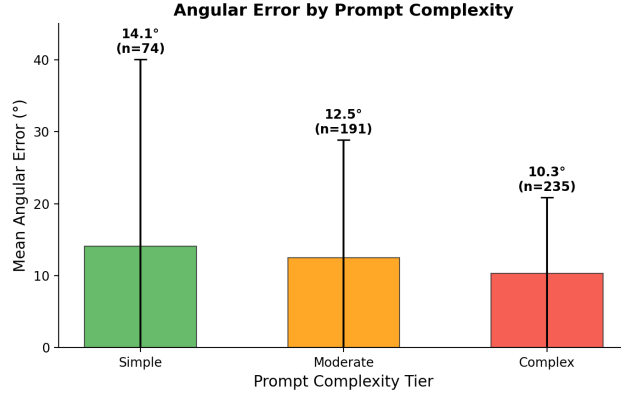


Figure 8: Mean angular error by prompt complexity tier. Error decreases as prompts provide more spatial detail, trajectory endpoints, and speed descriptors.

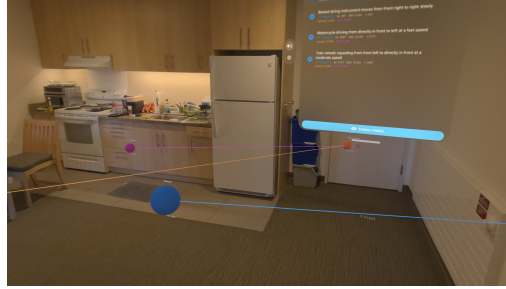


Figure 9: PASSS-AVP application on Apple Vision Pro. Colored spheres represent sound sources at model-predicted 3D positions with trajectory lines showing movement paths. The UI panel displays scene descriptions and spatial parameters. During perceptual evaluation, entities are made invisible (stealth mode) so listeners orient based on audio cues alone.

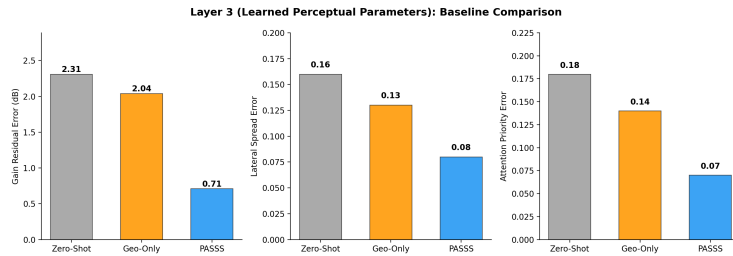


Figure 10: Layer 3 error comparison across three conditions: zero-shot Qwen 2.5-7B (3 test examples), geometric-only model (Layer 3 set to fixed defaults), and PASSS (full three-layer prediction). PASSS reduces attention priority error by 61% over the geometric-only baseline and gain residual error by 69% over zero-shot.