Spatial Audio Question Answering and Reasoning on Dynamic Source Movements

Anonymous ACL submission

Abstract

Spatial audio understanding aims to enable machines to interpret complex auditory scenes, particularly when sound sources move over time. In this work, we study Spatial Audio Question Answering (SAQA) with a focus on movement reasoning, where a model must infer object motion, position and directional changes directly from stereo audio. First, we introduce a movement-centric spatial audio augmentation framework that synthesizes diverse motion patterns from isolated mono audio events, enabling controlled and scalable training data generation. Second, we propose an end-to-end multimodal finetuning approach with a thinking mode, which allows audio-language models to produce explicit intermediate reasoning steps before predicting an answer. Third, we investigate the impact of query-conditioned source separation as a preprocessing stage and compare three inference regimes: no masking, an audio grounding model (AGM), and ground-truth masks. Our results show that reasoning amplifies the benefits of source separation, with thinking mode showing significant improvement of +5.1% when a single event is present in the question. These findings highlight the relationship between movement modeling, reasoning, and separation quality, offering new insights for advancing spatial audio understanding.

1 Introduction

Recent advances in audio question answering (AQA) have led to the development of state-of-the-art models capable of generating relevant answers from a given question and an audio clip. Building on this progress, we believe that spatial audio reasoning remains underexplored in audio-language models, particularly for dynamic motion cues. SAQA extends AQA to multichannel audio, implicitly integrating several traditional audio understanding tasks, including audio event detection (AED), audio captioning, audio localization and detection (Shimada et al., 2025), source separation (Nugraha et al., 2016), and object tracking (Li et al., 2025). SAQA has applications in smart surveillance, robotics, autonomous vehicles, and augmented reality. To interpret complex auditory scenes in these domains, SAQA systems must recognize diverse audio events and reason about dynamic source movements. This work advances SAQA with a focus on movement reasoning, bridging the gap between static spatial representations and real-world dynamics. The DCASE challenge Sound Event Localization and Detection (SELD) task (Shimada et al., 2025) has significantly advanced research in spatial audio by promoting models that jointly estimate event classes and their spatial attributes. (Sudarsanam and Politis, 2025) propose a spatial audio question answering model that predicts audio events, temporal order of events and spatial attributes ordering and spatial positions as a classification task. Although these models have achieved strong performance on benchmark datasets, they do not support open ended question answering and reasoning on top of the attributes. We develop an end-to-end spatial audio question answering model that includes a query conditioned audio source separation with the help of an Audio Grounding Model (AGM) to reduce the complexity of the audio depending on the question. Text-to-audio grounding is the task of detecting audio events in an audio clip based on natural language (Yasuda et al., 2022). From here on, we refer to this model as the Audio Grounding Model (AGM) for brevity. Similar to CLAP (Yuan et al., 2024), ELSA (Devnani et al., 2024) consists of a spatial audio encoder and a text encoder that learns a joint embedding of spatial attributes and semantic information via contrastive learning. Although the spatial audio representation learned by ELSA is well-suited for tasks like semantic retrieval and spatial audio captioning, it does not perform explicit question answering and lacks fine-grained temporal

reasoning.

BAT (Zheng et al., 2024) introduced the first end-to-end SAQA model, combining a spatial audio encoder with a large language model (LLM) to process binaural audio and project it into a text-aligned representation for question answering. However, BAT supports only static sources and does not provide reasoning for it's answers. (Sakshi et al., 2025) develop a plug and play spatial adapter that takes FOA input and is finetuned on top of a frozen LALM to introduce general spatial perception. During the time of this writing, we don't have access to their model implementation. Therefore, we only compare our model performance against BAT as a baseline.

To the best of our knowledge, we are the first to make the following contributions:

- **Movement-centric Spatial Audio Dataset Augmentation:** We introduce a dataset generation pipeline that creates spatial audio clips and corresponding question–answer pairs from the AudioSet strong-labeled corpus (Gemmeke et al., 2017). We isolate high-quality mono audio events, filter them using quality heuristics, and introduce controlled motion by assigning predefined trajectories before spatializing them into stereo scenes.

- **End-to-end Spatial Audio Reasoning with Thinking Mode:** We build an end-to-end spatial audio-language model that uses a thinking mode to produce intermediate reasoning steps before answering. This improves both interpretability and performance over non-thinking variants. As shown in Figure 1, the end-to-end system comprises of a temporally aware spatial audio encoder, a Q-former projector (Zhang et al., 2023), and a language model and is trained jointly with component-specific learning rates.

- **Query-conditioned Source Separation with Temporal Masking:** We evaluate source separation as a preprocessing step and show that extracting query-relevant audio sources leads to better attention allocation and improved accuracy. We compare no masking, an Audio Grounding Model (AGM), and ground truth masks, and demonstrate that separation synergizes strongly with thinking mode.

2 Related Work

2.1 Audio Language Models with Reasoning

Audio Language Models (ALMs) such as Mellow (Deshmukh et al., 2025), GAMA (Ghosh et al., 2024), and Audio Flamingo 3 (Goel et al., 2025) commonly follow a three-stage design that includes an audio encoder, an audio-to-text projection module, and a large language model. Training these components jointly has been shown to improve audio-text alignment, which in turn enhances downstream audio reasoning capabilities. The survey by (Cui et al., 2024) provides a broader perspective on the development of audio and speech language models, including architectural trends, pretraining strategies, and evaluation methods. (Sridhar et al., 2025) demonstrate that a curriculum-based finetuning approach can improve a model's ability to capture temporal structure and event order without degrading its overall performance on general audio tasks. Audio-CoT (Ma et al., 2025) explored chain-of-thought reasoning with self-consistency for LALMs across music, speech, and sound, and demonstrated performance improvements on the MMAU benchmark (Sakshi et al., 2024). Audio-Reasoner (Zhifei et al., 2025) introduced structured reasoning involving planning, caption grounding, step-by-step reasoning, and summarization. Inspired by these developments, our work introduces a structured reasoning framework specifically designed for spatial audio source-movement understanding. Our method employs a three-stage process, namely, step-by-step planning and solution derivation, answer generation, and rationale articulation, to enhance interpretability and robustness in spatial audio reasoning tasks.

2.2 Spatial Audio Understanding

(Tang et al., 2024) develop a spatial audio language model operating on first-order ambisonics (FOA) to assess its capability for localization, far-field recognition, and spatially informed extraction on the Spatial LibriSpeech dataset. (You et al., 2026) propose a unified hierarchical auditory scene analysis framework composed of a semantic audio understanding branch and a spatial processing branch, enabling LALMs to incorporate explicit spatial reasoning mechanisms. (Biswas et al., 2025) present OWL, which uses a geometry-aware spatial audio encoder trained with room impulse response (RIR) prediction as an auxiliary task while requiring only audio at inference time. Building on these works, our

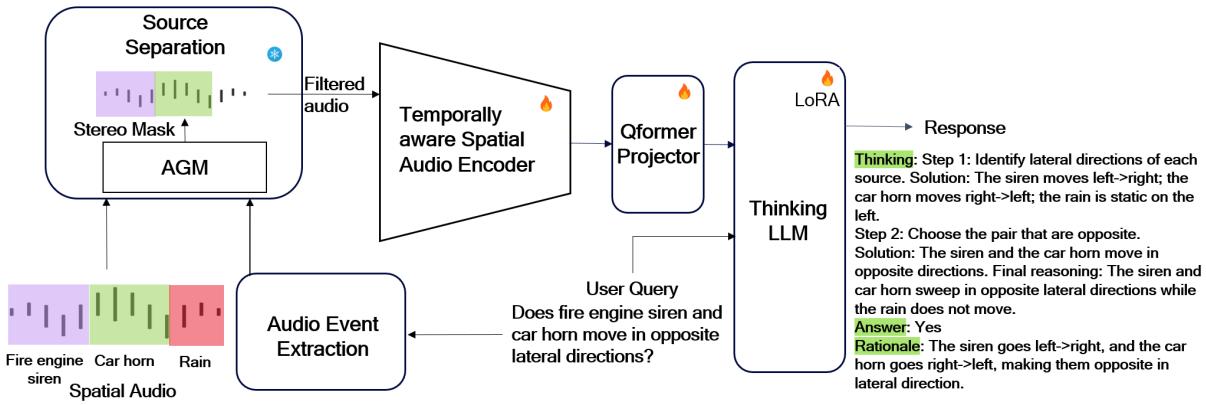

Figure 1: Our proposed spatial audio language model architecture with thinking mode. AGM refers to audio grounding model.

paper investigates spatial audio question answering with an explicit focus on movement reasoning.

3 Methodology

Our methodology has three components. First, we outline our data generation process. Second, we describe the multimodal spatial reasoning model. Third, we detail how query-conditioned source separation is applied using an Audio Grounding Model (AGM) to temporally mask stereo channels and isolate audio relevant to each query.

3.1 Movement–Centric Spatial Audio Dataset Augmentation

To study spatial audio reasoning with moving sound sources, we construct a controlled dataset of spatialized audio clips paired with movement–focused question–answer (Q/A) annotations. Our pipeline has three major steps: (1) isolating audio events and quality filtering (2) spatial scene synthesis with controlled motion, and (3) question-answer pair generation for movement reasoning.

3.1.1 Isolating Audio Events and Quality Filtering

We begin by extracting isolated audio events from the *AudioSet Strong–labeled* subset (Gemmeke et al., 2017), which provides frame–level annotations indicating when specific events occur. We apply two quality filters when selecting isolated audio events. First, each event must have a minimum duration of at least 3 seconds to ensure that the segment is long enough to support motion synthesis. Second, we require a CLAP audio–text similarity score of at least 0.45 between the event's audio embedding and the caption *"a sound of <event name>"*, which helps remove mislabeled or low-fidelity segments. After filtering, we obtain **224** high–quality mono audio events suitable for spatialization.

3.1.2 Spatial Scene Synthesis With Controlled Motion

Using the filtered events, we synthesize 10 second stereo spatial audio scenes. We adopt the spatial configuration from Dcase SELD task (Shimada et al., 2025) including stereo microphones, sources are positioned only in front of the listener, with azimuth angles sampled from $[-90°, 90°]$ where $0°$ is straight ahead, $-90°$ is left and $90°$ is right. Source distance ranges from $[0, 6]$ meters.

Each scene contains up to three events, and each event is randomly assigned either a static or a dynamic motion pattern. For dynamic sources, we allow several types of trajectories commonly used in spatial movement reasoning: **approach** motion, where the source moves closer to the listener (decreasing distance); **recede** motion, where the source moves away (increasing distance); **lateral movement**, involving left-to-right or right-to-left motion; and **arc trajectories**, where the source follows a curved path around the listener at a fixed radius, resulting in azimuth changes over time while the distance remains constant.

We render all scenes with Pyroomacoustics (Scheibler et al., 2018), using a maximum reflection order of 2 and surface absorption coefficient of 0.25, introducing modest reverberation while preserving salient spatial cues. For more information on parameters, refer to Appendix A.

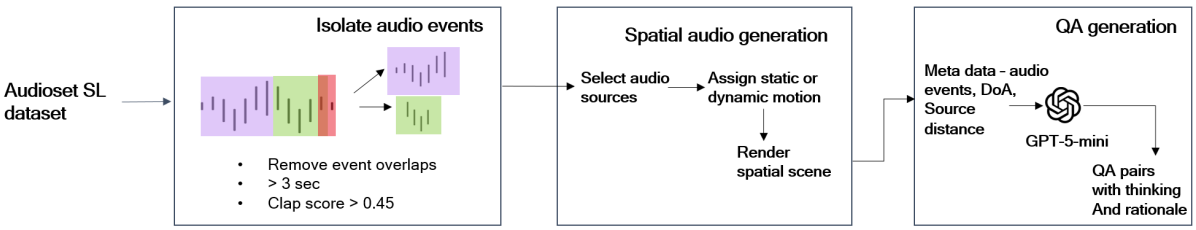

Figure 2: Spatial audio dataset generation

3.1.3 Question–Answer Pair Generation for Movement Reasoning

In the second stage of our pipeline, we construct question–answer (Q/A) pairs that evaluate movement understanding in Spatial AQA. For each synthesized audio clip, we store detailed metadata that includes framewise event identities, the per-frame azimuth and distance trajectories of each source, temporal boundaries and ordering such as event onsets and offsets, and a compact summary describing the overall movement patterns present in the scene. This metadata enables us to generate questions that require reasoning about spatio-temporal relationships rather than simply predicting numeric values.

While prior works such as BAT (Zheng et al., 2024) focuses on QA pairs that require direct estimation of DoA and distance, real-world queries often involve understanding spatial-temporal audio event relationships rather than purely numerical. We focus on generating diverse questions across 7 categories as shown in Table 2 including trajectories, radial changes, comparative diagnostics, temporal sequence, relative motion between events and natural perception.

To generate questions automatically, we provide `GPT-5-mini` (Microsoft Azure AI, 2026) with a structured scene description that includes the event list, azimuth and distance trajectories, onset/offset times, and a short summary of overall motion. To enable robustness to different question formats, the model is prompted to produce Boolean, multiple-choice, and open-ended questions. Apart from QA pairs, we also generate rationale to aid in interpretablity and thinking steps to improve the accuracy of the predicted answers. Depending on the difficulty of the question, we vary the thinking steps from 1 to 3. For more details on the QA data generation, please refer to the prompts in Appendix F.

We evaluated the quality of synthesized spatial QA pairs with a validated industrial expert in structured audio data collection. As shown in Table 1, the annotated scores are high, indicating that the generated questions are answerable, reasoning chains are coherent, and queries remain well grounded in their corresponding spatial audio scenes. We evaluate the boolean and multiple choice questions via keyword match. We utilize `GPT-5-mini` (Microsoft Azure AI, 2026) for evaluating thinking steps, rationale and open ended reasoning. For evaluating thinking steps, we require the LLM to score across the following dimensions: logical coherence, step completeness, factual accuracy and alignment with ground truth. We evaluate the rationale on conciseness, accuracy and clarity. Open ended questions are evaluated on factual accuracy and semantic similarity. For more details on the prompts, please refer to Appendix G.

Table 1: Human evaluation of the Spatial AQA dataset (1-5 scale; higher is better) across 200 randomly selected samples.

Metric	Score
Answer Accuracy	4.93
Reasoning Quality	4.94
Groundedness	4.97

3.2 Multimodal end-to-end training:

Our end-to-end Spatial Audio Language Model consists of three key components as show in the Figure 1: a temporally aware spatial audio encoder, a Q-Former projection module (Zhang et al., 2023), and a large language model (LLM) with thinking capabilities finetuned with LoRA. The Spatial audio encoder takes stereo audio as input and encodes them into audio tokens. We adopt the spatial audio encoder from BAT, modifying its penultimate layer to incorporate temporal awareness, enabling frame-level reasoning about audio events and spatial attributes. We utilize a Q-Former (Zhang et al., 2023) to bridge the acoustic feature embedding from spatial audio encoder to a Thinking LLM (Qwen 3 4B)

Table 2: Example QA pairs by category

Category	Question Type	Example Question	Answer
Trajectories	yes_no	Is the Croak sound moving from left to right?	Yes
Radial changes	yes_no	Are the waves moving closer to the listener as they sweep across?	Yes
Comparative diagnostics	multiple choice	Which source sweeps across a wider left→right range? *Options: [Blender, Car, Both equally, Neither]*	Blender
Temporal sequence	open	Narrate the overall choreography: where does each source start, how does it move, and where does it finish?	The crow starts on the left and moves in a moderate sweep across the center toward the right while approaching the listener. After the crow, a shout appears mainly on the right and executes a narrow left->right right-side sweep while receding. Later, the pigeon/dove begins on the left, sweeps widely across the center to the right, approaches then recedes as it passes, and finishes on the right.
Relative motion between events	yes_no	Do scissors and cattle sound move toward the same region (converge) as they progress?	Yes
Natural perception	open	Which source feels most noticeable near the middle of the scene and why?	The Applause feels most noticeable near the middle because it crosses the center with a moderate sweep, passing through that region, whereas the Camera stays on the left and does not pass through the middle.

(QwenTeam, 2025), enabling the model to generate internal thinking steps before arriving at the answer, and its rationale.

To enable robust spatial movement reasoning, we employ a variable learning framework that helps learn spatial audio-to-text. In the first stage, Foundational spatial grounding, the model is trained on spatial audioclips with frozen spatial audio encoder to allow the model to learn spatial audio to text mappings by learning the projection module weights from scratch. This stage helps the model in learning a clean mapping between the audio source signals and their corresponding trajectories in audio-text semantic space.

In the second stage, we train all the three components including the spatial audio encoder, Qformer projection and the thinking LLM. To ensure stable gradient updates and avoid catastrophic forgetting of clean signal from stage 1, we implement a balanced mini-batch sampling strategy with 1:1 ratio of audio clips with overlapping and non-overlapping audio events.

4 Audio source separation with temporal masking

When answering a question, humans naturally focus only on the few audio events that are relevant to the query rather than the entire soundscape. To mimic this behavior, we extract the key audio events mentioned in the question through keyword matching and pass this filtered set of event labels to

the source-separation module, which then isolates only the query-relevant audio from the spatial clip. Since the questions do not paraphrase the audio event names, we did not observe any errors in the keyword matching step.

To support diverse acoustic conditions and use cases, we develop an open vocabulary Sound Event Detection (SED) model that grounds free form text queries to corresponding audio events. Rather than relying on a fixed label set, the model localizes sound events conditioned on arbitrary textual descriptions of interest. Our Audio Grounding Model (AGM) is trained following the phrase level WSTAG framework in (Xu et al., 2024) using the AudioCaps (Kim et al., 2019) dataset, from which we derive phrase level supervision. AGM adopts the same architecture and hyperparameters as (Xu et al., 2024): a CRNN based audio encoder with eight convolutional layers followed by a bidirectional GRU, and a text encoder consisting of a word embedding layer with mean pooling. Frame level grounding scores are computed via cosine similarity between audio frame embeddings and text tag embeddings, followed by a sigmoid activation. To extract temporal event boundaries, we apply a global threshold of 0.8 to the framewise scores and refine them with a 0.3 second median filter. These boundaries are then used to mask the input audio by zeroing out waveform regions outside the detected event span; if no boundary is detected, no masking is applied. We utilize AGM as an inference time pre-processing step for filtering relevant audio events. Ground truth (GT) masking refers to using the oracle time periods for the filtered audio events. In the future, we plan to include the audio separation module in the end-to-end training to support time-frequency masking.

5 Experiments

5.1 Datasets, Models and Evaluation

We choose the BAT model as our baseline. We use Audioset Strong Labeled (Gemmeke et al., 2017) to construct the spatial audio clips using pyroomacoustics. Our dataset contains 40k spatial audio clips, evenly split between overlapping and non-overlapping spatial audio scenes, with an 80/10/10 train, dev and test splits. The generated question–answer dataset includes 559k total questions, comprising 195k yes/no, 111k multiple-choice, and 251k open-ended questions. For open ended answers, thinking, and rationale, we evaluate using GPT-5-mini. See Appendix G for prompt details. For open-ended answers, we utilize the factual accuracy for the results section 6.

5.2 Experiment Setup

We train the AQA architecture end-to-end using distributed data parallelism across 4 A100 GPUs for 5 epochs with gradient accumulation over 2 steps as described in the methodology section. Results are averaged over two runs with random seeds. We employ differentiated learning rates for different components: 5×10^{-6} for the spatial audio encoder, 1×10^{-4} for the Q-Former projection layer, and 2×10^{-5} for the Qwen2.5-3B-Thinking LLM with LoRA adapters. We have cross-entropy as the loss function. The audio encoder remains frozen for the first 0.5 epochs to stabilize training. We use AdamW optimizer with a weight decay of 0.01, warmup for 1000 steps, and cosine annealing learning rate schedule with a minimum learning rate of 10% of the initial value. Gradient clipping is applied with a maximum norm of 1.0. The LoRA configuration uses rank r=8, alpha=32 with a dropout of 0.05.

6 Results

Across all evaluations, we observe a consistent interaction between masking quality and thinking. We refer to the configuration without source separation as NoMask. While NoMask provides the lowest performance and GT masking provides an upper bound to the performance gains from time-based masking, AGM improves accuracy over NoMask and enabling some of the benefits associated with reasoning. From Table 3, BAT baseline provides poor performance due to overfitting to it's training format questions which are mostly predicting the spatial attributes with limited reasoning. In the next subsections, we discuss and analyze our models' performance across various categories.

Overall results: The Table 3 showcases the overall accuracy of the model. Overall, ground-truth (GT) masking yields the highest accuracy (56.1%), followed by AGM (55.0%) and NoMask (54.3%). Thinking improves performance most when masking is present: the improvement under GT masking is +2.0% compared to only +0.2% under NoMask, resulting in a Δ Interaction of +1.8 percentage point (pp). AGM shows an improvement of +1.0% in between NoMask and GT. This improvement

Table 3: Accuracy (%). ΔInteraction measures how much more (or less) thinking helps under a method compared to NoMask: $\Delta\text{Interaction} = (\text{Acc}_T^{\text{method}} - \text{Acc}_N^{\text{method}}) - (\text{Acc}_T^{\text{NoMask}} - \text{Acc}_N^{\text{NoMask}})$. Acc is accuracy, T refers to thinking and N refers to non-thinking

	BAT	**Thinking**			**Non-Thinking**			**Δ Interaction (pp)**	
Question Type		NoMask	AGM	GT	NoMask	AGM	GT	AGM	GT
Yes_No	41.7	72.1	73.8	**76.3**	72.6	72.8	72.3	**+1.5**	**+4.5**
Multiple Choice	0.3	**53.1**	52.9	52.8	50.4	50.3	50.3	−0.1	−0.2
Open	0.00	41.0	41.4	**41.8**	41.2	41.0	41.5	**+0.6**	**+0.5**
Overall	14.7	54.3	55.0	**56.1**	54.1	54.0	54.1	+0.8	+1.8

aligns with the intuition that the model benefits from both (i) cleaner evidence and (ii) an extended reasoning chain prior to generating the final answer. Please refer to Appendix B for human evaluation of open ended answers and their correlation with GPT-5-mini and Appendix E for qualitative examples.

Overall results by question type: Across the three major question formats, Yes/No, Multiple-Choice, and Open-Ended, we observe clear differences in how source separation interacts with the LLM's reasoning pipeline as seen in Table 3. For Yes/No questions, GT masking with thinking significantly improves accuracy. These questions tend to hinge on detecting a single dominant acoustic cue ("Does the Train whistle move from the right toward the left?", "Is the crowing approaching the listener over time?"). Masking isolates the relevant source and removes competing background audio, making the evidence sharper and allowing the model to commit to a more confident binary decision. For MCQ questions, AGM and GT masking perform on par or slightly lower than NoMask. We believe this could be due to two reasons: First, many MCQ questions require reasoning over multiple/all audio events. For example, "Which source spans a wider left->right range? choices: [Church bell, Electric shaver, They span equally wide, Can't tell]". Second, by only performing time-based masking, there is still some noise left in the audio clip from the irrelevant audio events hindering the LLM reasoning. In the future, we will work on performing time-frequency masking. For open-ended questions, we see a slight increase in accuracy for AGM and GT in thinking mode. These questions rely on a richer holistic audio scene (e.g., "Explain where each source ends and how its final position follows from its path.", "Do the chainsaw and tick draw closer together, stay the same distance from each other, or move apart as the scene progresses? Explain perceptually.").

Thinking over non-thinking: The Δ interaction measures how much thinking improves the performance of AGM or GT relative to NoMask. In Tables 3, 10, and 9, AGM delivers moderate gains in ΔInteraction, while GT masking provides substantial gains, showing that high-quality audio isolation most effectively supports text-based reasoning. Overall, these results demonstrate that thinking mode reliably outperforms non-thinking mode, and that its effectiveness is closely tied to the quality of source separation. Better-isolated audio allows the LLM's reasoning steps to operate on cleaner spatial evidence, resulting in stronger spatial AQA performance. For NoMask, thinking provides only small improvements for Yes/No and Open-Ended questions, with a slightly larger gain for Multiple-Choice questions (+2.7%). However, we also find that thinking incurs a notable latency cost: 6.19 seconds on average compared to 2.36 seconds for non-thinking inference (2.6× slower). Taken together, these results suggest that when masking is unavailable, the modest accuracy gains offered by thinking mode may not justify the substantially higher inference time.

Accuracy by overlap: The Table 4 highlights that even in thinking mode, filtering relevant audio sources continues to provide measurable benefits. For non overlapping audio event cases, GT masking yields highest gains of +2.9% while AGM provides moderate yield of +1.4% over NoMask. In overlapping scenes, performance decreases across all methods, but the relative ordering of *NoMask* < *AGM* < *GT* remains the same, indicating that reasoning is still limited by the residual noise left in the audio by irrelevant overlapped audio events

Table 4: Accuracy by overlap category (%) for thinking mode.

Overlap	NoMask	AGM	GT
No overlap	55.0	56.4	**57.9**
Overlap	53.7	53.7	**54.3**

Table 5: Accuracy by audio events (%) for thinking mode.

Audio Events	NoMask	AGM	GT
Complete match	48.4	48.6	**49.2**
1 Event	64.3	66.2	**68.7**
2 events	**54.1**	53.7	53.8

Error Type	**Non-Thinking**	**Thinking**
Lateral Direction	18.5	15.4
Attribute Swapping	15.7	14.4
Source Misidentification	14.5	13.3
Distance / Depth	14.2	10.3
Crossing Center	11.7	8.9
Temporal Sequence	7.8	5.5
Rationale–Answer Mismatch	9.5	4.1
Magnitude Errors	2.9	2.7

Table 6: Error category comparison for GT masking: Thinking vs Non-Thinking. Values are in %. The lower error percentage the better.

after time-based masking.

Accuracy by number of relevant audio events: The Table 5 summarizes performance across scenes containing a single event, two events, or all events (`Complete match`) with thinking mode. The `1 Event` condition yields the strongest improvements: GT+Thinking reaches 68.7% with a +4.4% gain over NoMask, and AGM also provides meaningful benefits (66.2%, +1.9%). This indicates that clean source isolation is crucial for leveraging the thinking mode. A similar but weaker trend appears in the `Complete match` setting, where both AGM and GT offer only modest performance gains, suggesting that mixed scenes with multiple active events are inherently harder to disentangle. Overall, the reduced gains for `Complete match` and `2 Events` highlight that multi-source spatial mixtures remain a key bottleneck for reasoning even when masking is applied.

Comparison of error categories: Table 6 shows the categorization of errors in GT-masked thinking and non-thinking responses. GT masking in thinking mode reduces errors consistently across all categories, with the largest drops occurring in lateral direction (azimuth) errors (–3.1%), distance/depth estimation errors (–3.9%), and rationale–answer mismatches (–5.4%). These improvements suggest that masking meaningfully enhances the precision of spatial geometry reasoning. Both azimuth and distance estimation rely heavily on clean directional cues, which are often blurred by overlapping sources in the non-masked audio; isolating the ground-truth region provides clearer spatial evidence and directly reduces these errors. The reduction in rationale–answer mismatches further indicates that thinking mode, when combined with masking, produces more stable reasoning and is less likely to generate rationales that contradict or diverge from the final answer. This aligns with the hypothesis that clearer audio inputs lead to more coherent and consistent reasoning trajectories.

7 Conclusion

In this work, we developed a spatial audio multimodal LLM capable of reasoning about dynamic audio scenes and tracking the movement of sound sources over time. Our architecture integrates spatial audio encoding, audio–language alignment, and LLM-based reasoning to support richer spatial understanding than prior static spatial AQA systems. A key finding is that incorporating query-relevant source separation meaningfully improves attention and reasoning by isolating the most informative regions of the audio clip. In future, we aim to perform temporal frequency masking and extend our framework to more complex real-world environments and multi-source recordings, and to explore stronger temporal and motion priors for spatial reasoning for long audio.

Limitations

Since masking is applied temporally, overlapping irrelevant events within the same time window are not completely suppressed, potentially introducing acoustic interference in densely populated soundscapes. Additionally, our model currently supports audio clips with a maximum duration of 10 seconds

(320,000 samples at 32 kHz) due to computational and memory constraints. This temporal limitation restricts the model's ability to capture long-term spatial trajectories and complex multi-event interactions. Our framework currently supports a limited set of motion patterns. In future work, we aim to extend it to handle longer audio sequences, incorporate more complex motion dynamics, and include recorded spatial audio clips to better capture object movements and spatial relationships in real-world acoustic environments.

8 Ethical Considerations

This work involves processing audio clips from a publicly available dataset, Audioset. The audioclips only contain non-speech audio events and does not contain any sensitive or personal information.

References

Subrata Biswas, Mohammad Nur Hossain Khan, and Bashima Islam. 2025. Owl: Geometry-aware spatial reasoning for audio large language models. *Preprint*, arXiv:2509.26140.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*.

Soham Deshmukh, Satvik Dixit, Rita Singh, and Bhiksha Raj. 2025. Mellow: a small audio language model for reasoning. *arXiv preprint arXiv:2503.08540*.

Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. 2024. Learning spatially-aware language and audio embeddings. In *NeurIPS*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313, Miami, Florida, USA. Association for Computational Linguistics.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*. Website: https://research.nvidia.com/labs/adlr/AF3/.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhangbin Li, Jinxing Zhou, Jing Zhang, Shengeng Tang, Kun Li, and Dan Guo. 2025. Patch-level sounding object tracking for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5075–5083.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*.

Microsoft Azure AI. 2026. Gpt-5 mini. `https://ai.azure.com/catalog/models/gpt-5-mini`.

Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. 2016. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664.

QwenTeam. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

S Sakshi, Vaibhavi Lokegaonkar, Neil Zhang, Ramani Duraiswami, Sreyan Ghosh, Dinesh Manocha, and Lie Lu. 2025. Spur: A plug-and-play framework for integrating spatial audio understanding and reasoning into large audio-language models. *arXiv preprint arXiv:2511.06606*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*. Benchmark: https://github.com/Sakshi113/MMAU.

Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. 2018. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 351–355.

Kazuki Shimada, Archontis Politis, Iran R Roman, Parthasaarathy Sudarsanam, David Diaz-Guerra, Ruchi Pandey, Kengo Uchida, Yuichiro Koyama,

Naoya Takahashi, Takashi Shibuya, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji. 2025. Stereo sound event localization and detection with onscreen/offscreen classification. *arXiv preprint arXiv:2507.12042*.

Arvind Krishna Sridhar, Yinyi Guo, and Erik Visser. 2025. Enhancing temporal understanding in audio question answering for large audio language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 1026–1035.

Parthasaarathy Sudarsanam and Archontis Politis. 2025. Towards spatial audio understanding via question answering. *Preprint*, arXiv:2507.09195.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Jun Zhang, Lu Lu, Zejun Ma, Yuxuan Wang, and 1 others. 2024. Can large language models understand spatial audio? In *Proc. Interspeech 2024*, pages 4149–4153.

Xuenan Xu, Ziyang Ma, Mengyue Wu, and Kai Yu. 2024. Towards weakly supervised text-to-audio grounding. *Preprint*, arXiv:2401.02584.

Masahiro Yasuda, Yasunori Ohishi, and Shoichiro Saito. 2022. Echo-aware adaptation of sound event localization and detection in unknown environments. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 226–230. IEEE.

Yuhuan You, Lai Wei, Xihong Wu, and Tianshu Qu. 2026. The world is not mono: Enabling spatial understanding in large audio-language models. *Preprint*, arXiv:2601.02954.

Yi Yuan, Zhuo Chen, Xubo Liu, Haohe Liu, Xuenan Xu, Dongya Jia, Yuanzhe Chen, Mark D. Plumbley, and Wenwu Wang. 2024. T-clap: Temporal-enhanced contrastive language-audio pretraining. *Preprint*, arXiv:2404.17806.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. 2023. Vision transformer with quadrangle attention. *arXiv preprint arXiv:2303.15105*.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. 2024. Bat: Learning to reason about spatial sounds with large language models. *arXiv preprint arXiv:2402.01591*.

Xie Zhifei, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23829–23851, Suzhou, China. Association for Computational Linguistics.

A Spatial Audio Generation Parameters

The spatial audio scenes were synthesized using the pyroomacoustics library (Scheibler et al., 2018) with the following acoustic parameters: Room dimensions were set to 7.0 × 5.0 × 3.0 meters (length × width × height) with random perturbations of ±0.5 meters per dimension to introduce variability across scenes. A stereo microphone array was positioned at the center coordinates [3.5, 1.0, 1.5] meters with an inter-microphone spacing of 0.18 meters, approximating typical binaural recording configurations. Room impulse responses (RIRs) were computed using the image source method with a maximum reflection order of 2 and uniform wall absorption coefficient of 0.25, simulating moderately reverberant indoor environments. Air absorption was enabled to model frequency-dependent attenuation. All audio was rendered at a sampling rate of 32 kHz. Each scene contained 3 simultaneous sound sources with durations of 10 seconds. During audio synthesis, we adopt an adaptive frame-based rendering approach setting the frame duration to be between 10-50 ms depending on the source velocity (10 ms for fast motion >5 m/s, 50 ms for slow motion <1 m / s).

B Human Evaluation and Correlation with GPT-5-mini as judge

To validate the reliability of our automated LLM-based evaluation, we conducted a human evaluation study on open-ended questions. We recruited three evaluators with graduate-level education in computer science from our organization. Each evaluator was provided with the same evaluation prompt used for GPT-5-mini (see Appendix G.3) and asked to assess 50 randomly selected question–answer pairs across factual accuracy and semantic similarity dimensions. Each QA pair included model responses from three configurations: NoMask, AGM, and GT thinking settings. Table 7 presents the mean evaluation scores from human annotators and GPT-5-mini across the three model configurations. Both human and automated evaluators show consistent trends, with the AGM and GT thinking configurations achieving higher scores than the NoMask configuration. The Pearson correlation between human and GPT-5-mini scores was 0.78 and 0.61 for factual accuracy and semantic similarity, respectively, indicating a strong positive correlation between human and automated evaluations. These results suggest that GPT-5-mini can serve as a rea-

sonable proxy for human evaluation, particularly for assessing factual correctness, though human judgment remains important for capturing semantic nuances.

C Further Results

C.1 Overall results

The overall results with thinking and rationale scores is provided in the Table 8.

C.2 Question Type x Audio Events

Question type × audio events extracted: Table 10 shows the performance of models for each question type and relevant audio events frequency. For `Yes/No` and `Open` questions, the `1 Event` condition produces the highest accuracies across all masking methods. Both question types rely on identifying the salient motion of a single audio event either via GT masking or partially via AGM, reduces scene clutter and supports stronger reasoning gains. This is reflected in the large interaction values for `Yes/No` under 1 Event (+7.8 for GT, +2.6 for AGM) and moderate but positive gains for Open. In contrast, the `Complete Match` condition offers smaller benefits from masking. MCQ questions often require 2 or more audio events to be compared across the spatial attributes. This reduces the necessity for masking in such scenarios. AGM shows this behavior most clearly as it helps in `Open` and `Yes/No` questions for 1 Event, but converges toward or below NoMask for MCQ in 1 Event and 2 Events conditions.

D Question Type x Overlap

The Table 10 showcases the performance of our model across question type and overlapping scenario.

E Qualitative Evaluation

Table11 shows that AGM and GT masking substantially improve answer correctness across all question types. This improvement comes not only from clearer audio input but also from more accurate and grounded reasoning steps. In contrast, the no-mask conditions frequently exhibit direction flips, incorrect sweep-span assumptions, and mistakes in approach–recede interpretation even when detailed reasoning is provided. These patterns reveal that small ambiguities in audio can distort the entire reasoning chain, underscoring the importance of audio source separation. With disentangled audio, the model's reasoning stabilizes and aligns with the true spatial cues, demonstrating that high-quality audio representations are essential for reliable audio-based reasoning.

Table 7: Comparison of human and GPT-5-mini mean evaluation scores for open-ended questions across model configurations

Model Type	Human Annotators		GPT-5-mini	
	Factual Acc.	**Semantic Sim.**	**Factual Acc.**	**Semantic Sim.**
NoMask Thinking	2.600	3.193	2.100	2.500
AGM Thinking	2.747	3.280	2.420	2.780
GT Thinking	2.727	3.267	2.460	2.760

Table 8: Overall core metrics. Accuracies are %; Thinking/Rationale are 1–5 scales. GT refers to ground-truth masking based source separation, NoMask refers to without using source separation, AGM refers to audio grounding model based source separation.

	Thinking			Non-Thinking		
Metric	NoMask	AGM	GT	NoMask	AGM	GT
Accuracy (%)	54.3	55.0	**56.1**	54.1	54.0	54.1
Thinking Avg (1-5)	3.507	3.522	**3.549**	0.0	0.0	0.0
Rationale Avg (1-5)	4.156	4.168	**4.179**	**4.182**	**4.182**	4.180

Table 9: Question Type × Audio Events. Accuracies are %; ΔInteraction is defined as $(\mathrm{Acc}_T^{\mathrm{method}} - \mathrm{Acc}_N^{\mathrm{method}}) - (\mathrm{Acc}_T^{\mathrm{NoMask}} - \mathrm{Acc}_N^{\mathrm{NoMask}})$.

		Thinking			Non-Thinking			Δ Interaction (pp)	
QType	**Audio Events**	NoMask	AGM	GT	NoMask	AGM	GT	AGM	GT
Multiple Choice	Complete Match	**53.2**	53.0	53.9	50.6	50.0	51.0	**+0.4**	**+0.3**
Multiple Choice	1 Event	**57.0**	56.3	52.7	52.7	54.2	50.2	−2.2	−1.8
Multiple Choice	2 Events	**51.1**	50.9	47.7	48.7	50.0	47.0	−1.5	−1.7
Open	Complete Match	**35.6**	36.1	36.2	36.1	35.5	36.3	**+1.1**	**+0.4**
Open	1 Event	52.8	53.2	**54.2**	52.3	52.3	52.7	**+0.4**	**+1.0**
Open	2 Events	**43.1**	43.0	43.6	43.3	43.7	43.2	−0.5	**+0.6**
Yes No	Complete Match	72.7	73.2	**74.2**	72.4	72.4	72.8	**+0.5**	**+1.1**
Yes No	1 Event	70.9	73.8	**77.5**	72.5	72.8	71.3	**+2.6**	**+7.8**
Yes No	2 Events	**75.8**	74.8	76.2	73.4	73.6	75.3	−1.2	−1.5

Table 10: Question Type × Overlap. Accuracies are %;

		Thinking			Non-Thinking			Δ Interaction (pp)	
QType	**Overlap**	NoMask	AGM	GT	NoMask	AGM	GT	AGM	GT
Yes No	No Overlap	72.2	74.4	**78.3**	72.6	73.0	72.2	**+1.8**	**+6.5**
Yes No	Overlap	72.0	73.2	**74.4**	72.6	72.6	72.4	**+1.2**	**+2.6**
Multiple Choice	No Overlap	53.2	55.3	54.2	49.2	47.6	48.4	**+3.7**	**+1.8**
Multiple Choice	Overlap	**53.0**	50.5	51.4	51.6	53.0	52.2	−3.9	−2.2
Open	No Overlap	42.5	43.0	**43.8**	41.8	41.2	41.9	**+1.1**	**+1.2**
Open	Overlap	39.6	39.8	**39.8**	40.6	40.8	41.0	**+0.0**	−0.2

F Prompt for Spatial Audio QA Pair Generation (with Thinking Steps)

The following is the full prompt used to generate spatial audio question–answer pairs with perceptual thinking steps.

```
You are an expert in spatial audio reasoning and question generation. Your task is
    to create high-quality QA pairs that evaluate a multimodal LLM's understanding
    of spatial audio scenes.

### Inputs:
- FrameTrends: A dictionary where each key is an event name and each value contains:
    - azimuth: {direction, span_category, crosses_center, start_side, end_side,
        is_arc}
    - distance: {trend_profile, variation_category}
    - temporal: {start_time, end_time, duration} if available
    - summary_text: A natural-language summary combining lateral, radial, and
        temporal info
    - Note: direction may include 'arc left->right' or 'arc right->left' for curved
        trajectories

This FrameTrends object is derived from all available scene data,
so treat it as the authoritative source for reasoning.

### Objectives:
Create diverse QA pairs that require reasoning about:
- Lateral trajectories (left->right, right->left, static)
- Arc trajectories (curved motion around the listener with constant radius)
- Radial changes (approach, recede, approach->recede)
- Relative motion between events (opposite directions, convergence/divergence)
- Sequencing and choreography (order, endpoints, overlap vs non-overlap if available
    )
- Comparative diagnostics (which spans wider azimuth? which ends farther?)
- Natural perceptual implications (e.g., "Which sound feels most noticeable near the
     middle?")

### Constraints:
- DO NOT ask for timestamps or exact numeric values.
- Use only qualitative info from FrameTrends (direction, span_category,
    trend_profile, overlaps).
- Use listener-centric, natural language: "left", "right", "middle", "closer", "
    farther".
- Avoid technical jargon like "binaural salience"; use intuitive terms like
"noticeable" or "prominent".
- Everything occurs within -90 degree to +90 degree in front of the listener,
so do not repeat "front stage" unnecessarily.
- Center crossing is optional; only mention if relevant.
- Avoid speculation beyond provided data; if ambiguous,
use cautious phrasing ("appears to", "likely") only when supported.
- **Thinking steps and rationale must sound like perceptual reasoning, not metadata
    citation.**
    - Do NOT mention FrameData, MetaInfo, or structured sources.
    - Use natural language based on the scene description and qualitative trends
    (e.g., "moves left to right and gets closer").
    - Avoid phrases like "according to data", "based on FrameTrends." or "described
        as"
    - Write as if you are analyzing the raw audio features present in the embedding.

### Output Format:
Return an array of QA objects in JSON:
{
  "id": "<unique_id>",
  "question": "<string>",
  "type": "yes_no | multiple_choice | open",
  "choices": [optional for multiple_choice],
  "answer": "<string>",
   "thinking": "<think>\nStep 1: [Describe step]\nSolution:
```

```
    [Provide scene-specific reasoning for this step]\n[Optional Step 2: ...]\
        nSolution: [...]\n[Optional Step 3: ...]
    \nSolution: [...]\nFinal reasoning: [Summarize why the answer is correct]\n</
        think>",
  "rationale": "<one-sentence perceptual reasoning>",
  "tags": ["lateral", "radial", "relative_motion", "sequencing", "comparative",
      ...],
}

### Dynamic Thinking Steps Rules:
- For simple questions (single attribute, , one event, yes/no):
    - Use 1 step + solution +Final reasoning.
- For moderate questions (two attributes OR involves more than one event OR simple
    comparison):
    - Use 2 steps  + solutions + Final reasoning.
- For complex questions
(comparative, multi-hop, choreography, relative motion, temporal reasoning):
    - Use 3 steps + solutions + Final reasoning.
- Always include Final reasoning summarizing why the answer is correct.
- Steps must be short, perceptual, and qualitative and followed by reasoning
    solution for that step
(e.g., "Check if the sound moves laterally").
- Do NOT include numeric values or cite structured data.

### Example Thinking Sections:
**Simple Yes/No Question**
<think>
Step 1: Check if the smoke detector moves laterally.
Solution: The smoke detector stays fixed on the right without sweeping.
Final reasoning: It remains on the right, so it is static in azimuth.
</think>

**Moderate Question (two events)**
<think>
Step 1: Compare lateral spans.
Solution: The train sweeps widely; the detector stays fixed.
Step 2: Compare radial behavior.
Solution: The train approaches then recedes; the detector remains steady.
Final reasoning: The train is dynamic both laterally and radially,
while the detector is static.
</think>

**Complex Question**
<think>
Step 1: Identify lateral behavior of both sources.
Solution: The smoke detector is static on the right; the train moves right->left
    across the center.
Step 2: Identify radial behavior of both sources.
Solution: The smoke detector maintains a steady distance; the train approaches then
    recedes.
Step 3: Contrast them to explain differences.
Solution: One is stationary and steady; the other sweeps widely and varies strongly
    in distance.
Final reasoning: The train is dynamic both laterally and radially, while the
    detector is static.
</think>

### Required Mix:
- At least 5 Yes/No questions (fundamentals)
- At least 3 Multiple Choice questions (comparatives/diagnostics)
- At least 6 Open-ended questions (expanded reasoning),
including at least 3 focused on RELATIVE MOTION between events
- At least 4 questions must involve radial reasoning (approach/recede)
- At least 2 questions should summarize overall choreography or sequencing
- At least 2 questions should compare lateral and radial trends together
- If temporal info exists, include at least 2 questions about order or overlap

### Open-Ended Reasoning Templates (adapt naturally):
1) **Qualitative Trend Summary**
   - "Summarize how [Event] changes in distance while moving laterally."
```

```
2) **Relative Motion Contrast**
   - "Compare the movement of [Event A] and [Event B] in both direction and distance
       ."
3) **Stage Choreography**
   - "Narrate the overall motion: where each source starts, how it moves, and where
       it finishes."
4) **Convergence/Divergence**
   - "Do [Event A] and [Event B] move toward the same region or away from each other
       ? Explain."
5) **Endpoint Reasoning**
   - "Explain where each source ends and how its final position follows from its
       path."
6) **Prominence Near Middle**
   - "Which source feels most noticeable near the middle and why?"
7) **Comparative Span**
   - "Which source spans a wider left->right range, and how does that affect the
       scene's balance?"
8) **Temporal Order or Overlap**
   - "Which event starts first?" or "Do [Event A] and [Event B] overlap in time?"
9) **Path Characteristics**
   - "Describe the path that [Event] takes as it moves.
   How does its distance change relate to its lateral movement?"
10) **Motion Pattern Comparison**
   - "Compare how [Event A] and [Event B] move through space.
   What makes their paths different?"

### Procedure (internal; do not output steps):
1) Use FrameTrends.summary_text and structured fields for reasoning.
2) Decide question complexity and adjust number of thinking steps dynamically:
    - If question involves >1 event -> at least moderate (minimum 2 steps).
3) For each step, provide both the step description and the solution
for that step based on the scene.
4) Generate <think> section first, then answer, then rationale.
5) Ensure coverage of lateral, radial, relative motion, sequencing, and comparative
    reasoning.
6) Validate for internal consistency; if ambiguity exists, use cautious wording.
```

G Evaluation of QA pairs

G.1 Prompt for Evaluating Thinking Steps

```
       "You are evaluating the reasoning process (thinking) for a spatial audio
           question.

Question: {question}
Ground Truth Answer: {ground_truth_answer}

Ground Truth Thinking:
{ground_truth_thinking}

Predicted Thinking:
{predicted_thinking}

Evaluate the predicted thinking on four dimensions:

1. **Logical Coherence** (0-5):
Does the reasoning flow logically from step to step?
   - 5: Perfect logical flow, each step follows naturally
   - 4: Mostly coherent with minor logical gaps
   - 3: Generally logical but some unclear connections
   - 2: Multiple logical gaps or unclear reasoning
   - 1: Poor logical structure
   - 0: Incoherent or contradictory reasoning

2. **Step Completeness** (0-5):
Are all necessary reasoning steps present?
   - 5: All necessary steps present and well-developed
   - 4: Most steps present, minor omissions
   - 3: Key steps present but some missing
```

```
   - 2: Several important steps missing
   - 1: Many critical steps missing
   - 0: Minimal or no reasoning steps

3. **Factual Accuracy** (0-5):
Are the spatial facts and observations correct?
   - 5: All spatial facts correct (directions, distances, movements)
   - 4: Minor factual errors that don't affect conclusion
   - 3: Some factual errors but core understanding present
   - 2: Multiple factual errors affecting reasoning
   - 1: Major factual errors throughout
   - 0: Completely incorrect facts

4. **Alignment with Ground Truth** (0-5):
How similar is the reasoning approach to the ground truth?
   - 5: Nearly identical reasoning approach and steps
   - 4: Very similar approach with minor differences
   - 3: Similar overall approach but different execution
   - 2: Different approach but reaches similar conclusions
   - 1: Significantly different approach
   - 0: Completely different reasoning

Also identify any specific errors in the thinking process. Categorize each error as:
- **Spatial reasoning errors**: lateral_direction, distance_depth, crossing_center,
temporal_sequence
- **Logical errors**: missing_steps, incorrect_flow, contradictions,
    incomplete_analysis
- **Factual errors**: source_misidentification, attribute_swapping, magnitude_errors
- **Consistency errors**: thinking_answer_mismatch, internal_contradictions

For each error, specify severity: minor, moderate, or major.

Provide your evaluation in this exact JSON format:
{{
    "logical_coherence": <score 0-5>,
    "step_completeness": <score 0-5>,
    "factual_accuracy": <score 0-5>,
    "alignment_with_ground_truth": <score 0-5>,
    "explanation": "<brief explanation of scores>",
    "errors": [
        {{"type": "<error_type>", "category": "<category>", "severity": "<severity
            >",
        "description": "<description>"}},
        ...
    ]
}}"
```

G.2 Prompt for Evaluating Rationale

```
"You are evaluating the rationale (brief explanation) for a spatial audio answer.

Question: {question}
Ground Truth Answer: {ground_truth_answer}
Predicted Answer: {predicted_answer}

Ground Truth Rationale:
{ground_truth_rationale}

Predicted Rationale:
{predicted_rationale}

Evaluate the predicted rationale on three dimensions:

1. **Conciseness** (0-5): Is the rationale appropriately brief yet complete?
   - 5: Perfect balance of brevity and completeness
   - 4: Mostly concise with minor verbosity or slight incompleteness
   - 3: Acceptable length but could be more concise or complete
   - 2: Too verbose or too brief, missing key points
   - 1: Significantly too long or too short
```

```
   - 0: Extremely verbose or minimal content

2. **Accuracy** (0-5): Does it correctly summarize the spatial relationships?
   - 5: All spatial facts correct and well-summarized
   - 4: Minor inaccuracies that don't affect understanding
   - 3: Some inaccuracies but core message correct
   - 2: Multiple inaccuracies affecting the summary
   - 1: Major inaccuracies throughout
   - 0: Completely incorrect summary

3. **Clarity** (0-5): Is it easy to understand?
   - 5: Crystal clear and easy to understand
   - 4: Clear with minor ambiguities
   - 3: Generally clear but some confusing parts
   - 2: Somewhat unclear or confusing
   - 1: Very unclear or hard to follow
   - 0: Incomprehensible

Also identify any specific errors in the rationale. Categorize each error as:
- **Spatial reasoning errors**: lateral_direction, distance_depth, crossing_center,
temporal_sequence
- **Factual errors**: source_misidentification, attribute_swapping, magnitude_errors
- **Consistency errors**: rationale_answer_mismatch, contradictions

For each error, specify severity: minor, moderate, or major.

Provide your evaluation in this exact JSON format:
{{
    "conciseness": <score 0-5>,
    "accuracy": <score 0-5>,
    "clarity": <score 0-5>,
    "explanation": "<brief explanation of scores>",
    "errors": [
        {{"type": "<error_type>", "category": "<category>", "severity": "<severity
            >",
        "description": "<description>"}},
        ...
    ]
}}"
```

G.3 Evaluation of Open ended answers

```
"You are evaluating a spatial audio reasoning answer.
Compare the predicted answer against the ground truth.

Question: {question}

Ground Truth Answer: {ground_truth}

Predicted Answer: {predicted}

Evaluate the predicted answer on two dimensions:

1. **Factual Accuracy** (0-5): Are the spatial facts correct (directions, distances,
     movements,
source identifications)?
   - 5: All facts are correct
   - 4: Minor factual errors that don't change the main conclusion
   - 3: Some factual errors but core understanding is present
   - 2: Multiple factual errors affecting the conclusion
   - 1: Major factual errors throughout
   - 0: Completely incorrect facts

2. **Semantic Similarity** (0-5): How well does the predicted answer capture the
    meaning and
completeness of the ground truth?
   - 5: Captures all key points with equivalent meaning
   - 4: Captures most key points with minor omissions
   - 3: Captures main idea but misses some important details
```

```
    - 2: Partially captures the idea with significant gaps
    - 1: Minimal semantic overlap
    - 0: Completely different meaning

Provide your evaluation in this exact JSON format:
{{
    "factual_accuracy": <score 0-5>,
    "semantic_similarity": <score 0-5>,
    "explanation": "<brief explanation of scores>"
}}"
```

Table 11: Selected samples to illustrate model reasoning and responses across masking regimes

Question	Model Responses
Which sound spans a wider left–right range? Choices : [Whimper (dog), Tick-tock, They span the same range]	**Ground truth answer:** Whimper (dog) **NoMask without thinking mode** **Thinking:** **Answer:** Tick-tock **Rationale:** Tick-tock is described as a wide sweep while the dog's movement is moderate, so Tick-tock spans more left–right range. **NoMask with thinking mode** **Thinking:** Step 1: Compare the span categories of both sounds. Solution: Both the whimper and tick-tock are described with a wide lateral span. Step 2: Decide which is wider based on those descriptions. Solution: Since both are labeled wide, they span the same wide left–right range. Final reasoning: Both sources have wide sweeps, so neither spans a wider left–right range than the other. **Answer:** They span the same range **Rationale:** Both the whimper and tick-tock are described as having wide lateral spans, so they cover the same wide left–right range. **AGM with thinking** **Thinking:** Step 1: Compare the azimuth span of the two sounds. Solution: The whimper has a wide left->right sweep crossing the center, while the tick-tock is static on the right. Step 2: Decide which is wider. Solution: The whimper clearly spans a wider left–right range than the static tick-tock. Final reasoning: The whimper's wide sweep contrasts with the tick-tock's fixed position, so the whimper spans more azimuth. **Answer:** Whimper (dog) **Rationale:** The whimper performs a wide left->right sweep, whereas the tick-tock remains fixed on the right, so the whimper spans more azimuth. **GT with thinking**

Question	Model Responses (continued)
	Thinking: Step 1: Compare the lateral span of the two sounds. Solution: The whimper has a wide sweep across the scene, while the tick-tock is static on the right. Step 2: Conclude which spans more left–right range. Solution: The whimper clearly spans a wider range than the static tick-tock. Final reasoning: Because the whimper sweeps widely from left to right while the tick-tock stays fixed on the right, the whimper spans the wider range. **Answer:** Whimper (dog) **Rationale:** The whimper performs a wide left->right sweep whereas the tick-tock remains static on the right, so the whimper spans more azimuth.
Does the wind chime sweep all the way across the center to the left side? (Yes/No)	**Ground truth answer:** no **NoMask without thinking mode** **Thinking:** **Answer:** Yes **Rationale:** The wind chime follows an arc that crosses the center and ends on the left side. **NoMask with thinking mode** **Thinking:** Step 1: Check the wind chime's lateral path. Solution: The wind chime moves from the right toward the left and is described as crossing the center. Final reasoning: Because it travels right->left and crosses the center, it goes all the way across to the left side. **Answer:** yes **Rationale:** The wind chime moves from right to left and crosses the center, so it reaches the left side. **AGM with thinking mode** **Thinking:** Step 1: Check the lateral span of the wind chime. Solution: The wind chime moves from right to left and is described as a narrow sweep that does not cross center. Final reasoning: Because its sweep is narrow and does not cross the center, it does not reach the left side and thus does not sweep across the middle to the left. **Answer:** no **Rationale:** The wind chime has a narrow right->left motion that stays on the right and does not cross center. **gt_thinking** **Thinking:** Step 1: Check the wind chime's lateral sweep and whether it crosses center. Solution: The wind chime moves from right to left but is described as a narrow sweep that stays on the right and does not cross center. Final reasoning: Because its sweep is narrow and it remains on the right side, it does not cross the center to reach the left. **Answer:** No

Question	Model Responses (continued)
	Rationale: The wind chime's narrow right->left sweep remains on the right and does not cross the center.
Which source both moves right->left and shows a receding distance? choices: [Power tool, Coin (dropping), Bicycle bell, Both Power tool and Coin (dropping)	**Ground truth answer:** Power tool **NoMask without thinking mode** **Thinking:** **Answer:** Both Power tool and Coin (dropping) **Rationale:** Both the power tool and the coin move right->left and are described as receding, unlike the bicycle which approaches. **NoMask with thinking mode** **Thinking:** Step 1: Check lateral directions for each source. Solution: Power tool and Coin both move right->left; Bicycle bell moves left->right. Step 2: Check radial trends for each source. Solution: Power tool recedes, Coin recedes, Bicycle bell approaches. Step 3: Combine lateral and radial. Solution: Both Power tool and Coin move right->left and recede, so they match the description. Final reasoning: Since both Power tool and Coin move right->left and are receding, they are the correct pair. **Answer:** Both Power tool and Coin (dropping) **Rationale:** Both the Power tool and Coin travel right->left and are described as receding, so they fit the condition. **AGM with thinking mode** **Thinking:** Step 1: Check lateral direction for each source. Solution: Both Power tool and Coin move right->left. Step 2: Check radial trends for each source. Solution: Power tool is receding; Coin is approaching. Final reasoning: Only the Power tool both moves right->left and recedes. **Answer:** Power tool **Rationale:** The Power tool moves right->left and recedes, while the Coin moves right->left but approaches. **GT with thinking mode** **Thinking:** Step 1: Check lateral directions for both candidates. Solution: Both the Power tool and the Coin move right->left. Step 2: Check radial trends for both candidates. Solution: The Power tool recedes while the Coin approaches. Final reasoning: Only the Power tool matches both right->left lateral motion and receding distance. **Answer:** Power tool **Rationale:** The Power tool is the only source that combines a right->left sweep with a receding distance trend.

Question	Model Responses (continued)
Summarize how Oink changes in distance while it moves from left to right. (Open ended)	**Ground truth answer:** Oink moves left to right while approaching the listener. **NoMask without thinking mode** **Thinking:** **Answer:** Oink moves from left to right while first approaching the listener and then receding, so it feels closer as it nears the center and then moves away toward the right. **Rationale:** Oink's left->right path includes an approach then a recede, so it feels closer near the center and then farther as it continues right. **NoMask with thinking mode** **Thinking:** Step 1: Identify Oink's lateral motion. Solution: Oink moves left->right in a wide sweep that crosses the center. Step 2: Describe its radial trend during that sweep. Solution: Its distance follows an approach->recede pattern, so it gets closer as it nears the center and then moves away. Final reasoning: The combination of a wide left->right sweep and an approach->recede distance profile means Oink comes in toward the listener near the middle and then recedes toward the right. **Answer:** Oink moves from left to right while first approaching the listener and then receding, producing a strong sense of coming closer near the middle before moving away. **Rationale:** Oink's left->right sweep is paired with an approach->recede distance profile, so it draws nearer near the center and then moves away. **AGM with thinking mode** **Thinking:** Step 1: Describe Oink's lateral motion. Solution: Oink travels left->right across a wide arc and crosses the center. Step 2: Describe Oink's radial trend during that lateral motion. Solution: Its distance profile is an approach with large variation, so it gets closer as it moves. Final reasoning: Combining the wide left->right sweep with a clear approach trend explains how Oink moves closer while sweeping across the scene. **Answer:** Oink moves from left to right while approaching the listener; its distance decreases as it sweeps across, producing a clear approach during the left->right motion. **Rationale:** Oink performs a wide left->right sweep while approaching, so it becomes closer as it moves across the listener. **GT with thinking** **Thinking:** Step 1: Identify Oink's lateral movement. Solution: Oink moves left->right with a moderate sweep and crosses the center. Step 2: Identify Oink's radial trend while moving. Solution: Oink follows an approach trend, meaning it becomes closer over its path. Final reasoning: Together, Oink sweeps across the listener and approaches, so it grows closer as it travels from left to right.

Question	Model Responses (continued)
	Answer: Oink moves from left to right and steadily approaches the listener, becoming noticeably closer as it travels across the scene. **Rationale:** Oink's left->right sweep combined with an approaching distance profile means it becomes closer as it moves across the listener.
Do Humming and Purr move in the same lateral direction? (Yes/No)	**Ground truth answer:** yes **NoMask without thinking mode** **Thinking:** **Answer:** no **Rationale:** Humming goes left->right while Purr goes right->left, so their lateral directions differ. **NoMask with thinking mode** **Thinking:** Step 1: Compare the lateral directions of both sounds. Solution: Humming moves left->right while Purr moves right->left, so they move in opposite directions. Final reasoning: Their lateral directions are opposite, not the same. **Answer:** No **Rationale:** Humming goes left->right while Purr goes right->left, so they move oppositely. **AGM with thinking mode** **Thinking:** Step 1: Compare the lateral directions of both sounds. Solution: Both Humming and Purr are described as moving left->right across the scene. Final reasoning: Since both are moving left->right, they share the same lateral direction. **Answer:** yes **Rationale:** Both Humming and Purr sweep from left to right, so their lateral directions match. **GT with thinking mode** **Thinking:** Step 1: Compare the lateral directions of Humming and Purr. Solution: Both are described as moving right->left. Final reasoning: Since both travel from right to left, they share the same lateral direction. **Answer:** yes **Rationale:** Both Humming and Purr sweep right->left, so their lateral directions are the same.