

Egocentric Spatial Reference Resolution under ASR Noise in Traditional Chinese Conversational Navigation

Yu-Ti Huang

Trans-disciplinary Bachelor Degree Program,
National Taiwan University
kvoyuti.93@gmail.com

Abstract

Resolving egocentric spatial expressions in conversational navigation requires mapping speaker-centric references to allocentric orientations. In practice, this process is often mediated by automatic speech recognition (ASR), where transcription errors, linguistic variation, and referential ambiguity obscure spatial relations. Despite progress in navigation and multimodal reasoning, egocentric spatial reference resolution under ASR-transcribed, non-English conversational input remains underexplored. We introduce Conversational Orientation Reasoning (COR), a diagnostic benchmark for egocentric spatial reference resolution in Traditional Chinese conversational navigation. COR consists of 10,000 instances pairing ASR-transcribed utterances with structured landmark coordinates derived from real-world environments projected onto a 50×50 grid. COR isolates the effects of spoken language noise from geometric computation, enabling evaluation of transcription errors, linguistic variation, and referential ambiguity beyond end-to-end navigation success. Using COR, we study an interpretable decomposition of orientation inference into three steps: extracting spatial relations, mapping landmark coordinates to absolute directions, and inferring the speaker's orientation. Models following this decomposition achieve 99.6% accuracy on clean text and retain 95.7% accuracy under ASR-transcribed input. COR supports analysis of failure modes in spoken spatial reference resolution, providing insights into how language grounding degrades under noisy conversational input. The code¹ and dataset² are publicly available under the MIT license.

¹<https://github.com/yu-ti-huang/Conversational-Orientation-Reasoning>

²<https://huggingface.co/datasets/yu-ti-huang/Conversational-Orientation-Reasoning>

1 Introduction

Egocentric spatial reference resolution concerns mapping speaker-centric expressions such as “the exit is on my right” to allocentric orientations including north, south, east, and west. In realistic settings, this process is often mediated by automatic speech recognition (ASR), where transcription errors, disfluencies, and referential ambiguity can obscure spatial relations and perspective cues. As a result, failures in navigation systems frequently arise not from geometric reasoning itself, but from misinterpretation of egocentric language under noisy spoken input.

Despite growing interest in conversational navigation systems (Sundar et al., 2024; Pareek et al., 2024; Sheshadri and Hara, 2024; Kaniwa et al., 2024; Liu et al., 2024; Levi and Kadar, 2025), most existing approaches assume access to GPS signals, detailed maps, or fixed global reference frames (de Vries et al., 2018; Chen et al., 2020), primarily focus on English-language inputs, and emphasize end-to-end navigation success (Ghosh et al., 2024; Tang et al., 2023). While effective in aggregate performance, such formulations provide limited diagnostic insight into how ASR-induced noise and linguistic ambiguity affect the interpretation of egocentric spatial references, particularly in non-English conversational settings. In conversational navigation, spoken Traditional Chinese commonly exhibits fillers, discourse particles, and diverse realizations of spatial expressions, which can amplify referential ambiguity under ASR.

Related progress in embodied AI, multimodal chain-of-thought reasoning (Mu et al., 2023; Sun et al., 2024; Liu et al., 2025; Shen et al., 2025), and large audio-language models (Zhang et al., 2023; Xie and Wu, 2024; Fu et al., 2025; Défossez et al., 2024) has further advanced navigation, planning, and speech processing, but typically assumes known orientations or primarily focuses on

transcription, summarization, or intent-level tasks (yu Huang et al., 2024; Yang et al., 2024; Wang et al., 2025; Shi et al., 2025). Even recent speech-based reasoning efforts such as Audio-CoT (Ma et al., 2025) do not directly study egocentric spatial reference resolution under ASR-mediated noise.

To address this gap, we introduce Conversational Orientation Reasoning (COR), a diagnostic benchmark for egocentric spatial reference resolution in Traditional Chinese conversational navigation. COR pairs ASR-transcribed egocentric utterances with structured landmark coordinates derived from real-world urban layouts in Taiwan and projected onto a 50×50 grid representation. By controlling spatial geometry, COR enables targeted analysis of uncertainty introduced by spoken language, including transcription errors, linguistic variation, and referential ambiguity. Using COR, we analyze an interpretable decomposition of orientation inference into spatial relation extraction, coordinate-based direction identification, and final orientation inference. We evaluate performance under clean text, ASR-transcribed inputs, and multiple evaluation conditions, including unseen environments and ambiguous references. Our study is guided by three research questions:

- **RQ1 (Orientation inference performance).** How accurately can models resolve egocentric spatial references into allocentric orientations under ASR-transcribed input?
- **RQ2 (Ablation studies).** How do ASR-transcribed input, spatial coordinates, and structured output formats contribute to performance under transcription noise?
- **RQ3 (Evaluation under variation).** How does egocentric spatial reference resolution behave under linguistic variation, referential ambiguity, and unseen spatial environments?

Our contributions are threefold:

1. **Benchmark.** We introduce COR, a diagnostic benchmark comprising 10,000 instances derived from real-world Taiwanese urban layouts for egocentric-to-allocentric spatial reference resolution under ASR-transcribed Traditional Chinese conversational input.
2. **Analysis framework.** We use a structured decomposition of orientation inference to examine how transcription noise and linguistic ambiguity affect spatial language grounding.

3. **Empirical analysis.** Through evaluations across linguistic variation, referential ambiguity, and unseen spatial domains, we analyze failure modes that are obscured by end-to-end navigation metrics.

2 Related Work

2.1 Egocentric Spatial Language and Conversational Navigation

Egocentric spatial reference resolution concerns interpreting speaker-centric expressions into allocentric orientations with respect to a world-centric reference frame. This problem is embedded in natural language navigation tasks, where agents follow instructions or dialogues grounded in spatial environments. Early benchmarks such as Room-to-Room (R2R) (Anderson et al., 2018) and Cooperative Vision-and-Dialog Navigation (CVDN) (Thomason et al., 2019) focus on visually grounded instruction following, while subsequent work extends navigation to urban and conversational settings, including Talk the Walk (de Vries et al., 2018) and Touchdown (Chen et al., 2020).

In these benchmarks, egocentric spatial language is typically evaluated only through end-to-end navigation success, making it difficult to isolate errors arising from linguistic misinterpretation. Other studies incorporate spatial representations, such as panoramic orientation in Speaker-Follower models (Fried et al., 2018), 3D scene graphs in SpatialRGPT (Cheng et al., 2024), and large-scale egocentric visual data in Ego4D (Grauman et al., 2022). Existing approaches generally assume known agent orientations or operate over high-level action spaces, rather than evaluating the intermediate task of resolving egocentric spatial references into allocentric orientations. As a result, failures in egocentric spatial interpretation caused by spoken language noise are often entangled with downstream perception or action errors, and remain insufficiently characterized as a standalone language understanding problem, particularly in non-English conversational settings.

2.2 Multimodal Chain-of-Thought Reasoning

Chain-of-thought (CoT) prompting encourages models to generate intermediate reasoning steps and has been shown to improve performance across a wide range of reasoning tasks (Jin et al., 2024; Wang et al., 2023; Prystawski et al., 2023; Tang et al., 2023). Extensions to multimodal chain-

of-thought (MCoT) enable reasoning over language, vision, and spatial inputs, allowing embodied agents to decompose complex tasks into interpretable subgoals. Representative systems include EmbodiedGPT (Mu et al., 2023), E-CoT (Lin et al., 2024), Emma-X (Sun et al., 2024), SpatialCoT (Liu et al., 2025), and MCoCoNav (Shen et al., 2025). Prior work employs MCoT to improve planning, action execution, or multimodal perception. In contrast, our work uses structured reasoning to examine how egocentric spatial relations expressed in language are transformed into allocentric orientation labels. This perspective enables our analysis of spoken language understanding failures under ASR-induced corruption. COR serves as a controlled diagnostic setting for studying how spoken language noise affects spatial reference resolution.

2.3 Spoken Language Understanding and Audio-Language Models

Large audio-language models (LALMs) extend language models with speech inputs and demonstrate strong performance on transcription, classification, and dialogue tasks (Zhang et al., 2023; Xie and Wu, 2024; Fu et al., 2025; Défossez et al., 2024). Benchmarks such as SALMONN (Tang et al., 2024), AudioBench (Wang et al., 2025), AirBench (Yang et al., 2024), and Dynamic-SUPERB (Yu Huang et al., 2024) primarily evaluate speech perception quality and end-to-end reasoning performance. Audio-CoT (Ma et al., 2025) further explores reasoning over speech-derived representations. In contrast to these efforts, our work focuses on ASR-mediated language understanding rather than acoustic modeling or speech recognition itself. We treat ASR transcripts as a noisy textual modality and study how transcription errors and linguistic ambiguity affect egocentric spatial reference resolution when combined with structured spatial inputs. This setting isolates downstream language grounding errors introduced by ASR, complementing existing audio-language benchmarks.

3 Method

This section defines the orientation reasoning task and the controlled setting used in COR. We describe our stage-wise curriculum training procedure and the structured output format used for inference and analysis under ASR-transcribed inputs.

3.1 Overview

We study egocentric-to-allocentric orientation inference in a controlled conversational navigation setting. Given a clean egocentric description A , we generate an ASR transcript A' using a text-to-speech (TTS) and ASR pipeline. This introduces controlled transcription noise while keeping the underlying spatial semantics unchanged. The transcript A' is paired with structured spatial coordinates T that describe the user position and nearby landmarks. A model is trained to produce a structured reasoning trace and a final allocentric orientation prediction $D^* \in \{\text{north, east, south, west}\}$.

As illustrated in Figure 1, our setup consists of three components: (1) *Speech synthesis and transcription*, which converts clean descriptions into ASR transcripts; (2) *Multimodal input preparation*, which serializes transcripts and spatial coordinates into a unified input; and (3) *Orientation inference*, which predicts the allocentric orientation and optionally produces intermediate steps for analysis. Unlike prior audio-language models that emphasize transcription quality, we treat ASR output as a noisy textual modality. This setup yields ASR-transcribed inputs with a well-characterized noise profile, enabling analysis of how transcription errors affect downstream spatial reference resolution without introducing additional variability from acoustic recording conditions.

3.2 Task Formulation

We study egocentric-to-allocentric orientation reasoning in a structured setting derived from the Gongguan MRT area in Taiwan (Figure 2). Let A denote a clean egocentric description, A' its ASR transcript, T the spatial coordinate input, and D^* the ground-truth allocentric orientation.

Environment. To evaluate the robustness of spatial reasoning in a high-resolution context, the real-world Gongguan MRT area in Taipei is projected into a discrete 50×50 Manhattan grid $\mathcal{G}$. The user position is $u = (x_u, y_u) \in \mathcal{G}$, and physical landmarks form a sparse set $\mathcal{L} = \{\ell_i\}$ with coordinates $p(\ell_i) \in \mathcal{G}$ occupying valid locations on the grid. This 50×50 representation expands the state space to 2,500 potential positions, thereby mitigating the risk of the model memorizing specific coordinate-to-landmark mappings while intentionally removing geometric ambiguity (e.g., diagonal relations). By fixing spatial geometry across instances, COR isolates uncertainty arising from language interpre-

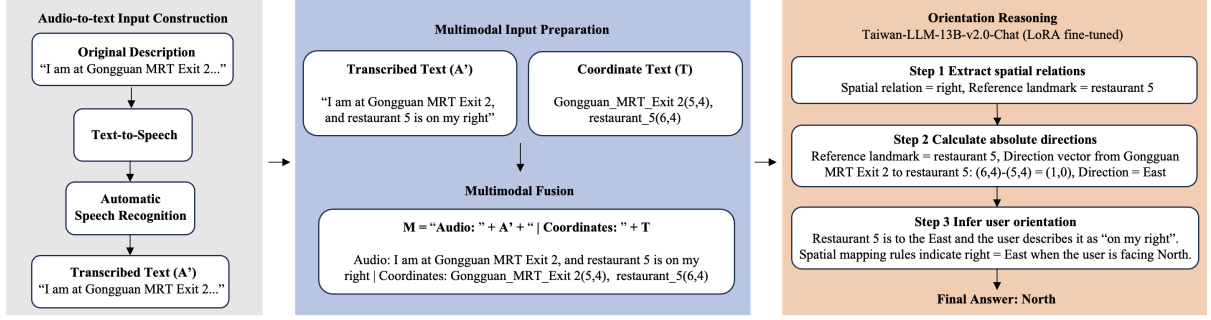


Figure 1: Pipeline of our setup: (1) speech synthesis and transcription, (2) multimodal input preparation and fusion, and (3) orientation inference.

tation and transcription noise rather than spatial computation.

Egocentric description. Each instance contains an egocentric spatial relation expressed in natural language and a reference landmark ℓ_r . For clear directional expressions, the underlying egocentric relation $q \in \{\text{FRONT, BACK, LEFT, RIGHT}\}$ can be uniquely identified. In addition, COR includes vague directional expressions commonly observed in conversational navigation, such as expressions indicating approximate or imprecise orientation (e.g., “slightly to the left front”, “roughly on the left side”). For such cases, the egocentric relation does not correspond to a single discrete direction.

After TTS and ASR, we obtain the transcript A' . For example, the utterance “*I am at Exit 2, and restaurant 5 is on my right*” implies that the user is facing north, i.e., $D^* = \text{north}$. In contrast, utterances describing vague directions may admit multiple plausible allocentric orientations.

Mapping rule. The mapping rules (Table 1) define the labeling protocol used to derive allocentric orientations in COR. They are not intended as a proposed solution. Instead, they specify how geometric relations are fixed in order to study language understanding under noise. These rules are used to derive ground-truth labels and are never executed or enforced within the model during inference.

Given the reference landmark ℓ_r , we compute the relative vector $\Delta = p(\ell_r) - u = (\Delta_x, \Delta_y)$. The absolute direction of the landmark is determined by the dominant axis: east or west if $|\Delta_x| > |\Delta_y|$, and north or south otherwise, with the sign given by Δ_x or Δ_y .

For clear egocentric relations, the user orientation is derived by rotating the absolute landmark direction according to the relation q , yielding a single allocentric orientation. For vague directional

expressions, however, this procedure yields a set of plausible allocentric orientations rather than a unique label. Accordingly, the ground-truth orientation is annotated as a set

$$D^* \subseteq \{\text{north, east, south, west}\},$$

reflecting intrinsic ambiguity in directional interpretation.

Since the environment is a Manhattan grid with only four cardinal neighbors, diagonal cases do not occur. All ground-truth orientation sets are derived from the same geometric protocol. While the labeling rules are deterministic given a fixed interpretation, the presence of vague directional expressions results in inherently ambiguous orientation annotations. Each instance is associated with a step-by-step reasoning trace used to supervise the structured output format. All automatically generated instances are verified by human annotators for correctness.

Facing	Front	Back	Right	Left
North	N	S	E	W
East	E	W	S	N
South	S	N	W	E
West	W	E	N	S

Table 1: Relative-to-absolute direction mapping rules.

3.3 Training Procedure

We train Taiwan-LLM-13B-v2.0-Chat with a stage-wise curriculum to stabilize learning of the decomposed inference steps. All training stages use clean text inputs, while ASR-transcribed inputs are introduced only at evaluation time to examine the effects of transcription noise.

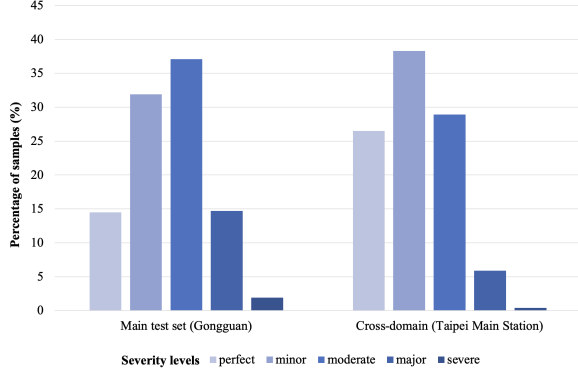


Figure 3: ASR error severity distribution in the two evaluation sets.

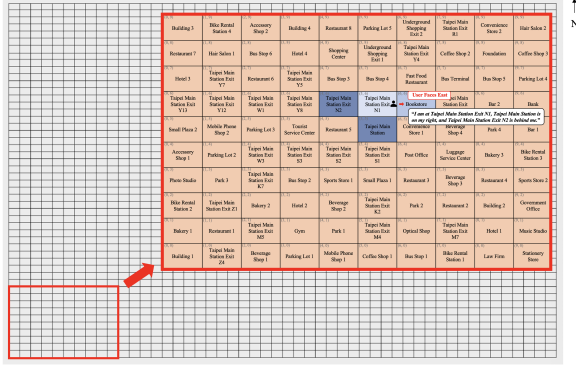


Figure 4: Cross-domain evaluation environment. The Taipei Station area is projected into a 50×50 grid to verify the model’s spatial reasoning logic in an unseen environment. Only representative landmarks are labeled for readability.

Evaluation subsets. Beyond the main test set (1,500 examples from the Gongguan area), we construct two additional evaluation sets for RQ3: (1) Cross-domain (1,200 examples), projected into a 50×50 grid from the unseen Taipei Station area (Figure 4); and (2) Referential ambiguity (400 examples), which include ambiguous references, disfluent or incomplete utterances, and semantically underspecified mentions.

ASR error profile. Figure 3 shows the ASR error severity distribution across the main and cross-domain test sets. Exact counts are provided in Appendix 5 (Table 7).

Example. “I am at Gongguan MRT Exit 2, and restaurant 5 is on my right” $\rightarrow$ Label: North.

4.2 Orientation inference performance (RQ1)

We report *orientation accuracy* as the primary metric, defined as the proportion of predictions matching the ground-truth orientation. For methods that

ID	Method	Input	Acc.	Err.
B0	Rule-based	Clean	99.86	–
		ASR	39.51	–
B1	Zero-shot	Clean	25.0	5.2
B2	Few-shot (no CoT)	Clean	25.9	4.7
B3	Few-shot (CoT)	Clean	21.1	39.2
B4	Fine-tuned	Clean	12.8	50.4
B5	MCoT (ours)	Clean	99.6	0.0
		ASR	95.7	0.0

Table 2: Orientation accuracy (%). **Err.** denotes format error rate (%).

ASR Severity	CER (%)	MCoT Acc. (%)
Perfect	0	99.6
Minor	$0 < \text{CER} \leq 5$	97.8
Moderate	$5 < \text{CER} \leq 15$	95.4
Severe	$\text{CER} > 15$	92.1

Table 3: MCoT performance across different ASR error severity levels (CER).

generate reasoning traces, we additionally report *format error rate*, defined as the proportion of outputs violating the expected schema.

Rule-based and Fuzzy Matching baseline. We implement a deterministic baseline that (i) extracts relation–landmark pairs using pattern matching and (ii) applies the fixed mapping rules in Table 1. While this baseline achieves 99.86% accuracy on clean text, its performance collapses to 39.51% under ASR noise. Crucially, even when augmented with fuzzy matching techniques (e.g., Levenshtein distance and pinyin-based normalization), the rule-based approach remains fragile against Traditional Chinese ASR corruptions that involve phonetic homophones and orthographic shifts. These errors frequently result in valid but incorrect character sequences, which cause 40.8% of cases to fail at the extraction stage. This gap proves that the bottleneck is not geometric calculation but the failure of symbolic systems to reconcile noisy linguistic references with structured spatial data. Table 2 summarizes results across baselines and our method. See Appendix A.6 for a detailed analysis of fuzzy matching ineffectiveness.

Quantitative Results. The curriculum-trained model achieves 99.6% accuracy on clean inputs and 95.7% accuracy on ASR-transcribed inputs. As shown in Table 2, its performance exceeds that of non-CoT baselines under both settings. The ex-

Error	Description
Direction	Incorrect relative-to-absolute mapping
Relation	Landmark or relation misidentified
ASR	Errors caused by ASR transcription

Table 4: Error taxonomy used in our analysis.

perimental design intentionally isolates linguistic grounding under ASR noise by fixing the geometric environment and excluding ambiguous spatial configurations. As evidenced by the rule-based baseline, even with access to the same geometric protocol, performance degrades sharply under ASR corruption, suggesting that the dominant challenge lies in resolving noisy spoken references rather than spatial computation. Our model does not have access to the labeling rules used to generate orientation annotations. It learns to map noisy utterances to egocentric spatial relations through decomposition and constrained reasoning. This design accounts for the observed performance while preserving failure modes under ASR noise. We observe that direct classification baselines perform poorly under ASR noise; notably, the fine-tuned baseline (B4) suffers catastrophic formatting failures when forced to simultaneously resolve phonetically corrupted landmarks and map them across the 50×50 grid without a reasoning scaffold, further motivating our decomposition-based approach.

4.3 Qualitative Analysis of Failure Modes

To characterize the linguistic bottleneck, we examine the reasoning traces produced by our MCoT model. We categorize failures into three primary modes: Direction, Relation, and ASR errors. Table 4 summarizes their definitions, and we provide the representative example for each type with English translation output in Appendix A.1.

ASR Misrecognition Error. This mode represents the most significant challenge introduced by spoken input, where ASR misrecognition causes the model to incorrectly map multiple spatial relations to the same landmark, leading to inconsistent reasoning. As shown in Figure 5, while the model incorrectly extracts relations in Step 1 due to phonetic corruption (e.g., yin-liu-dian), it still correctly performs coordinate-based direction computation in Step 2 and maintains internal mapping logic in Step 3. This highlights that the bottleneck lies in language grounding under noise.

ID	Setting	Acc.	Err.
A1	Clean text only	25.0	0.7
A2	ASR text only	16.2	35.8
A3	ASR + coordinates	26.4	3.0
A4	MCoT (Clean)	99.6	0.0
	MCoT (ASR)	95.7	0.0

Table 5: Ablation results. Non-MCoT settings (A1–A3) use direct prediction without reasoning traces. **Err.** denotes format error rate (%).

4.4 Ablation studies (RQ2)

We conduct ablation studies to assess the contributions of ASR transcripts, spatial coordinates, and structured chain-of-thought reasoning. Results are reported in Table 5.

Findings. Adding spatial coordinates to ASR transcripts (A2→A3) improves accuracy while reducing format errors. Introducing structured MCoT on top of multimodal inputs (A3→A4) yields a gain of over 65 points in accuracy and eliminates format errors. These results indicate that the structured decomposition is critical for stabilizing egocentric spatial interpretation under transcription noise, even as the spatial resolution increases.

5 Conclusion

This paper examined egocentric-to-allocentric orientation reasoning in conversational navigation under ASR-mediated input and introduced the Conversational Orientation Reasoning (COR) benchmark, a controlled diagnostic setting for studying the effects of transcription noise and linguistic ambiguity in Traditional Chinese. Based on a 50×50 grid and a dataset of 10,000 instances, we decomposed orientation inference into relation extraction, coordinate-based direction identification, and final orientation inference. Results show that models trained with this structured format achieve 99.6% accuracy on clean text and maintain 95.7% accuracy under ASR-transcribed input. Beyond aggregate performance, COR supports granular error analysis, revealing that failures in spoken navigation often stem from linguistic grounding failures rather than geometric computation.

We further compared learning-based models with a deterministic rule-based solver that directly applies the benchmark’s construction logic. While the rule-based solver is effective on clean text, its accuracy degrades substantially under ASR noise. In contrast, learning-based models exhibit more

Original Traditional Chinese	English Translation
Model Input Audio: "我現在在公園4，前面是引溜電4，後面是酒吧1，左邊是念包電3，右邊是酒吧2" Coordinates: 公園4(8,5), 飲料店4(7,5), 酒吧1(9,5), 麵包店3(8,4), 酒吧2(8,6) Questions: 使用者面朝哪個方向？請逐步推理：	Model Input Audio: "I am at park 4, yin-liu-dian 4 is in front of me, bar 1 is behind me, nian-bao-dian 3 is on my left, and bar 2 is on my right" Coordinates: park 4(8,5), drink_shop 4(7,5), bar 1(9,5), bakery 3(8,4), bar 2(8,6) Questions: Which direction is the user facing? Please reason step by step.
Model Output 第一步：提取空間關係 關係1：空間關係=前面，參考地標=引溜電4 關係2：空間關係=後面，參考地標=引溜電4 關係3：空間關係=左邊，參考地標=念包電3 關係4：空間關係=右邊，參考地標=酒吧2 第二步：計算絕對方向 參考地標1=引溜電4，從公園4到引溜電4方向向量：(8,5)-(7,5)=(1,0)，方向：東方 參考地標2=飲料店4，從公園4到飲料店4方向向量：(7,5)-(8,5)=(-1,0)，方向：西方 參考地標3=麵包店3，從公園4到麵包店3方向向量：(8,4)-(8,5)=(0,-1)，方向：南方 參考地標4=酒吧2，從公園4到酒吧2方向向量：(8,6)-(8,5)=(0,1)，方向：北方 第三步：推理朝向 引溜電4在東方且在使用者前面，空間映射：面朝東方時，前面=東方 飲料店4在西方且在使用者的後面，空間映射：面朝東方時，後面=西方 麵包店3在南方且在使用者的左邊，空間映射：面朝東方時，左邊=南方 酒吧2在北方且在使用者的右邊，空間映射：面朝東方時，右邊=北方 結論：使用者面朝東方 Final answer: 東方 X	Model Output Step 1: Extract spatial relations Spatial relation 1 = front, reference landmark = yin-liu-dian 4 Spatial relation 2 = back, reference landmark = yin-liu-dian 4 Spatial relation 3 = left, reference landmark = nian-bao-dian 3 Spatial relation 4 = right, reference landmark = bar 2 Step 2: Calculate absolute directions Reference landmark 1 = yin-liu-dian 4, Direction vector from park 4 to yin-liu-dian 4: (8,5)-(7,5) = (1,0), Direction = East Reference landmark 2 = drink shop 4, Direction vector from park 4 to drink shop 4: (7,5)-(8,5) = (-1,0), Direction = West Reference landmark 3 = nian-bao-dian 3, Direction vector from park 4 to nian-bao-dian 3: (8,4)-(8,5) = (0,-1), Direction = South Reference landmark 4 = bar 2, Direction vector from park 4 to bar 2: (8,6)-(8,5) = (0,1), Direction = North Step 3: Infer user orientation Yin-liu-dian 4 is to the East and the user describes it as "in front of me". Spatial mapping rules indicate front = East when the user is facing East. Drink shop 4 is to the West and the user describes it as "behind me". Spatial mapping rules indicate back = West when the user is facing East. Nian-bao-dian 3 is to the South and the user describes it as "on my left". Spatial mapping rules indicate left = South when the user is facing East. Bar 2 is to the North and the user describes it as "on my right". Spatial mapping rules indicate right = North when the user is facing East. Therefore, the user is facing East. Final answer: East X

Figure 5: Example of an ASR misrecognition error with original output. The input contains phonetic substitutions common in Traditional Chinese ASR, such as “yin-liu-dian” for “drink shop” and “nian-bao-dian” for “bakery”.

robust error patterns and produce interpretable intermediate outputs that facilitate error attribution, highlighting COR’s value as a diagnostic tool for spoken language understanding.

6 Limitations

COR adopts a discrete 50 × 50 Manhattan grid with axis-aligned landmarks, simplifying spatial geometry and excluding diagonal or continuous relations. ASR noise is introduced via a TTS-to-ASR pipeline for controlled analysis, but does not capture the full variability of spontaneous speech, such as prosody, accent variation, or diverse disfluencies. Additionally, while COR focuses on Traditional Chinese to address an underexplored non-English setting, broader multilingual coverage remains an open challenge. Future work will extend COR to richer spatial configurations, more realistic speech conditions, evaluations across larger general-purpose language models, and interaction-centric setups such as conversational clarification under ambiguous references.

Acknowledgements

The author would like to express sincere gratitude to Cheng-Yuan Ho and Chien-Wen Yuan at National Taiwan University for their invaluable guidance and co-advising throughout this research.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. [Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments](#). *Preprint*, arXiv:1711.07280.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2020. [Touchdown: Natural language navigation and spatial reasoning in visual street environments](#). *Preprint*, arXiv:1811.12354.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. [Spatialrgpt: Grounded spatial reasoning in vision language models](#). *Preprint*, arXiv:2406.01584.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. [Talk the walk: Navigating new york city through grounded dialogue](#). *Preprint*, arXiv:1807.03367.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. [Moshi: a speech-text foundation model for real-time dialogue](#). *Preprint*, arXiv:2410.00037.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. [Speaker-follower models for vision-and-language navigation](#). *Preprint*, arXiv:1806.02724.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiwu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. 2025. [Vita-1.5: Towards gpt-4o level real-time vision and speech interaction](#). *Preprint*, arXiv:2501.01957.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraismwami, and Dinesh Manocha. 2024. [Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities](#). *Preprint*, arXiv:2406.11768.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu

- Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, and 66 others. 2022. [Ego4d: Around the world in 3,000 hours of egocentric video](#). *Preprint*, arXiv:2110.07058.
- Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenye Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. [The impact of reasoning step length on large language models](#). *Preprint*, arXiv:2401.04925.
- Yuka Kaniwa, Masaki Kuribayashi, Seita Kayukawa, Daisuke Sato, Hironobu Takagi, Chieko Asakawa, and Shigeo Morishima. 2024. [Chitchatguide: Conversational interaction using large language models for assisting people with visual impairments to explore a shopping mall](#). *Proc. ACM Hum.-Comput. Interact.*, 8(MHCI).
- Elad Levi and Ilan Kadar. 2025. [Intelligent: A multi-agent framework for evaluating conversational ai systems](#). *Preprint*, arXiv:2501.11067.
- Ming-Yi Lin, Ou-Wen Lee, and Chih-Ying Lu. 2024. [Embodied ai with large language models: A survey and new hri framework](#). In *2024 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 978–983.
- Shuijing Liu, Aamir Hasan, Kaiwen Hong, Runxuan Wang, Peixin Chang, Zachary Mizrahi, Justin Lin, D. Livingston McPherson, Wendy A. Rogers, and Katherine Driggs-Campbell. 2024. [Dragon: A dialogue-based robot for assistive navigation with visual language grounding](#). *IEEE Robotics and Automation Letters*, 9(4):3712–3719.
- Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, Helong Huang, Guangjian Tian, Weichao Qiu, Xingyue Quan, Jianye Hao, and Yuzheng Zhuang. 2025. [Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning](#). *Preprint*, arXiv:2501.10074.
- 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](#). *Preprint*, arXiv:2501.07246.
- Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. [Embodiedgpt: Vision-language pre-training via embodied chain of thought](#). *Preprint*, arXiv:2305.15021.
- Ritawari Pareek, Divyansh Chauhan, Sonal Tuteja, and Kapil Madan. 2024. [Enhancing campus navigation: A conversational ai agent for location assistance](#). IC3-2024, New York, NY, USA. Association for Computing Machinery.
- Ben Prystawski, Michael Y. Li, and Noah D. Goodman. 2023. [Why think step by step? reasoning emerges from the locality of experience](#). *Preprint*, arXiv:2304.03843.
- Zhixuan Shen, Haonan Luo, Kexun Chen, Fengmao Lv, and Tianrui Li. 2025. [Enhancing multi-robot semantic navigation through multimodal chain-of-thought score collaboration](#). *Preprint*, arXiv:2412.18292.
- Smitha Sheshadri and Kotaro Hara. 2024. [Conversational localization: Indoor human localization through intelligent conversation](#). *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(4).
- Jiatong Shi, Hye jin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, Dareen Safar Alharthi, Yichen Huang, Koichi Saito, Jionghao Han, Yiwen Zhao, Chris Donahue, and Shinji Watanabe. 2025. [Versa: A versatile evaluation toolkit for speech, audio, and music](#). *Preprint*, arXiv:2412.17667.
- Qi Sun, Pengfei Hong, Tej Deep Pala, Vernon Toh, U-Xuan Tan, Deepanway Ghosal, and Soujanya Poria. 2024. [Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning](#). *Preprint*, arXiv:2412.11974.
- Rahul Sundar, Shreyash Gadgil, Tankala Satya Sai, Sathi Sai Krishna Reddy, Gautam B, Ishita Mittal, Jyotsna Sree Guduguntla, and Shanmukesh Pujala. 2024. [Innoguiidegpt: Integrating conversational interface and command interpretation for navigation robots](#). In *Proceedings of the Third International Conference on AI-ML Systems, AIMLSystems '23*, New York, NY, USA. Association for Computing Machinery.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. [Salmonn: Towards generic hearing abilities for large language models](#). *Preprint*, arXiv:2310.13289.
- Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. [Large language models are in-context semantic reasoners rather than symbolic reasoners](#). *Preprint*, arXiv:2305.14825.
- Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. [Vision-and-dialog navigation](#). *Preprint*, arXiv:1907.04957.
- Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. [Audiobench: A universal benchmark for audio large language models](#). *Preprint*, arXiv:2406.16020.
- Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. [Towards understanding chain-of-thought prompting: An empirical study of what matters](#). In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), Toronto, Canada. Association for Computational Linguistics.

Zhifei Xie and Changqiao Wu. 2024. [Mini-omni: Language models can hear, talk while thinking in streaming](#). *Preprint*, arXiv:2408.16725.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. [Air-bench: Benchmarking large audio-language models via generative comprehension](#). *Preprint*, arXiv:2402.07729.

Chien yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung yi Lee. 2024. [Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech](#). *Preprint*, arXiv:2309.09510.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. [Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities](#). *Preprint*, arXiv:2305.11000.

A.1 Error taxonomy and Examples

Note on reasoning traces. The intermediate reasoning steps shown in this section are generated by the model and are not externally verified or enforced at inference time. We include these reasoning traces to facilitate qualitative analysis and error attribution, rather than to claim strict faithfulness between the intermediate steps and the final prediction. In particular, a correct final orientation does not require all intermediate steps to be correct, and conversely, errors may arise even when earlier steps appear plausible.

A.1.1 Direction Understanding Error

Error Analysis. The model correctly extracted relations and computed absolute directions but incorrectly applied spatial mapping rules in the final step, concluding the user faces South.

Input.

Audio: I am at Taipei Main Station Exit S2, restaurant 5 is in front of me, and Taipei Main Station Exit S3 is on my left | Coordinates: Taipei_Main_Station_Exit_S2(4,4), restaurant_5(4,5), Taipei_Main_Station_Exit_S3(3,4)

Model Output.

1. Step 1: Extract spatial relations

Spatial relation 1 = front, landmark = restaurant 5

Spatial relation 2 = left, landmark = Taipei Main Station Exit S3

2. Step 2: Calculate absolute directions

Reference landmark 1 = restaurant 5, Direction vector from Taipei Main Station Exit S2 to restaurant 5: $(4,5)-(4,4) = (0,1)$, Direction = North

Reference landmark 2 = Taipei Main Station Exit S3, Direction vector from Taipei Main Station Exit S2 to Taipei Main Station Exit S3: $(3,4)-(4,4) = (-1,0)$, Direction = West

3. Step 3: Infer user orientation

Restaurant 5 is to the North and the user describes it as “on my left”. Spatial mapping rules indicate left = North when the user is facing South.

Taipei Main Station Exit S3 is to the West and the user describes it as “on my left”. Spatial mapping rules indicate left = West when the user is facing South.

Final Answer. The user is facing South.

A.1.2 Relation Extraction Error

Error Analysis. The model incorrectly extracted spatial relations from the input, misidentifying which landmark corresponds to which direction.

Input.

Audio: I am at bus stop 2, Taipei Main Station Exit S3 is in front of me, Taipei Main Station Exit K7 is on my left, and sports store 1 is on my right | Coordinates: bus_stop_2(3,3), Taipei_Main_Station_Exit_S3(3,4), Taipei_Main_Station_Exit_K7(2,3), sports_store_1(4,3)

Model Output.

1. Step 1: Extract spatial relations

Spatial relation 1 = left, landmark = Taipei Main Station Exit S3

Spatial relation 2 = right, landmark = sports store 1

2. Step 2: Calculate absolute directions

Reference landmark 1 = Taipei Main Station Exit S3, Direction vector from bus stop 2 to Taipei Main Station Exit S3: $(3,4)-(3,3) = (0,1)$, Direction = North

Reference landmark 2 = sports store 1, Direction vector from bus stop 2 to sports store 1: $(4,3)-(3,3) = (1,0)$, Direction = East

3. Step 3: Infer user orientation

Taipei Main Station Exit S3 is to the North and the user describes it as “on my left”. Spatial mapping rules indicate left = North when the user is facing South.

Sports store 1 is to the East and the user describes it as “on my right”. Spatial mapping rules indicate right = East when the user is facing South.

Final Answer. The user is facing South.

A.1.3 ASR Misrecognition Error

Error Analysis. ASR misrecognition caused the model to incorrectly map multiple spatial relations to the same landmark (e.g., confusing “drink shop” with “yin-liu-dian”), leading to inconsistent internal reasoning even when the absolute directions are correctly computed.

Input.

Audio: I am at park 4, yin-liu-dian 4 is in front of me, 981 is behind me, nian-bao-dian 3 is on my left, and 982 is on my right | Coordinates: park_4(8,5), drink_shop_4(7,5), bar_1(9,5), bakery_3(8,4), bar_2(8,6)

Model Output.

1. Step 1: Extract spatial relations

Spatial relation 1 = front, reference landmark = yin-liu-dian 4

Spatial relation 2 = back, reference landmark = yin-liu-dian 4

Spatial relation 3 = left, reference landmark = nian-bao-dian 3

Spatial relation 4 = right, reference landmark = bar 2

2. Step 2: Calculate absolute directions

Reference landmark 1 = yin-liu-dian 4, Direction vector from park 4 to yin-liu-dian 4: $(8,5)-(7,5) = (1,0)$, Direction = East

Reference landmark 2 = drink shop 4, Direction vector from park 4 to drink shop 4: $(7,5)-(8,5) = (-1,0)$, Direction = West

Reference landmark 3 = nian-bao-dian 3, Direction vector from park 4 to nian-bao-dian 3: $(8,4)-(8,5) = (0,-1)$, Direction = South

Reference landmark 4 = bar 2, Direction vector from park 4 to bar 2: $(8,6)-(8,5) = (0,1)$, Direction = North

3. Step 3: Infer user orientation

Yin-liu-dian 4 is to the East and the user describes it as “in front of me”. Spatial mapping rules indicate front = East when the user is facing East.

Drink shop 4 is to the West and the user describes it as “behind me”. Spatial mapping rules indicate back = West when the user is facing East.

Nian-bao-dian 3 is to the South and the user describes it as “on my left”. Spatial mapping rules indicate left = South when the user is facing East.

Bar 2 is to the North and the user describes it as “on my right”. Spatial mapping rules indicate right = North when the user is facing East.

Final Answer. East

A.2 Implementation Details for the Experiments

Model Architecture. We use Taiwan-LLM-13B-v2.0-Chat as the base model with LoRA fine-tuning under 4-bit quantization (rank $r = 16$, $\alpha = 32$, dropout rate 0.05).

Training Parameters. All experiments run for 5 epochs with batch size 1 and gradient accumulation steps of 32. The learning rate is set to 5×10^{-5} , with a maximum sequence length of 768 tokens. Training is conducted in bfloat16 mixed precision on NVIDIA A100 GPUs.

Data Format. Inputs follow the LLaMA chat template, combining ASR transcripts and spatial coordinates. The model is trained to output structured three-step reasoning traces aligned with the MCoT design.

ASR Preprocessing. Speech inputs were transcribed with Whisper-base (zh) using greedy decoding (temperature 0, no beam search) on 16 kHz mono audio. We report character error rate (CER) on the TTS→ASR loop as a proxy of noise, rather than Whisper’s internal probabilities. TTS was used solely to synthesize controlled variants for evaluation experiments.

Table 6: Training and ASR preprocessing details.

Model	Taiwan-LLM-13B-v2.0-Chat
Fine-tuning	LoRA (4-bit)
Rank / α / Dropout	16 / 32 / 0.05
Optimizer	AdamW (5×10^{-5})
Batch / Accum.	1 / 32
Precision	bfloat16
Hardware	NVIDIA A100
ASR	Whisper-base (zh), greedy
Audio	16 kHz mono
Metric	CER (TTS→ASR)

A.3 Full Prompts

A.3.1 Training Prompts

Step 0: Spatial Relation Extraction

Task. Extract all spatial relations and reference landmarks from the following sentence.

Input. “I am at Gongguan MRT Exit 1, Restaurant 4 is in front of me, Gongguan MRT Exit 2 is behind me, Bus Stop 1 is on my left, and Bike Rental Station 3 is on my right.”

Output.

Spatial relation 1 = front, Reference landmark 1 = Restaurant 4
 Spatial relation 2 = behind, Reference landmark 2 = Gongguan MRT Exit 2
 Spatial relation 3 = left, Reference landmark 3 = Bus Stop 1
 Spatial relation 4 = right, Reference landmark 4 = Bike Rental Station 3

Step 1: Vector Direction Learning

Input.

From = (0,0), To = (1,0)

Output.

Direction vector: (1,0) - (0,0) = (1,0),
 Direction = East

Step 2: Orientation Inference

Task. Infer user orientation based on the landmark absolute direction and spatial relation.

Input.

Landmark absolute direction = East
 Spatial relation = behind

Output.

Analysis: The landmark is to the East, and the user describes it as “behind me.” Spatial mapping rules indicate behind = East when the user is facing West. Therefore, the user is facing West.

Step 3: Complete Chain-of-Thought Reasoning

Input.

Audio: I am at Academic Building A, Student Activity Center 1 is in front of me, Academic Building B is behind me, Parking Lot 2 is on my left, and Small Plaza 2 is on my right | Coordinates: Academic Building A(7,1), Student Activity Center 1(6,1), Academic Building B(8,1), Parking Lot 2(7,0), Small Plaza 2(7,2)

Output.

1. Step 1: Extract spatial relations

Spatial relation 1 = front, Reference landmark 1 = Student Activity Center 1
 Spatial relation 2 = behind, Reference landmark 2 = Academic Building B
 Spatial relation 3 = left, Reference landmark 3 = Parking Lot 2
 Spatial relation 4 = right, Reference landmark 4 = Small Plaza 2

2. Step 2: Calculate absolute directions

Reference landmark 1 = Student Activity Center 1, Direction vector: (6,1) - (7,1) = (-1,0), Direction = West
 Reference landmark 2 = Academic Building B, Direction vector: (8,1) - (7,1) = (1,0), Direction = East
 Reference landmark 3 = Parking Lot 2, Direction vector: (7,0) - (7,1) = (0,-1), Direction = South
 Reference landmark 4 = Small Plaza 2, Direction vector: (7,2) - (7,1) = (0,1), Direction = North

3. Step 3: Infer user orientation

Student Activity Center 1 is to the West, and the user describes it as “in front of me.” Spatial mapping rules indicate front = West when the user is facing West.
 Academic Building B is to the East, and the user describes it as “behind me.” Spatial mapping rules indicate behind = East when the user is facing West.
 Parking Lot 2 is to the South, and the user describes it as “on my left.” Spatial mapping rules indicate left = South when the user is facing West.
 Small Plaza 2 is to the North, and the user describes it as “on my right.” Spatial mapping rules indicate right = North when the user is facing West.
 Therefore, the user is facing West.

A.3.2 Baseline Evaluation Prompts

Training data format. All training uses the LLaMA chat template with Traditional Chinese inputs and outputs, as Taiwan-LLM-13B-v2.0-Chat is optimized for Traditional Chinese. All training and evaluation were conducted in Traditional Chinese. The prompts below are English translations for clarity and reproducibility.

B1: Zero-shot baseline

Question: Audio: I am at Gongguan MRT Exit 3, and Dormitory 2 is on my right | Coordinates: Gongguan_MRT_Exit_3(5,8), Dormitory_2(6,8)
Which direction is the user facing?
Please answer North, South, East, or West.
Answer:

B2: Few-shot prompting (no CoT)

Instruction: Based on the audio description and coordinate information, determine which direction the user is facing.

Example: Audio: I am at the gym, and the pharmacy is in front of me | Coordinates: gym(4,6), pharmacy(4,7)
Answer: North

Example: Audio: I am at the park, and the water park is behind me | Coordinates: park(0,0), water_park(0,1)
Answer: South

Example: Audio: I am at the foundation, and the high school is on my right | Coordinates: foundation(0,7), high_school(0,6)
Answer: East

Example: Audio: I am at the cooperative store, and the theater is on my left | Coordinates: cooperative_store(8,8), theater(8,7)
Answer: West

Question: {user_input}
Answer:

B3: Few-shot prompting with CoT

Instruction: Use three-step reasoning to determine the user's facing direction given the audio description and coordinates.

Example 1

Input: Audio: I am at the gym, and the pharmacy is in front of me | Coordinates: gym(4,6), pharmacy(4,7)
Output:
Step 1: Extract spatial relations

Spatial relation = front
Reference landmark = pharmacy
Step 2: Calculate absolute directions
Direction vector from gym to pharmacy:
 $(4,7) - (4,6) = (0,1)$
Direction = North
Step 3: Infer user orientation
The pharmacy is to the North, and the user describes it as "in front of me."
Spatial mapping rules indicate front = North when the user is facing North.
Therefore, the user is facing North.

Example 2

Input: Audio: I am at the park, and the water park is behind me | Coordinates: park(0,0), water_park(0,1)
Output:
Step 1: Extract spatial relations
Spatial relation = behind
Reference landmark = water park
Step 2: Calculate absolute directions
Direction vector from park to water park:
 $(0,1) - (0,0) = (0,1)$
Direction = North
Step 3: Infer user orientation
The water park is to the North, and the user describes it as "behind me."
Spatial mapping rules indicate behind = North when the user is facing South.
Therefore, the user is facing South.

Now use the same three-step reasoning:

Input: {user_input}
Output:

B4: Fine-tuned direct classification

USER: {user_input}
ASSISTANT:

A.4 Controlled Linguistic Variation Details

A.4.1 Linguistic Variation (R1)

To analyze model behavior under natural linguistic variations, we constructed test sets in Traditional Chinese with diverse expression patterns, while ensuring spatial semantics and orientation outputs.

Variation types

- **Word order variations:** sentence inversion, argument permutation, and syntactic paraphrasing.
- **Synonym substitutions:** spatial term substitution, position verb substitution, and landmark term substitution.

A.4.2 Referential Ambiguity (R3)

We test the model's ability to handle ambiguous or underspecified references commonly encountered in natural conversational navigation. In addition to the variation types listed below, R3 includes

a subset of cases with vague directional expressions (e.g., approximate or imprecise orientation descriptions). For this subset, ground-truth orientations are annotated as sets and evaluated using set-based accuracy, while the remaining R3 cases retain single-label evaluation.

Variation types

- **Referential ambiguity:** generic references (“this building”), and demonstrative pronouns (“that place”).
- **Incomplete utterances:** disfluency (“I am at... um...”), uncertainty markers (“should be”), and hesitation patterns.
- **Semantic underspecification:** vague location terms (“some building”) and imprecise references (“over there”).

Example of R3 test cases

Original: Audio: I am at security office, and Dormitory 6 is behind me | Coordinates: security_office(7,3), dormitory_6(7,4)

Referential ambiguity: Audio: I am at this building, that dormitory is behind me | Coordinates: security_office(7,3), dormitory_6(7,4)

Incomplete utterance: Audio: I am at... um... security office, Dormitory 6 should be behind | Coordinates: security_office(3,5), dormitory_6(3,4)

Semantic underspecification: Audio: I am at some place, that building over there is behind me | Coordinates: security_office(3,5), dormitory_6(3,4)

A.5 ASR Error Severity Statistics

Table 7 reports the distribution of ASR error severity in both evaluation sets.

A.6 Rule-based Solver Analysis

To contextualize performance on COR, we implement a deterministic rule-based solver as a reference system. This solver provides a point of comparison under clean text and helps illustrate how symbolic approaches behave when language input is affected by ASR noise.

Algorithm sketch. Given an ASR transcript and landmark coordinates, the solver proceeds:

Table 7: ASR error severity distribution with exact counts.

Set	Severity	Count	%
Main (Gongguan)	perfect	218	14.5
	minor	478	31.9
	moderate	556	37.1
	major	220	14.7
	severe	28	1.9
Cross-domain (Taipei)	perfect	318	26.5
	minor	460	38.3
	moderate	347	28.9
	major	71	5.9
	severe	4	0.4

1. Extract egocentric relations (front, back, left, right) and landmark names using pattern-based and exact string matching.
2. For each extracted landmark, compute the relative coordinate vector and determine its absolute direction using a dominant-axis heuristic.
3. Infer the user orientation by applying fixed relative-to-absolute mapping rules (Table 1).

If any required landmark cannot be matched exactly, the instance is treated as a failed extraction.

Observed behavior under ASR noise. Although the solver attains high accuracy on clean text (99.86%), its performance degrades substantially on ASR-transcribed input. The primary source of error is not geometric computation, but failures in matching linguistic references under transcription noise. Common issues include: (i) homophones and character substitutions in landmark names; (ii) truncation or merging of multi-character landmark tokens; and (iii) inconsistent transcription of alphanumeric identifiers (e.g., numeric exits or store indices). These mismatches often prevent landmark resolution, leading to errors in subsequent orientation inference.

Discussion. The behavior of this solver highlights a central challenge addressed by COR. In speech-mediated settings, errors frequently originate from noisy language input rather than from spatial reasoning itself. As a result, approaches that rely solely on exact string matching are brittle under ASR transcription, motivating models that can integrate linguistic context with spatial reasoning when interpreting spoken navigation instructions.

Ineffectiveness of Fuzzy Matching. To address potential improvements via fuzzy matching, we

further evaluated the solver using Levenshtein distance and pinyin-based normalization for landmark extraction. However, the performance gain was marginal (reaching approximately 40.2%). This is primarily because ASR errors in Traditional Chinese often involve semantic shifts that fuzzy matching cannot bridge: (i) homophones that are phonetically identical but orthographically distinct, making character-level edit distance ineffective, and (ii) fragmented syntax where spatial relations are separated from landmarks by disfluencies. These findings underscore that the challenge in COR goes beyond simple string corruption, requiring the model to maintain referential integrity through deep linguistic understanding.

A.7 Stage-wise Supervision and Curriculum Design

We employ stage-wise supervision to stabilize the learning of decomposed reasoning components, rather than to encode or enforce deterministic mapping rules. This approach acts as a structural prior that encourages the model to ground its final orientation in the extracted linguistic and geometric cues, effectively reducing the noise-induced hallucination of orientations. In preliminary experiments, training all reasoning steps jointly from scratch often resulted in unstable behavior, such as frequent schema violations or premature guessing of final orientations without forming consistent intermediate representations. Stage-wise supervision affects only the sequence in which subtasks are introduced, while keeping the data distribution and supervision signals unchanged. Earlier stages focus on learning localized components (e.g., relation extraction and coordinate-based direction identification), which helps reduce interference between linguistic processing and geometric reasoning when later stages are trained end to end.

Crucially, this curriculum does not introduce additional rules or constraints beyond those used to define the evaluation labels. The mapping rules in Table 1 specify the target orientation but are neither executed nor enforced during inference. The model must still interpret noisy language input, resolve landmark references under ASR transcription errors, and reconcile multiple spatial cues. Accordingly, stage-wise supervision should be understood as a means of stabilizing intermediate representations, not as a mechanism for memorizing rule templates. In practice, removing early-stage super-

vision primarily increases schema violations and inconsistency in intermediate steps, while having limited effect on final orientation accuracy.