

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

Anonymous ACL submission

Abstract

Resolving egocentric spatial expressions in conversational navigation requires mapping speaker-centric references such as “on my right” to allocentric orientations (e.g., north, east, south, west). 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 pairs ASR-transcribed utterances with structured landmark coordinates derived from real-world environments, while controlling spatial geometry to isolate the effects of spoken language noise. This enables 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 high performance on clean text and retain high accuracy under ASR-transcribed input, while unstructured baselines degrade substantially. COR supports analysis of failure modes in spoken spatial reference resolution, providing insights into how language grounding degrades under noisy conversational input.

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, and is fundamental to conversational navigation and spoken lan-

guage understanding. 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.

085	To address this gap, we introduce Conversational	failure modes that are obscured by end-to-end	133
086	Orientation Reasoning (COR), a diagnostic bench-	navigation metrics.	134
087	mark for egocentric spatial reference resolution		
088	in Traditional Chinese conversational navigation.	2 Related Work	135
089	COR pairs ASR-transcribed egocentric utterances	2.1 Egocentric Spatial Language and	136
090	with structured landmark coordinates derived from	Conversational Navigation	137
091	real-world urban layouts in Taiwan and projected	Egocentric spatial reference resolution concerns	138
092	onto a grid representation. By intentionally con-	interpreting speaker-centric expressions into allo-	139
093	trolling spatial geometry, COR enables targeted	centric orientations with respect to a world-centric	140
094	analysis of uncertainty introduced by spoken lan-	reference frame. This problem is embedded in nat-	141
095	guage, including transcription errors, linguistic	ural language navigation tasks, where agents fol-	142
096	variation, and referential ambiguity, rather than	low instructions or dialogues grounded in spatial	143
097	approximating full navigation complexity. Using	environments. Early benchmarks such as Room-	144
098	COR, we analyze an interpretable decomposition of	to-Room (R2R) (Anderson et al., 2018) and Co-	145
099	orientation inference into spatial relation extrac-	operative Vision-and-Dialog Navigation (CVDN)	146
100	tion, coordinate-based direction identification, and	(Thomason et al., 2019) focus on visually ground-	147
101	final orientation inference, enabling fine-grained	instruction following, while subsequent work ex-	148
102	examination of where and how errors arise. We	tends navigation to urban and conversational	149
103	evaluate performance under clean text, ASR-trans-	settings, including Talk the Walk (de Vries et al.,	150
104	cribed inputs, and multiple evaluation conditions,	2018) and Touchdown (Chen et al., 2020).	151
105	including unseen environments and ambiguous	In these benchmarks, egocentric spatial language	152
106	references. Our study is guided by three research	is typically evaluated only through end-to-end	153
107	questions:	navigation success, making it difficult to isolate	154
108		errors arising from linguistic misinterpretation.	155
109		Other studies incorporate spatial representations,	156
110		such as panoramic orientation in Speaker-Follower	157
111		models (Fried et al., 2018), 3D scene graphs in	158
112		Spatial-RGPT (Cheng et al., 2024), and large-	159
113		scale egocentric visual data in Ego4D (Grauman	160
114		et al., 2022). Despite these advances, existing	161
115		approaches generally assume known agent	162
116		orientations or operate over high-level action	163
117		spaces, rather than evaluating the intermediate	164
118		task of resolving egocentric spatial refer-	165
119		ences into allocentric orientations. As a result,	166
120		failures in egocentric spatial interpreta-	167
121		tion caused by spoken language noise are often	168
122		entangled with downstream perception or action	169
123		errors, and remain insufficiently characterized	170
124		as a standalone language understanding prob-	171
125		lem, particularly in non-English conversational	
126		settings.	
127		2.2 Multimodal Chain-of-Thought Reasoning	172
128		Chain-of-thought (CoT) prompting encourages	173
129		models to generate intermediate reasoning	174
130		steps and has been shown to improve perfor-	175
131		mance across a wide range of reasoning	176
132		tasks (Jin et al., 2024; Wang et al., 2023;	177
		Prystawski et al., 2023; Tang et al., 2023).	178
		Extensions to multimodal chain-	179
		of-thought (MCoT) enable reasoning over	180
		language, vision, and spatial inputs, allow-	181
		ing embodied agents to decompose complex	
		tasks into inter-	

pretable 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 primarily employs MCoT to improve planning, action execution, or multimodal perception. In contrast, our work uses structured reasoning as an analytic scaffold to examine how egocentric spatial relations expressed in language are transformed into allocentric orientation labels. This perspective enables fine-grained 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

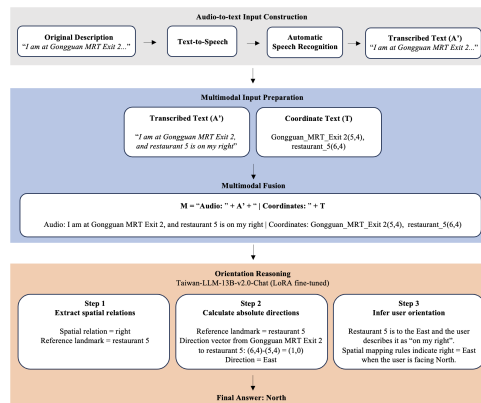


Figure 1: Pipeline of our setup: (1) speech synthesis and transcription, (2) multimodal input preparation and fusion, and (3) orientation inference.

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. The real-world area is projected into a discrete 10×10 Manhattan grid \mathcal{G} . The user position is $u = (x_u, y_u) \in \mathcal{G}$, and landmarks form a set $\mathcal{L} = \{\ell_i\}$ with coordinates $p(\ell_i) \in \mathcal{G}$. This representation intentionally removes geometric ambiguity (e.g., diagonal relations). By fixing spatial geometry across instances, COR isolates uncertainty arising from language interpretation 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 (label protocol). 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 only 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 orien-

Figure 2: Task environment. Gongguan MRT area projected into a 10×10 grid map.

tation 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.

Stage-wise fine-tuning. Let $f_{\theta(i)}$ denote the model at stage i . Stage-wise supervision is used to

337 stabilize learning and to expose intermediate struc- 377
338 tures, rather than to encode or enforce the mapping 378
339 rules themselves. The curriculum controls the order 379
340 in which subtasks are introduced. 380

341 **(S1) Relation extraction.** The model extracts 381
342 the egocentric relation and reference landmark 382
343 from clean descriptions: 383

$$344 \quad r_1 = f_{\theta(0)}(A) \rightarrow (q, \ell_r).$$

345 **(S2) Coordinate-based direction identifica- 384
346 tion.** The model infers the absolute direction from 385
347 user and landmark coordinates: 386

$$348 \quad r_2 = f_{\theta(1)}(u, p(\ell_r)) \rightarrow d_{\text{abs}} \in \{\text{N,E,S,W}\}.$$

349 **(S3) Orientation inference.** The final orien- 387
350 tation is inferred from the absolute direction and 388
351 egocentric relation: 389

$$352 \quad r_3 = f_{\theta(2)}(d_{\text{abs}}, q) \rightarrow \hat{D},$$

353 where \hat{D} denotes a ranked list of allocentric orien- 390
354 tations. For clear directional expressions, \hat{D} contains 391
355 a single dominant orientation. For vague direc- 392
356 tional expressions, evaluation considers whether \hat{D} 393
357 intersects with the ground-truth orientation set D^* . 394

358 **(S4) End-to-end integration.** ASR transcripts 395
359 and coordinates are serialized into a single input: 396

$$360 \quad M = \text{"Audio: " + } A' + \text{" | Coordinates: " + } T,$$

361 where T serializes $(u, \{\ell_i, p(\ell_i)\})$ into tokens. The 397
362 model then generates a full reasoning trace: 398

$$363 \quad S = f_{\theta(3)}(M),$$

364 which includes intermediate steps (r_1, r_2, r_3) and 399
365 the final prediction D^* . At inference time, the 400
366 model is not constrained to follow the intermedi- 401
367 ate steps. The structured output format is used to 402
368 facilitate analysis and error attribution. 403

369 **Objective.** Training supervision is applied to 404
370 both intermediate steps and final predictions: 405

$$371 \quad \mathcal{L} = - \sum_{t=1}^{|S|} \log P(s_t | s_{<t}, M).$$

372 4 Experimental Results 410

373 4.1 Dataset 411

374 We evaluate on the Conversational Orientation 412
375 Reasoning (COR) benchmark, which targets 413
376 egocentric-to-allocentric orientation inference in 414

377 conversational navigation. COR contains 4,600 378
378 instances. Each instance consists of (i) an egocen- 379
379 tric utterance in Traditional Chinese, (ii) structured 380
380 landmark coordinates on a discrete grid, and and 381
381 (iii) a cardinal orientation label or set of labels. 382
382 A step-by-step reasoning trace is additionally pro- 383
383 vided to supervise the structured output format. 384

384 To simulate speech-driven conditions while 385
385 maintaining experimental control, we synthesize 386
386 speech from clean utterances and transcribe it using 387
387 an ASR system to obtain noisy transcripts. This 388
388 ASR-mediated setup introduces transcription noise 389
389 without altering the underlying spatial semantics. 390
390 All instances are generated from the grid-based la- 391
391 beling protocol (Table 1) and verified by human 392
392 annotators for correctness. 393

393 **Data splits.** COR is divided into 3,216 training, 394
394 688 validation, and 696 test examples. Models are 395
395 fine-tuned using clean text only, while ASR tran- 396
396 scripts are introduced at evaluation time. To ensure 397
397 coverage across spatial relations, the training set is 398
398 balanced over combinations of egocentric relations: 399
399 each single relation (front/back/left/right) contains 400
400 320 utterances, while each double-, triple-, and 401
401 quadruple-relation combination contains 280 utter- 402
402 ances. A subset of 400 examples introduces con- 403
403 trolled linguistic variations (e.g., synonym substitu- 404
404 tions and word-order changes), distributed across 405
405 splits. The test sets reflect multilingual characteris- 406
406 tics common in Taiwan: English landmark names 407
407 appear in 4.7% of the main test samples and 46.5% 408
408 of the cross-domain samples, with occasional sim- 409
409 plified Chinese variants introduced by ASR. 410

410 **Evaluation subsets.** Beyond the main test set 411
411 (696 examples from the Gongguan area), we con- 412
412 struct two additional evaluation sets for RQ3: (1) 413
413 Cross-domain (540 examples), projected into a 414
414 10×10 grid from the unseen Taipei Station area 415
415 (Figure 4); and (2) Referential ambiguity (200 ex- 416
416 amples), which include ambiguous references, dis- 417
417 fluent or incomplete utterances, and semantically 418
418 underspecified mentions. 419

419 **ASR error profile.** Figure 3 shows the ASR er- 420
420 ror severity distribution across the main and cross- 421
421 domain test sets. Exact counts are provided in 422
422 Appendix 5 (Table 7). 423

423 **Example.** “*I am at Gongguan MRT Exit 2, and 424
424 restaurant 5 is on my right*” \rightarrow Label: North. 425

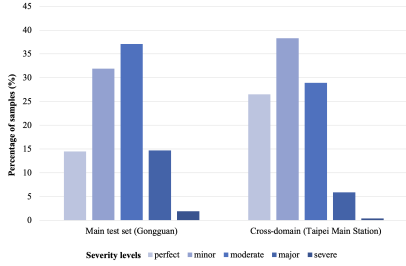


Figure 3: ASR error severity distribution in the two evaluation sets.

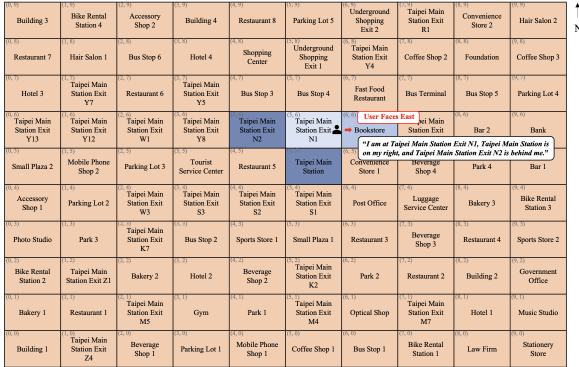


Figure 4: Cross-domain evaluation environment. Taipei Station area projected into a 10×10 grid.

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 generate reasoning traces, we additionally report *reasoning quality*, measured as the match rate of intermediate steps (range 0–1, higher is better), and *format error rate*, defined as the proportion of outputs violating the expected schema.

Rule-based baseline. We implement a deterministic baseline that (i) extracts relation–landmark pairs using pattern matching over the transcript and (ii) applies the fixed mapping rules in Table 1. On clean text, this baseline achieves 99.86% accuracy, indicating that orientation labels can be recovered when linguistic input is error-free. Under ASR-transcribed input, however, accuracy drops to 39.51%, with 40.80% of cases failing at the extraction stage, primarily due to landmark string corruption such as homophones and character substitutions. This contrast illustrates the sensitivity of string-based extraction to transcription noise and motivates the use of learning-based models for handling noisy spatial language. Table 2 summarizes results across baselines and our method.

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	100.0	0.0
		ASR	98.1	0.0

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

Quantitative Results. The curriculum-trained model achieves 100.0% accuracy on clean inputs and 98.1% accuracy on ASR-transcribed inputs. As shown in Table 2, its performance exceeds that of non-CoT baselines under both settings. In contrast, the rule-based baseline performs reliably on clean text but fails in a large fraction of ASR-transcribed cases due to its sensitivity to landmark string corruption. A detailed analysis of the rule-based solver is provided in Appendix A.6.

The experimental 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, including fine-tuned variants, perform poorly under ASR noise due to frequent output format violations and confusion between relational and orientation supervision, further motivating our decomposition-based approach that separates relation extraction from constrained orientation inference.

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 3 summarizes their definitions, and we provide

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

Table 3: Error taxonomy used in our analysis.

Direction Understanding Error	
Model Input	Audio: “我現在在台北車站S2出口，前面是餐廳5，左邊是台北車站S3出口” Coordinates: 台北車站S2出口(4,4), 餐廳5(4,5), 台北車站S3出口(3,4) 問題: 使用者面朝哪個方向？請逐步推理：
Model Output	<p>第一步：提取空間關係</p> <p>關係1：空間關係=前面，參考地標=餐廳5</p> <p>關係2：空間關係=左邊，參考地標=台北車站S3出口</p> <p>第二步：計算絕對方向</p> <p>參考地標1=餐廳5，從台北車站S2出口到餐廳5方向向量：$(4,5)-(4,4)=(0,1)$，方向：北方</p> <p>參考地標2=台北車站S3出口，從台北車站S2出口到台北車站S3出口方向向量：$(3,4)-(4,4)=(-1,0)$，方向：西方</p> <p>第三步：推理朝向</p> <p>餐廳5在北方且在使用者的左邊，空間映射：面朝南方時，左邊=北方</p> <p>台北車站S3出口在西方且在使用者的左邊，空間映射：面朝南方時，左邊=西方</p> <p>結論：使用者面朝南方</p> <p>Final answer: 南方 X</p>

Figure 5: Direction understanding error with original Traditional Chinese model output (ASR-transcribed).

one representative example for each type below.

Direction Understanding. This is the most frequent mode (9 cases), where the model correctly extracts relations and computes directions but fails the final mapping logic. As shown in Figure 5, the model correctly identifies that the landmark is North of the user (Step 2) but incorrectly applies the rotation rule in Step 3, concluding that facing South would place North on the user’s “left”. This suggests that egocentric-to-alloentric transformation remains a challenge for LLMs even when geometric inputs are accurate.

Relation Extraction. These errors (2 cases) occur when the model fails to extract all mentioned landmarks or misattributes them to the wrong spatial relations. As illustrated in Figure 6, the model omits a key landmark from the noisy transcript and incorrectly pairs the remaining ones with the wrong egocentric directions (e.g., assigning “left” to a landmark that was described as “in front”). This highlights the challenge of maintaining referential integrity when the model must parse long and disfluent ASR outputs.

ASR Misrecognition. Phonetic corruption can lead to semantic inconsistencies that propagate through the reasoning trace (3 cases). In Figure 7,

Relation Extraction Error	
Model Input	Audio: “我現在在公車站2，前面是台北車站S3出口，左邊是台北車站推騎出口，右邊是運動用品店1” Coordinates: 公車站2(3,3), 台北車站S3出口(3,4), 台北車站K7出口(2,3), 運動用品店1(4,3) 問題: 使用者面朝哪個方向？請逐步推理：
Model Output	<p>第一步：提取空間關係</p> <p>關係1：空間關係=左邊，參考地標=台北車站S3出口</p> <p>關係2：空間關係=右邊，參考地標=運動用品店1</p> <p>第二步：計算絕對方向</p> <p>參考地標1=台北車站S3出口，從公車站2到台北車站S3出口方向向量：$(3,4)-(3,3)=(0,1)$，方向：北方</p> <p>參考地標2=運動用品店1，從公車站2到運動用品店1方向向量：$(4,3)-(3,3)=(1,0)$，方向：東方</p> <p>第三步：推理朝向</p> <p>台北車站S3出口在北方且在使用者的左邊，空間映射：面朝南方時，左邊=北方</p> <p>運動用品店1在東方且在使用者的右邊，空間映射：面朝南方時，右邊=東方</p> <p>結論：使用者面朝南方</p> <p>Final answer: 南方 X</p>

Figure 6: Relation extraction error with original Traditional Chinese model output (ASR-transcribed).

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)	100.0	0.0
	MCoT (ASR)	98.1	0.0

Table 4: Ablation results. **Err.** denotes format error rate (%).

the ASR engine converts “drink shop” to the near-homophone “yin-liu-dian” and “bakery” to “nian-bao-dian”. As a result, the model incorrectly maps multiple distinct spatial relations (e.g., both “front” and “back”) to the same corrupted landmark name in Step 1. While this introduces a logical contradiction, the model often attempts to reconcile these by assigning different coordinates to the same corrupted string in Step 2 to maintain a plausible orientation inference.

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 4.

Findings. Adding spatial coordinates to ASR transcripts (A2→A3) improves accuracy by 10.2 points while substantially reducing format errors. Introducing structured chain-of-thought reasoning on top of multimodal inputs (A3→A4) yields over 70 points in accuracy and eliminates format errors. These results indicate that the structured decomposition is critical for stabilizing egocentric spatial

ASR Misrecognition Error	
Model Input	
Audio: “我現在在公園4，前面是引溜電4，後面是981，左邊是念包電3，右邊是982”	
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)	
問題: 使用者面朝哪個方向? 請逐步推理:	
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=念包電3, 從公園4到念包電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在東方且在使用者的前面, 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	
念包電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	

Figure 7: ASR misrecognition error with original Traditional Chinese model output (ASR-transcribed).

interpretation under transcription noise.

4.5 Evaluation under variation (RQ3)

Unlike referential ambiguity caused by underspecified landmarks, vague directional expressions introduce intrinsic ambiguity in orientation. In COR, such vague directional expressions constitute a subset of the referential ambiguity setting (R3). For this subset, ground-truth annotations are represented as orientation sets and evaluated using set-based accuracy, while the remaining referential ambiguity cases retain single-label evaluation.

We evaluate model behavior under three sources of variation: linguistic variation (R1), unseen spatial environments (R2), and referential ambiguity (R3). Results are summarized in Table 5, with detailed experimental configurations provided in Appendix A.4. Across R1–R3, the model produces well-formed outputs with no format violations, and reasoning traces remain consistent across conditions. In the cross-domain setting (R2), 29 errors are observed, the majority of which involve direction understanding (22/29), followed by ASR misrecognition effects (6/29) and relation extraction errors (1/29). Performance under linguistic variation (R1) and referential ambiguity (R3) exhibits similar error patterns but with fewer total errors.

5 Conclusion and Limitations

Conclusion. This paper examined egocentric-to-allothetic orientation reasoning in conversational

ID	Setting	Acc.	Err.
R1	Linguistic variation	100.0	0.0
R2	Cross-domain	94.6	0.0
R3	Referential ambiguity	99.5	0.0

Table 5: Generalization under variation. **Err.** denotes format error rate (%).

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. 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 high accuracy on clean text and maintain high accuracy under ASR-transcribed input. Beyond end-to-end accuracy, COR supports fine-grained error analysis, revealing that failures in spoken navigation often stem from language grounding issues.

We further compared learning-based models with a deterministic rule-based solver that directly applies the benchmark’s construction rules. While the rule-based solver performs reliably on clean text, its accuracy degrades substantially under ASR noise. In contrast, learning-based models exhibit different error patterns and produce interpretable intermediate outputs that facilitate error attribution, highlighting COR’s value as a diagnostic benchmark beyond aggregate accuracy.

Limitations and future work. COR adopts a discrete 10×10 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 disfluencies. Additionally, while COR focuses on Traditional Chinese to address an under-explored non-English setting, broader multilingual coverage remains open. Future work will extend COR to richer spatial configurations, more realistic speech conditions, and broader linguistic and interaction-centric evaluations, including conversational clarification under ambiguous references.

605
606
607
608
609
610
611

612
613
614
615

616
617
618
619
620

621
622
623
624

625
626
627
628
629

630
631
632
633
634
635

636
637
638
639
640
641

642
643
644
645
646
647

648
649
650
651
652
653
654
655
656

657
658
659
660
661

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 Snaveley, 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, Xiawu 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 Duraiswami, 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, Wenyue 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). 662
663
664
665
666
667
668

Elad Levi and Ilan Kadar. 2025. Intelligent: A multi-agent framework for evaluating conversational ai systems. *Preprint*, arXiv:2501.11067. 669
670
671

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. 672
673
674
675
676

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. 677
678
679
680
681
682
683

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. 684
685
686
687
688
689
690
691
692

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. 693
694
695
696

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. 697
698
699
700
701

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. 702
703
704
705
706

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. 707
708
709
710

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. 711
712
713
714

Smitha Sheshadri and Kotaro Hara. 2024. Conversational localization: Indoor human localization through intelligent conversation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(4). 715
716
717
718

719	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. <i>Versa: A versatile evaluation toolkit for speech, audio, and music</i> . <i>Preprint</i> , arXiv:2412.17667.	775
720		776
721		777
722		778
723		779
724		780
725		781
726		782
727	Qi Sun, Pengfei Hong, Tej Deep Pala, Vernon Toh, U-Xuan Tan, Deepanway Ghosal, and Soujanya Poria. 2024. <i>Emma-x: An embodied multimodal action model with grounded chain of thought and look-ahead spatial reasoning</i> . <i>Preprint</i> , arXiv:2412.11974.	783
728		784
729		785
730		786
731		787
732	Rahul Sundar, Shreyash Gadgil, Tankala Satya Sai, Sathi Sai Krishna Reddy, Gautam B, Ishita Mittal, Jyotsna Sree Guduguntla, and Shanmukesh Pujala. 2024. <i>InnoGUIDEgpt: Integrating conversational interface and command interpretation for navigation robots</i> . In <i>Proceedings of the Third International Conference on AI-ML Systems, AIMLSystems '23</i> , New York, NY, USA. Association for Computing Machinery.	788
733		789
734		790
735		791
736		792
737		793
738		794
739		795
740	Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. <i>Salmonn: Towards generic hearing abilities for large language models</i> . <i>Preprint</i> , arXiv:2310.13289.	796
741		797
742		798
743		799
744		800
745	Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. <i>Large language models are in-context semantic reasoners rather than symbolic reasoners</i> . <i>Preprint</i> , arXiv:2305.14825.	801
746		802
747		803
748		804
749		805
750	Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. <i>Vision-and-dialog navigation</i> . <i>Preprint</i> , arXiv:1907.04957.	806
751		807
752		808
753	Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. <i>Audiobench: A universal benchmark for audio large language models</i> . <i>Preprint</i> , arXiv:2406.16020.	809
754		810
755		811
756		812
757		813
758	Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. <i>Towards understanding chain-of-thought prompting: An empirical study of what matters</i> . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , Toronto, Canada. Association for Computational Linguistics.	814
759		815
760		816
761		817
762		818
763		819
764		820
765		821
766	Zhifei Xie and Changqiao Wu. 2024. <i>Mini-omni: Language models can hear, talk while thinking in streaming</i> . <i>Preprint</i> , arXiv:2408.16725.	822
767		823
768		824
769	Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. <i>Air-bench: Benchmarking large audio-language models via generative comprehension</i> . <i>Preprint</i> , arXiv:2402.07729.	825
770		826
771		827
772		828
773		829
774		830
775	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. <i>Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech</i> . <i>Preprint</i> , arXiv:2309.09510.	831
776		832
777		833
778		834
779		835
780		836
781		837
782		838
783	Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. <i>Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities</i> . <i>Preprint</i> , arXiv:2305.11000.	839
784		840
785		841
786		842
787		843
788		844
789		845
790		846
791		847
792		848
793		849
794		850
795		851
796		852
797		853
798		854
799		855
800		856
801		857
802		858
803		859
804		860
805		861
806		862
807		863
808		864
809		865
810		866
811		867
812		868
813		869
814		870
815		871
816		872
817		873
818		874
819		875
820		876
821		877
822		878
823		879
824		880

825	Reference landmark 2 = Taipei Main Station	Sports store 1 is to the East and the user de-	872
826	Exit S3, Direction vector from Taipei Main	scribes it as “on my right”. Spatial mapping	873
827	Station Exit S2 to Taipei Main Station Exit	rules indicate right = East when the user is	874
828	S3: $(3,4)-(4,4) = (-1,0)$, Direction = West	facing South.	875
829	3. Step 3: Infer user orientation	Final Answer. The user is facing South.	876
830	Restaurant 5 is to the North and the user de-	A.1.3 ASR Misrecognition Error	877
831	scribes it as “on my left”. Spatial mapping	Error Analysis. ASR misrecognition caused the	878
832	rules indicate left = North when the user is	model to incorrectly map multiple spatial relations	879
833	facing South.	to the same landmark, leading to inconsistent rea-	880
834	Taipei Main Station Exit S3 is to the West and	soning. Figure 7 shows the original output.	881
835	the user describes it as “on my left”. Spatial	Input.	882
836	mapping rules indicate left = West when the	Audio: I am at park 4, yin-liu-dian	883
837	user is facing South.	4 is in front of me, 981 is behind	884
838	Final Answer. The user is facing South.	me, nian-bao-dian 3 is on my left, and	885
839	A.1.2 Relation Extraction Error	982 is on my right Coordinates:	886
840	Error Analysis. The model incorrectly extracted	park_4(8,5), drink_shop_4(7,5),	887
841	spatial relations from the input, misidentifying	bar_1(9,5), bakery_3(8,4), bar_2(8,6)	888
842	which landmark corresponds to which direction.	Model Output.	889
843	Figure 6 shows the original output.	1. Step 1: Extract spatial relations	890
844	Input.	Spatial relation 1 = front, reference landmark	891
845	Audio: I am at bus stop 2, Taipei	= yin-liu-dian 4	892
846	Main Station Exit S3 is in front of	Spatial relation 2 = back, reference landmark	893
847	me, Taipei Main Station Exit K7 is on	= yin-liu-dian 4	894
848	my left, and sports store 1 is on my	Spatial relation 3 = left, reference landmark =	895
849	right Coordinates: bus_stop_2(3,3),	nian-bao-dian 3	896
850	Taipei_Main_Station_Exit_S3(3,4),	Spatial relation 4 = right, reference landmark	897
851	Taipei_Main_Station_Exit_K7(2,3),	= bar 2	898
852	sports_store_1(4,3)	2. Step 2: Calculate absolute directions	899
853	Model Output.	Reference landmark 1 = yin-liu-dian 4, Di-	900
854	1. Step 1: Extract spatial relations	rection vector from park 4 to yin-liu-dian 4:	901
855	Spatial relation 1 = left, landmark = Taipei	$(8,5)-(7,5) = (1,0)$, Direction = East	902
856	Main Station Exit S3	Reference landmark 2 = drink shop 4, Direc-	903
857	Spatial relation 2 = right, landmark = sports	tion vector from park 4 to drink shop 4: $(7,5)-$	904
858	store 1	$(8,5) = (-1,0)$, Direction = West	905
859	2. Step 2: Calculate absolute directions	Reference landmark 3 = nian-bao-dian 3, Di-	906
860	Reference landmark 1 = Taipei Main Station	rection vector from park 4 to nian-bao-dian 3:	907
861	Exit S3, Direction vector from bus stop 2 to	$(8,4)-(8,5) = (0,-1)$, Direction = South	908
862	Taipei Main Station Exit S3: $(3,4)-(3,3) =$	Reference landmark 4 = bar 2, Direction vec-	909
863	$(0,1)$, Direction = North	tor from park 4 to bar 2: $(8,6)-(8,5) = (0,1)$,	910
864	Reference landmark 2 = sports store 1, Direc-	Direction = North	911
865	tion vector from bus stop 2 to sports store 1:	3. Step 3: Infer user orientation	912
866	$(4,3)-(3,3) = (1,0)$, Direction = East	Yin-liu-dian 4 is to the East and the user de-	913
867	3. Step 3: Infer user orientation	scribes it as “in front of me”. Spatial mapping	914
868	Taipei Main Station Exit S3 is to the North	rules indicate front = East when the user is	915
869	and the user describes it as “on my left”. Spa-	facing East.	916
870	tial mapping rules indicate left = North when	Drink shop 4 is to the West and the user de-	917
871	the user is facing South.	scribes it as “behind me”. Spatial mapping	918
		rules indicate back = West when the user is	919
			920

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 \rightarrow ASR)

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 \rightarrow ASR loop as a proxy of noise, rather than Whisper’s internal probabilities. TTS was used solely to synthesize controlled variants for evaluation experiments.

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.

- 1005 1. **Step 1: Extract spatial relations**
- 1006 Spatial relation 1 = front, Reference landmark
- 1007 1 = Student Activity Center 1
- 1008 Spatial relation 2 = behind, Reference land-
- 1009 mark 2 = Academic Building B
- 1010 Spatial relation 3 = left, Reference landmark
- 1011 3 = Parking Lot 2
- 1012 Spatial relation 4 = right, Reference landmark
- 1013 4 = Small Plaza 2
- 1014 2. **Step 2: Calculate absolute directions**
- 1015 Reference landmark 1 = Student Activity Cen-
- 1016 ter 1, Direction vector: $(6,1) - (7,1) = (-1,0)$,
- 1017 Direction = West
- 1018 Reference landmark 2 = Academic Building
- 1019 B, Direction vector: $(8,1) - (7,1) = (1,0)$, Di-
- 1020 rection = East
- 1021 Reference landmark 3 = Parking Lot 2, Direc-
- 1022 tion vector: $(7,0) - (7,1) = (0,-1)$, Direction =
- 1023 South
- 1024 Reference landmark 4 = Small Plaza 2, Direc-
- 1025 tion vector: $(7,2) - (7,1) = (0,1)$, Direction =
- 1026 North
- 1027 3. **Step 3: Infer user orientation**
- 1028 Student Activity Center 1 is to the West, and
- 1029 the user describes it as “in front of me.” Spa-
- 1030 tial mapping rules indicate front = West when
- 1031 the user is facing West.
- 1032 Academic Building B is to the East, and the
- 1033 user describes it as “behind me.” Spatial map-
- 1034 ping rules indicate behind = East when the
- 1035 user is facing West.
- 1036 Parking Lot 2 is to the South, and the user
- 1037 describes it as “on my left.” Spatial mapping
- 1038 rules indicate left = South when the user is
- 1039 facing West.
- 1040 Small Plaza 2 is to the North, and the user
- 1041 describes it as “on my right.” Spatial mapping
- 1042 rules indicate right = North when the user is
- 1043 facing West.
- 1044 Therefore, the user is facing West.

1045 A.3.2 Baseline Evaluation Prompts

1046 **Training data format.** All training uses the

1047 LLaMA chat template with Traditional Chinese in-

1048 puts and outputs, as Taiwan-LLM-13B-v2.0-Chat

1049 is optimized for Traditional Chinese. All training

1050 and evaluation were conducted in Traditional Chi-

1051 nese. The prompts below are English translations

1052 for clarity and reproducibility.

1053 **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:

1062 **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:

1092 **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").

Table 7: ASR error severity distribution with exact counts.

Set	Severity	Count	%
Main (Gongguan)	perfect	101	14.5
	minor	222	31.9
	moderate	258	37.1
	major	102	14.7
	severe	13	1.9
Cross-domain (Taipei)	perfect	143	26.5
	minor	207	38.3
	moderate	156	28.9
	major	32	5.9
	severe	2	0.4

- **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.

1206 **Algorithm sketch.** Given an ASR transcript and
1207 landmark coordinates, the solver proceeds:

- 1208 1. Extract egocentric relations (front, back, left,
1209 right) and landmark names using pattern-
1210 based and exact string matching.
- 1211 2. For each extracted landmark, compute the rel-
1212 ative coordinate vector and determine its abso-
1213 lute direction using a dominant-axis heuristic.
- 1214 3. Infer the user orientation by applying fixed
1215 relative-to-absolute mapping rules (Table 1).

1216 If any required landmark cannot be matched ex-
1217 actly, the instance is treated as a failed extraction.

1218 **Observed behavior under ASR noise.** Although
1219 the solver attains high accuracy on clean text
1220 (99.86%), its performance degrades substantially
1221 on ASR-transcribed input. The primary source of
1222 error is not geometric computation, but failures in
1223 matching linguistic references under transcription
1224 noise. Common issues include: (i) homophones
1225 and character substitutions in landmark names; (ii)
1226 truncation or merging of multi-character landmark
1227 tokens; and (iii) inconsistent transcription of al-
1228 phanumeric identifiers (e.g., numeric exits or store
1229 indices). These mismatches often prevent land-
1230 mark resolution, leading to errors in subsequent
1231 orientation inference.

1232 **Discussion.** The behavior of this solver high-
1233 lights a central challenge addressed by COR. In
1234 speech-mediated settings, errors frequently origi-
1235 nate from noisy language input rather than from
1236 spatial reasoning itself. As a result, approaches that
1237 rely solely on exact string matching are brittle un-
1238 der ASR transcription, motivating models that can
1239 integrate linguistic context with spatial reasoning
1240 when interpreting spoken navigation instructions.

1241 **A.7 Stage-wise Supervision and** 1242 **Curriculum Design**

1243 We employ stage-wise supervision to stabilize the
1244 learning of decomposed reasoning components,
1245 rather than to encode or enforce deterministic map-
1246 ping rules. In preliminary experiments, training all
1247 reasoning steps jointly from scratch often resulted
1248 in unstable behavior, such as frequent schema vio-
1249 lations or premature guessing of final orientations
1250 without forming consistent intermediate represen-
1251 tations. Stage-wise supervision affects only the
1252 sequence in which subtasks are introduced, while

keeping the data distribution and supervision sig- 1253
nals unchanged. Earlier stages focus on learning 1254
localized components (e.g., relation extraction and 1255
coordinate-based direction identification), which 1256
helps reduce interference between linguistic pro- 1257
cessing and geometric reasoning when later stages 1258
are trained end to end. 1259

Crucially, this curriculum does not introduce ad- 1260
ditional rules or constraints beyond those used to 1261
define the evaluation labels. The mapping rules in 1262
Table 1 specify the target orientation but are nei- 1263
ther executed nor enforced during inference. The 1264
model must still interpret noisy language input, re- 1265
solve landmark references under ASR transcription 1266
errors, and reconcile multiple spatial cues. Accord- 1267
ingly, stage-wise supervision should be understood 1268
as a means of stabilizing intermediate represen- 1269
tations, not as a mechanism for memorizing rule 1270
templates. In practice, removing early-stage super- 1271
vision primarily increases schema violations and 1272
inconsistency in intermediate steps, while having 1273
limited effect on final orientation accuracy. 1274