Time-Aware Assistive Navigation

Masaki Kuribayashi*, Zhongkai Shangguan*, Eshed Ohn-Bar

Boston University, Waseda University

ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION 2026 VIENNA

Motivation

Assistive vision-and-language agents must go beyond describing surroundings: they need to understand users' needs, anticipate risks, and determine when to intervene or remain silent.

Problem with Current MLLMs:

✗ Mixes up "left" / "right" — a common yet dangerous error
✗ Provides instructions while user crosses intersections
✗ Not trained for blind users; ignores safety-critical timing

TIMELI Agent:

✓ Reasons about when to speak and when to stay silent
✓ Aligned with professional mobility guide practices
✓ Trained with time-aware safety supervision

Task Overview & Contributions

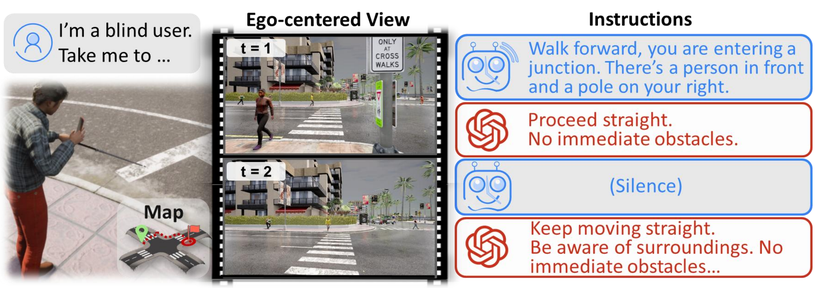


Figure 1: Step-by-step video instruction generation. TIMELI provides temporally-aware, concise, and safe navigation guidance for blind users.

1. **TIMELI Benchmark:** First large-scale, video-based, time-aware benchmark for safety-critical assistive navigation.
2. **Reason Prediction:** Direct supervision to predict the underlying rationale for each instruction.
3. **Comprehensive Evaluation:** Open-loop, closed-loop, and sim-to-real analysis revealing persistent challenges.

TIMELI Benchmark Data

Designed with mobility guides & blind users: Pilot study with 10 blind participants and Orientation & Mobility guides informed instruction timing and safety rules.

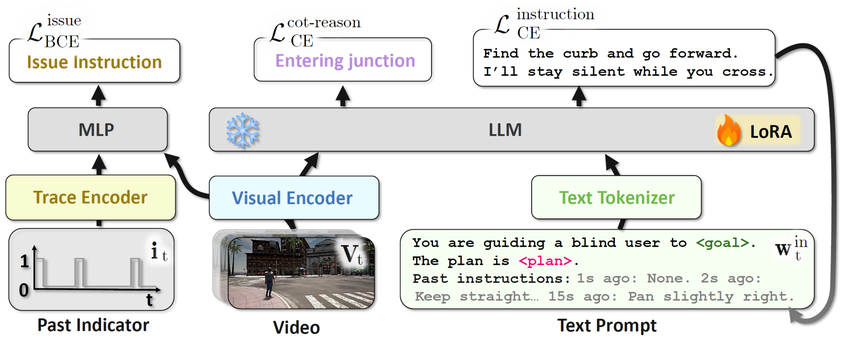
79.1% of frames → model should remain silent.
81.0% forward, **17.6%** turns, **21.6%** at intersections.
Annotated **YouTube real-world** set for sim-to-real evaluation.

Method

Time-aware instruction generation is a sequential **decision-making task**. At each time step t, the goal is to generate an output instruction $\mathbf{w}_t^{\text{out}} = (w_t^1, \ldots, w_t^{N_w})$.

The model receives observations $\mathbf{o}_t = (\mathbf{V}_t, \mathbf{w}_t^{\text{in}})$:

Ego-Video: current + past *T−1* frames $\mathbf{V}_t \in \mathbb{R}^{W\times H\times 3\times T}$.
Navigation Plan: goal coordinate + high-level command (e.g., "turn left") from A* path planner.
Task Specification: user context, safety rules, instruction history with timestamps.
Explicit "None": model must output "None" when no instruction is necessary.



Reason Prediction

We construct $\mathbf{w}_t^{\text{out}}$ so the model first predicts the rationale before generating the instruction, which enhances the model's ability to reason over instruction timing and enables interpretable prediction diagnostics.

Timing Supervision

A **binary classifier head** determines whether an instruction should be issued:
Instruction history encoded as 1D binary vector
Combined with visual encoder features via MLP;
Serves as **gating mechanism** for early exit during inference → reduced latency and computation.

Overall Loss Function:

$$\mathcal{L}_{\text{out}} = \underbrace{\mathcal{L}_{\text{CE}}^{\text{cot-reason}}}_{\text{Why to speak}} + \underbrace{\mathcal{L}_{\text{CE}}^{\text{instruction}}}_{\text{What to say}} + \beta \underbrace{\mathcal{L}_{\text{BCE}}^{\text{issue}}}_{\text{When to speak}}$$

where $\mathcal{L}_{\text{CE}} = \exp\left[-\frac{1}{N_w}\sum_{n=1}^{N_w}\log P(w^n \mid w^1, \ldots, w^{n-1})\right]$

Experiment Results

Open-Loop Model Evaluation

Model	Size	Modality	BLEU-4	ROUGE-L	Timing F1	Timing AUC	Conciseness
Off-the-Shelf Models							
VILA [26]	3B	Image	0.626	6.545	0.579	0.778	0.554
TinyLLaVA [106]	3B	Image	0.000	5.728	0.361	0.500	0.264
VILA [26]	3B	Video	0.000	4.377	0.465	0.651	0.552
TimeChat [21]	7B	Video	0.000	7.423	0.379	0.536	0.026
NavGPT [28]	-	Image	2.536	12.998	0.042	0.508	0.167
GPT-4o [8]	-	Image	7.395	20.059	0.451	0.654	0.220
GPT-4o [8]	-	Video	2.536	12.998	0.042	0.508	0.245
Finetuned Models on TIMELI Benchmark							
TinyLLaVA	3B	Image	0.000	0.000	0.000	0.500	0.000
TinyLLaVA+	3B	Image	13.294	27.730	0.254	0.573	0.325
VILA	3B	Video	8.618	20.063	0.540	0.726	0.048
VILA+	3B	Video	10.916	23.789	0.410	0.629	0.300
TimeChat	7B	Video	0.000	6.788	0.030	0.508	0.067
TimeChat+	7B	Video	11.891	29.127	0.101	0.525	0.316
LLaVA-v1.6	7B	Image	12.398	24.765	0.572	0.749	0.064
LLaVA-v1.6+	7B	Image	20.390	38.250	0.776	0.869	0.395

Closed-Loop Model Evaluation

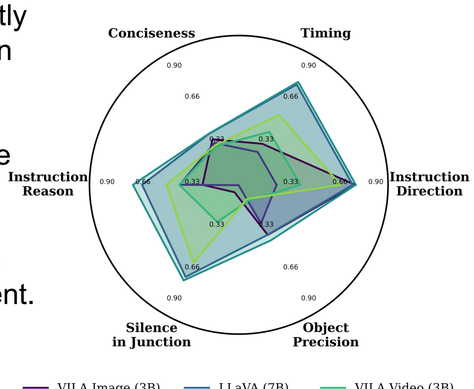
Model	Size	Modality	Success Rate	Route Completion	Navigation Score	Navigation Error (m)	Collision/Min	Instructions/Min
TinyLLaVA	3B	Image	0.00	0.01	0.01	0.30	0.00	0.00
TinyLLaVA+	3B	Image	0.58	0.91	0.91	1.28	0.24	19.02
VILA	3B	Video	0.08	0.52	0.41	46.37	0.63	7.52
VILA+	3B	Video	0.33	0.85	0.68	2.38	0.59	16.58
LLaVA-v1.6	7B	Image	0.00	0.36	0.28	47.60	0.35	13.45
LLaVA-v1.6+	7B	Image	0.42	0.93	0.92	1.48	0.58	19.94
TimeChat	7B	Video	0.00	0.01	0.01	0.38	0.00	0.00
TimeChat+	7B	Video	0.25	0.74	0.59	9.19	0.55	3.54

Sim-to-Real Transfer

Model	Size	Modality	BLEU-4	ROUGE-L	Timing F1	Timing AUC	Conciseness
TinyLLaVA+	3B	Image	0.065	1.592	0.454	0.500	0.019
VILA+	3B	Video	7.005	20.359	0.186	0.545	0.253
LLaVA-v1.6+	7B	Image	8.478	22.587	0.732	0.815	0.278
TimeChat+	7B	Video	0.001	22.995	0.025	0.506	0.383

Key Findings

1. Existing models frequently fail across key navigation dimensions: timing, conciseness, reasoning, junction silence, obstacle precision, and direction correctness.
2. Reason prediction is the most impactful component.
3. Sim-to-real transfer is promising, validating CARLA synthetic training.



Check Our Website *timeli-icra.github.io*