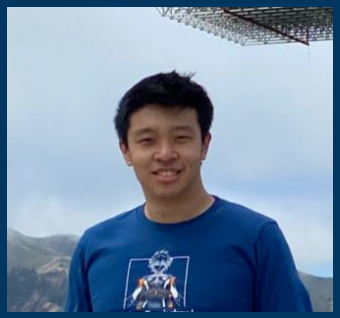
ULN: Towards Underspecified Vision-and-Language Navigation

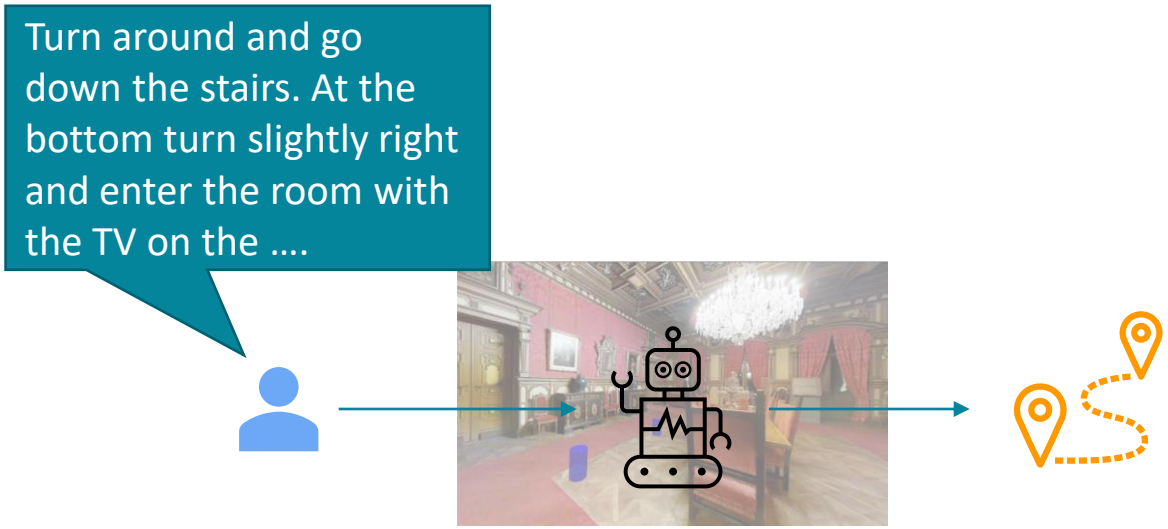



Weixi Feng Tsu-jui Fu Yujie Lu William Wang

UC Santa Barbara

Vision-and-Language Navigation

- The embodied agent navigates to a target location by following human language instructions.
  - Instruction understanding
  - Alignment between linguistic semantics and visual observation



Problems & Motivation

- Many existing VLN datasets consist of low-level instructions that describe every single step for the agent.
  - R2R, R4R, RxR
- Agents are usually trained on a single type of instructions and may not generalize well in real applications
  - Low-level datasets
  - High-level datasets: REVERIE, SOON

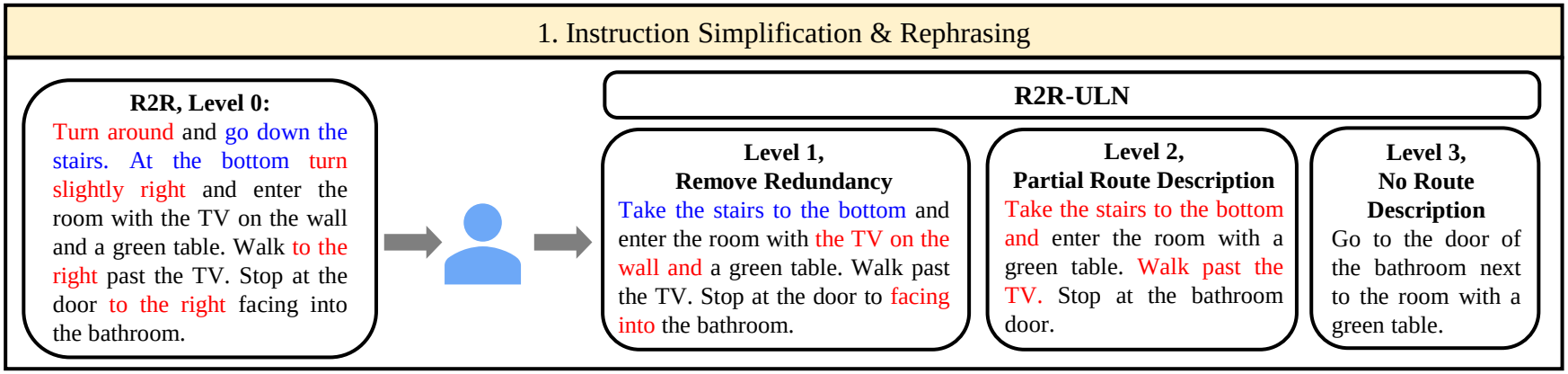
- In reality, users tend to omit some details when instructing robot in an indoor environment.
- Evaluate agents with **multi-level underspecified** instructions to probe their generalization to language variations.

Contributions

- A new evaluation dataset, i.e., R2R-ULN
- A model-agnostic framework to improve performance on underspecified instructions.

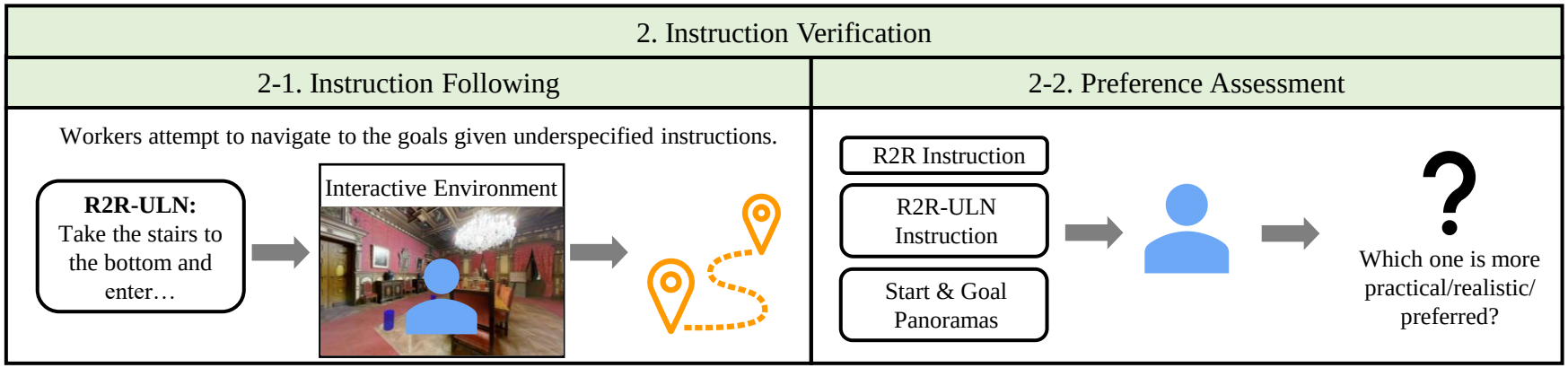
Approach

- Data Collection



Approach

- Data Collection

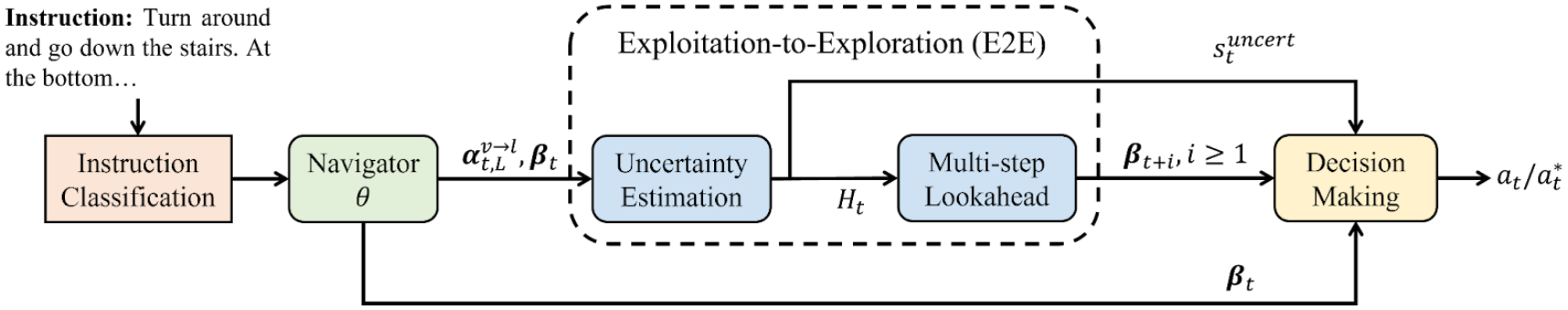


Dataset Statistics

Number of:	R2R	R2R-ULN		
	L0	L1	L2	L3
Instructions	1622	3282	3282	3282
Paths	917	917	917	917
Tokens	38.6	27.3	18.9	8.7
Direction Tokens	2.5	1.1	0.7	0.2
Object Tokens	8.5	6.6	4.6	2.6

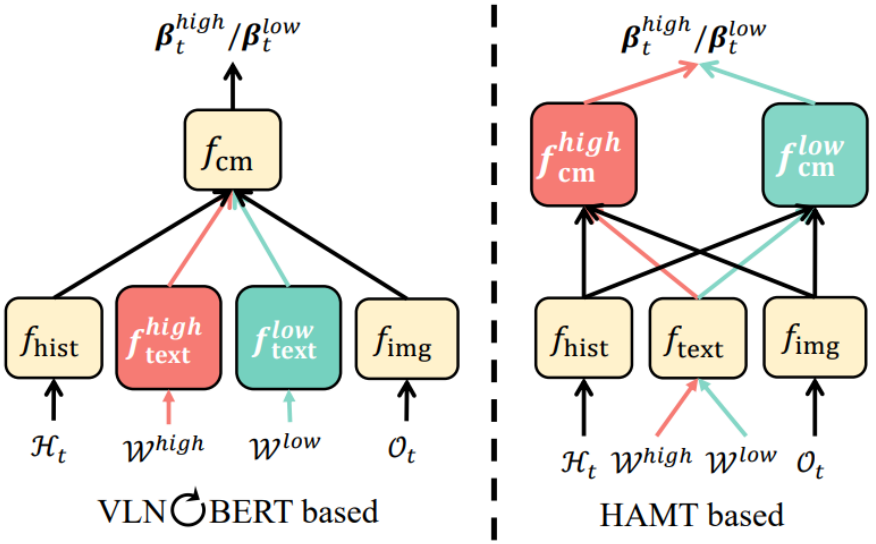
Level	R2R-ULN Val-Unseen			
	Instr. Following		Instr. Preference	
	SR↑	SPL↑	Practicality	Efficiency
L_0	86	72	-	-
L_1	82	68	55%	57%
L_2	82	65	63%	59%
L_3	75	58	68%	66%

Model-Agnostic Framework



- A hybrid navigation agent
- Exploitation-to-Exploration: multi-step lookahead

Agent: Granularity-Specific Subnetworks

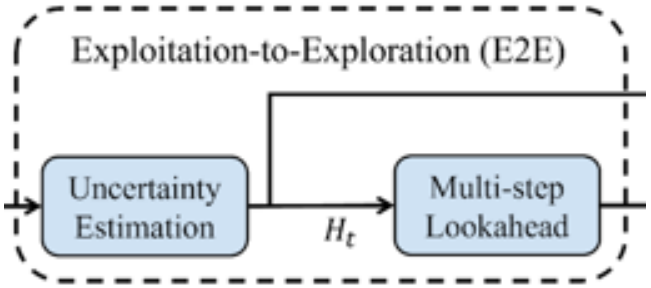


Instruction Classification

Turn around and go down the stairs. At the bottom turn slightly right and enter the room with the TV on the wall… (L_0, L_1, L_2)

Go to the living room. (L_3)

E2E: Multi-Step Lookahead



- Uncertainty Estimation: A two-layer network to predict step-wise uncertainty score based on cross-attention weights $\boldsymbol{\alpha}_{t,L}^{v\rightarrow l}$ and decision logits $\boldsymbol{\beta}_t$

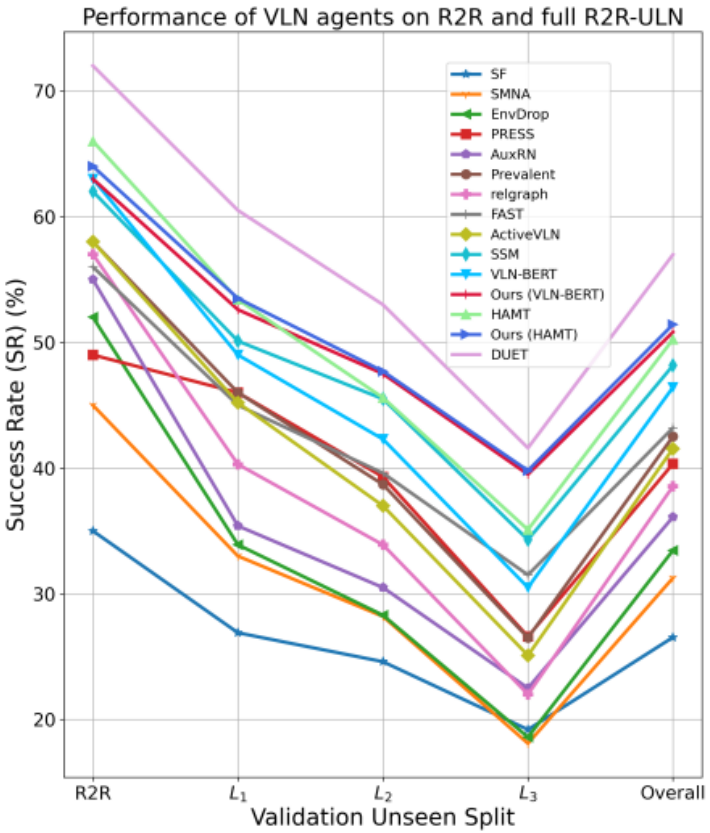
$$s_t^{\text{uncert}} = f_{\text{uncert}}([\boldsymbol{\alpha}_{t,L}^{\text{v}\rightarrow\text{t}}; \boldsymbol{\beta}_t]).$$

- Multi-step Lookahead: Active explore the future steps when the agent is uncertain:

$$a_t^* = \arg\max_c \left[\beta_{t,c} + \sum_{i=1}^{K} \gamma^i \max_{c'}(\beta_{t+i,c'})\right].$$

Experimental Results

	Methods	R2R-ULN Val-Unseen			
		TL	NE↓	SR↑	SPL↑
0	Human	14.97	2.94	77.4	61.7
	Greedy-Decoding Agents				
1	Speaker-Follower (Fried et al., 2018)	14.86	8.43	22.0	17.4
3	EnvDrop (Tan et al., 2019)	8.74	8.26	24.6	23.3
4	PREVALENT (Hao et al., 2020)	11.91	7.28	33.8	31.1
	Exploration-based Agents				
6	FAST-Short (Ke et al., 2019)	22.89	6.78	36.8	26.4
7	Active VLN (Wang et al., 2020)	19.40	7.08	32.2	21.2
8	SSM (Wang et al., 2021)	26.64	6.70	39.8	26.1
9	VLN↻BERT (Hong et al., 2021)	13.00	6.47	39.3	35.0
11	Ours (VLN↻BERT-based, w/ E2E)	23.02	**6.13**	**44.7**	29.7
12	HAMT (Chen et al., 2021a)	12.98	6.33	41.7	37.6
14	Ours (HAMT-based, w/ E2E)	28.31	**6.05**	**44.6**	25.9



- Paper: https://arxiv.org/abs/2210.10020
- Dataset & Code: https://github.com/weixi-feng/ULN

UC SANTA BARBARA