Environment-agnostic Multitask Learning for Natural Language Grounded Navigation

Xin (Eric) Wang*, Vihan Jain*, Engene Ie, William Yang Wang,
Zornitsa Kozareva, Sujith Ravi

ECCV'20 ONLINE
23-28 AUGUST 2020

Natural Language Grounded Navigation

Command embodied agents to navigate in the 3D world with **natural language,** such as coarse-/fine-grained instructions, questions, dialog

Person:

Can you grab the plant for me?

Sure. Where is it?

Get out of the room and go towards the kitchen. The plant is on the window near the kitchen.

Gotcha.

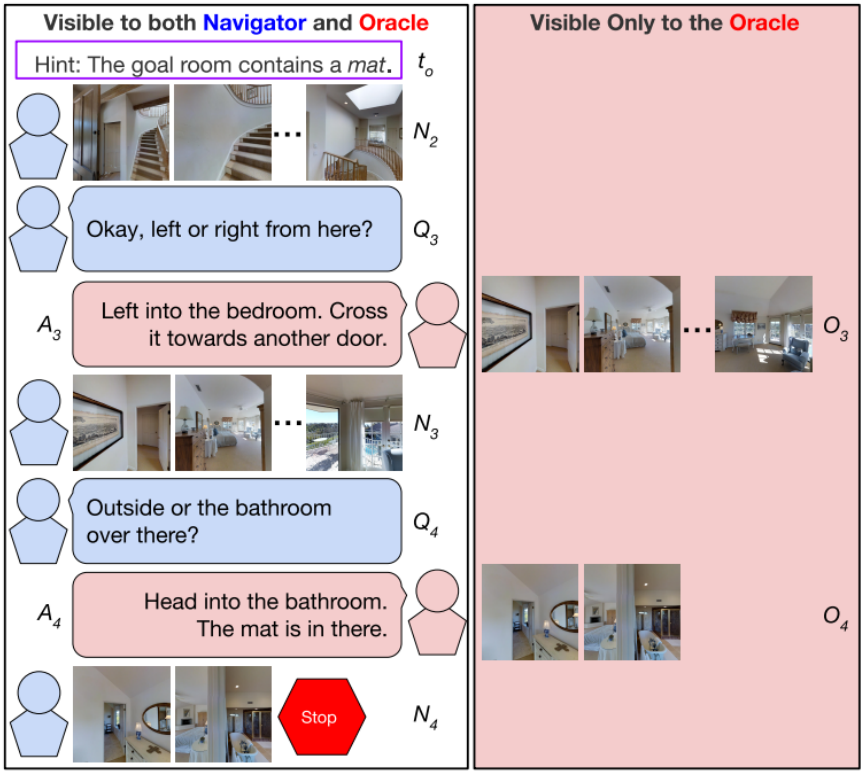
Vision-and-Language Navigation (VLN)

- Given fine-grained instruction and a starting location
- Agent must reach the target location by following the natural language instruction
- Room-to-Room (R2R) Dataset

Walk beside the outside doors and behind the chairs across the room.

Anderson et al., CVPR 2018

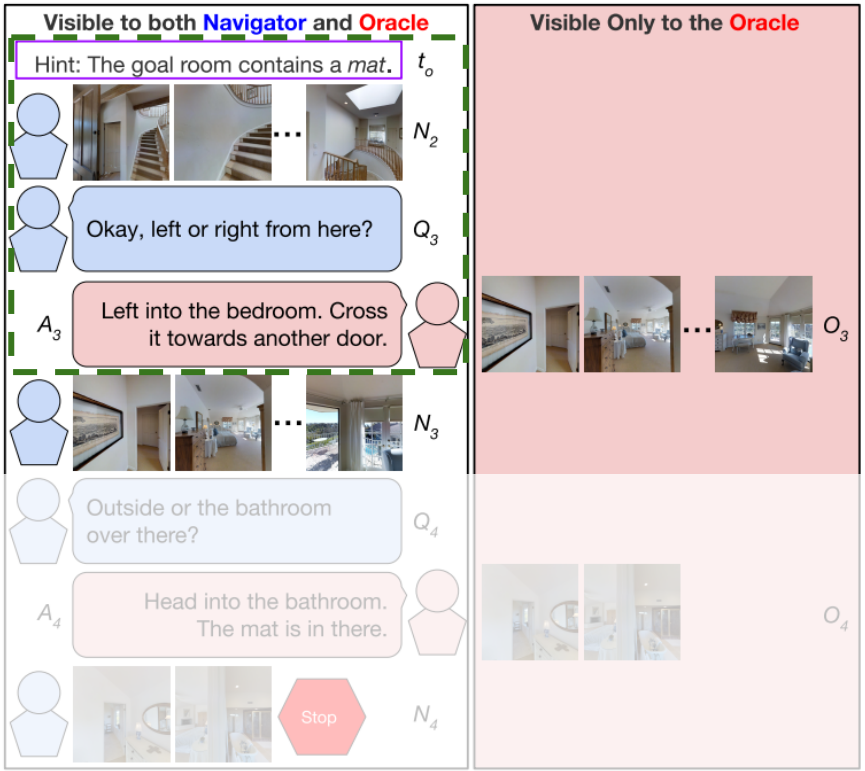
Cooperative Vision-and-Dialog Navigation (CVDN)



- Both Navigator and Oracle are given a **hint** (e.g., the goal room contains a mat)
- **Navigator**: go towards the goal room and can stop anytime to ask a question
- **Oracle**: foresee the next best steps and answer the questions

Thomason et al., CoRL 2019

Sub-task: Navigation from Dialog History (NDH)



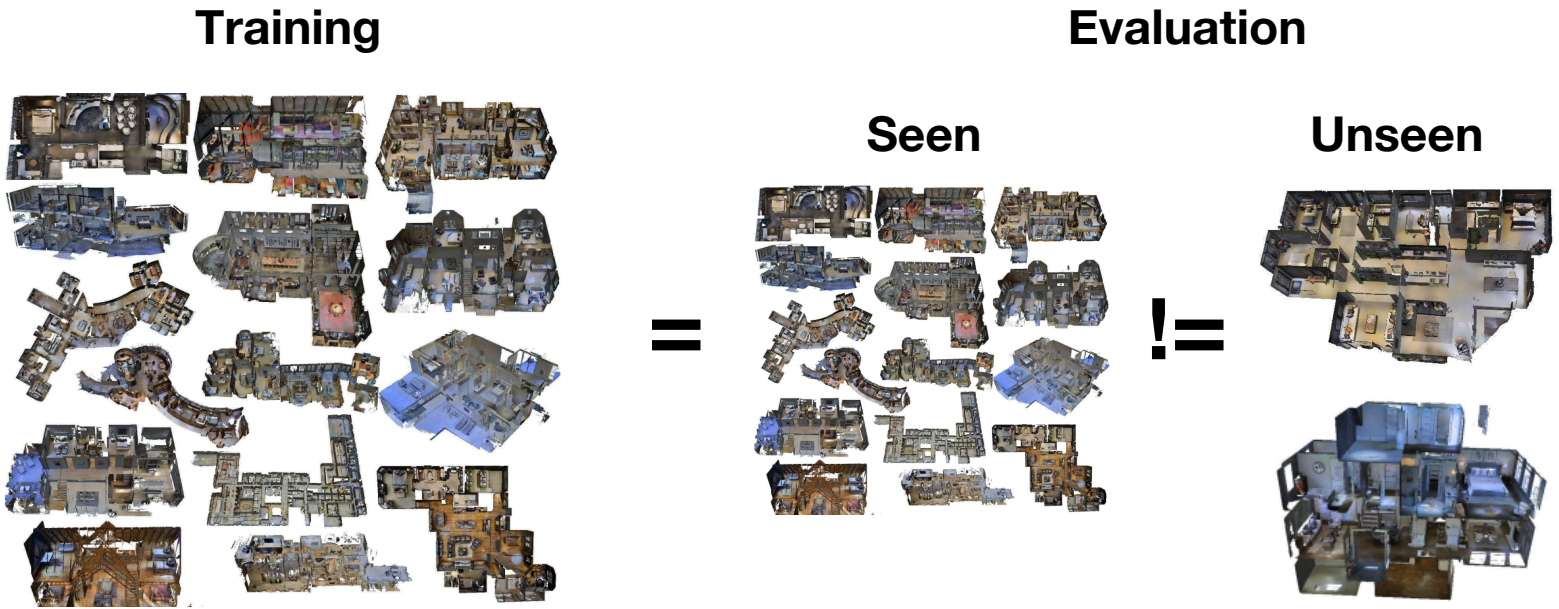
- Given the **dialogue history**, predict the **navigation actions** that bring the agent closer to the goal room

Thomason et al., CoRL 2019

Challenge

Poor Generalization Issue

- Navigation models tend to overfit seen environments and perform poorly on unseen environments



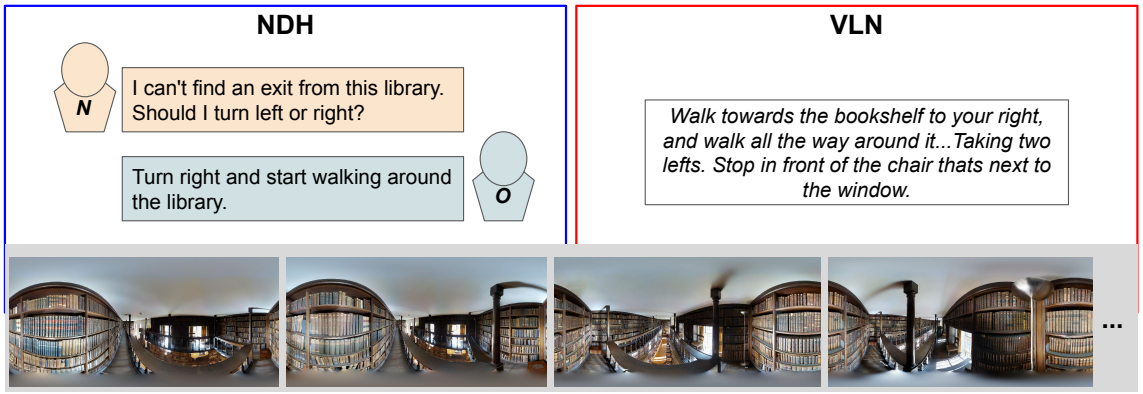
Data Scarcity is A Big Problem

- Real-world experiments are **NOT scalable**
- Data collection is **prohibitively expensive and time-consuming**
- Models break under **distribution shift**

Environment-agnostic Multitask Navigation

Towards Generalizable Navigation

- **Multitask learning**: transfer knowledge across tasks



- **Environment-agnostic learning**: invariant representations that can be better generalized on unseen environments

A Strong Baseline for VLN: RCM

Leave the living room. Go through the hallway with paintings on the wall and head to the kitchen. Stop next to the wooden dining table.

VLN Instruction

Word Embedding

Language Encoder

Paired Demo Path

Panoramic Features $\{\boldsymbol{v}_{t,j}\}_{j=1}^{m}$

Trajectory Encoder

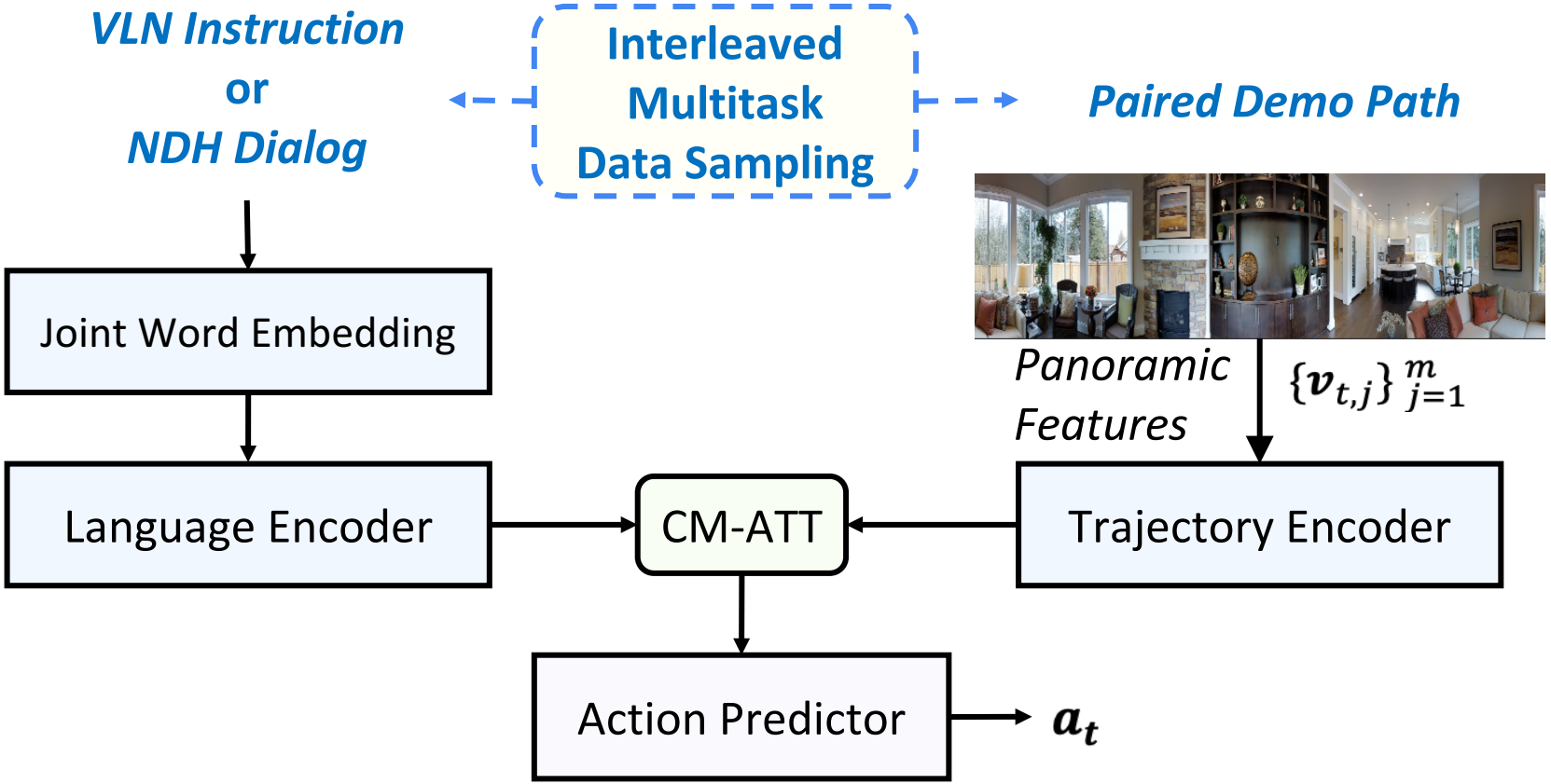
CM-ATT

Action Predictor

$\boldsymbol{a}_t$

Wang et al., CVPR 2019

Multitask RCM



Multitask Reinforcement Learning

- Navigation Loss: Reinforcement Learning + Supervised Learning

$$\mathcal{L}_{nav} = -\mathbb{E}_{a_t \sim \pi}[R(s_t, a_t) - b] - \mathbb{E}[\log \pi(a_t^* | s_t)]$$

- Reward shaping:
  - VLN: Distance to Goal

$$R(s_t, a_t) = \sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'}),$$

$$\text{where } r(s_{t'}, a_{t'}) = \begin{cases} d(s_{t'}, v_{tar}) - d(s_{t'+1}, v_{tar}) & \text{if } t' < T \\ \mathbb{1}[d(s_T, v_{tar}) \leq d_{th}] & \text{if } t' = T \end{cases}$$

  - NDH: Distance to Room

$$d(s_t, \{v_i\}_1^N) = \min_{1 \leq i \leq N} d(s_t, v_i)$$

Effect of Multitask RL

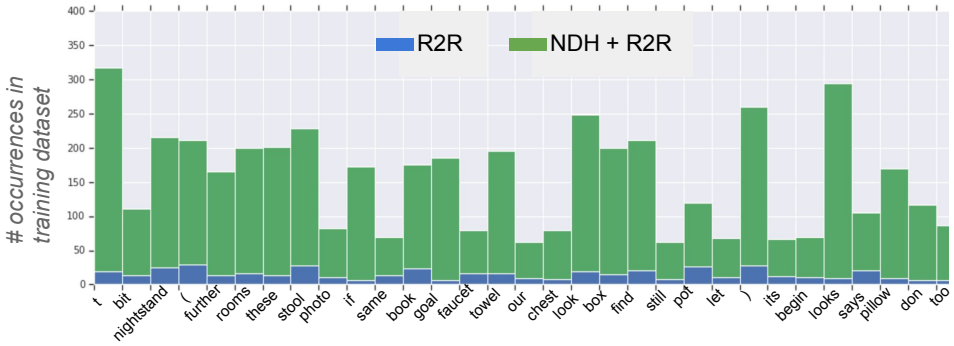
Fold	Model	NDH Evaluation: Inputs for NDH t_o	A_i	Q_i	$A_{1:i-1}; Q_{1:i-1}$	NDH Evaluation: Progress ↑	VLN Evaluation: PL	NE ↓	SR ↑	SPL ↑	CLS ↑
Val Seen	NDH-RCM	✓				**6.97**					
		✓	✓			**6.92**					
		✓	✓	✓		**6.47**					
		✓	✓	✓	✓	**6.49**					
	VLN-RCM						10.75	5.09	52.39	48.86	63.91
	MT-RCM	✓				3.00	11.73	4.87	54.56	52.00	65.64
		✓	✓			5.92	11.12	4.62	54.89	**52.62**	66.05
		✓	✓	✓		5.43	10.94	**4.59**	54.23	52.06	66.93
		✓	✓	✓	✓	5.28	10.63	5.09	**56.42**	49.67	**68.28**
Val Unseen	NDH-RCM	✓				1.25					
		✓	✓			2.69					
		✓	✓	✓		2.69					
		✓	✓	✓	✓	2.64					
	VLN-RCM						10.60	6.10	42.93	38.88	54.86
	MT-RCM	✓				**1.69**	13.12	5.84	42.75	38.71	53.09
		✓	✓			**4.01**	11.06	5.88	42.98	40.62	54.30
		✓	✓	✓		**3.75**	11.08	5.70	44.50	39.67	54.95
		✓	✓	✓	✓	**4.36**	10.23	**5.31**	**46.20**	**44.19**	**54.99**

- NDH benefits from VLN
- VLN benefits from NDH with more fine-grained information about paths
  - Extending visual paths alone is NOT helpful
- Multitask RL improves generalization
  - Seen-unseen gap is narrowed

Effect of Multitask RL

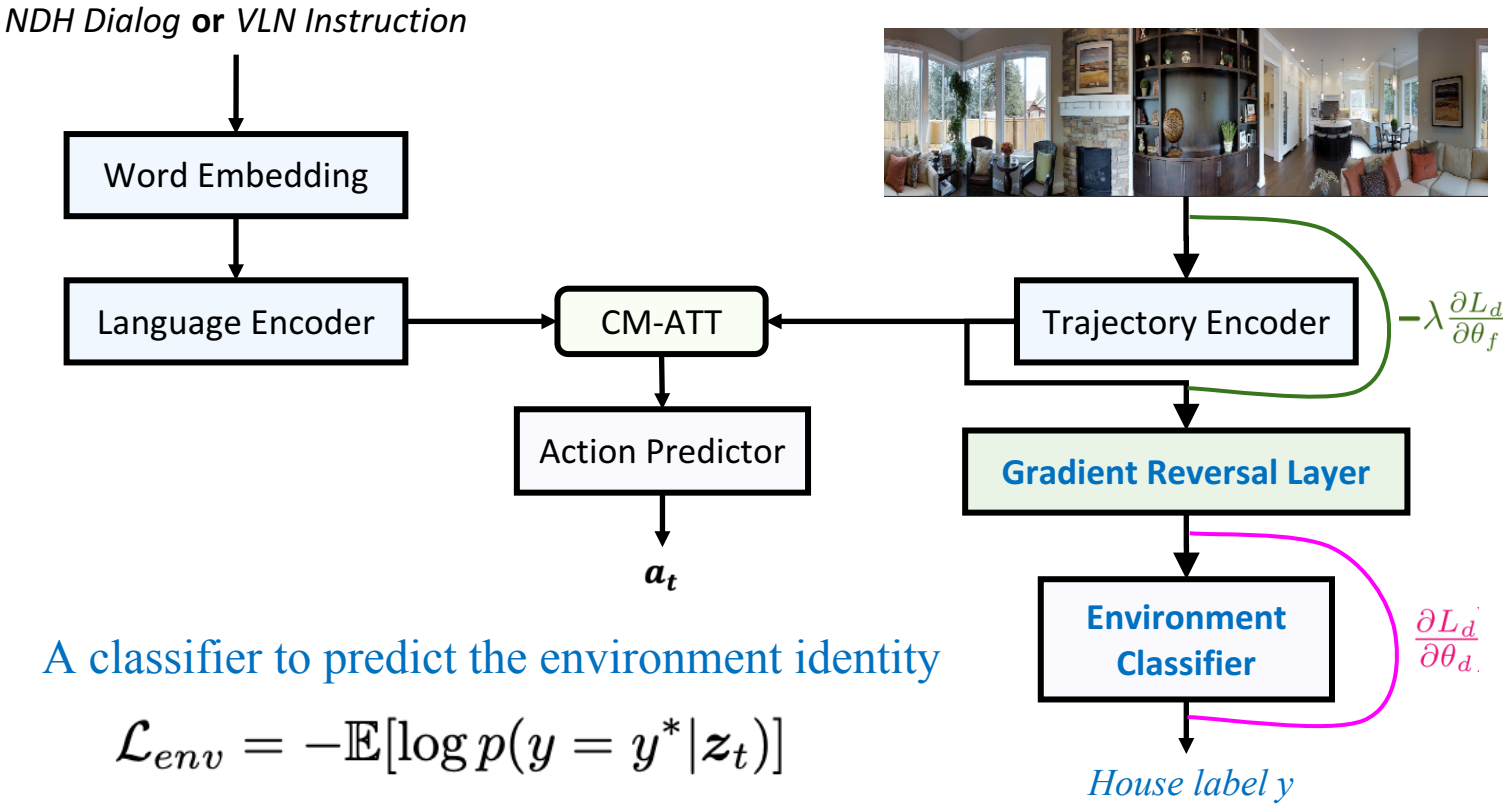
Multitask learning benefits from

- More appearances of unrepresented words
- Shared semantic encoding of the whole sentences



Language Encoder	Val Seen						Val Unseen					
	NDH	VLN					NDH	VLN				
	Progress ↑	PL	NE ↓	SR ↑	SPL ↑	CLS ↑	Progress ↑	PL	NE ↓	SR ↑	SPL ↑	CLS ↑
Shared	**5.28**	10.63	5.09	**56.42**	**49.67**	**68.28**	**4.36**	10.23	**5.31**	**46.20**	**44.19**	**54.99**
Separate	5.17	11.26	**5.02**	52.38	48.80	64.19	4.07	11.72	6.04	43.64	39.49	54.57

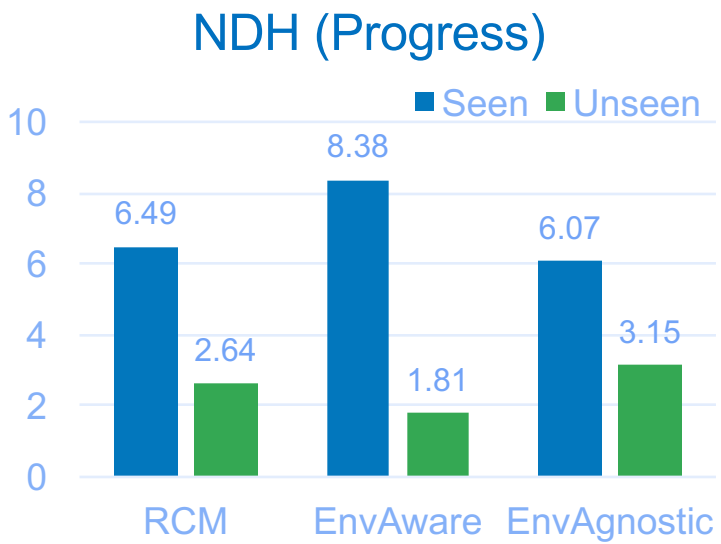
Environment-agnostic Representation Learning

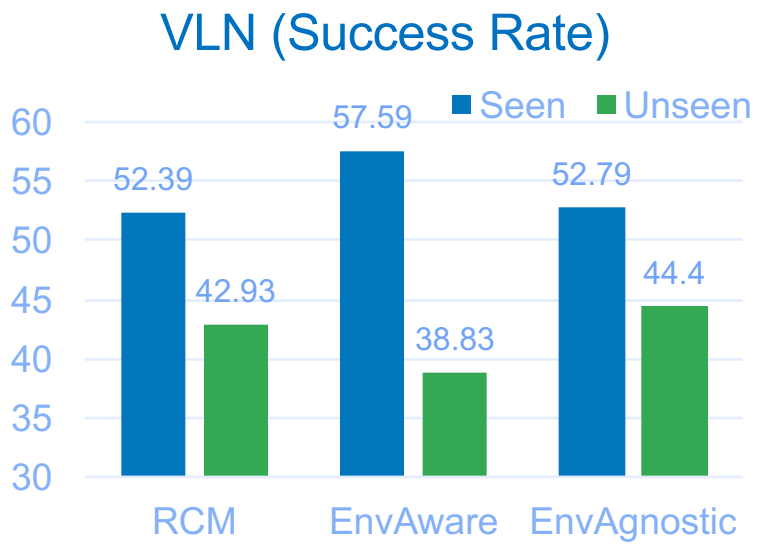


- A classifier to predict the environment identity

$$\mathcal{L}_{env} = -\mathbb{E}[\log p(y = y^* | \boldsymbol{z}_t)]$$

Environment-Aware *versus* Environment-Agnostic

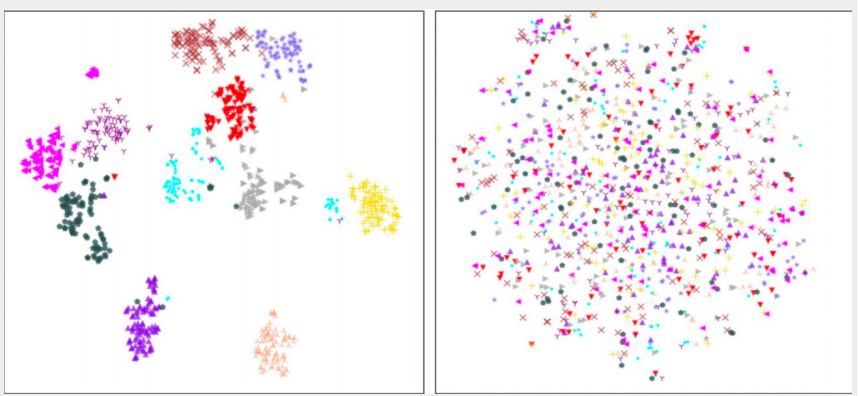




- Env-aware learning tends to overfit seen environments
- Env-agnostic learning generalizes better on unseen environments
- (Potential) Meta-learning with env-aware & env-agnostic may benefit from both worlds

Environment-Aware *versus* Environment-Agnostic

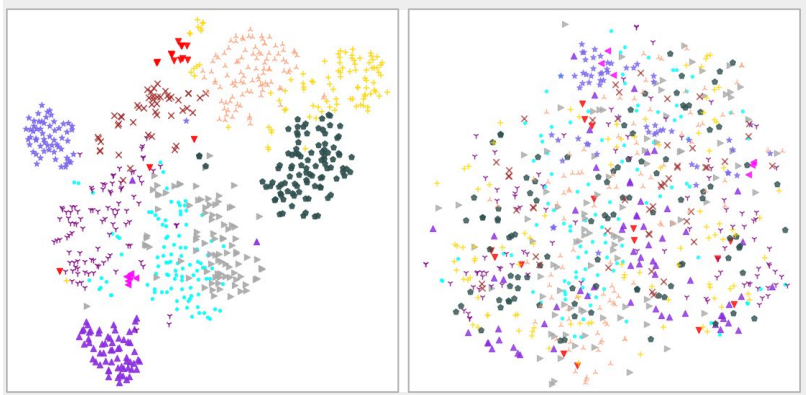
Seen



(a) *(b)*

EnvAware **EnvAgnostic**

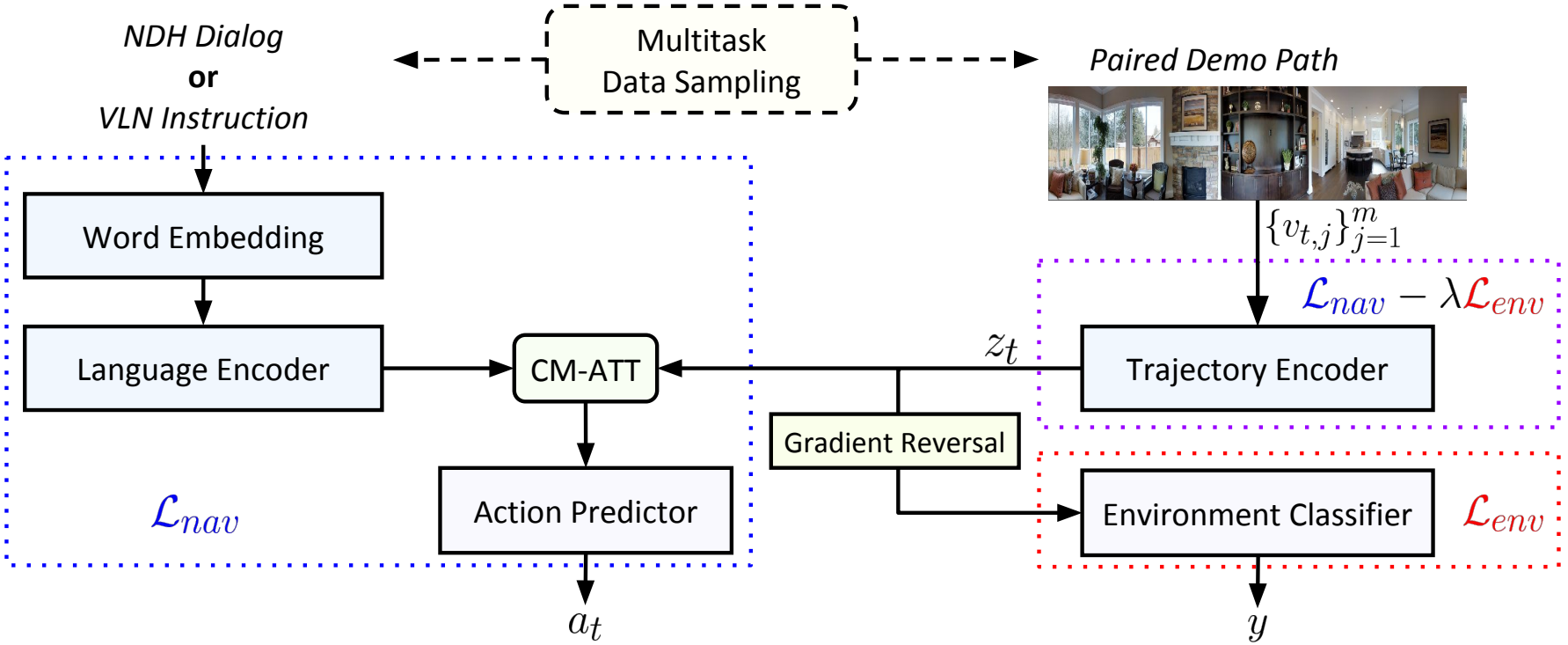
Unseen



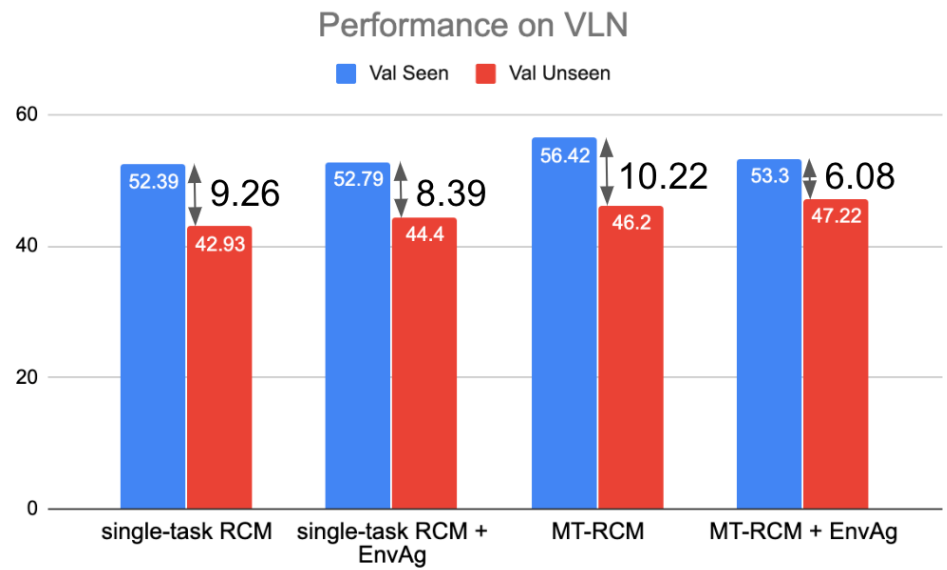
(a') *(b')*

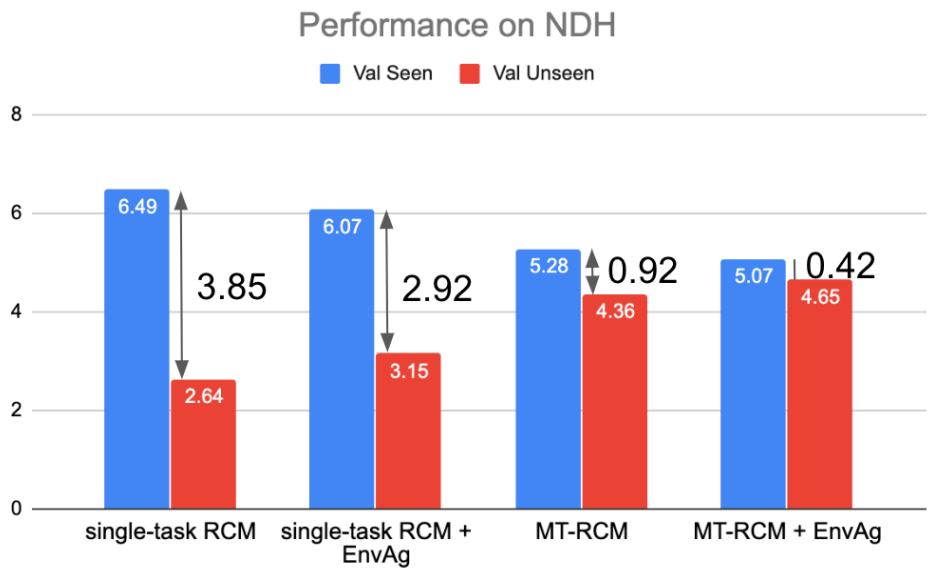
EnvAware **EnvAgnostic**

Environment-agnostic Multitask Learning Framework



Effect of Environment-agnostic Multitask Learning





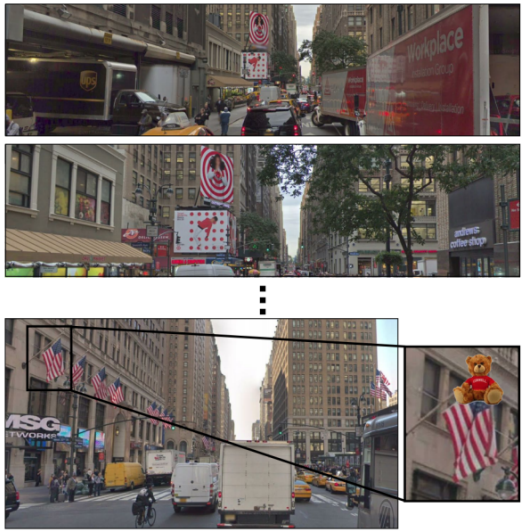
Ranking 1st on CVDN Leaderboard

Rank	Participant team	spl	dist_to_end_reduction	Last submission at
1	Environment-agnostic Multitask Learning	0.17	3.91	4 months ago
2	Test2-NDH	0.09	3.44	4 months ago
3	ed	0.14	3.30	2 months ago
4	lbq1991	0.19	3.00	1 month ago
5	CMN (Cross-modal Memory Network)	0.14	2.95	7 months ago
6	UVLN (aux)	0.11	2.75	7 months ago
7	PREVALENT	0.24	2.44	8 months ago
8	CVDN Seq2Seq Baseline (Seq2Seq Baseline)	0.16	2.35	9 months ago
9	Pansy	0.15	1.76	4 months ago

https://evalai.cloudcv.org/web/challenges/challenge-page/463/leaderboard/1292

Future Work

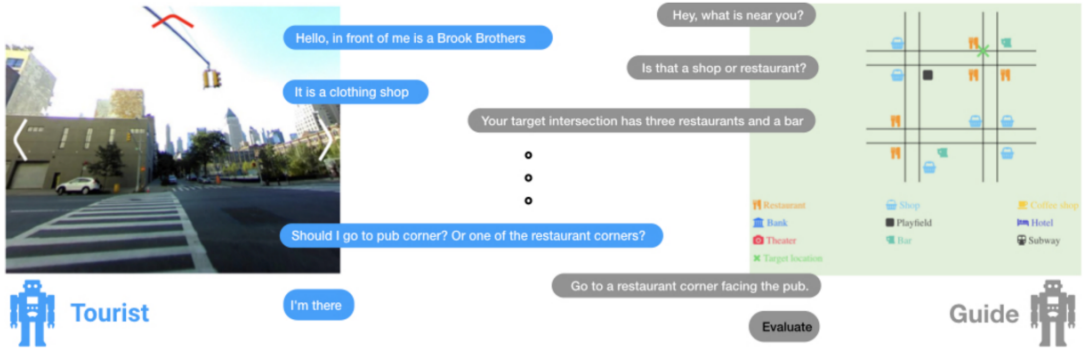
Generalized Navigation on Street View



Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

TouchDown (Chen, et al. 2019)

StreetLearn
(Mirowski, et al. 2018)



TalkTheWalk (de Vries, et al. 2018)

Thanks!

Paper: https://arxiv.org/abs/2003.00443

Code: https://github.com/google-research/valan