Generalizable VLN Methods

Xin (Eric) Wang
Assistant Professor @ UCSC

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Contributors to Deep Learning Success

- Algorithms (*what we'd like to believe*)
- Computation
- **Data**

http://tensorlab.cms.caltech.edu/users/anima/slides/data-scarcity-active-learning.pdf

Data Scarcity is A Big Problem

- DL requires **big data**, often prohibitively expensive to collect
- DRL is impressive but brittle, suffering from **high sample complexity**
- In general, these models break under **distribution shift**

Vision-and-Language Navigation (VLN)

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.

Data Scarcity in VLN

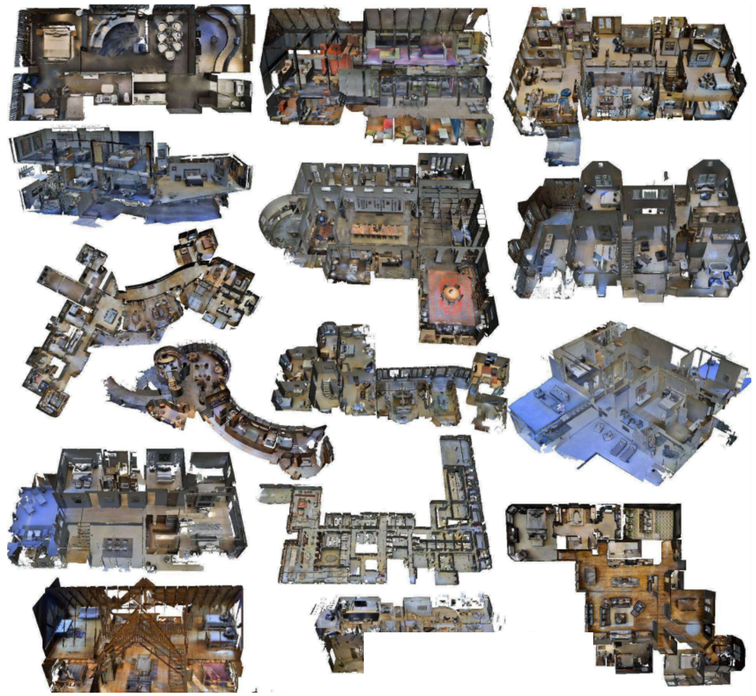
- Real-world experiments are not scalable
- 3D room scan and reconstruction is limited and expensive
- Human instructions are hard to collect in interactive environments

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Poor Generalization Issue

- Models tend to overfit seen environments and perform poorly on unseen environments

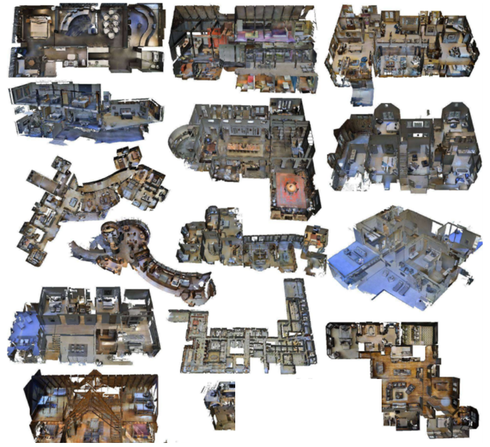
Training



=

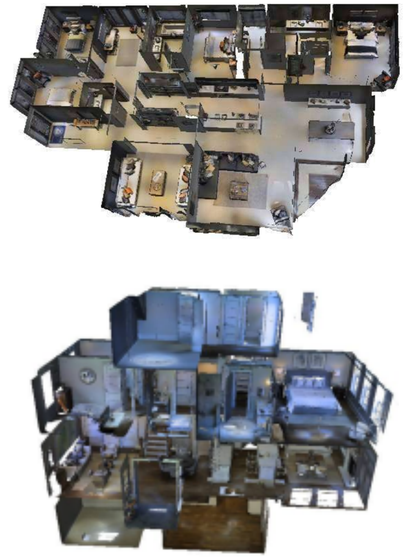
Evaluation

Seen



!=

Unseen



Generalizable VLN Methods

1. Data augmentation
2. Evaluation of generated navigation instructions
3. Multitask learning
4. Unseen environment adaptation

Generalizable VLN Methods

1. **Data augmentation**
2. Evaluation of generated navigation instructions
3. Multitask learning
4. Unseen environment adaptation

UCSB
UNIVERSITY OF CALIFORNIA
SANTA BARBARA

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Counterfactual Vision-and-Language Navigation via Adversarial Path Sampling

Tsu-Jui Fu, Xin Eric Wang, Matthew F. Peterson, Scott T. Grafton,
Miguel P. Eckstein, William Yang Wang, ECCV 2020

ECCV'20
ONLINE
23-28 AUGUST 2020

Counterfactual VLN

- **Counterfactual thinking** is a concept in psychology that involves the human tendency to create possible alternatives to life events that have already occurred

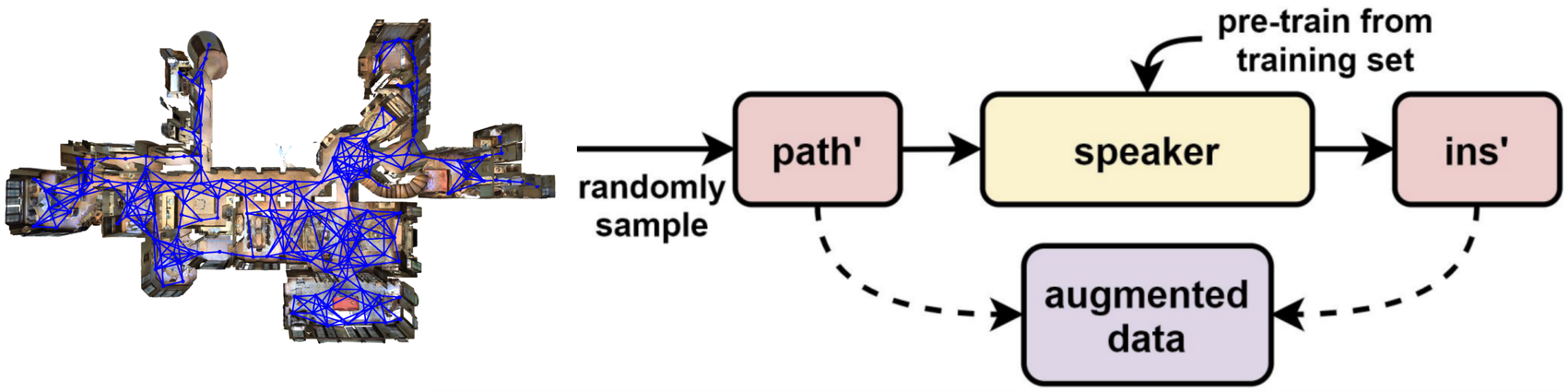
- "*If we stop in front of the dining table instead of walking away from it, what should the instruction be?*"

- **Adversarial Path Sampling (APS)** takes counterfactual actions and samples increasingly challenging paths

- Speaker to augment those paths with new instructions → a more effective VLN agent

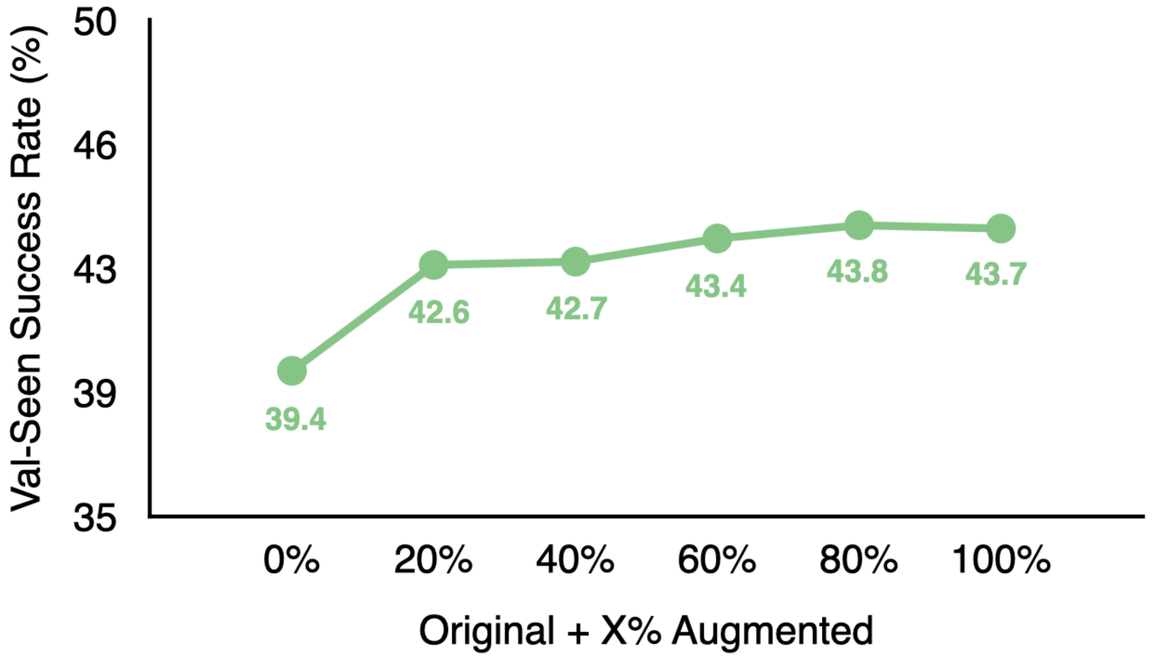
Speaker-Driven Data Augmentation

- Data augmentation with **Speaker**:
  - Training set D only covers limited path-instruction pairs
  - Randomly sample paths from training environments
  - Use a Speaker to synthesize instructions on paths → D'
  - Train the agent on $D \cup D'$



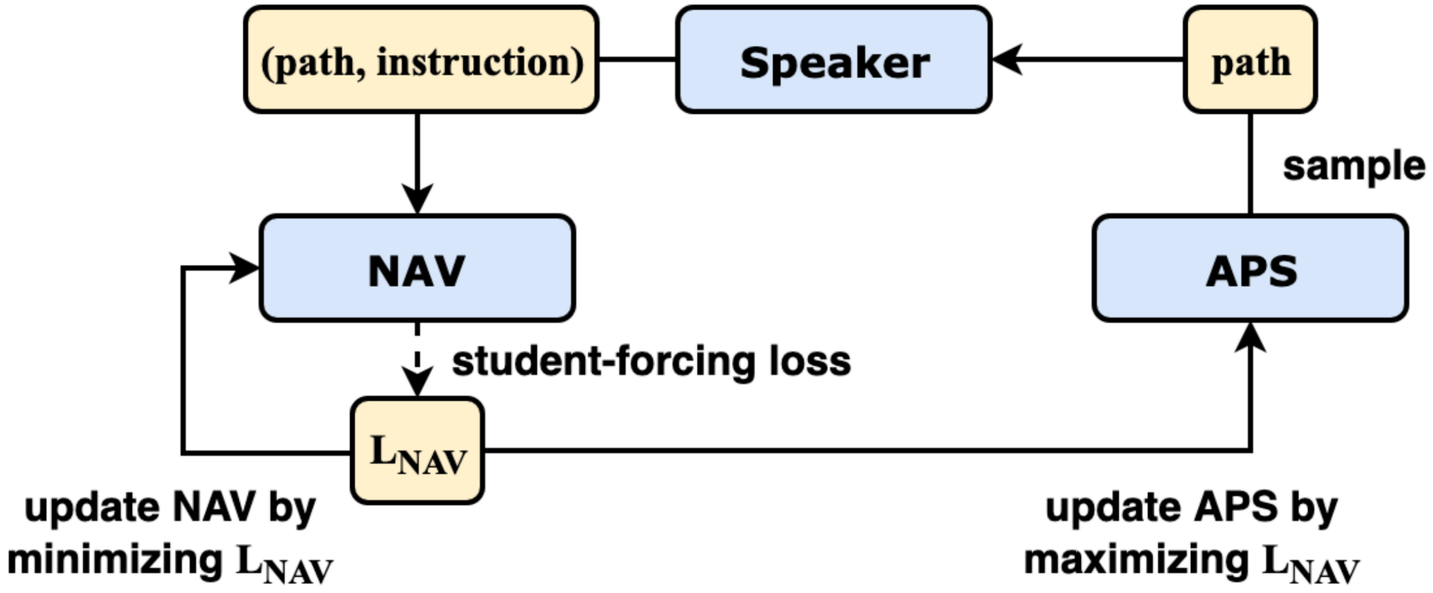
Fried et al., “Speaker-Follower Models for Vision-and-Language Navigation”, NeurIPS 2018

Randomly-Sampled Paths Are Inefficient

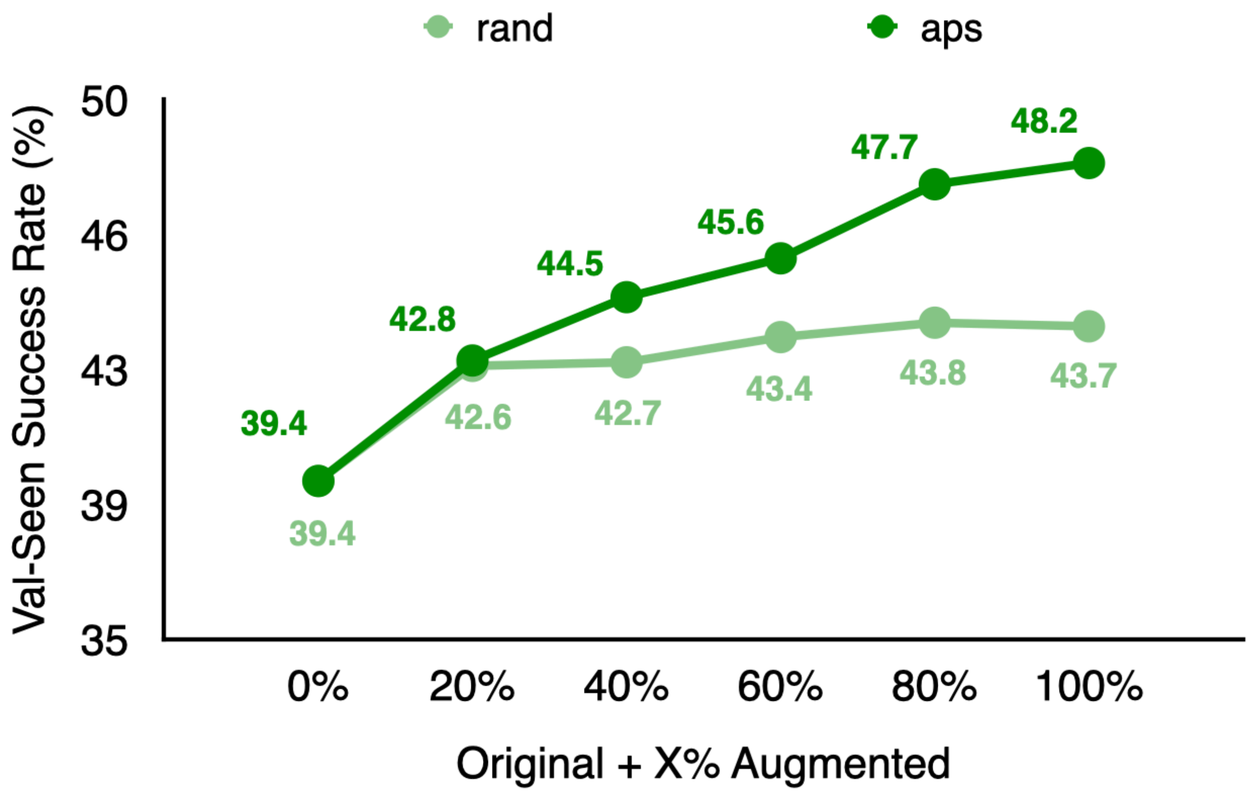


Adversarial Path Sampling (APS)

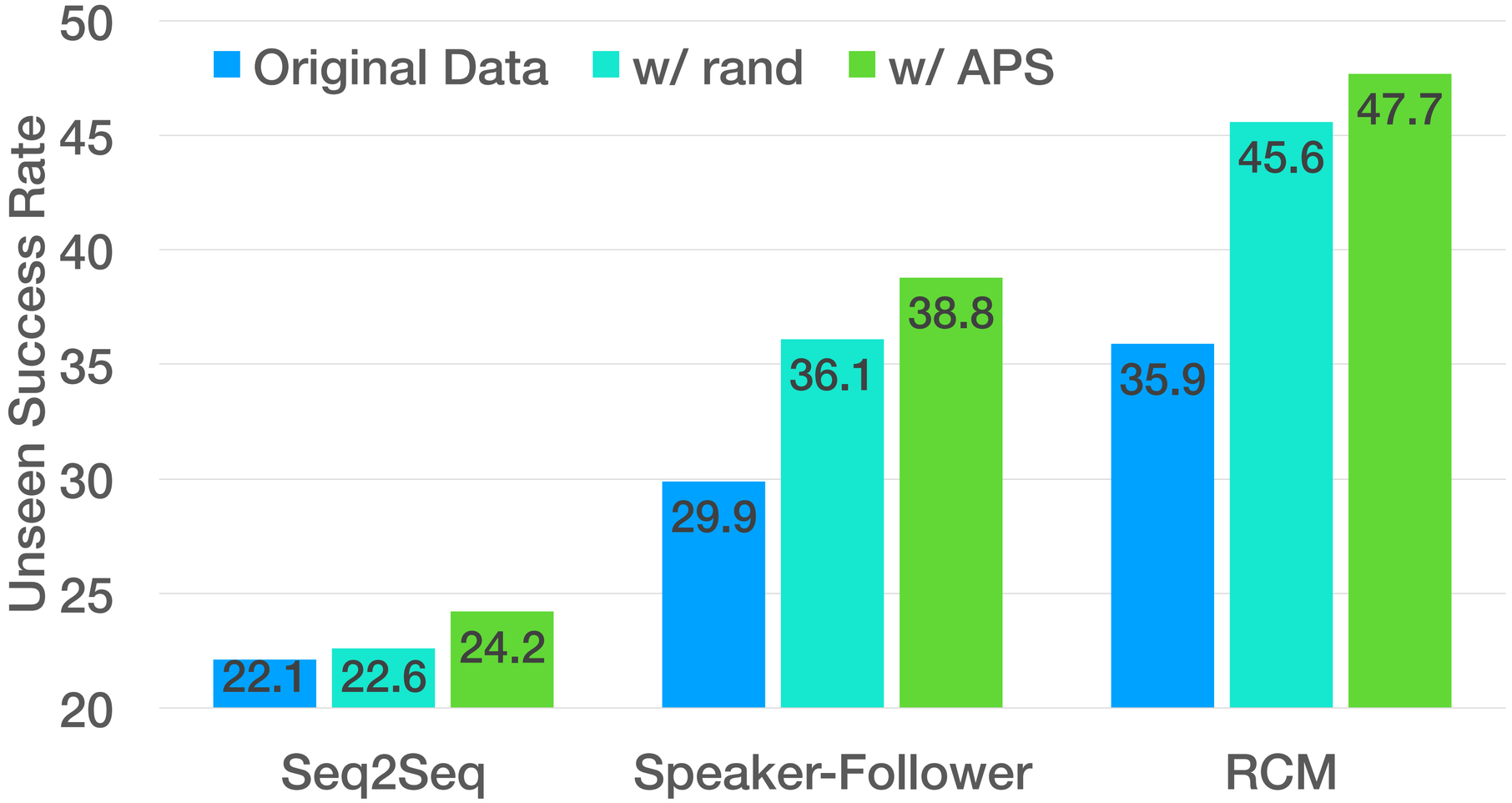
- **APS** learns to sample **increasingly challenging paths**, which the navigator (NAV) cannot navigate easily
- **NAV** is trying to accomplish augmented paths from APS and thus **optimized for a better navigation policy**



APS *vs.* Random Sampling

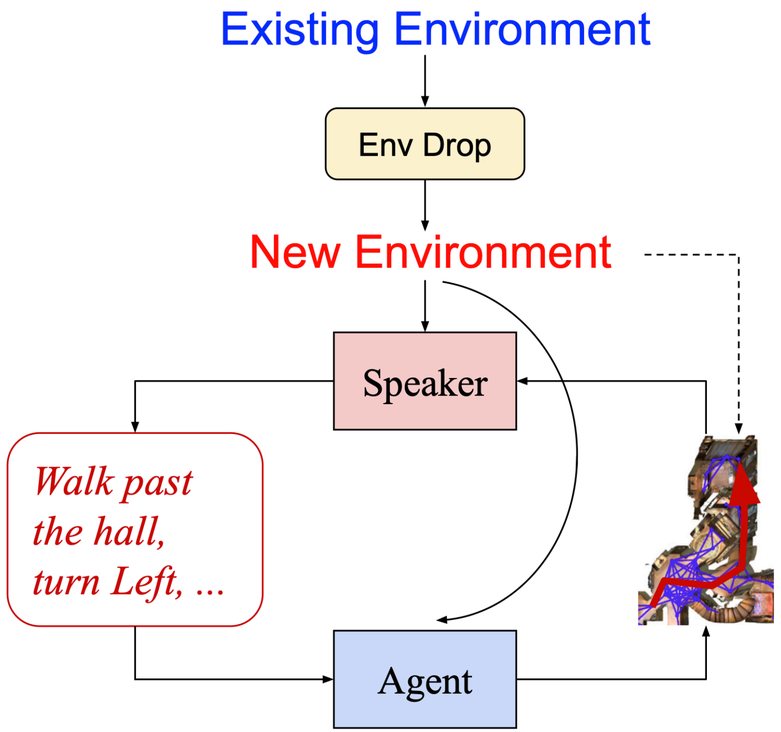


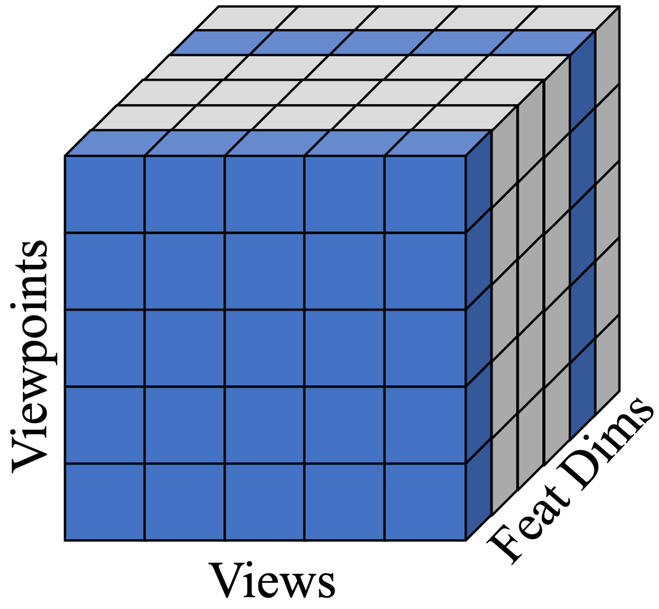
Navigation Results



Environment Dropout

- Augment “new” environments by modifying existing visual environments along the feature dimension





Environmental Dropout

Tan et al., “Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout”, NAACL 2019

Limitations of Speaker

- Pre-trained as a vision-to-language generation model
- Expected to generate instructions from scratch
- Particularly challenging in more complicated environments than indoor environments

Outdoor VLN Challenges

- More complicated and diverse visual perception
- Longer navigation paths
- More linguistically-complex instructions
- Limited number of human annotations

Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right. Touchdown is on the back of the dinosaur.

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

Multimodal Text Style Transfer for Outdoor Vision-and-Language Navigation

Wanrong Zhu, Xin Eric Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana,
Kazoo Sone, Sugato Basu, William Yang Wang, EACL 2021

EACL 2021

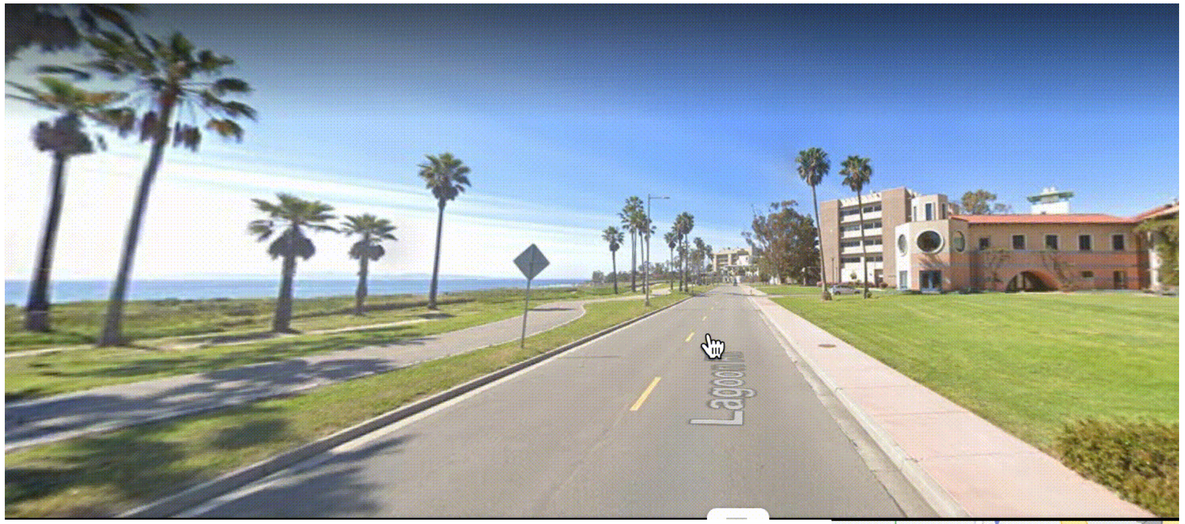
Synthesizing Instructions is Hard Outdoors

- **Red tokens:** contradiction with ground truth
- **Blue tokens:** alignment with ground truth

Ground Truth	**Head northwest** on W 35th St toward Hudson Blvd E. **Turn right** at the 1st cross street onto Hudson Blvd E.
Speaker	Turn so the red construction is on your left and the red brick building is on your right. Go forward to the intersection and turn right. You'll have a red brick building with a red awning on your right.

External Resource

- Google Street Views
- API-generated instructions (Google Map API)
- **Google Map instruction style is different from human description**



University of California, Santa Barbara
Santa Barbara, CA 93106

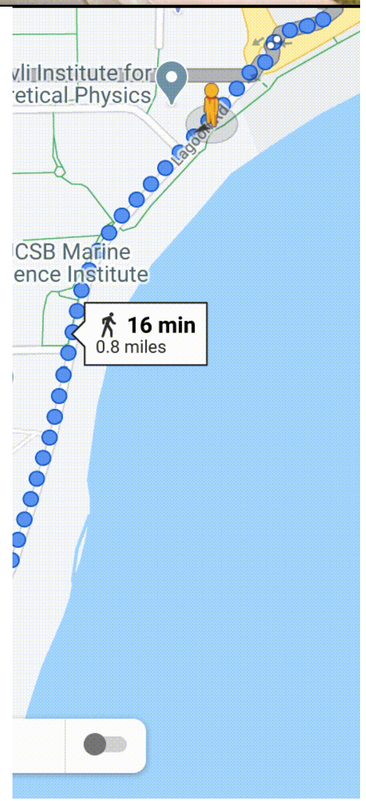
Head southwest on CA-217 W/Ward Memorial Blvd
95 ft

Slight left onto Ward Memorial Blvd
207 ft

At the traffic circle, take the 2nd exit onto Lagoon Rd
0.7 mi

Turn left
69 ft

Goleta Point
California 93106



Creating Pre-training Datasets for VLN

Visual Path

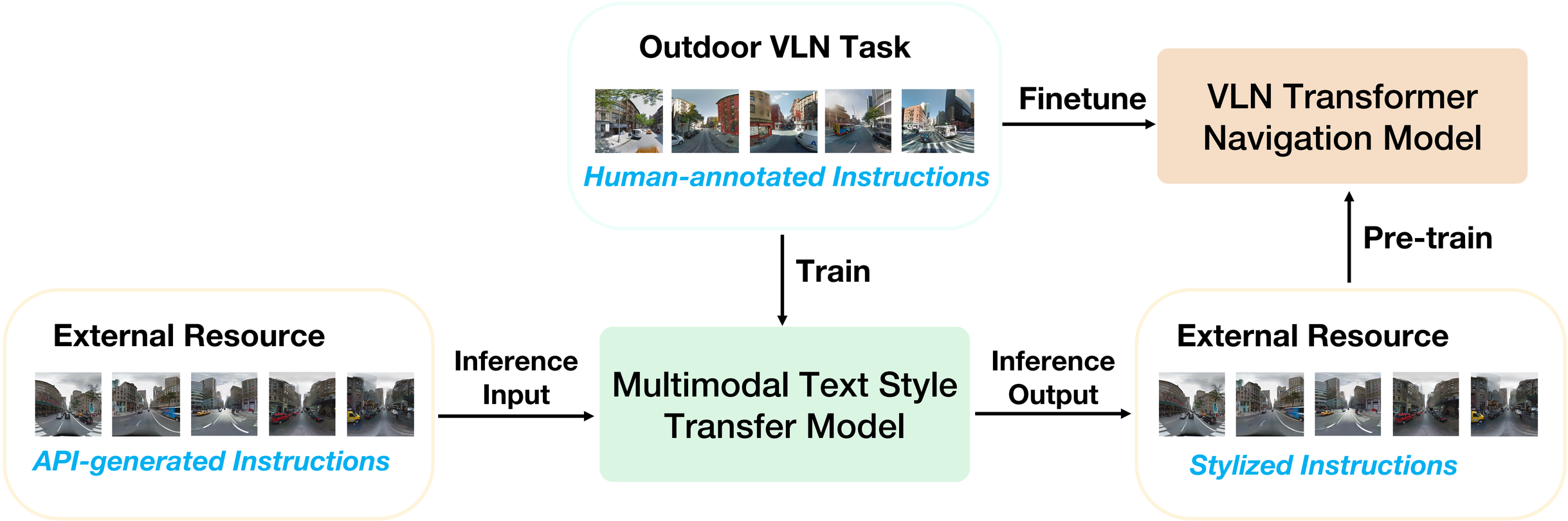
Head northwest on E 23rd St toward 2nd Ave. Turn left at the 2nd cross street onto 3rd Ave.

Orient yourself so you are facing the same as the traffic on the 4 lane road. Travel down this road until the first intersection. Turn left and go down this street with the flow of traffic. You'll see a black and white stripped awning on your right as you travel down the street.

Google Map API Instruction

Human Instruction

Multimodal Text Style Transfer



Objective: transfer to human-style instructions while keeping the guiding information provided by Google Map API.

Multimodal Text Style Transfer: Training

Human-annotated Instruction

Go straight. There will be a red wall to your right. Take a right. Stop at the intersection.

Masking

[MASK]. There will be a [MASK] to [MASK] right. Take a right. Stop at the [MASK].

Recovering

Multimodal Text Style Transfer Model

Go straight. There will be a red wall to your right. Take a right. Stop at the intersection.

Multimodal Text Style Transfer: Inference

API-generated Instruction

Head southwest on 5th Ave toward E 49th St. Turn right onto W 47th St.

Masking

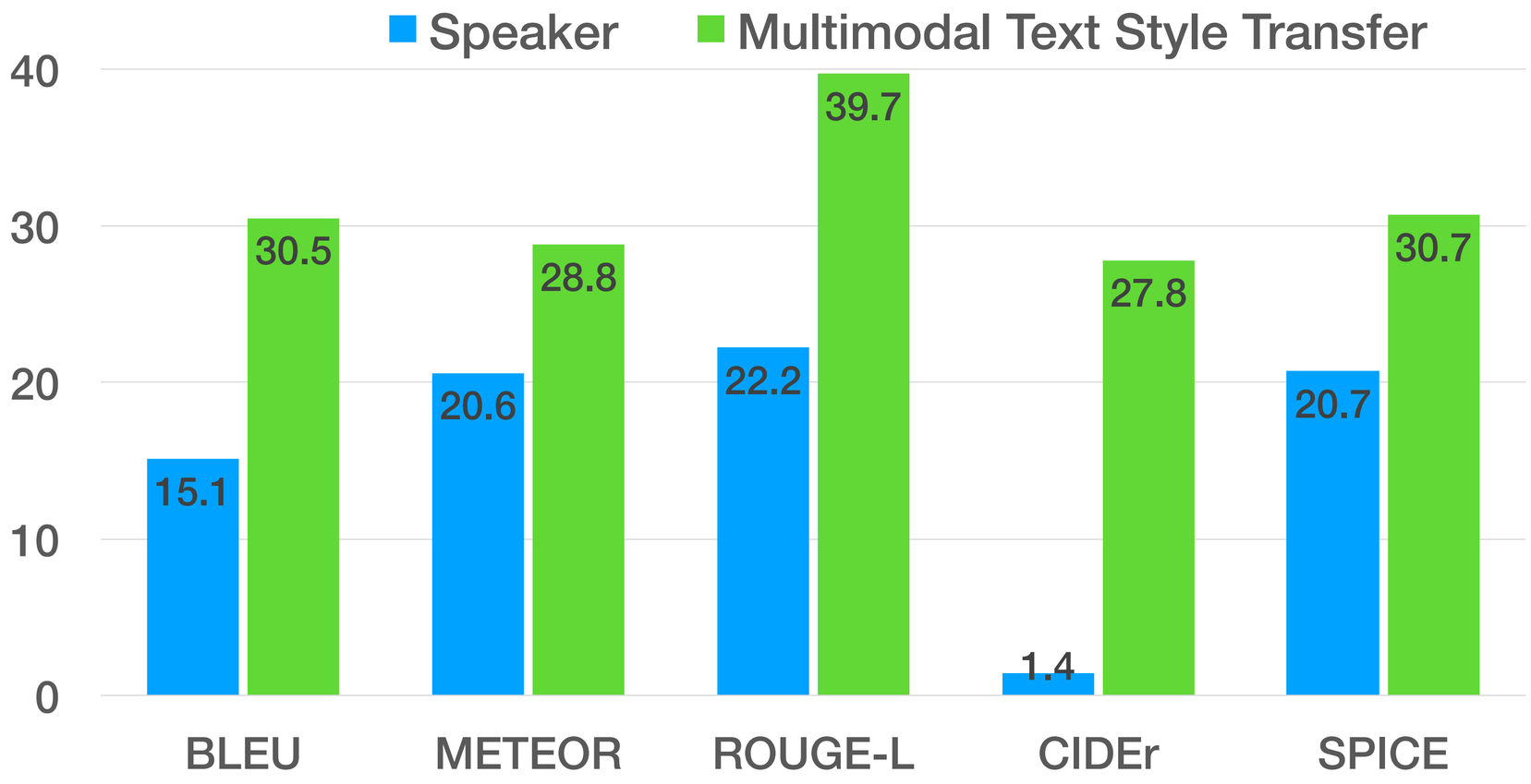
[MASK] on [MASK] toward [MASK]. [MASK] right onto [MASK].

Transferring Text Style

Multimodal Text Style Transfer Model

Head down the street with traffic on your right. Turn right onto the street.

Stylized Instruction Quality Is Much Higher



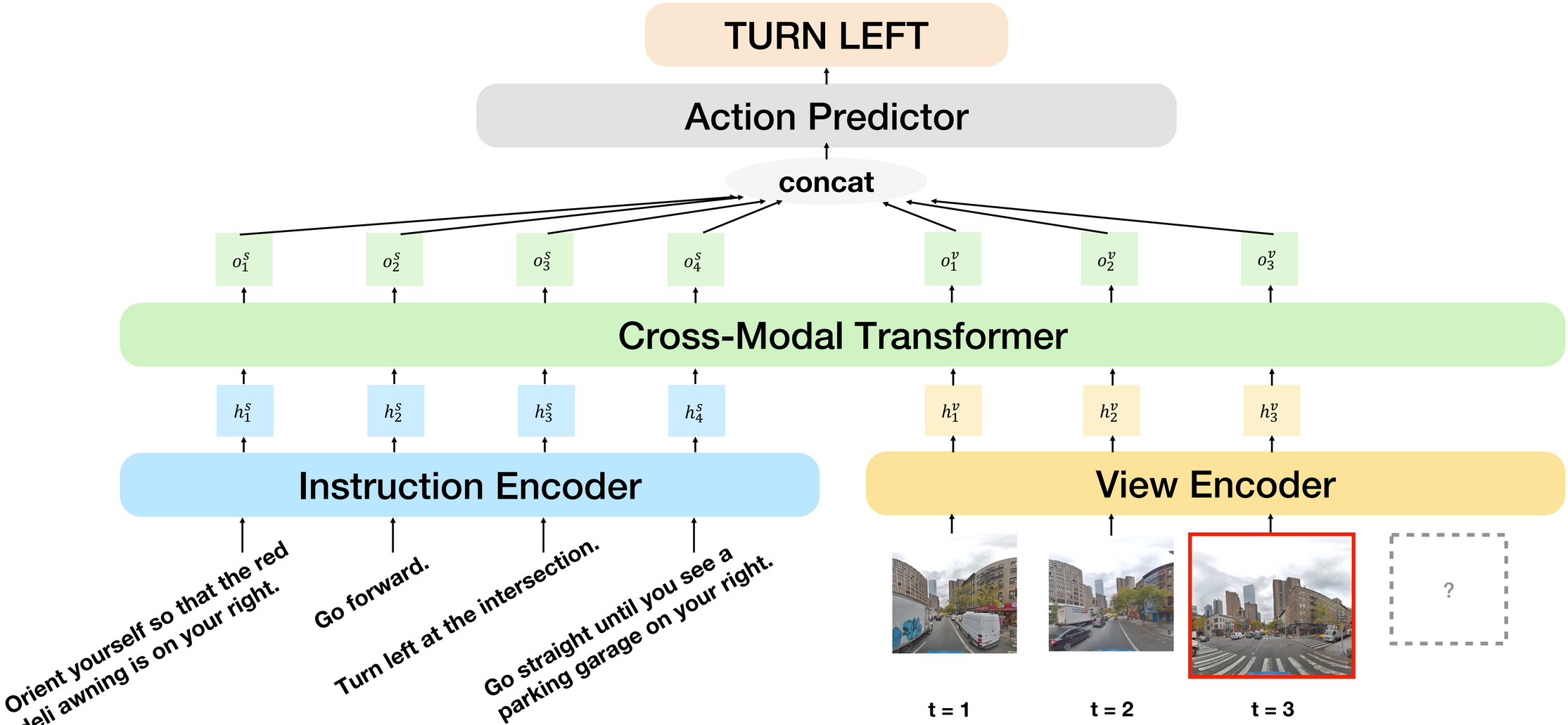
	#infill
Speaker	160
MTST	**308**

Stylized Instruction Quality Is Much Higher

- **Red tokens:** contradictions with ground truth
- **Blue tokens:** alignments with ground truth

Ground Truth	**Head northwest** on W 35th St toward Hudson Blvd E. **Turn right** at the 1st cross street onto Hudson Blvd E.
Speaker	Turn so the red construction is on your left and the red brick building is on your right. Go forward to the intersection and turn right. You'll have a red brick building with a red awning on your right.
Multimodal Text Style Transfer	Move forward with traffic on the right turn right at the light. Continue straight.

Navigation Model: VLN Transformer



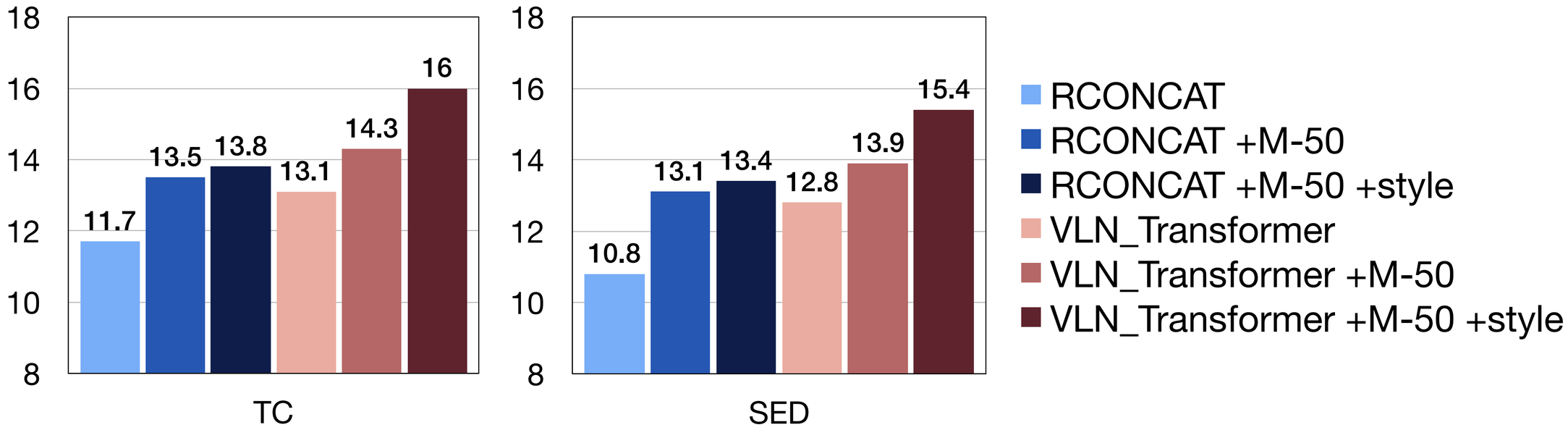
Datasets

- **Outdoor VLN Dataset:** Touchdown Dataset
- **External Resource:** StreetLearn Dataset
- **Stylized, Augmented Dataset**: **M-50 Dataset**

Dataset	Trajectory Source	Instruction Source
Touchdown	Google Street Views	Human Annotator
StreetLearn	Google Street Views	Google Map API

Navigation Results

- **Baseline model:** RCONCAT
- **+M-50:** pre-train on a StreetLearn subset with API-generated instructions
- **+M-50 +style:** pre-train on a StreetLearn subset with stylized instructions
- **TC:** task completion rate (= success rate)
- **SED:** success weighted by edit distance



Takeaways

- Data is the key to DL success, and DO NOT limit your models to existing labeled datasets
- Counterfactual reasoning is an essential ability for model generalization and enriches data augmentation approaches for VLN
- Pre-training can help, but be aware of distribution shift between pre-training and fine-tuning data
- Multimodal style transfer enables effective utilization of (potentially all) available data on Google Maps

Generalizable VLN Methods

1. Data augmentation
2. **Evaluation of generated navigation instructions**
3. Multitask learning
4. Unseen environment adaptation

Generalizable VLN Methods

1. Data augmentation
2. Evaluation of generated navigation instructions
3. **Multitask learning**
4. Unseen environment adaptation

Environment-agnostic Multitask Learning for Natural Language Grounded Navigation

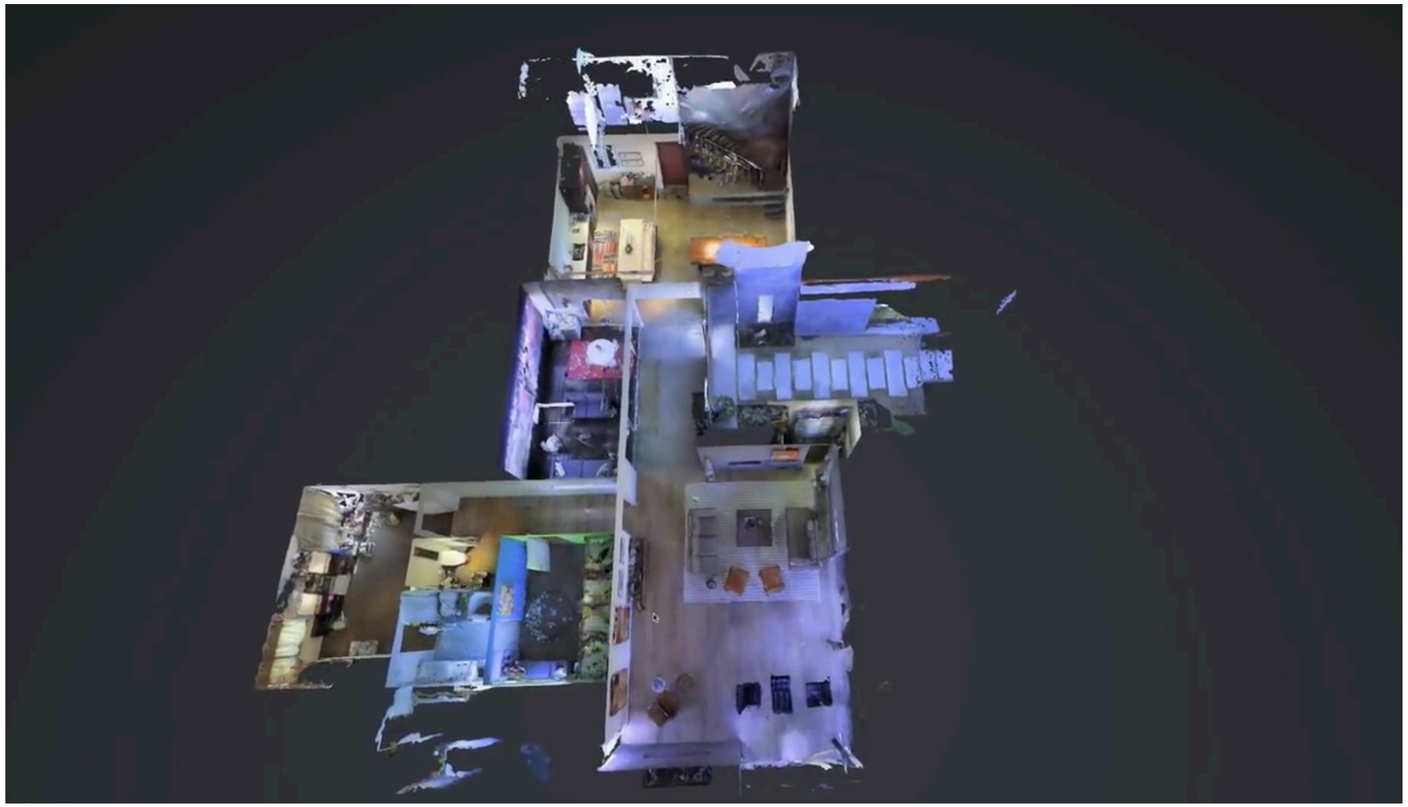
Xin Eric Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi, ECCV 2020

Vision-and-Language Navigation (VLN)

Input:

- **First-person view visual scene** (no GPS, no map)
- **Language instruction:**

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.

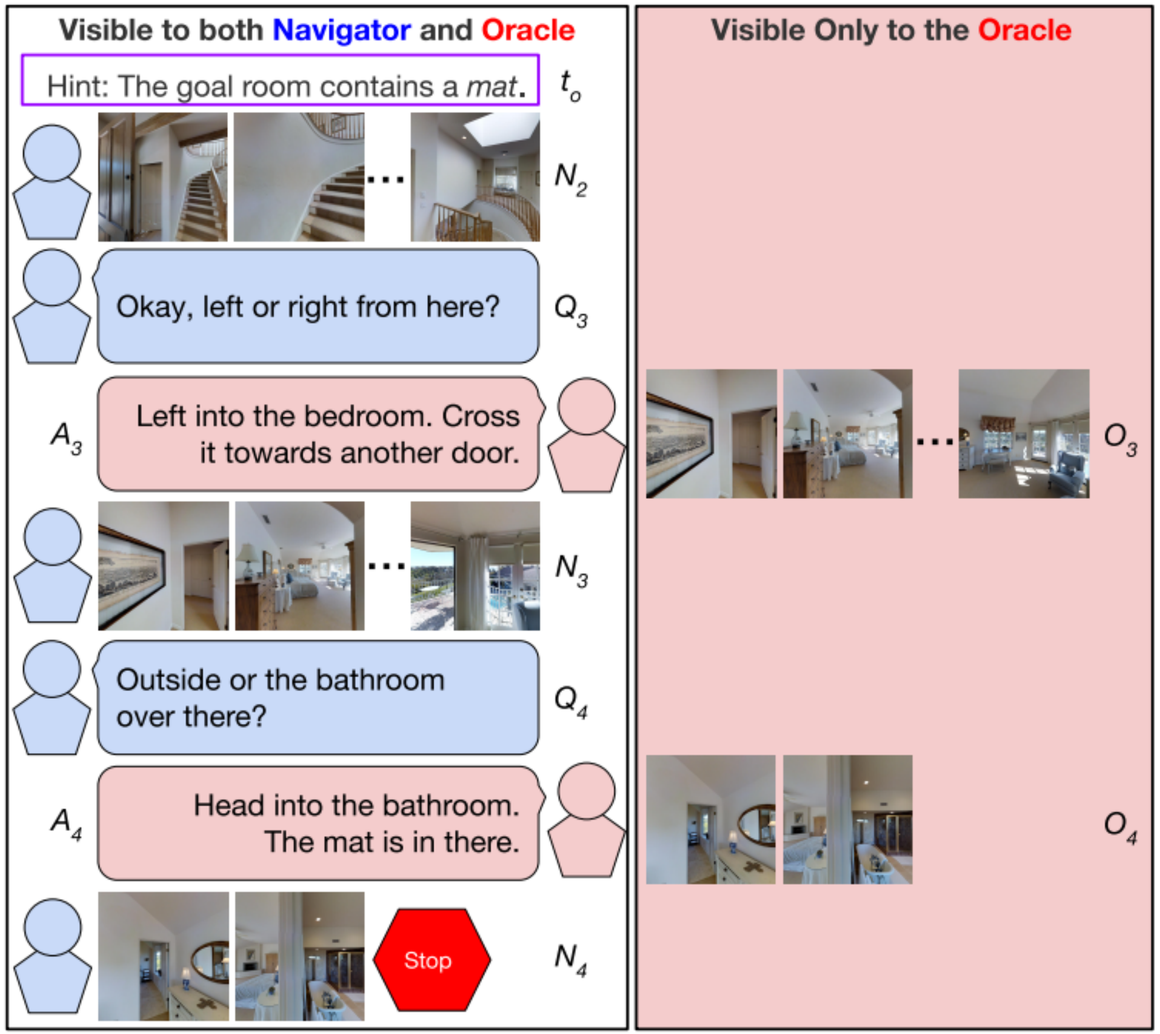


Output:

Actions: *Turn left, turn right, look up, look down, go forward, stop*

Anderson et al., “Vision-and-Language Navigation”, CVPR 2018

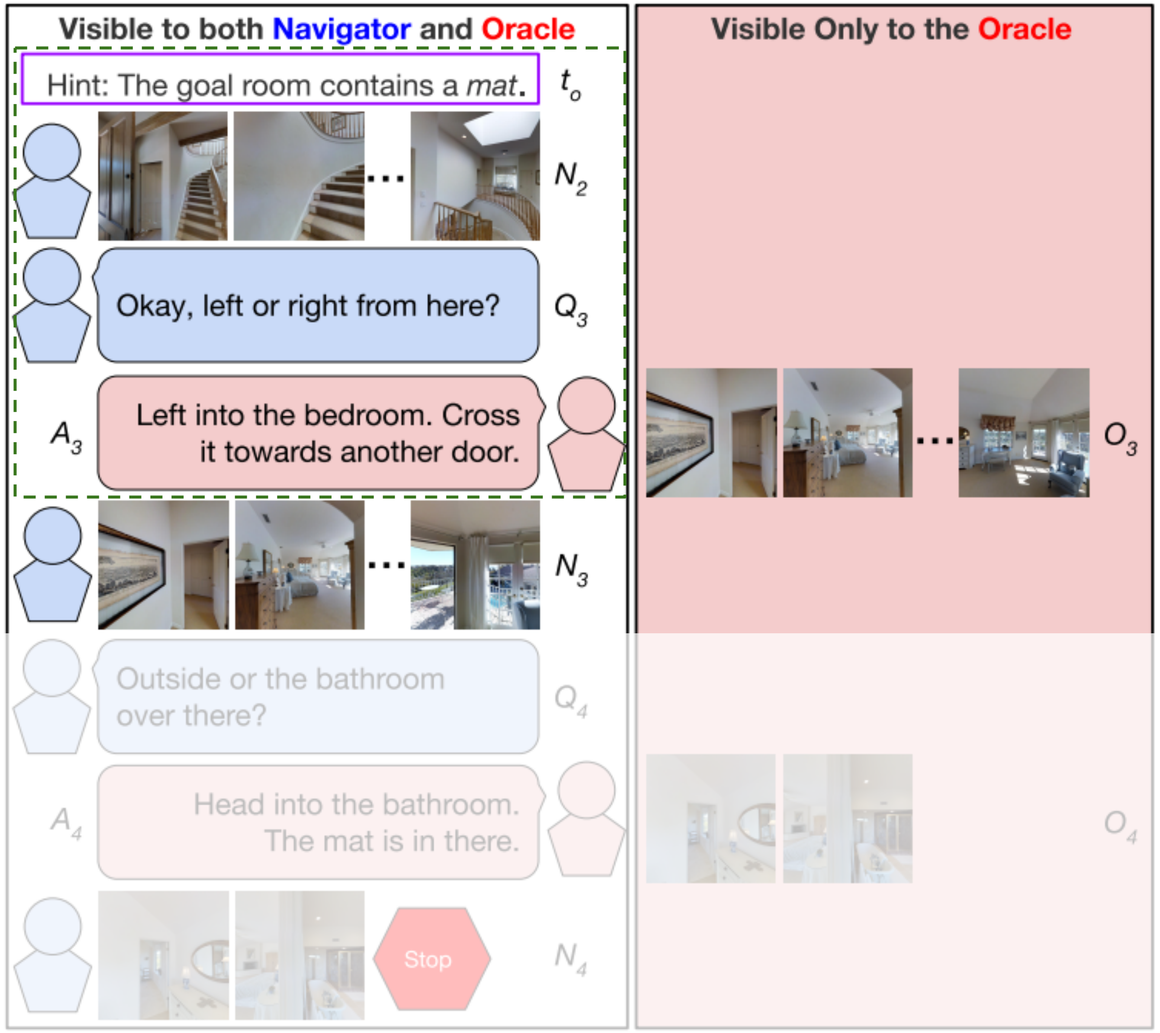
Cooperative Vision-and-Dialog Navigation (CVDN)



- 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., “Vision-and-Dialog Navigation”, CoRL 2019

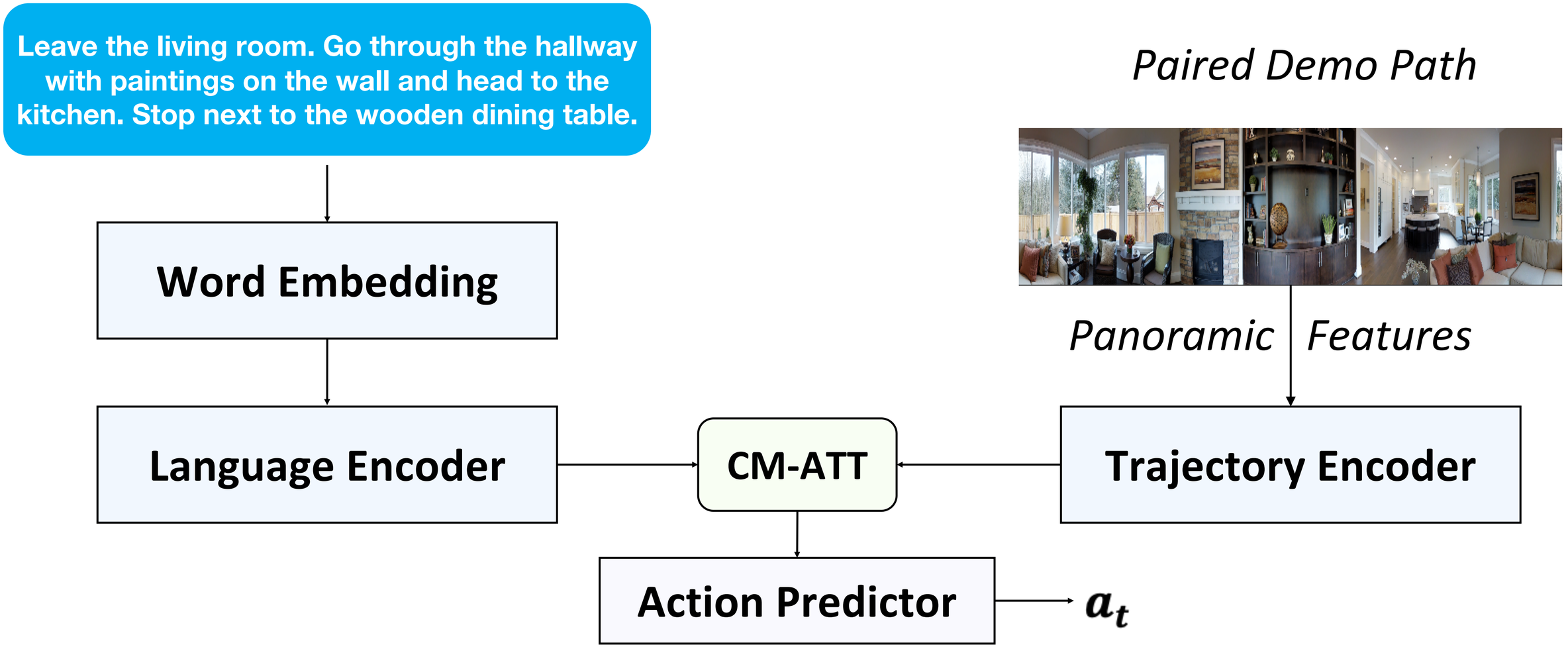
Sub-task: Navigation from Dialog History (NDH)



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

Thomason et al., “Vision-and-Dialog Navigation”, CoRL 2019

A Strong Baseline for VLN: RCM



Wang et al., “Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation”, CVPR 2019 Best Student Paper

Multitask RCM

VLN Instruction
or
NDH Dialog

Interleaved Multitask Data Sampling

Paired Demo Path

Panoramic | *Features*

Word Embedding

Language Encoder

CM-ATT

Trajectory Encoder

Action Predictor

a_t

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

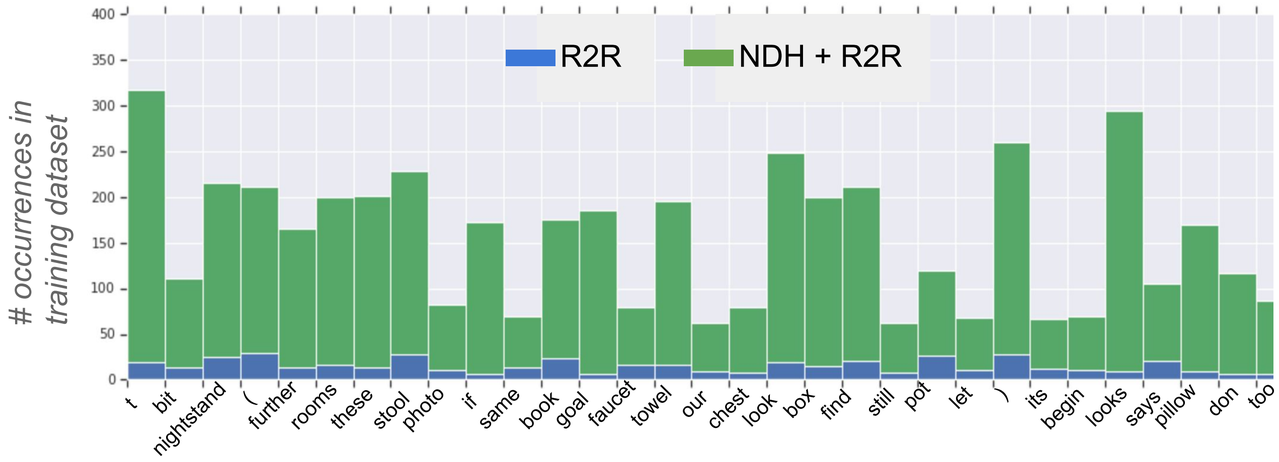
Fold	Model	NDH Evaluation: Inputs for NDH (t_o A_i Q_i $A_{1:i-1}; Q_{1:i-1}$)	NDH Evaluation: Progress ↑	VLN Evaluation: PL	NE ↓	SR ↑	SPL ↑	CLS ↑
Val Seen	NDH-RCM	✓	**6.97**					
		✓ ✓	**6.92**					
		✓ ✓ ✓	**6.47**					
		✓ ✓ ✓ ✓	**6.49**					
	VLN-RCM			10.75	5.09	52.39	48.86	63.91
	MT-RCM	✓	3.00	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

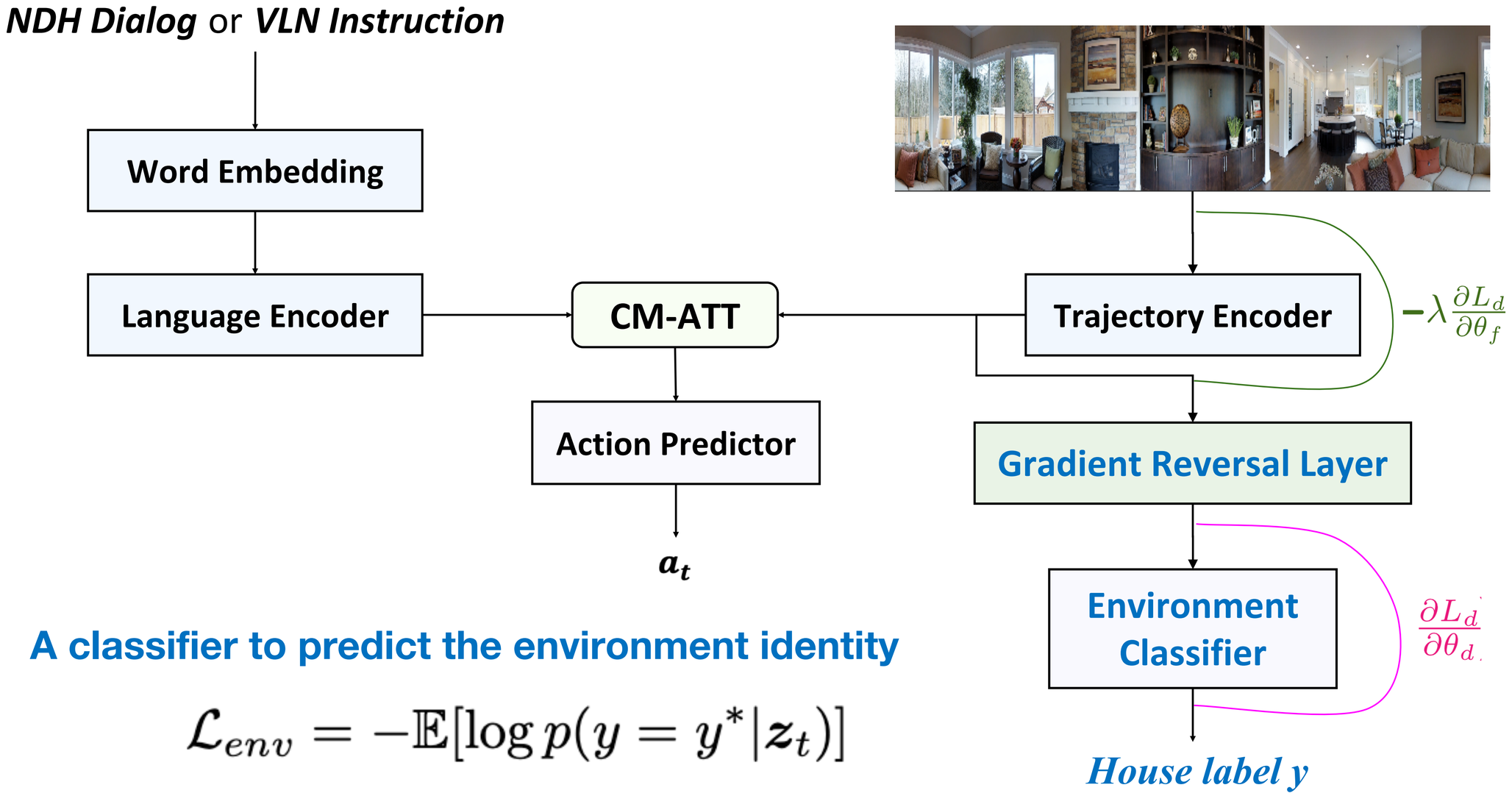
Multitask learning benefits from

- More appearances of under-represented 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

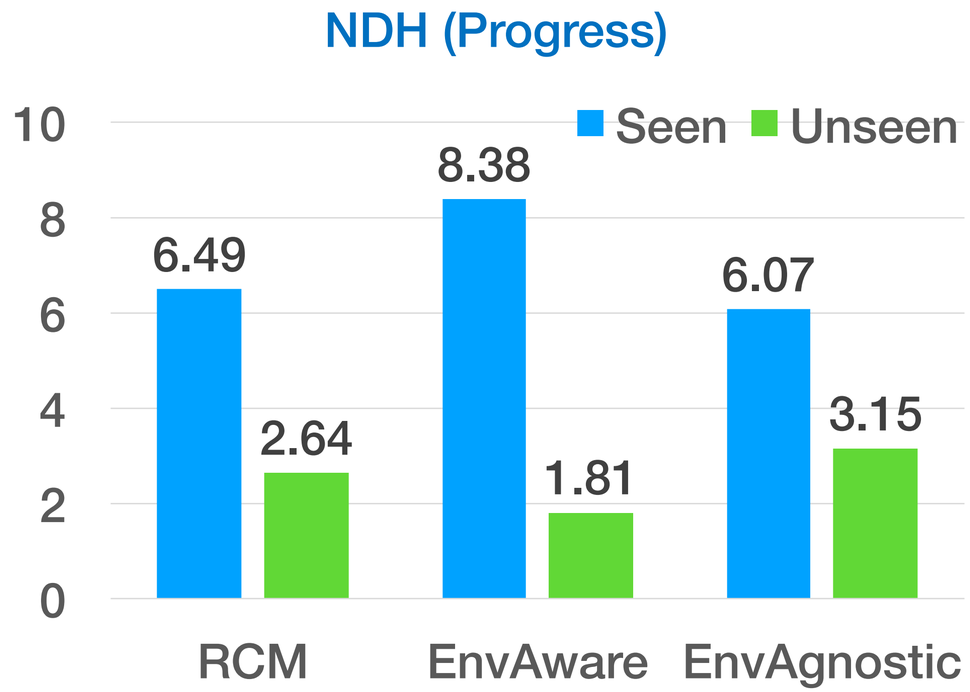
Environment-agnostic Representation Learning

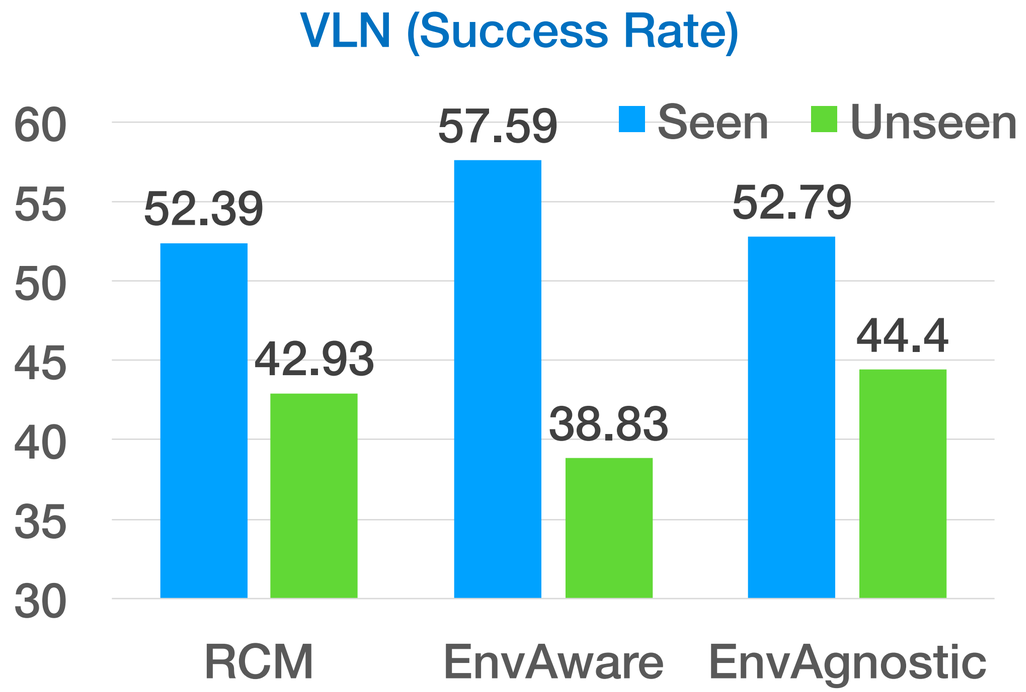


- **A classifier to predict the environment identity**

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

Environment-Aware *vs.* Environment-Agnostic

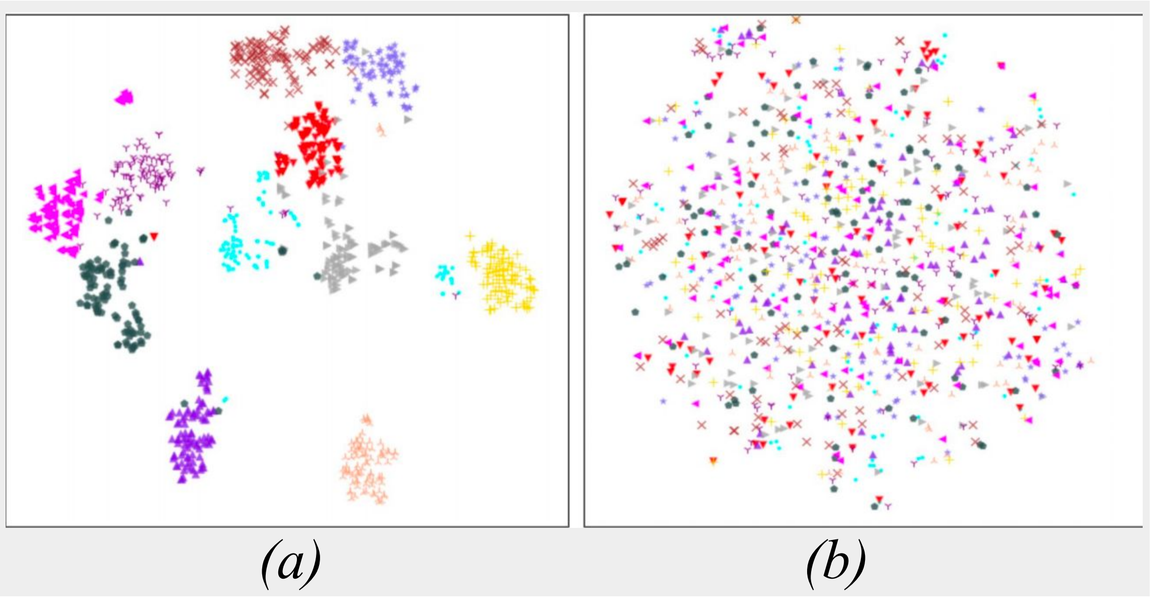




- Env-aware learning tends to overfit seen environments
- Env-agnostic learning generalizes better on unseen environments
- (Potential) Meta-learning with env-aware & env-agnostic may benefit from both worlds

Environment-Aware *vs.* Environment-Agnostic

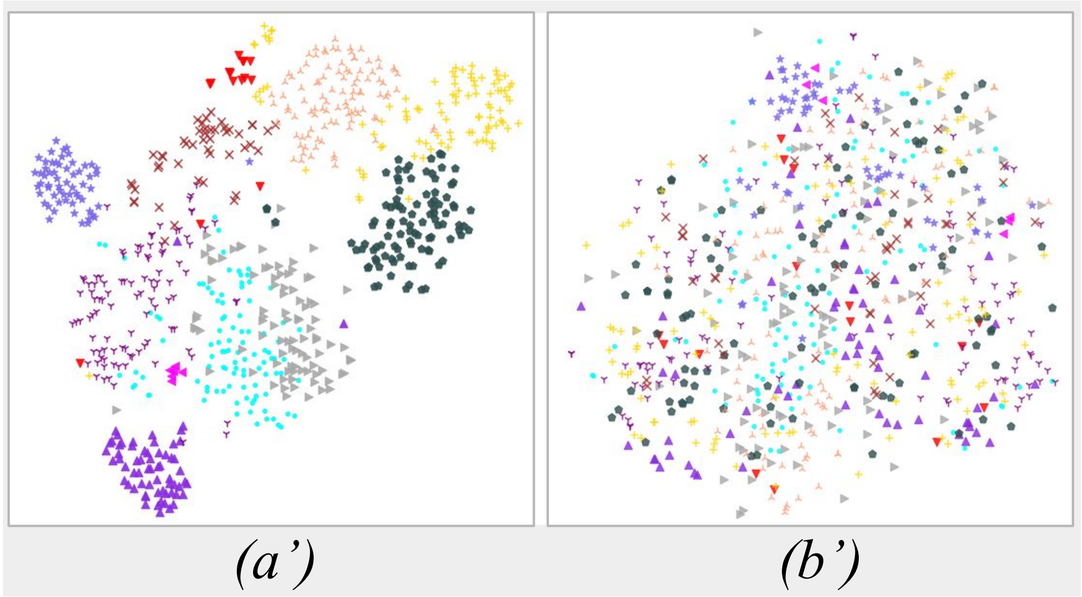
Seen



(a) *(b)*

EnvAware **EnvAgnostic**

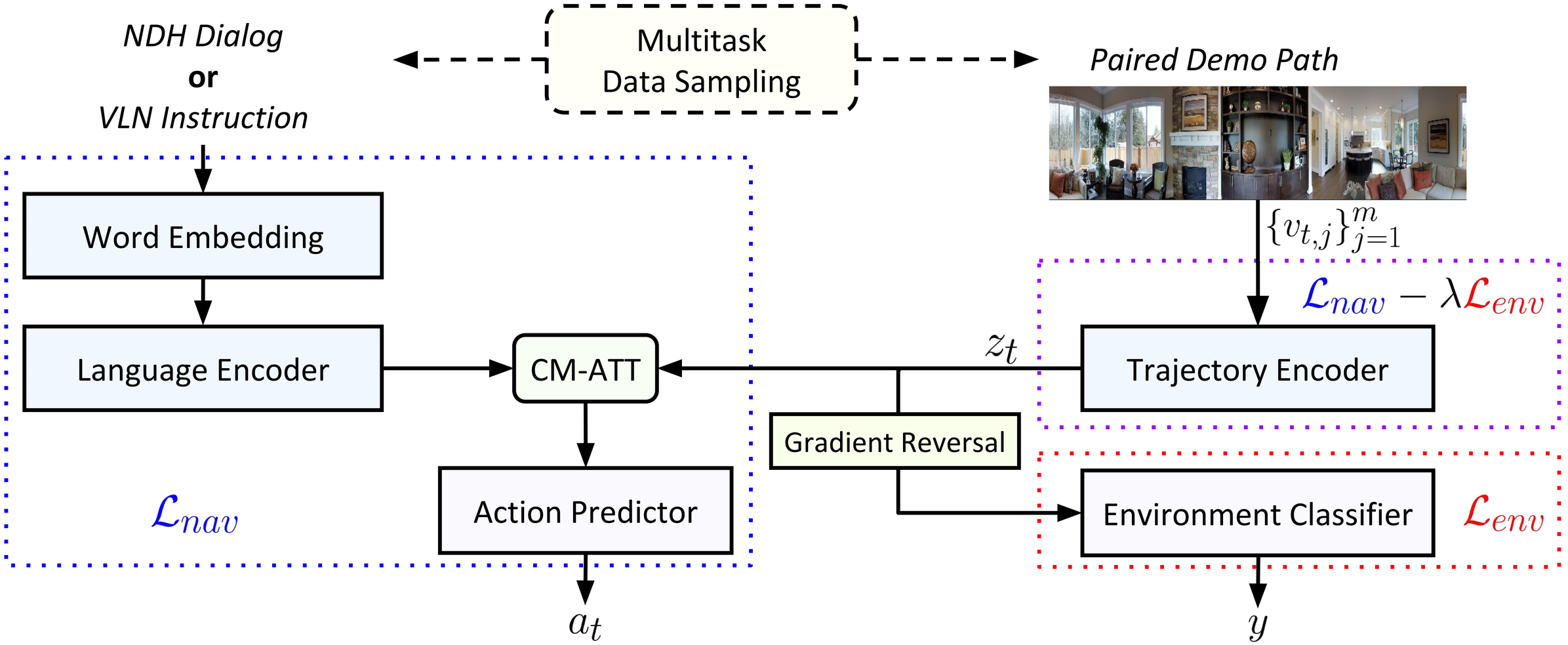
Unseen



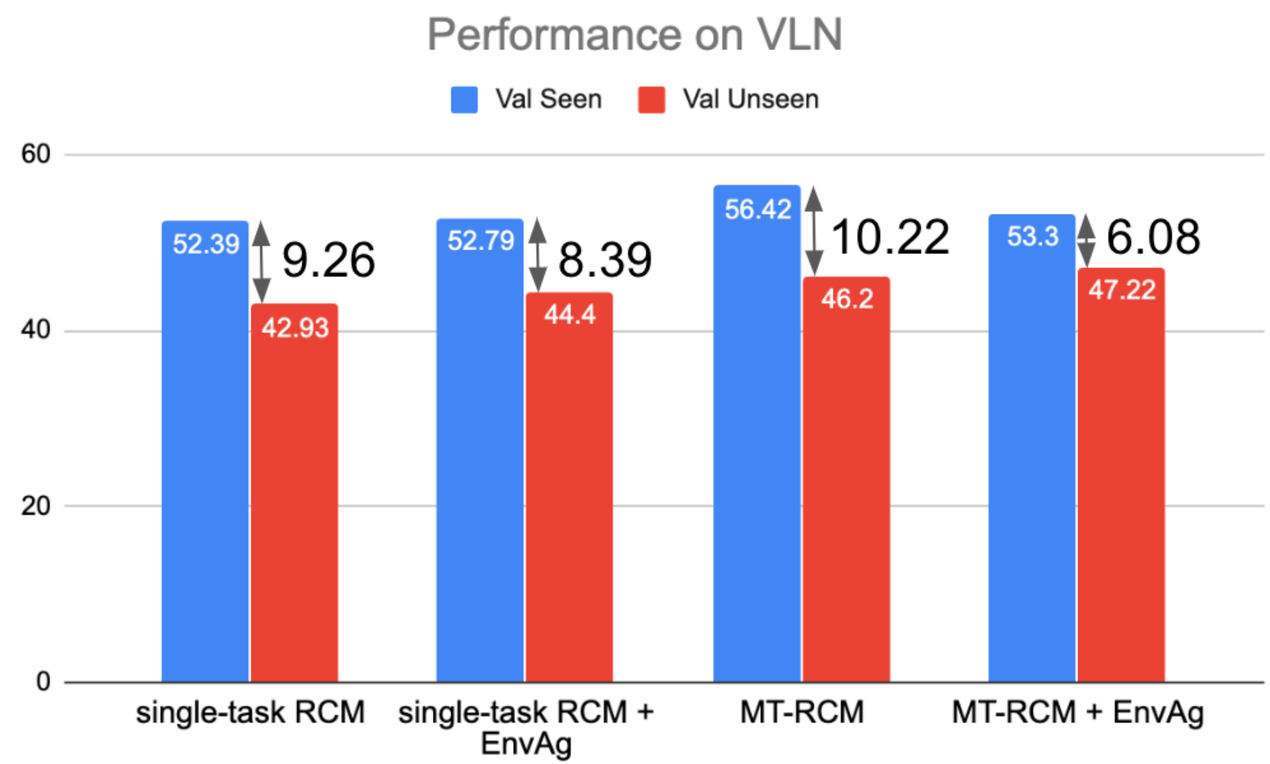
(a') *(b')*

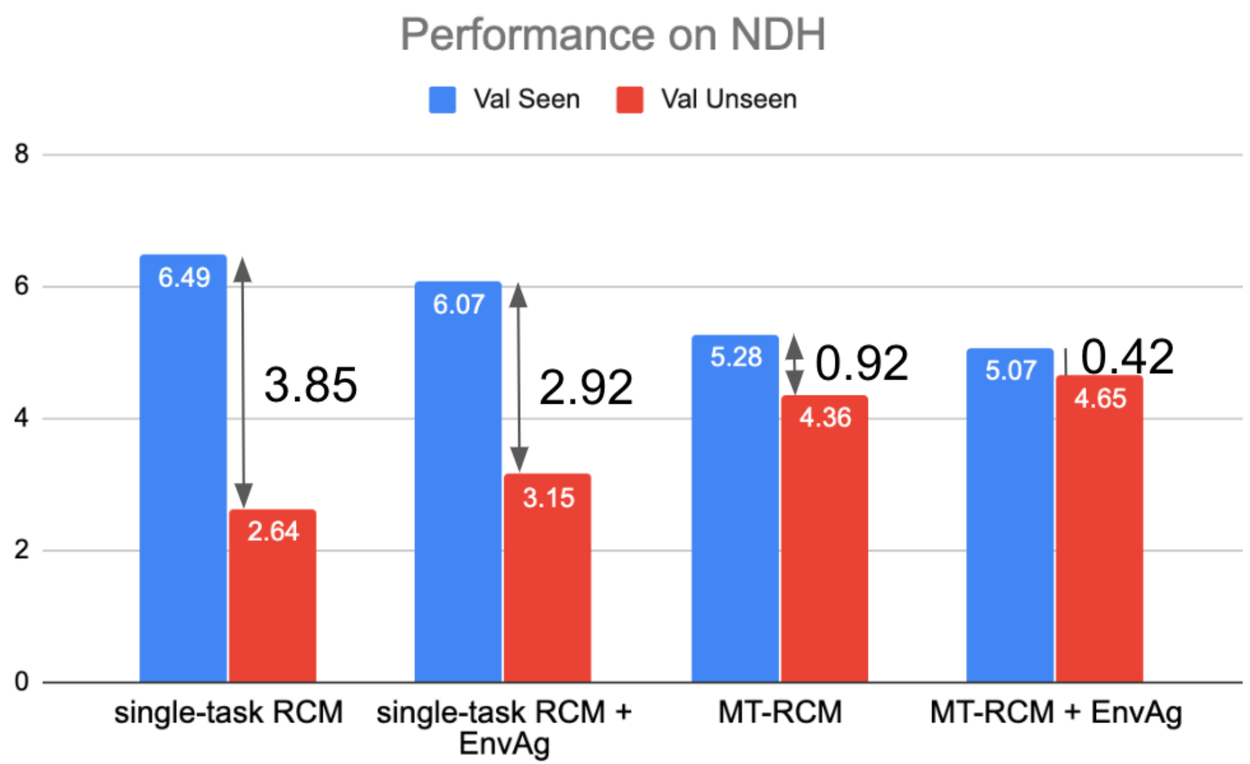
EnvAware **EnvAgnostic**

Environment-agnostic Multitask Learning



Results


Performance on VLN
Val Seen
Val Unseen
60
40
20
0
52.39
9.26
42.93
52.79
8.39
44.4
56.42
10.22
46.2
53.3
6.08
47.22
single-task RCM
single-task RCM + EnvAg
MT-RCM
MT-RCM + EnvAg


Performance on NDH
Val Seen
Val Unseen
8
6
4
2
0
6.49
3.85
2.64
6.07
2.92
3.15
5.28
0.92
4.36
5.07
0.42
4.65
single-task RCM
single-task RCM + EnvAg
MT-RCM
MT-RCM + EnvAg

Takeaways

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

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

Generalizable VLN Methods

1. Data augmentation
2. Evaluation of generated navigation instructions
3. Multitask learning
4. **Unseen environment adaptation**

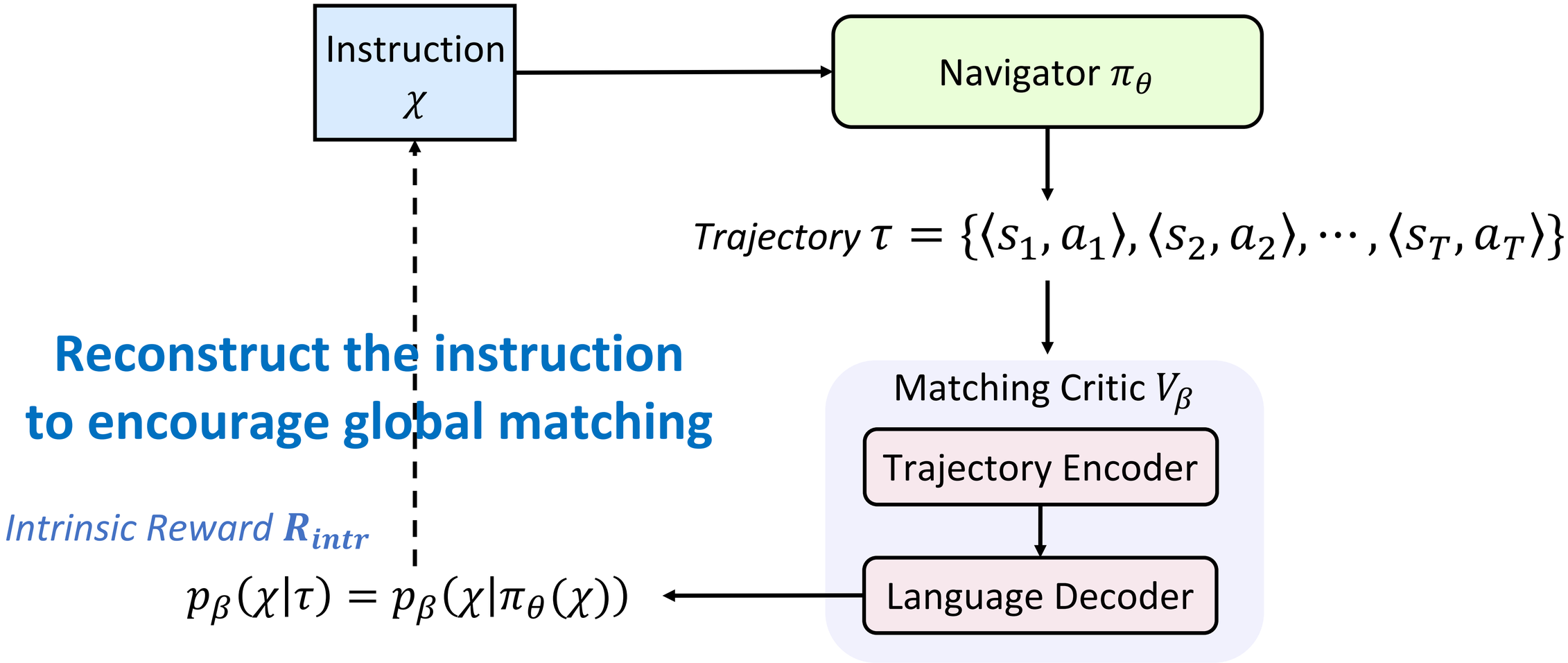
Self-supervised Imitation Learning (SIL)

Unseen Environment

- Explore unseen environments with self-supervision
- Adapt to new environments

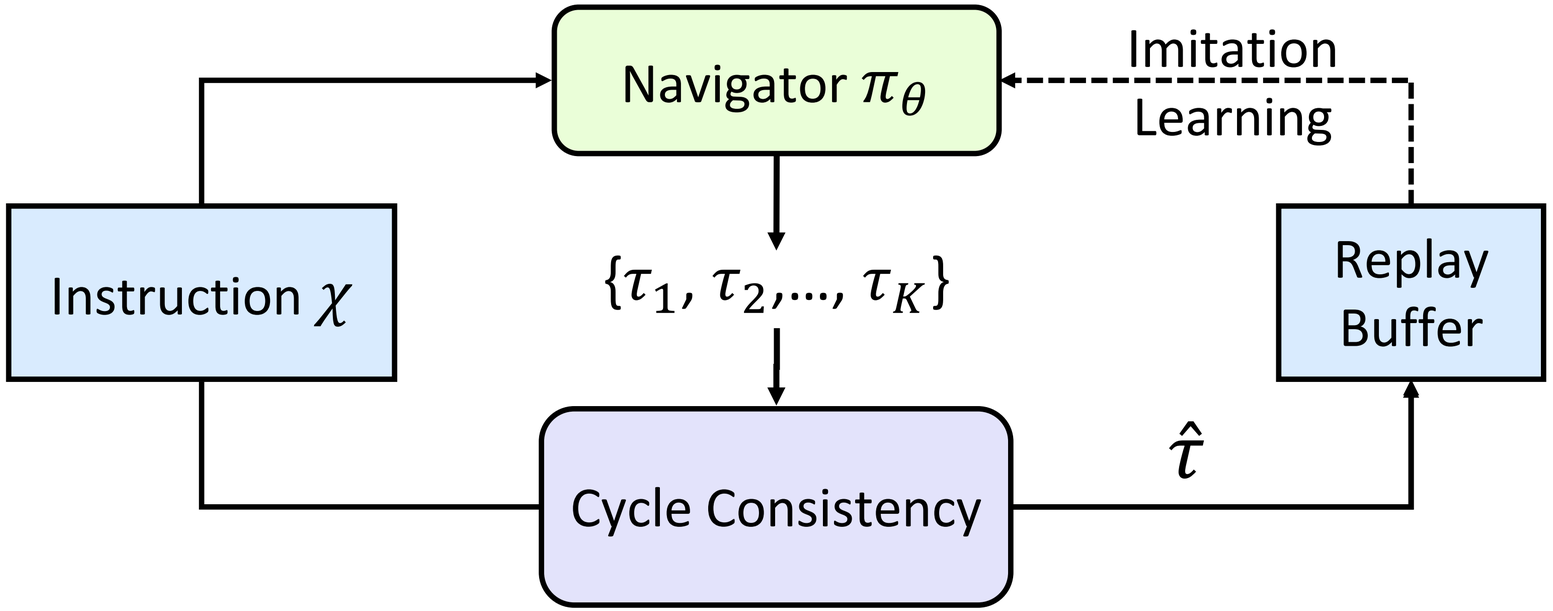
Wang et al., “Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation”, CVPR 2019 Best Student Paper

Cycle Consistency → Self-supervision



SIL: Adapt to New Environment

Unseen Environment



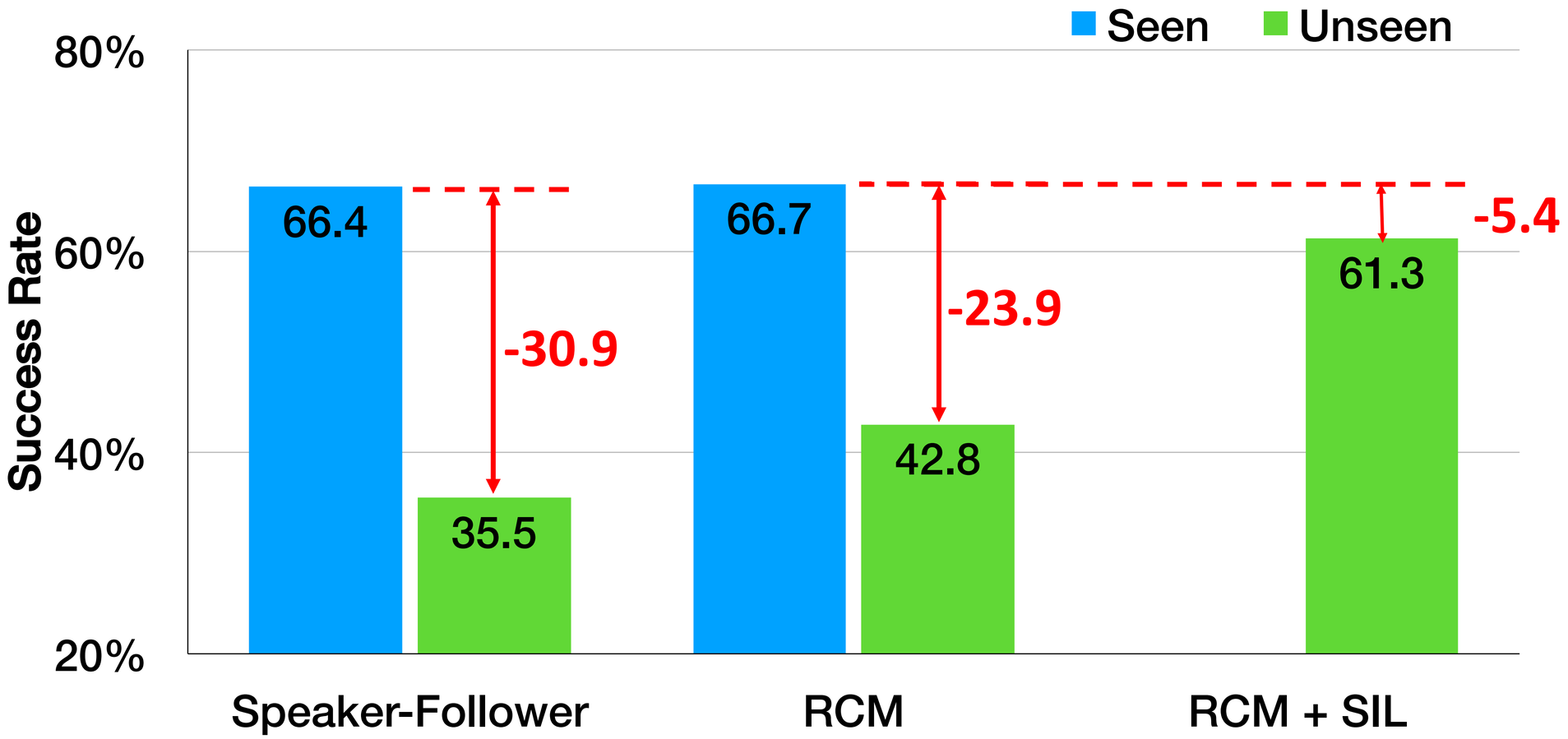
- PG with Off-policy Monto-Carlo return == supervised learning with $\hat{\tau}$ as the "GT"
- Learning from its previous good behaviors → better policy that adapts to new environments

Instruction: Head towards the kitchen. ... Walk forward and stop beside the bottom of the steps *facing the double white doors*.

Before SIL

After SIL

Results: Seen-Unseen Gap Narrowed



Efficiency: average path length (14.84 meters → 9.12 meters)

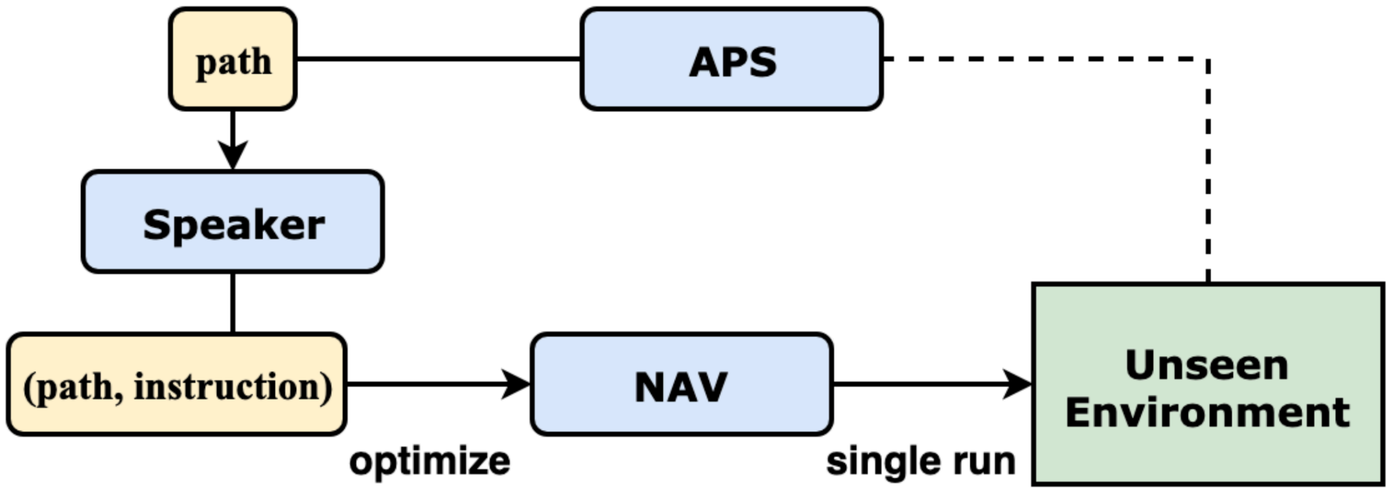
Pre-exploration with EnvDrop

- No access to GT instructions
- Augment instructions for shortest paths in unseen environments
- However, shortest paths in unseen environments are not available in the real world

Tan et al., “Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout”, NAACL 2019

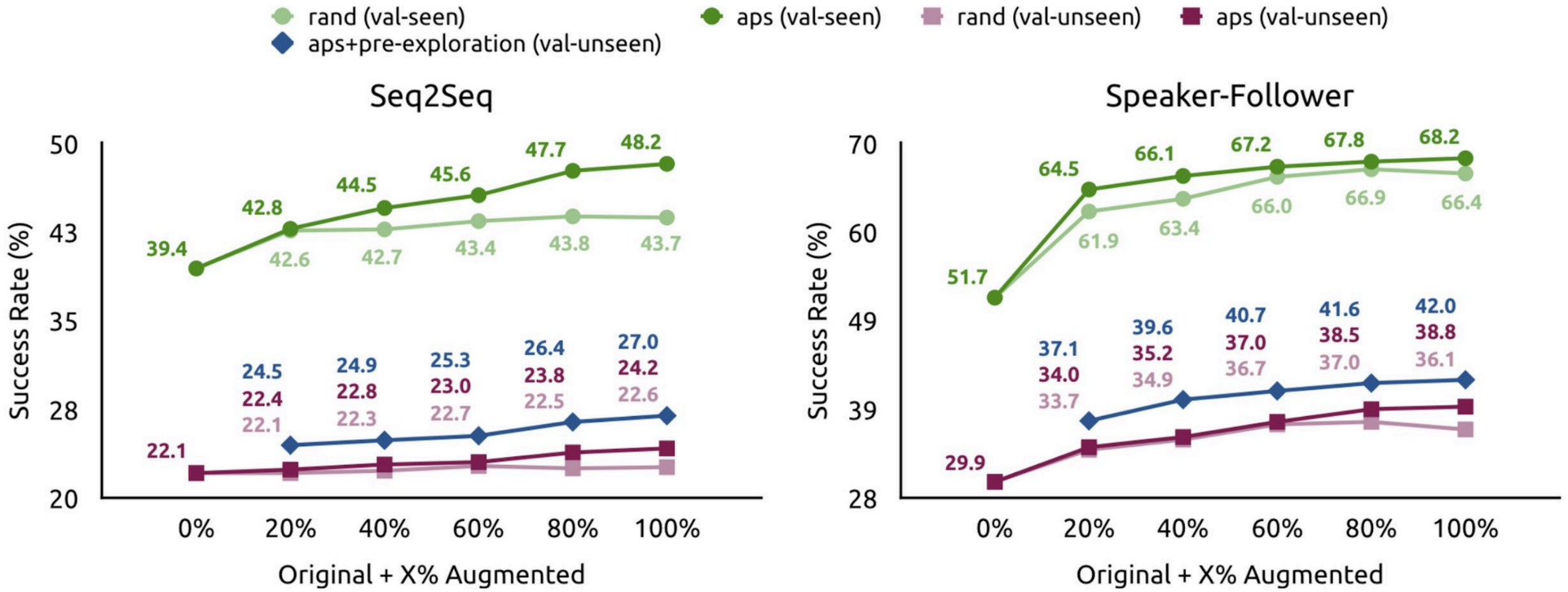
Environment-based Pre-exploration with APS

- No access to GT instructions (SIL) or shortest paths (EnvDrop)
- Use APS to sample paths while exploring and adapt NAV to a certain unseen environment
- After pre-exploration, NAV will execute the human instruction within this environment in a single run



Fu et al., “Counterfactual Vision-and-Language Navigation via Adversarial Path Sampling”, ECCV 2020

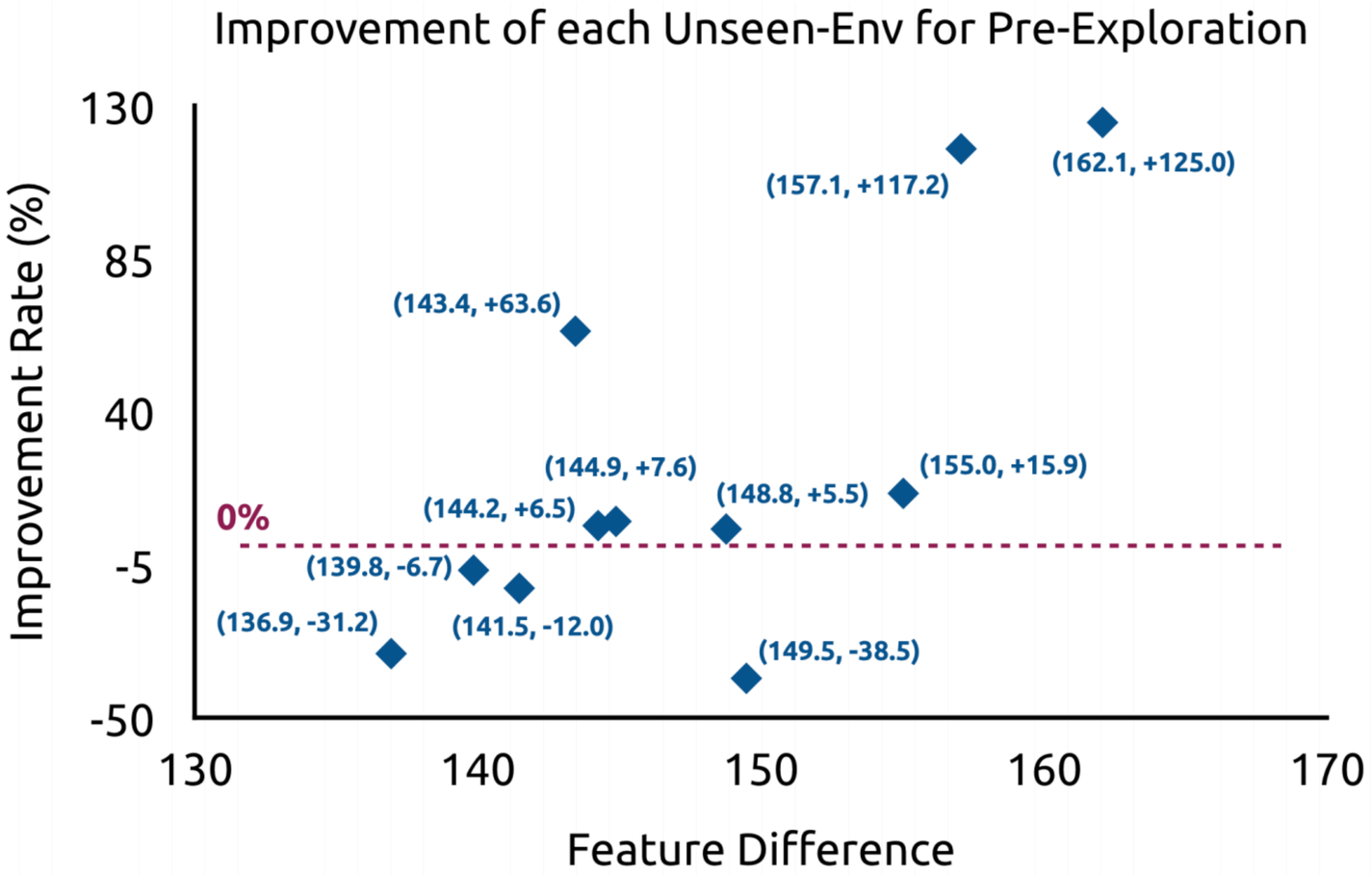
Results



- APS sampled paths help on both seen and unseen environments
- **Pre-exploration with APS** further improves navigation on unseen environments

Pre-exploration in Individual Environments

- **X: Feature difference** between training environments and a particular unseen environment
- **Y: Improvement rate** after pre-exploring the unseen environment
- **More different** the unseen environment is, **more effective** the pre-exploration is



Takeaways

- Pre-exploring unseen environments with self-supervision greatly alleviate the poor generalization issue
- Environment-based pre-exploration is more practical, where no GT instructions or meta-information of the environments are used
- Future work towards more efficient adaptation

Thank you!

References

- Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments, Anderson et al. CVPR 2018
- Vision-and-Dialog Navigation, Thomason et al. CoRL 2019
- Speaker-Follower Model for Vision-and-Language Navigation, Fried et al. NeurIPS 2018
- Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation, Wang et al. CVPR 2019
- Learning to Navigate Unseen Environment: Back Translation with Environment Dropout, Tan and Bansal, NAACL 2019
- Counterfactual Vision-and-Language Navigation via Adversarial Path Sampling, Fu et al. ECCV 2020
- Environment-agnostic Multitask Learning for Natural Language Grounded Navigation, Wang et al. ECCV 2020
- Multimodal Text Style Transfer for Outdoor Vision-and-Language Navigation, Zhu et al. EACL 2021