Vision Language Navigation

Weiwen Chen 2019.12.29

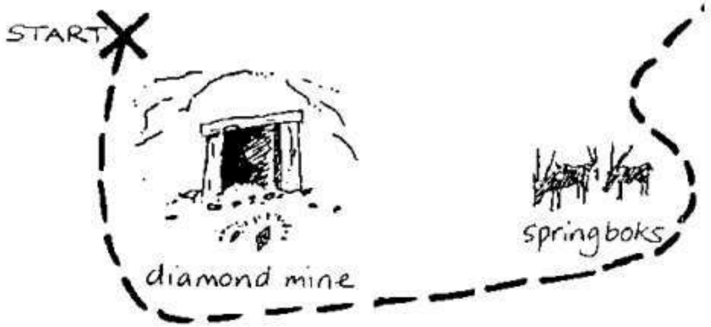
Outline

- Short history of VLN(from sim & data)
- Problem Definition
- Methods
  - Baseline: seq-to-seq
  - Look Before You Leap (model-based & model-free)
  - Speaker-Follower (Data Augmentation, Action Space)
  - Reinforced Cross-Modal Matching (Align matching, SIL)
  - Self-Monitoring (Align progress)
  - Environmental Dropout (Data Augmentation, IL)
  - Auxiliary Reasoning Tasks :)

History of VLN

Simulator & data

Follow Navigational Directions



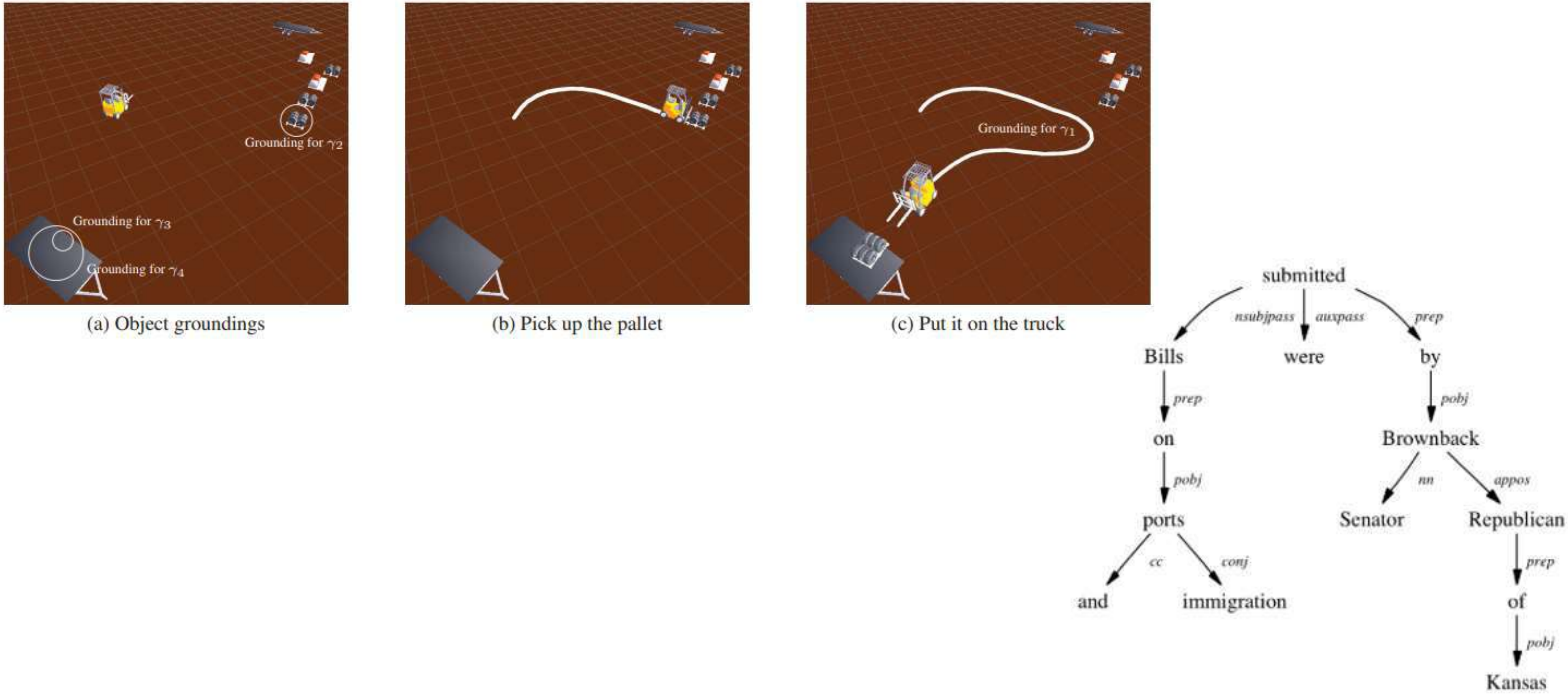
1. go vertically down until you're underneath eh diamond mine
2. then eh go right until you're
3. you're between springbok and highest viewpoint

[redacted]
[redacted]

Figure 2: The instruction giver and instruction follower face each other, and cannot see each others maps.

Vogel, et al. Learning to Follow Navigational Directions. ACL2010

Mobile Manipulation



(a) Object groundings

(b) Pick up the pallet

(c) Put it on the truck

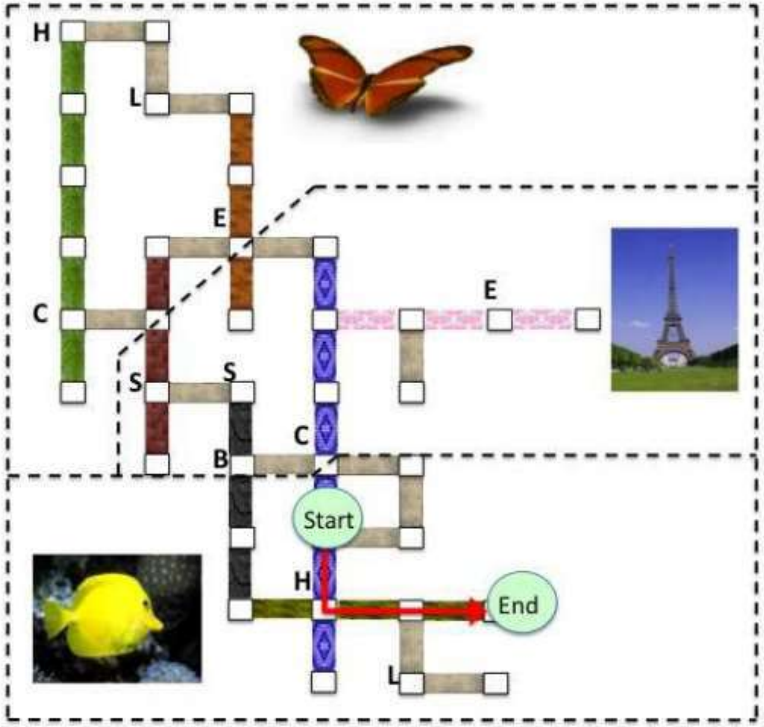
Stefanie Tellex, et al. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. AAAI2011

Language Navigation

Instruction: "Go away from the lamp to the intersection of the red brick and wood"

Basic: Turn ( ) ,
Travel (steps: 1)

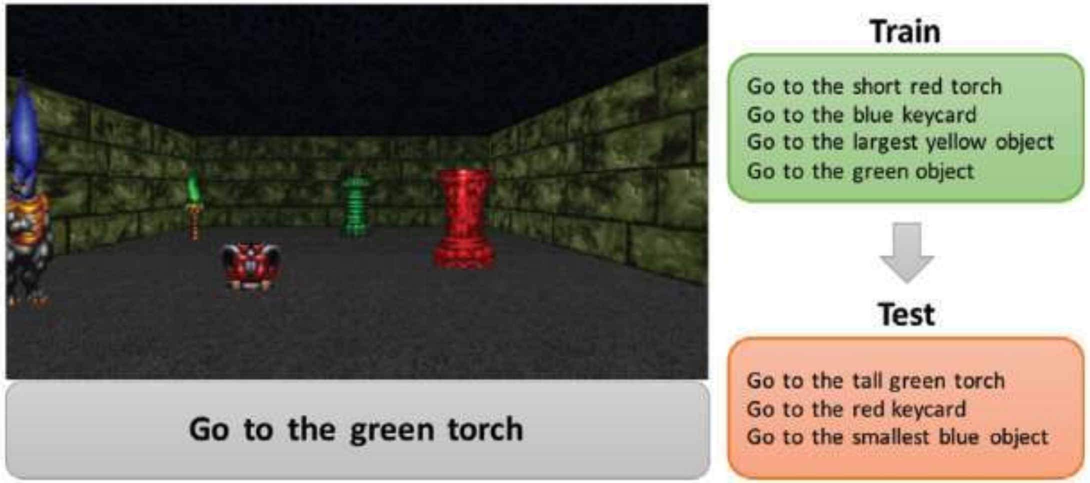
Landmarks: Turn ( ) ,
Verify (left: WALL , back: LAMP , back: HATRACK , front: BRICK HALL) ,
Travel (steps: 1) ,
Verify (side: WOOD HALL)



- "Go towards the coat rack and take a left at the coat rack. go all the way to the end of the hall and this is 4."
- "turn so that the wall is on your right side. walk forward once. turn left. walk forward twice."

Chen, David L, et al. Learning to interpret natural language navigation instructions from observations. AAAI2011

Gated-Attention



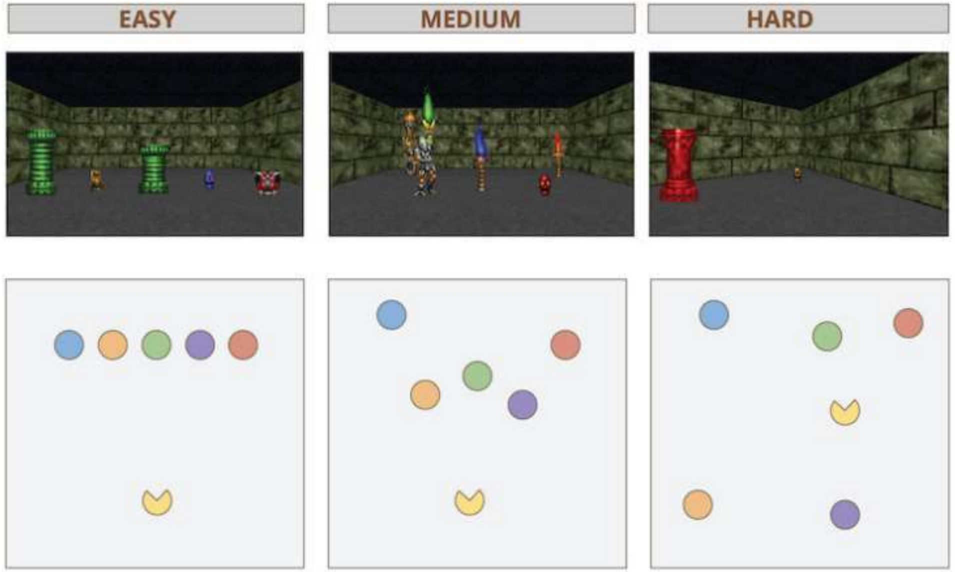
Go to the green torch

Train

Go to the short red torch
Go to the blue keycard
Go to the largest yellow object
Go to the green object

Test

Go to the tall green torch
Go to the red keycard
Go to the smallest blue object



Chaplot, et al. Gated-Attention Architectures for Task-Oriented Language Grounding. AAAI2018
Sinha, et al. Attention Based Natural Language Grounding by Navigating Virtual Environment. WACV2019

Problem Def

costly

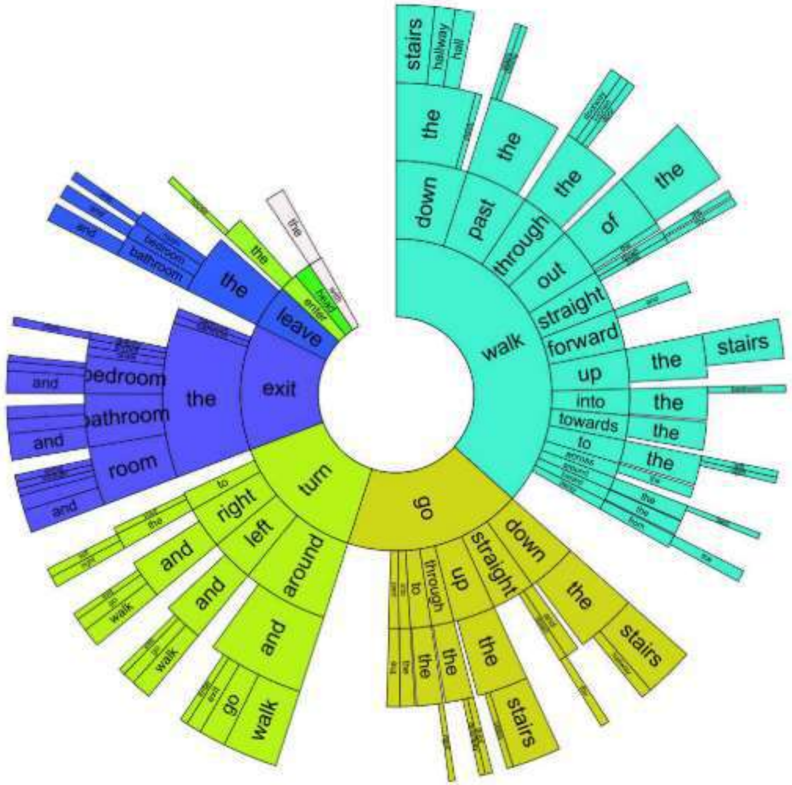
Matterport3D

- Big：
- 10,800 panoramic views
- from 194,400 RGB-D images
- Of 90 building-scale scenes

Chang, A,et al. Matterport3D: Learning from RGB-D data in indoor environments. 3DV 2017.

Vision-and-Language Navigation

- Over 400 workers
- 1,600 hours

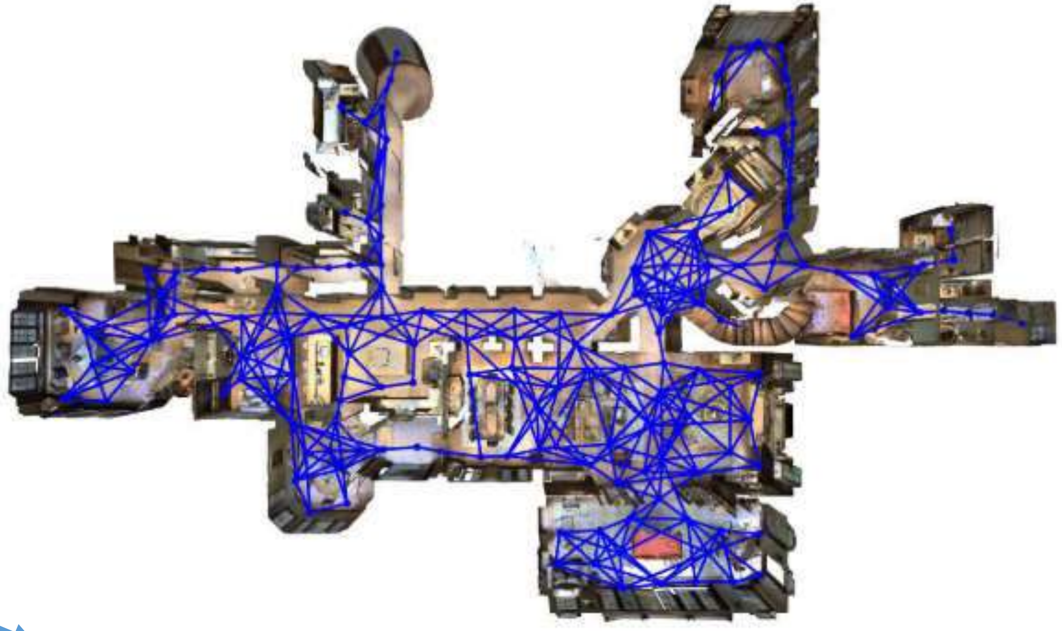


Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

Anderson, et al. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. CVPR2018

Action Space

- 6 actions
- View: left, right, up, down
- Move: forward
- End: stop



$$W_{t+1} = \{v_t\} \cup \{v_i \in V \mid \langle v_t, v_i \rangle \in E \wedge v_i \in P_t\}$$

Anderson, et al. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. CVPR2018

Challenges

- cross-modal grounding
- ill-posed feedback
- Generalization

Methods

CV & NLP & RL

Baseline seq-to-seq

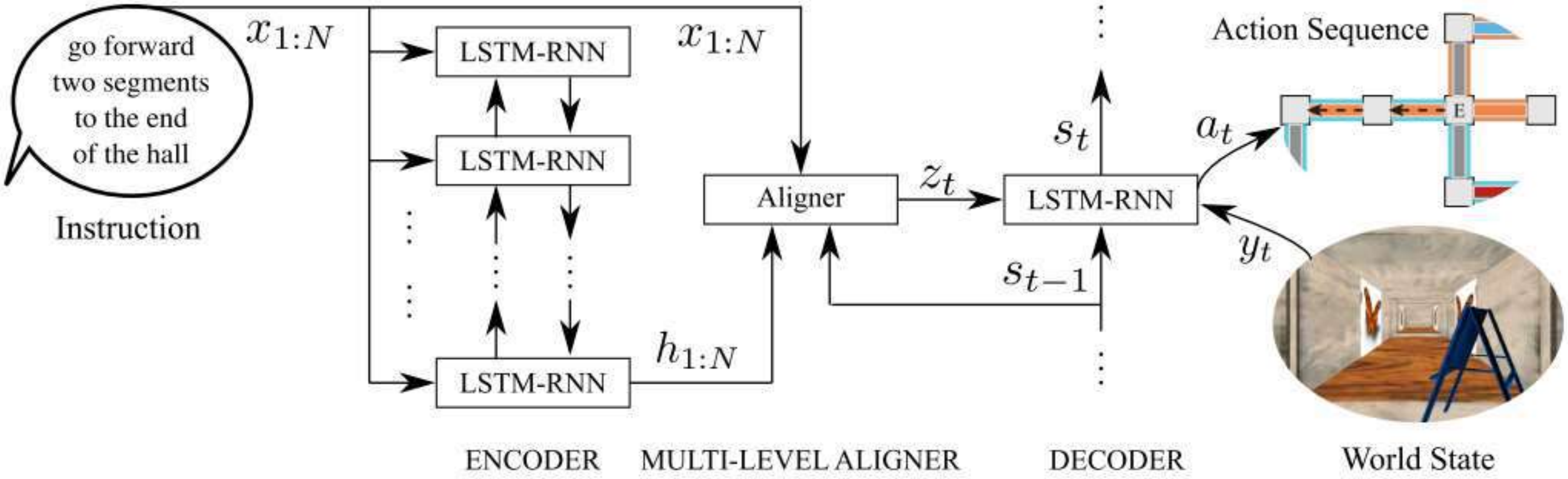
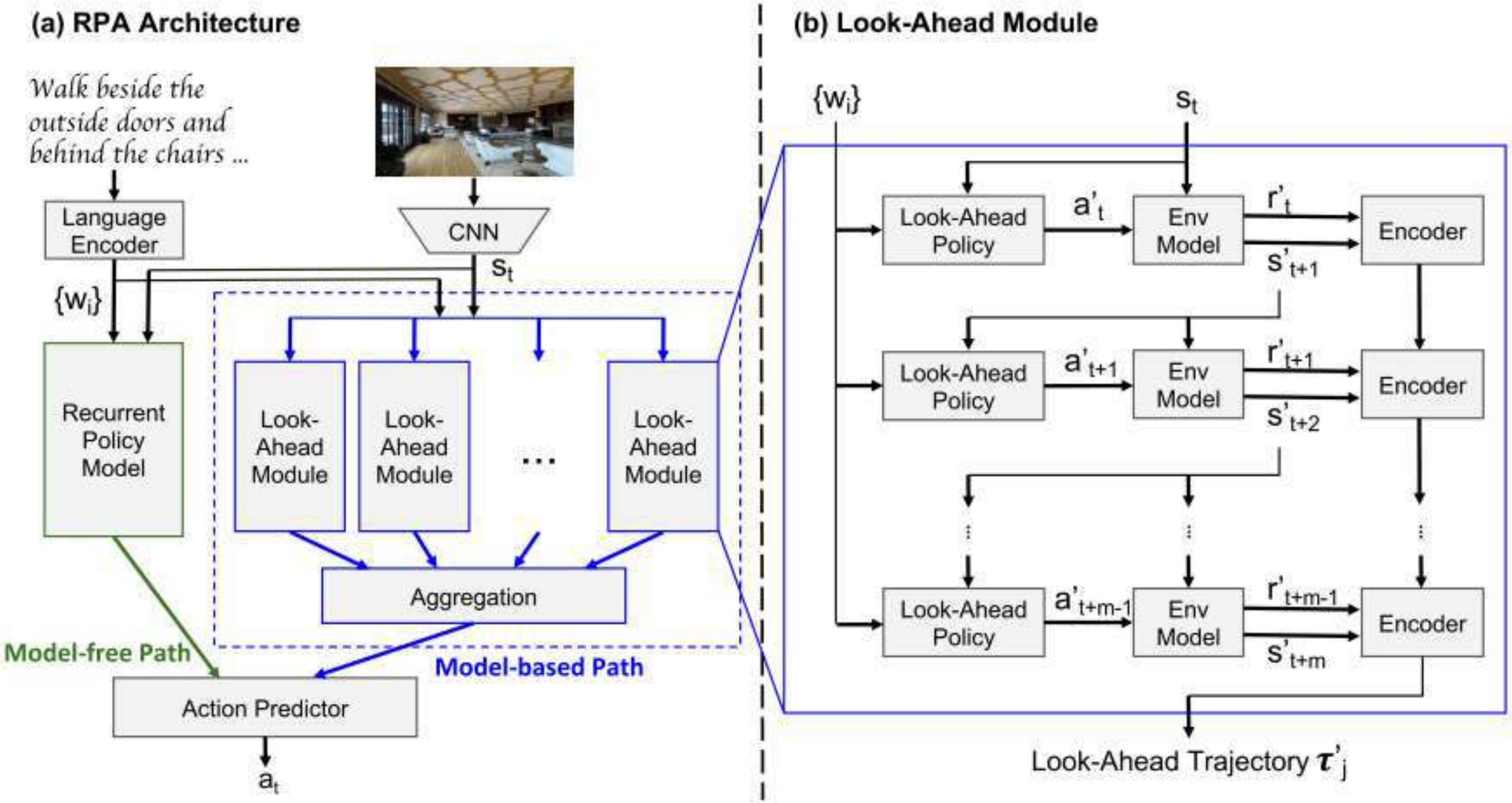


Figure 2: Our encoder-aligner-decoder model with multi-level alignment

Mei, et al. Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences. AAAI 2016.

Look Before You Leap



Wang, Xin, et al. Look Before You Leap: Bridging Model-Free and Model-Based Reinforcement Learning for Planned-Ahead Vision-and-Language Navigation. ECCV2018.

Look Before You Leap

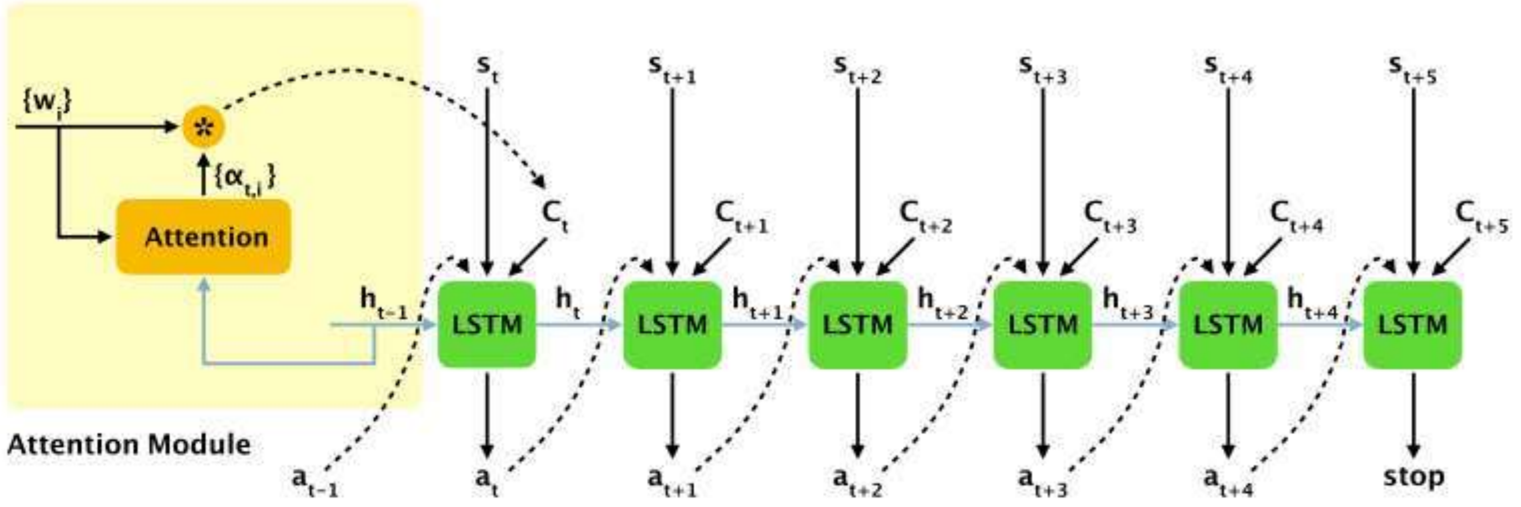


Fig. 4: An example of the unrolled recurrent policy model (from t to $t+5$). The left-side yellow region demonstrates the attention mechanism at time step t.

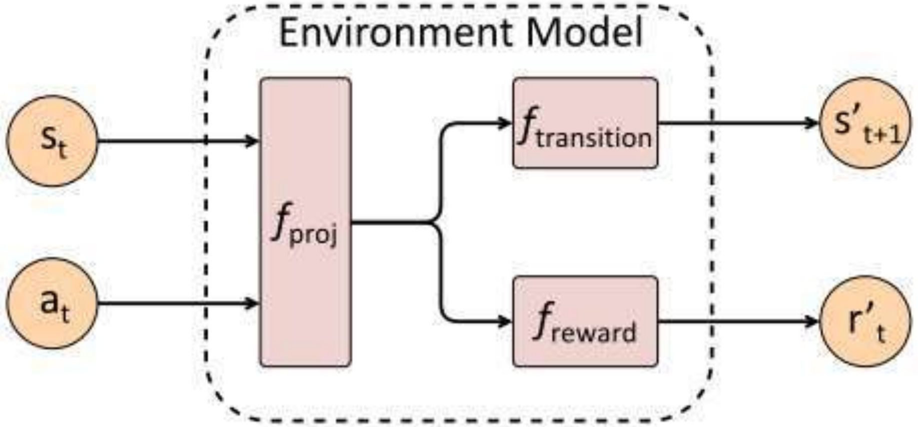
$$c_t = \sum \alpha_{t,i} w_i$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{n} \exp(e_{t,k})} \quad , \quad \text{where } e_{t,i} = h_{t-1}^{\top} w_i$$

$$h_t = LSTM(h_{t-1}, [c_t, s_t, a_{t-1}])$$

Wang, Xin, et al. Look Before You Leap: Bridging Model-Free and Model-Based Reinforcement Learning for Planned-Ahead Vision-and-Language Navigation. ECCV2018.

Look Before You Leap

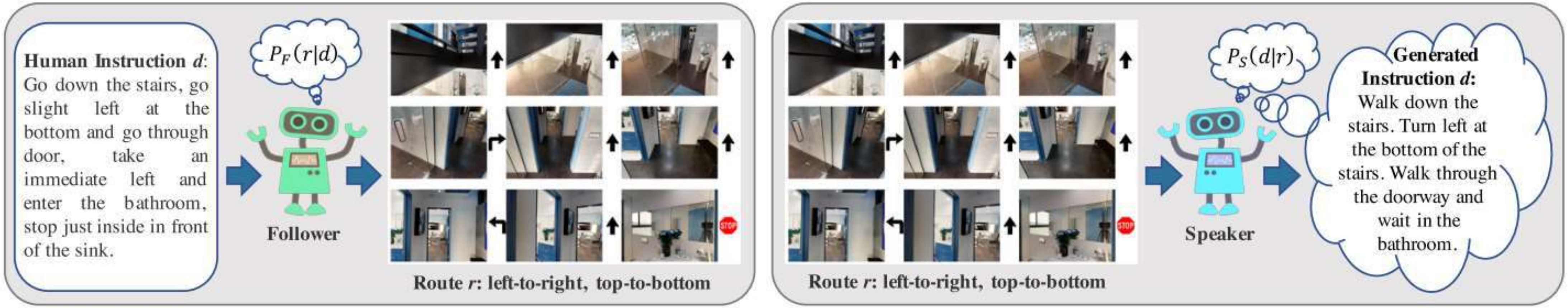


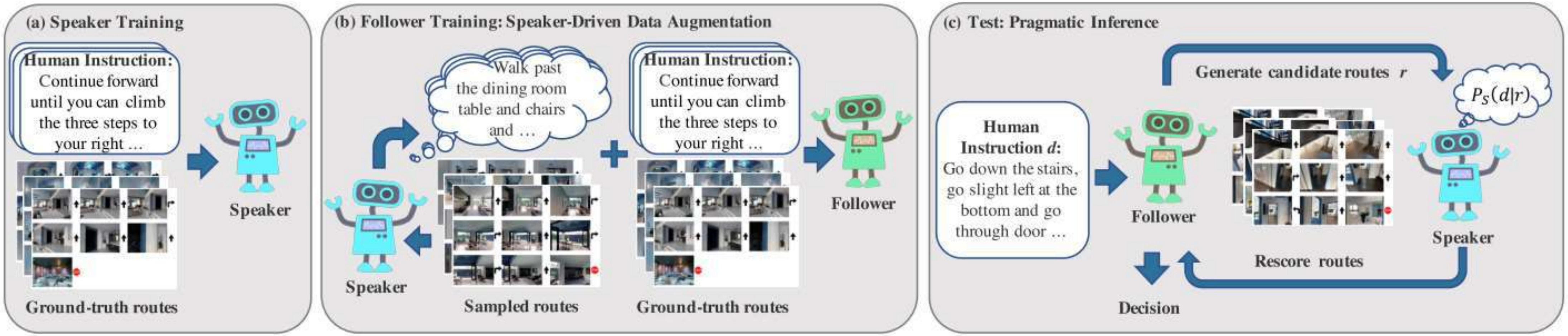
$$s'_{t+1} = f_{\text{transition}}(f_{\text{proj}}(s_t, a_t))$$
$$r'_t = f_{\text{reward}}(f_{\text{proj}}(s_t, a_t))$$

$$\max_{\theta} \mathcal{J}^{\pi} = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r(a_t, s_t) | \pi(o_t; \theta)\right]$$

Wang, Xin, et al. Look Before You Leap: Bridging Model-Free and Model-Based Reinforcement Learning for Planned-Ahead Vision-and-Language Navigation. ECCV2018.

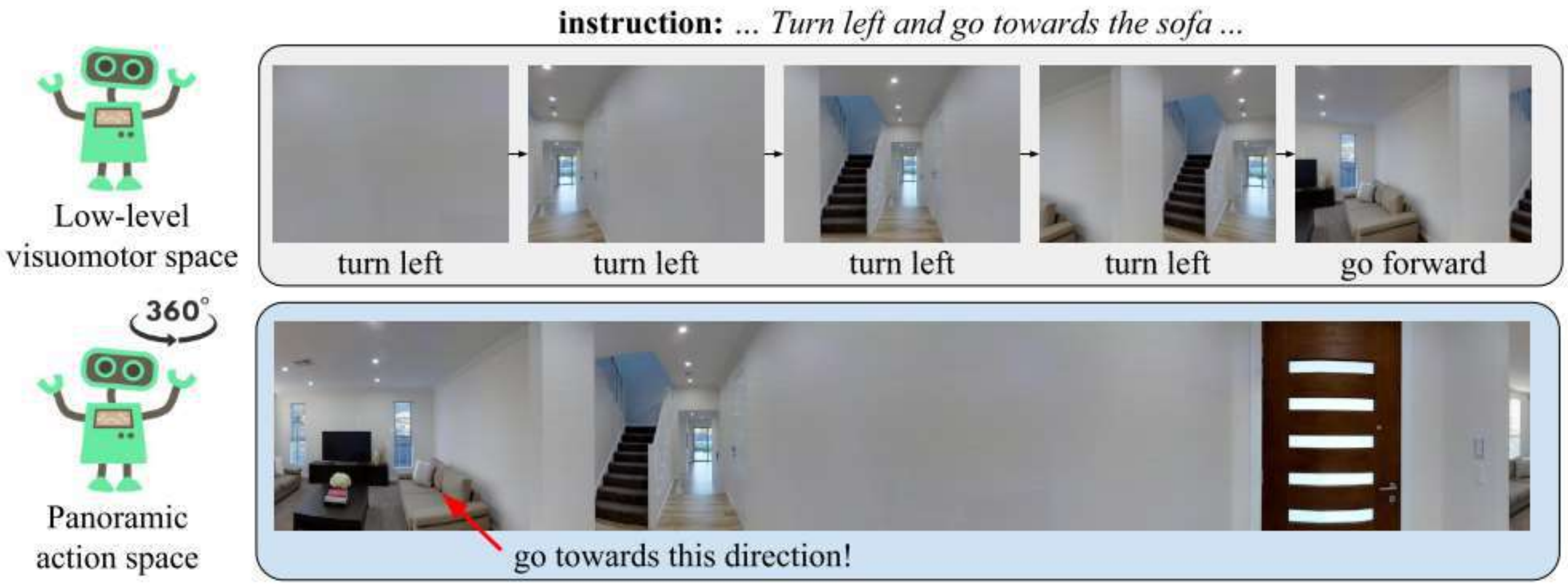
Speaker-Follower


Human Instruction *d*: Go down the stairs, go slight left at the bottom and go through door, take an immediate left and enter the bathroom, stop just inside in front of the sink.
$P_F(r|d)$
Follower
Route *r*: left-to-right, top-to-bottom
Route *r*: left-to-right, top-to-bottom
$P_S(d|r)$
Speaker
Generated Instruction *d*: Walk down the stairs. Turn left at the bottom of the stairs. Walk through the doorway and wait in the bathroom.


(a) Speaker Training
Human Instruction: Continue forward until you can climb the three steps to your right ...
Speaker
Ground-truth routes
(b) Follower Training: Speaker-Driven Data Augmentation
Walk past the dining room table and chairs and ...
Human Instruction: Continue forward until you can climb the three steps to your right ...
Follower
Speaker
Sampled routes
Ground-truth routes
(c) Test: Pragmatic Inference
Generate candidate routes *r*
$P_S(d|r)$
Human Instruction *d*: Go down the stairs, go slight left at the bottom and go through door ...
Follower
Speaker
Rescore routes
Decision

Fried, Daniel, et al. Speaker-Follower Models for Vision-and-Language Navigation. NIPS2018

Speaker-Follower


instruction: ... *Turn left and go towards the sofa ...*
Low-level visuomotor space
turn left
turn left
turn left
turn left
go forward
360°
Panoramic action space
go towards this direction!

Reinforced Cross-Modal Match

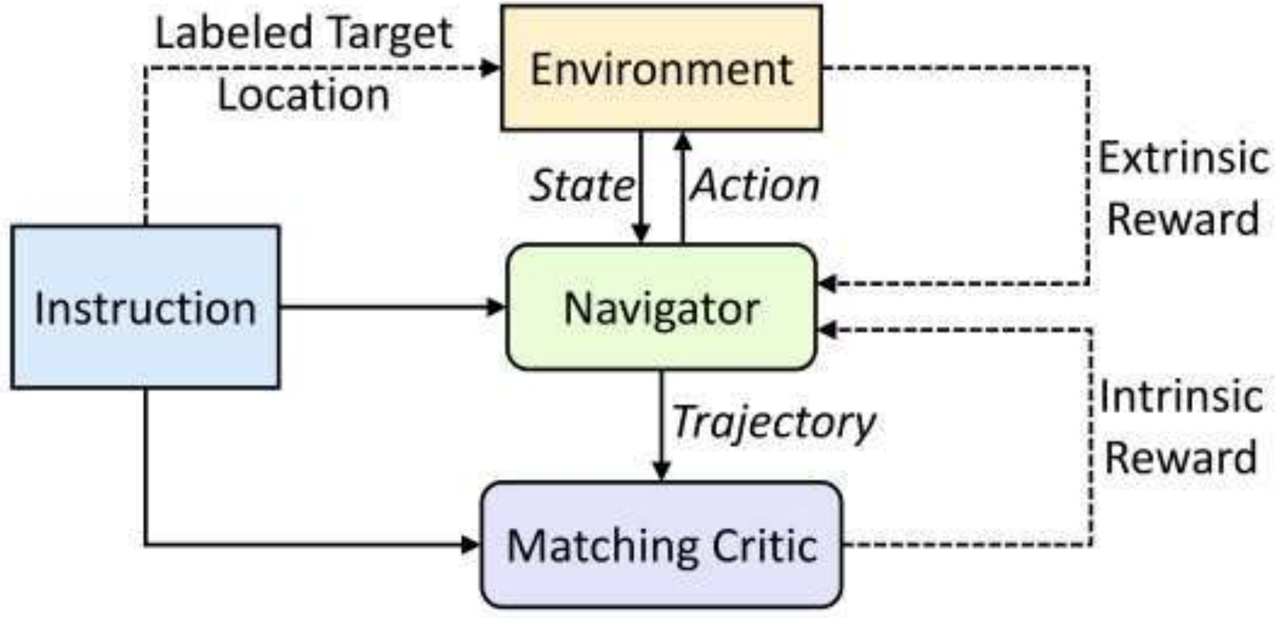


Figure 2: Overview of our RCM framework.

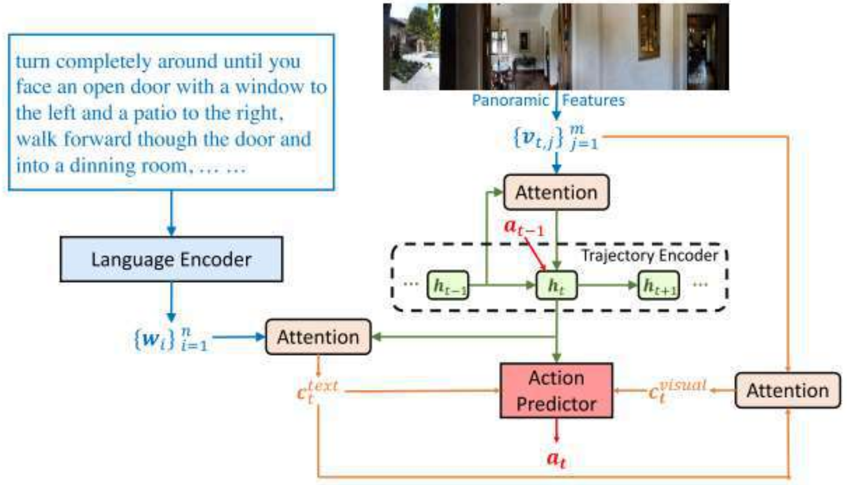
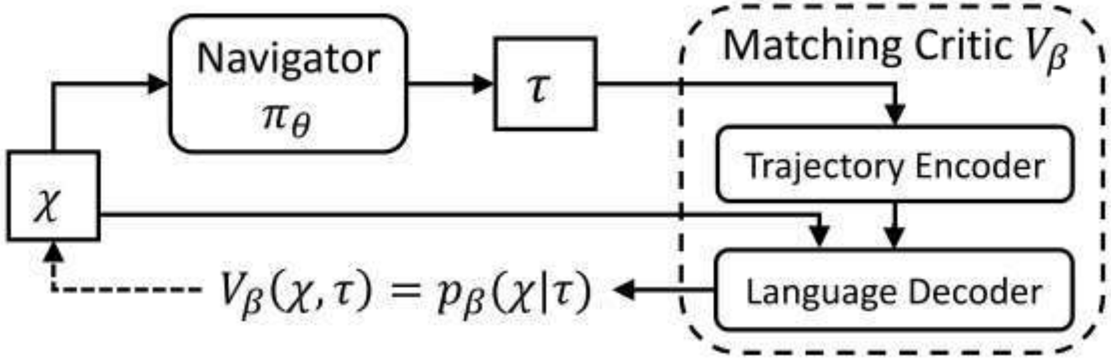
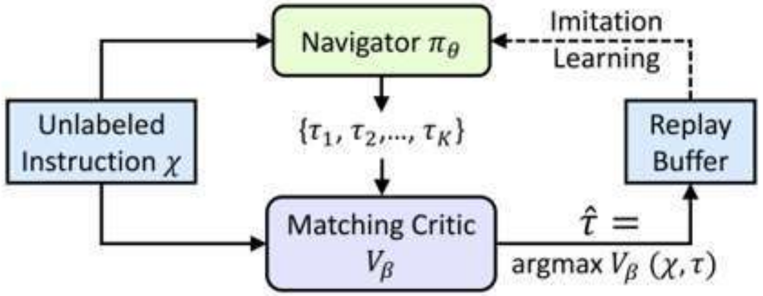


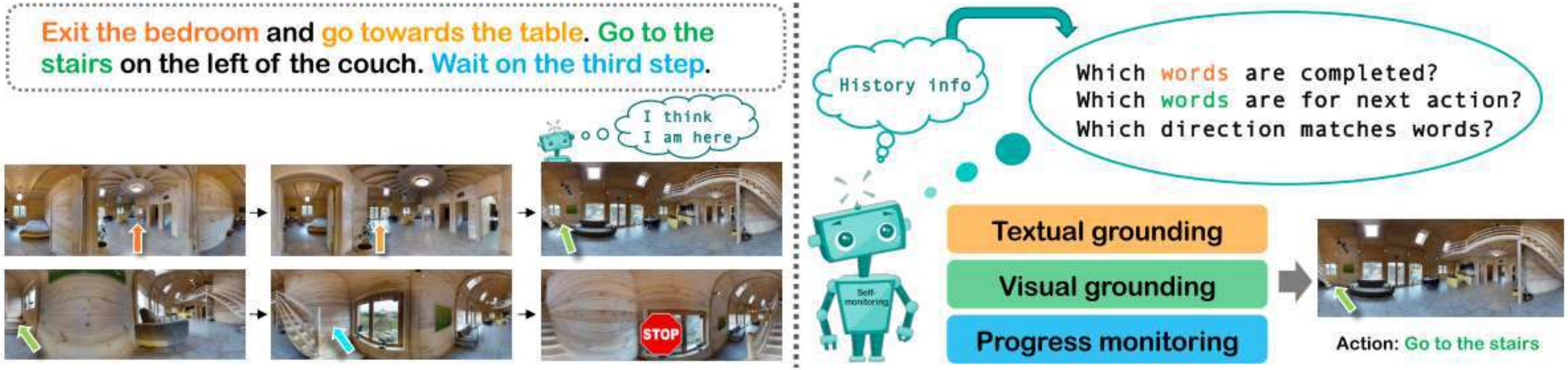
Figure 3: Cross-modal reasoning navigator at step t.




Navigator π_θ
Imitation Learning
Unlabeled Instruction χ
$\{\tau_1, \tau_2, \ldots, \tau_K\}$
Replay Buffer
Matching Critic V_β
$\hat{\tau} =$ argmax $V_\beta(\chi, \tau)$

Wang Xin, et al. Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation
CVPR2019 best stu paper

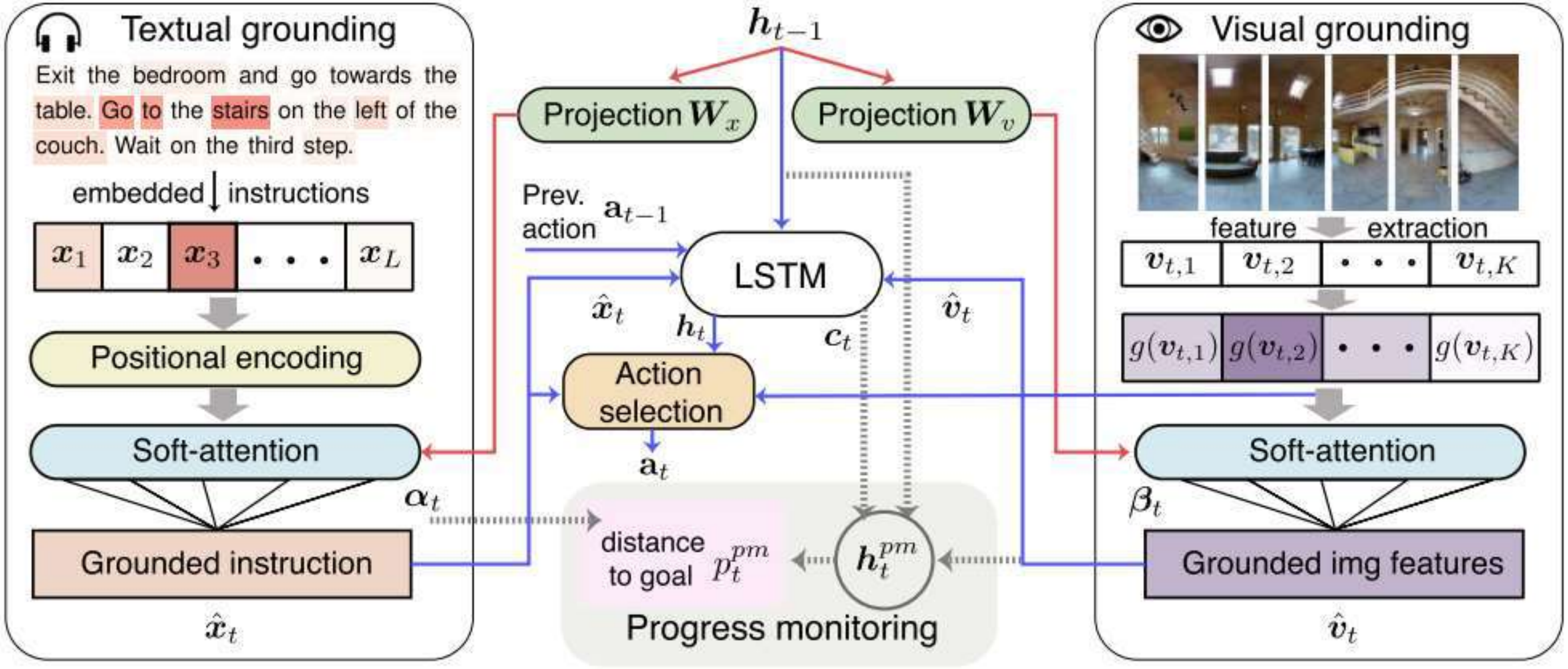
Self Monitoring


Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.
I think I am here
STOP
History info
Which words are completed?
Which words are for next action?
Which direction matches words?
Self-monitoring
Textual grounding
Visual grounding
Progress monitoring
Action: Go to the stairs

Ma, Chih-Yao, et al. SELF-MONITORING NAVIGATION AGENT VIA AUXIL-IARY PROGRESS ESTIMATION. ICLR2019.

Self Monitoring

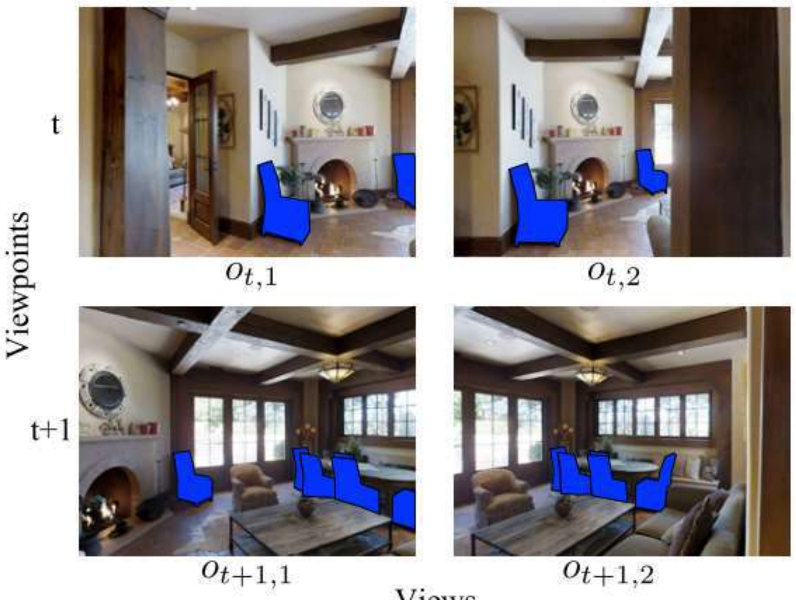
$$z_{t,l}^{\text{textual}} = (\boldsymbol{W}_x \boldsymbol{h}_{t-1})^\top PE(\boldsymbol{x}_l) \qquad o_{t,k} = (\boldsymbol{W}_a[\boldsymbol{h}_t, \hat{\boldsymbol{x}}_t])^\top g(\boldsymbol{v}_{t,k}) \qquad z_{t,k}^{\text{visual}} = (\boldsymbol{W}_v \boldsymbol{h}_{t-1})^\top g(\boldsymbol{v}_{t,k})$$



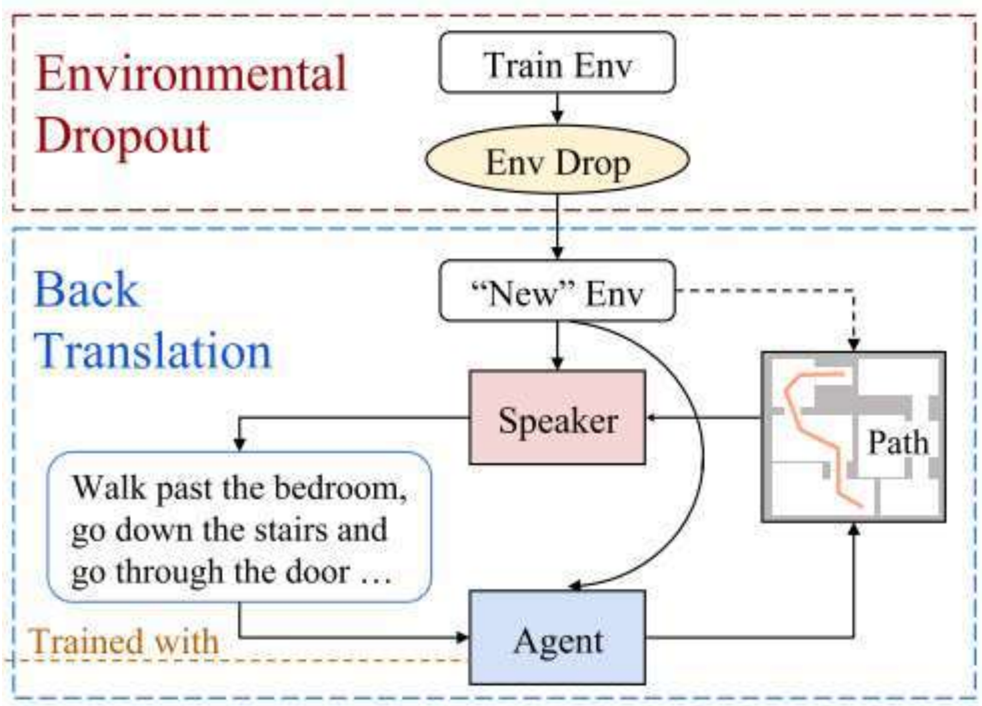
$$\boldsymbol{h}_t^{pm} = \sigma(\boldsymbol{W}_h([\boldsymbol{h}_{t-1}, \hat{\boldsymbol{v}}_t]) \otimes \tanh(\boldsymbol{c}_t))$$

Ma, Chih-Yao, et al. SELF-MONITORING NAVIGATION AGENT VIA AUXIL-IARY PROGRESS ESTIMATION. ICLR2019.

Environmental Dropout

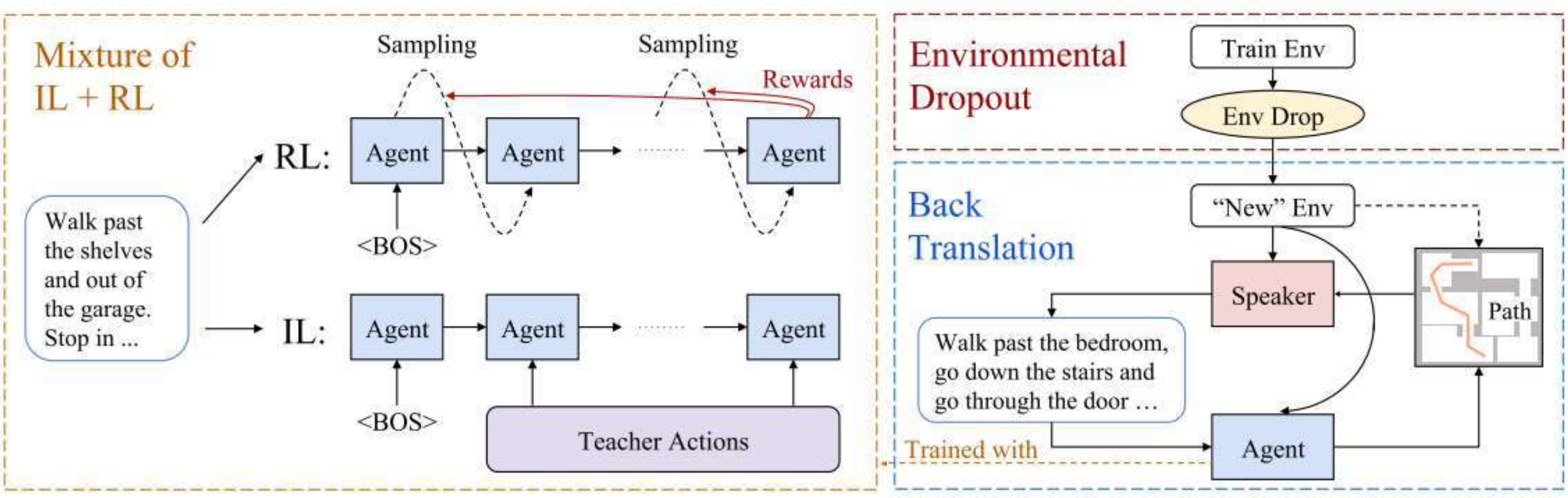


(b) Environmental dropout



Tan Hao, et al. Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. ACL2019

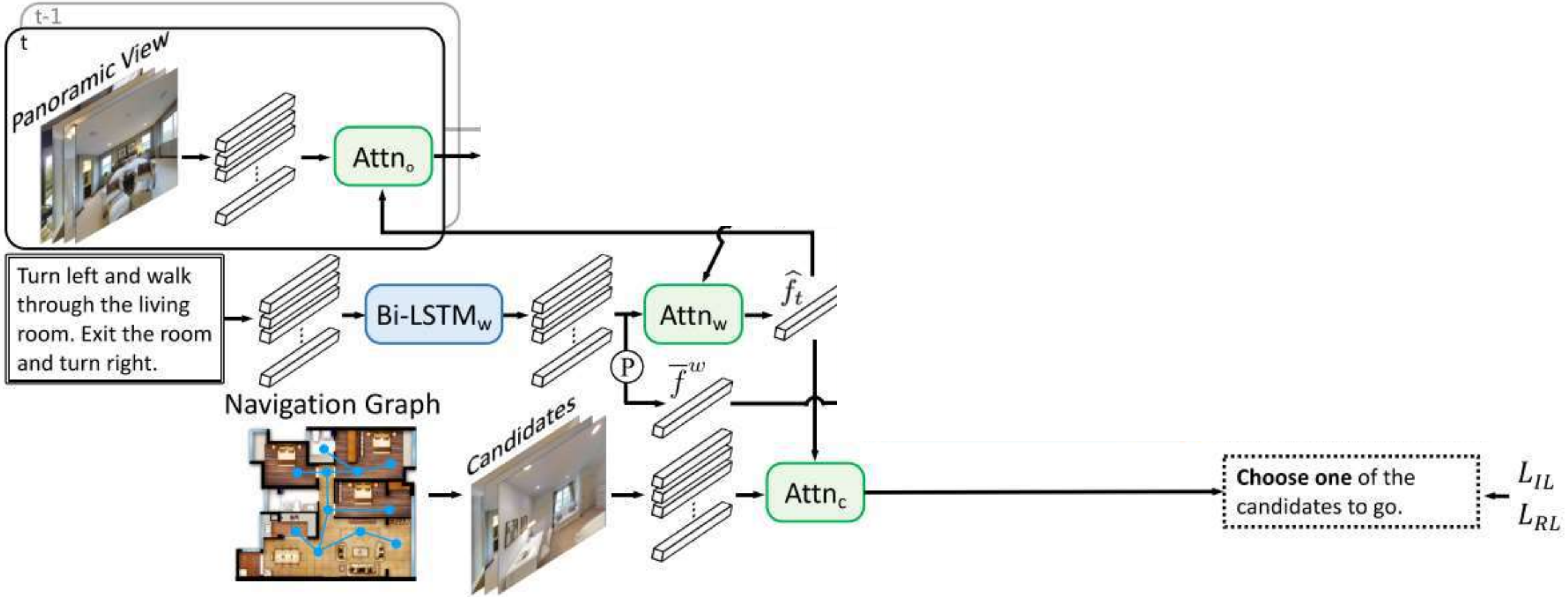
Environmental Dropout



$$\mathcal{L}^{\text{MIX}} = \mathcal{L}^{\text{RL}} + \lambda_{\text{IL}}\mathcal{L}^{\text{IL}}$$

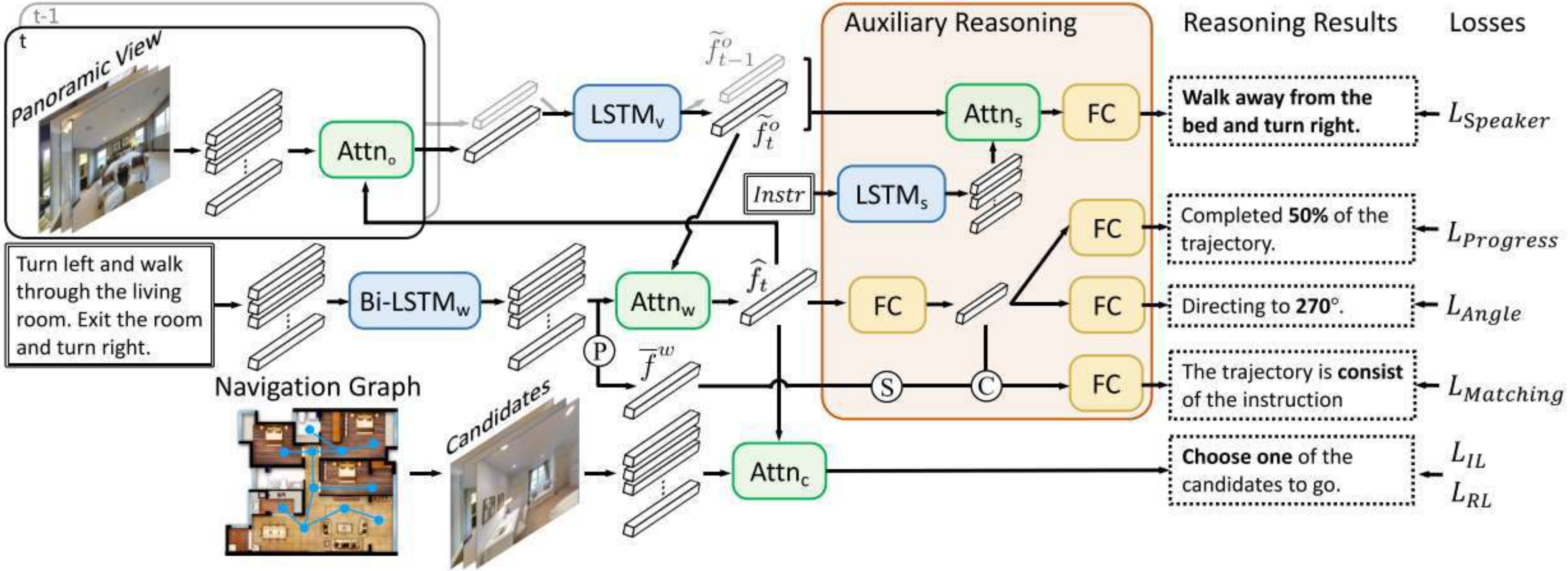
Tan Hao, et al. Learning to Navigate Unseen Environments: Back Translation with Environmental Dropout. ACL2019

Auxiliary Reasoning Tasks



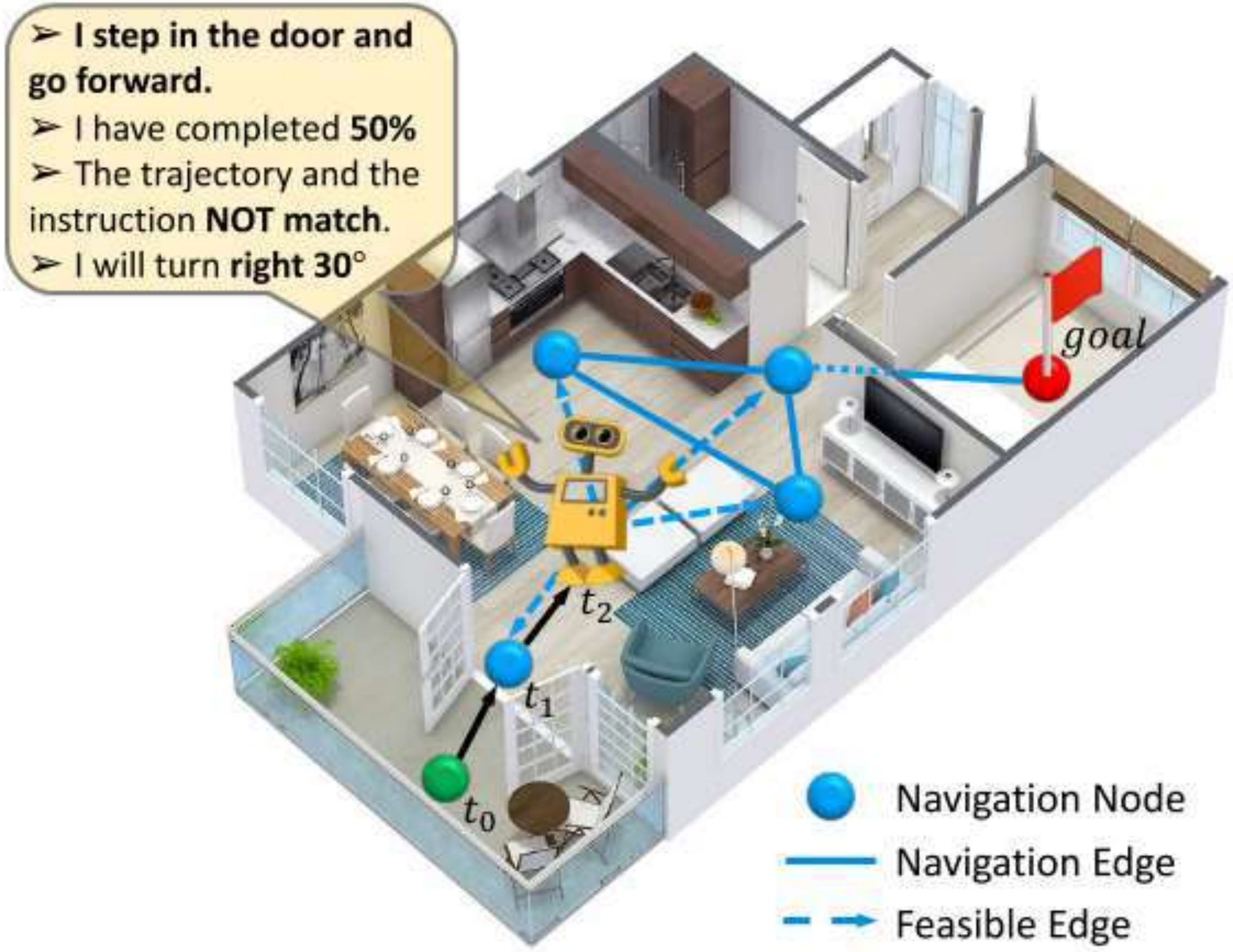
Zhu fd, Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks. CVPR2020

Auxiliary Reasoning Tasks



Zhu fd, Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks. CVPR2020

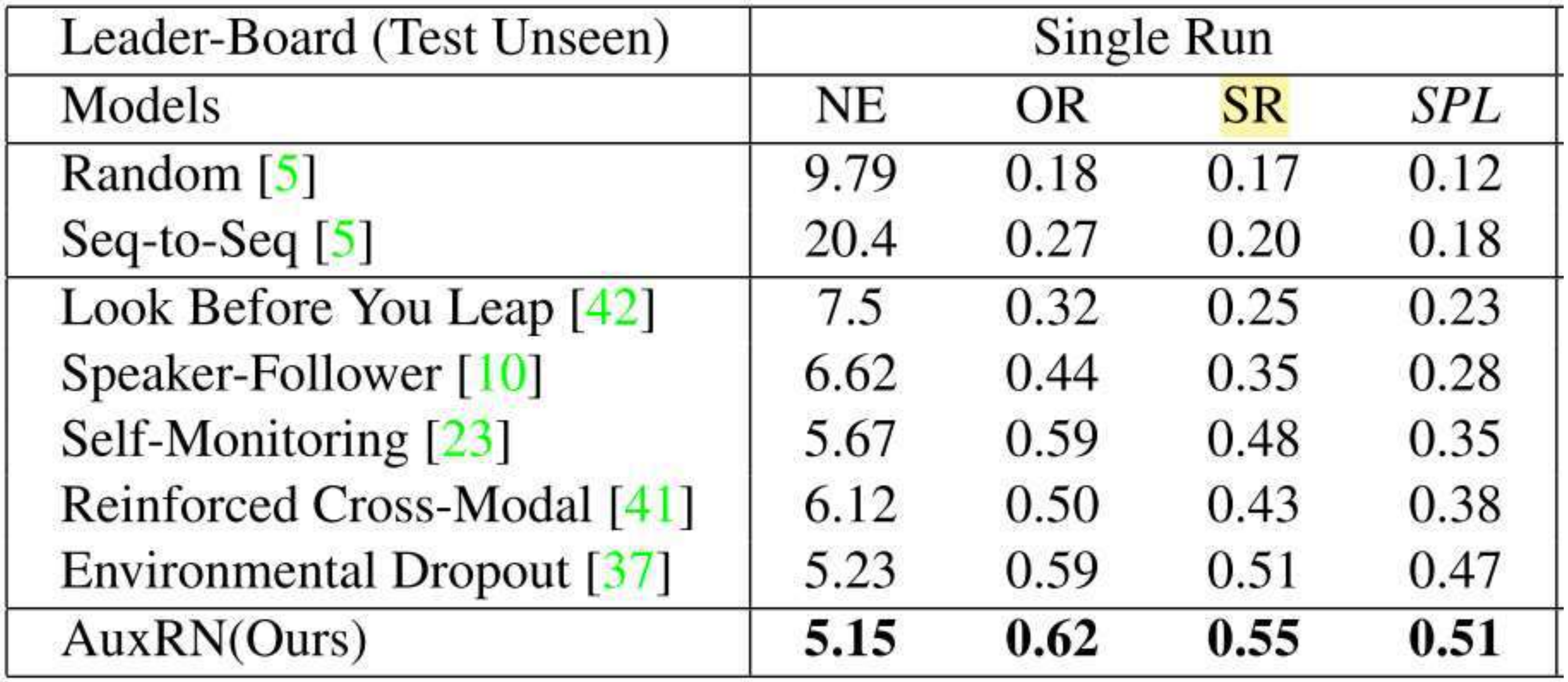
Auxiliary Reasoning Tasks



- I step in the door and go forward.
- I have completed **50%**
- The trajectory and the instruction **NOT match.**
- I will turn **right 30°**

Zhu fd, Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks. CVPR2020

Experimental results



Leader-Board (Test Unseen)	Single Run			
Models	NE	OR	SR	*SPL*
Random [5]	9.79	0.18	0.17	0.12
Seq-to-Seq [5]	20.4	0.27	0.20	0.18
Look Before You Leap [42]	7.5	0.32	0.25	0.23
Speaker-Follower [10]	6.62	0.44	0.35	0.28
Self-Monitoring [23]	5.67	0.59	0.48	0.35
Reinforced Cross-Modal [41]	6.12	0.50	0.43	0.38
Environmental Dropout [37]	5.23	0.59	0.51	0.47
AuxRN(Ours)	**5.15**	**0.62**	**0.55**	**0.51**

Zhu fd, Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks. CVPR2020

Conclusion

- Base: Seq-to-seq

- Align(Matching, Progress)
- Data augmentation(Self/semi-supervise)
- RL(Model-based/free, IL, SIL)

- How to get the data

Thank you

- End