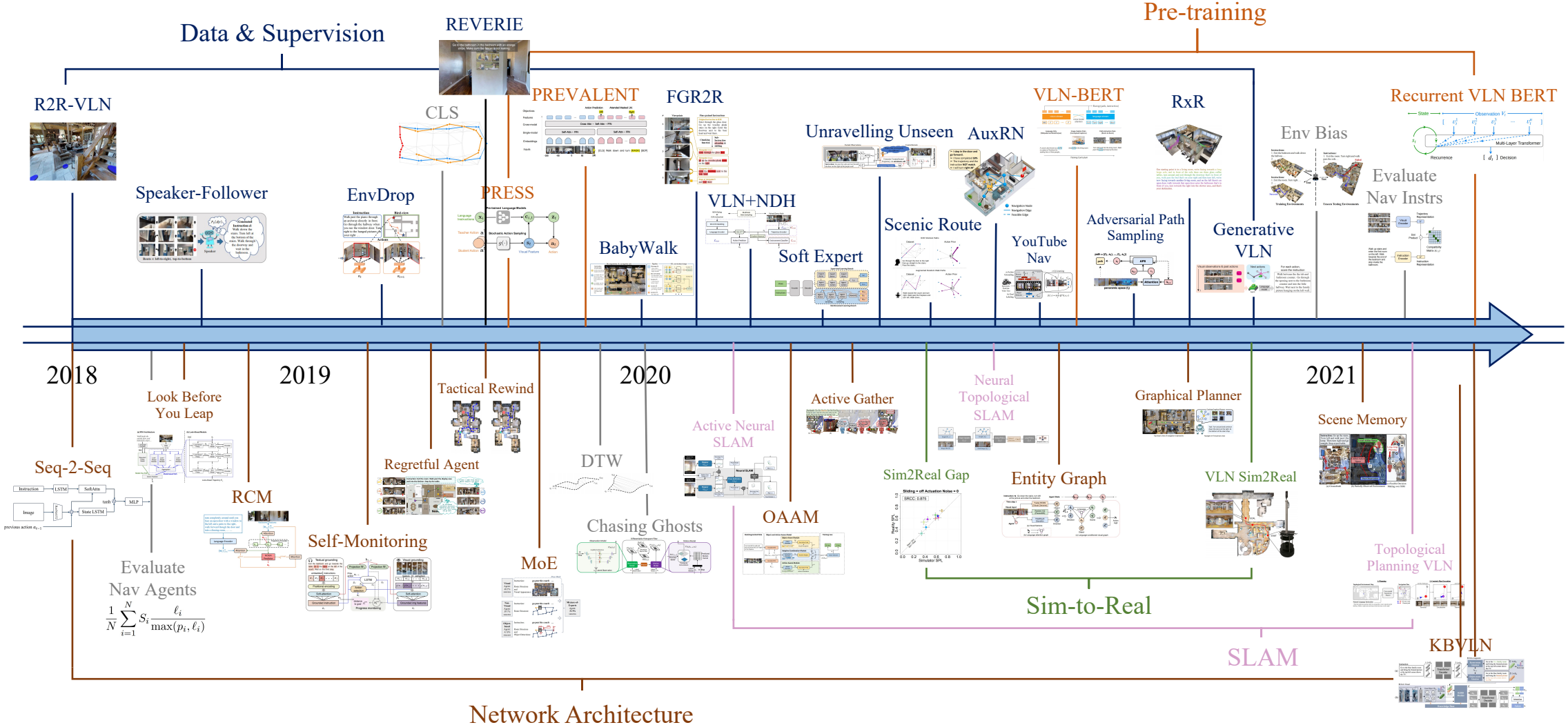
VLN Tutorial Summarisation

Qi Wu

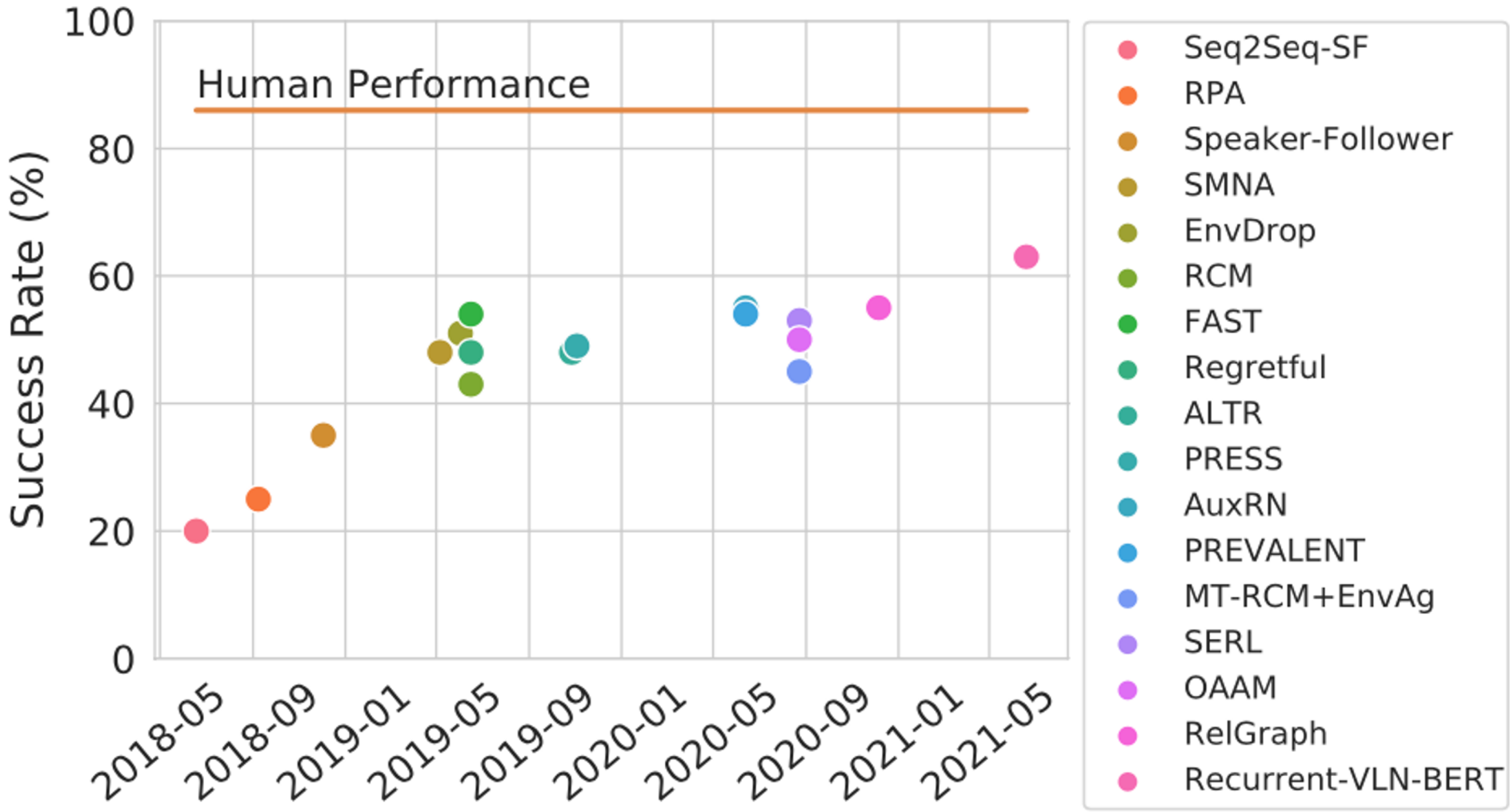
University of Adelaide

VLN Timeline



- Evaluation methods, Analysis, and others

Some studies



Research Questions

- What can the agents learn from the instructions?
  - Do they pay more attention to object tokens or directions tokens?
- What do the agents capture from the visual environment?
  - Are they staring at the closely surrounded objects or also browsing further layout?
  - Do they focus on individual visual instances or perceive the overall outline?
- Can agents match textual tokens to visual entities?
  - How reliable are such connections?

Diagnosing Vision-and-Language Navigation: What Really Matters?

Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel Eckstein, William Yang Wang

https://arxiv.org/abs/2103.16561

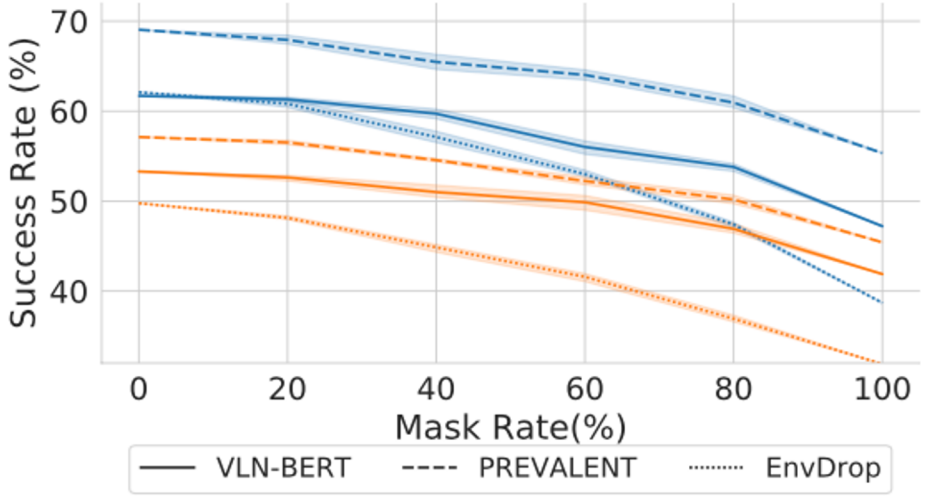
The Effect of Object-related Tokens

Setting	Instruction
Original	Go left down the hallway toward the exit sign. Turn right and go down the hallway. Go into the door on the left and stop by the table.
Mask out 50% object tokens	Go left down the [MASK] toward the exit sign. Turn right and go down the [MASK]. To into the door on the left and stop by the [MASK].
Randomly mask baseline	Go left [MASK] the hallway toward the exit sign. [MASK] right and go down the hallway. To into the door on [MASK] left and stop by the table.

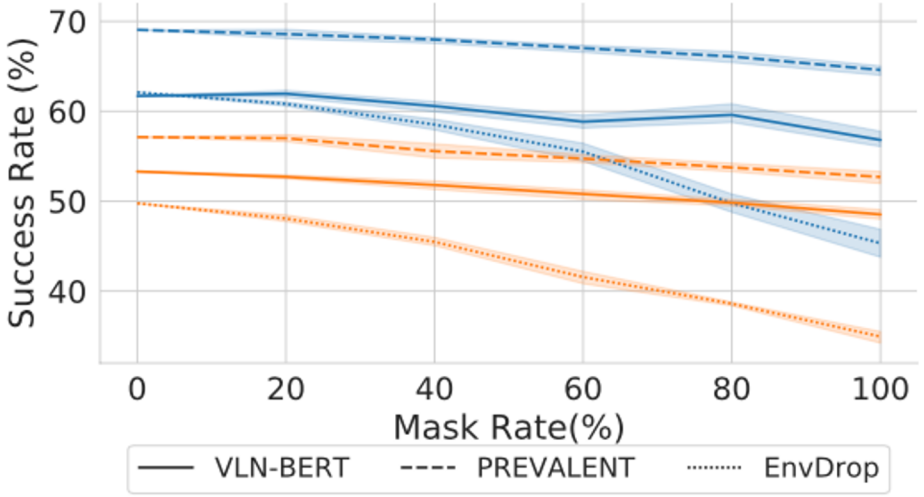
The Effect of Object-related Tokens

Blue lines: results on validation seen set
Orange lines: results on validation unseen set

- ➔ Indoor VLN
  - ◆ R2R

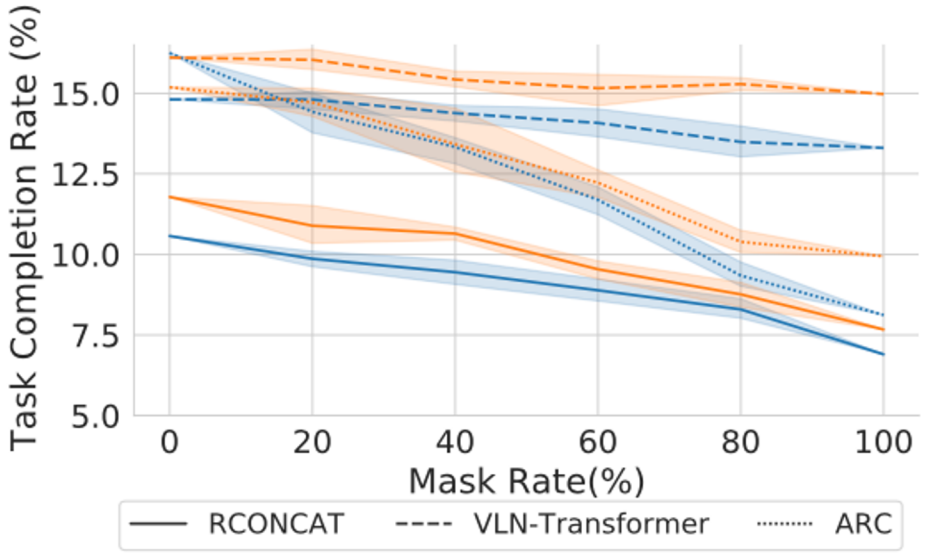


(a) [R2R] Mask Object Tokens

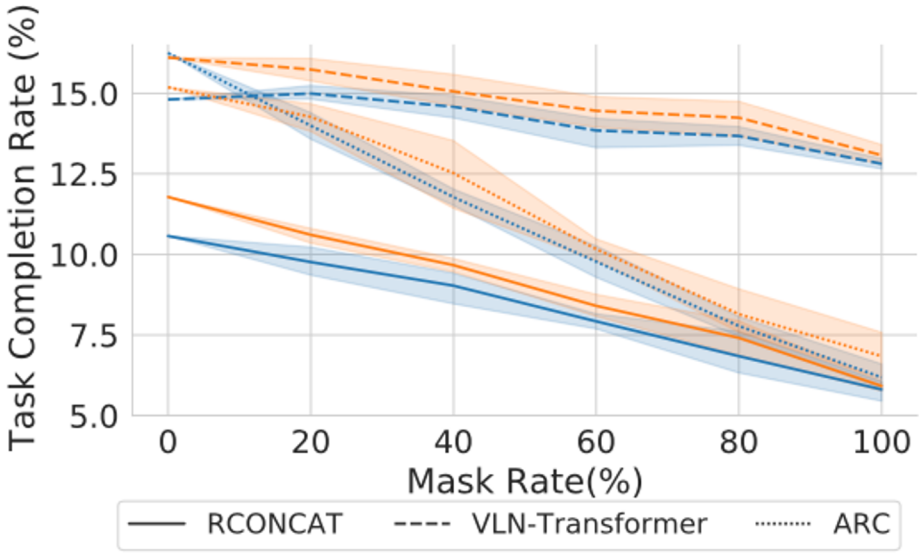


(b) [R2R] Randomly Mask Out

- ➔ Outdoor VLN
  - ◆ Touchdown



(d) [TD] Mask Object Tokens

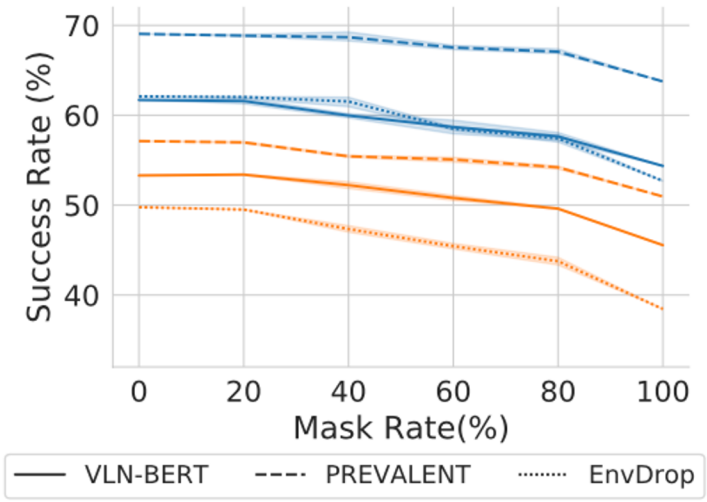


(e) [TD] Randomly Mask Out

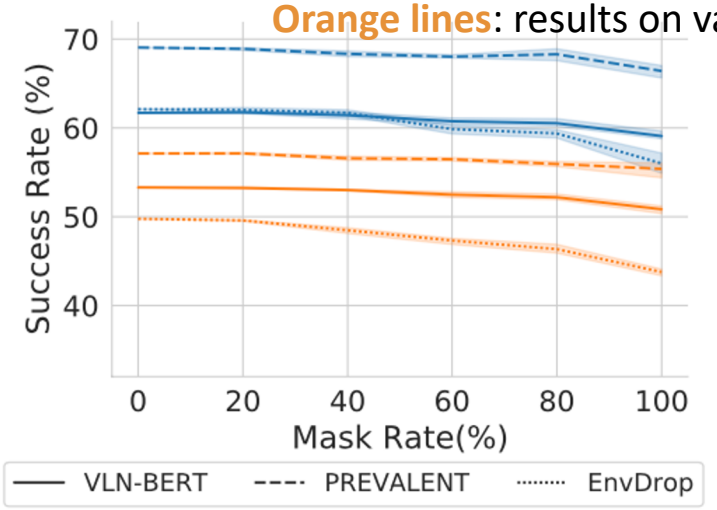
The Effect of Direction-related Tokens

Blue lines: results on validation seen set
Orange lines: results on validation unseen set

- ➔ Indoor VLN
  - ◆ R2R

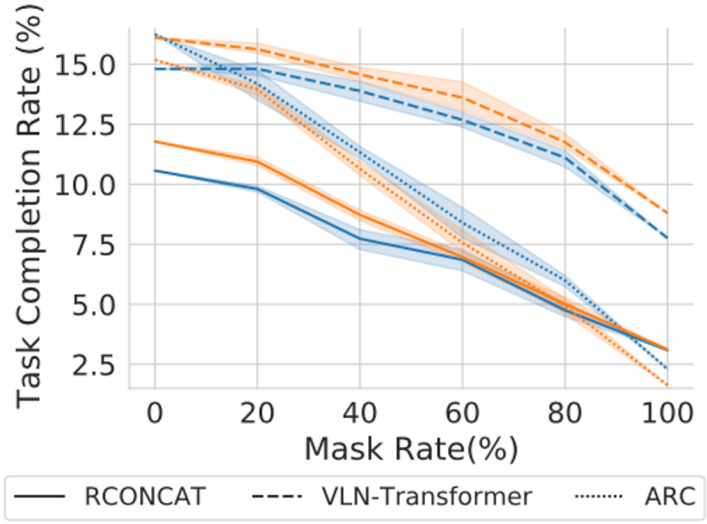


(a) [R2R] Mask Directions

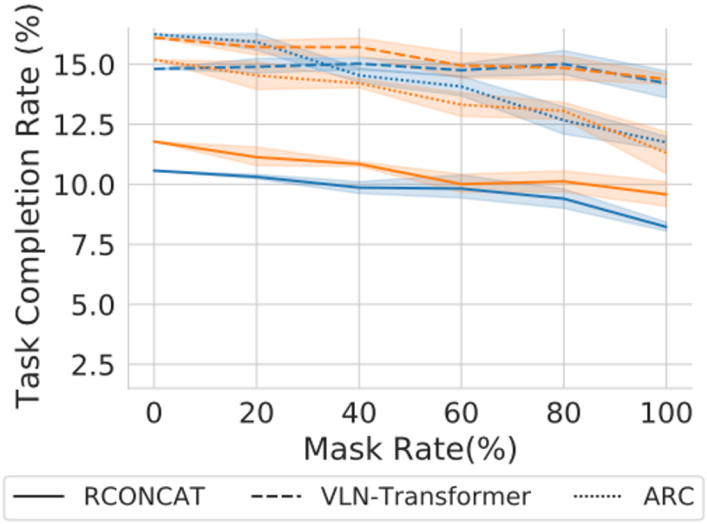


(b) [R2R] Randomly Mask Out

- ➔ Outdoor VLN
  - ◆ Touchdown



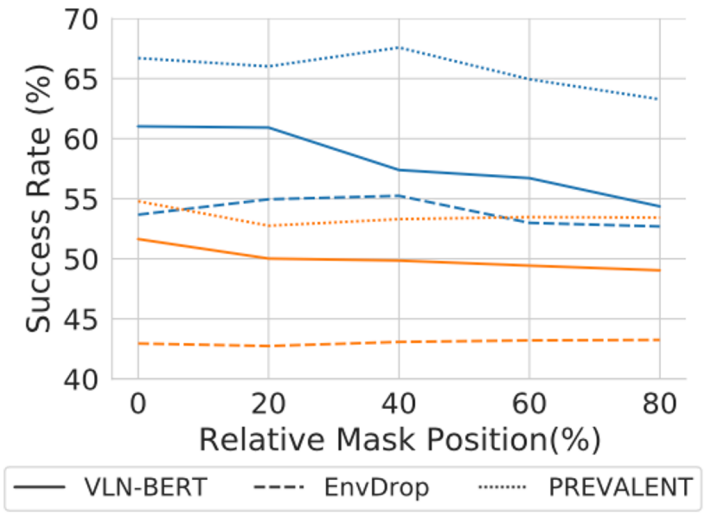
(c) [TD] Mask Directions



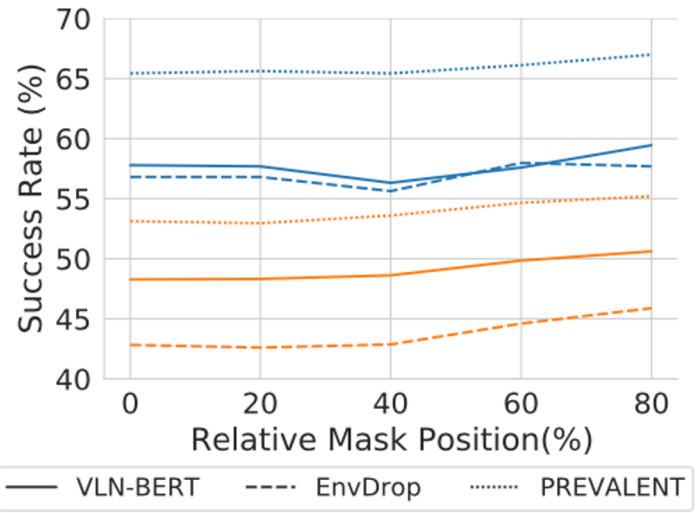
(d) [TD] Randomly Mask Out

The Effect of Objects and Directions at Different Position

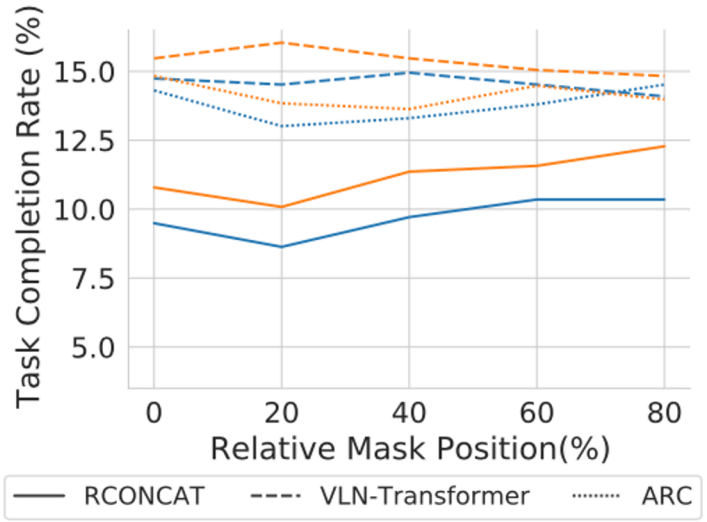
- ➔ Indoor VLN
  - ◆ R2R
- ➔ Outdoor VLN
  - ◆ Touchdown



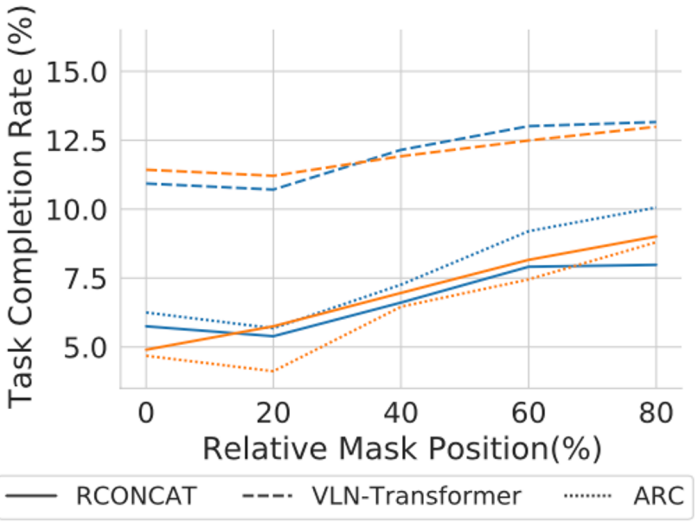
(a) [R2R] Mask Objects



(b) [R2R] Mask Directions



(c) [TD] Mask Objects

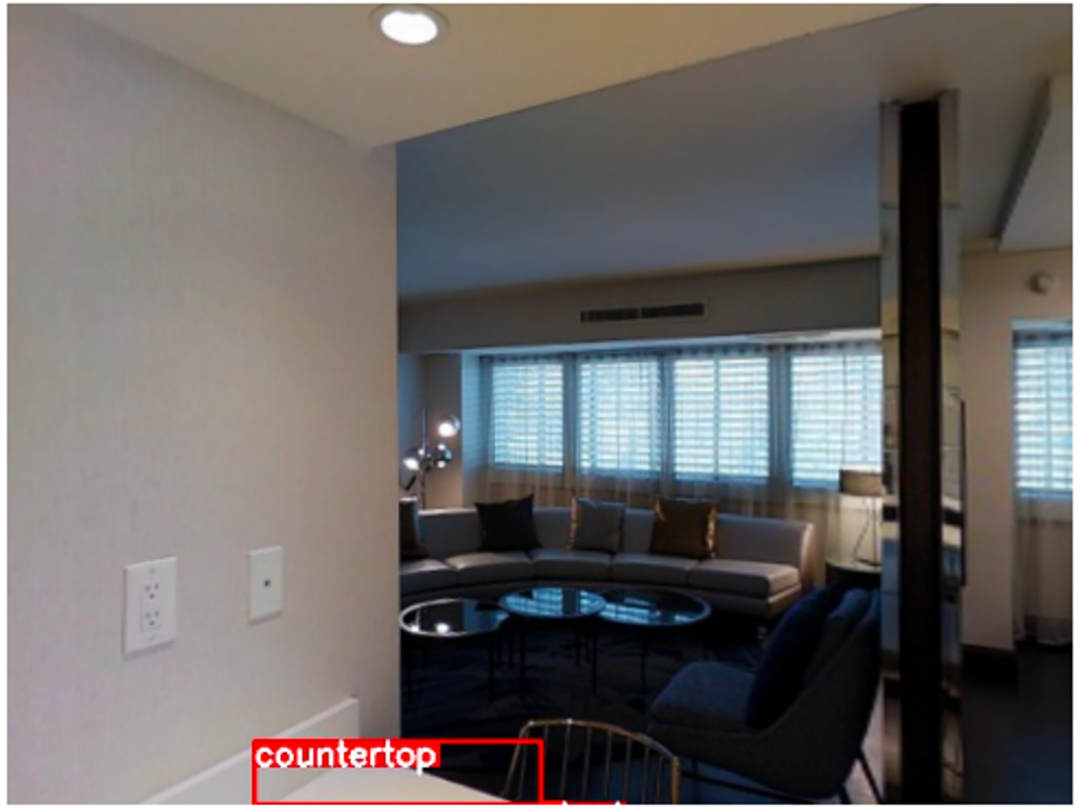


(d) [TD] Mask Directions

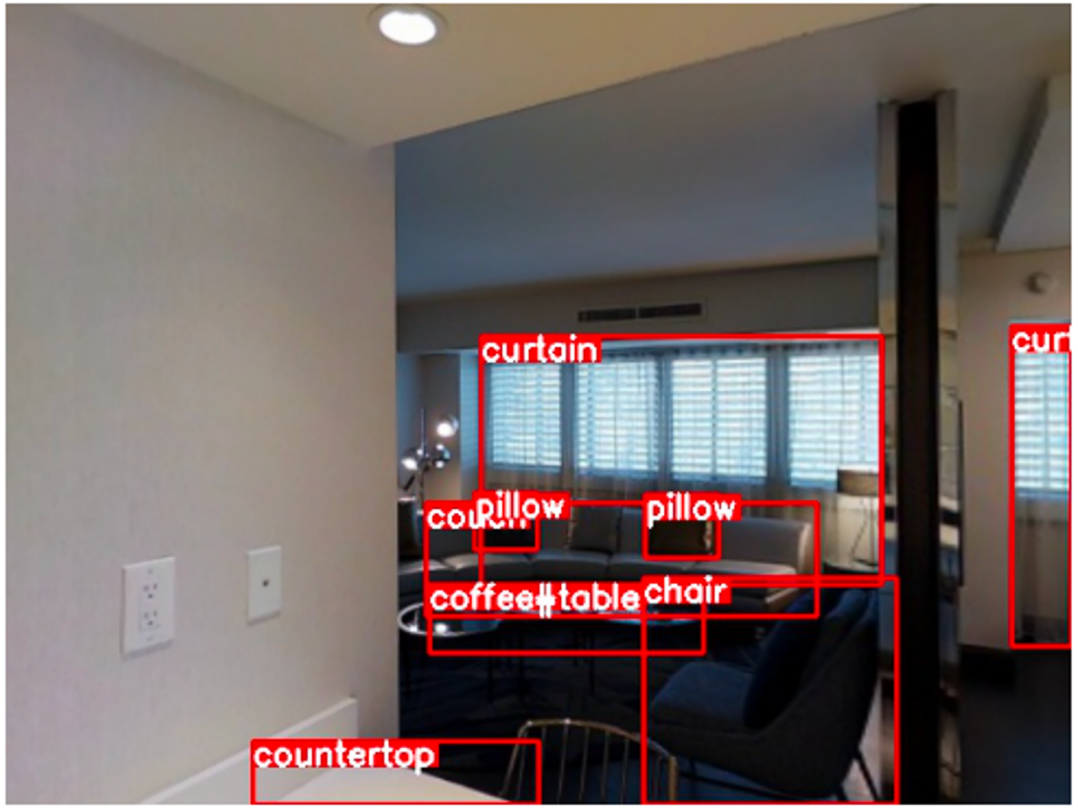
Takeaways

- Indoor agents will refer to object tokens in R2R instructions during navigation, while outdoor agents do not pay much attention to the object tokens.
- Indoor agents will take direction tokens into consideration when navigating, while outdoor agents heavily rely on direction tokens when making decisions.
- Object tokens at different positions are almost equally important to indoor navigation agents.

The Effect of Environment Objects



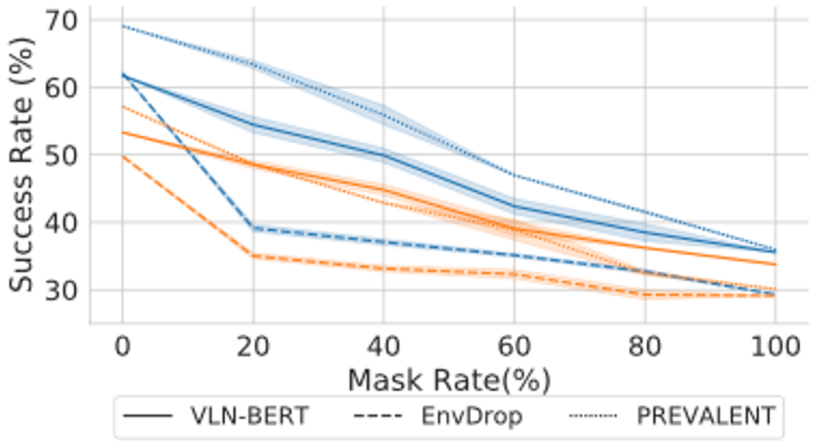
(a) Visual objects within 3m



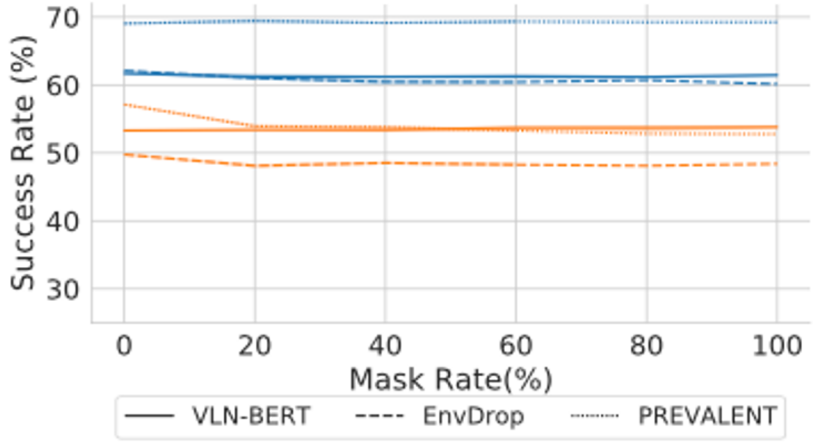
(b) All visual objects

The Effect of Environment Objects

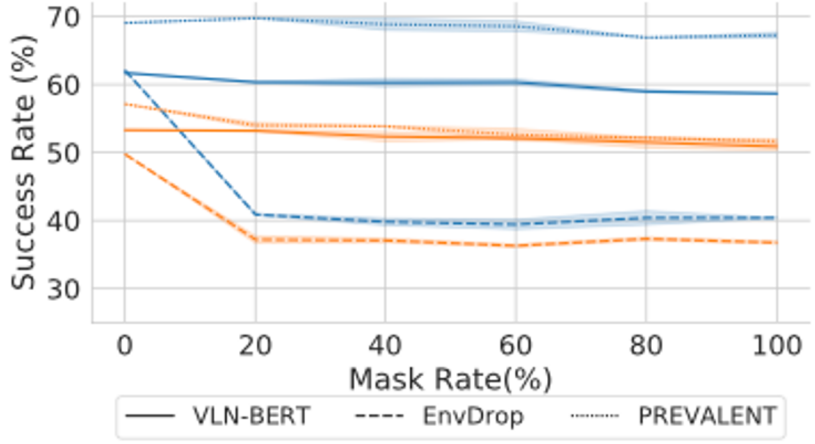
Blue lines: results on validation seen set
Orange lines: results on validation unseen set



(a) Mask Objects Regardless of Distance



(b) Mask Objects Within 3 meters

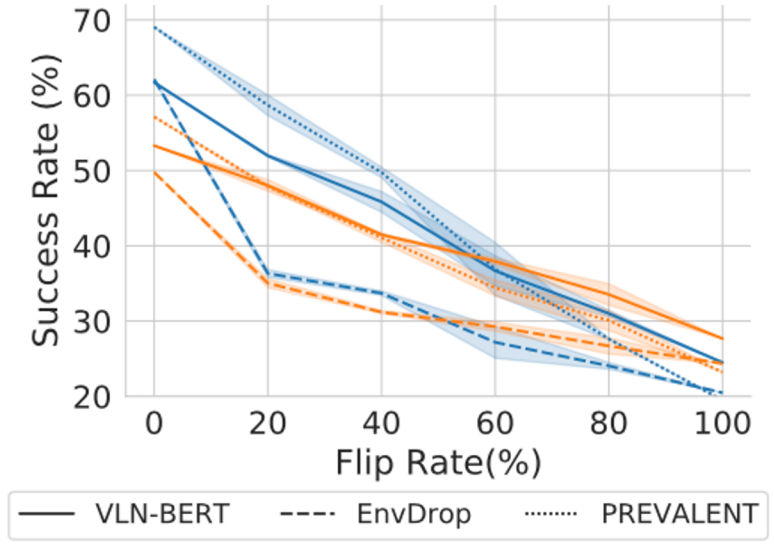


(c) Mask the same number of Objects as (b) but Regardless of Distance

Blue lines: results on validation seen set
Orange lines: results on validation unseen set

The Effect of Directions in the Environment

- Randomly flip some of the viewpoints horizontally in R2R.

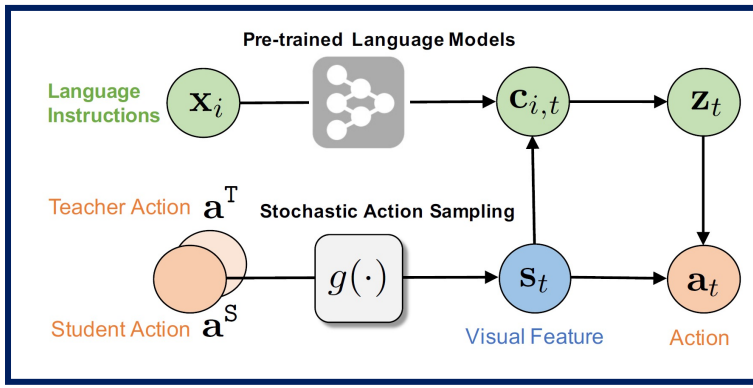


Takeaways

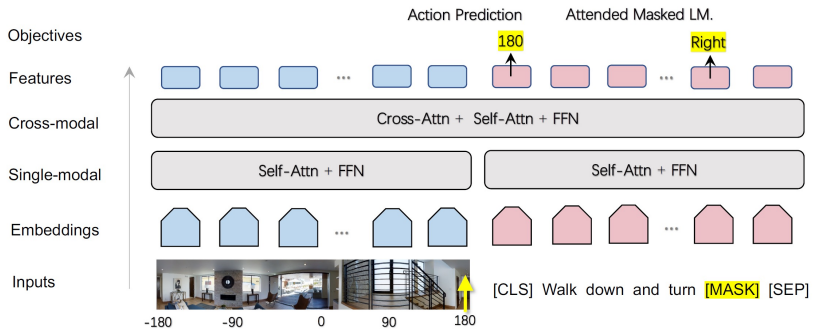
- Instead of merely staring at surrounding objects, indoor agents can set their sights on objects further from current viewpoint.
- Aside from recognizing each individual object, Transformer-based agents also form a better understanding of the global features.

Trend

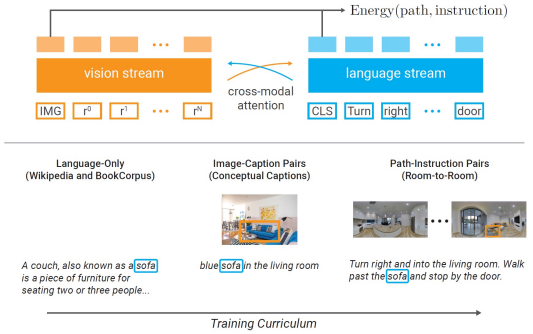
PRESS (Li et al., 2019)



PREVALENT (Hao et al., 2020)



VLN-BERT (Majumdar et al., 2020)

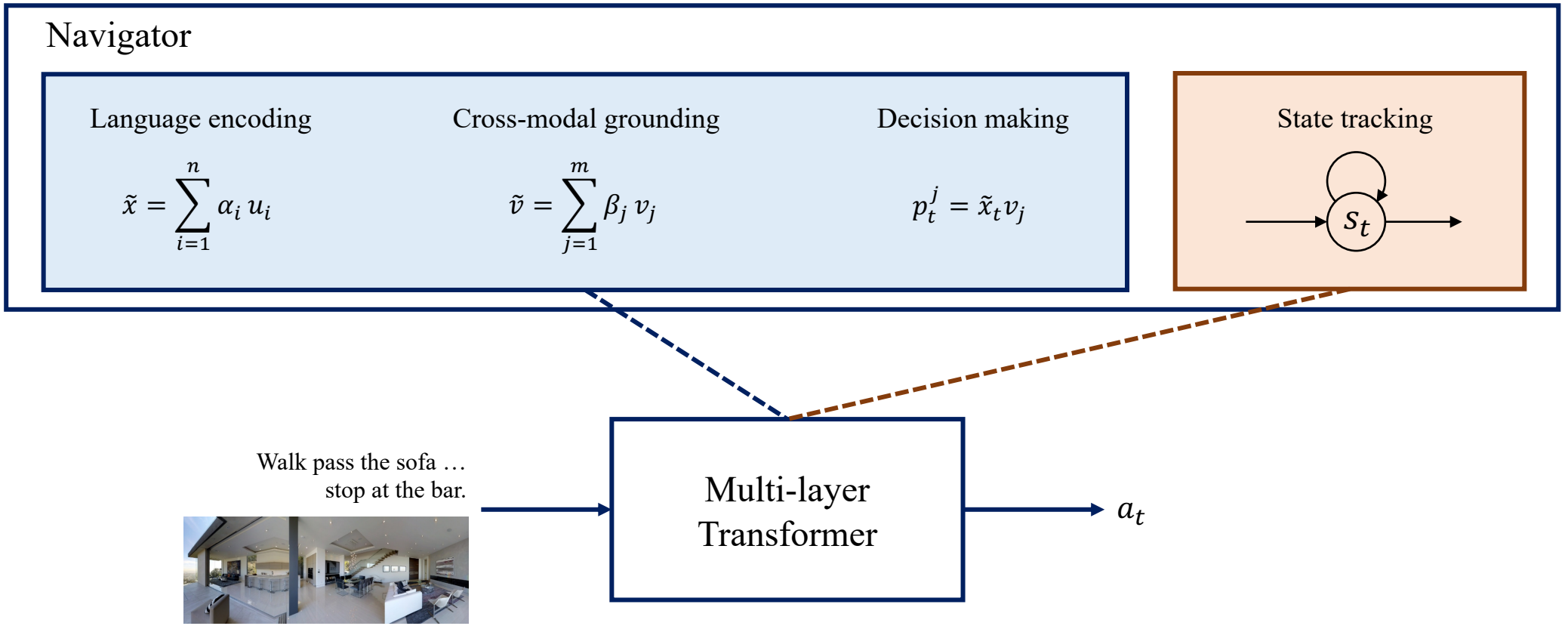


Walk pass the sofa …
stop at the bar.

V&L Encoder → Navigator Network → a_t

Previous Vision-Language BERT models for VLN.

- Towards Transformer-based navigator

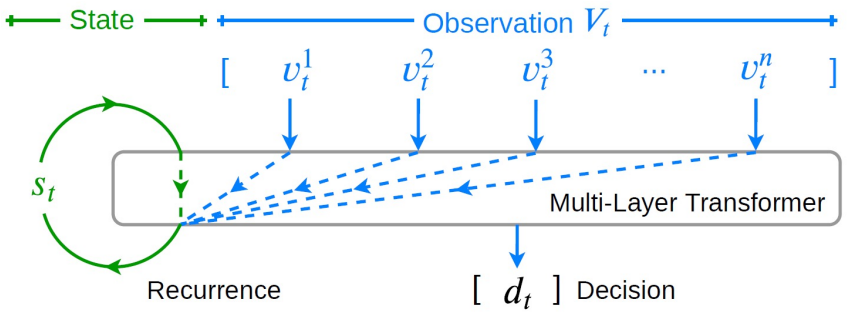


Towards Transformer-based navigator.

VLN↻BERT: A Recurrent Vision-and-Language BERT for Navigation

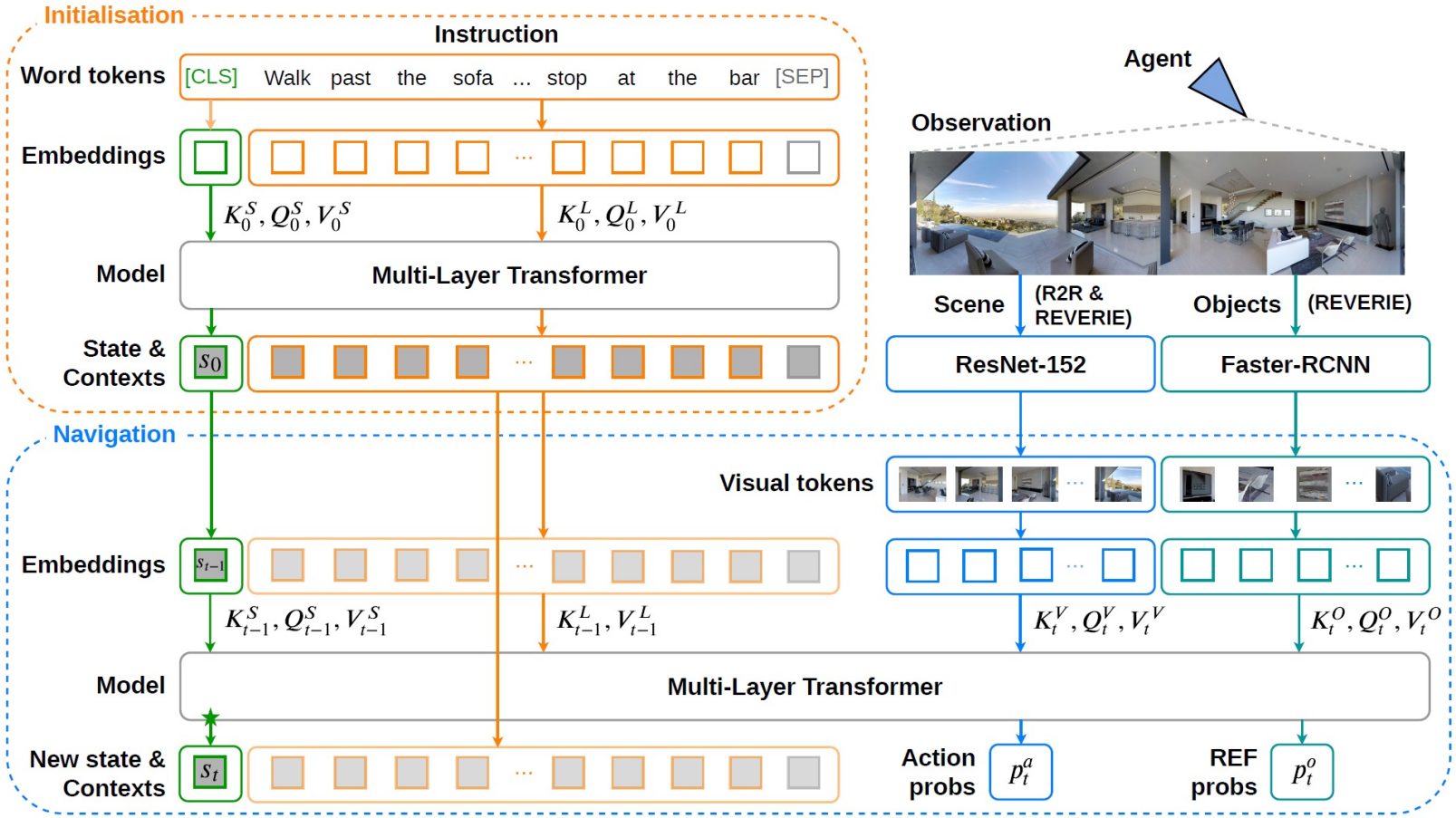
Yicong Hong[1] Qi Wu[2] Yuankai Qi[2] Cristian Rodriguez-Opazo[1,2] Stephen Gould[1]
[1]The Australian National University [2]The University of Adelaide
[1,2]Australian Centre for Robotic Vision
{yicong.hong, cristian.rodriguez, stephen.gould}@anu.edu.au
qi.wu01@adelaide.edu.au, qykshr@gmail.com

Project URL: https://github.com/YicongHong/Recurrent-VLN-BERT



Monday, June 21, 2021 10:00 PM – 12:30 AM
Paper ID:784

VLN↻BERT



Recurrent Vision-and-Language BERT

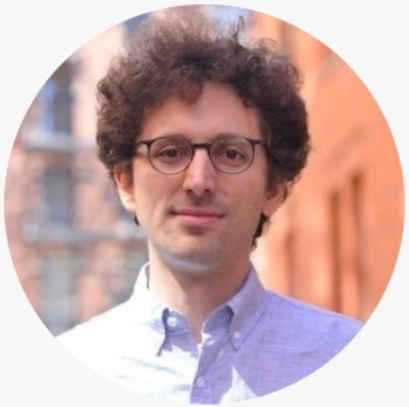
Tutorial Speakers

Qi Wu

University of Adelaide

Xin (Eric) Wang

UC Santa Cruz



Yoav Artzi

Cornell University

Peter Anderson

Google Research