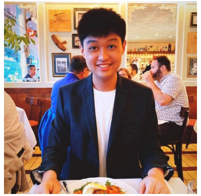
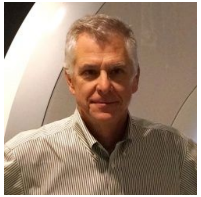
Counterfactual Vision-and-Language Navigation via Adversarial Path Sampling



Tsu-Jui Fu

Xin Wang

Matthew Peterson



Scott Grafton

Miguel Eckstein

William Wang

UC Santa Barbara

Vision-and-Language Navigation (VLN)

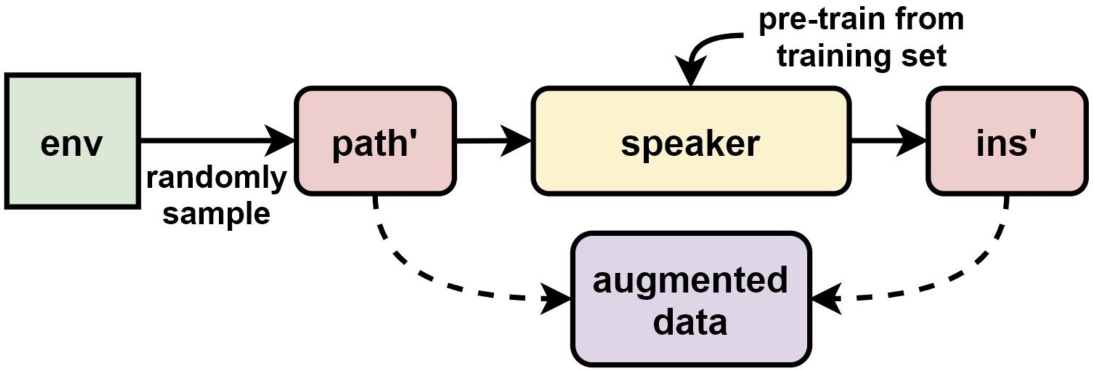
- Achieve the goal based on the instruction in a room
  - learns to align the **linguistic semantic** and **visual understanding**

- Difficult to collect (instruction, path) pairs
  - the **data scarcity** makes learning the optimal match challenging

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.

[NeurIPS’18] Data Augmentation with Speaker

- Expand the training set
  - a **speaker** to back-translate path into instruction
  - randomly sample paths as augmented data

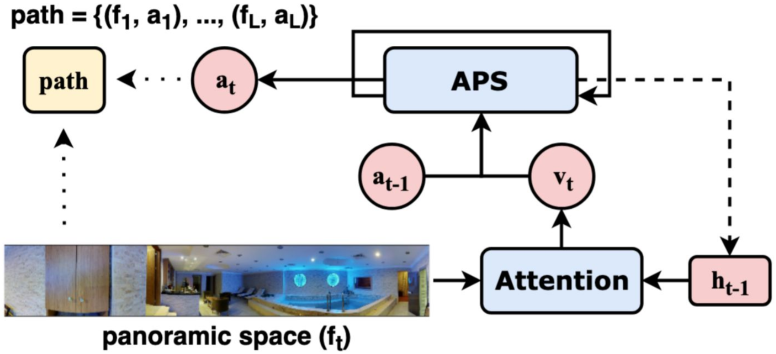


  - however, the **help is limited** since the augmented path are arbitrary

https://arxiv.org/abs/1806.02724

Adversarial Path Sampling (APS)

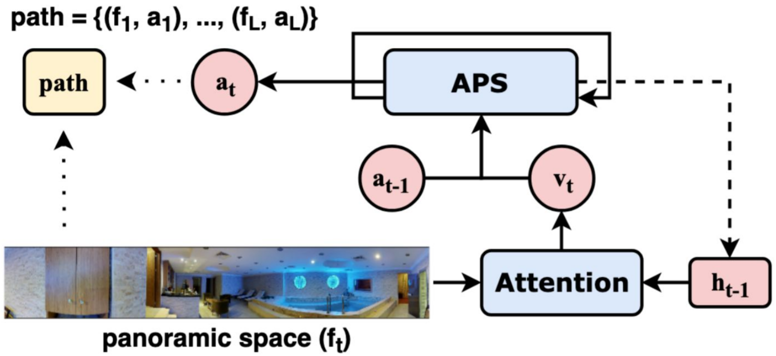
- To make the sampled path more useful
  - APS learns to sample **challenging paths** that NAV cannot navigate easily
  - NAV tries to solve the paths from APS



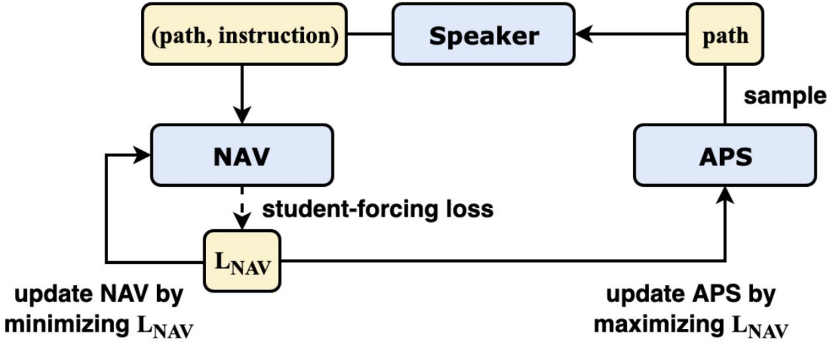
Adversarial Path Sampler (APS)

Adversarial Path Sampling (APS)

- To make the sampled path more useful
  - APS learns to sample **challenging paths** that NAV cannot navigate easily
  - NAV tries to solve the paths from APS



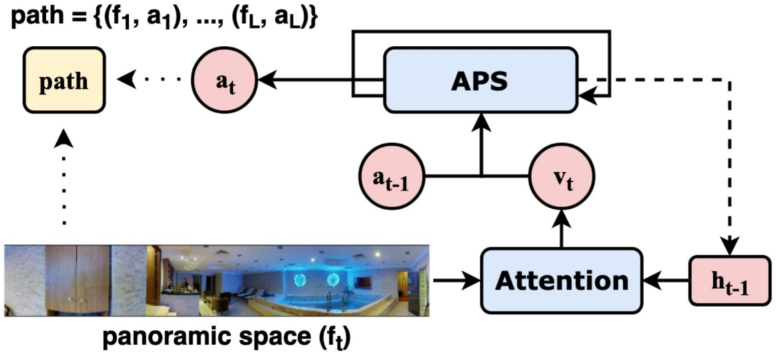
Adversarial Path Sampler (APS)



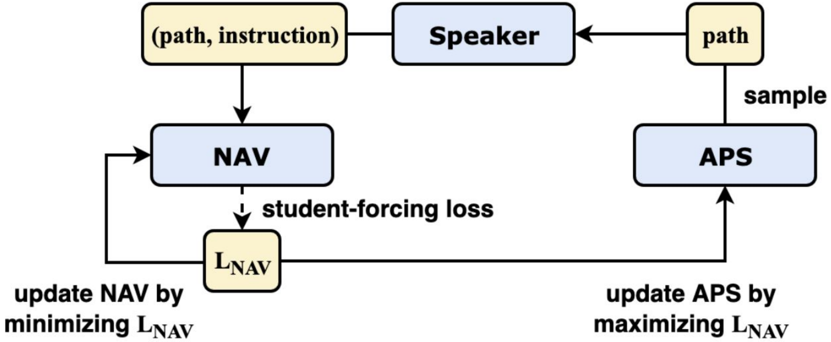
Adversarial Training

Adversarial Path Sampling (APS)

- To make the sampled path more useful
  - APS learns to sample **challenging paths** that NAV cannot navigate easily
  - NAV tries to solve the paths from APS



Adversarial Path Sampler (APS)

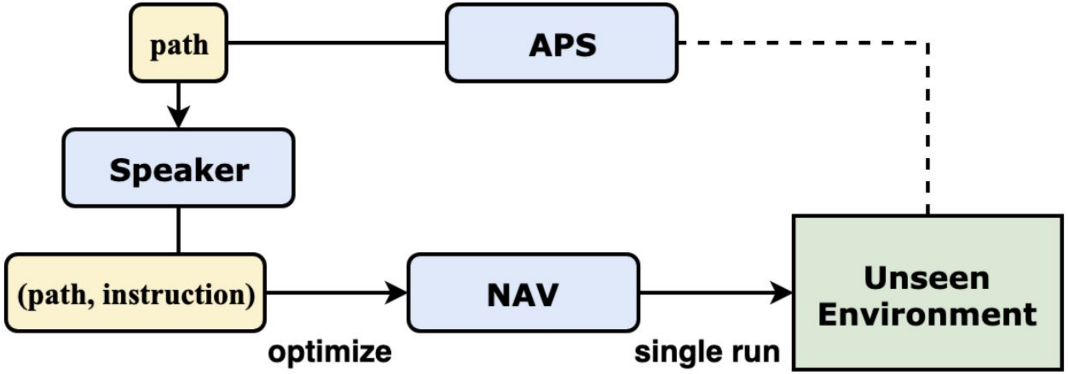


Adversarial Training

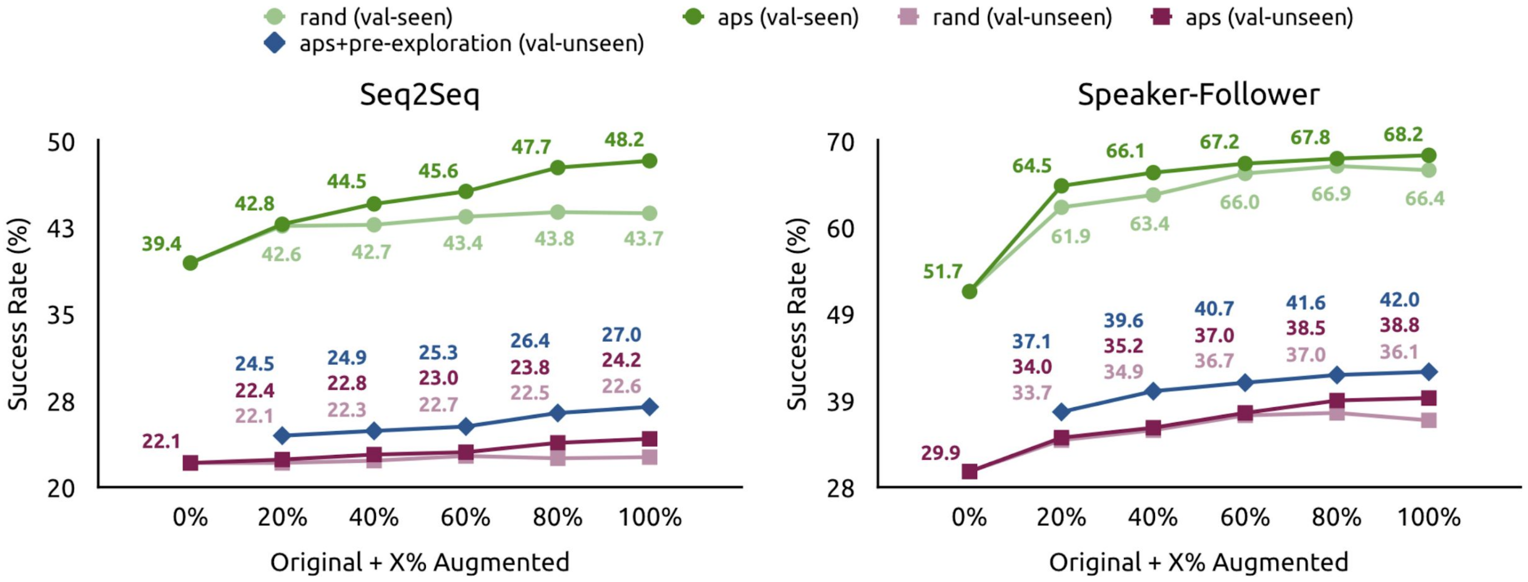
$$\min_{\text{NAV}} \max_{\text{APS}} \mathcal{L}_{\text{NAV}}.$$

Pre-Exploration with APS

- Under **unseen environments**, we can do **pre-exploration** to make NAV more robust
    - use APS to sample paths and optimize NAV for unseen adaption
    - then, NAV can run each instruction in a **single turn**

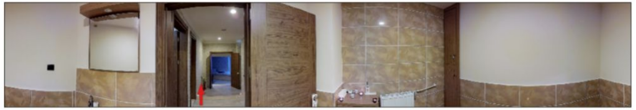


Result

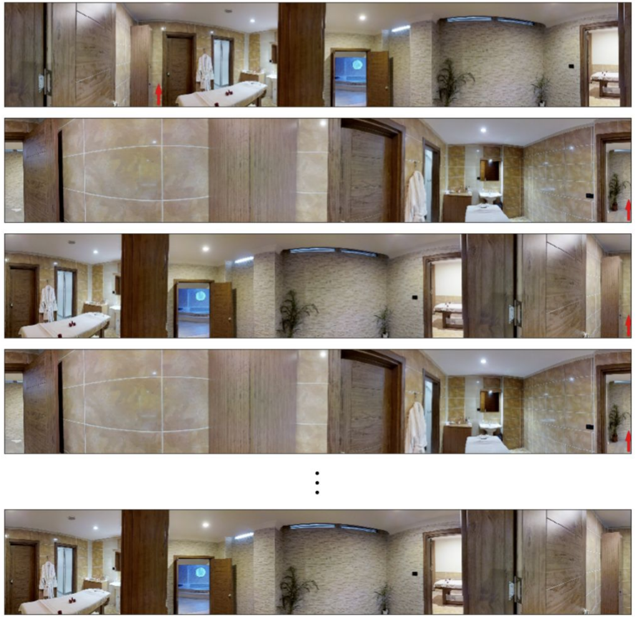


- Randomly sampled **stop improving** when using more than **60%**
- **APS sampled** helps both seen and unseen environments
- **Pre-Exploration** further helps unseen environments

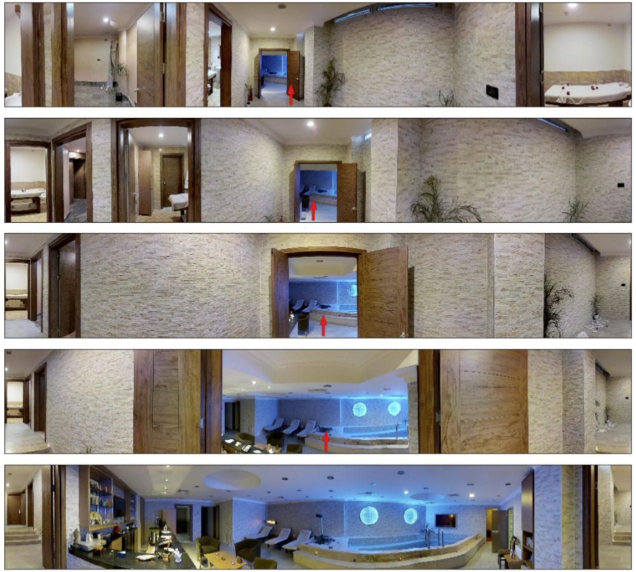
Result



*Walk **out of the bathroom** and straight across the hall.*
Walk down the steps and stop.



(a) without Pre-Exploration



(b) with Pre-Exploration