Evolving Graphical Planner: Contextual Global Planning for Vision-and-Language Navigation

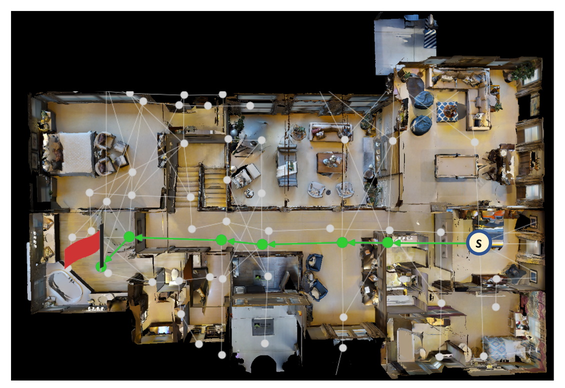
Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky

Princeton University

Overview

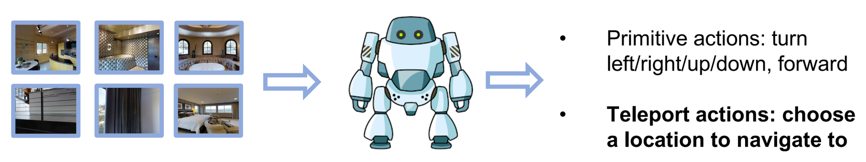
Problem setting

Goal: building an autonomous agent that navigates through an unseen environment by following instructions and dynamically planning the path to reach the goal location.



Turn around and exit the bedroom. Walk along the corridor and keep straight. Walk pass the sofa and the painting on the bedroom wall. Enter the bathroom and stop in front of the tub.

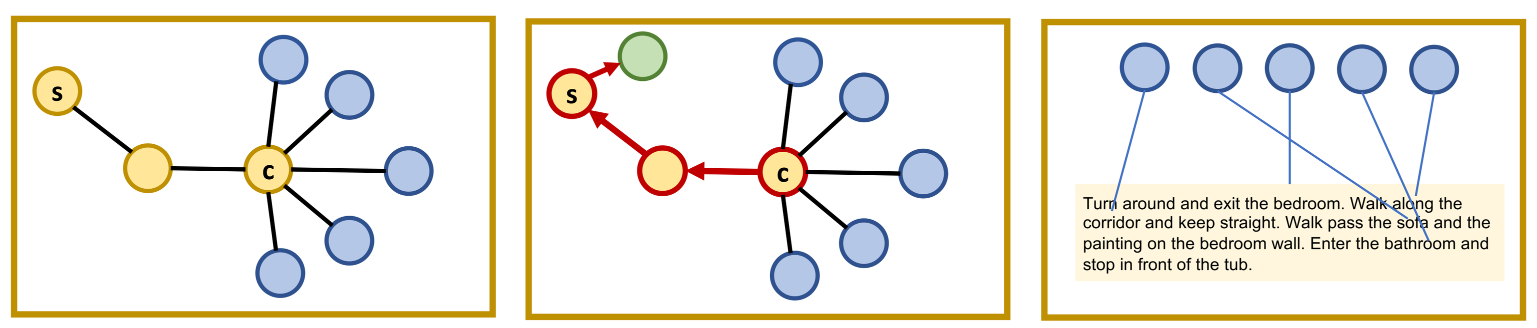
Discrete topological connections



Decision making with instructions and observations

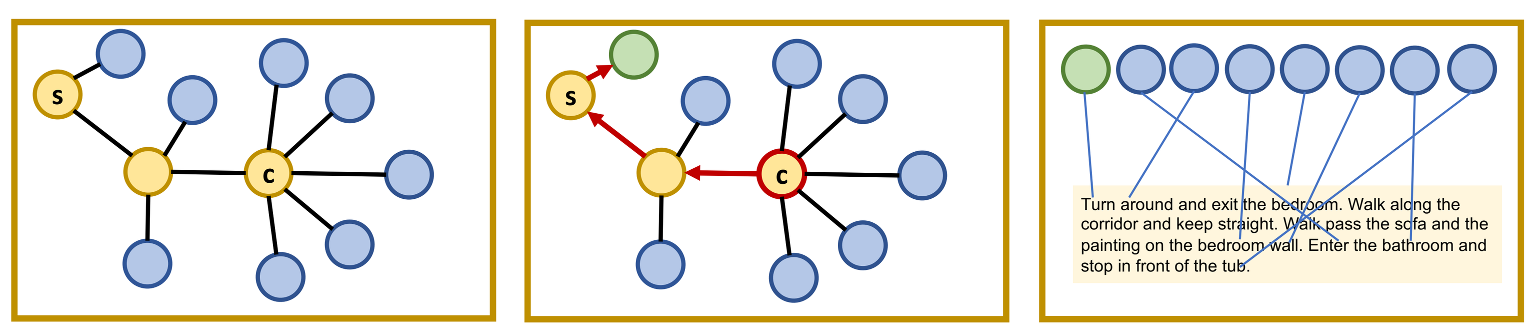
Our navigation agent

Standard navigation – local decision space



Local decision making | Multi-step error correction | Limited observation features for grounding

Evolving Graphical Planner – global decision space



Global decision making | Single-step error correction | Full observation features for grounding

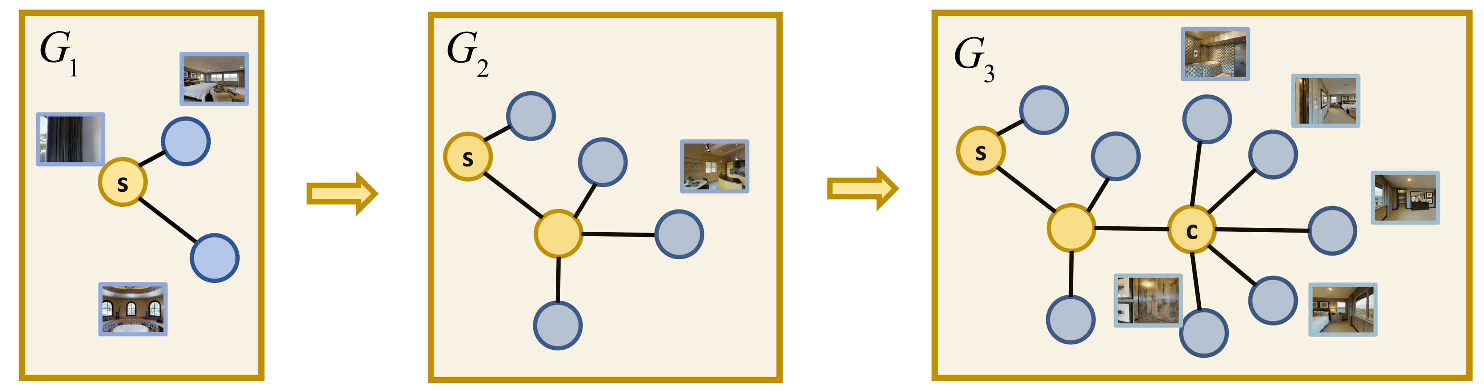
Contributions

- We propose a differentiable Evolving Graphical Planner that expands the **decision-making** process to a **global space**
- The planner can achieve **efficient planning** over the ever-expanding graph memory along with the navigation
- A new **graph-based imitation supervision** is proposed to alleviate the mismatch issue in student sampling training
- We show superior performance compared to previous backbone navigation architectures

Evolving Graphical Planner

Graphical memory

The EGP keeps a dynamic graphical memory that stores the **raw observations** of each location. The connectivity of nodes is determined by the topological connections of the environments.



Step 1 | Step 2 | Step 3

$G_t = (V_t, E_t)$ $\quad v_t^i = (visual_t^i, angle_t^i)$ $\quad e_t^{ij} = (v_t^i, v_t^j)$

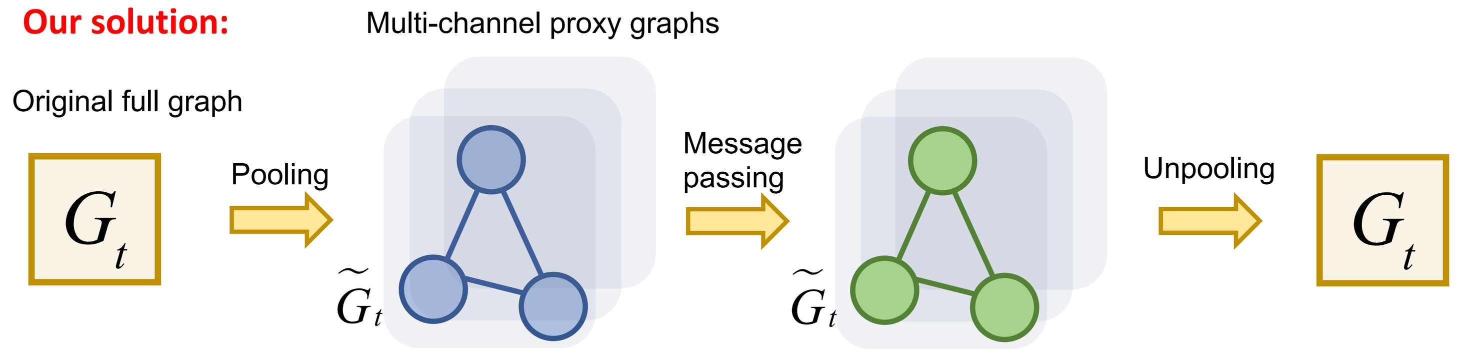
Proxy graphs

The ever-expanding sizes of the graphical memory will lead to two problems:

(1) The memory cost grows rapidly (memory issue);

(2) The topological long-range nodes make communication harder (performance issue)

Our solution:



Pooling matrix: $A_t = GNN(G_t)$ $\quad \widetilde{V}_t = A_t^T V_t$ $\quad \widetilde{E}_t = A_t^T E_t A_t$ $\quad$ Proxy graphs: $\widetilde{G}_t = (\widetilde{V}_t, \widetilde{E}_t)$

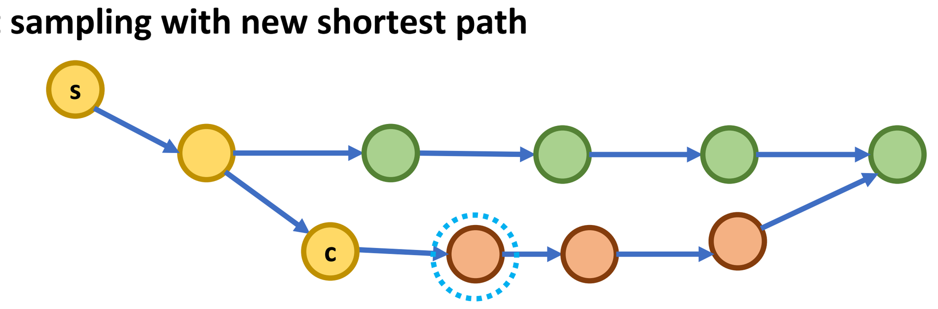
Graph-augmented imitation supervision

Current widely used supervision strategy for student sampling in training navigation imitation agent requires re-computing a new shortest path to the goal location. This leads to:

(1) A potential mismatch issue between the new path and the given instruction;

(2) The need to access more information in the environment

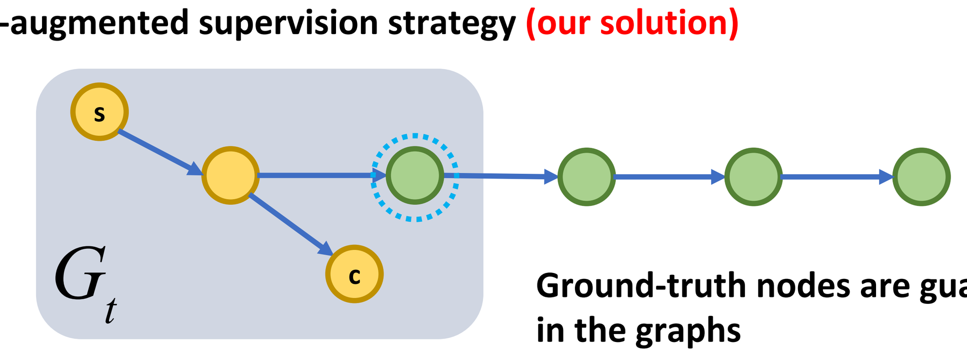
Student sampling with new shortest path



… Walk pass the sofa and the painting on the bedroom wall. Enter the bathroom and stop in front of the tub.

Blue circled nodes are used in the loss function as the supervision

Graph-augmented supervision strategy (our solution)



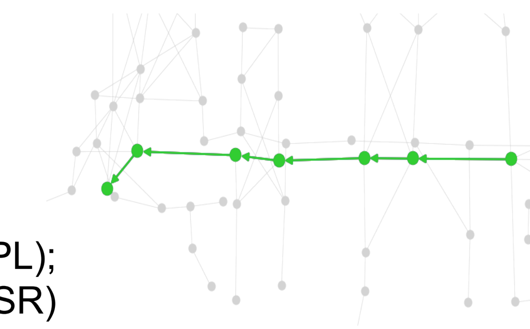
Ground-truth nodes are guaranteed to exist in the graphs

Experiments

Room-to-room (R2R)

R2R is a Matterpord3D-based 3D photorealistic dataset with human generated instructions as guidance for navigation. Paths are generated by shortest-path algorithm

Metrics: success rate(SR); navigation error(NE); path length(PL); success rate per path length unit(SPL); oracle success rate(OSR)



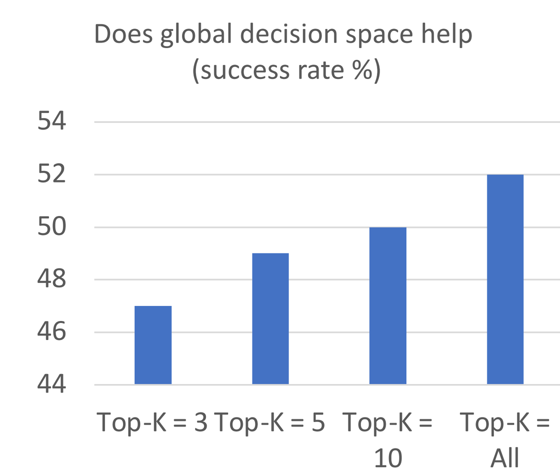
Compare with previous navigation backbone architectures

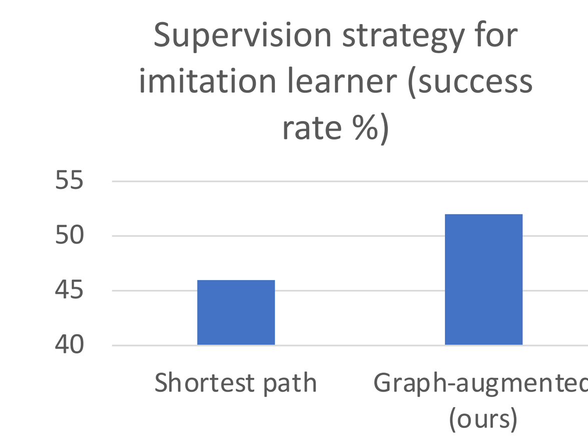
Models	Type	Val Unseen NE ↓	Val Unseen SR% ↑	Val Unseen SPL% ↑	Val Unseen OSR% ↑	Test NE ↓	Test SR% ↑	Test SPL% ↑	Test OSR% ↑
SF* [1]	IL	6.62	36	-	45	6.62	35	28	44
RCM* [2]	IL+RL	5.88	43	-	52	6.12	43	38	50
Monitor* [3]	IL	5.52	45	32	56	5.67	48	35	59
Regretful* [4]	IL	5.32	50	41	59	5.69	48	40	56
FAST* [5]	IL	4.97	56	43	-	5.14	54	41	-
Baseline agent	IL	6.20	43	36	52	-	-	-	-
EGP (ours)	IL	5.34	52	41	**65**	-	-	-	-
EGP* (ours)	IL	**4.83**	**56**	**44**	64	**5.34**	**53**	**42**	**61**

Compared to FAST, EGP:

1. **doesn't need hand-crafted search procedure, extra info (speaker, self-monitor, etc.)**
2. **Is differentiable and have much shorter path length**

The contribution of each component





Room-for-room (R4R)

R4R is a Matterpord3D-based 3D photorealistic dataset that extends R2R by concatenating two paths to create long and twisted paths for testing path following.

Metrics: success rate(SR); navigation error(NE); path length(PL); dynamic time warping(DTW); coverage weighted by length score(CLS)

Models	Type	PL	NE ↓	SR% ↑	CLS↑	nDTW↑	SDTW↑
Speaker-Follower [1]	IL+RL	19.9	8.47	23.8	29.6	-	-
RCM + goal-oriented [6]	IL+RL	32.5	8.45	28.6	20.4	26.9*	11.4*
RCM + fidelity-oriented [6]	IL+RL	28.5	8.08	26.1	34.6	30.4*	12.6*
PTA low-level [7]	IL+RL	10.2	8.19	27.0	35.0	20.0	8.0
PTA high-level [7]	IL+RL	17.7	8.25	24.0	37.0	32.0	10.0
EGP (ours)	IL	18.3	**8.0**	**30.2**	**44.4**	**37.4**	**17.5**

We are the first model that uses pure imitation learning to train and achieves the state-of-the-art performance.

References

[1] Fried, Daniel, et al. "Speaker-follower models for vision-and-language navigation." *NeurIPS*. 2018.
[2] Wang, Xin, et al. "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation." *CVPR*. 2019.
[3] Ma, Chih-Yao, et al. "Self-monitoring navigation agent via auxiliary progress estimation." *ICLR 2019*
[4] Ma, Chih-Yao, et al. "The regretful agent: Heuristic-aided navigation through progress estimation." *CVPR* 2019.
[5] Ke, Liyiming, et al. "Tactical rewind: Self-correction via backtracking in vision-and-language navigation." *ICCV* 2019.
[6] Jain, Vihan, et al. "Stay on the path: Instruction fidelity in vision-and-language navigation." *arXiv preprint arXiv:1905.12255* (2019).
[7] Landi, Federico, et al. "Perceive, Transform, and Act: Multi-Modal Attention Networks for Vision-and-Language Navigation." *arXiv preprint arXiv:1911.12377* (2019).