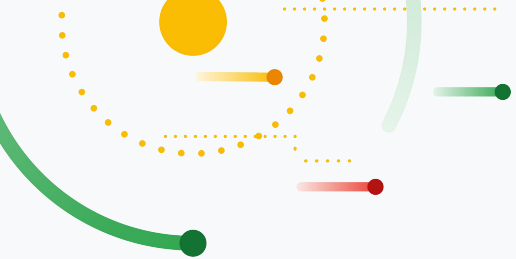


Google AI

Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation

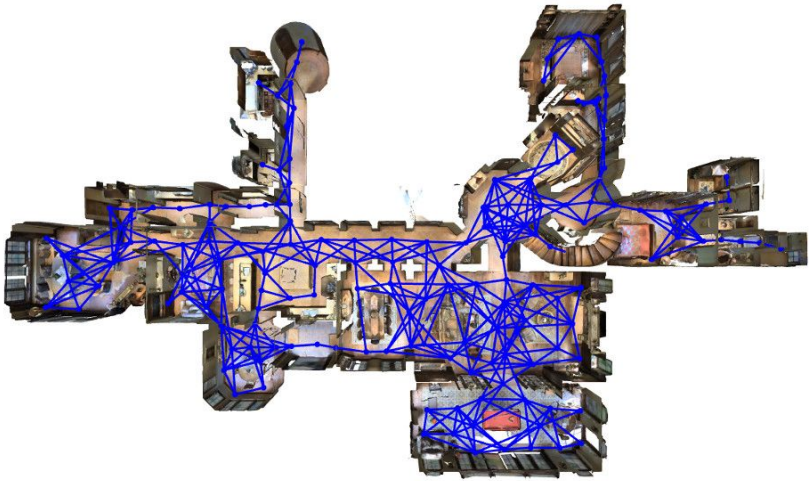
Google Research

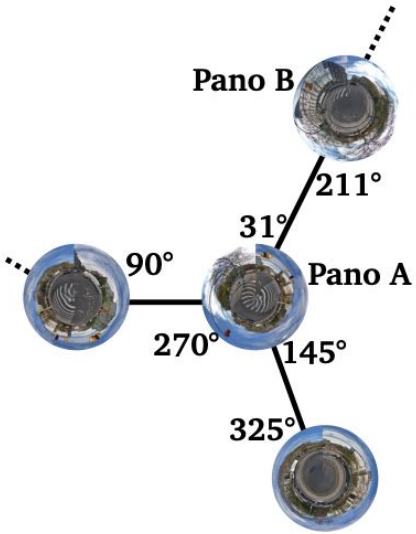
Vihan Jain, **Gabriel Ilharco, Alexander Ku**,
Ashish Vaswani, Eugene Ie, Jason Baldridge

ACL Florence, 29th July 2019

Vision-and-Language Navigation (VLN)

- Language
- Perception
- Action
- Planning





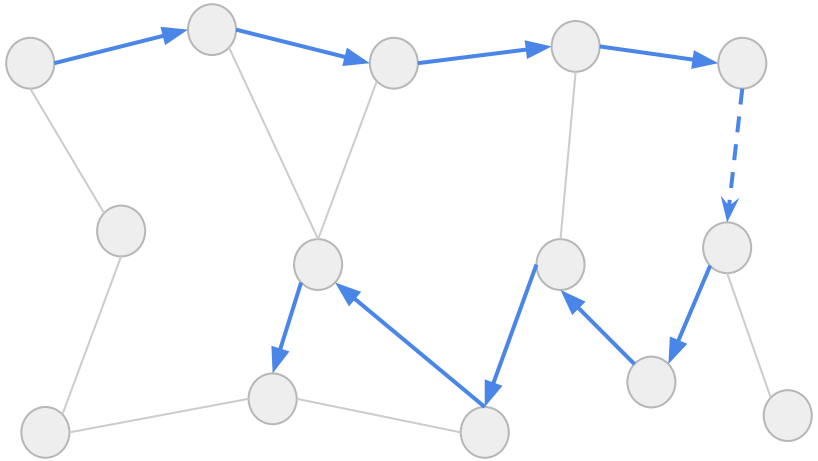
Vision-and-Language Navigation (VLN)

Example from Room-to-Room (R2R)[1] dataset

[1] Anderson et al. *Vision-and-language Navigation: Interpreting visually grounded navigation instructions in real environments*, **CVPR, 2018.**

Key Contributions

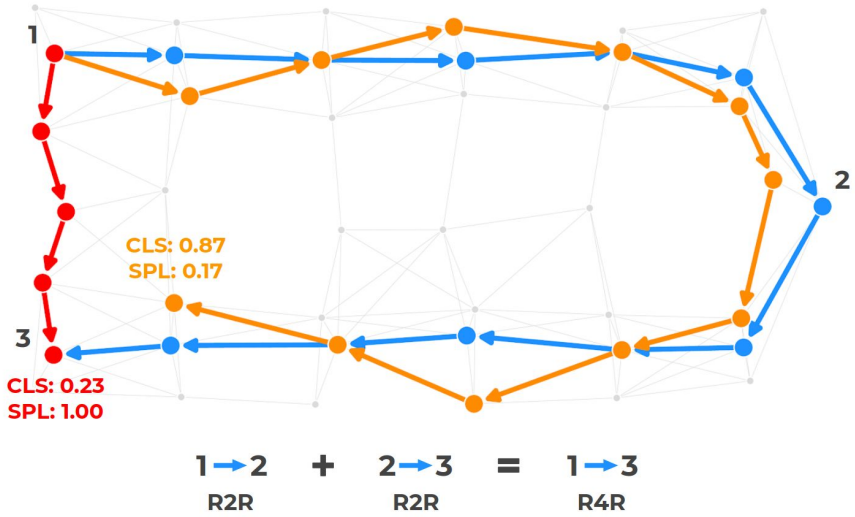
- Data



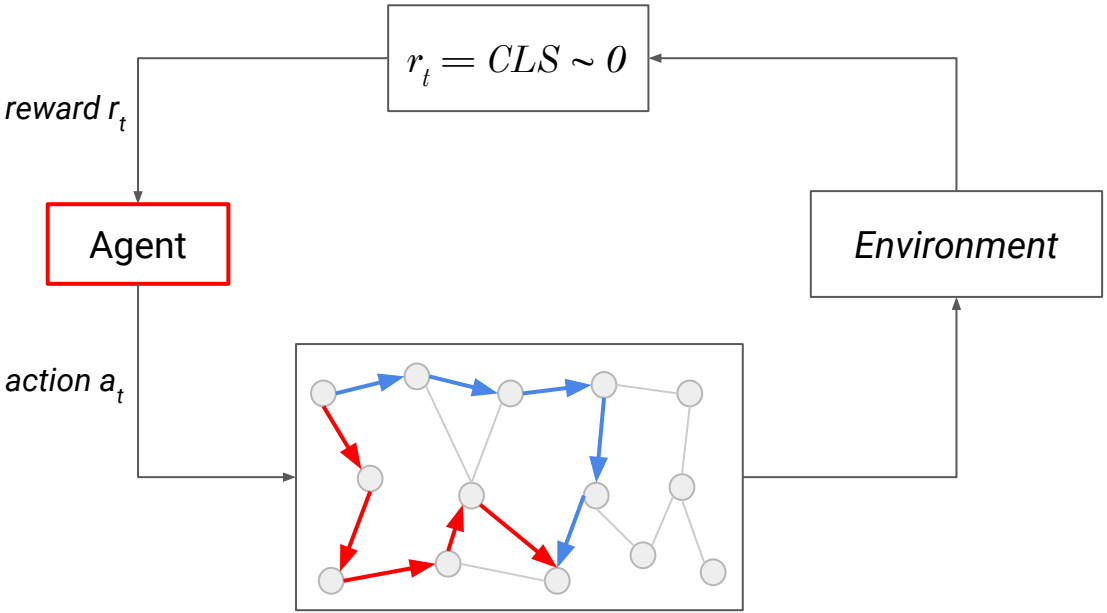
Make a left down at the narrow hall... Go out the door and wait. Turn around and enter the bedroom... Walk into the doorway and stop

Key Contributions

- Data
- Evaluation



Key Contributions

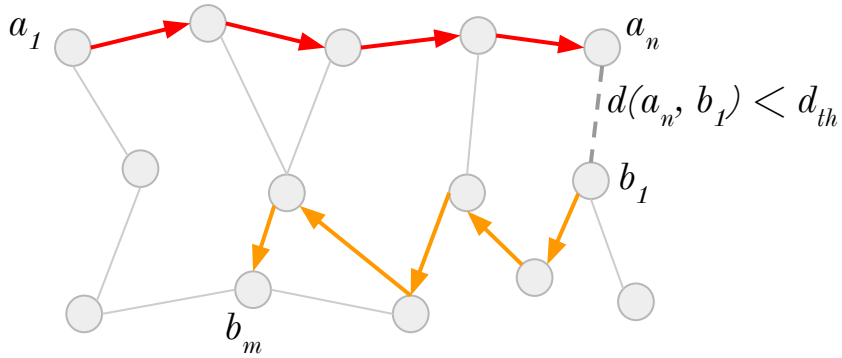


- Data
- Evaluation
- Agent training

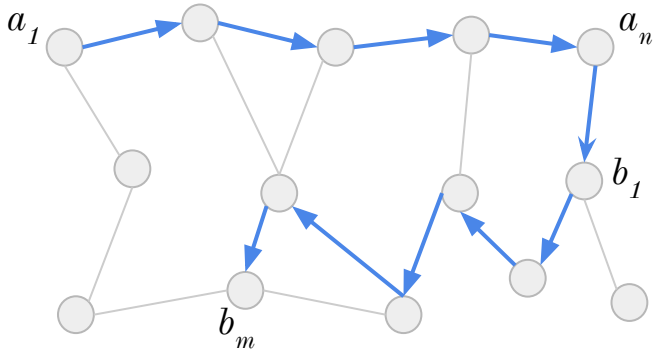
Google AI

R2R → R4R

Make a left down at the narrow hall... Go out the door and wait



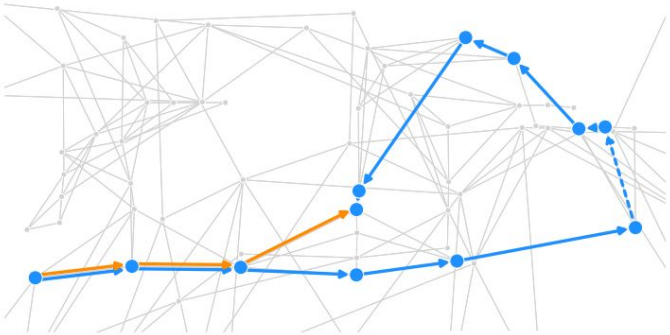
Turn around and enter the bedroom... Walk into the doorway and stop

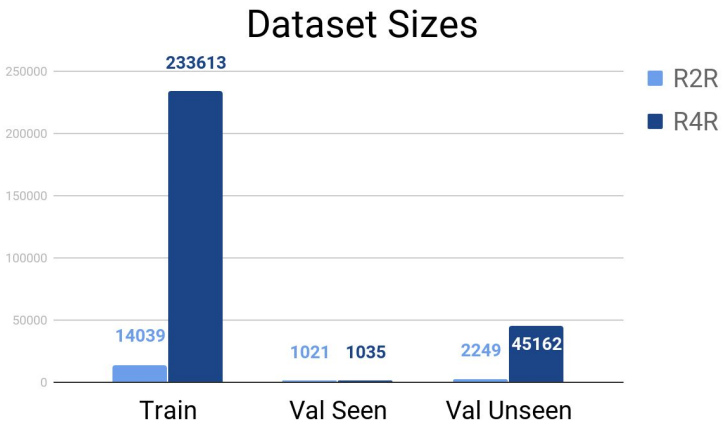


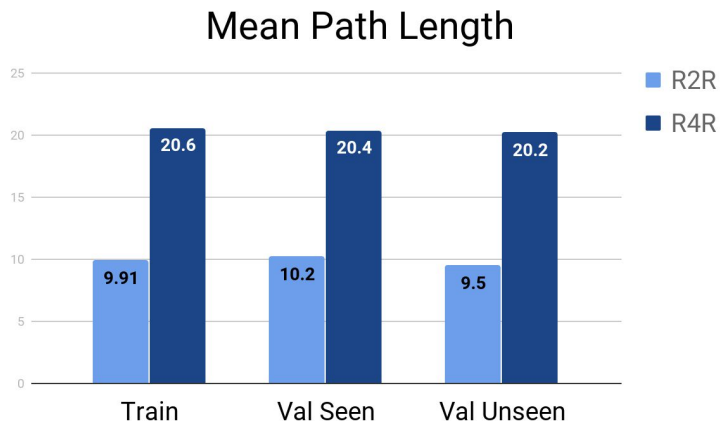
Make a left down at the narrow hall... Go out the door and wait. Turn around and enter the bedroom... Walk into the doorway and stop

R2R-to-R4R code is at https://github.com/googleresearch/google-research/tree/master/r4r

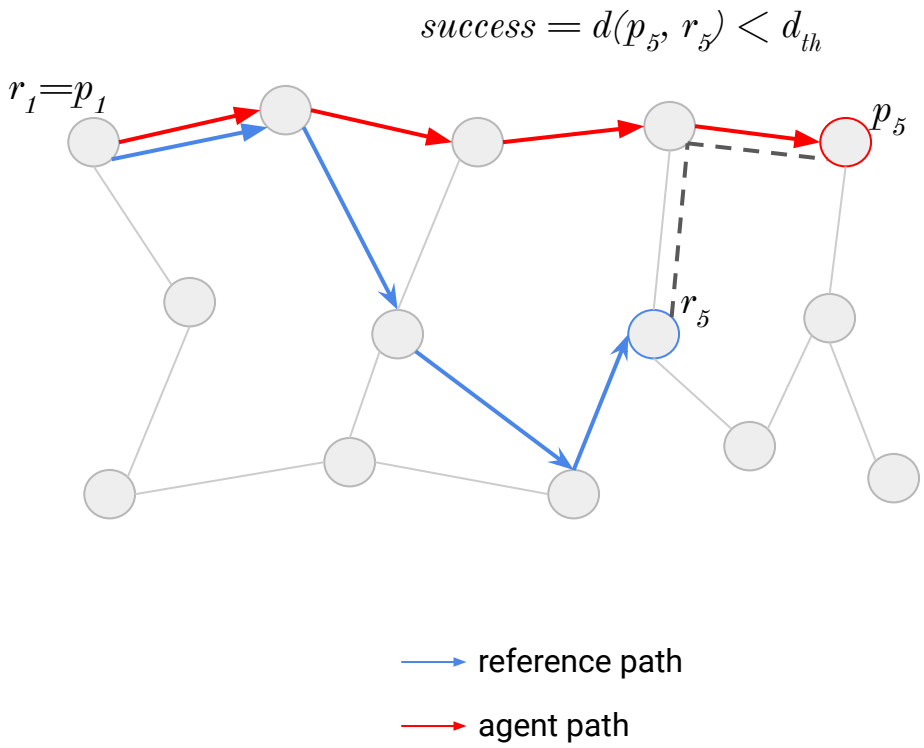
R2R v/s R4R




Dataset Sizes
250000
200000
150000
100000
50000
0
233613
14039
1021
1035
2249
45162
Train
Val Seen
Val Unseen
R2R
R4R

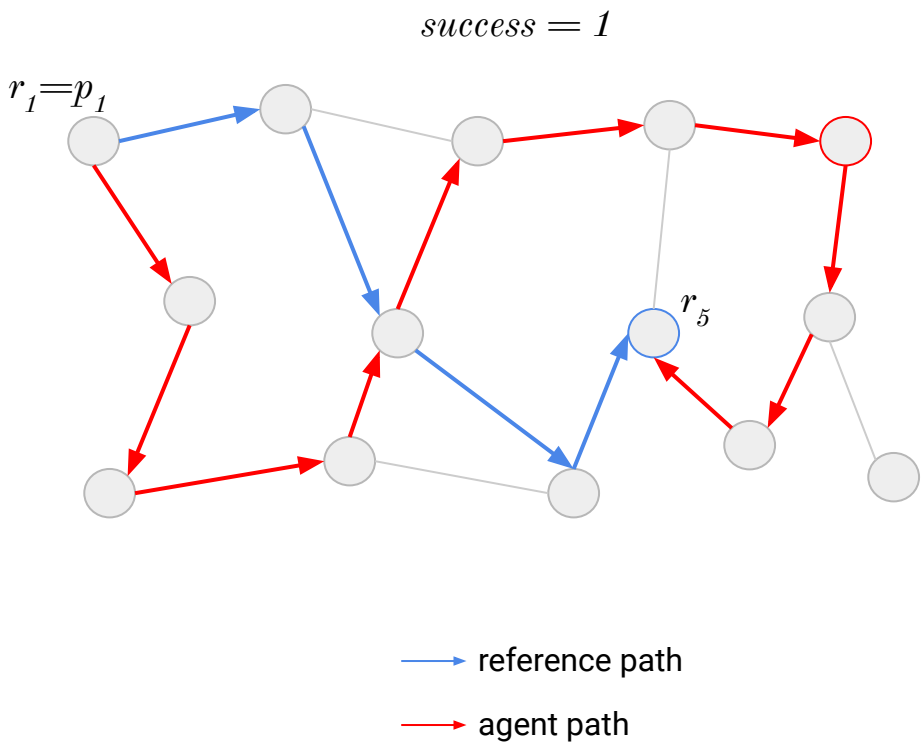

Mean Path Length
25
20
15
10
5
0
20.6
20.4
20.2
9.91
10.2
9.5
Train
Val Seen
Val Unseen
R2R
R4R

VLN Evaluation: Success Rate (SR)



Google AI

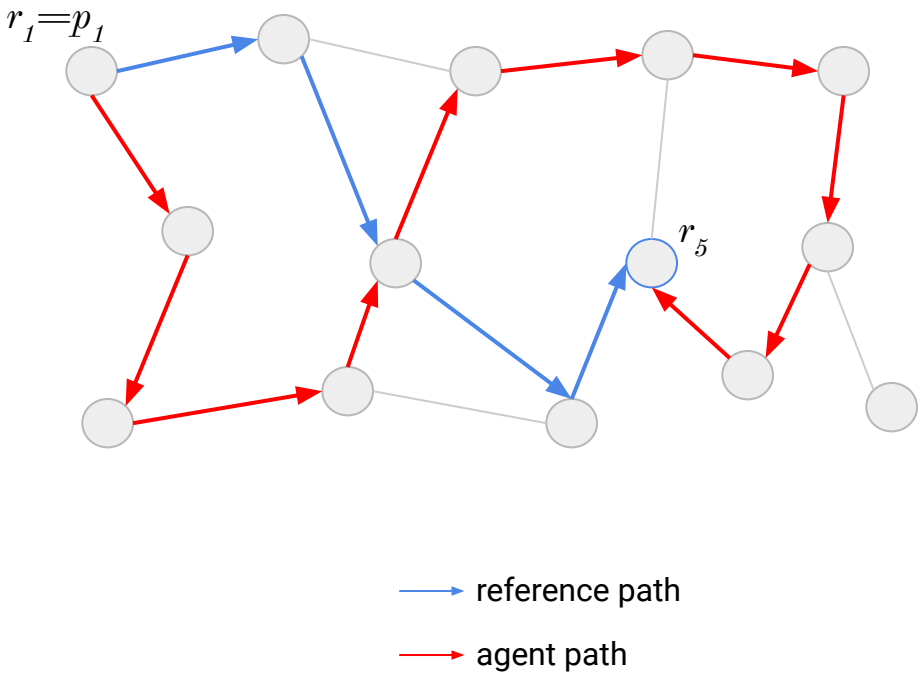
VLN Evaluation: Success Rate (SR)



VLN Evaluation: SPL

Success weighted by Path Length[1]

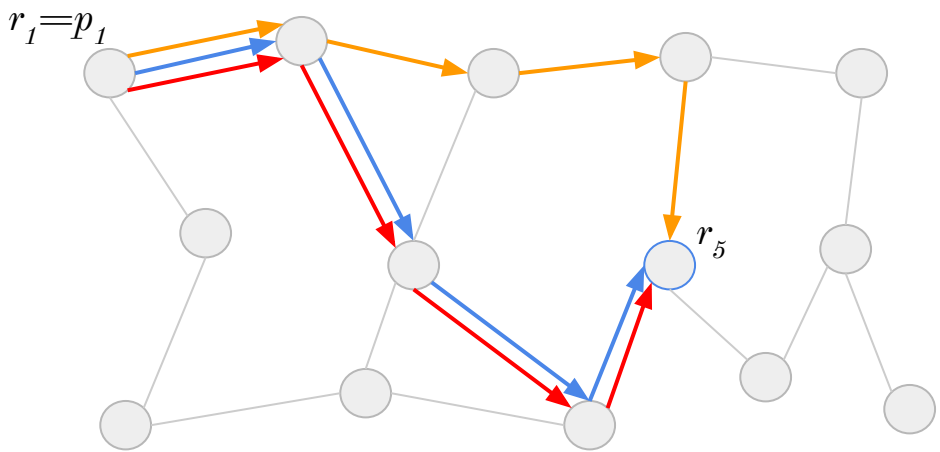
$$spl = 4/10 = 0.4$$



[1] **Anderson et al.** *On Evaluation of Embodied Navigation Agents* **arXiv, 2018.**

VLN Evaluation: SPL

$spl = 1$
$spl = 1$



reference path

agent path 1

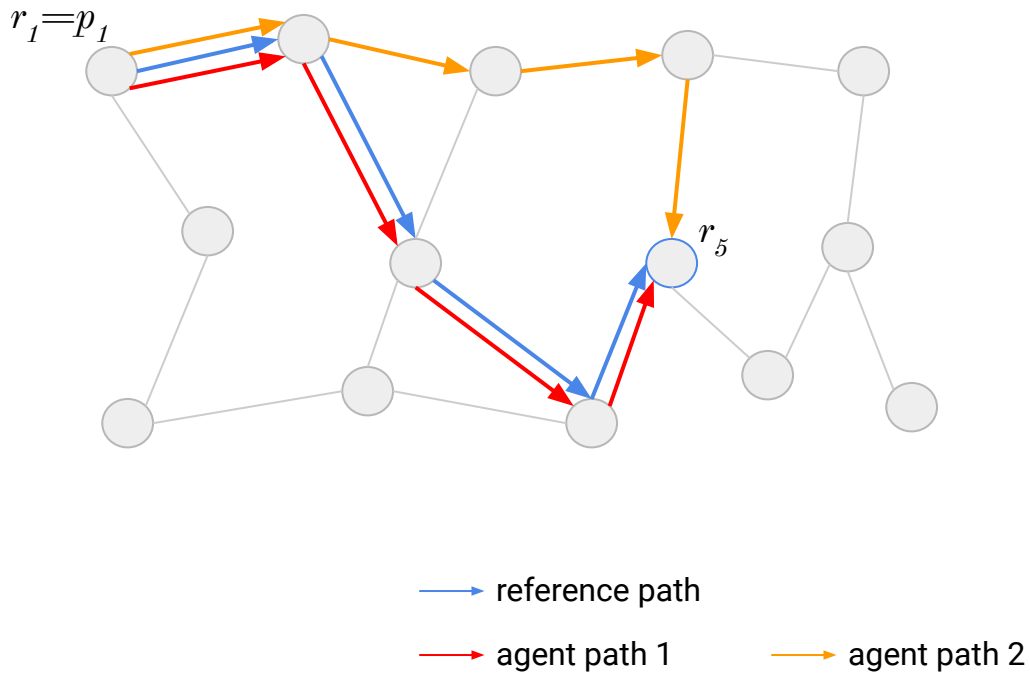
agent path 2

[1] Anderson et al. *On Evaluation of Embodied Navigation Agents* **arXiv, 2018.**

VLN Evaluation: SED

Success weighted by Edit Distance[1]

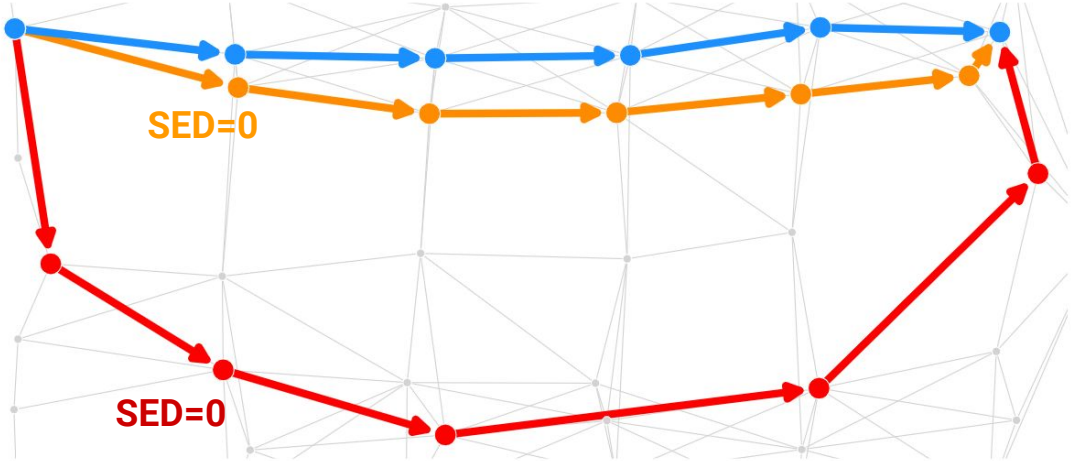
$sed = 1 - 0 = 1$
$sed = 1 - 3/4 = 0.25$



[1] Chen et al. *Touchdown: Natural language navigation and spatial reasoning in visual street environments* **CVPR, 2019**

Google AI

VLN Evaluation: SED



CLS: New VLN Evaluation Metric

- **C**overage weighted by **L**ength **S**core (CLS): product of Path Coverage (*PC*) and Length Score (*LS*)

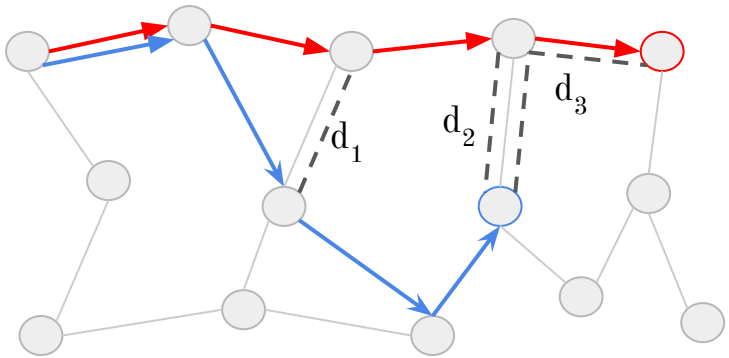
$$\mathrm{CLS}(P, R) = \mathrm{PC}(P, R) \cdot \mathrm{LS}(P, R)$$

R: reference path
P: agent's predicted path

CLS: New VLN Evaluation Metric

- Path Coverage (PC): average coverage of each node in reference path with respect to the predicted path

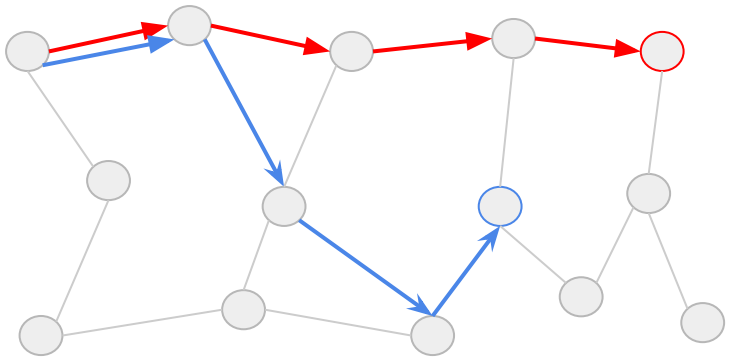
$$PC \propto exp(-\frac{d_1 + d_2 + d_3}{d_{th}})$$



CLS: New VLN Evaluation Metric

- Expected optimal path length (EPL) is a function of path coverage
- Length Score (LS): compares path length of predicted path P to EPL

$$LS \propto \frac{1}{|EPL - PL(P)|}$$



reference path

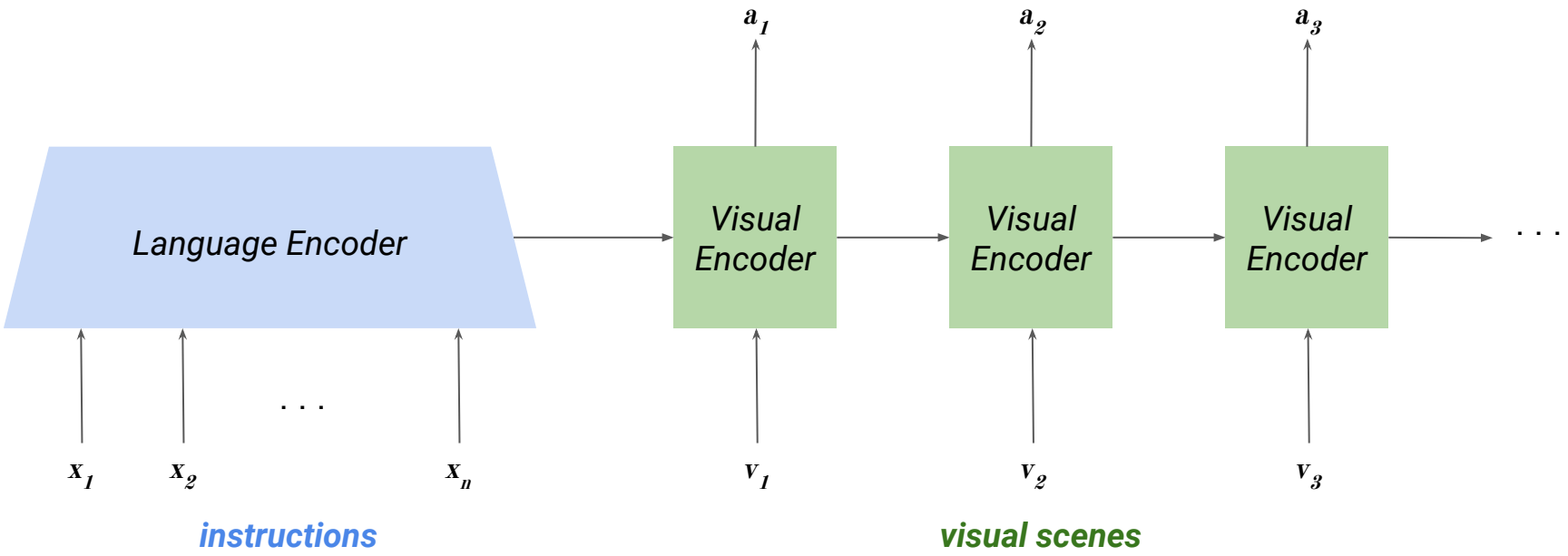
agent's predicted path *P*

CLS: Desirable Properties

	Path Similarity Measure	Soft Penalties	Unique Optimum	Scale Invariance	Tractability
CLS	PC measures how well the predicted path covered the nodes of reference path ✓	Both PC and LS are continuous measures ✓	A predicted path achieves the maximum score if and only if it is equal to reference path ✓	Both PC and LS are invariant due to graph invariant constant d_{th} ✓	Computation Time: PC - $O(\|P\|.\|R\|)$ LS - $O(\|P\|+\|R\|)$ ✓

Training VLN Agents

- Architecture similar to RCM[1] model



[1] **Wang et al.** *Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation* **CoRR, 2018.**

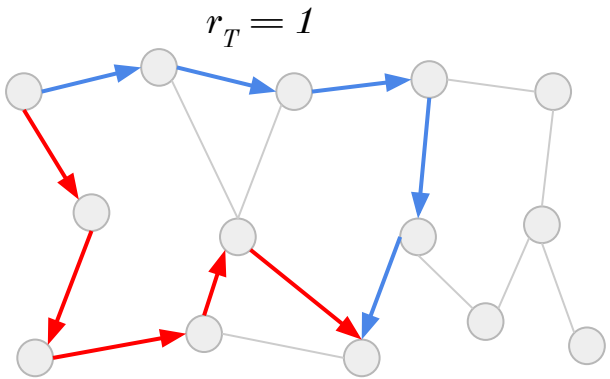
Training VLN Agents

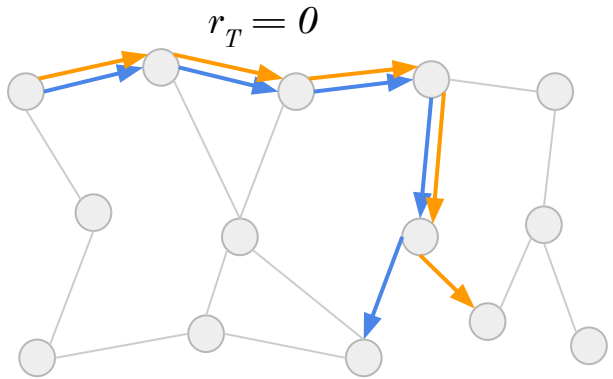
Goal-oriented agents

- encouraged to pursue the goal node only

The immediate reward after taking action a_t at time step t in an episode of length T

$$r(s_t, a_t) = \begin{cases} \textbf{+ve} \text{ if closer to goal, } \textbf{-ve} \text{ otherwise} & \text{if } t < T \\ \textbf{1} \text{ if reached goal, } \textbf{0} \text{ otherwise} & \text{if } t = T \end{cases}$$



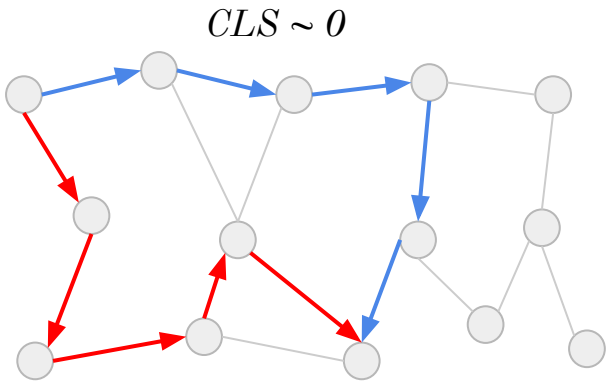


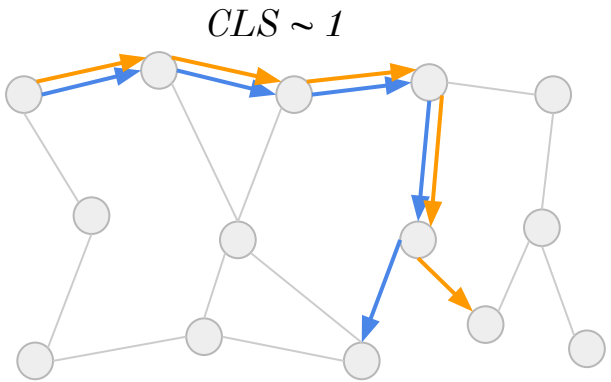
Training VLN Agents

Fidelity-oriented agents

- reach the goal node + **conform to the reference path R**

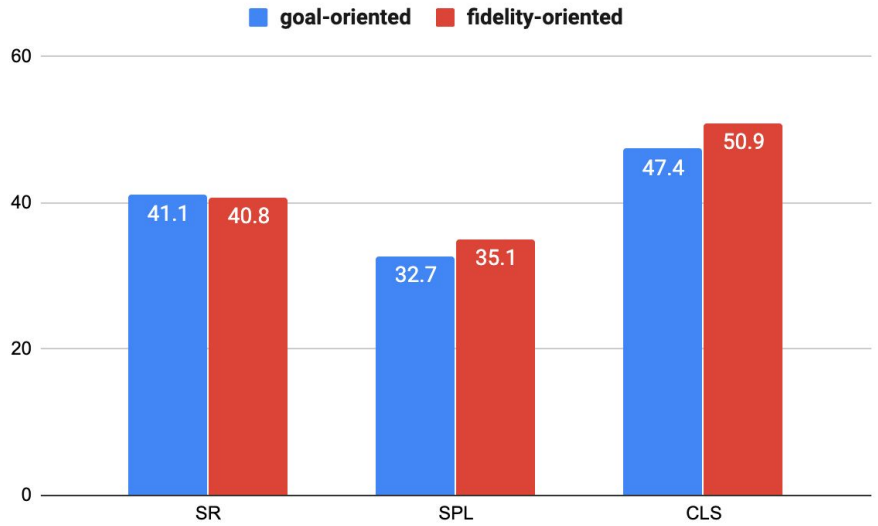
$$r(s_t, a_t) = \begin{cases} 0 & \text{if } t < T \\ (\mathbf{1} \text{ if reached goal, } \mathbf{0} \text{ otherwise }) + CLS(s_{1\ldots T}, R) & \text{if } t = T \end{cases}$$





R2R Performance

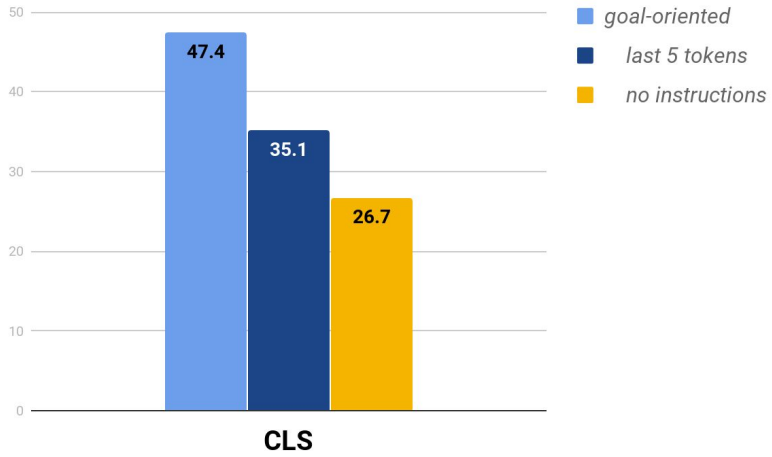
- Fidelity-oriented agents perform slightly better on SPL, CLS
- SPL appears consistent with CLS

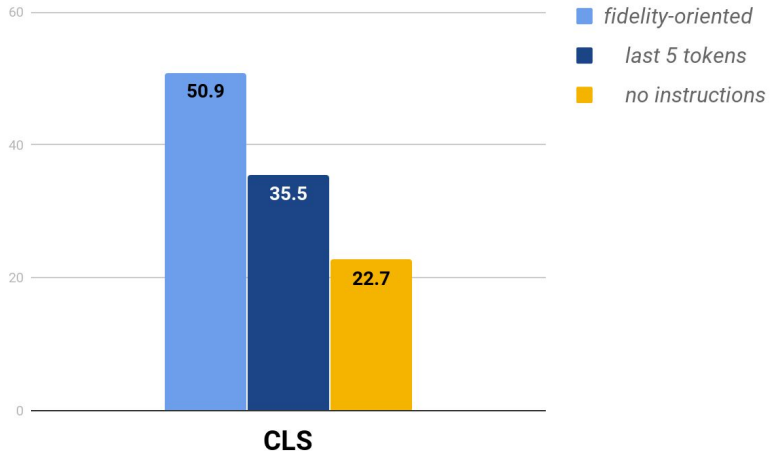


Results on Validation Unseen dataset

R2R Performance

- Ablation Studies
  - Agent optimized to reach the goal may incidentally appear to be conforming to the instructions

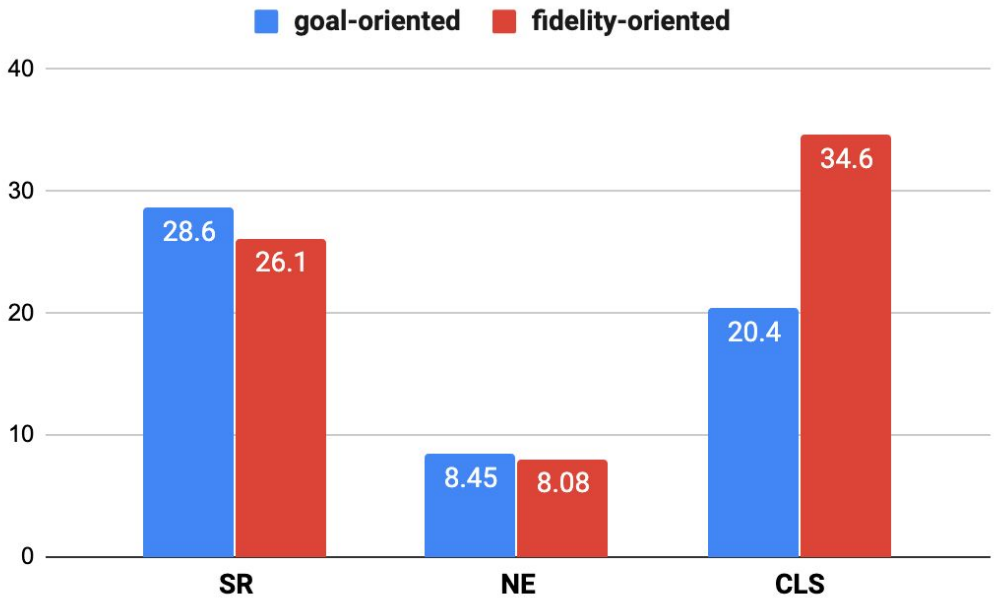




Results on Validation Unseen dataset

R4R Performance

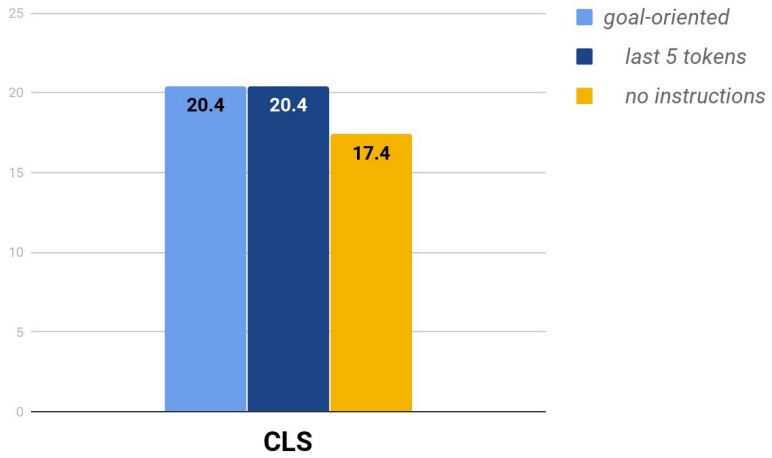
- Fidelity-oriented agents outperform goal-oriented agents

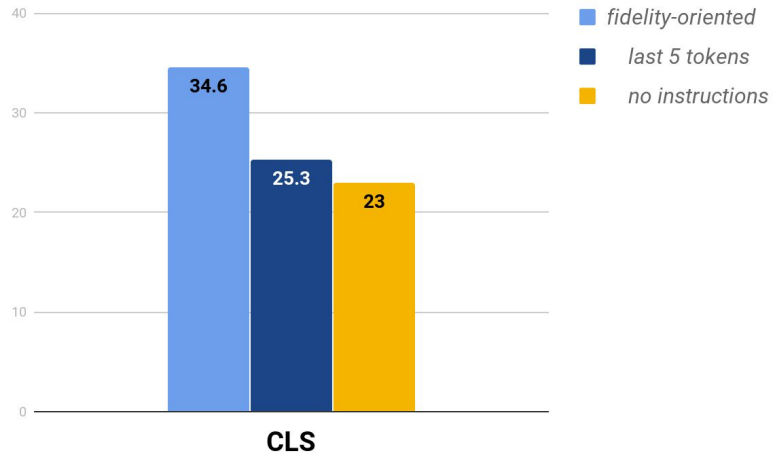


Results on Validation Unseen dataset

R4R Performance

- Ablation Studies
  - Fidelity-oriented agents attend more carefully to the instructions

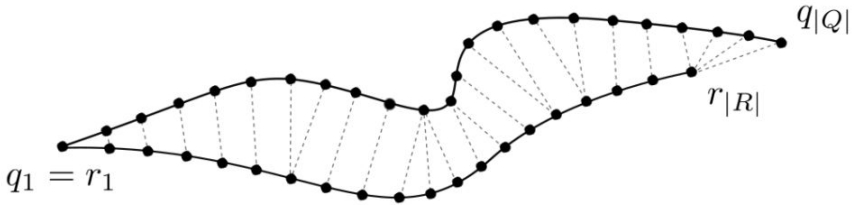


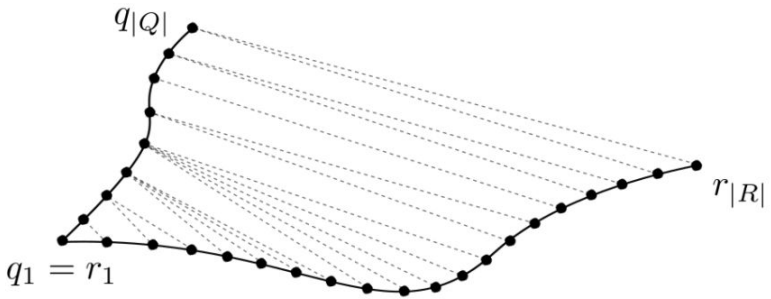


Results on Validation Unseen dataset

Recent Work

- Effective and General Evaluation for Instruction Conditioned Navigation using Dynamic Time Warping - https://arxiv.org/abs/1907.05446
- Suite of DTW[1] based evaluation metrics for general instruction conditioned robotic tasks including VLN



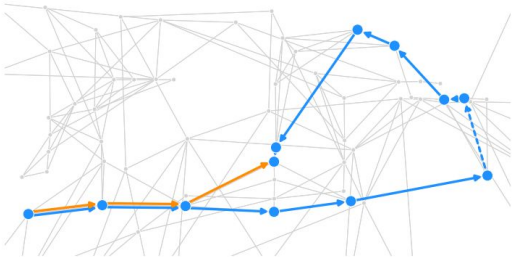


[1] Berndt et al. *Using Dynamic Time Warping to Find Patterns in Time Series* **AAAIWS'94.**

Conclusion

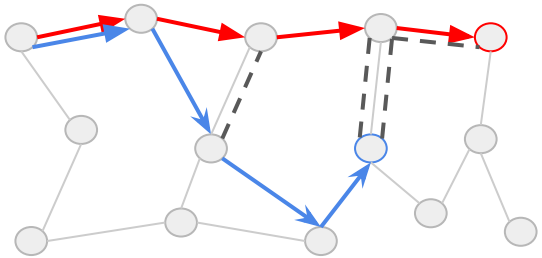
Data ✓

R4R



Evaluation ✓

CLS

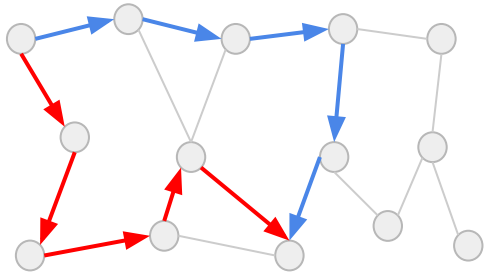


$$PC \propto exp(-\frac{d_1 + d_2 + d_3}{d_{th}})$$

$$LS \propto \frac{1}{|EPL - PL(P)|}$$

✓ Agent training

Fidelity-oriented agents



$r_T \sim 0$

Google AI

Thank You!

Questions?

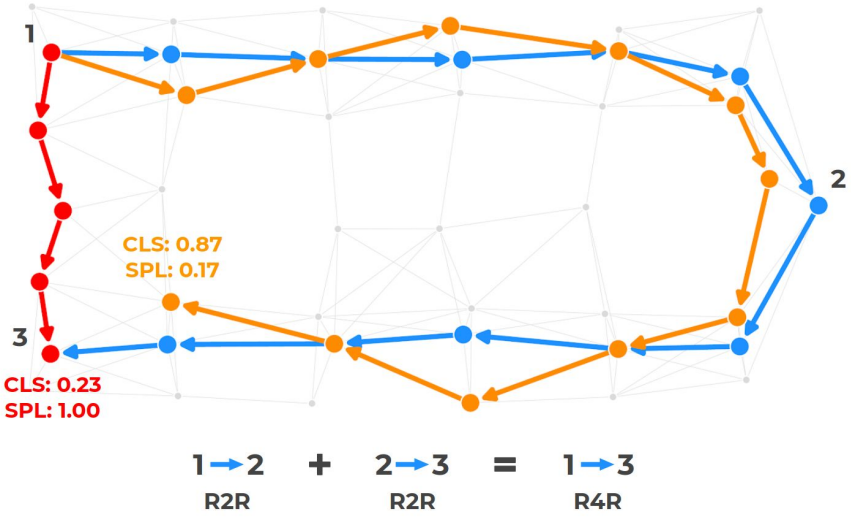
Appendix

Vision-and-Language Navigation (VLN)

- Interpret natural **language** instruction
- Combine with **spatio-temporal** and visual scene understanding
- Taking **action** in environments with dynamically changing **visual** percepts
- **Plan** sequence of actions to reach a goal

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

Key Contributions



- Biases in R2R Dataset
  - direct-to-goal shortest paths
  - primary evaluation metrics are based on goal completion
- To better gauge an agent's ability to stick to the path
  - R4R - general paths
  - CLS - measure of path fidelity
- Navigation agents trained using CLS as reward are more path conformant

Room-to-Room (R2R) Dataset[1]

- Data Collection
  - Sample start and goal nodes from the same house
  - Compute **shortest path** from start → goal
    - reject if path length is <5m or number of edges ≠ [4, 6]

- Dataset Statistics
  - **21,567** total <path, instruction> pairs
  - Average instruction length: 29 words; Vocabulary size: ~3.1k
  - Average path length: 10m
  - Success Criteria: success if the navigation error is less than 3m

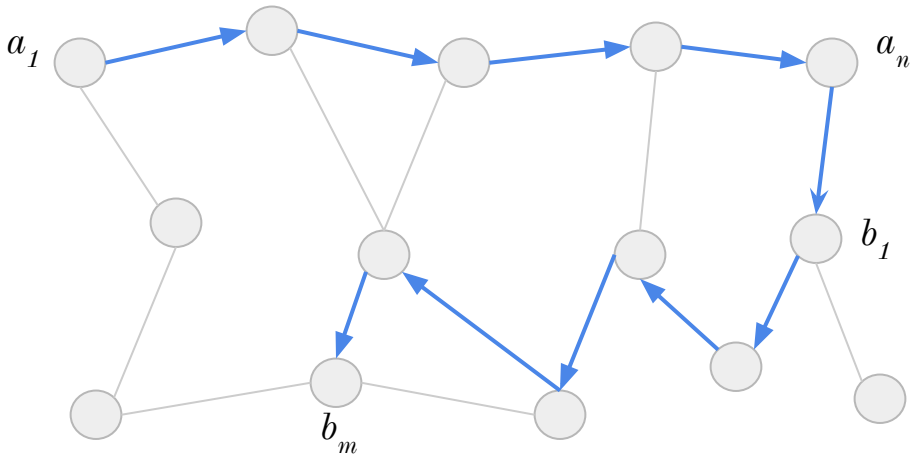
[1] Anderson et al. *Vision-and-language Navigation: Interpreting visually grounded navigation instructions in real environments*, **CVPR, 2018.**
*Amazon Mechanical Turk

Room-for-Room (R4R) Dataset

- Shortcomings of R2R dataset
  - All paths are direct-to-goal shortest paths
  - Largest path has only 6 edges
  - Agents maximizing success rate may incidentally appear to be maximizing path conformity

Google AI

Room-for-Room (R4R) Dataset



Make a left down at the narrow hall... Go out the door and wait. Turn around and enter the bedroom... Walk into the doorway and stop

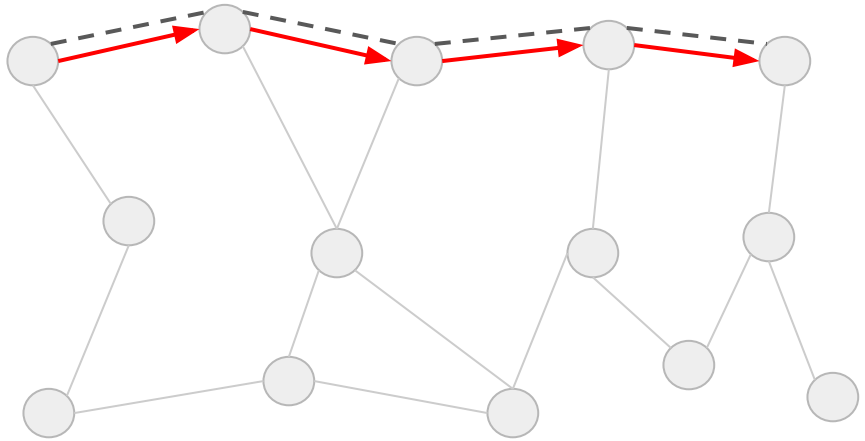
Existing VLN Evaluation Metrics

- Path Length (PL)
- Navigation Error (NE)
- Oracle Navigation Error (OSR)
- Success Rate (SR)
- Oracle Success Rate (OSR)
- Success weighted by Path Length (SPL)[1]
- Success weighted by Edit Distance (SED)[2]

[1] Anderson et al. *On Evaluation of Embodied Navigation Agents* **arXiv, 2018.**
[2] Chen et al. *Touchdown: Natural language navigation and spatial reasoning in visual street environments* **CVPR, 2019**

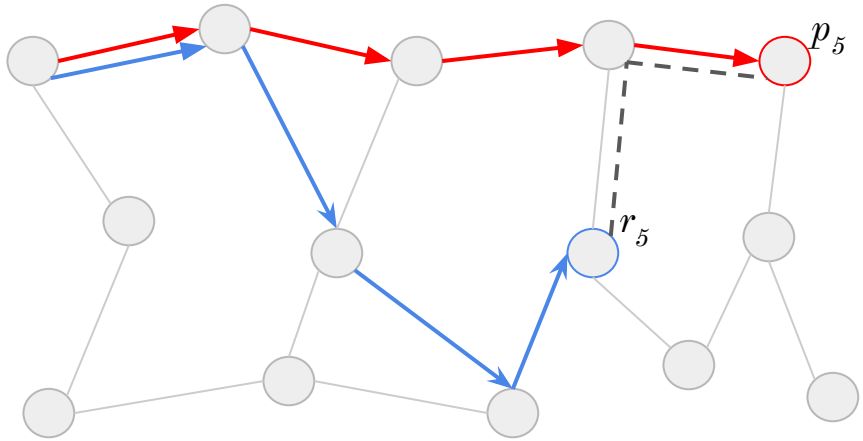
Existing VLN Evaluation Metrics

- Path Length (PL)
- Navigation Error
- Oracle Navigation Error
- Success Rate
- Oracle Success Rate
- Success weighted by Path Length (SPL)
- Success weighted by Edit Distance (SED)



Existing VLN Evaluation Metrics

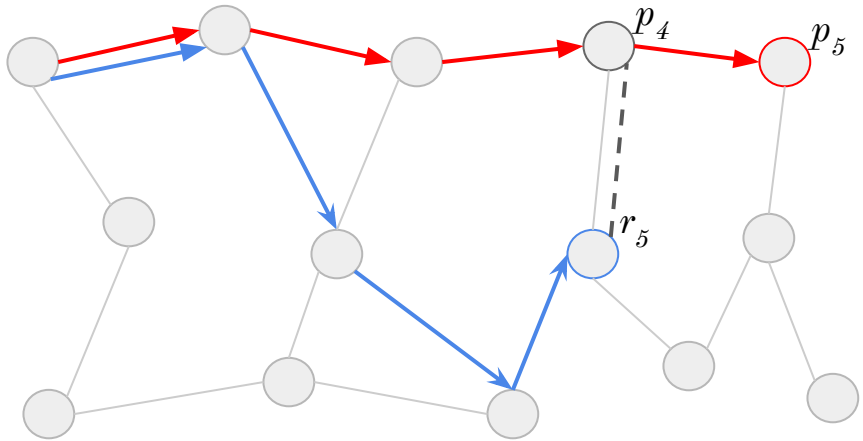
- Path Length
- Navigation Error (NE)
- Oracle Navigation Error
- Success Rate
- Oracle Success Rate
- Success weighted by Path Length (SPL)
- Success weighted by Edit Distance (SED)



Google AI

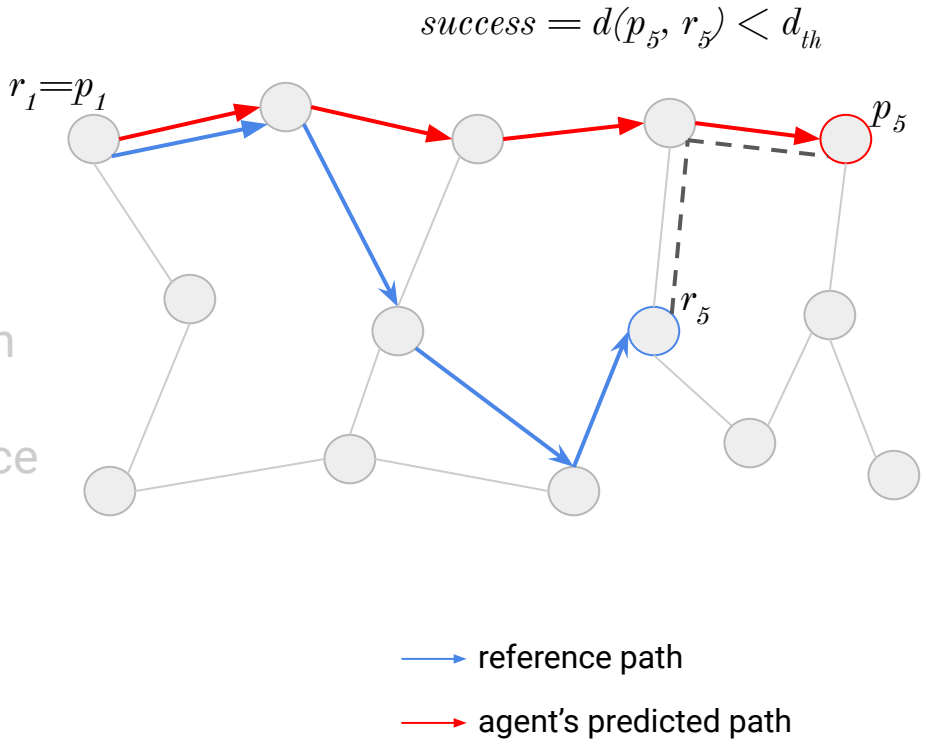
Existing VLN Evaluation Metrics

- Path Length
- Navigation Error
- Oracle Navigation Error (ONE)
- Success Rate
- Oracle Success Rate
- Success weighted by Path Length (SPL)
- Success weighted by Edit Distance (SED)



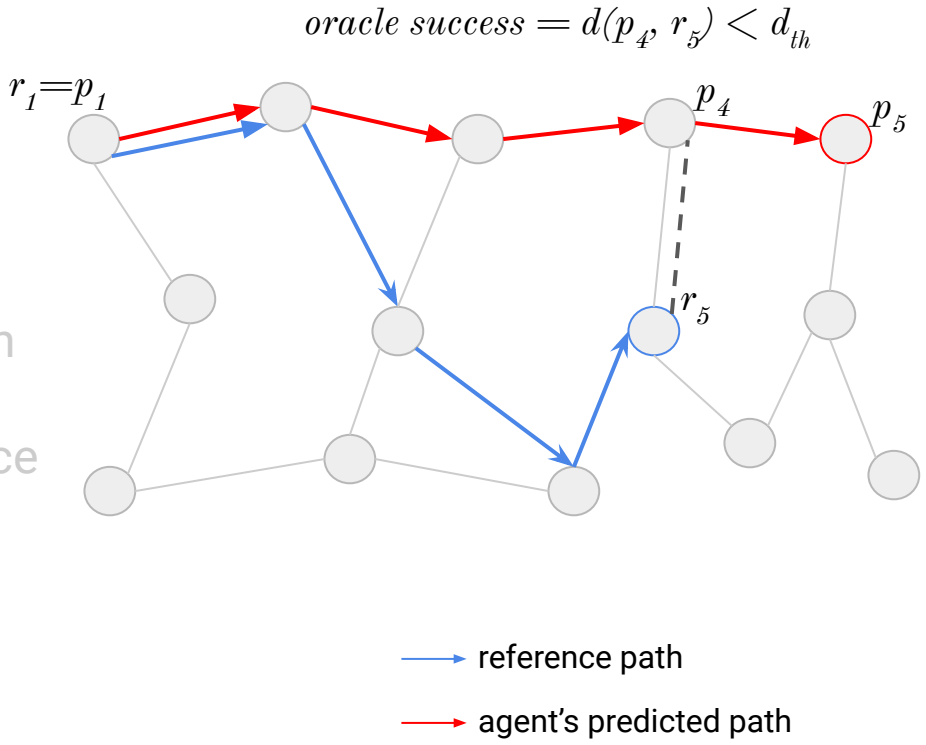
Existing VLN Evaluation Metrics

- Path Length
- Navigation Error
- Oracle Navigation Error
- Success Rate (SR)
- Oracle Success Rate
- Success weighted by Path Length (SPL)
- Success weighted by Edit Distance (SED)



Existing VLN Evaluation Metrics

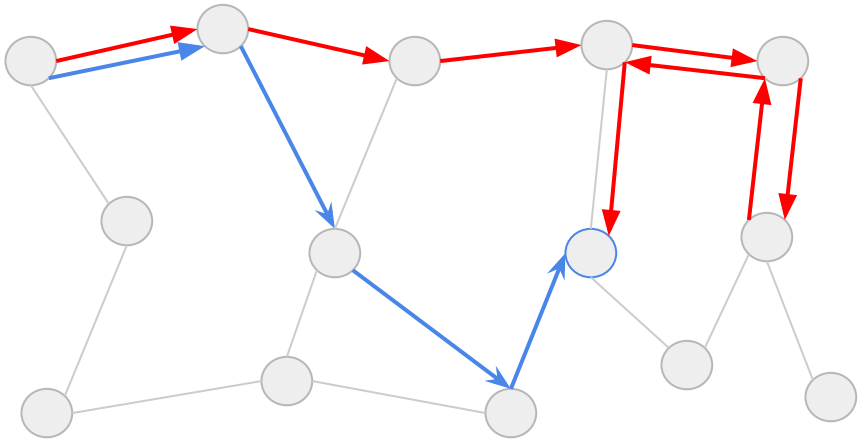
- Path Length
- Navigation Error
- Oracle Navigation Error
- Success Rate
- Oracle Success Rate (OSR)
- Success weighted by Path Length (SPL)
- Success weighted by Edit Distance (SED)



Existing VLN Evaluation Metrics

- Path Length
- Navigation Error
- Oracle Navigation Error
- Success Rate
- Oracle Success Rate
- Success weighted by Path Length (SPL)[1]
- Success weighted by Edit Distance (SED)

$$spl = 4 \;/\; 8 = 0.5$$



$$\mathrm{SR}(P,R) \cdot \frac{d(p_1, r_{|R|})}{\max\{\mathrm{PL}(P), d(p_1, r_{|R|})\}}$$

reference path

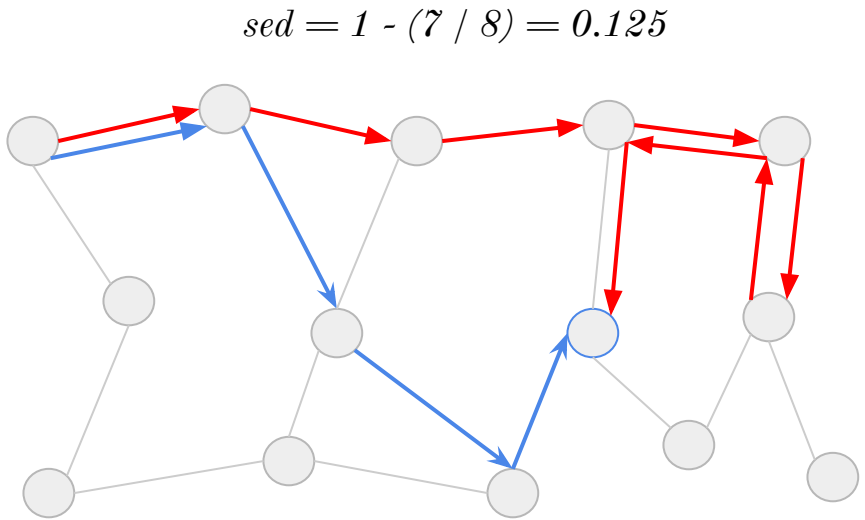
agent's predicted path

[1] **Anderson et al.** *On Evaluation of Embodied Navigation Agents* **arXiv, 2018.**

Existing VLN Evaluation Metrics

- Path Length
- Navigation Error
- Oracle Navigation Error
- Success Rate
- Oracle Success Rate
- Success weighted by Path Length (SPL)
- Success weighted by Edit Distance (SED)[2]

$$\mathrm{SR}(P,R)\cdot\left(1-\frac{\mathrm{ED}(P,R)}{\max\{|P|,|R|\}-1}\right)$$



[2] Chen et al. *Touchdown: Natural language navigation and spatial reasoning in visual street environments* **CVPR, 2019**

Desiderata

Path Similarity Measure	Soft Penalties	Unique Optimum	Scale Invariance	Tractability
Penalize deviations from reference path even if they lead to the same goal	Soft notion of dissimilarity that depends on distances in the graph	Perfect score if and only if the reference and predicted paths are an exact match	Can consistently be used for multiple datasets	Fast, automated evaluation of performance

Desiderata Coverage of Existing Metrics

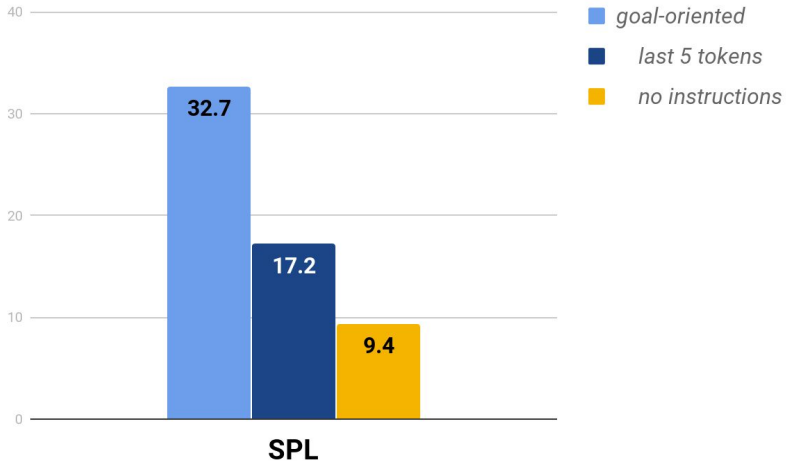
	Path Similarity Measure	Soft Penalties	Unique Optimum	Scale Invariance	Tractability
Path Length		✓			✓
Navigation Error		✓			✓
Oracle Navigation Error		✓			✓
Success Rate				✓	✓
Oracle Success Rate				✓	✓
SPL		✓		✓	✓
SED	✓		✓	✓	✓

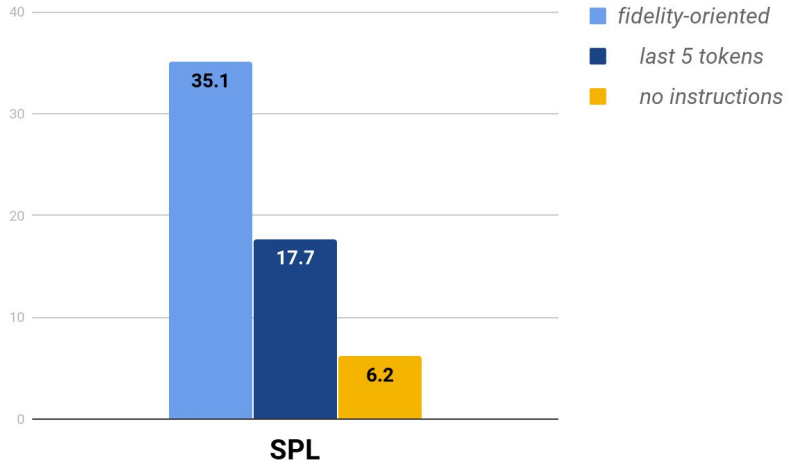
Conclusion

- Following instructions is important in VLN
  - going straight to the goal can often be deadly, e.g., games, search-and-rescue
- R4R has more general paths: better dataset for VLN
- CLS
  - metric for computing an agent's path fidelity to reference path
  - can also be used as reward function to incentivize the agent to better conform to the reference path
- Future Work: new benchmarks for R4R, more datasets using the toolkit provided in our work

R2R Performance

- Ablation Studies
  - Agent optimized to reach the goal may incidentally appear to be conforming to the instructions





Results on Validation Unseen dataset