Generative Language-Grounded Policy in Vision-and-Language Navigation with Bayes’ Rules

Shuhei Kurita
RIKEN AIP
JST PRESTO

Kyunghyun Cho
NYU Courant Institute
NYU Center for Data Science
CIFAR Fellow

Task: Vision-and-Language Navigation

An agent is embodied in the photorealistic indoor 3D modeling. With a textual instruction, the agent navigates in the indoor environment and reaches the goal place.

At initial time of $t = 0$, the agent receives the instruction X and the current visual observations s_0 . The agent doesn't know the entire environment information at first.
For time-step t, the agent obtains the visual observations s_t of the current place and chooses a next action a_t from the set of the possible actions A. By iteratively choosing the next action a_t , the agent is required to reach the goal place specified in X.

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*. 2017]

Two possible ways to VLN

Two possible approaches to VLN agents:
1. the ***discriminative*** approaches as most previous studies.
2. the ***generative*** approaches: a language model to navigate

Most previous studies adapt the modeling of $p(a_t|X, h_t)$ to prefict next action a_t from the instruction X, visual observations s_t and actions a_t , where $h_t = \{s_{:t}, a_{:t-1}\}$.
However, there are in fact two possible approaches to building such a VLN agent: ***discriminative*** and ***generative***.
We propose the generative modeling in Bayes' that directly utilize the conditional language model $p(X|a_t, h_t)$ for the navigation.

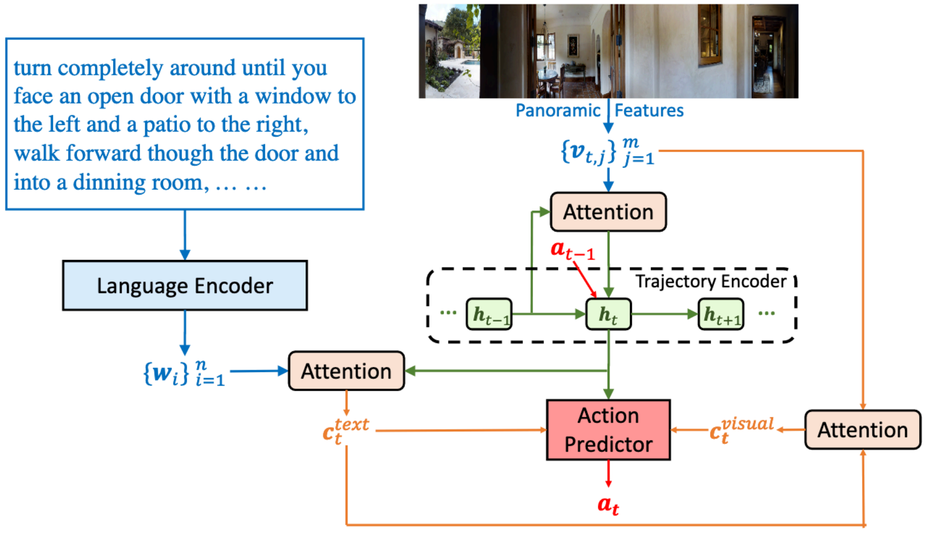


Figure 3: Cross-modal reasoning navigator at step t.

Reinforced cross-modal matching [Wang *et al*. 2019] as an example of the previous studies.

Proposed: Generative Language-Grounded Policy

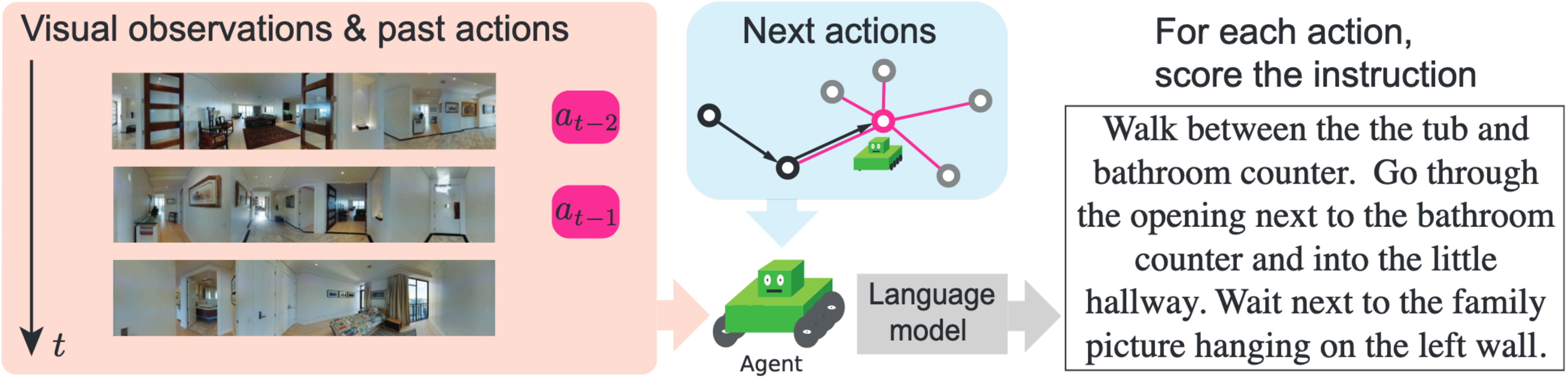
We propose the first model that directly uses a conditional language model $p(X|a_t, h_t)$ for navigation.

From the Bayes'

$$p(a_t|h_t, X) = \frac{p(X|a_t, h_t)p'(a_t|h_t)}{\sum_{a'_t \in \mathcal{A}} p(X|a'_t, h_t)p'(a'_t|h_t)} = \frac{p(X|a_t, h_t)}{\sum_{a'_t \in \mathcal{A}} p(X|a'_t, h_t)}$$

under $p'(a_t|h_t) = 1/|\mathcal{A}|$. Learning is:

$$L = -\sum_{t=1}^{T} \left\{ \log p(X|a_t, h_t) + \log \sum_{a'_t \in A} p(X|a'_t, h_t) \right\}$$



Results: Our *Gen*. policy vs *Disc*. policies of previous models

Model	Validation (Seen) PL↓	NE↓	SR↑	SPL↑	CLS↑	nDTW↑	SDTW↑	Validation (Unseen) PL↓	NE↓	SR↑	SPL↑	CLS↑	nDTW↑	SDTW↑
Disc.	10.69	5.40	0.519	0.482	0.619	0.588	0.445	12.88	6.52	0.380	0.335	0.488	0.458	0.304
Disc. +Aug.(A)	10.60	5.15	0.525	0.489	0.633	0.596	0.445	12.05	6.22	0.431	0.392	0.528	0.496	0.356
Gen.	11.23	5.53	0.481	0.451	0.625	0.579	0.427	12.98	6.17	0.434	0.371	0.514	0.478	0.344
Gen. +Aug. (B)	11.45	4.78	**0.563**	**0.531**	**0.664**	**0.630**	**0.505**	13.92	4.78	**0.476**	**0.405**	**0.539**	**0.503**	**0.379**
Gen.+Disc.(A+B)	10.18	4.67	0.568	0.540	0.680	0.640	0.510	12.06	5.42	0.489	0.437	0.570	0.533	0.403
Gen.+Disc.(A+B)*	11.30	4.58	0.575	0.541	0.678	0.636	0.509	14.65	5.19	0.518	0.439	0.564	0.515	0.397

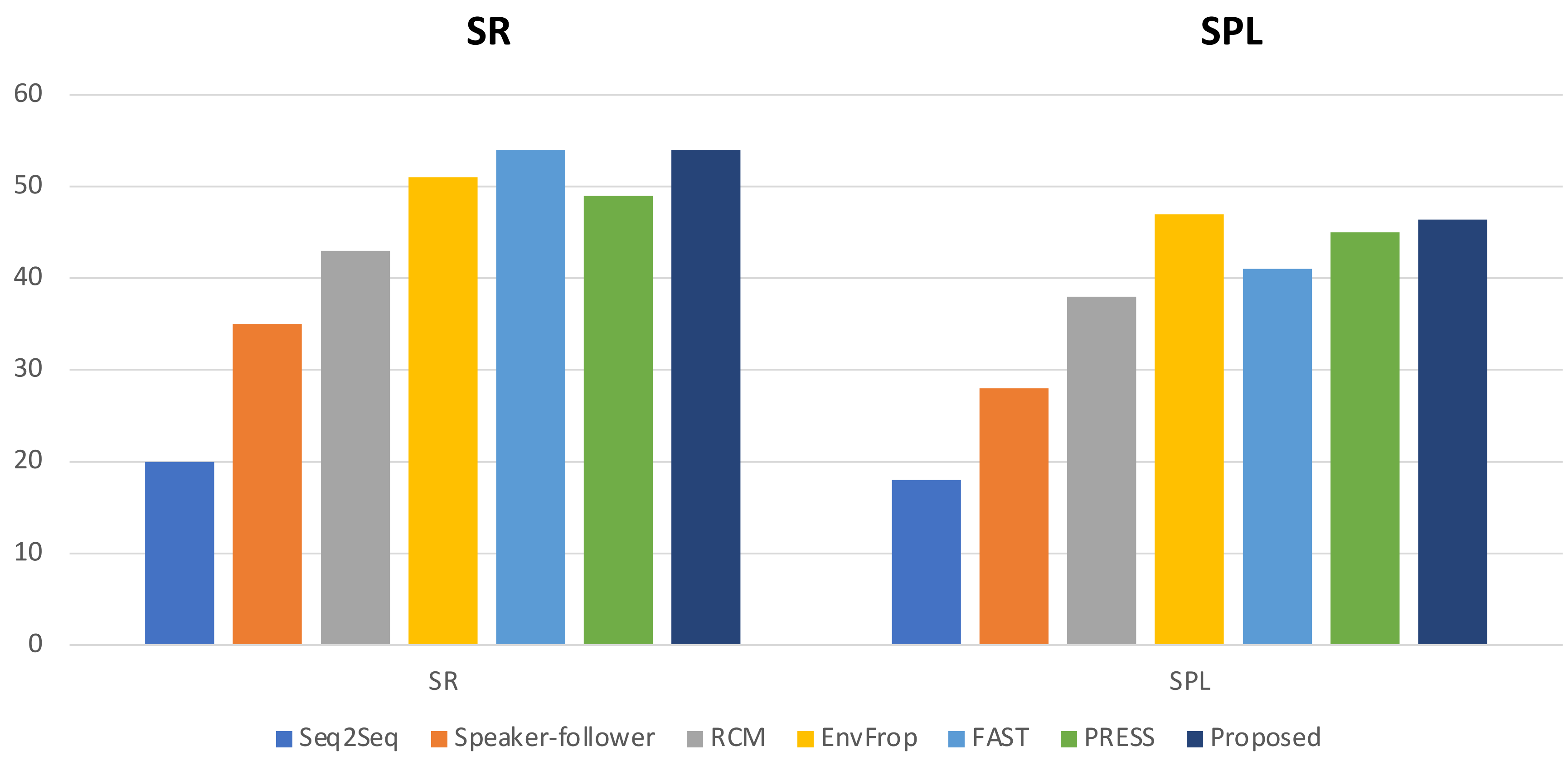
R2R experiments: generatiive vs discriminative policies

Model		Validation (Seen) PL↓	NE↓	SR↑	CLS↑	nDTW↑	SDTW↑	Validation (Unseen) PL↓	NE↓	SR↑	CLS↑	nDTW↑	SDTW↑
RCM	fidelity-oriented	18.8	5.4	0.526	0.553	-	-	28.5	5.4	0.261	0.346	-	-
nDTW	fidelity-oriented	-	-	-	-	-	-	-	-	0.285	0.354	0.304	0.126
BabyWalk	IL+RL	-	-	-	-	-	-	22.8	8.6	0.250	0.455	0.344	0.136
BabyWalk	IL+RL+Cur.	-	-	-	-	-	-	23.8	7.9	0.296	0.478	0.381	**0.181**
Disc.	supervised	20.1	7.0	0.386	0.622	0.512	0.305	20.0	9.8	0.172	0.446	0.305	0.101
Disc.	fidelity-oriented	21.1	6.6	0.449	0.644	0.530	0.360	29.2	9.2	0.211	0.385	0.282	0.116
Gen.	supervised	19.8	8.8	0.316	0.563	0.442	0.246	19.7	9.8	0.193	**0.479**	0.325	0.121
Gen.	fidelity-oriented	21.0	6.9	0.448	0.629	0.517	0.349	22.8	8.7	0.255	0.471	0.348	0.162

R4R experiments (R4R = R2R + R2R i.e. long and complicated trajectories)

R2R Test set result

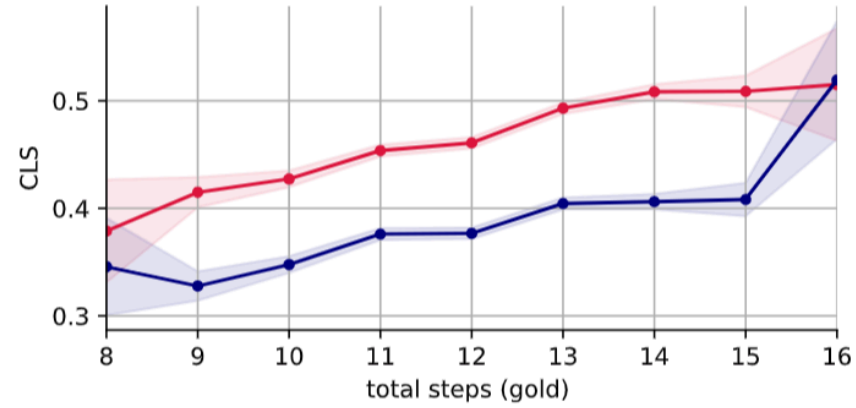
With the combination of the generative and discriminative policy, we achieve the competitive or better results in both the success rate (SR) and SPL in the R2R test set.

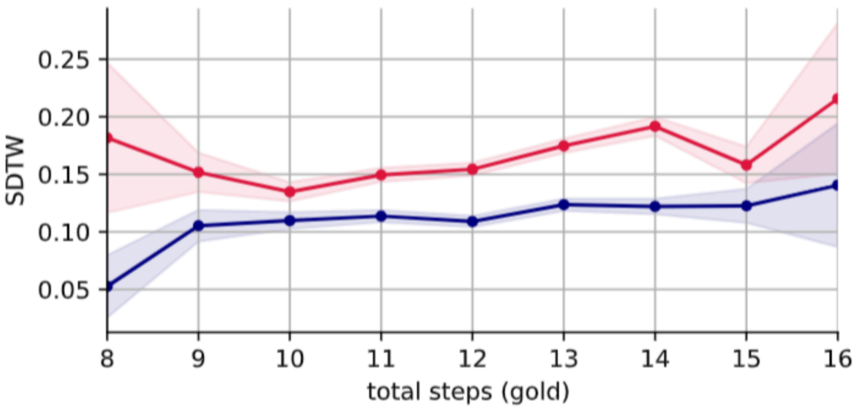


(Ref: Human scores SR: 86, SPL: 76)

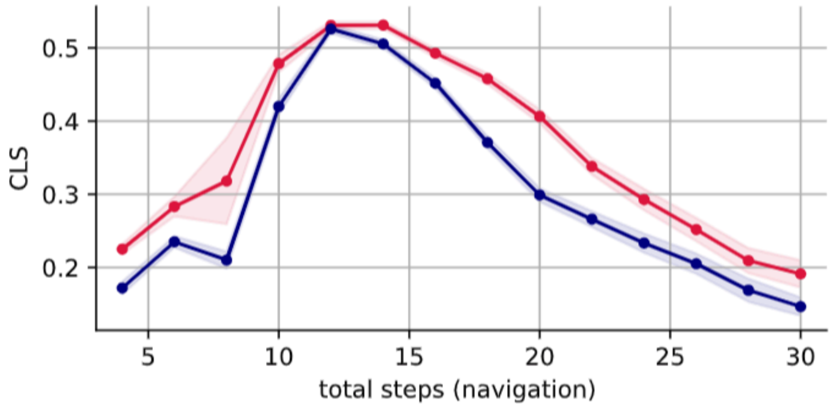
Analyses for trajectory-length and model accuracies

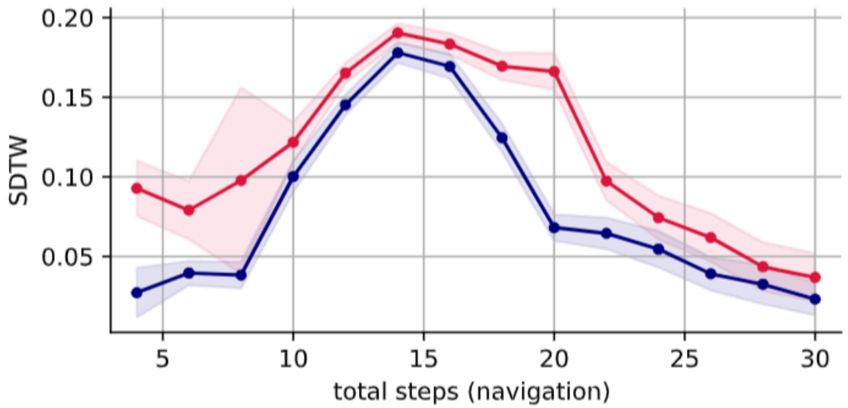
Performance comparisons on the *reference* trajectory steps





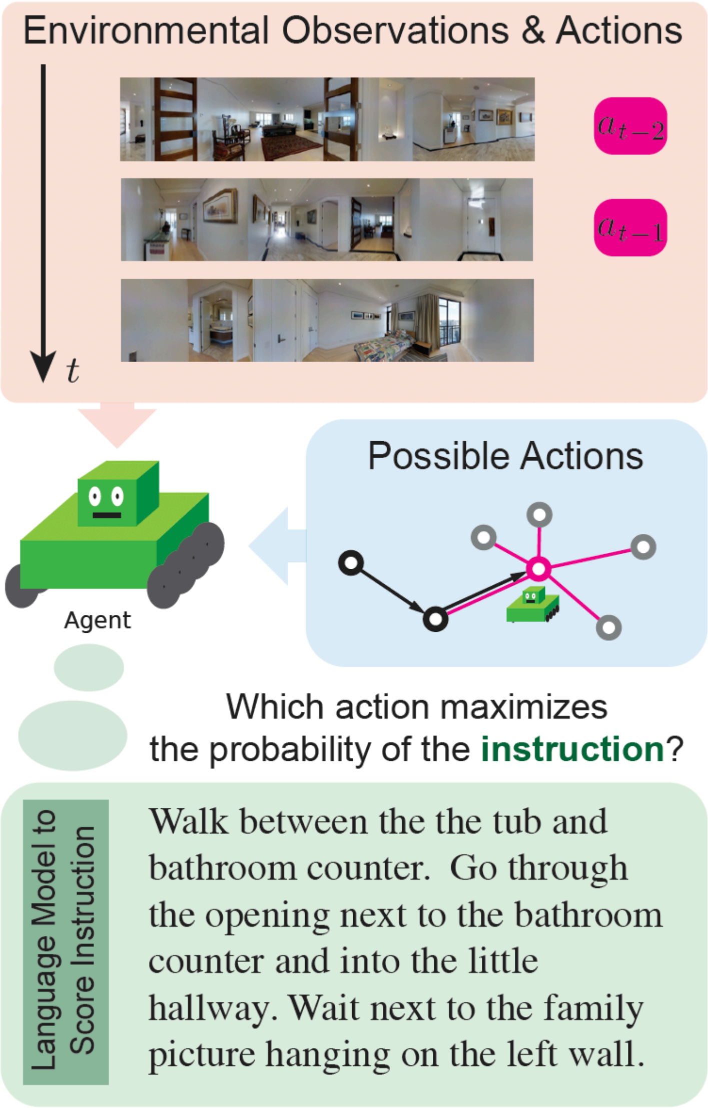
Performance comparisons on the *navigation* trajectory steps





On the R4R unseen validation set. Generative：red, Discriminative: blue

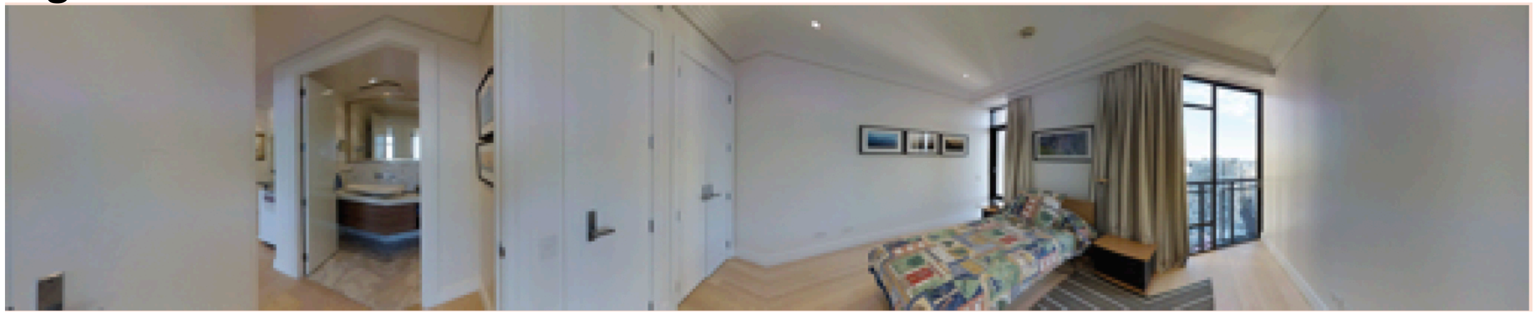
1-TENT Visualization

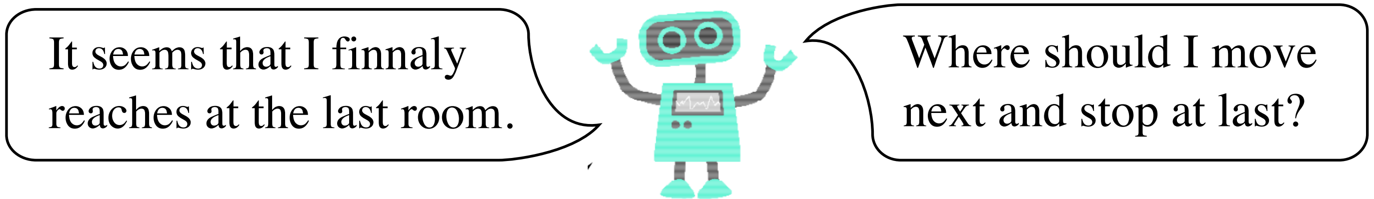


Instruction

Walk between the the tub and bathroom counter. Go through the opening next to the bathroom counter and into the little hallway. Wait next to the family picture hanging on the left wall.

Agent view



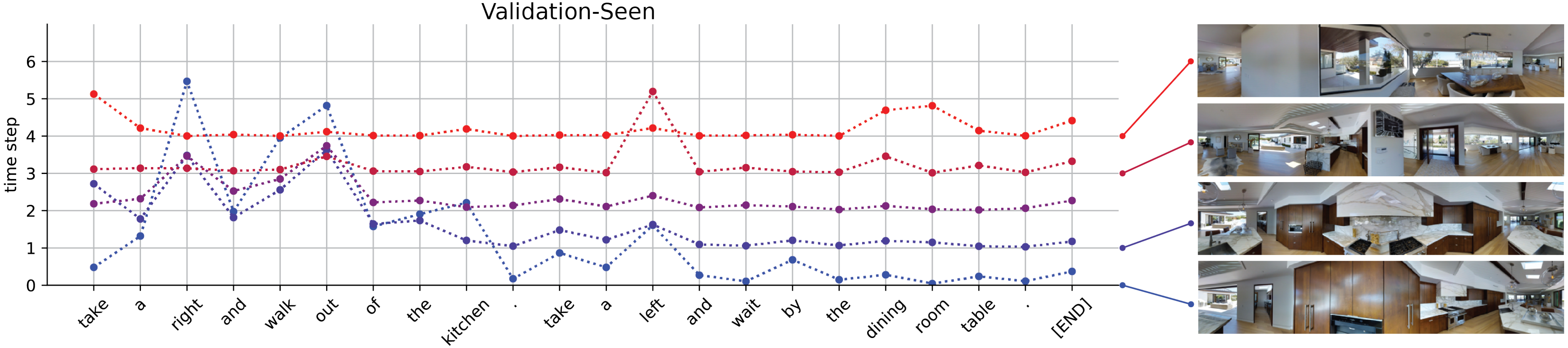


1-TENT (1 - Token-wise prediction ENTropy) to visualize agent's decision:

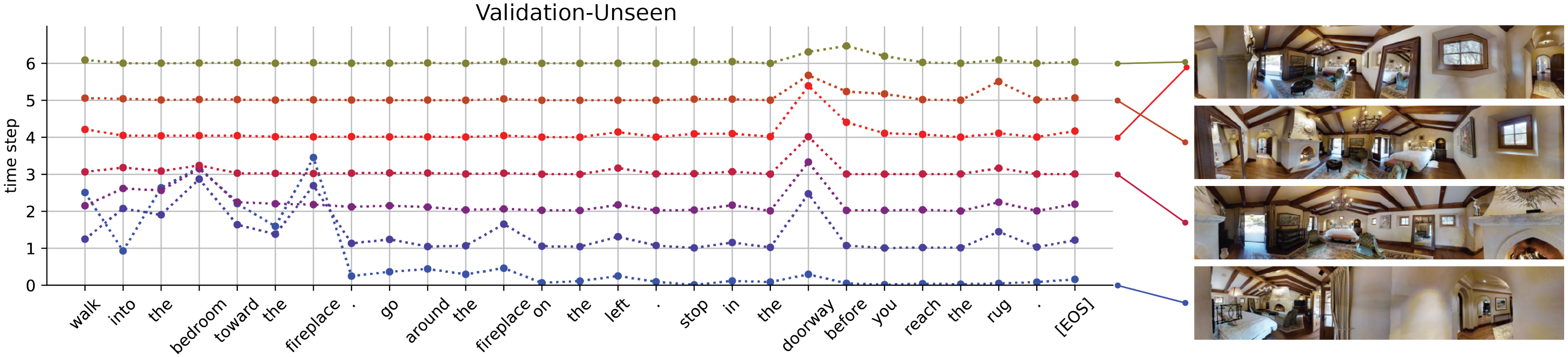
$$S(w_t) = -\sum_{a_t \in \mathcal{A}} q(a_t, w_t) \log_{|\mathcal{A}|} q(a_t, w_t),$$

$$q(a_t, w_t) = \frac{p(w_t | a_t, h_t, w_{:t-1})}{\sum_{a_t \in \mathcal{A}} p(w_t | a_t, h_t, w_{:t-1})}$$

1-TENT Visualization



Take a right and walk out of the kitchen. Take a left and wait by the dining room table.



Walk into the bedroom toward the fireplace. Go around the fireplace on the left.
Stop in the doorway before you reach the rug.

Conclusion

Two possible approaches for Vision-and-language navigation agents:
1. the ***discriminative*** approaches as most previous studies.

$$p(a_t|X, h_t)$$

2. the ***generative*** approaches: a language model to navigate

$$p(X|a_t, h_t)$$

We proposed the ***generative language-grounded policy***, which utilizes a *vision-and-action-conditioned language model* to follow the given textual instruction in the navigation.

Our experimental results shows that our *generative* approach is more effective than previous *discriminative* approaches especially in unseen validation and test sets in VLN.