Explore and Explain: Self-supervised Navigation and Recounting

R. Bigazzi, F. Landi, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara
{roberto.bigazzi, federico.landi, marcella.cornia, silvia.cascianelli, lorenzo.baraldi, rita.cucchiara}@unimore.it

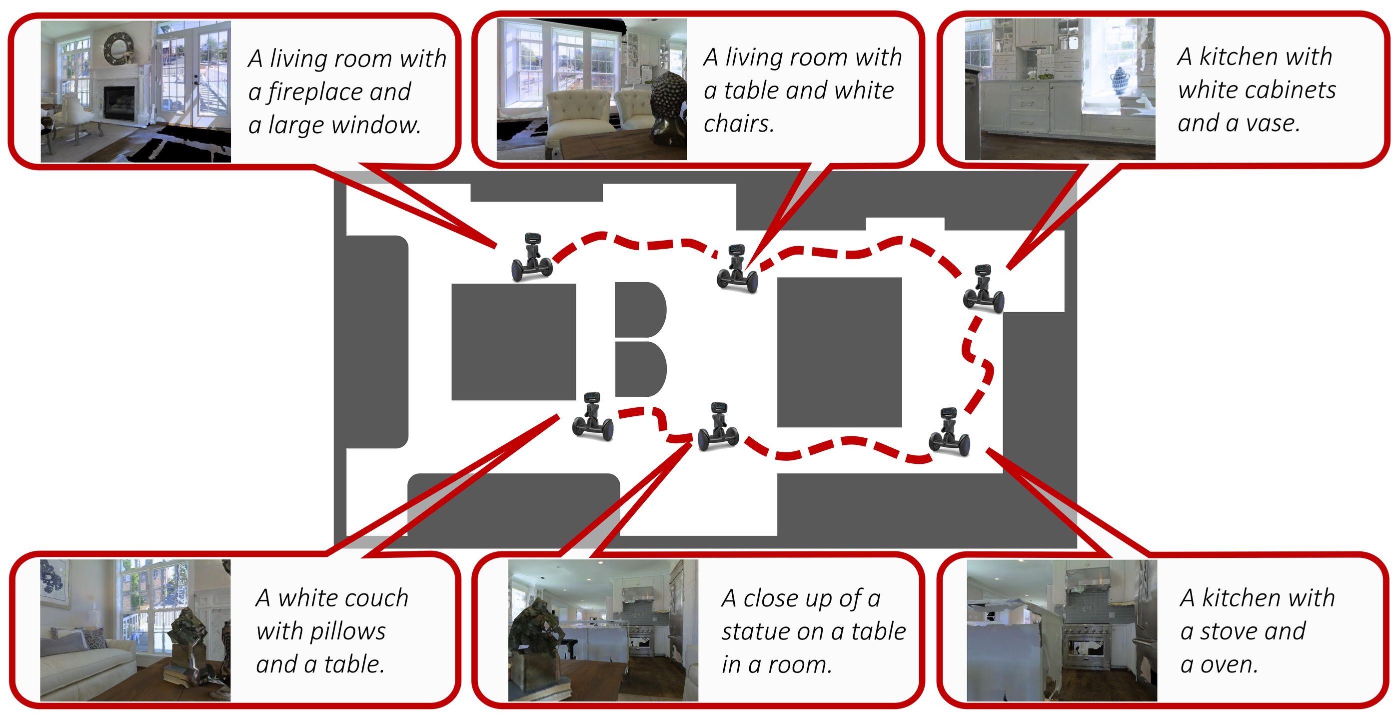
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Overview

Embodied AI has been recently gaining attention as it aims to foster the development of autonomous and intelligent robots combining decision making with computer vision.
In this work, we aim to enrich the traditional embodied exploration setting with a new task involving a third modality: describing the most relevant features of the environment [1] via natural language.

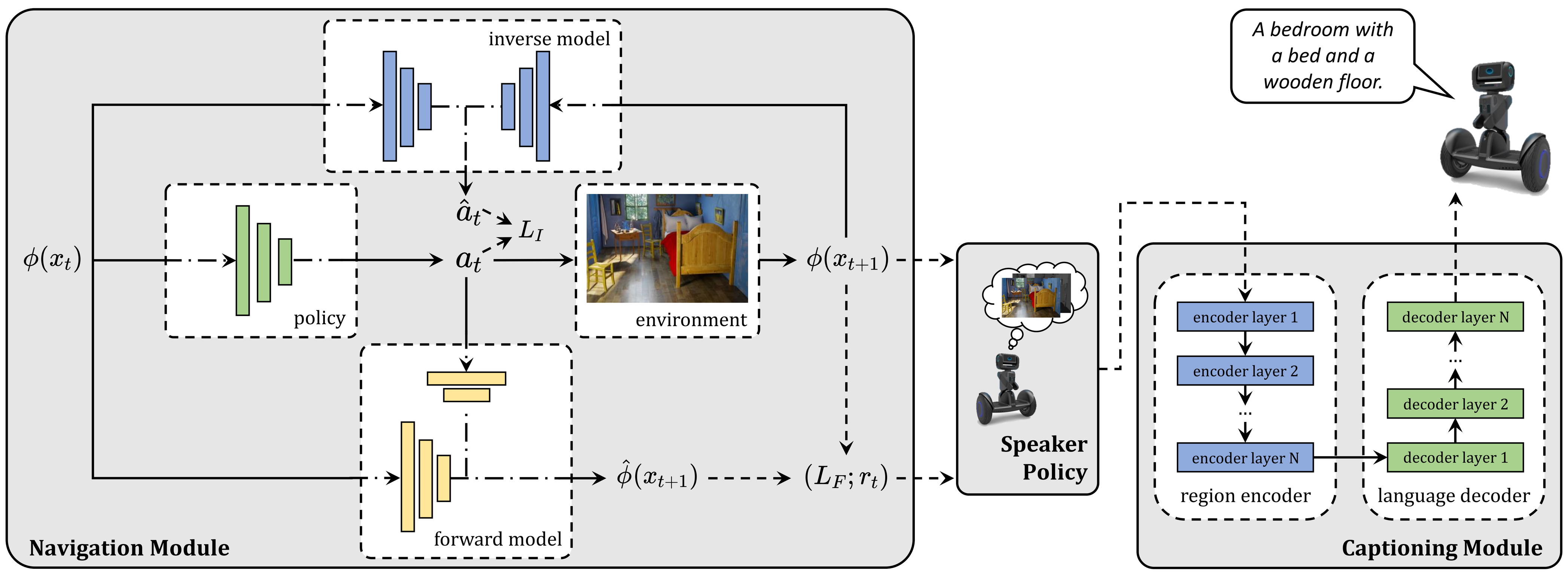
The proposed task presents three main challenges:
- How to maximize the relevance of seen objects?
- How to describe what the agent sees in its trajectory?
- How can the agent know when to talk? What is the best *speaker policy*?

We call this new setting for embodied AI ***Explore and Explain***.



eX^2 Architecture

We call our model **eX^2**, from the name of the task, and it consists of three main components: a navigation module, a captioner and a speaker policy:
- The **navigation module** is based on curiosity-driven exploration [2] and is in charge of exploring the environment.
- The **captioning module** is built on the Transformer model [3] and produces textual sentences describing the agent point of view.
- The **speaker policy** connects the previous modules and activates the captioner based on the information collected during the navigation.



Navigation Reward

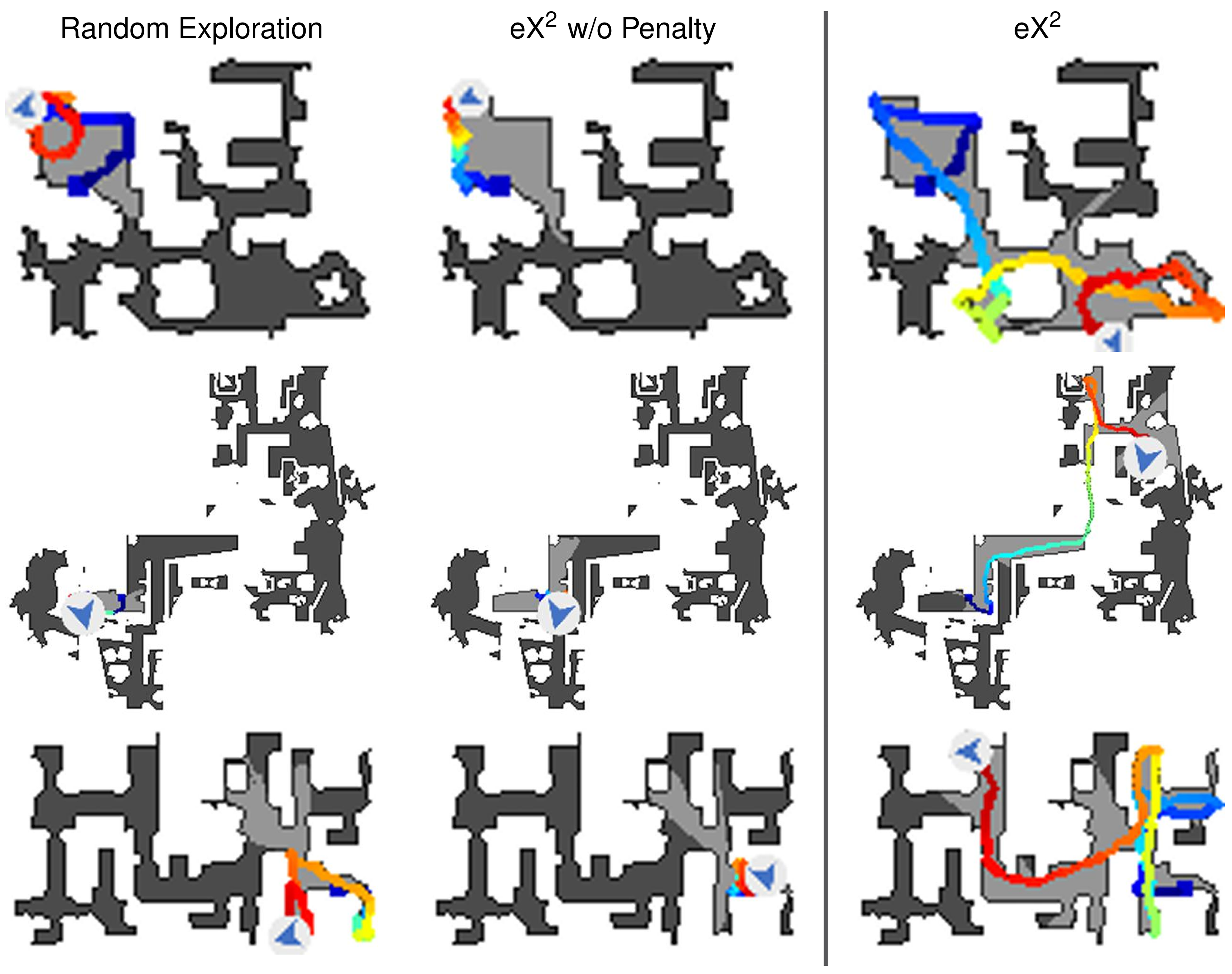
The navigation policy is trained with PPO [4] to maximize the sum of a two-component reward. Those components are the surprisal of the curiosity-driven exploration and a penalty p_t for repeated actions.

$$r_t = \frac{\eta}{2} \left\| f\left(\phi(x_t), a_t\right) - \phi(x_{t+1}) \right\|_2^2 - p_t$$

Speaker Policy

- **Object-driven:** the captioner is triggered if the number of objects (**O**) in the scene is above a certain threshold.
- **Depth-driven:** the mean depth value of the observation (**D**) is used for the activation.
- **Curiosity-driven:** the captioner is triggered using the surprisal reward (**S**).

Explore

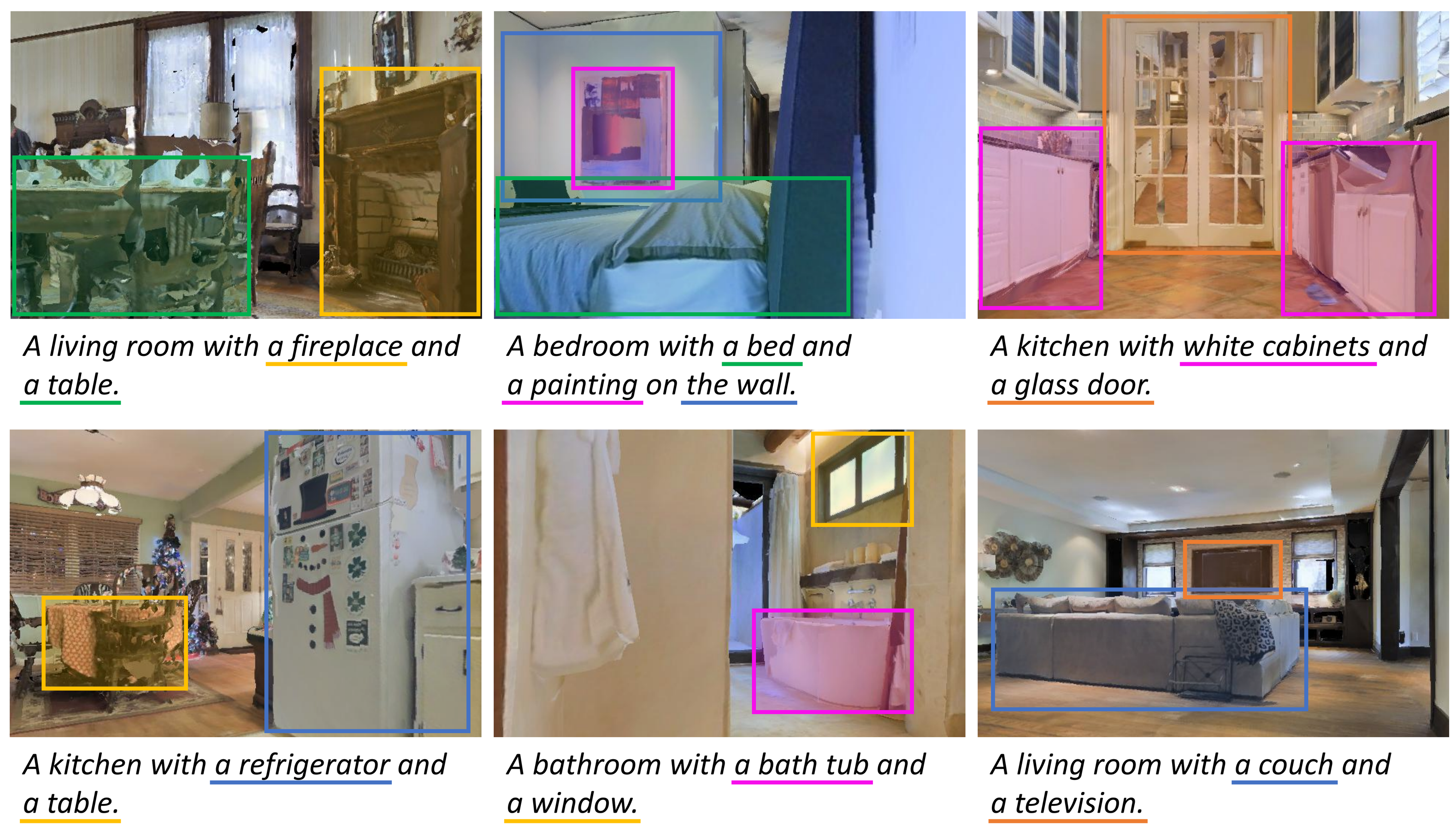


Qualitative results of the navigation module compared to a random explorer and to eX^2 without the penalty component in the reward.

Navigation Module	Surprisal
Random Exploration	0.333
eX^2 w/o Penalty for repeated actions (RGB only)	0.193
eX^2 w/o Penalty for repeated actions (Depth only)	0.361
eX^2 w/o Penalty for repeated actions (RGB + Depth)	0.439
eX^2	**0.697**

Surprisal scores for different navigation policies. Higher values indicate better exploration.

Explain



Sentences generated on images extracted from eX^2 exploration.

	eXplore and eXplain (2 layers)				
Speaker Policy	Cov$_{>1\%}$	Cov$_{>3\%}$	Cov$_{>5\%}$	Cov$_{>10\%}$	Div
Object-driven (**O** $\geq$ **5**)	0.373	0.497	0.577	0.713	0.340
Depth-driven (**D** $>$ **0.75**)	0.425	0.525	0.595	0.715	0.325
Curiosity-driven (**S** $>$ **1.0**)	**0.448**	**0.545**	**0.610**	**0.723**	**0.349**

Coverage (Cov): assesses how the predicted caption covers all the ground-truth objects.
Diversity (Div): diversity in terms of described objects of two consecutively generated captions.

References

[1] Savva, Manolis, et al. *Habitat: A platform for embodied ai research*, in ICCV 2019
[2] Pathak, Deepak, et al. *Curiosity-driven exploration by self-supervised prediction*, in CVPR Workshops 2017
[3] Vaswani, Ashish, et al. *Attention is all you need*, in NeurIPS 2017
[4] Schulman, John, et al. *Proximal policy optimization algorithms*, in arXiv 2017