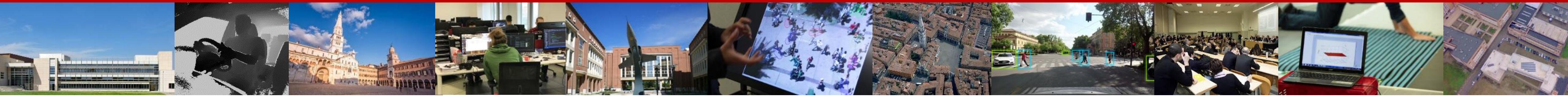
ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Explore and Explain: Self-supervised Navigation and Recounting



Roberto Bigazzi, Federico Landi, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, Rita Cucchiara

{name.surname}@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

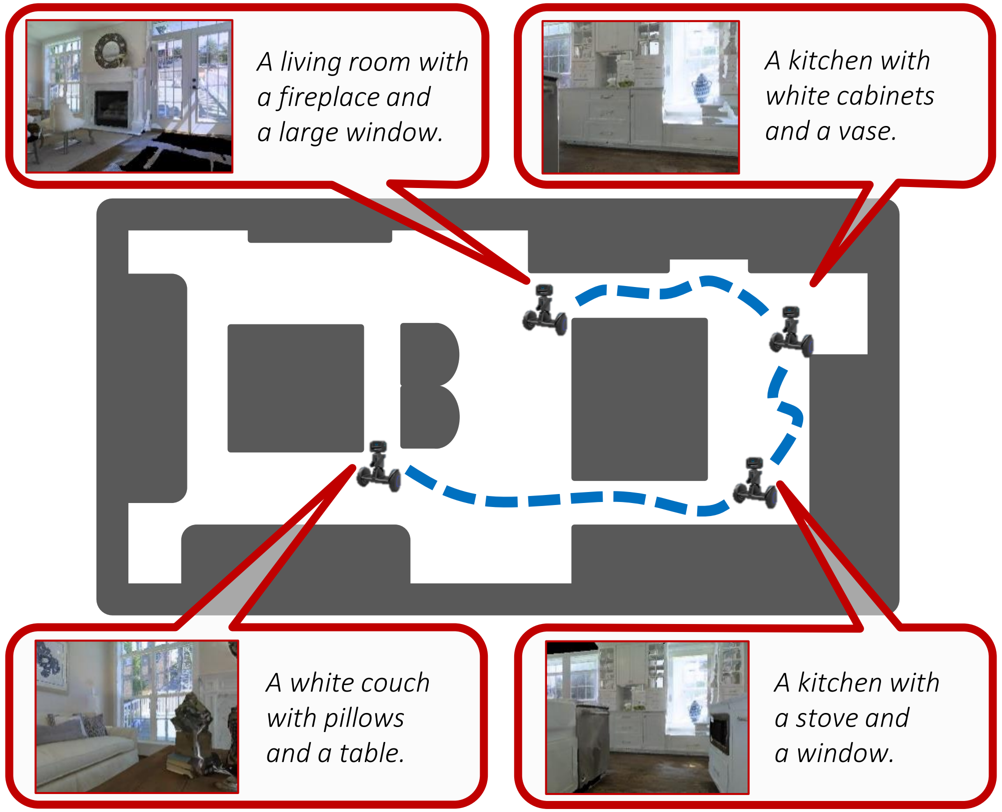
A NEW SETTING FOR EMBODIED AI

Current research on embodied AI mainly focuses on stand-alone tasks. Instead, we aim at bridging recent findings on **embodied exploration** and **image captioning**.

We devise a new setting involving:

Exploration of the environment

Description of the current view



We call this new task **Explore and Explain**

NEW TASK, NEW CHALLENGES

- How to maximize the relevance of seen objects?

  Exploration must be driven by curiosity towards novel elements [1]

- How to describe what the agent sees in its trajectory?

  The agent should integrate a State-of-the-Art model for image captioning

- How can the agent know when to talk?

  Not all that the agent sees is interesting: we need a Speaker Policy to activate the description module

[1] Pathak et al., CVPR Workshops 2017

CURIOSITY-DRIVEN EXPLORATION

Given a representation $\phi(x_t)$ for the rgb-d observation x_t, we sample an action a_t from the policy:

$$a_t \in \Big\{\textbf{0.25m ahead, 15}^\circ\textbf{ left, 15}^\circ\textbf{ right}\Big\}$$

Forward dynamics: predict $\phi(x_{t+1})$ given $\phi(x_t)$ and a_t:

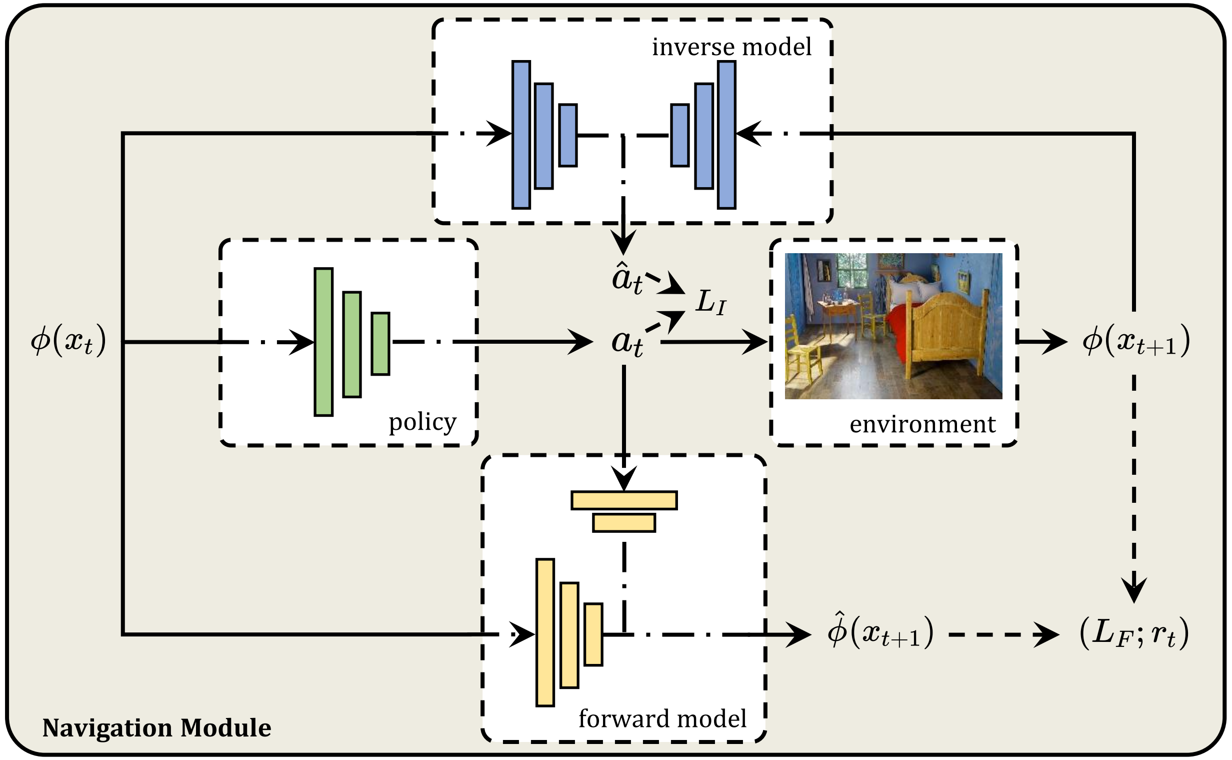
$$\hat{\phi}(x_{t+1}) = f\Big(\phi(x_t), a_t; \theta_F\Big)$$

Inverse dynamics: infer a_t given $\phi(x_t)$ and $\phi(x_{t+1})$:

$$\hat{a}_t = g\Big(\phi(x_t), \phi(x_{t+1}); \theta_I\Big)$$

CURIOSITY-DRIVEN EXPLORATION

The agent is trained with PPO [2]. The reward is proportional to the error of the forward model (**surprisal**), minus a **penalty p_t for repeated actions**.



$$L_F = \frac{1}{2}\left\|\hat{\phi}(x_{t+1}) - \phi(x_{t+1})\right\|_2^2$$

$$r_t = \eta L_F - p_t$$

[2] Schulman et al., arXiv 2017

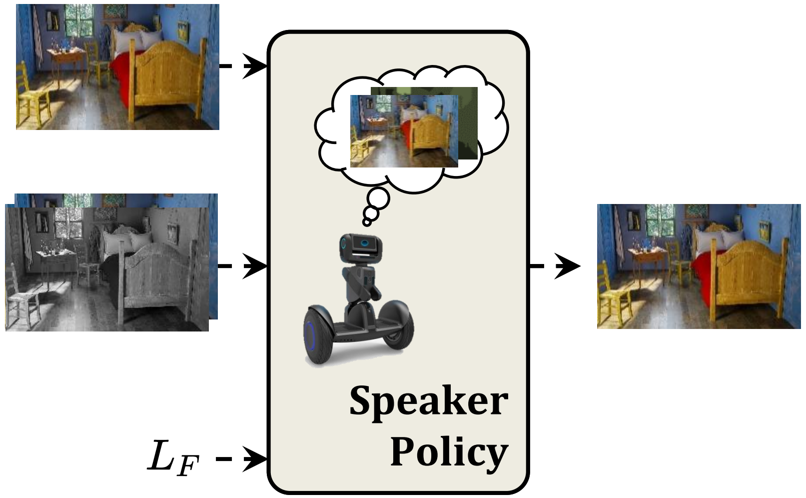
SPEAKER POLICY

The **speaker policy**, basing on the current observation, decides when to generate a sentence:

Object-driven: at least *N* objects (O) are observed in the scene

Depth-driven: the mean depth value (D) is above a fixed threshold

Curiosity-driven: the surprisal (S) is above a fixed threshold



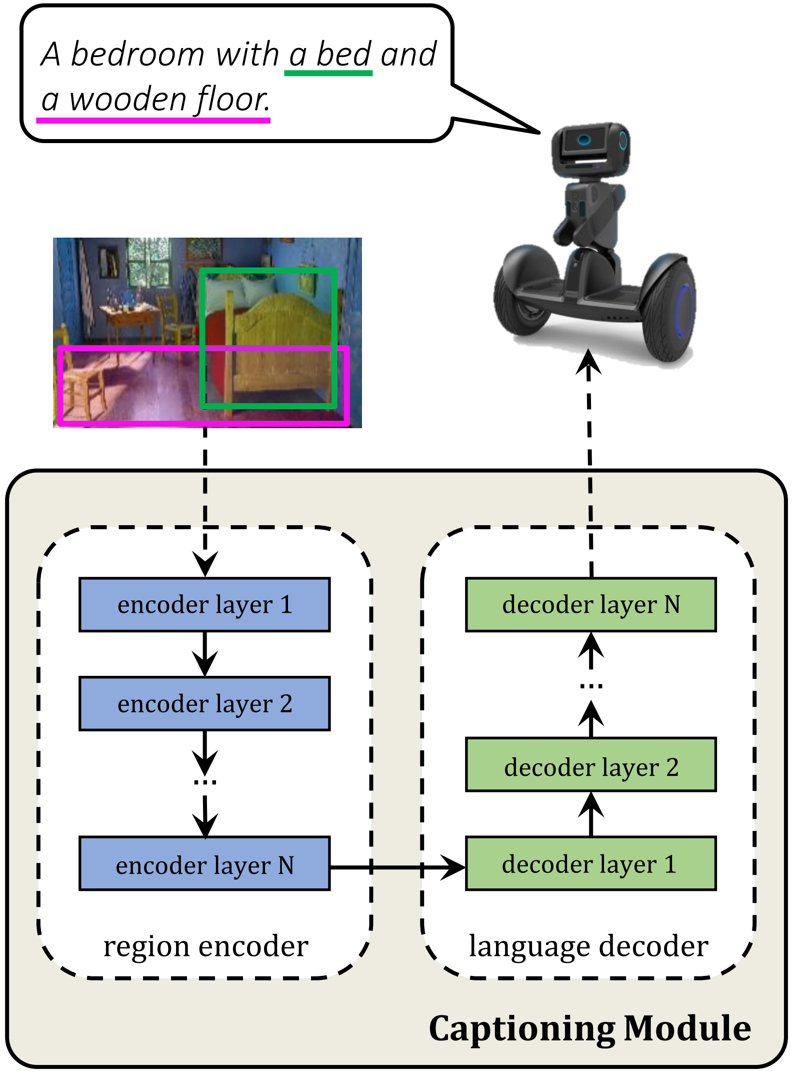
FULLY-ATTENTIVE CAPTIONING MODEL

Fully-attentive encoder-decoder architecture [3]

Detection of object regions with Faster R-CNN [4]

Two-phase training on the COCO dataset [5]

Evaluated with available information from the scene [6]



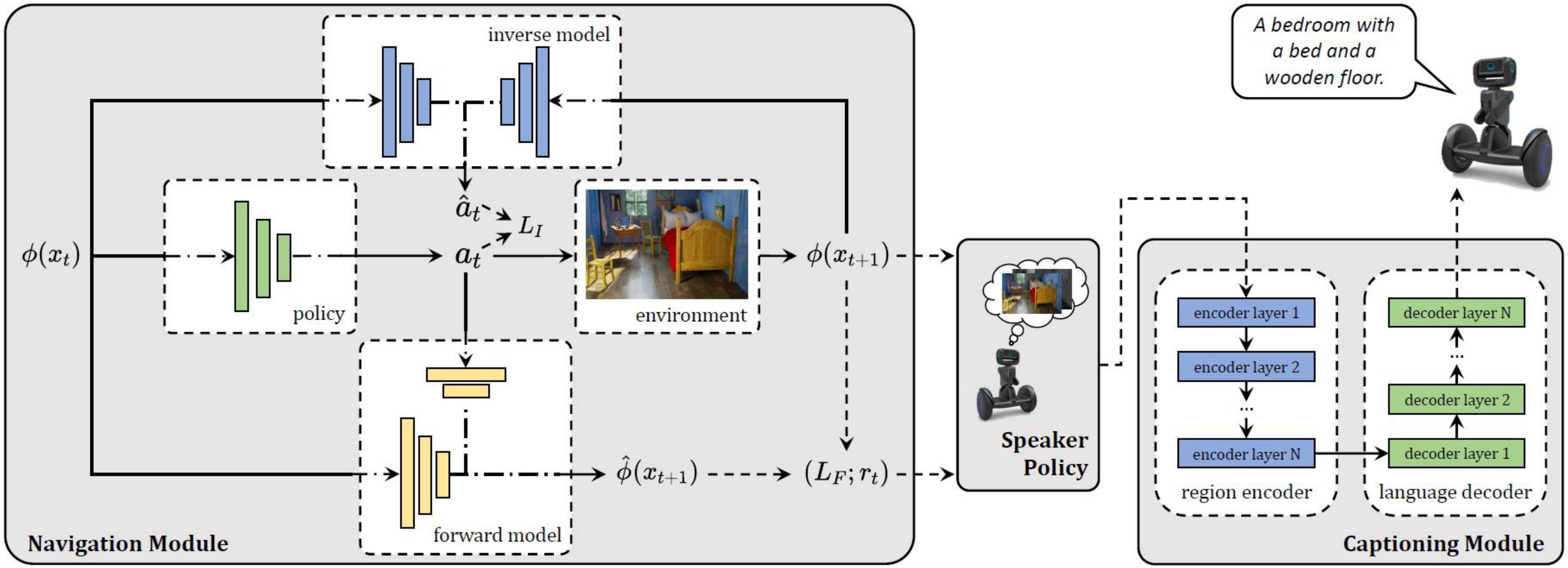
[3] Vaswani et al., NeurIPS 2017

[4] Shaoqing et al., NeurIPS 2015

[5] Lin et al., ECCV 2014

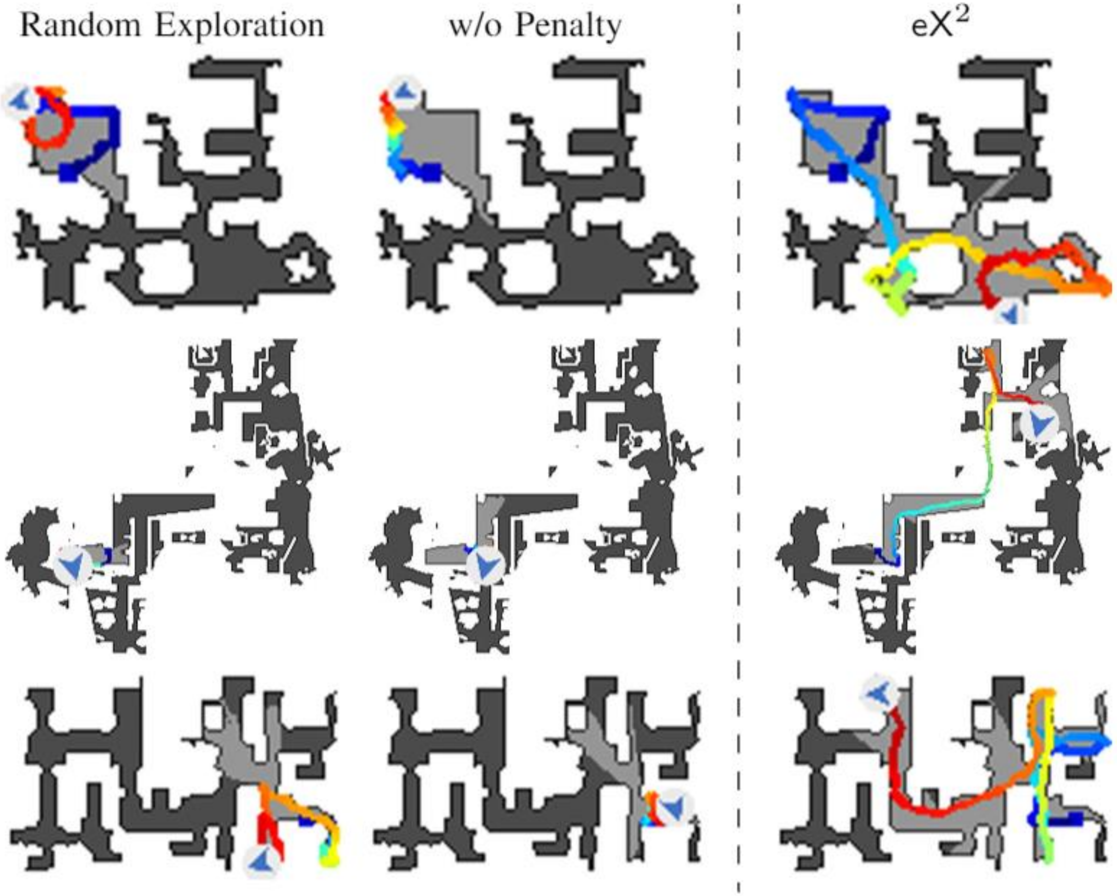
[6] Chang et al., 3DV 2017

EX² ARCHITECTURE



We call our model **eX²** (**Ex**plore and **Ex**plain), from the name of the task.

EX²PERIMENTAL RESULTS (NAVIGATION)



Navigation Module	Surprisal
Random Exploration	0.333
eX² w/o Penalty for repeated actions (RGB only)	0.193
eX² w/o Penalty for repeated actions (Depth only)	0.361
eX² w/o Penalty for repeated actions (RGB + Depth)	0.439
eX²	**0.697**

Our final agent **outperforms the baselines**

EX²PERIMENTAL RESULTS (CAPTIONING)

	Object-driven policy ($O \geq 1$) Loquacity = 43.3					Object-driven policy ($O \geq 3$) Loquacity = 27.4					Object-driven policy ($O \geq 5$) Loquacity = 15.8				
Captioning Module	$\text{Cov}_{>1\%}$	$\text{Cov}_{>3\%}$	$\text{Cov}_{>5\%}$	$\text{Cov}_{>10\%}$	Div	$\text{Cov}_{>1\%}$	$\text{Cov}_{>3\%}$	$\text{Cov}_{>5\%}$	$\text{Cov}_{>10\%}$	Div	$\text{Cov}_{>1\%}$	$\text{Cov}_{>3\%}$	$\text{Cov}_{>5\%}$	$\text{Cov}_{>10\%}$	Div
eX^2 (6 lay.)	0.456	0.550	0.609	0.706	0.386	0.387	0.502	0.576	0.696	0.363	0.348	0.468	0.549	0.691	0.352
eX^2 (3 lay.)	0.474	0.558	0.612	0.701	0.372	0.384	0.497	0.571	0.691	0.350	0.347	0.467	0.546	0.688	0.338
eX^2 (2 lay.)	**0.485**	**0.579**	**0.637**	**0.727**	0.368	**0.416**	**0.534**	**0.607**	**0.721**	0.349	**0.373**	**0.497**	**0.577**	**0.713**	0.340
eX^2 (1 lay.)	0.468	0.564	0.623	0.720	**0.394**	0.400	0.519	0.593	0.713	**0.377**	0.356	0.479	0.560	0.702	**0.373**

	Depth-driven policy ($D > 0.25$) Loquacity = 38.5					Depth-driven policy ($D > 0.5$) Loquacity = 31.1					Depth-driven policy ($D > 0.75$) Loquacity = 14.8				
Captioning Module	$\text{Cov}_{>1\%}$	$\text{Cov}_{>3\%}$	$\text{Cov}_{>5\%}$	$\text{Cov}_{>10\%}$	Div	$\text{Cov}_{>1\%}$	$\text{Cov}_{>3\%}$	$\text{Cov}_{>5\%}$	$\text{Cov}_{>10\%}$	Div	$\text{Cov}_{>1\%}$	$\text{Cov}_{>3\%}$	$\text{Cov}_{>5\%}$	$\text{Cov}_{>10\%}$	Div
eX^2 (6 lay.)	0.433	0.532	0.600	0.705	0.360	0.420	0.519	0.585	0.701	0.346	0.399	0.497	0.566	0.691	0.339
eX^2 (3 lay.)	0.427	0.524	0.588	0.700	0.349	0.413	0.511	0.577	0.695	0.335	0.394	0.491	0.559	0.685	0.330
eX^2 (2 lay.)	**0.463**	**0.562**	**0.625**	**0.730**	0.341	**0.449**	**0.550**	**0.612**	**0.726**	0.330	**0.425**	**0.525**	**0.595**	**0.715**	0.325
eX^2 (1 lay.)	0.448	0.548	0.613	0.723	**0.371**	0.434	0.536	0.603	0.719	**0.359**	0.412	0.513	0.583	0.708	**0.355**

	Curiosity-driven policy ($S > 0.7$) Loquacity = 27.2					Curiosity-driven policy ($S > 0.85$) Loquacity = 18.2					Curiosity-driven policy ($S > 1.0$) Loquacity = 6.4				
Captioning Module	$\text{Cov}_{>1\%}$	$\text{Cov}_{>3\%}$	$\text{Cov}_{>5\%}$	$\text{Cov}_{>10\%}$	Div	$\text{Cov}_{>1\%}$	$\text{Cov}_{>3\%}$	$\text{Cov}_{>5\%}$	$\text{Cov}_{>10\%}$	Div	$\text{Cov}_{>1\%}$	$\text{Cov}_{>3\%}$	$\text{Cov}_{>5\%}$	$\text{Cov}_{>10\%}$	Div
eX^2 (6 lay.)	0.425	0.523	0.588	0.703	0.356	0.421	0.515	0.581	0.699	0.360	0.422	0.518	0.583	0.702	0.364
eX^2 (3 lay.)	0.418	0.514	0.578	0.694	0.348	0.413	0.506	0.571	0.691	0.350	0.413	0.506	0.570	0.690	0.361
eX^2 (2 lay.)	**0.453**	**0.552**	**0.617**	**0.726**	0.340	**0.448**	**0.545**	**0.611**	**0.724**	0.342	**0.448**	**0.545**	**0.610**	**0.723**	0.349
eX^2 (1 lay.)	0.438	0.539	0.604	0.719	**0.370**	0.433	0.530	0.597	0.716	**0.373**	0.434	0.532	0.597	0.717	**0.380**

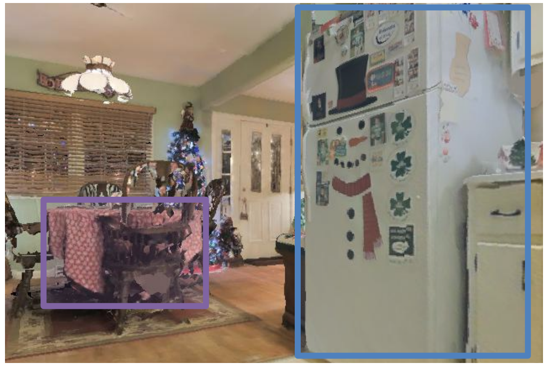
Surprisal is good criterion for the speaker policy. We obtain the best results with two transformer layers.

EX²PERIMENTAL RESULTS (CAPTIONING)

A living room with a fireplace and a table.

A bedroom with a bed and a painting on the wall.

A kitchen with white cabinets and a glass door.



A kitchen with a refrigerator and a table.

A bathroom with a bathtub and a window.

A living room with a couch and a television.

eX² describes the main objects in the scene and produces a suitable description even with partial occlusion

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Thank you for your attention

Explore and Explain: Self-supervised Navigation and Recounting


Roberto Bigazzi


Federico Landi


Marcella Cornia


Silvia Cascianelli


Lorenzo Baraldi


Rita Cucchiara

{name.surname}@unimore.it

University of Modena and Reggio Emilia, Italy

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImageLab