Evidential Active Recognition: Intelligent and Prudent Open-World Embodied Perception

Northwestern University

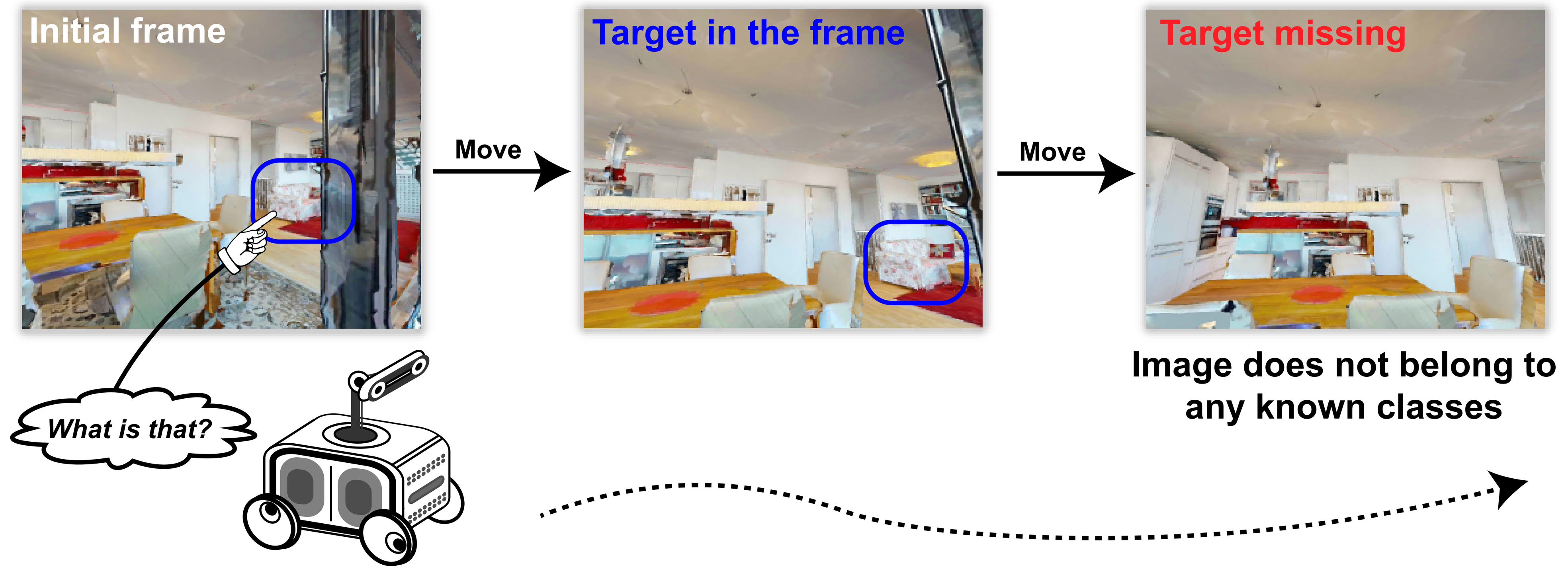
WORMPEX AI RESEARCH LLC

Lei Fan, Mingfu Liang, Yunxuan Li, Gang Hua and Ying Wu

{leifan, mingfuliang2020, yunxuanli2019}@u.northwestern.edu, ganghua@gmail.com, yingwu@northwestern.edu

CVPR SEATTLE, WA JUNE 17-21, 2024

Motivations



Unexpected visual observations often occur when the robot is exploring the environment, which bring negative impacts to embodied agents.

Negative impacts of unexpected observations to embodied agents

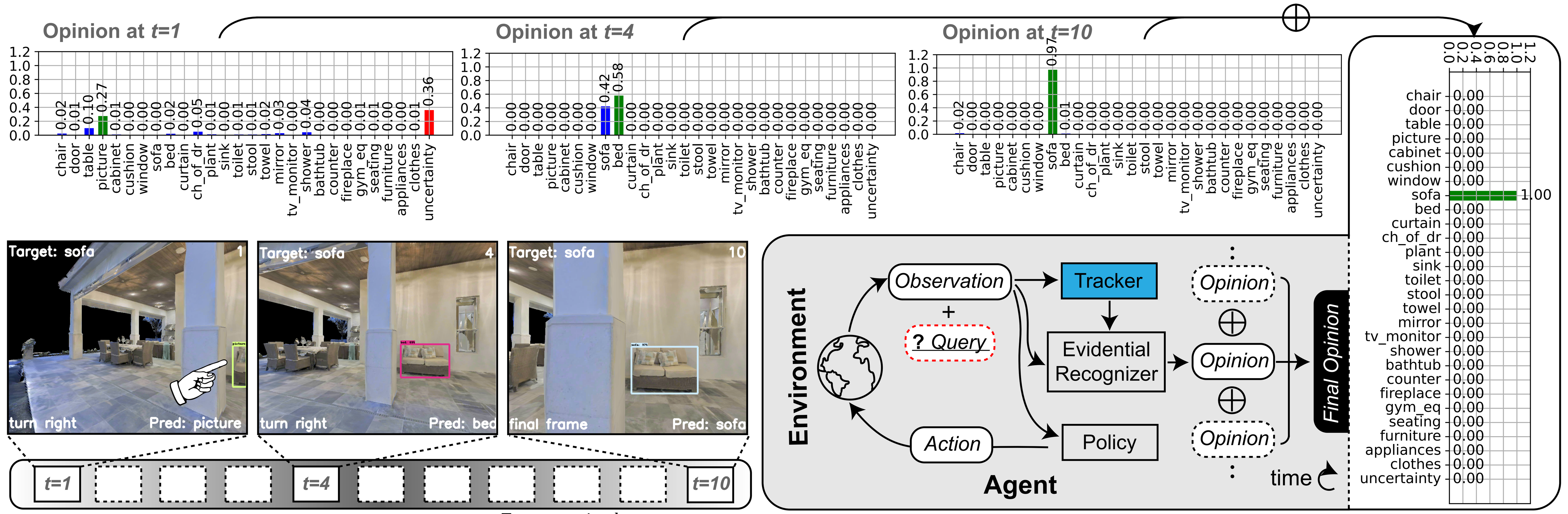
- **During training:** unexpected observations could mislead the policy learning.
  If the target is out-of-view, the recognition may fail to provide rewards that accurately represent the worth of actions being taken.
- **During testing:** unexpected observations could impede reasonable action selection while poisoning the final category prediction.

Preliminaries

We apply evidential deep learning for single frame uncertainty estimation.

- For a K-class recognition task, the frame of discernment $\Theta = \{k, 1 \leq k \leq K\}$ contains K exclusive singletons.
- Considering the visual observation v^t at timestep t, the method estimates $K+1$ mass values. Besides K belief terms b^t, the additional one is the uncertainty u^t.
- These $K+1$ mass values satisfy
$$\sum^K b_k^t + u^t = 1,\ 0 \leq u^t, b_k^t \leq 1.$$

Method



Type equation here.

The overview of proposed active recognition agent.

The bars for top prediction and uncertainty are colored green and red, respectively.

Note that uncertainty arises when the target is partially out of view at the first step.

Despite this, the final result is accurate due to the fusion of evidence.

Multi-view evidence combination rule between frame t and j

$$b_k = b_k^t \oplus b_k^j = \frac{1}{\sum_{P^t \cap P^j = \emptyset} b_{P^t}^t b_{P^j}^j} \sum_{P^t \cap P^j = k} b_{P^t}^t b_{P^j}^j$$
$$= \frac{b_k^t b_k^j + b_k^t u^j + b_{P^t}^t u^t}{1 - \sum_{i \neq q} b_i^t b_q^j}$$

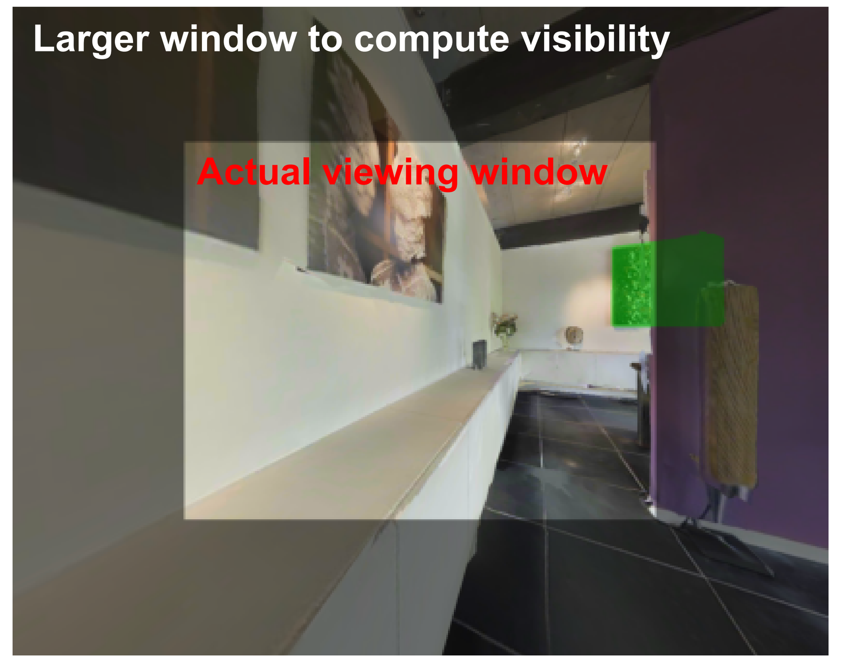
Reward design

We adopts our uncertainty estimation result into the reward design of training the recognition policy.

Specifically, the reward is straightforwardly defined as the estimated belief for the correct class y , which is $r^t = b_y^t$.

Difficulty-designated Dataset for Active Recognition

Target: *picture* (under green mask)
Difficulty level: "moderate"



Target: *chest of drawers* (under green mask)
Difficulty level: "hard"

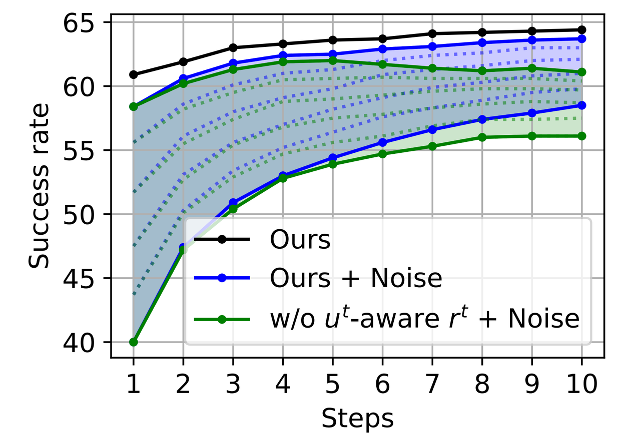
Why did we build this dataset?

Active recognition is supposed to address recognition challenges that cannot be resolved by passive recognition.

To better facilitate evaluation of active recognition in indoor simulator, we collect and propose this dataset.

We assign the difficulty level considering three aspects, i.e., visibility, relative distance and observed pixels.

Result



When the agent is trained without an uncertainty-aware reward, the performance is more likely to be negatively affected by noise.