Vision and Cognitive Systems

SCQ1097939 - LM CS,DS,CYB,PD

Context-aware motion prediction and embodied AI

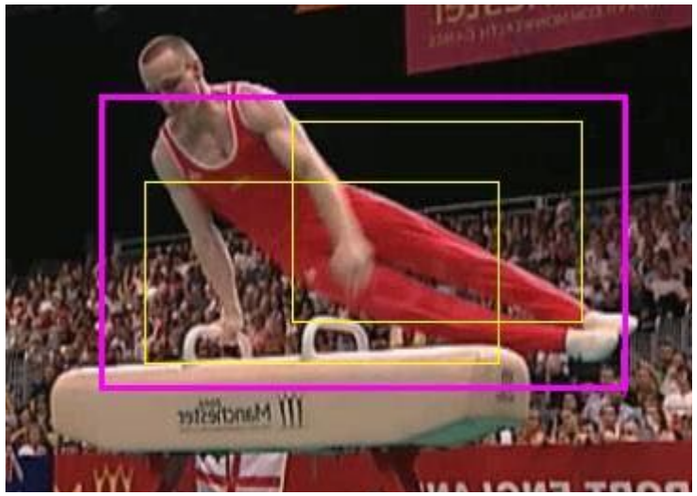
Prof. Lamberto Ballan

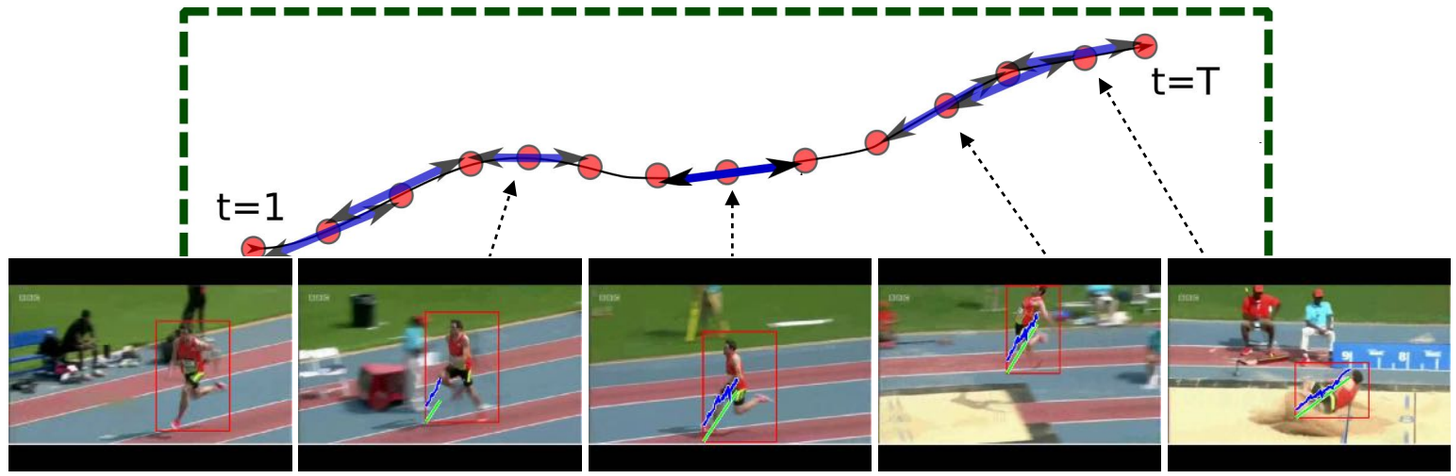
University of Padova, A.Y. 2023/24 January 16, 2024

DIPARTIMENTO MATEMATICA
Department of Mathematics “Tullio Levi-Civita”

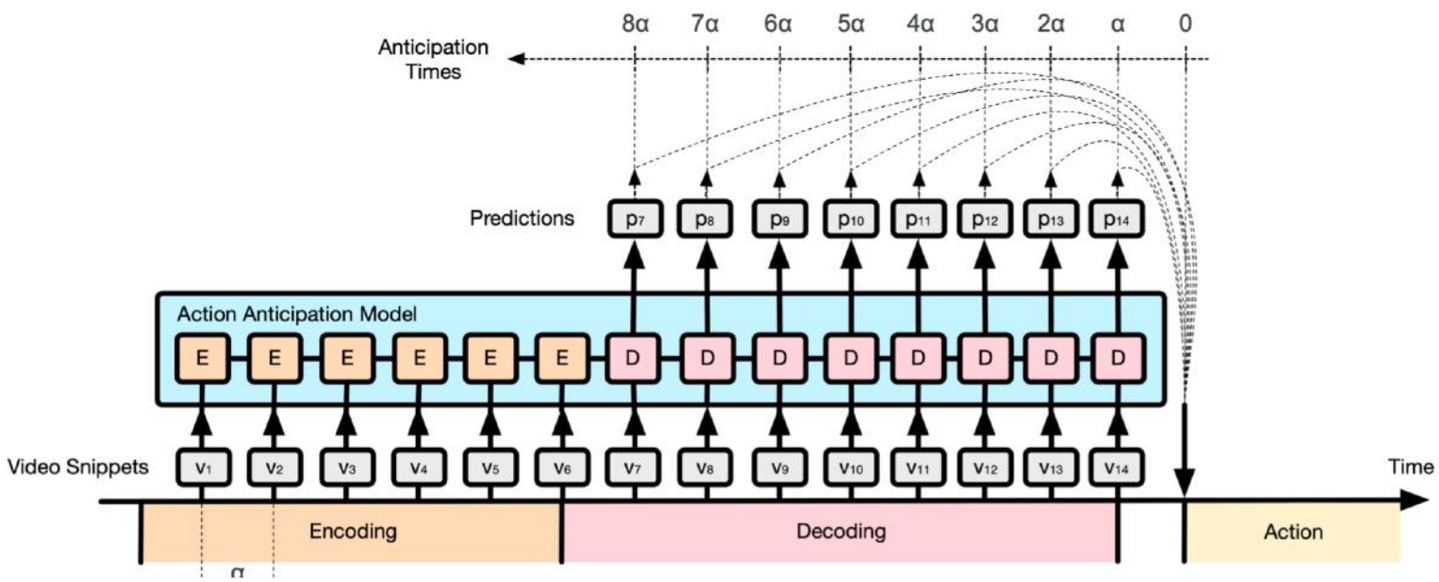
Action recognition meets predictive vision

- Predicting action progress/completion in videos





- Action anticipation and (pedestrian) intent prediction



Context-aware motion prediction and embodied social navigation

EPIC KITCHENS

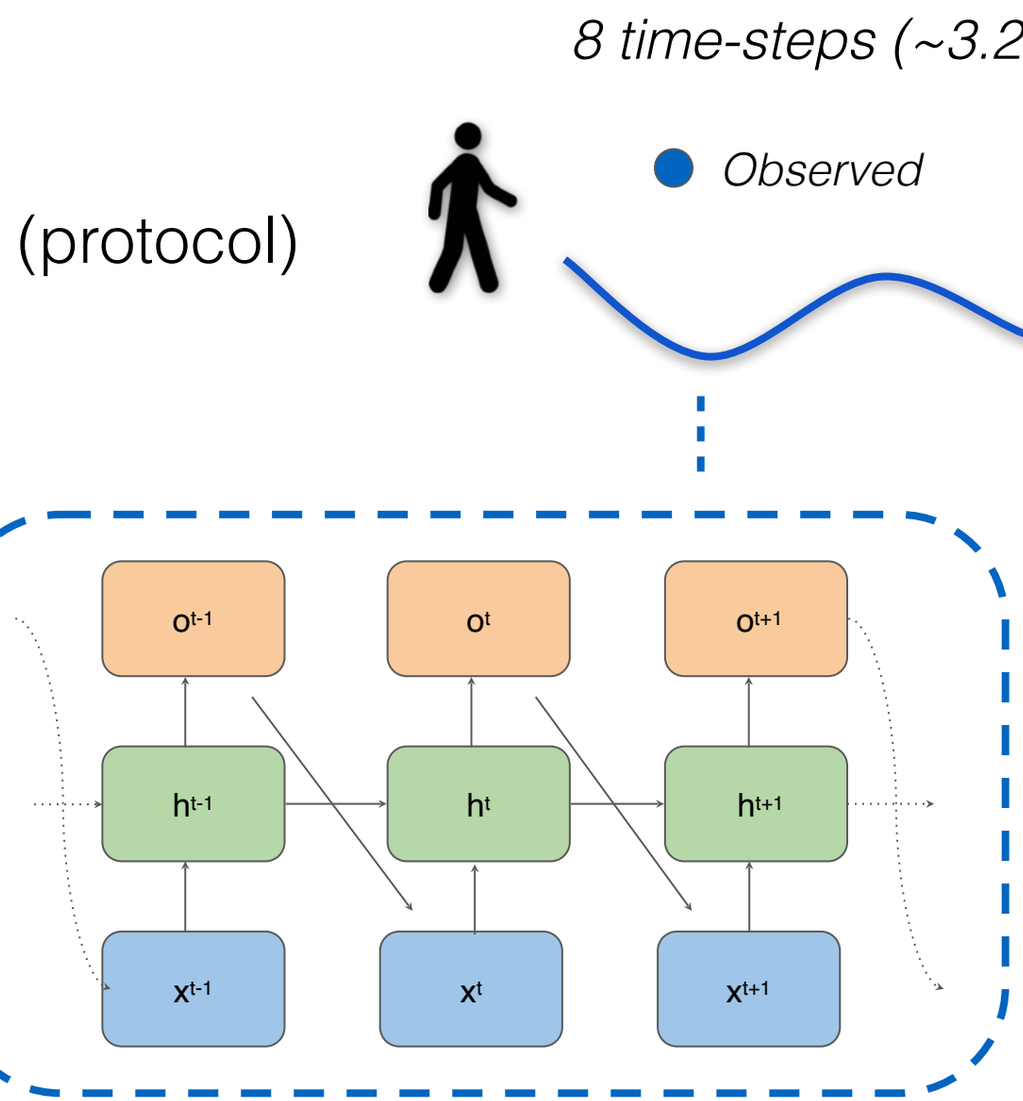
55 hours

2513 actions

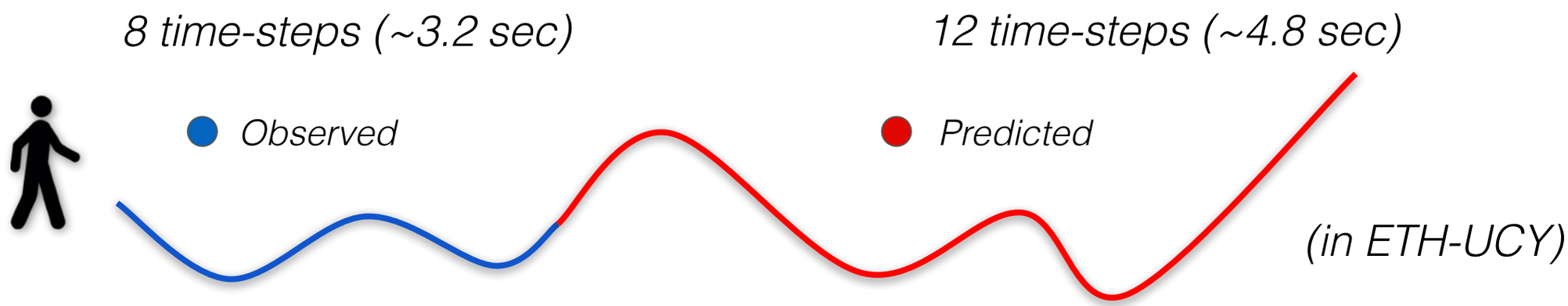
(125 verbs, 352 nouns)

Context-aware trajectory prediction

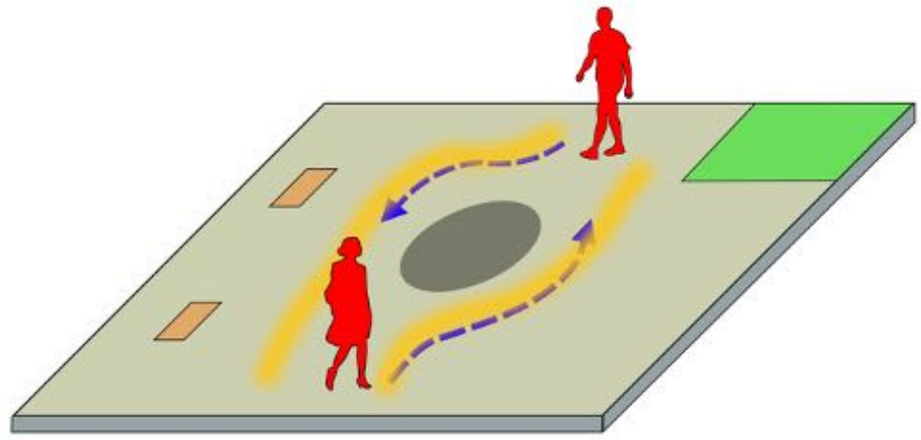
- Trajectories are modelled using RNNs/Transformers



(protocol)



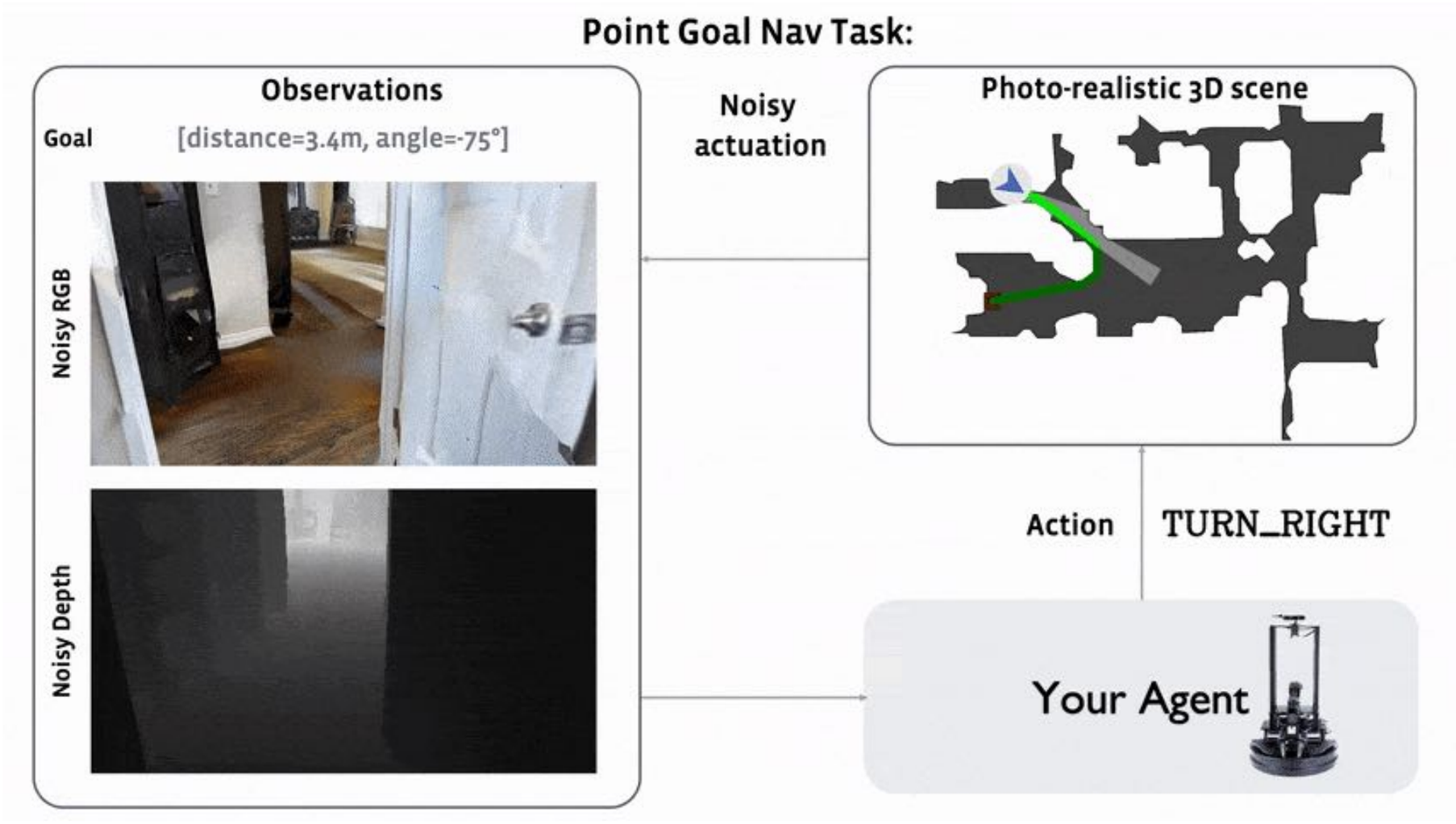
(in ETH-UCY)



- Human-space interactions
- Human-human (social) interactions

Embodied visual navigation

- Point Goal Navigation, Object Goal navigation, …



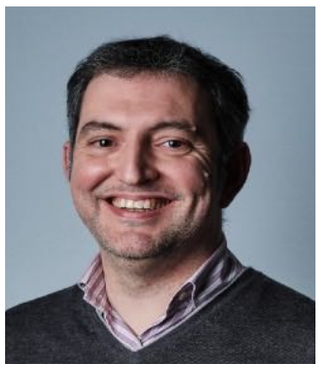
Exploiting Proximity-Aware Tasks for Embodied Social Navigation

Enrico Cancelli*
UniPD

Tommaso Campari*
UniPD, FBK

Luciano Serafini
FBK

Angel X. Chang
SFU



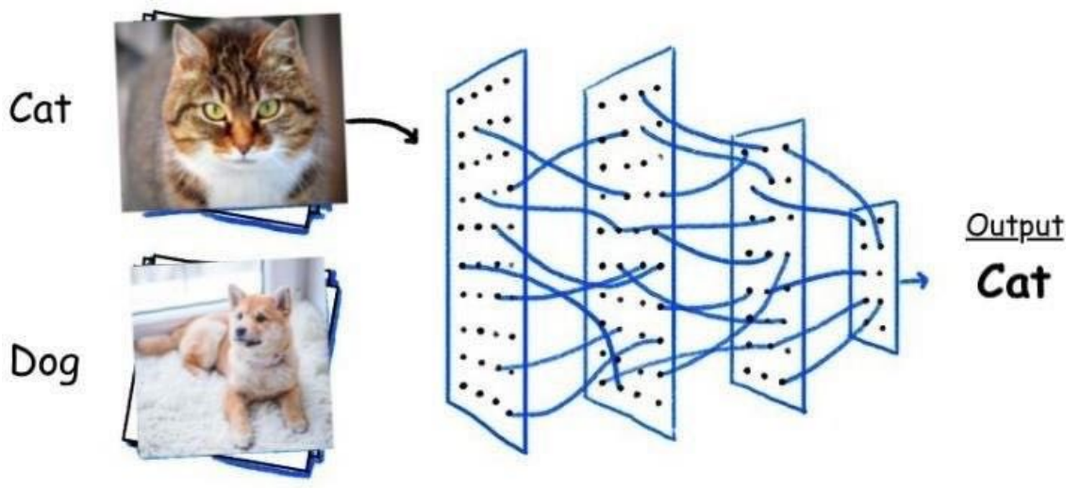
Lamberto Ballan
UniPD

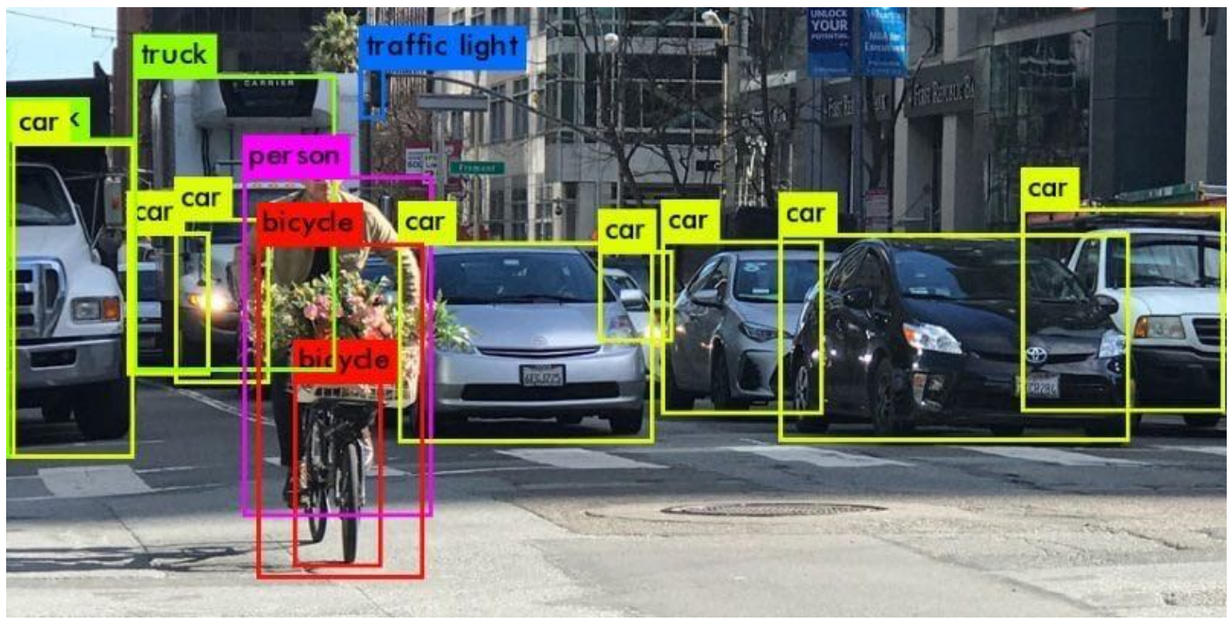
Context-aware motion prediction and embodied social navigation

[Cancelli*, Campari*, Serafini, Chang, Ballan - ICCV'23]

What is Embodied AI?

- Traditional setting in computer vision problems:
  - Image / Video samples ⟶ Classification / Regression
  - A fundamentally static scenario
    (observing and extracting information from samples)

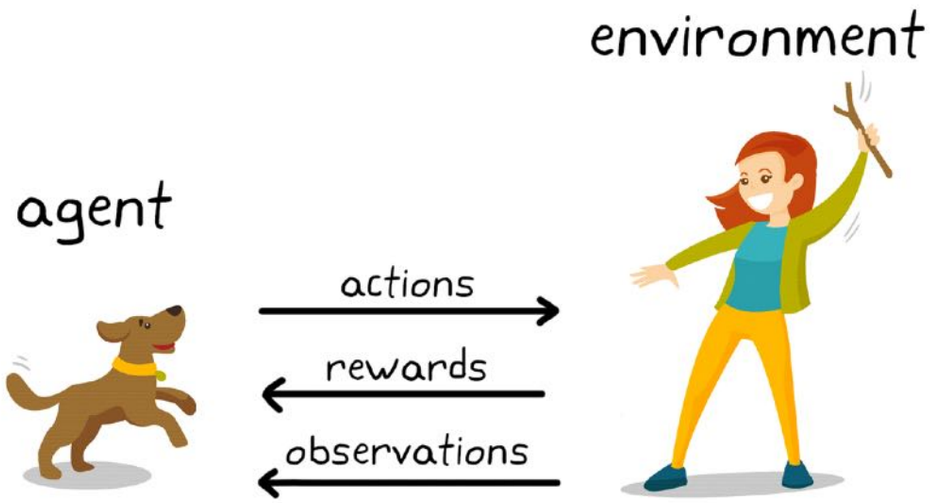


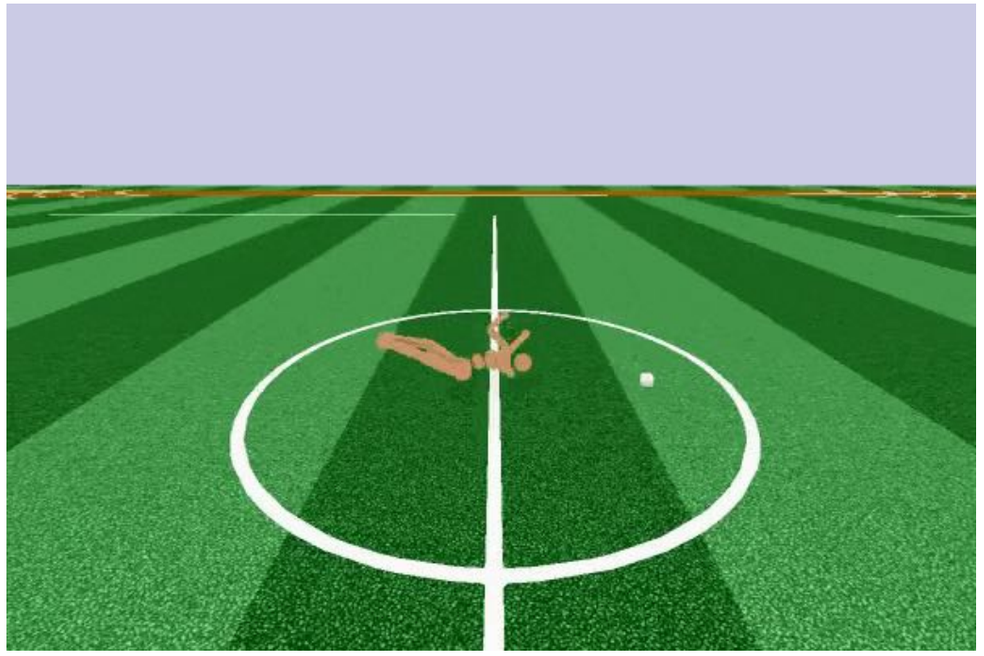


Context-aware motion prediction and embodied social navigation

What is Embodied AI?

- **Learning through interaction**: in Embodied AI, we learn by observing and acting
  - Not just an algorithm/model, but an agent that exists in a “physical” space
  - The agent explores and navigate a partially observable environment and performs actions in it



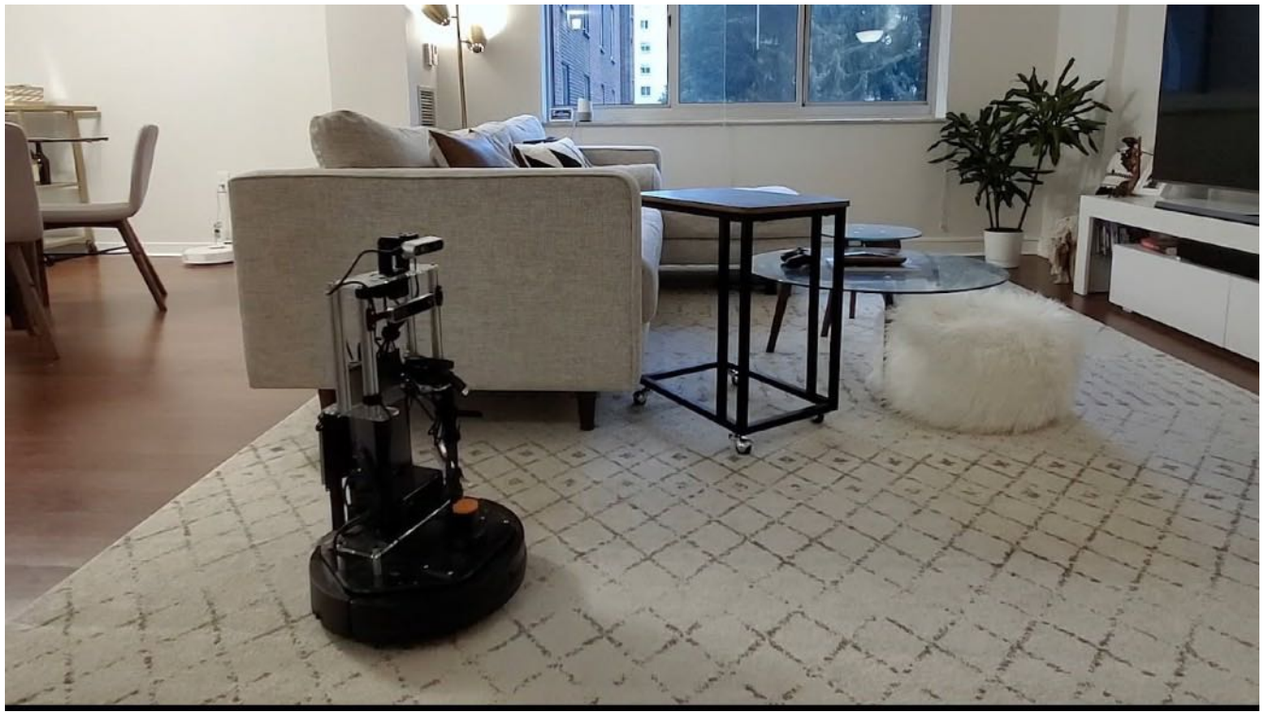


(usually deep reinforcement learning)

Context-aware motion prediction and embodied social navigation

What is Embodied AI?

- Pretty much we are talking about… Robots!
  - Main focus: **robot navigation** tasks
  - Robots are equipped with sensors such as RGB and RGB-D cameras and can navigate the scene



Context-aware motion prediction and embodied social navigation

What is Embodied AI?

- Embodied AI Workshop / community:

Context-aware motion prediction and embodied social navigation

https://embodied-ai.org/

Embodied AI: datasets/simulators

Tasks

EmbodiedQA (Das et al., 2018)

Language grounding (Hill et al., 2017)

Interactive QA (Gordon et al., 2018)

Vision-Language Navigation (Anderson et al., 2018)

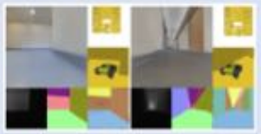
Visual Navigation (Zhu et al., 2017, Gupta et al., 2017)

Simulators

House3D (Wu et al., 2017)

AI2-THOR (Kolve et al., 2017)

MINOS (Savva et al., 2017)



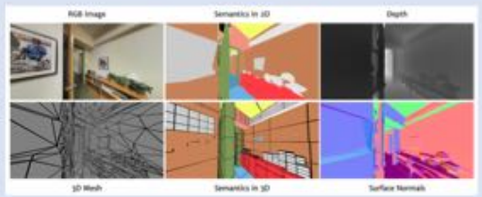
Gibson (Zamir et al., 2018)

CHALET (Yan et al., 2018)

Datasets

Replica (Straub et al., 2019)

Matterport3D (Chang et al., 2017)



2D-3D-S (Armeni et al., 2017)

Habitat Platform

Habitat API

Habitat Sim

Generic Dataset Support

Habitat

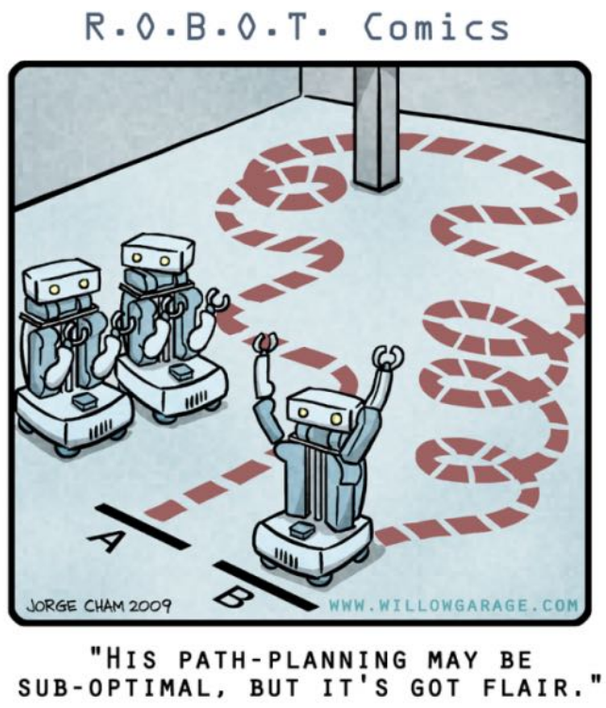
Context-aware motion prediction and embodied social navigation

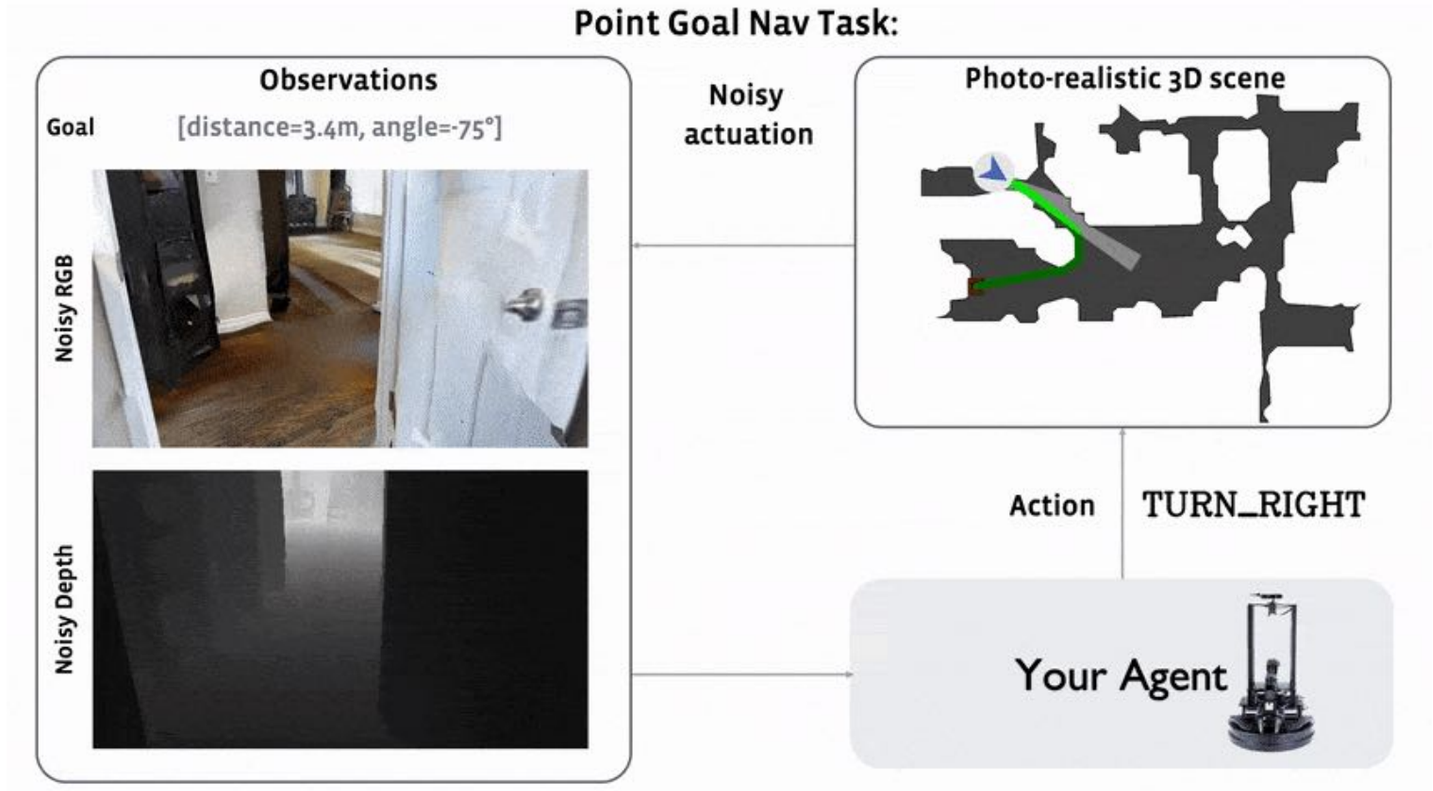
Common setting

- An episode is defined as a sequence of actions (e.g. 500), concluded by a STOP action
- A step is defined as a single action
  - MOVE_FORWARD by 25cm
  - TURN_LEFT by 30°
  - TURN_RIGHT by 30°
  - STOP concludes the episode
- GPS & Compass sensors are relative to the starting position (i.e. coordinates[0, 0, 0°])

Embodied AI: tasks

- **Point Goal navigation**: reach a given coordinate
- **Object Goal navigation**: find a specific object instance in the environment
- **Vision and language navigation**: follow some instructions in natural language

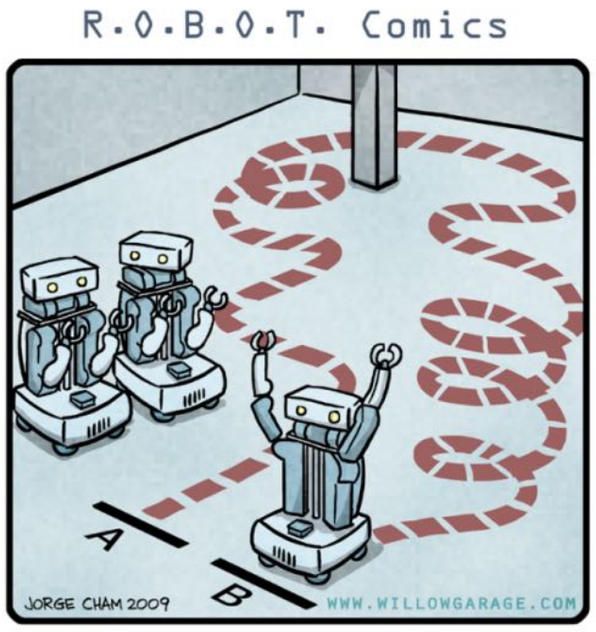


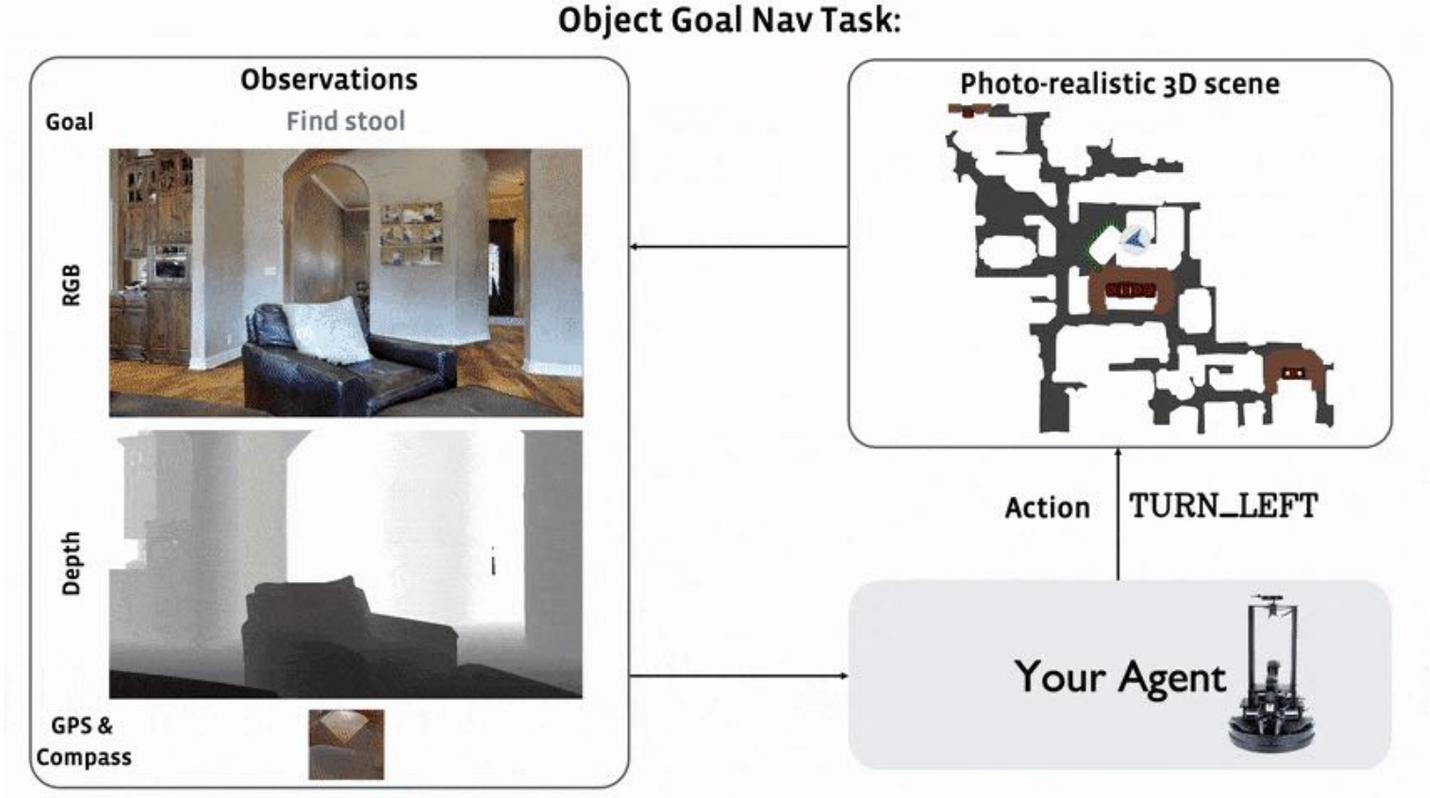


Context-aware motion prediction and embodied social navigation

Embodied AI: tasks

- **Point Goal navigation**: reach a given coordinate
- **Object Goal navigation**: find a specific object instance in the environment
- **Vision and language navigation**: follow some instructions in natural language





Context-aware motion prediction and embodied social navigation

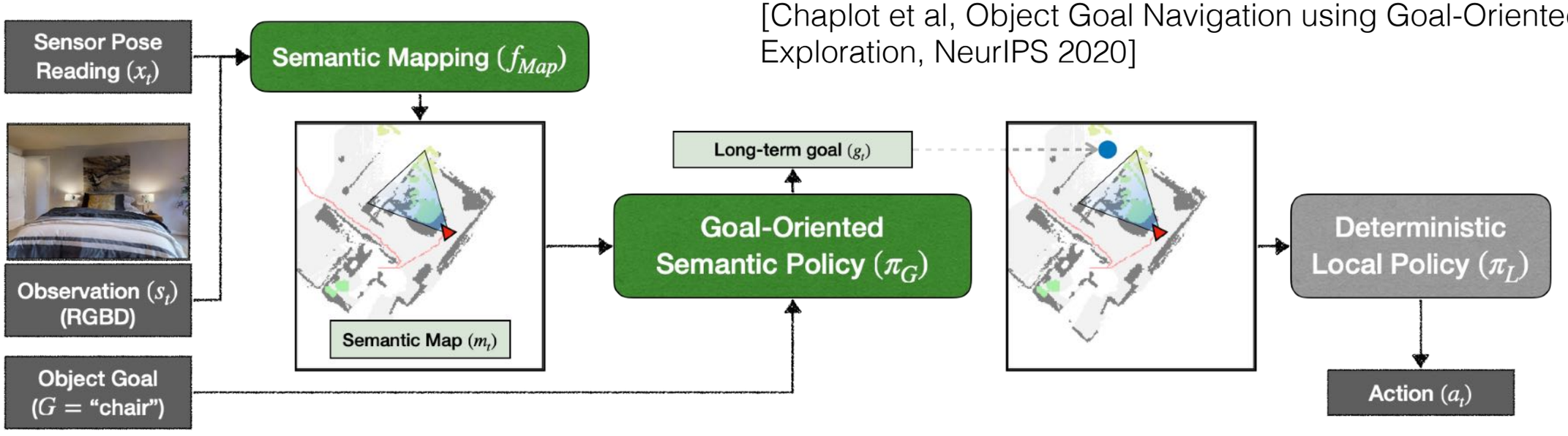
Embodied AI: evaluation / metrics

- **Success Rate**: measures the number of episodes in which the agent reached the target without collisions
- **SPL** (Success weighted by Path Length): measures the quality of the path taken by the agent by comparing the agent's path with the optimal path length
  - If the episode fails its value is 0; it's 1 if the path is optimal

Embodied AI: ~~our~~ prior work

- How can we exploit high-level semantic information to help an agent better navigate the environment?

- A couple of recent attempts:



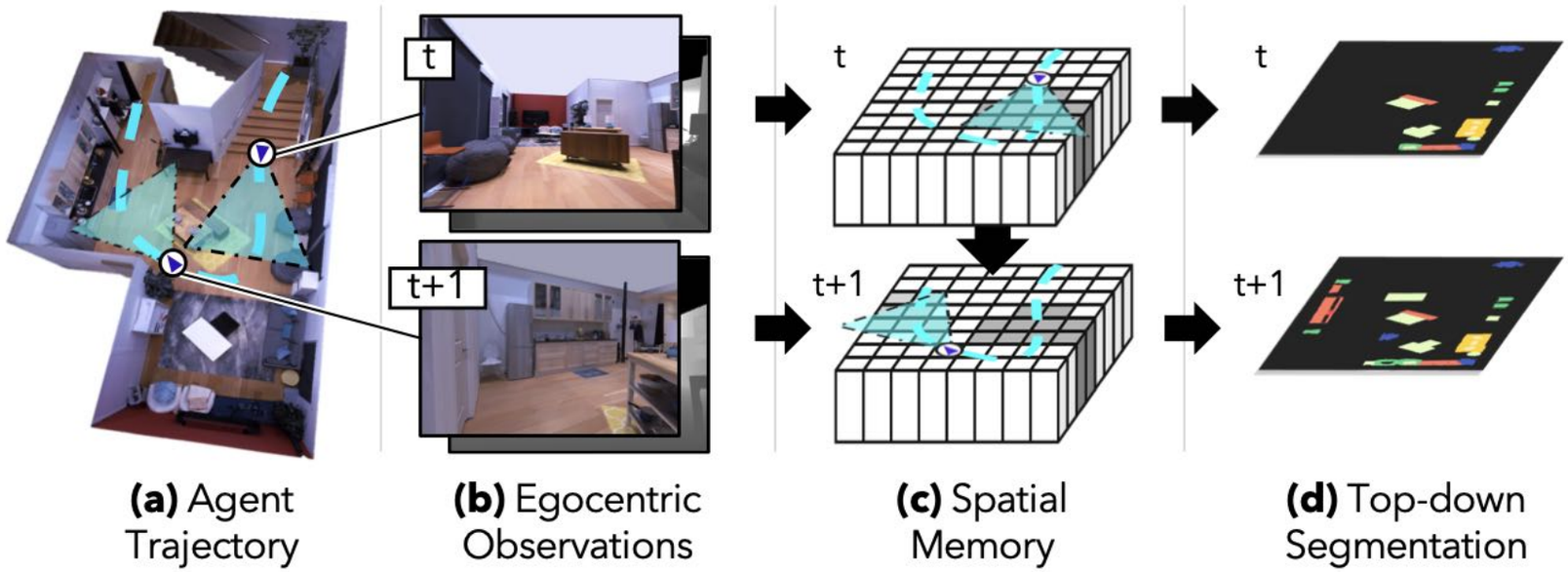
[Chaplot et al, Object Goal Navigation using Goal-Oriented Semantic Exploration, NeurIPS 2020]

- Build a semantic map
- Leverage a trained policy to find objects
- The model then uses path planning to build action plans

Embodied AI: ~~our~~ prior work

- How can we exploit high-level semantic information to help an agent better navigate the environment?
- A couple of recent attempts:

[Cartillier et al, Building Allocentric Semantic Maps and Representations from Egocentric Views, AAAI 2021]



- Build semantic maps in an end-to-end fashion
- Maps are saved and reused for Object Goal Navigation

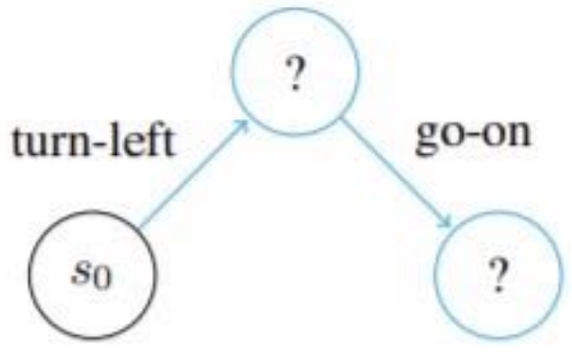
Embodied AI: our prior work

- How can we exploit high-level semantic information to help an agent better navigate the environment?
- Our approach aims at better exploiting:
  - High-level semantic information
  - Common-sense priors (pre-defined and, eventually, learned / updated in a data-driven fashion)
  - Reuse of previously acquired knowledge

Embodied AI: our prior work

- Core idea / intuition:
  - New environment: explore and build a knowledge base
  - Otherwise, let's exploit the knowledge base and navigate

No

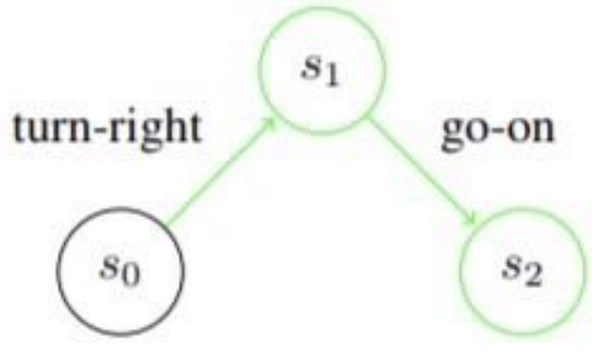


ObjectGoal Nav: "where is the oven?"

Embodied AI: our prior work

- Core idea / intuition:
  - New environment: explore and build a knowledge base
  - Otherwise, let's exploit the knowledge base and navigate

Yes

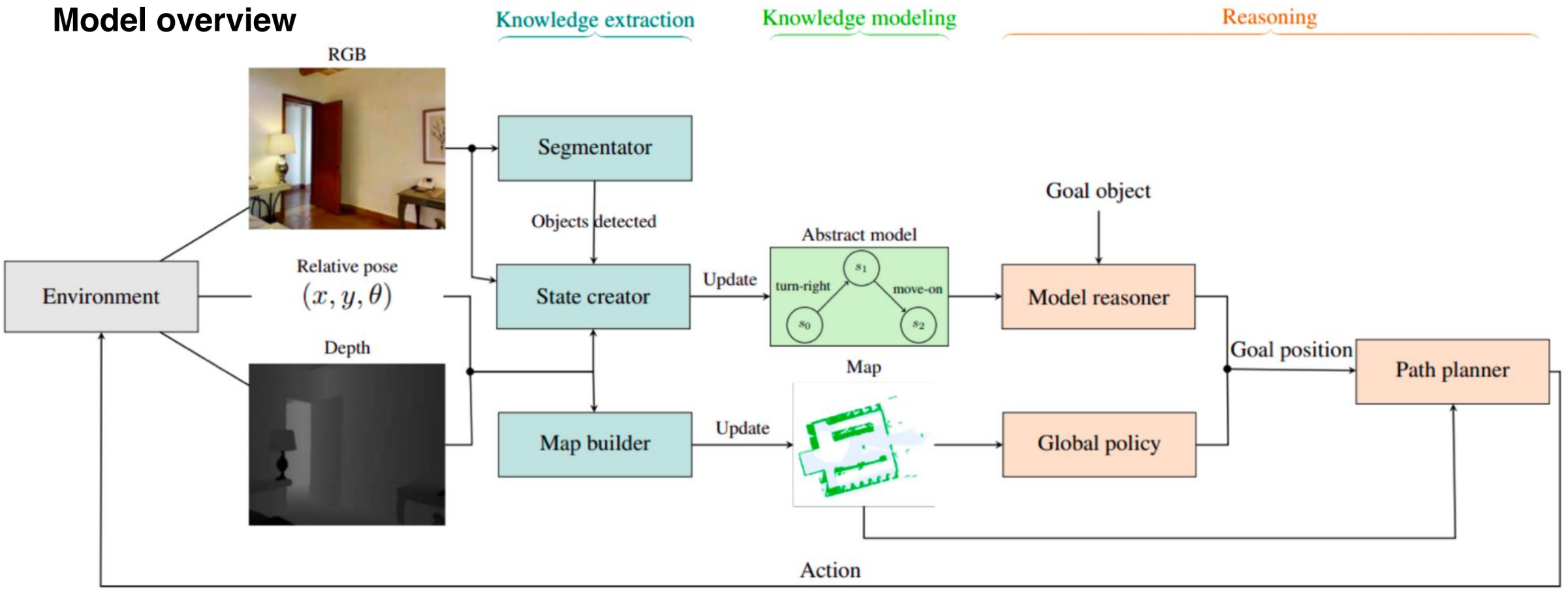


ObjectGoal Nav: "where is the oven?"

Embodied AI: our prior work

- Core idea / intuition:
  - New environment: explore and build a knowledge base
  - Otherwise, let's exploit the knowledge base and navigate

Model overview

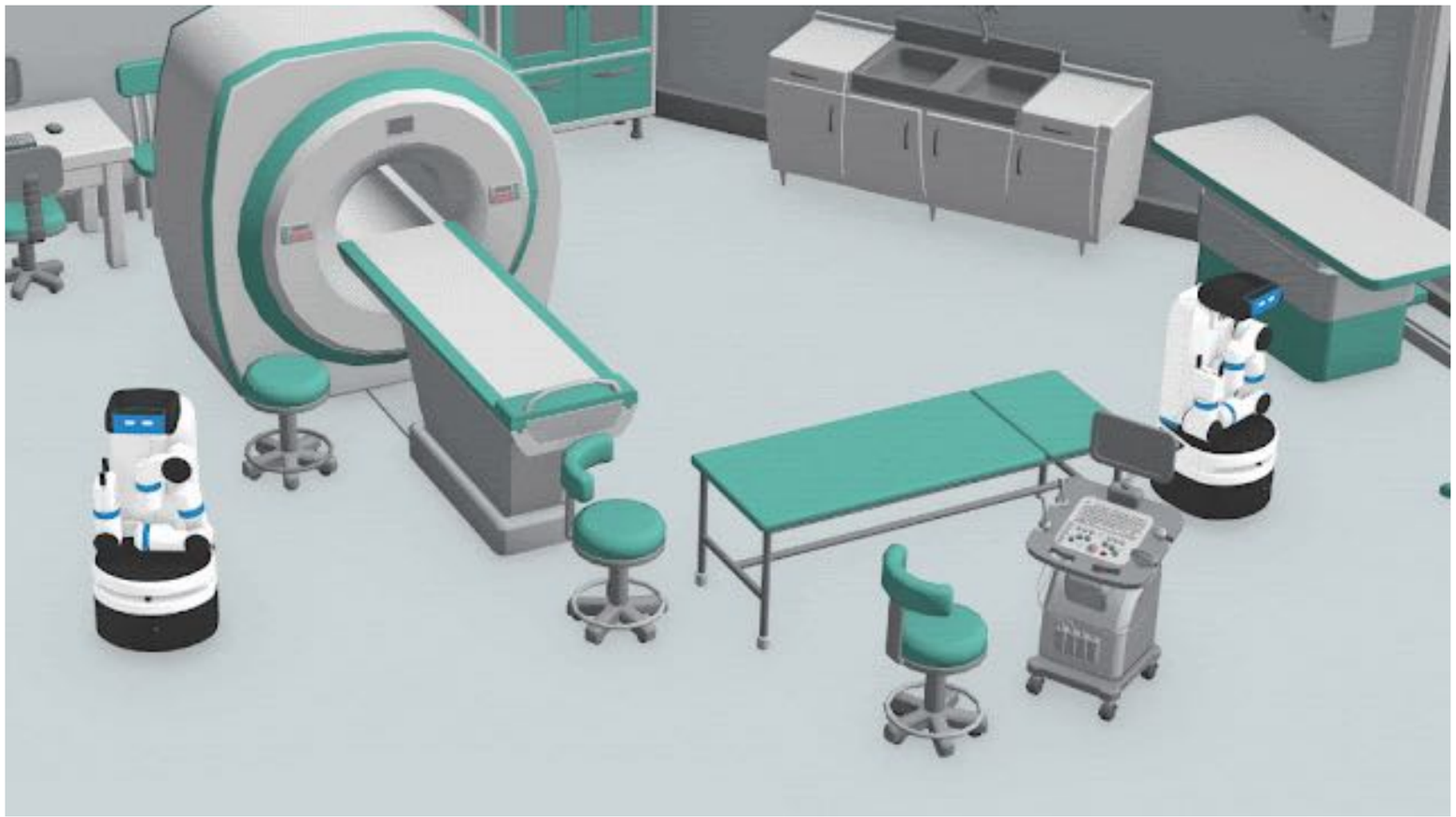


Context-aware motion prediction and embodied social navigation

[Campari, Lamanna, Traverso, Serafini, Ballan - CVPR'22]

What is missing?

- Although challenging, these scenarios are still **static**
- What about humans and other active agents?



Human-aware / Social Navigation

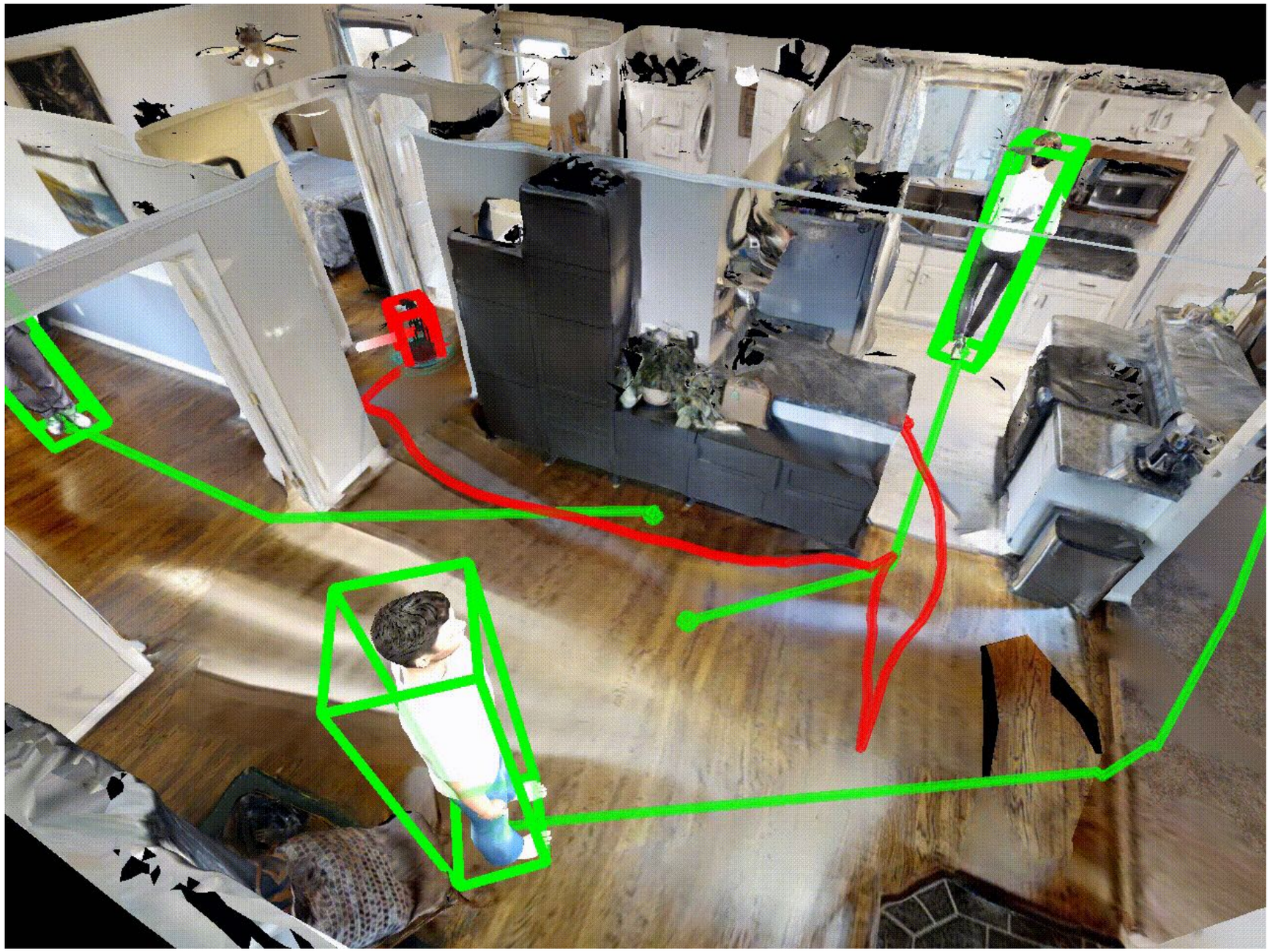
- Robots navigate in crowded environments
  - When designing an agent we should take into account humans' safety and interactions

- A new task: **Social Navigation**
  - Similar to Point Goal navigation
  - A certain number of people are randomly spawn in the area
  - If the agent hits a person the episode ends

Social Navigation benchmarks

- Gibson 4+ dataset:
  - 64 training scenes (50k episodes per env) + 8 validation and 14 testing scenes
- We introduce the HM3D-S (large-scale) dataset:
  - 800 training scenes (10k episodes per env) + 30 validation and 70 testing scenes
  - It is built on top of HM3D

What we would like to see…



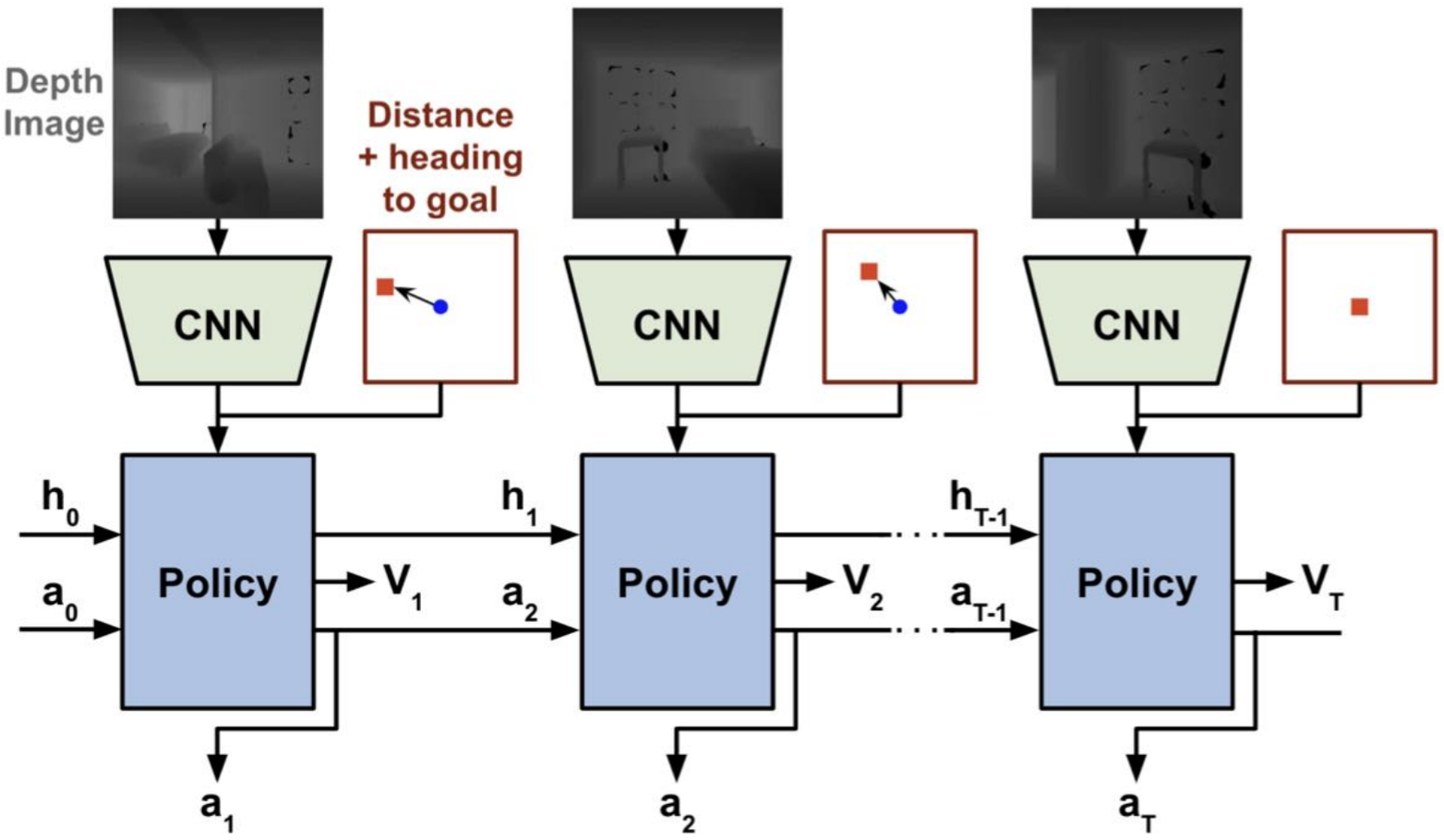
SocialNav

(a concrete example)

(*actually this is a concrete example obtained with our model :)

SocialNav “baseline”

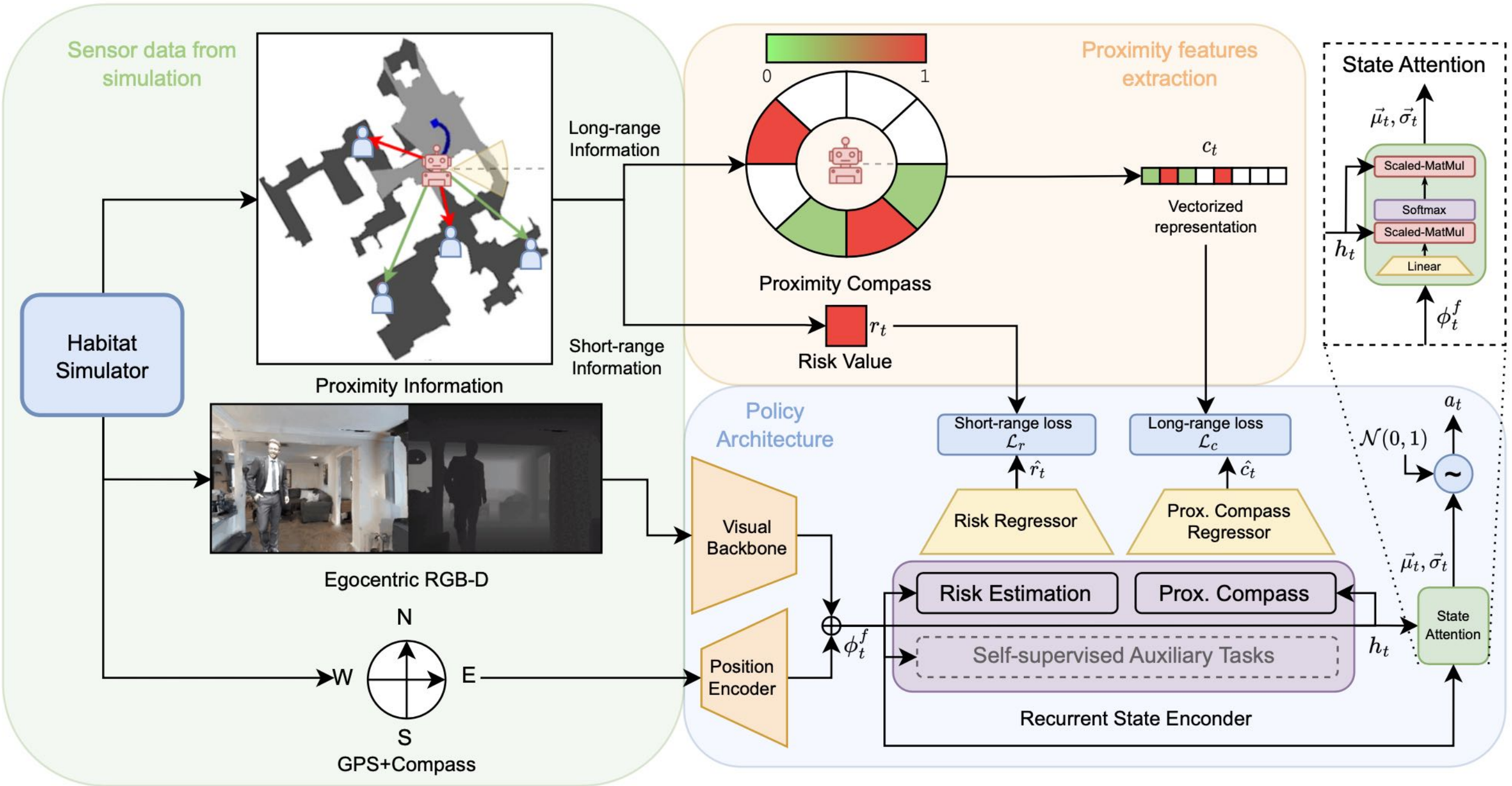
- Simple baseline recently presented at IROS’22 (and winner of the last SocialNav challenge)



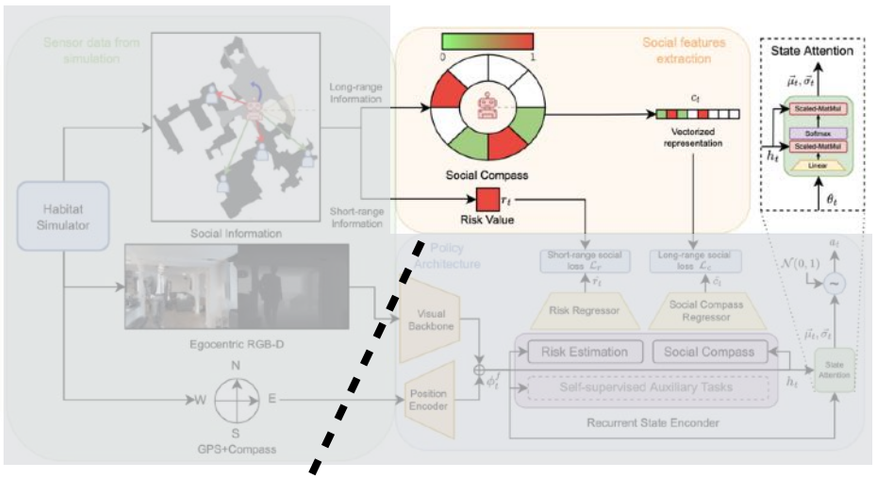
- Input: Depth + GPS and compass sensors
- No social information
- Evaluated only in terms of success rate

[Yokoyama et al, IROS 2022]

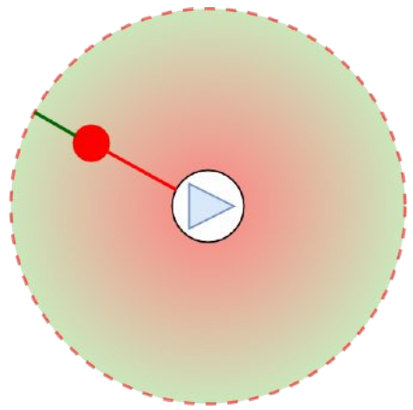
Our model


Sensor data from simulation
Habitat Simulator
Proximity Information
Long-range Information
Short-range Information
Egocentric RGB-D
N
W
E
S
GPS+Compass
Proximity features extraction
0
1
Proximity Compass
r_t
Risk Value
c_t
Vectorized representation
State Attention
$\vec{\mu}_t, \vec{\sigma}_t$
Scaled-MatMul
Softmax
Scaled-MatMul
Linear
h_t
ϕ_t^f
Policy Architecture
Visual Backbone
Position Encoder
Short-range loss $\mathcal{L}_r$
Long-range loss $\mathcal{L}_c$
$\hat{r}_t$
$\hat{c}_t$
Risk Regressor
Prox. Compass Regressor
Risk Estimation
Prox. Compass
Self-supervised Auxiliary Tasks
Recurrent State Enconder
a_t
$\mathcal{N}(0,1)$
~
$\vec{\mu}_t, \vec{\sigma}_t$
State Attention

Social features

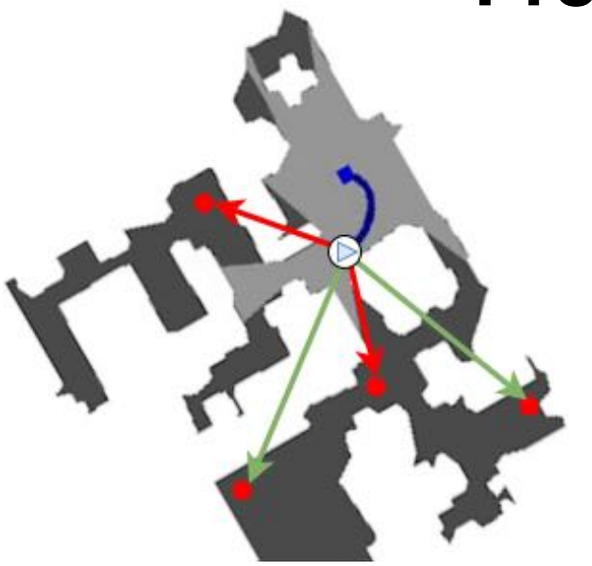


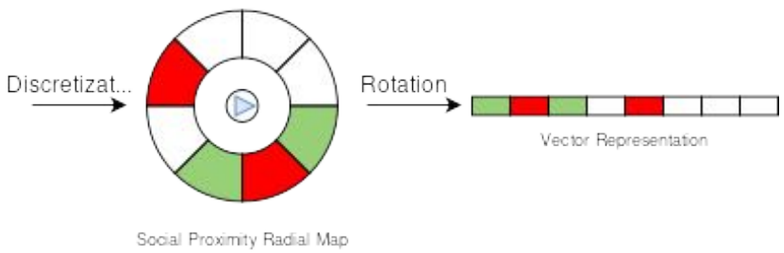
Risk value



- Short ranged
- From distance with closest person
- 0 if more than k meters
- 1 when colliding

Proximity compass

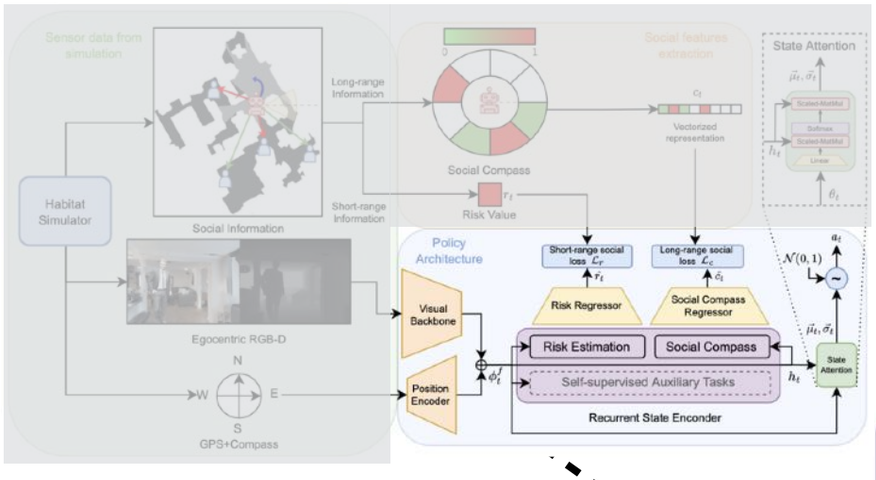


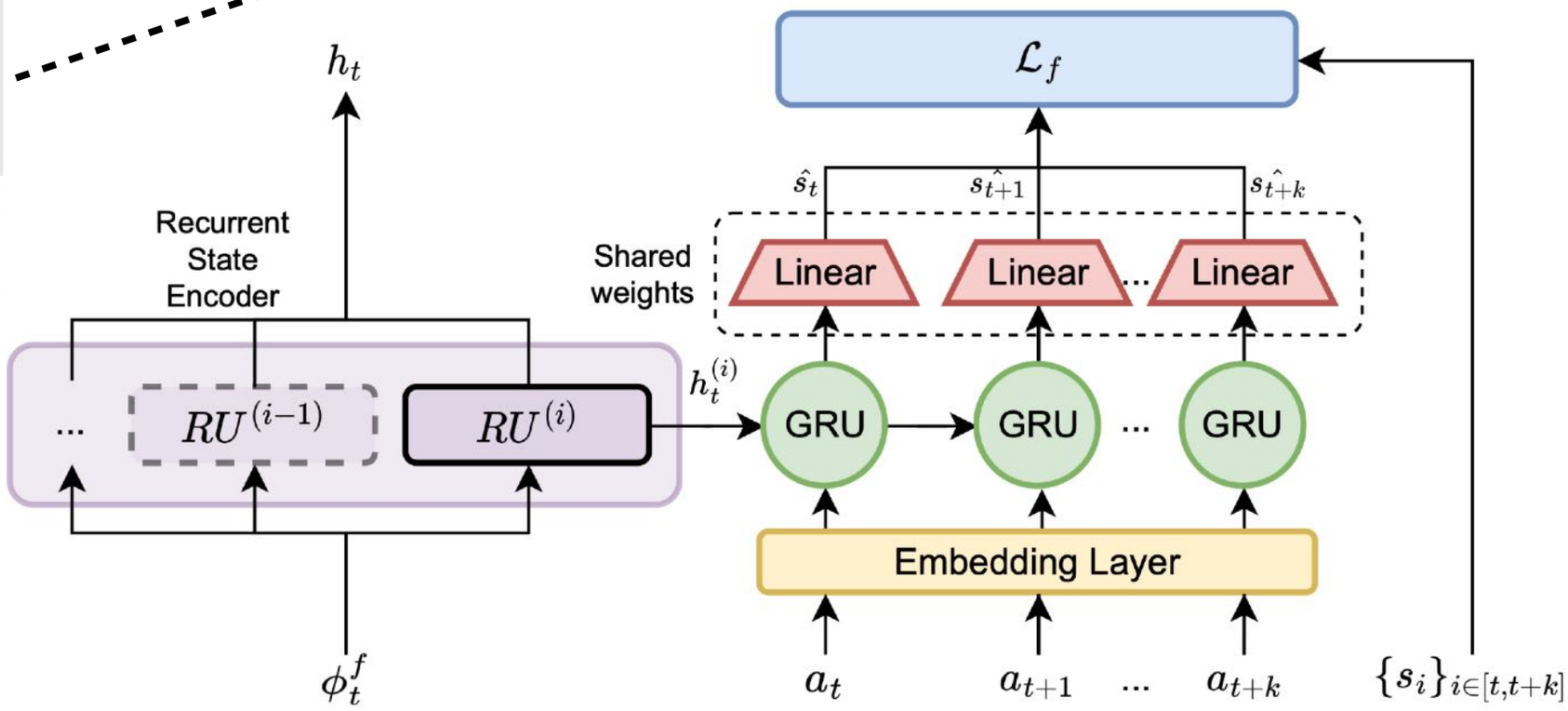


- Long ranged
- Divided in sectors
- Each sector has a value
- Distance+direction

Proximity-aware tasks

(based on **risk** or **proximity compass**)





- Predicting social features k steps in the future
- Using beliefs from RUs and the sequence of actions a from t to t+k

- Additionally, we might have also “generic” self-supervised tasks *

* [Ye et al, ICCV 2021, CoRL 2021]

Experiments: evaluation / metrics

- **Success Rate**: measures the number of episodes in which the agent reached the target without collisions
- **SPL** (Success weighted by Path Length): measures the quality of the path taken by the agent by comparing the agent's path with the optimal path length
  - If the episode fails its value is 0; it's 1 if the path is optimal
- **Human-collisions**: measures the number of episodes where the agent collided with an human

Experimental results

- Comparison with the state-of-the-art:

Name	Sensors		Aux Tasks			Social Tasks		Metrics (Gibson4+)			Metrics (HM3D-S)		
	RGB	Depth	CPCA	GID	CPCA/B	Risk	Compass	Success	SPL	H-Collisions	Success	SPL	H-Collisions
Baseline [44]		✓						72.65±1.6	47.43±1.2	24.35±1.9	62.76±2.2	36.69±1.1	29.29±2.2
Baseline+RGB [44]	✓	✓						74.28±1.8	44.84±0.7	23.78±1.3	61.43±0.5	34.84±0.6	29.23 ± 0.7
Aux tasks [43]	✓	✓	✓	✓	✓			73.4±2.0	52.08±1.4	23.40±1.5	63.62±1.6	42.27±1.2	24.79±2.2
Risk only	✓	✓				✓		74.90±1.7	50.25±1.1	22.56±1.2	66.22±1.2	45.26±0.8	24.47±1.7
Compass only	✓	✓					✓	75.08±1.5	50.55±1.0	22.49±1.1	67.32±1.7	45.74±1.0	23.54±1.7
Aux + risk	✓	✓	✓	✓	✓	✓		75.61±1.8	51.43±0.2	21.04±1.4	68.16±0.8	45.64±0.2	22.00±1.6
Aux+compass	✓	✓	✓	✓	✓		✓	75.63±1.2	52.60±1.6	23.17±1.2	67.94±1.4	45.76±1.0	23.78±2.0
Social tasks	✓	✓				✓	✓	76.6±1.8	52.81±1.2	20.47±0.4	68.35±0.5	45.83±0.5	**21.72±1.2**
Social + Aux tasks	✓	✓	✓	✓	✓	✓	✓	**77.24±1.1**	**55.23±1.4**	**19.50±1.0**	**70.16±1.1**	**47.60±1.0**	22.09±1.3

Table 1. Social Navigation evaluation on Gibson4+ and HM3D-S. For each model are listed the type of input data it uses (*Sensors* column) and, eventually, what kind of self-supervised *Aux tasks* or *Social tasks* the model employs. The metrics reported are *Success* rate, *SPL* and Human-Collisions Rate (*H-collisions*).

[44] Yokoyama et al, IROS 2022

[43] Ye et al, CoRL 2021

Evaluation protocol for SocialNav

- The common setting has some major limitations
  - Success and SPL are not sufficient to properly evaluate Social Navigation
  - We need to better understand and evaluate why (and where) an agent collides with an human

- We introduce a new protocol for evaluating human-agent encounters (based on trajectories and visibility)*

* inspired by [Pirk et al, ICRA-W 2022]

Evaluation protocol for SocialNav

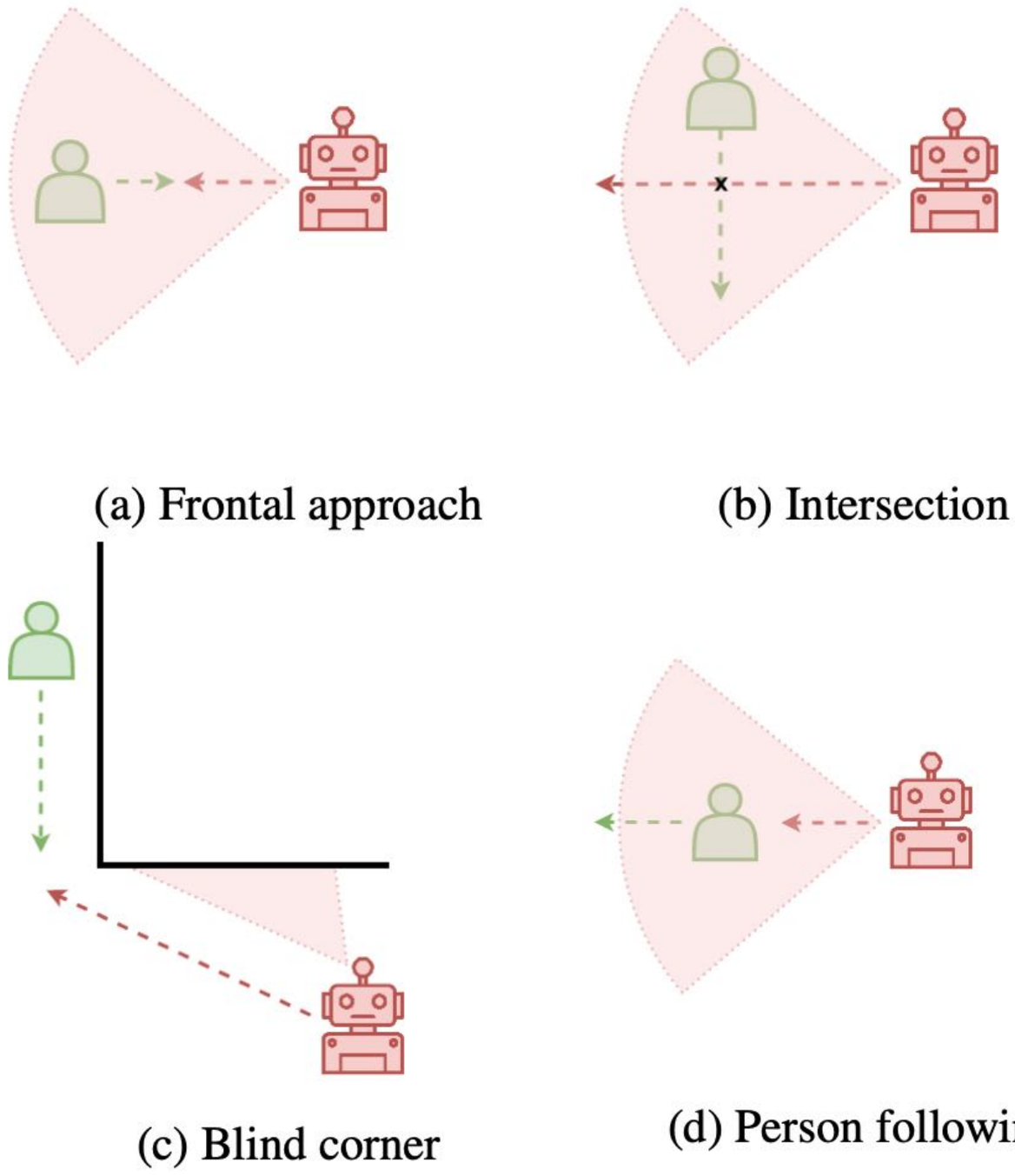
- We define 4 classes:

a) The agent and a person move towards each other

b) Paths intersect at 90 degrees

c) The agent approaches a corner, suddenly seeing a person

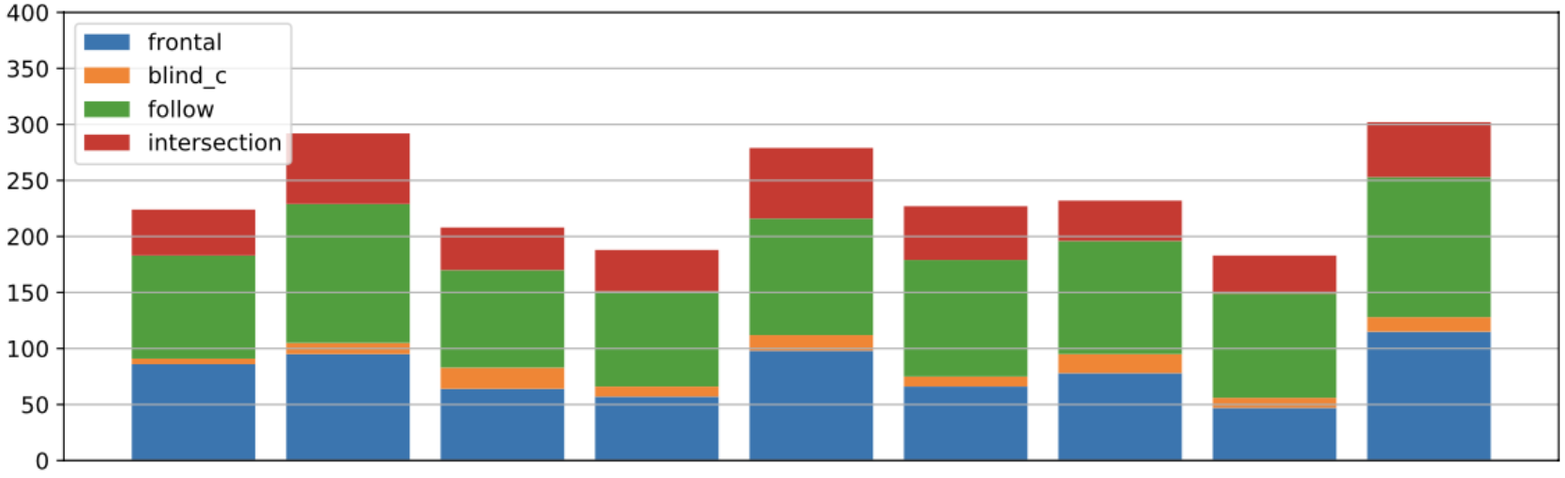
d) The agent follows a person



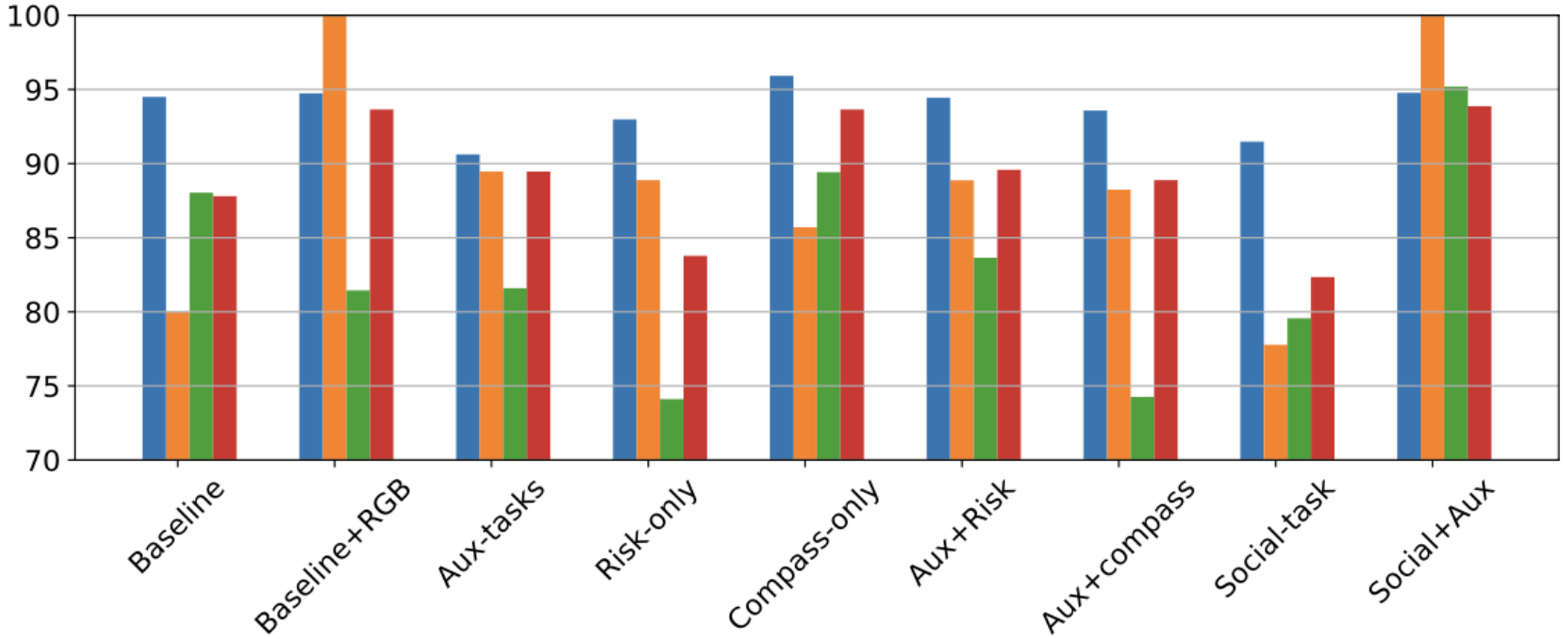
Evaluation protocol for SocialNav

- We then introduce the following (per-class) metrics:
  - **Number of encounters**
  - **ESR** (encounter survival rate): rate of encounters concluded with no human collisions
  - **ALV**: average linear velocity (also ALV@T for time-conditioned metrics)
  - **AD**: average agent-person distance (also AD@T for time-conditioned metrics)

“Fine-grained” results



(a) Number of encounters



(b) Encounter Survival Rate (ESR)

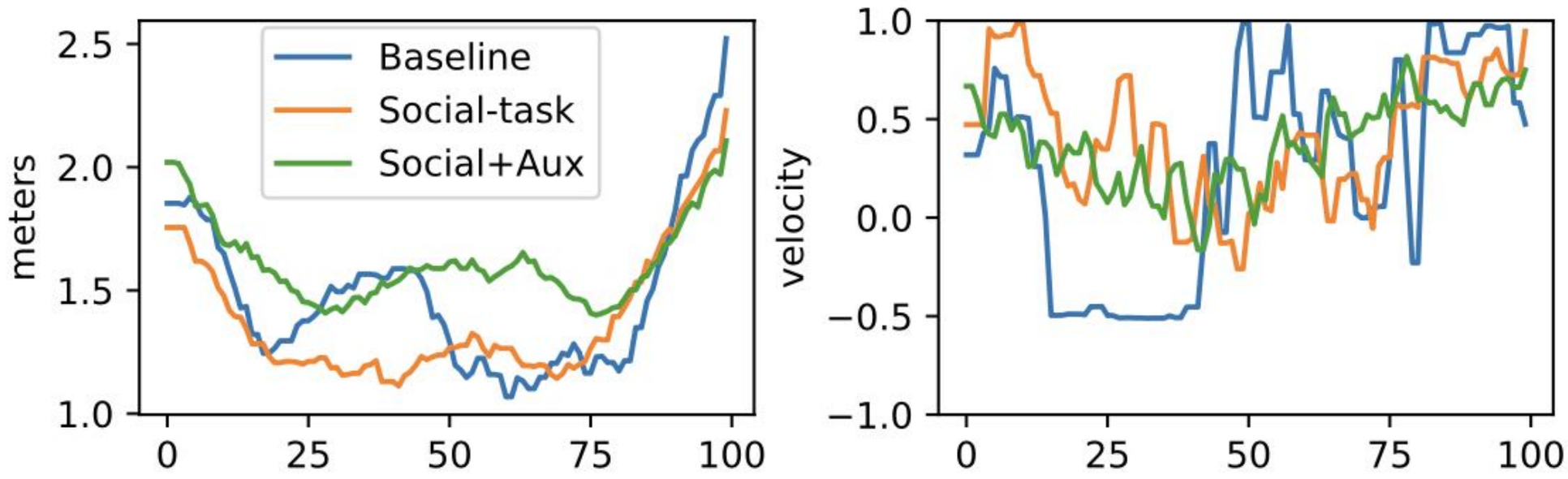
Two types of behaviour:

Encounter elusion (low #encounters, medium ESR) (Aux-tasks, Risk-only, Social-tasks)

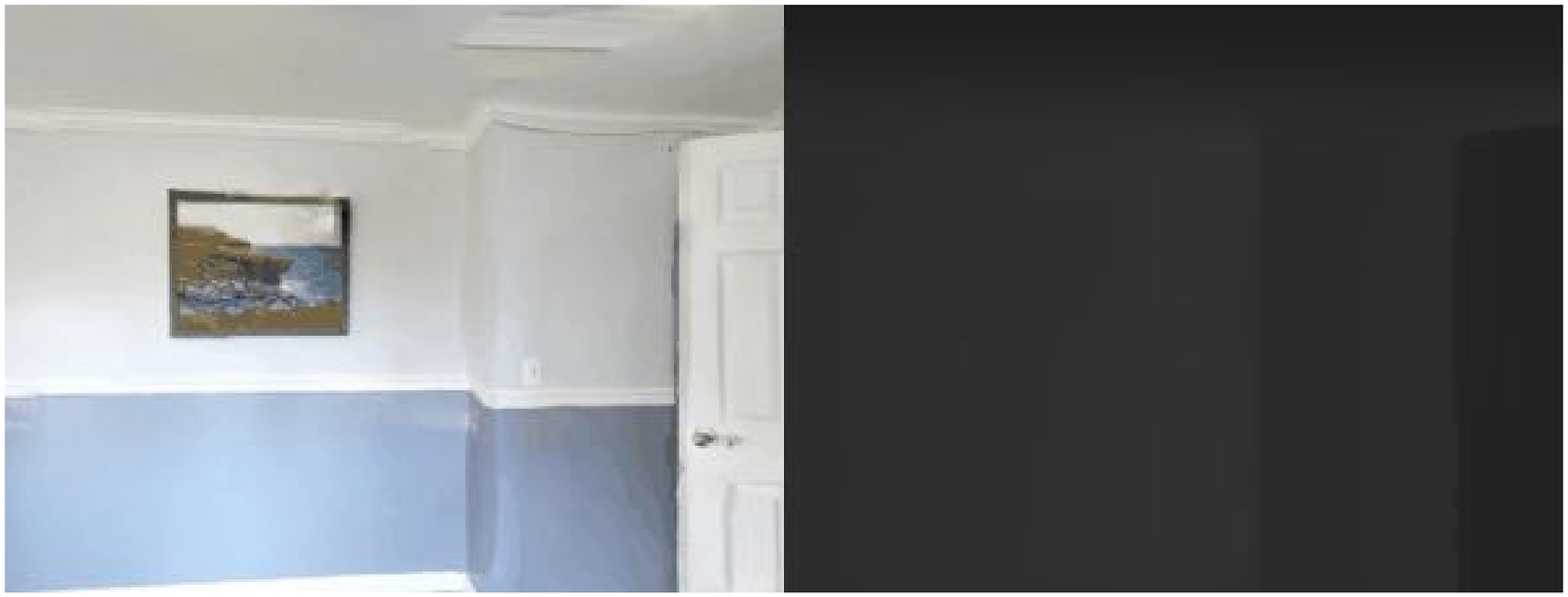
Encounter avoidance (high #encounters, high ESR) (Baseline+RGB, Compass-only, Social+Aux)

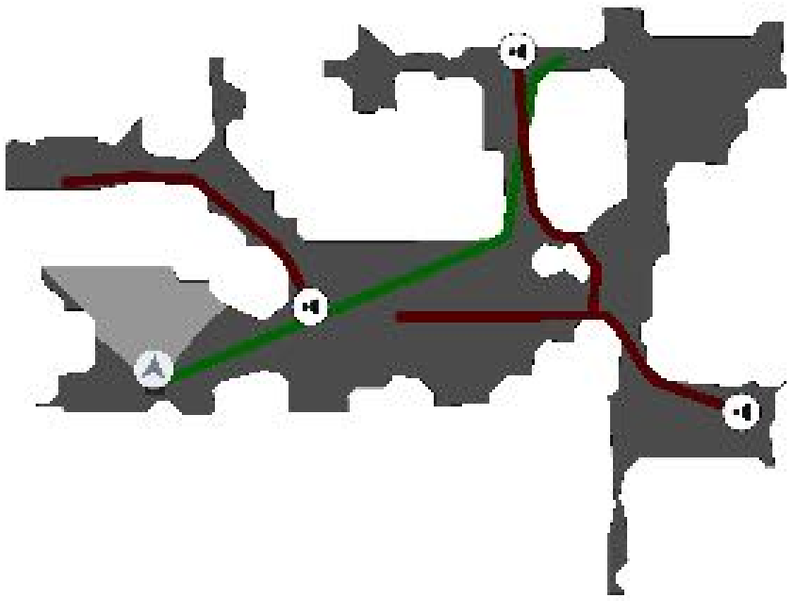
“Fine-grained” results

- Blind corner class (the most difficult one):
  - ALV and AD measured on normalized intervals (e.g., on 19 to 20% of all encounters)
  - Instability on ALV curves, Baseline is significantly less smooth than the others, with lower peaks
  - Also AD curve is less stable for Baseline, Social+Aux generally maintains an higher distance



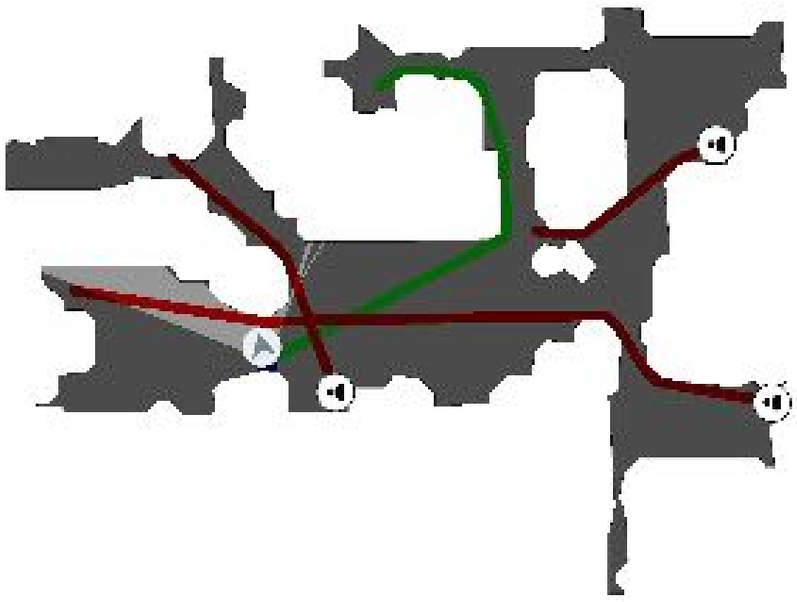
Qualitative results





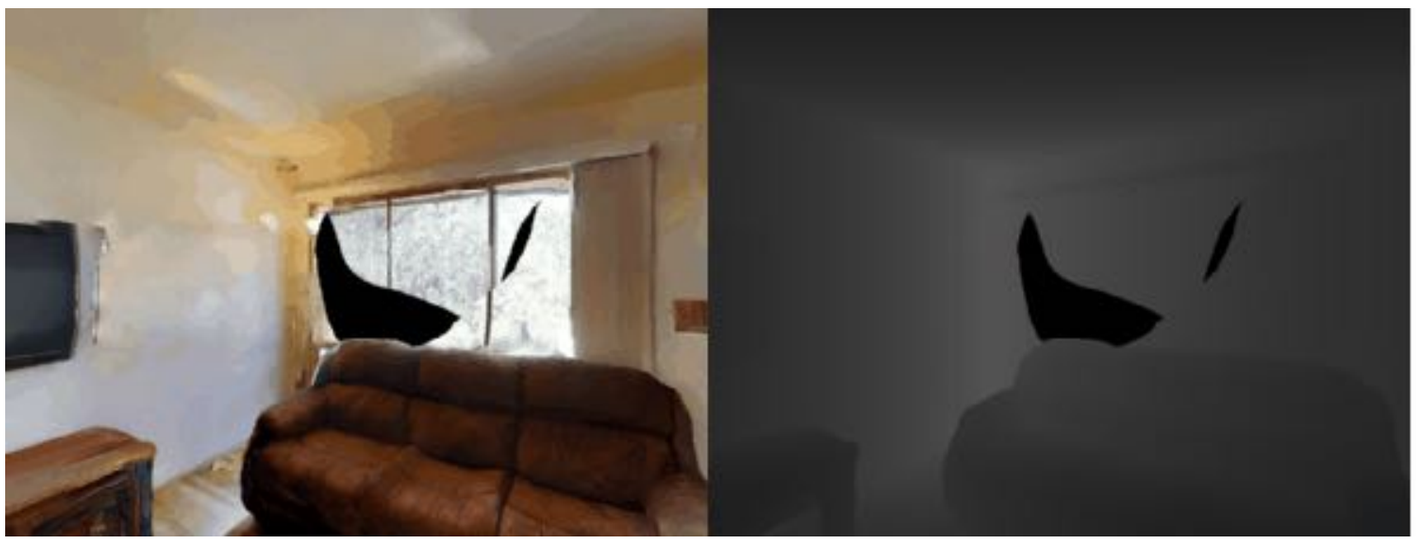
Successful episode

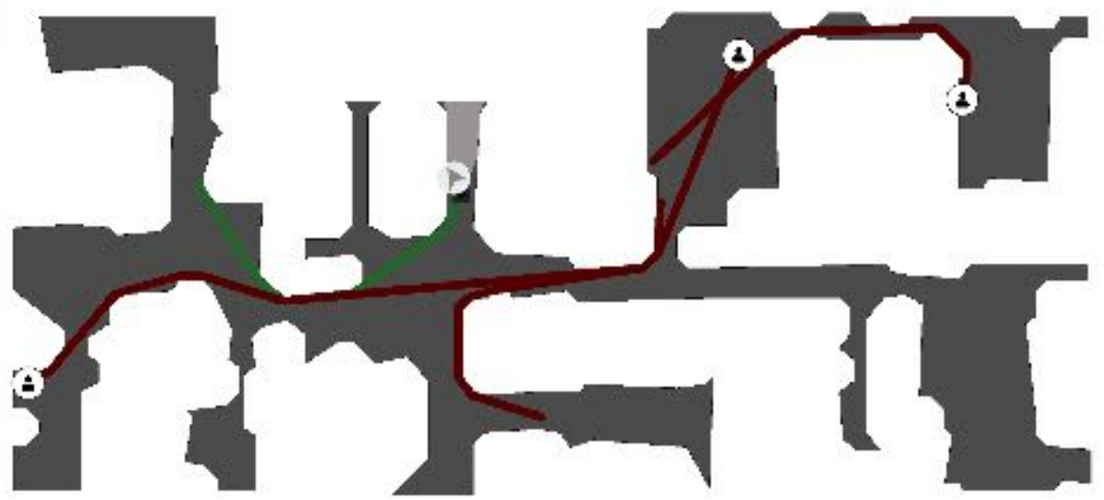
Qualitative results



“Suboptimal” but successful episode

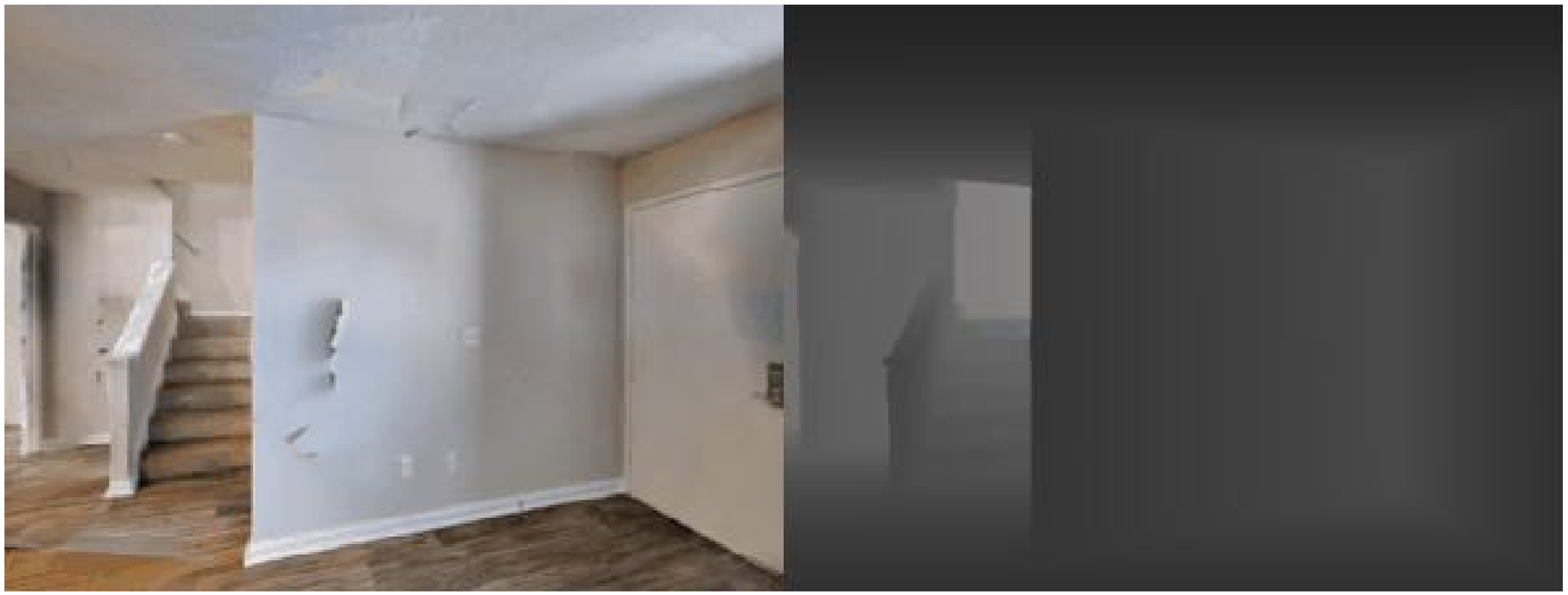
Qualitative results

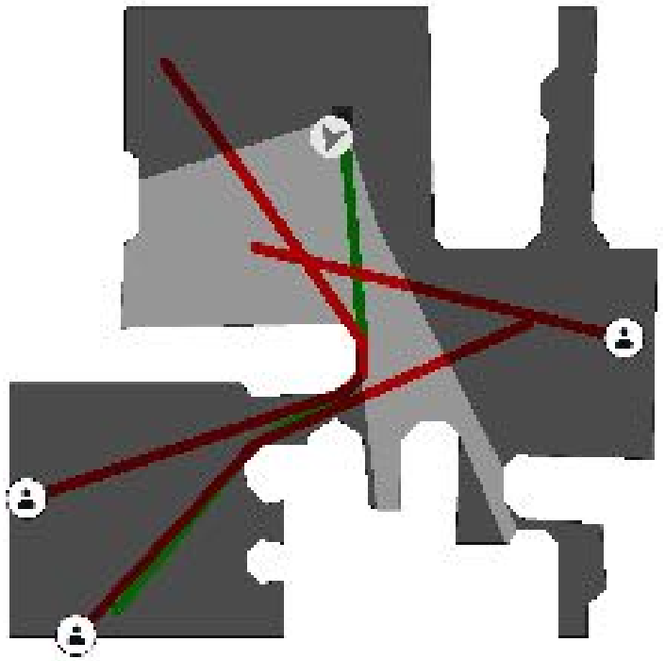




Failure case: the agent is stuck

Qualitative results





Failure case: the agent hits a person

Contact

✉ lamberto.ballan@unipd.it

🏠 http://www.lambertoballan.net

🏠 http://vimp.math.unipd.it

@ twitter.com/lambertoballan

Thanks!

Context-aware motion prediction and embodied social navigation

VCS Course 23-24
Embodied AI

Current Research Project

Filippo Ziliotto

What's Embodied AI ?

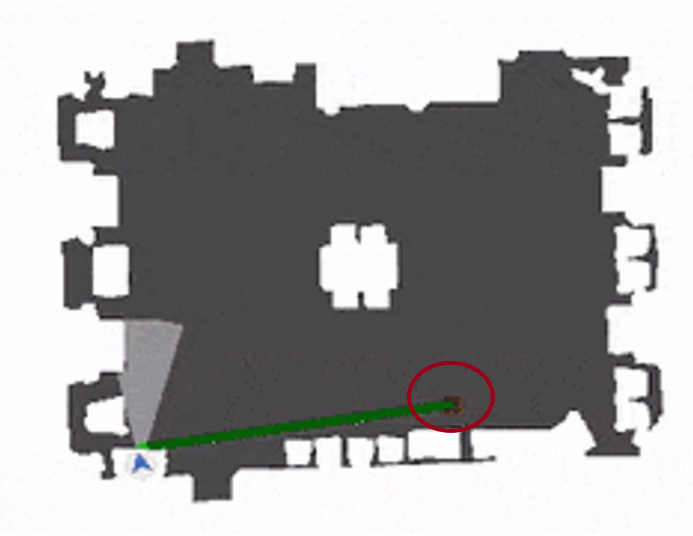
Agent that moves in environment:

- 3D Environment
- Robot(s)
- Human(s) (optional)

You also need a goal:

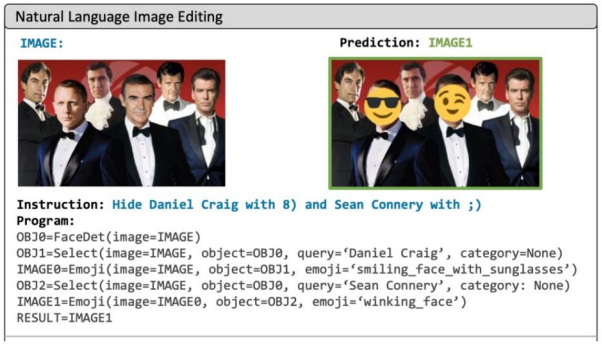
- (x,y,z) point
- object label
- Image

RGB **Depth** **Map** (optional)



What's my Research ?

The research is based on a paper of 2023



Visual Programming: Compositional visual reasoning without training

Tanmay Gupta, Aniruddha Kembhavi PRIOR @ Allen Institute for AI

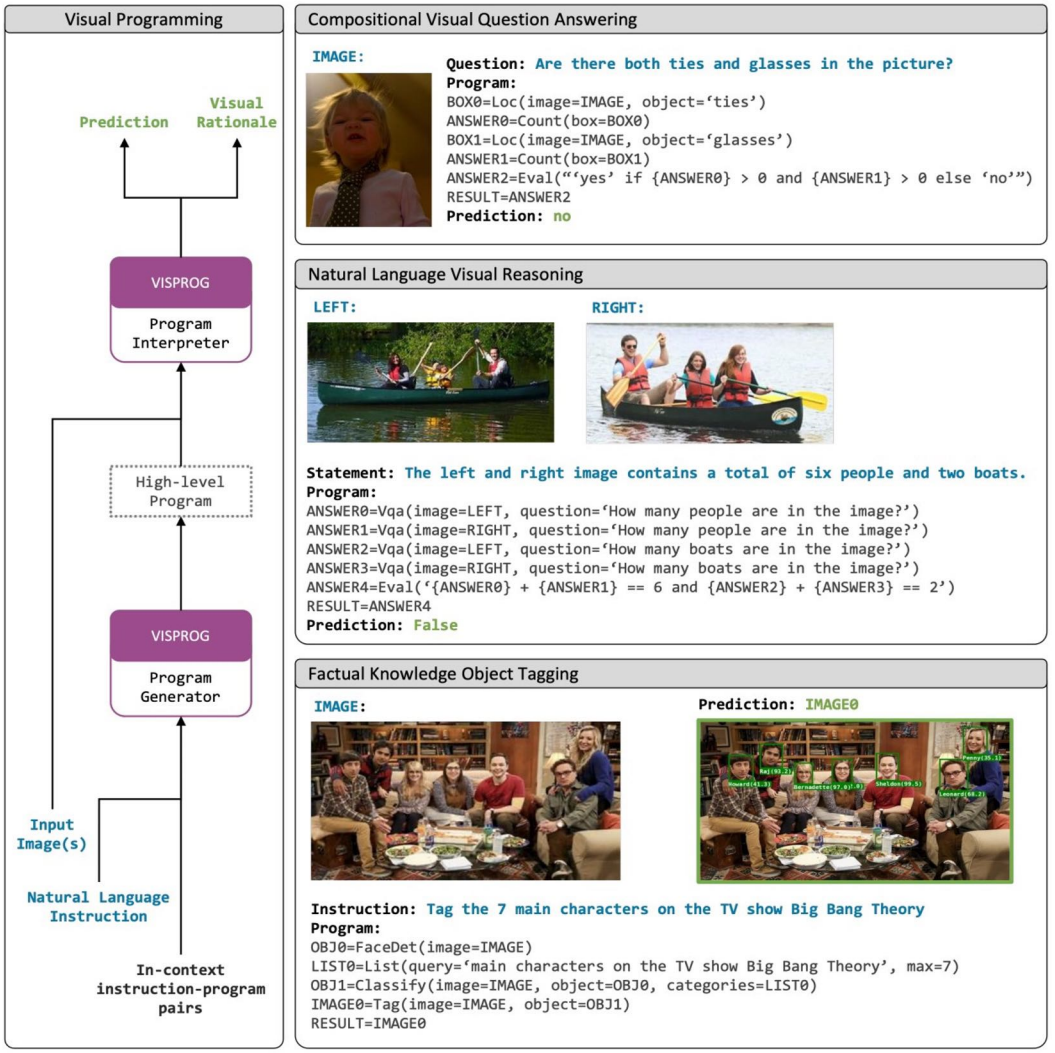
https://prior.allenai.org/projects/visprog

CVPR 2023 Best Paper Award!

What is VisProg ?

VisProg is a system for compositional visual reasoning:

- Modular architecture desing
- Highly interpretable
- Requires no training
- Neuro-symbolic through LLMs reasoning
- Easy extendable to new tasks



What can it do?

VisProg is able to achieve many tasks!

Instruction
Input

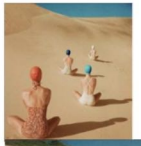
Output

Replace Leonardo DiCaprio with Leonardo DiCaprio wearing sunglasses

Create a color pop of the woman in blue and the woman in red and blur the background

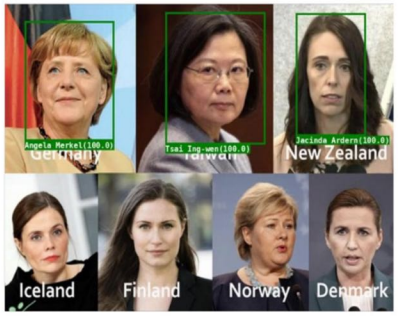
Replace Anne Hathaway with Emma Watson and Meryl Streep with Jennifer Lawrence

Replace the desert by sandy beach



Replace the couch with a plush blue couch

Tag the women leaders of Germany, Taiwan, and New Zealand



Tag the three female lead characters from Friends series

Tag these Scandinavian flags with their countries

Tag the painting of Girl with a Pearl Earring with its painter

How can it achieve this?

Idea: Using the **reasoning capability** of **LLMs** to generate smart ordered module instructions

In-context examples!

```
OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE,object=OBJ0,query='ground')
IMAGE0=Replace(image=IMAGE,object=OBJ1,prompt='white snow')
OBJ2=Seg(image=IMAGE0)
OBJ3=Select(image=IMAGE0,object=OBJ2,query='bear')
IMAGE1=Replace(image=IMAGE0,object=OBJ3,prompt='white polar bear')
RESULT=IMAGE1
```

In-context Examples

```
Instruction: Hide the face of Nicole Kidman with :p
Program:
OBJ0=Facedet(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='Nicole Kidman')
IMAGE0=Emoji(image=IMAGE, object=OBJ1, emoji='face_with_tongue')
RESULT=IMAGE0

Instruction: Create a color pop of the white Audi
Program:
OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='white Audi')
IMAGE0=ColorPop(image=IMAGE, object=OBJ1)
RESULT=IMAGE0

Instruction: Replace the red car with a blue car
Program:
OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='red car')
IMAGE0=Replace(image=IMAGE, object=OBJ1, prompt='blue car')
RESULT=IMAGE0

Instruction: Replace the BMW with an Audi and cloudy sky with clear sky
Program:
```

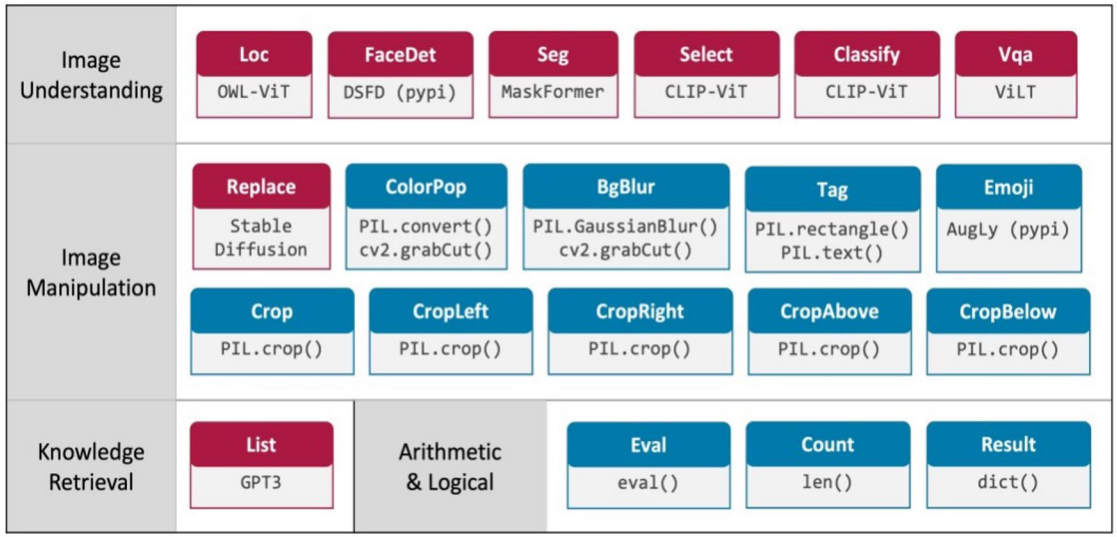
Prompt → GPT-3 → Program

```
OBJ0=Seg(image=IMAGE)
OBJ1=Select(image=IMAGE, object=OBJ0, query='BMW')
IMAGE0=Replace(image=IMAGE, object=OBJ1, prompt='Audi')
OBJ1=Seg(image=IMAGE0)
OBJ2=Select(image=IMAGE0, object=OBJ1, query='cloudy sky')
IMAGE1=Replace(image=IMAGE0, object=OBJ2, prompt='clear sky')
RESULT=IMAGE1
```

What are these modules?

VisProg has 20 integrated modules for different purposes!

- ➔ **Red modules** use vision pretrained model.
- ➔ **Blue modules** use image processing libraries and other python subroutines.
- ➔ Adding **new modules** to extend VISPROG's capabilities is straightforward



What are these modules?

Module are dependent on the task!

Task	Input	Output	Modules
Compositional Visual QA (GQA)	Image + Question	Text	Loc, Vqa, Eval, Count, Crop, CropLeft, CropRight, CropAbove, CropBelow
Reasoning on Image Pairs (NLVR)	Image Pair + Statement	True/False	Vqa, Eval
Factual Knowledge Object Tagging	Image + Instruction	Image	FaceDet, List, Classify, Loc, Tag
Image Editing with Natural Language	Image + Instruction	Image	FaceDet, Seg, Select, Replace, ColorPop, BgBlur, Emoji

Modules ordering matters in many cases….

Why it is Interpretable?

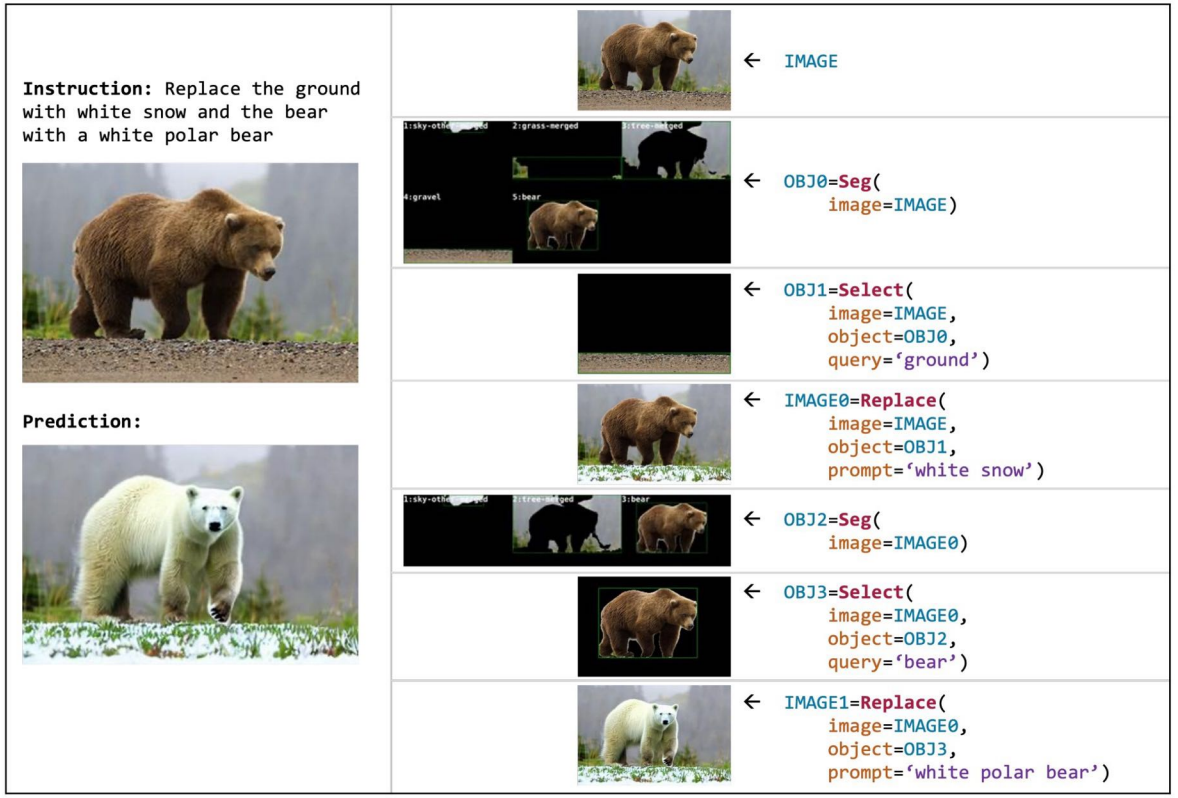
Basically every module outputs something….

Module A —--> Module B —--> Module C

Why is it Interpretable ?

The **program state** output is shown at every step (i.e. after every module)

This is called **Visual Rationale**



Does it fail?

As all models do, it has its problems unfortunately…..

Qualitative reasons of VisProg failure…

Original: Tag the CEO of IBM

Reason for failure:

```
# The knowledge query returns
one of the previous CEOs of IBM
LIST0 = List(query='CEO of IBM',
             max=1)
```

Modified: Tag the most recent CEO of IBM

Reason for success:

```
# The knowledge query returns the
current CEO of IBM
LIST0 = List(
  query='most recent CEO of IBM',
  max=1)
```

Original: Replace the coffee table with a glass-top modern coffee table

Reason for failure:

```
# The selection module selects
an incorrect region (rug)
OBJ1 = Select(
        query='coffee table',
        category=None)
```

Modified: Replace the coffee table (table-merged) with a glass-top modern coffee table

Reason for success:

```
# The category restricts the
search space
OBJ1 = Select(
        query='coffee table',
        category='table-merged')
```

Original: Tag the item that is used to make coffee

Reason for failure:

```
# Localization modules fails to
detect any objects
OBJ0 = Loc(
    image=IMAGE, object='item')
```

Modified: Tag the kitchen appliance that is used to make coffee

Reason for success:

```
# Localization modules detects
multiple appliances
which are then filtered by Select
OBJ0 = Loc(
  image=IMAGE,
  object='kitchen appliance that
          makes coffee')
```

Original: Tag the Triwizard Tournament Champions

Reason for failure:

```
# List restricts the output
length to 3
LIST0 = List(
  query='Triwizard Tournament
         Champions',
  max=3)
```

Modified: Tag the 4 Triwizard Tournament Champions

Reason for success:

```
# List outputs all 4 champions
LIST0 = List(
  query='Triwizard Tournament
         Champions',
  max=4)
```

What's Embodied AI again?

Agent that moves in environment:

- 3D Environment
- Robot
- Human (optional)

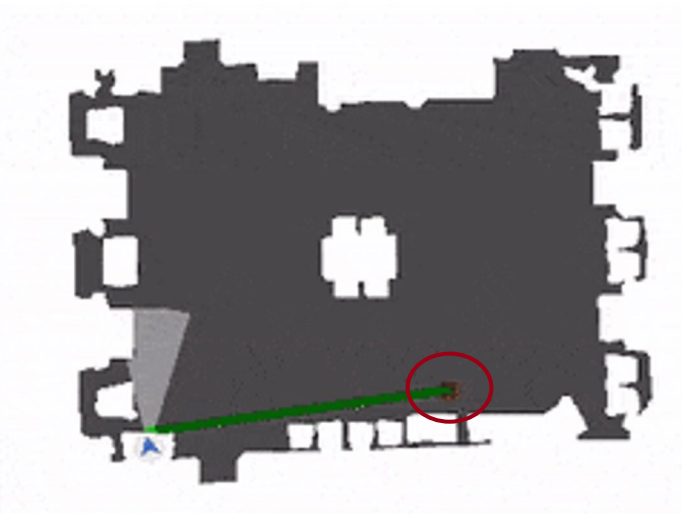
You also need a goal:

- (x,y,z) point
- object label
- Image

RGB

Depth

Map (optional)



How Embodied AI adapts to Visprog?

We would like to extend VisProg capability to Embodied AI!

VisProg → Embodied AI

Components:

1. LLM capable of reasoning
2. Agent in an Environment
3. A single Pointgoal Model
4. Pretrained CV models
5. Task to complete

Tasks & Modules ?

The idea is to have a **modular implementation**, to avoid some errors the LLM could output!

Tasks we want to achieve:

1. Pointgoal Navigation
2. Objectgoal Navigation
3. Multi-object goal Navigation
4. Visual Language Navigation (VLN)

Module we want to add:

1. Navigation module
2. Object Detector
3. Classifier
4. Next ??

Goal ?

The idea is to have a **modular implementation**, to avoid some errors the LLM could output!

Tasks we want to achieve:

1. Pointgoal Navigation → *"Navigate to (3,4)"*
2. Objectgoal Navigation → *"Navigate to the toilet in the bathroom"*
3. Multi-object goal Navigation → *"Navigate to the TV and then to the plant next to it"*
4. Visual Language Navigation (VLN)

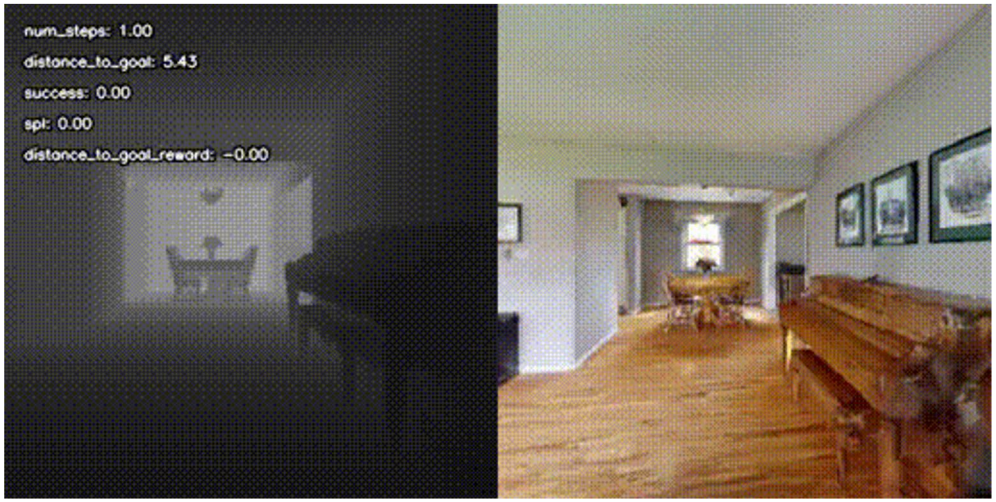
We want to achieve this **without any training!**

Failure Cases ?

Other modules has to be added to check the object

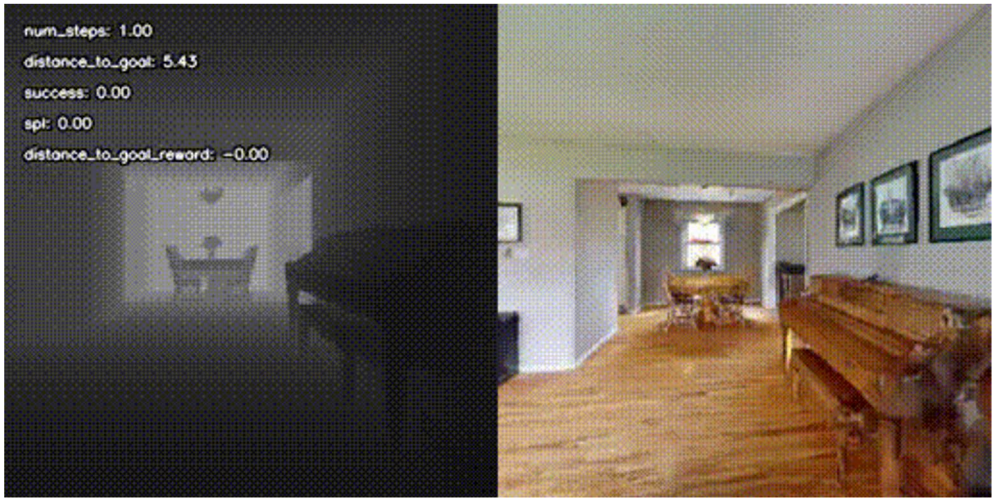
Instruction: Navigate to "**TV Monitor**"

➔ Fails to calculate the correct distance from depth



Instruction: Navigate to "**couch**"

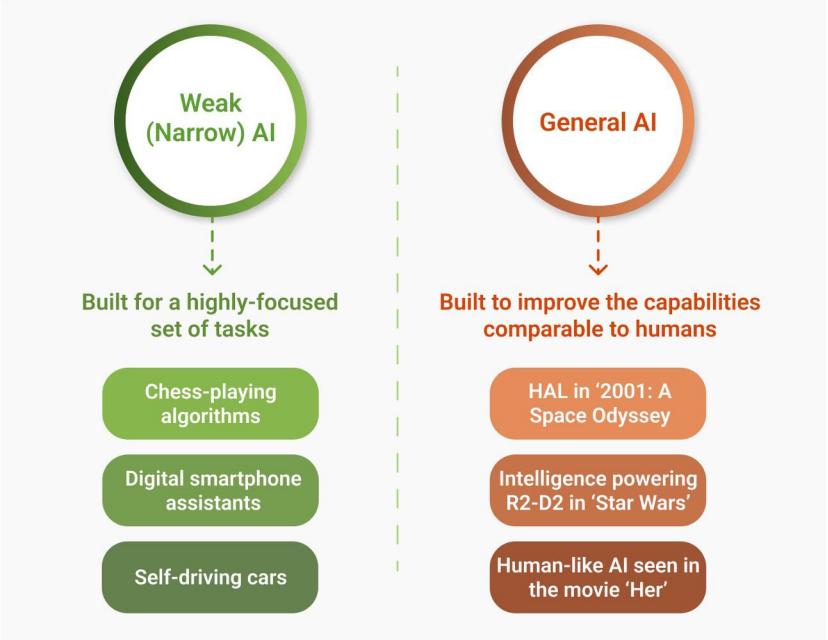
➔ Fails to understand the correct semantics of target goal



What's the Research Direction ?

VisProg is a step-forward to general AI systems:

1. **Highly interpretable** by the use of its *Visual Rationale*
2. Great at **zero-shot learning** (free-training system)
3. Adapts to **different tasks** (*VQA, NLVR, Image-tagging, image-editing*)
4. **Expands the scope of AI systems** to serve the long tail of complex tasks that people may wish to perform.



The End!