NRI: INT: COLLAB: Collaborative Task Planning and Learning through Language Communication in a Human-Robot Team

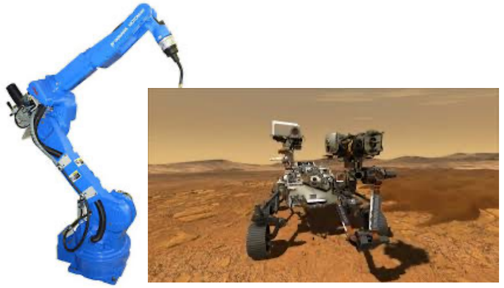
NSF IIS-1949634 (formerly 1830244)
NSF IIS-1830282

Poster # 25

Joyce Chai - University of Michigan

Julie Shah - Massachusetts Institute of Technology

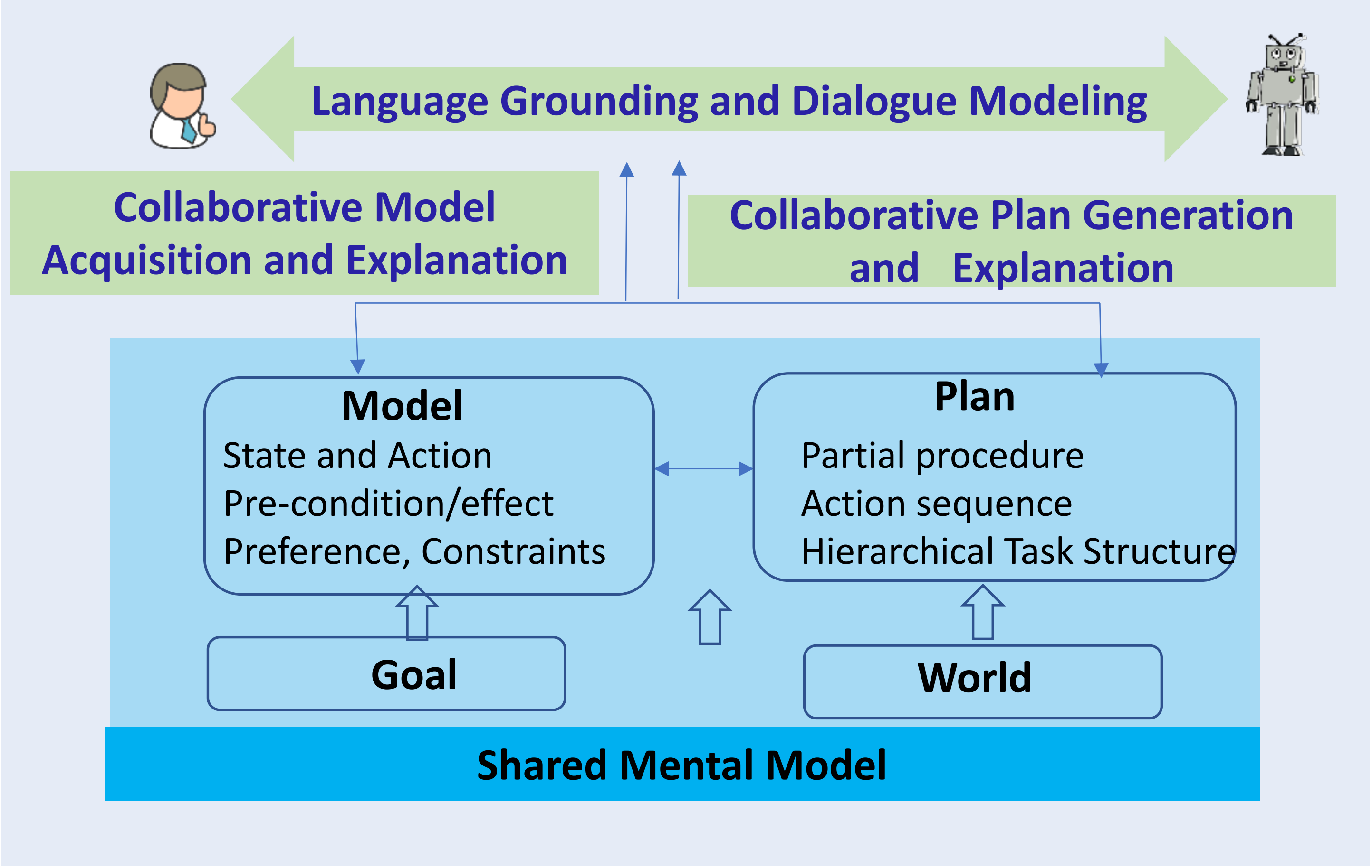
Motivation and Objectives



- No complete domain models for new situations
- Computationally expensive real-time planning

Empower robots with the ability to harness human knowledge and expertise to learn new states, actions, and plans.

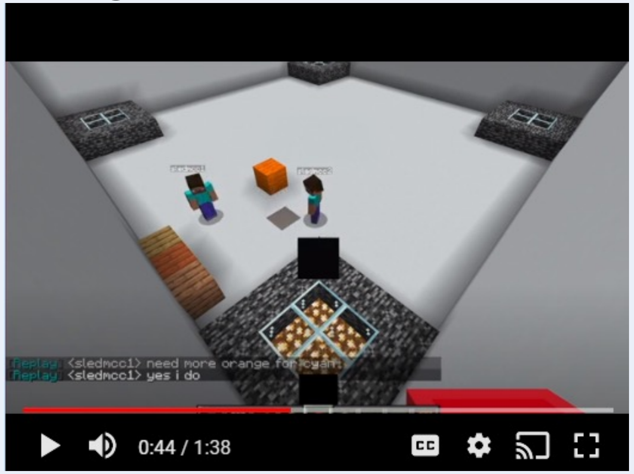
Link language and dialogue processing with the robot's underlying planning system to support collaborative task planning and learning in a human-robot team.



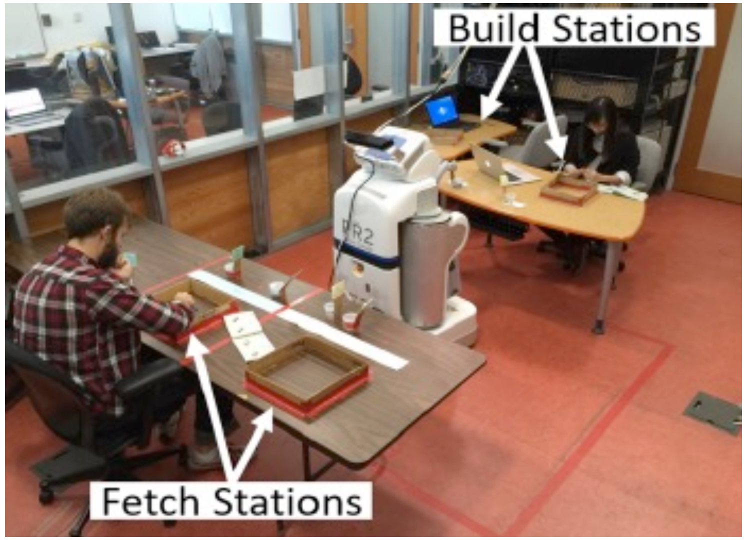
Research Progress

Learning new action and states through language interaction

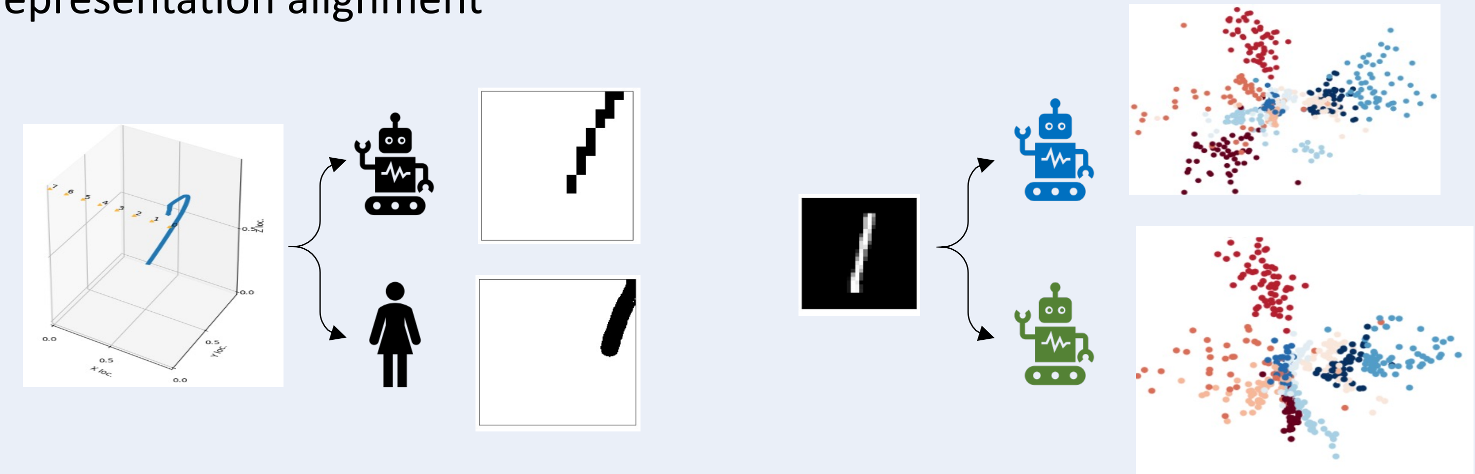
Mental model representation and learning in collaborative tasks



Plan acquisition from language instructions



General-purpose learning technique for efficient human-agent and agent-agent representation alignment



Action-Effect Prediction

Action
(squeeze-bottle)

V

X

X

V

X

Action
(peel-carrot)

X

Action
(mash-carrot)

X

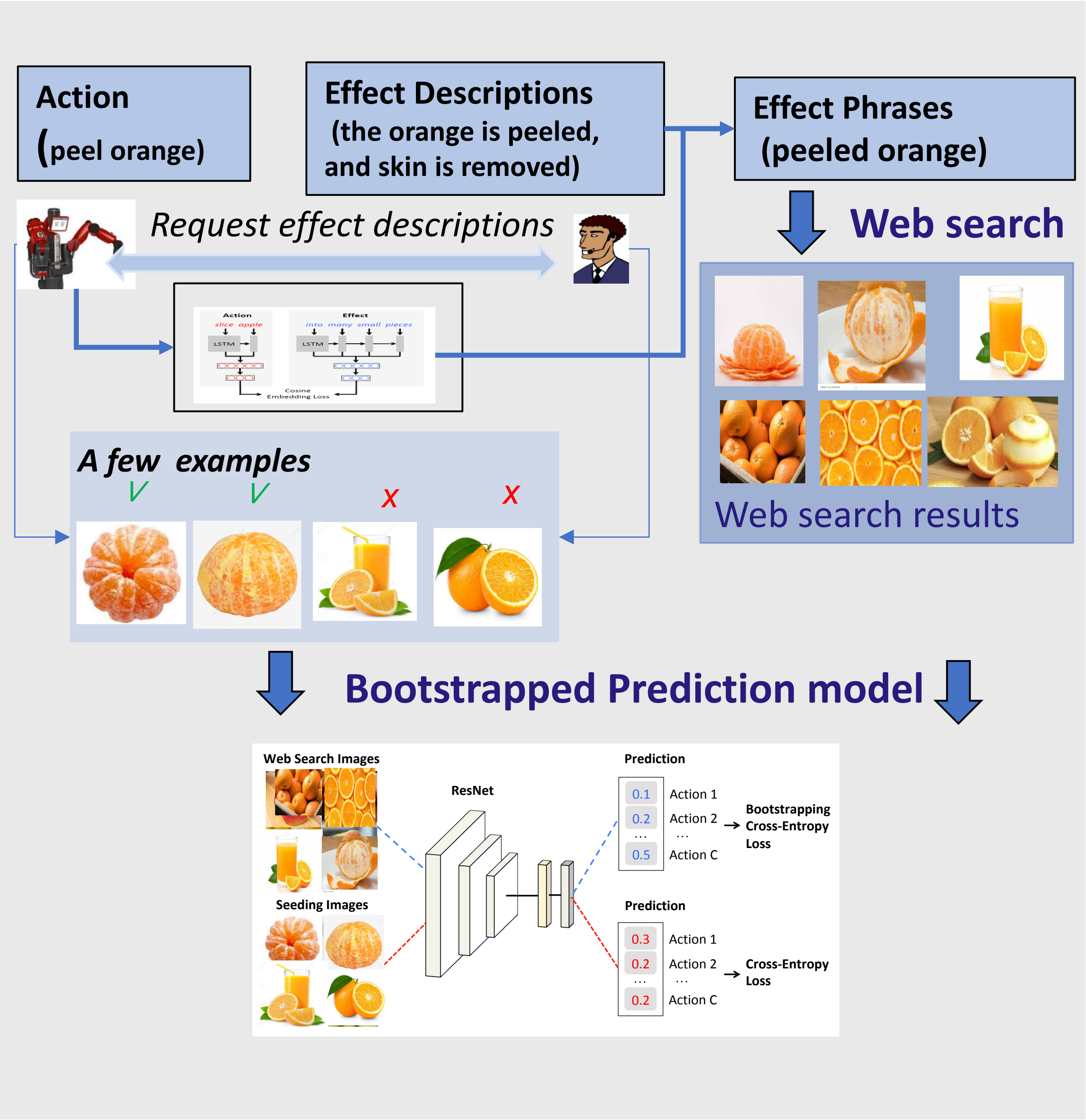
Action
(juice-carrot)

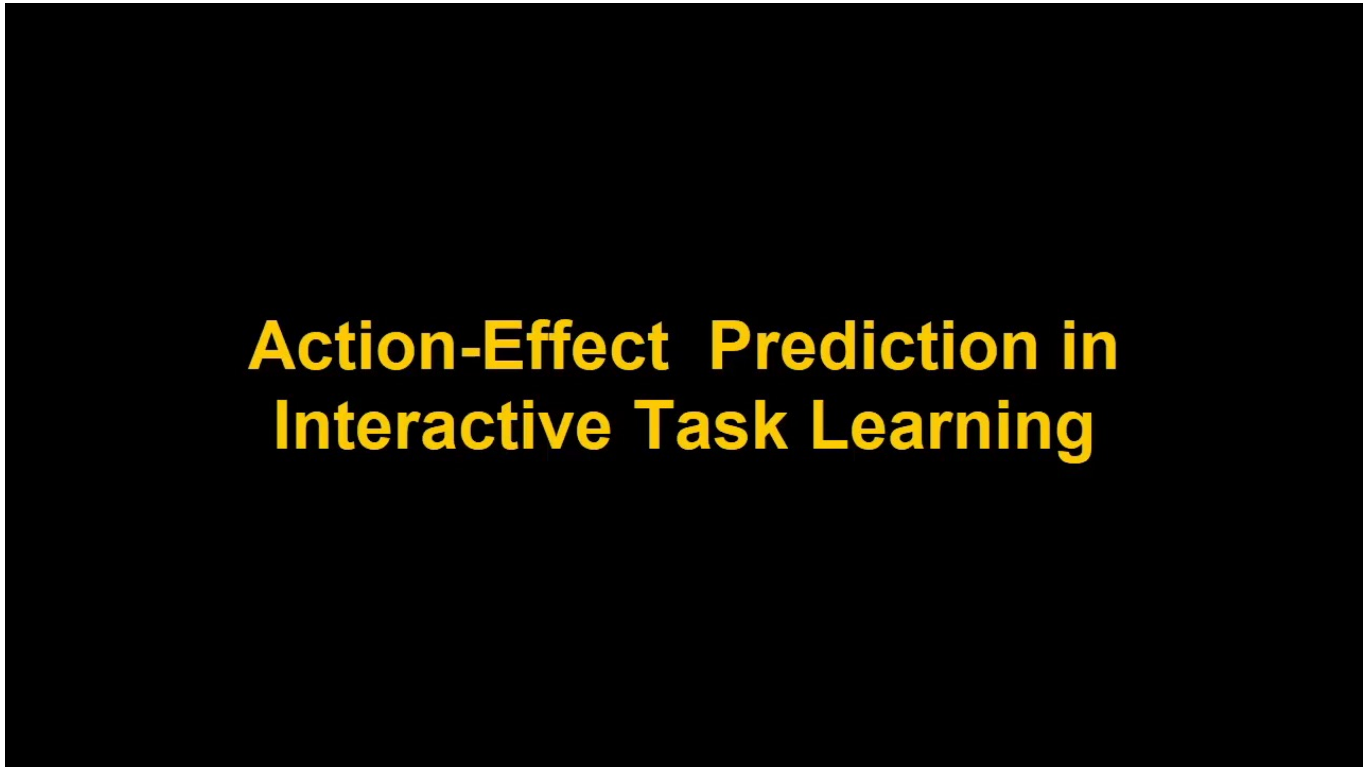
V

Action
(chop-carrot)

Q. Gao, S. Yang, J. Y. Chai, and L. Vanderwende. *What action causes this? Towards Naïve Physical Action-Effect Prediction, ACL 2018.*

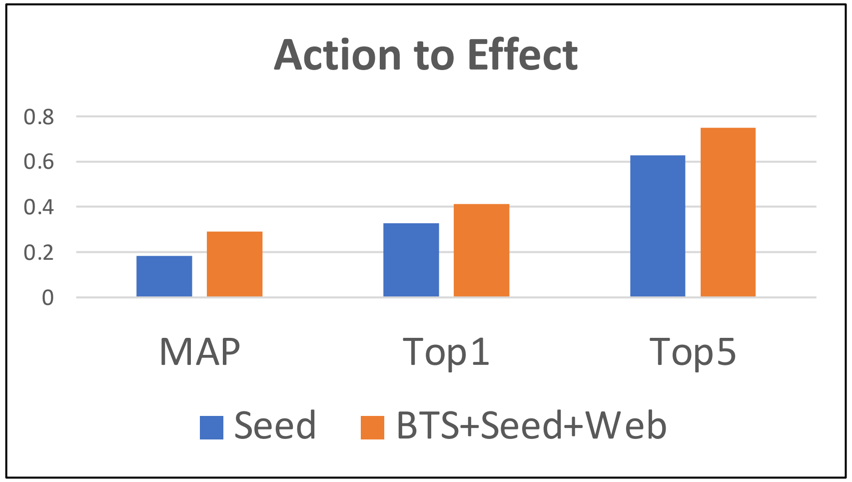
Learning Action-Effect

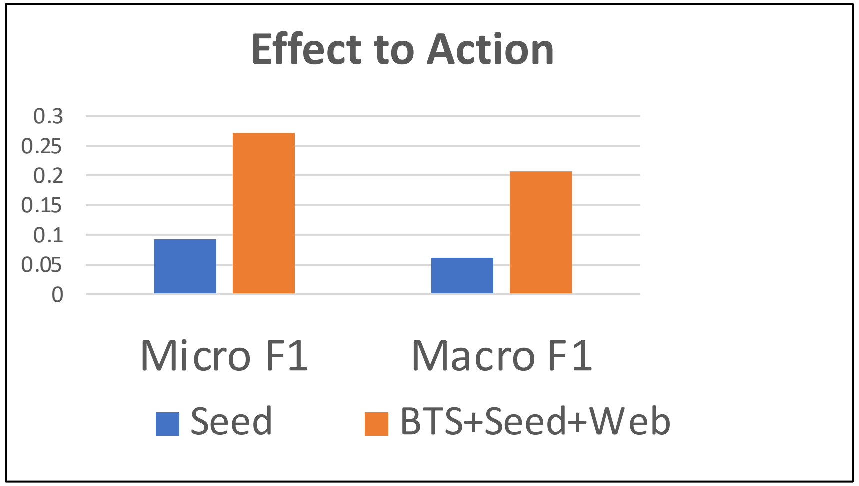




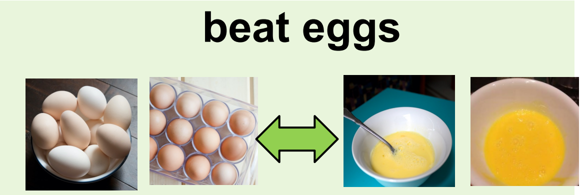
Dataset: 140 verb-noun pairs, 1400 effect descriptions, ~4200 annotated effect images, >60K web-searched images for training

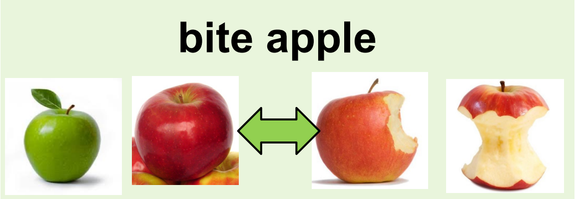
Learning Action-Effect





Action	AP
beat eggs	0.783
pile boxes	0.766
bite apple	0.484
slice onion	0.470





Action	AP
crack glass	0.047
lock drawer	0.037
stain shirt	0.023
close window	0.087

Natural Language Instructions to Actions

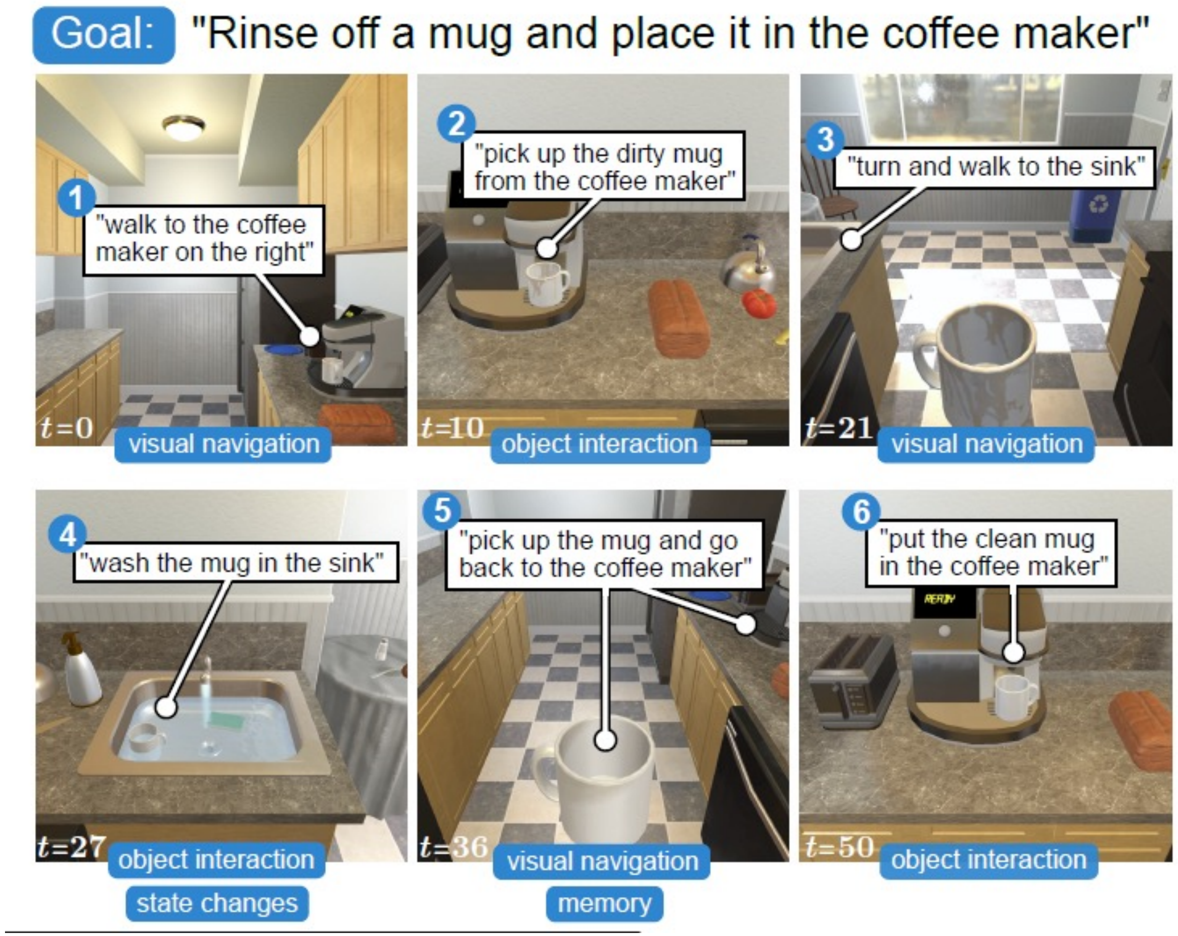
ALFRED

A Benchmark for Interpreting Grounded Instructions for Everyday Tasks

Mohit Shridhar[1] Jesse Thomason[1] Daniel Gordon[1] Yonatan Bisk[1,2,3]
Winson Han[3] Roozbeh Mottaghi[1,3] Luke Zettlemoyer[1] Dieter Fox[1,4]
AskForALFRED.com

Action **L**earning **F**rom **R**ealistic **E**nvironments and **D**irectives (ALFRED) (Shridhar et al., 2020)

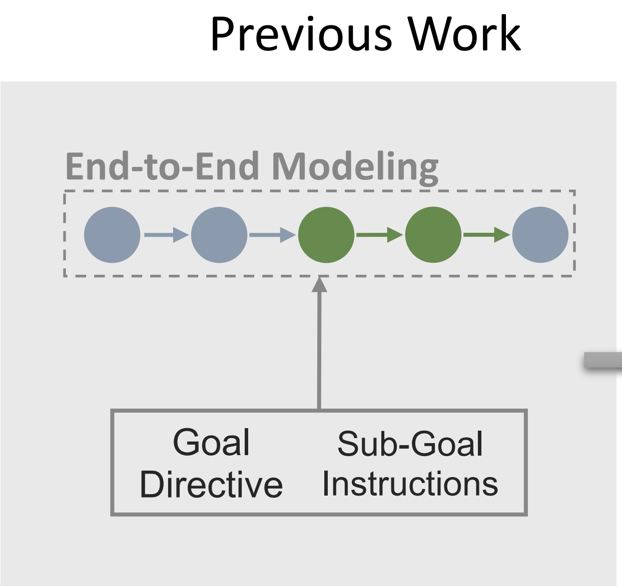
- Understand task goals
- Follow natural language instructions
- Ground language to perception
- Plan in the embodied environment



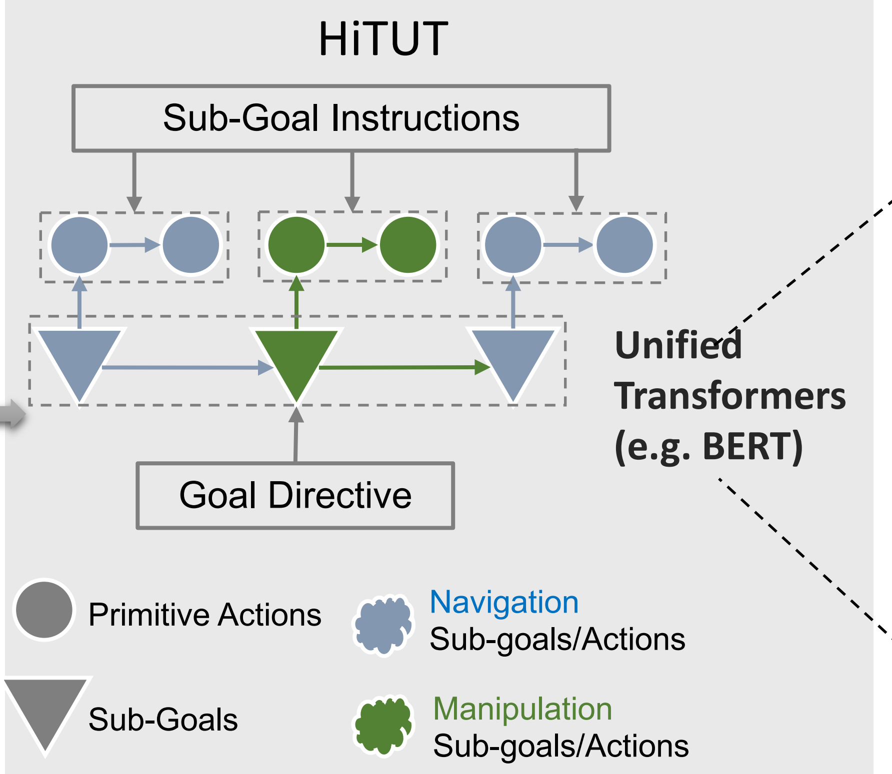
Hierarchical Task Learning

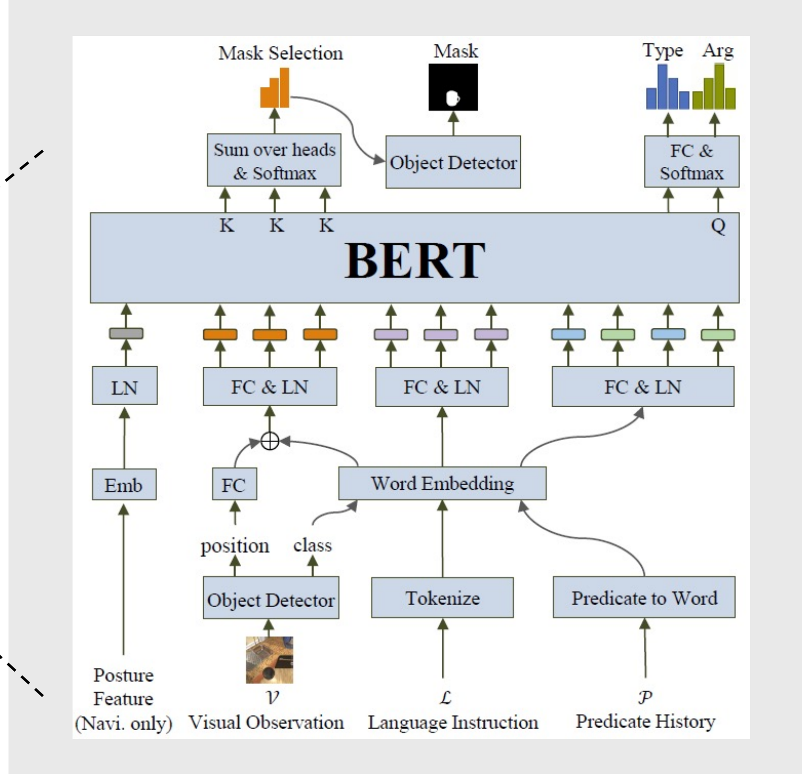
HiTUT (**Hi**erarchical **T**asks via **U**nified **T**ransformers)

- An ***explainable model*** achieving the new state-of-the-art performance
- A ***de-composable platform*** to support more in-depth evaluation and analysis



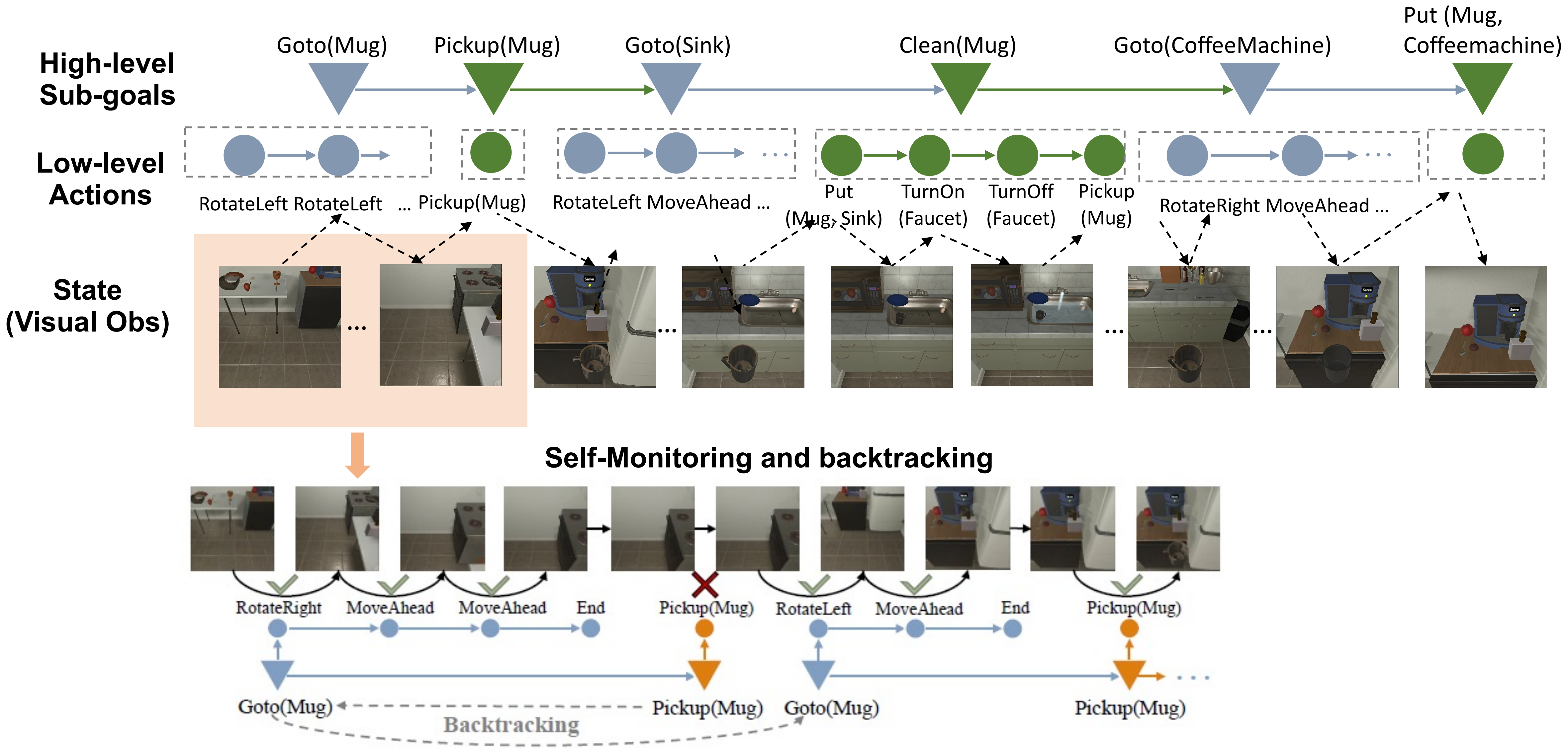
(Shridhar et al., 2020; Pratap Singh et al., 2020; Storks et al., 2021)





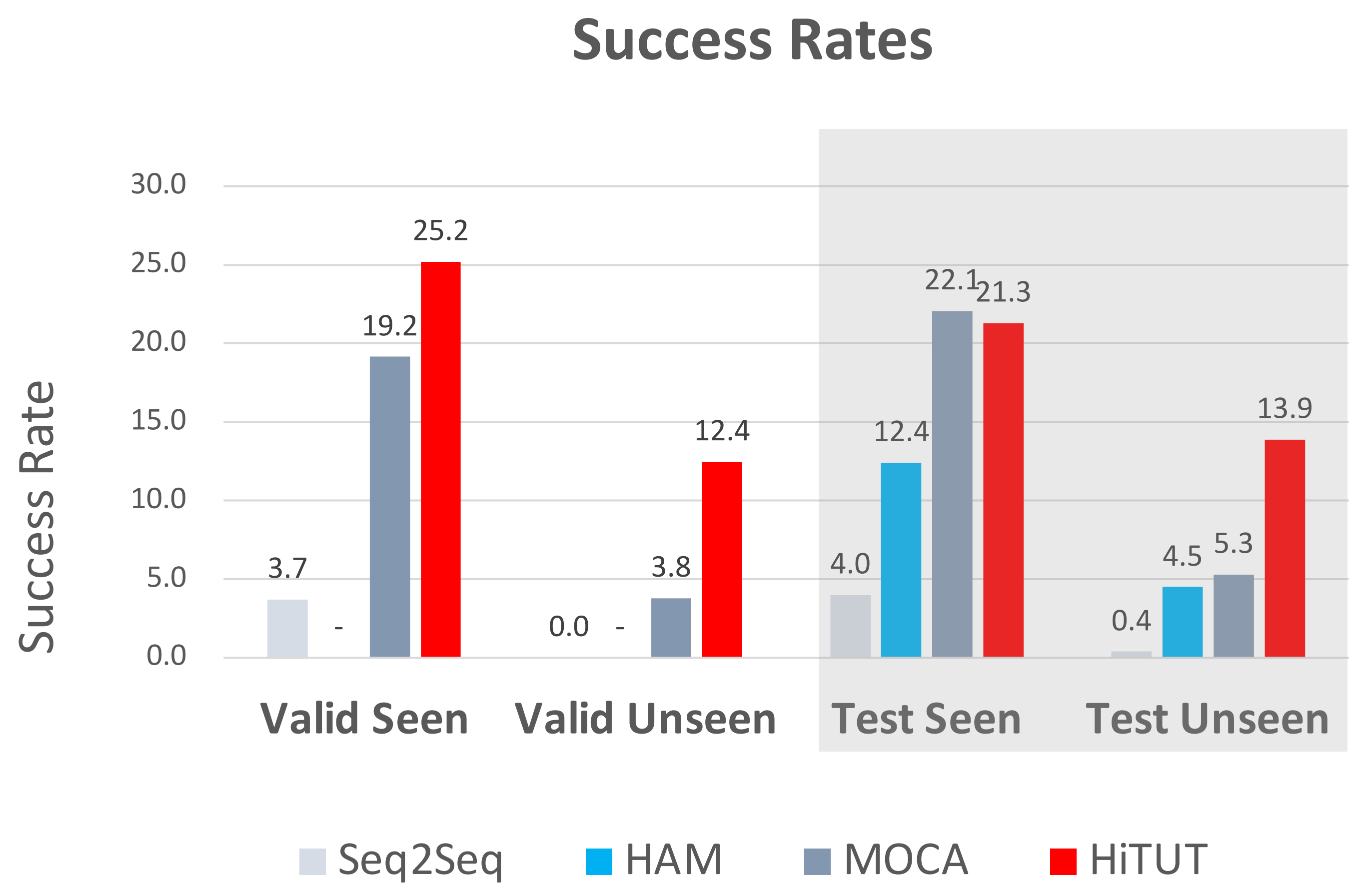
Hierarchical Structure with Self-Monitoring and Backtracking

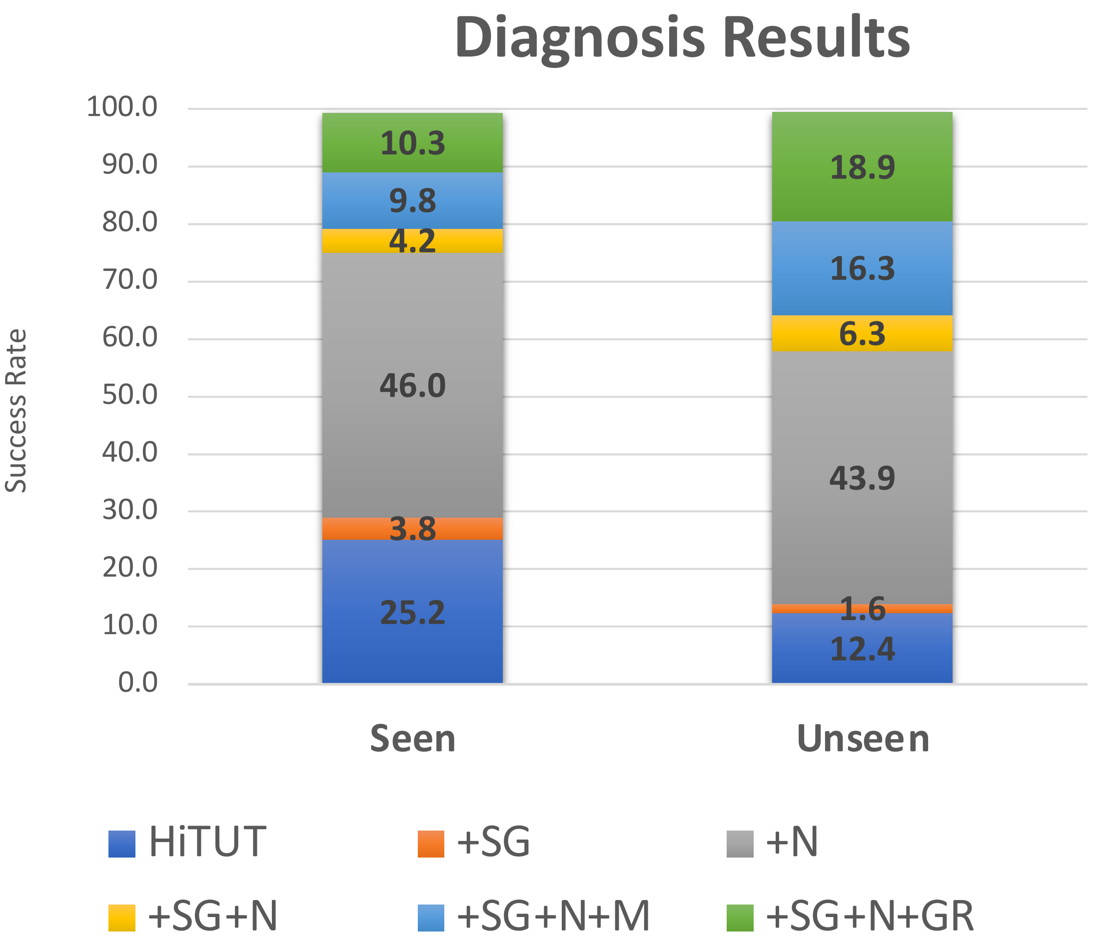
Goal Directive Place a cleaned mug in the coffee machine.



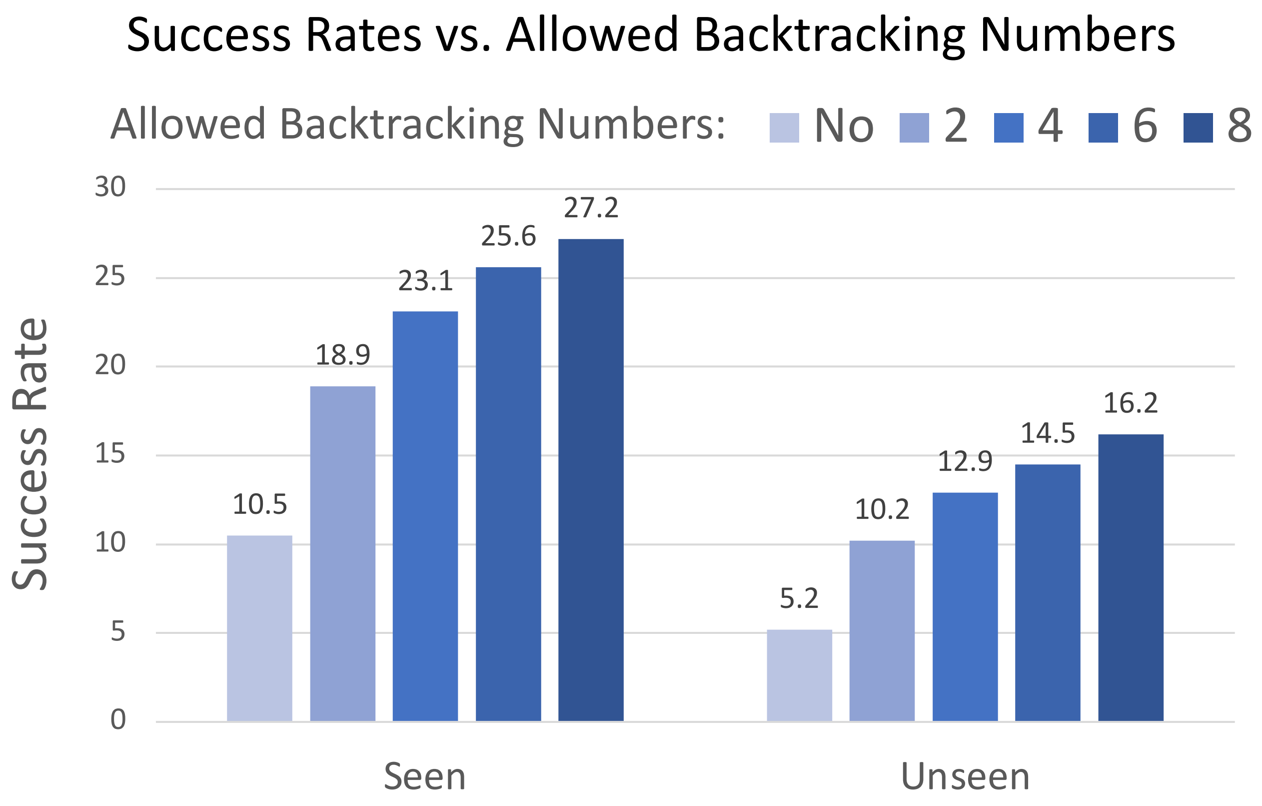
Results: Better Generalization in Unseen Environment

- Outperform previous STOA with a large margin (160% gain)





Results: backtracking improves performance

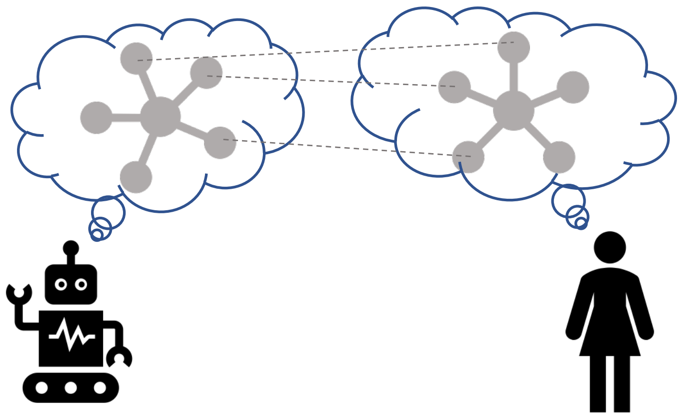


Task Goal:
Put two books on the desk.

Latent Space Alignment for Improved Partners

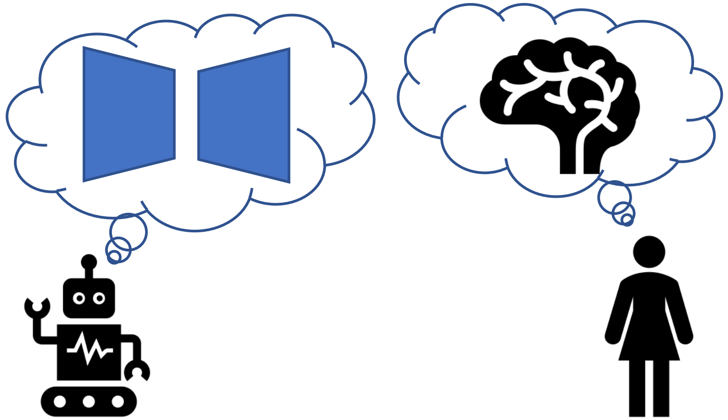
- Agents' *representation* interpretability complements interpretable input (e.g., language) and output domains (e.g., object grounding)
- Shared mental models are central to good human-only and human-robot teams [Mathieu et al. 2004, Nikolaidis and Shah 2012, Hiatt et al. 2017]
- Even in simple tasks like digit classification, many neural net models are confusing to people

"I dont know, I thought the task a little confused, its like our language and perception does not match with the machine language and perception. [sic]"



Aligning explicit mental models supports good team performance

We wish to create agents that can efficiently learn task-dependent, human-interpretable representations



Prior art focuses on model-based approaches; we seek to align neural net latent spaces

Scaling the quality of teammates' mental models: equifinality and normative comparisons, Mathieu et al. 2004
Human-robot teaming using shared mental models, Nikolaidis and Shah 2012
Human modeling for human–robot collaboration, Hiatt et al. 2017

Adversarially Guided Self-Play (ASP)

- Representations learned by neural nets may not align with human intuition

a) 2D encodings generated by a VAE of MNIST images

b) Humans might arrange encodings in more interpretable formats (e.g. dialpad)

c) Using ASP, we efficiently train models to learn the latent space from human preferences

- In Adversarially Guided Self-Play (ASP), we combine three training terms

Self-Play

Support high task performance

Trained via self-play

Learn a language

Adversarial Training

“Look like” the right sorts of representations

Trained via adversarial trained with large, unpaired corpus

Use English words

Supervised Training

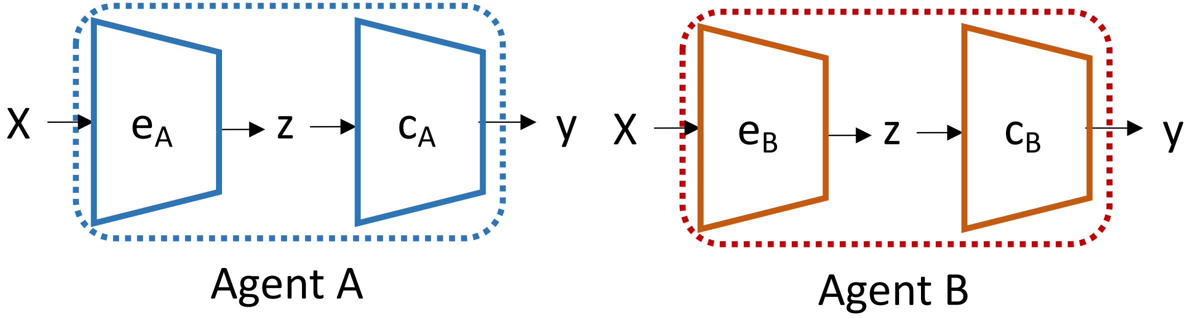
For some specific inputs, use specific representation

Trained via supervised loss

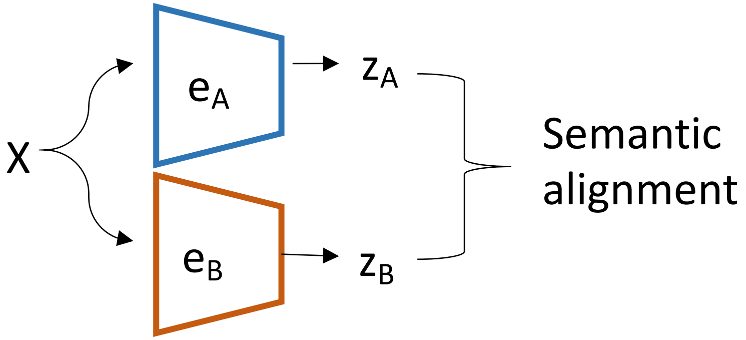
This is the meaning of some English words

Measures of Latent Space Alignment

- Assume agents where inputs X are encoded via *e* to a representation *z* that is classified via *c* to a label *y*

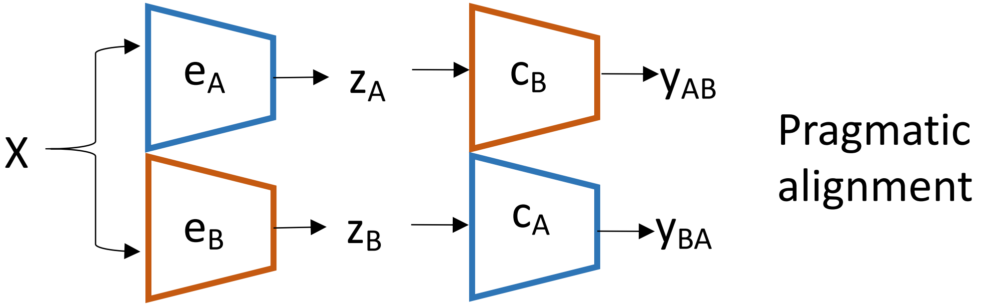


- Given two agents, A and B, *semantic alignment* is the normalized mean squared error between encodings for the same inputs, X



$$a_S(A,B,X) = -\frac{1}{Z|X|}\sum_{x\in X}(e_A(x)-e_B(x))^2 \qquad Z = \frac{1}{|X|^2}\sum_{i\in[1,\ldots,|X|]}\sum_{j\in[1,\ldots,|X|]}(e_A(x_i)-e_B(x_j))^2$$

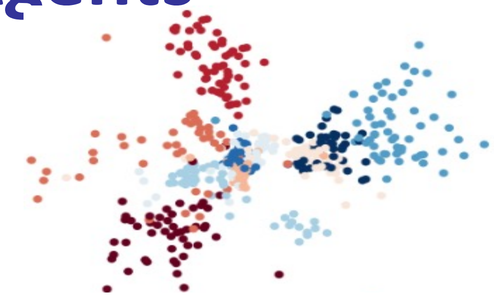
- For some agents like humans, accessing *z* is impossible: *pragmatic alignment* is the task performance when passing information via encodings.



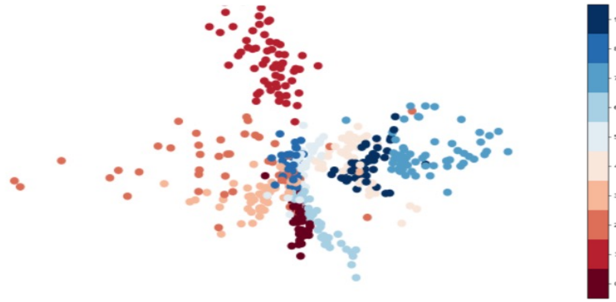
- With humans, further study *trust* of agents

Latent Space Alignment among Agents

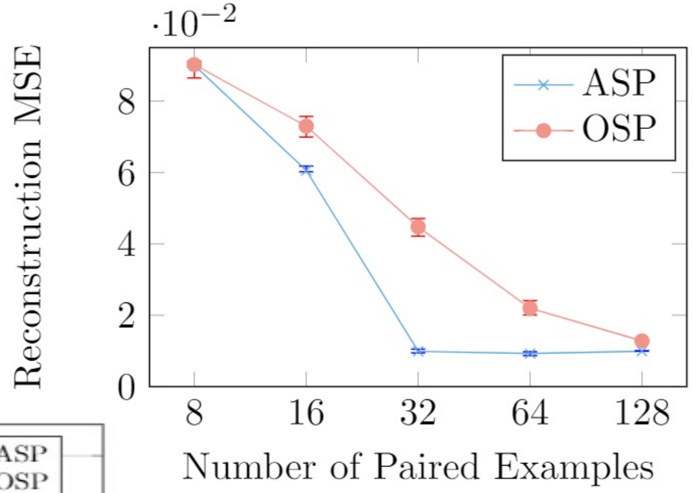
- Initial experiments in training models to align with pre-trained models
- Compared to prior art, ASP produces models with greater *semantic* and *pragmatic* alignment for the same amount of paired data, and measures are correlated across of a variety of tasks
- Greatest benefit shown for small (e.g. 32) amount of paired data.



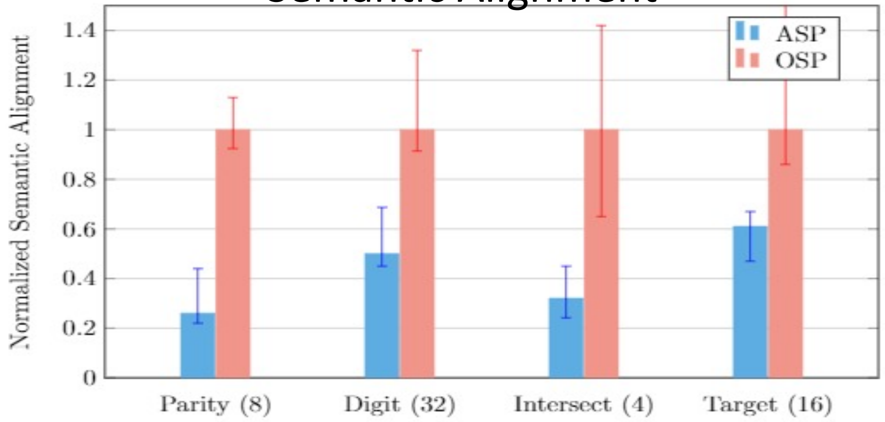
a) 2D encodings generated by a VAE of MNIST images



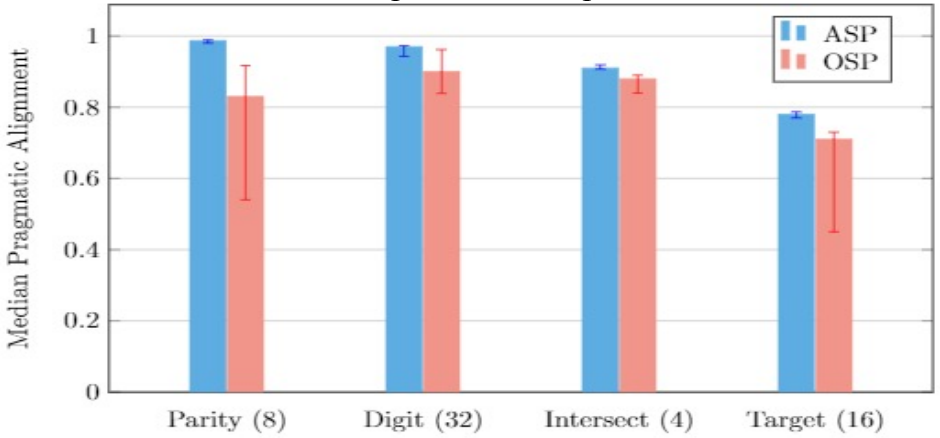
b) 2D encodings of an autoencoder trained to align with the first VAE, using 8 examples



Semantic Alignment



Pragmatic Alignment

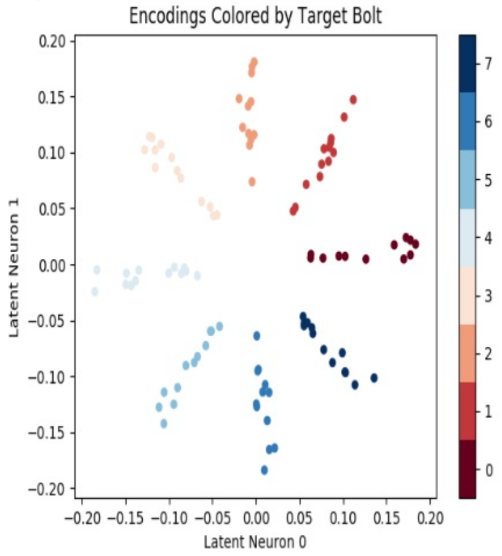


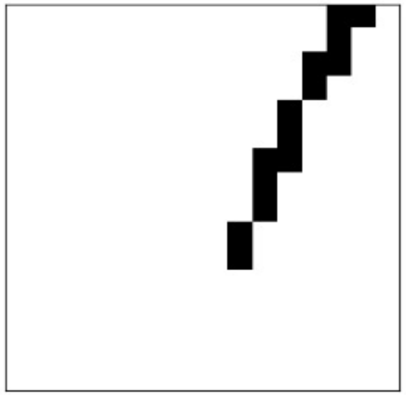
Latent Space Alignment with Humans

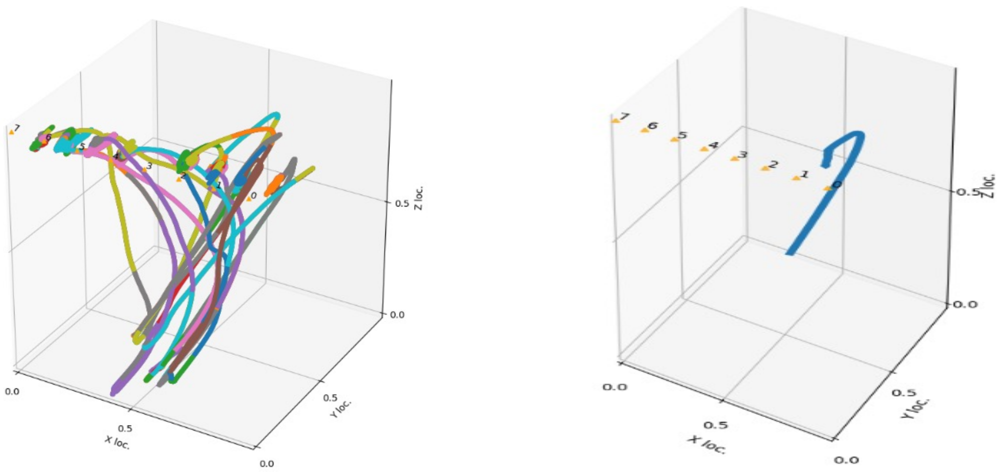
- Trained agents to align with “designs”
- Measure human → agent and agent → human performance (pragmatic alignment), as well as human trust calibration [Lee and See 2004]
    - Given encoding, human classifies
    - Given input, human encodes
    - Given encoding and input, human predicts classification correctness

a) Latent space “designs” b) Learned latent spaces using ASP

Different latent space designs for the same data







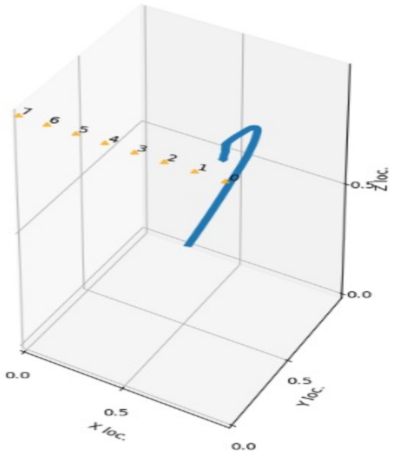
Trust in Automation: Designing for Appropriate Reliance, Lee and See 2004

Latent Space Alignment with Humans: Results

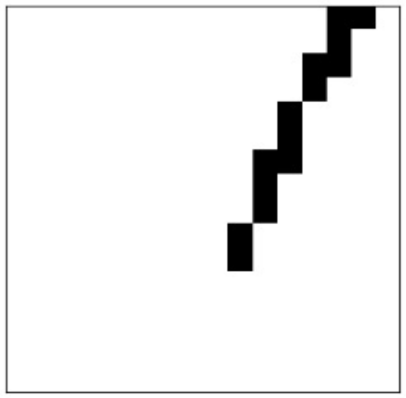
- ASP-trained models supported better classification accuracy (pragmatic alignment)
- Latent space design utility was task-dependent
- ASP-trained techniques results in better-calibrated trust: humans could better predict machine failures
- Pilot study established that humans could generate encodings, beyond merely choosing from options

Task	*ASP dig.*	*ASP par.*	*OSP dig.*	*OSP par.*	*PAE*
Parity	0.63 (200)	**0.82** (150)	-	0.70 (280)	0.75 (140)
Digit	**0.52**** (50)	0.30 (90)	0.33 (80)	-	0.24 (70)
	ASP 2D	*ASP sketch*	*OSP 2D*	*OSP sketch*	*PAE*
Inter.	**0.68*** (40)	0.61 (110)	0.62 (50)	0.53 (40)	0.56 (60)
Target	**0.53*** (130)	0.36 (110)	0.11 (90)	0.33 (130)	0.34 (150)

Classification accuracies when humans classified, given model encodings.
*, ** for $p < 0.1$, 0.05 for technique and design outperforming all others.



Trajectory (shown)



Machine's encoding of traj. (hidden)

Participant-generated sketch

Next step in the coming year

- Acquiring shared mental models based on collaborative discourse
  - Conducting empirical studies in physical interaction with robots
- Interactive learning to ground language instructions to plan structures
  - Developing dialogue strategies to support sample efficient learning and exception handling
  - Incorporating algorithms (from simulation) to physical world interaction
- Participant-guided latent space learning
  - Allow participants, rather than designers, to provide the latent space design
- Aligning emergent communication with semantic spaces (e.g., word embeddings)
- Integration and evaluation in physical world
  - Factorial design and hypothesis validation to measure the role of model/plan explanation, use of dialogue for model reconciliation, and incremental learning and refining models and plans