Mapping Navigation Instructions to Continuous Control Actions with Position-Visitation Prediction

Valts Blukis, Dipendra Misra, Ross A. Knepper, Yoav Artzi

Cornell CIS
Computer Science

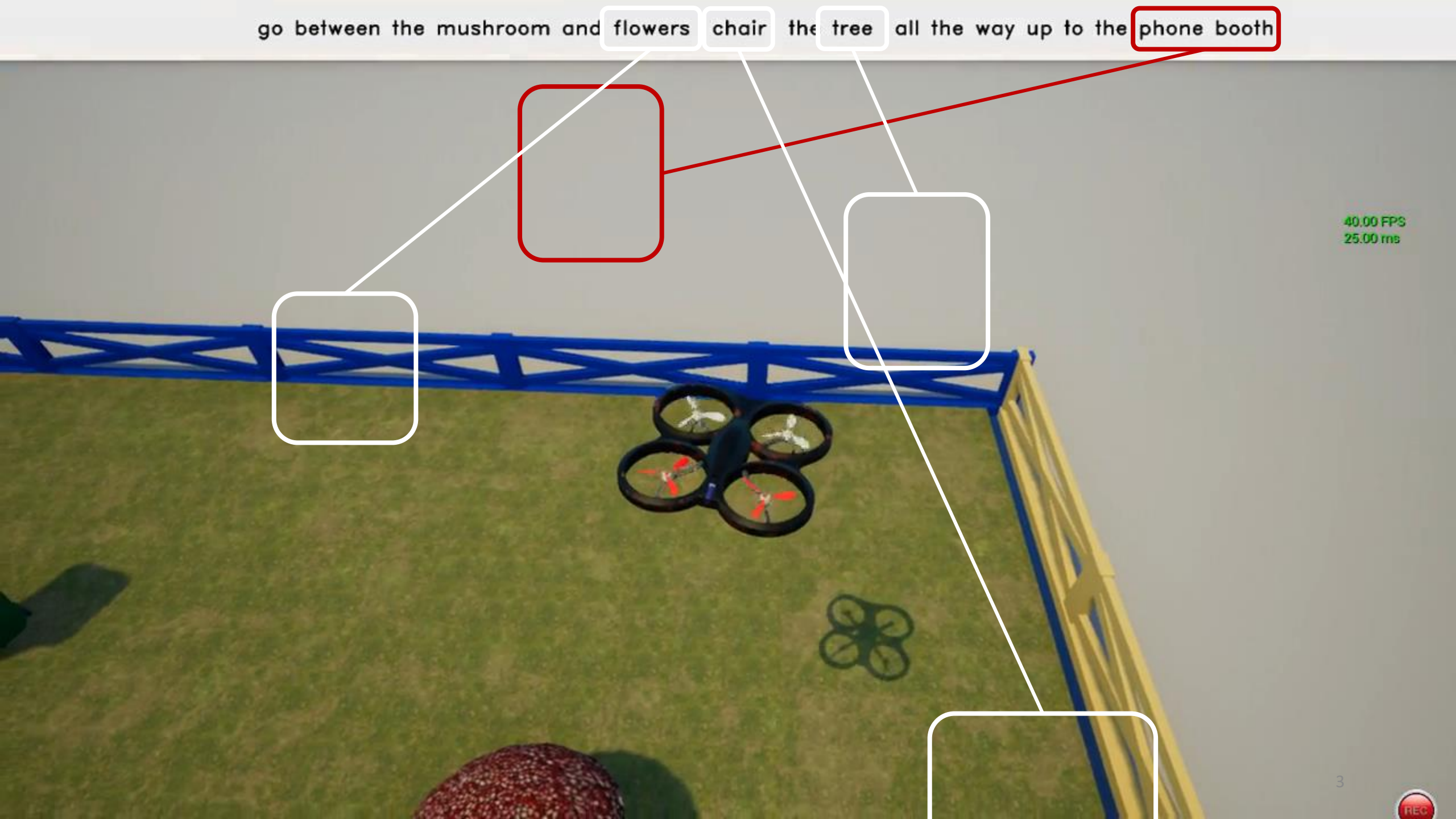
Motivation

- For wide adoption, robot control interfaces should be:
  - Accessible
  - Expressive

- **Natural language fulfills these criteria**

- **Combining natural language with**

Go towards the blue fence passing the anvil and tree on the right

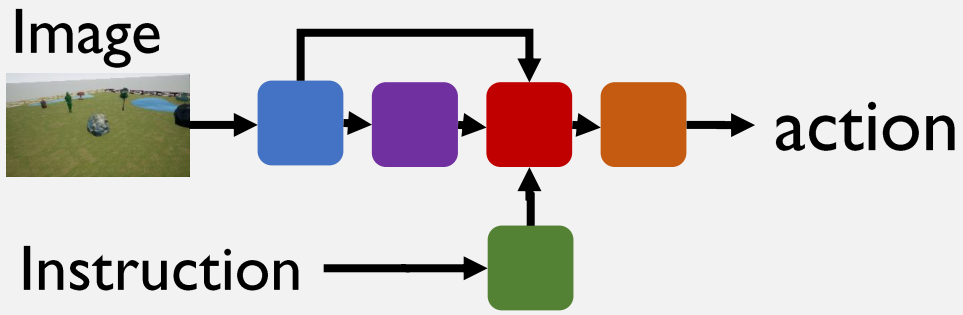

go between the mushroom and flowers chair the tree all the way up to the phone booth
40.00 FPS
25.00 ms
REC

Following Natural Language Instructions is Hard

It requires:

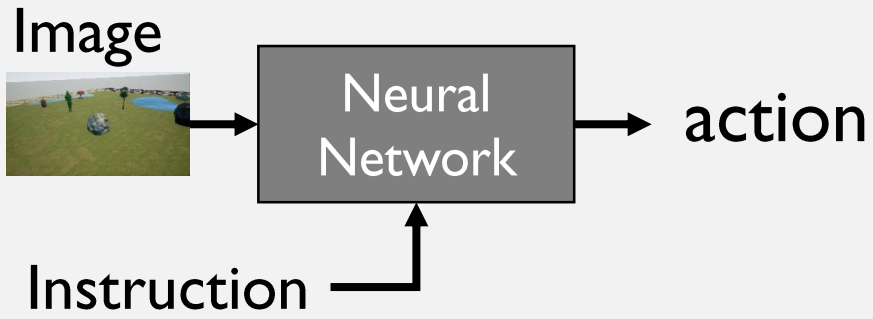
Visual Perception | Spatial Reasoning | Language Understanding | Language Grounding | Control

Modular approaches



Hard to scale knowledge representations

Single-model approaches



Difficult learning

Lacks interpretability

Our approach: Two-Stage Decomposition

Image

Instruction

Stage 1

Planning with Position-Visitation Prediction

Stopping Probability
Visiting Probability

Position-Visitation Distributions

Stage 2

Action generation

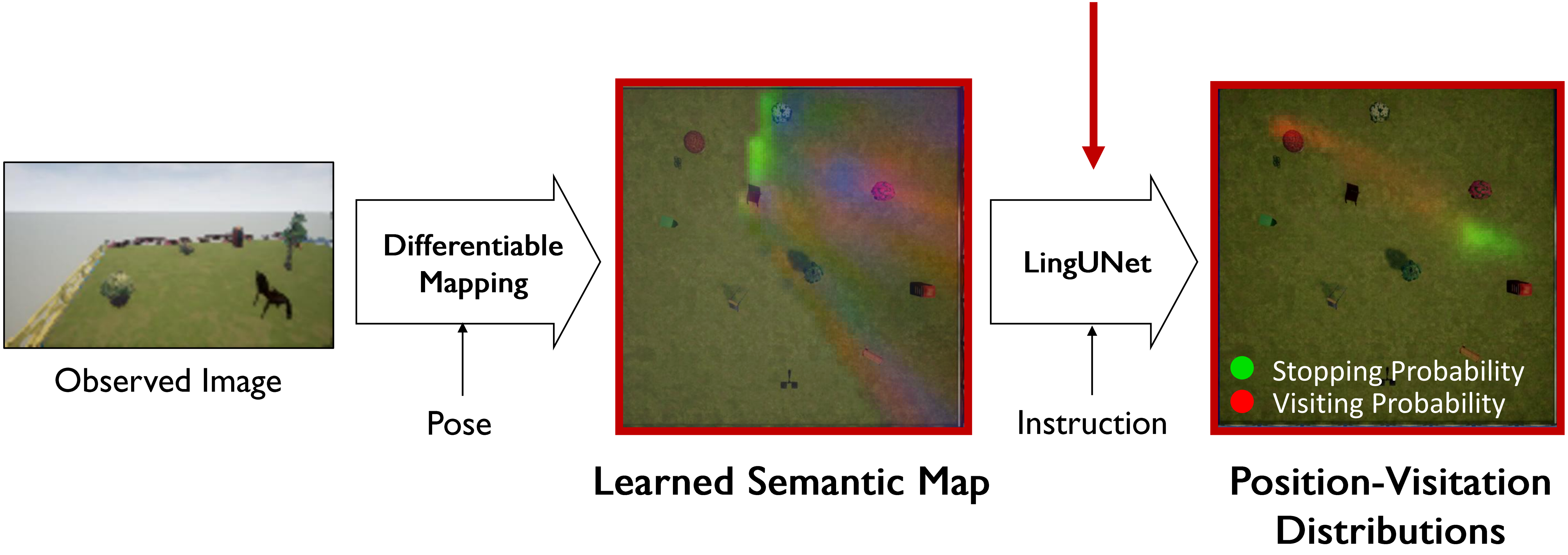
action

Stage 1: Position Visitation Prediction

Predicted ... in the map

Plan where to go with the LingUNet module

A convolutional image encoder-decoder architecture conditioned on natural language



Stage 2: Action Generation

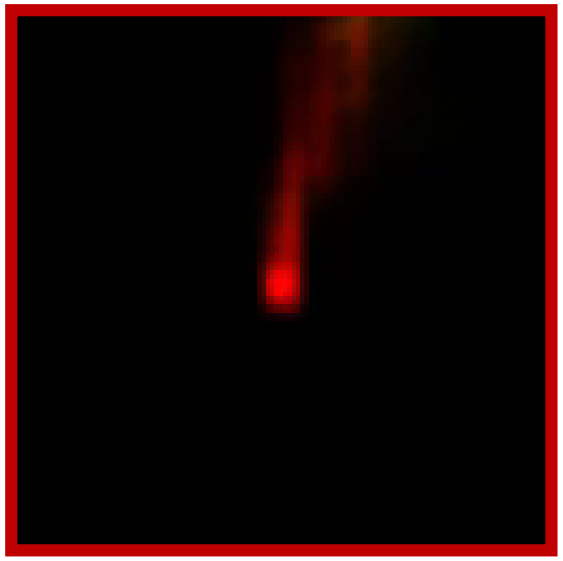
No dependence on language → simple control problem.

Training experience not limited by availability of natural language data.

Position-Visitation Distributions

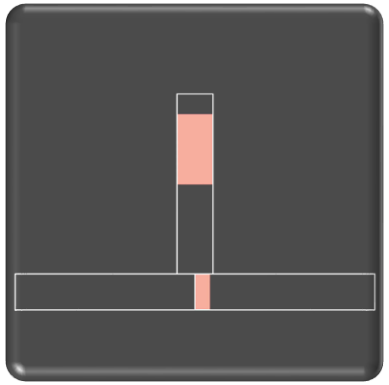
Transform to Egocentric

Pose



Egocentric Position-Visitation Distributions

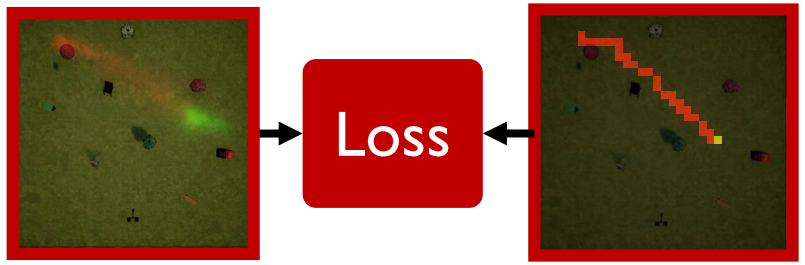
Fully Connected NN



Continuous Velocity Command

Learning

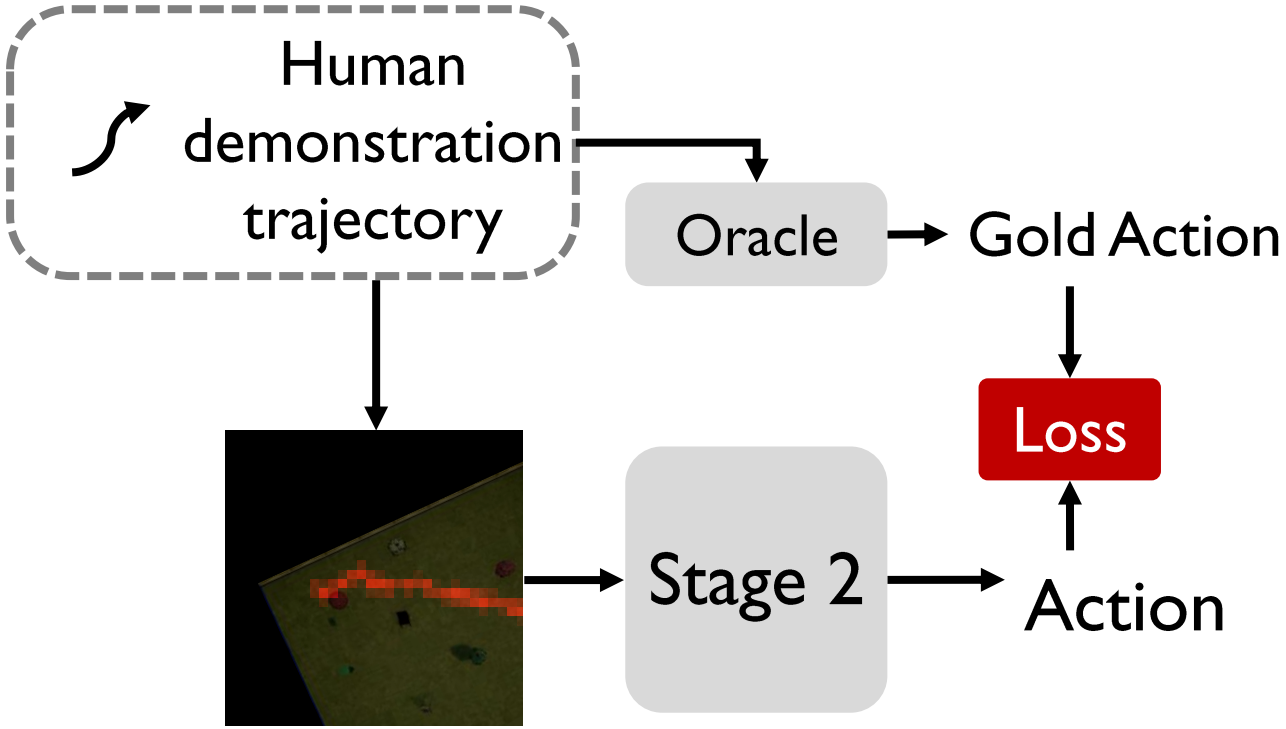
Stage 1
Supervised learning



Predicted Distributions

Human Demonstration Visitation Distributions

Stage 2
Imitation Learning



Human Demonstration Visitation Distributions

Evaluation & Results

We are releasing this benchmark!
https://github.com/clic-lab/drif

- Realistic simulator powered by Microsoft AirSim
- Real, crowdsourced natural language instructions from the LANI corpus
- We achieve state of the art: ~41% success rate

Generalization to Predicting State-Visitation Distributions

- Our approach generalizes to predicting state visitation distributions in an approximation of the true MDP.

- If the MDP approximation is good, then the learned policy has bounded suboptimality with regard to the true MDP.

Thank You!