From VQA to VLN: Recent Advances in Vision-and-Language Research

CVPR 2021 Tutorial

Part 2.3: Forward to Realistic VLN

Yoav Artzi
Cornell University

Peter Anderson
Google Research

Where are we?

Part I: Vision-and-Language Understanding

Part II: Vision-Language Navigation (VLN)

I. Introduction

II. Generalizable VLN Methods

III. Forward to Realistic VLN

Yoav Artzi
Cornell University

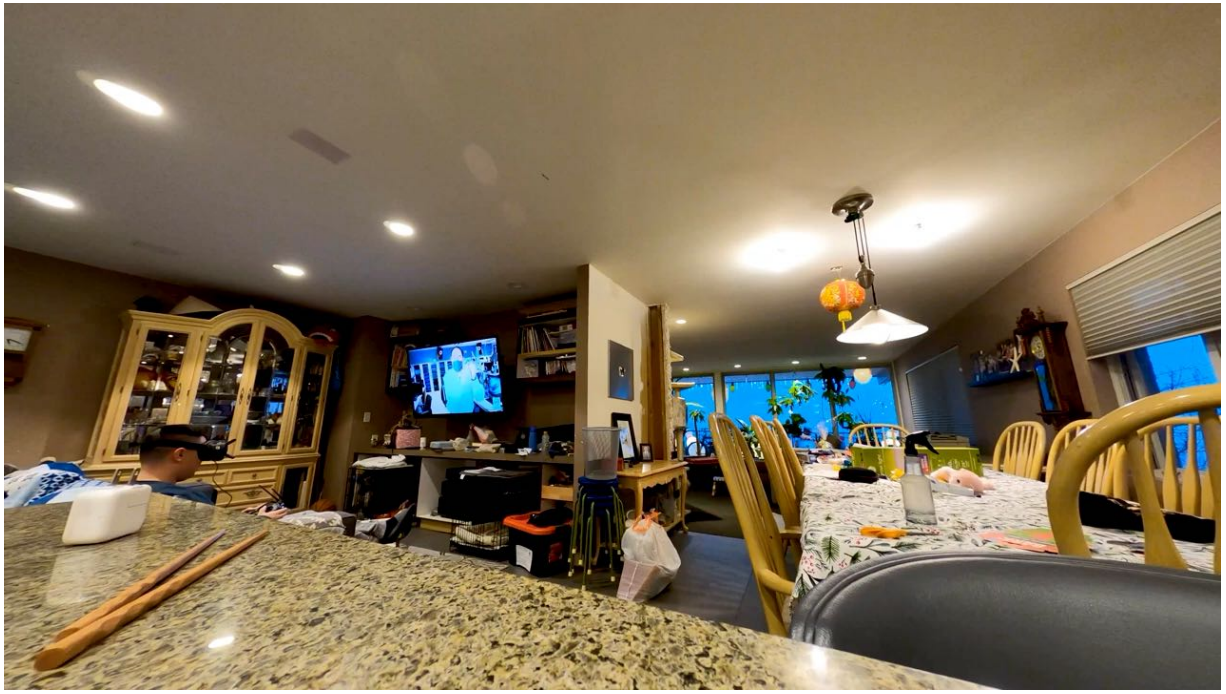
Peter Anderson
Google Research

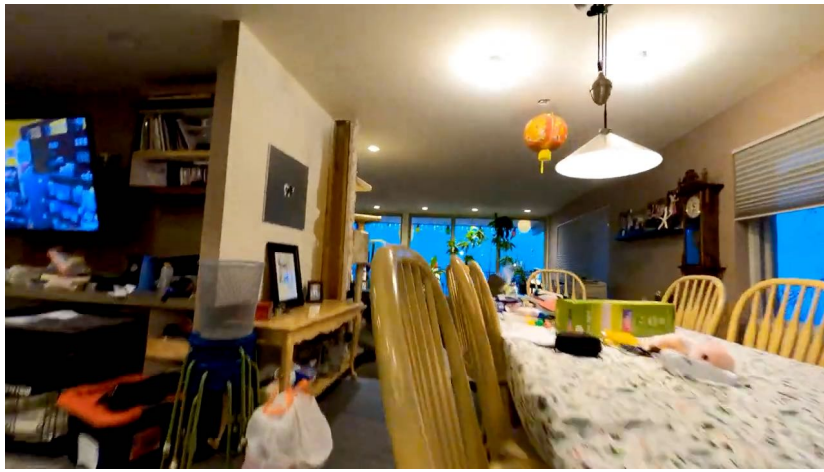

Matterport3D


Matterport3D

https://www.flickr.com/photos/9508280@N07/36994647773/in/photostream/

Real-life Observations

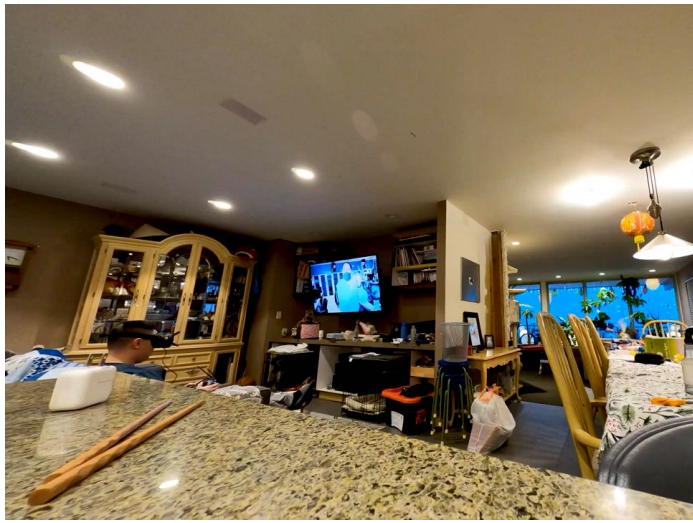




https://youtu.be/tphrhe9BRek

Fundamentally Different

- Not just more complex, but completely different
- More of the same is will not solve this shift
- Visual divergence → different language



Forward to Realistic VLN

Three aspects of the problem, three case studies:

- Real-life observations: urban navigation
- Real-life control: drone instruction following
- Sim2real: from 3D scans to physical buildings

Forward to Realistic VLN

Three aspects of the problem, three case studies:

- Real-life observations: urban navigation
- Real-life control: drone instruction following
- Sim2real: from 3D scans to physical buildings

Real-life Observations: Urban Navigation

- Touchdown: panorama-based VLN and spatial reasoning benchmark in an urban environment
- Today's focus: task and data collection considerations

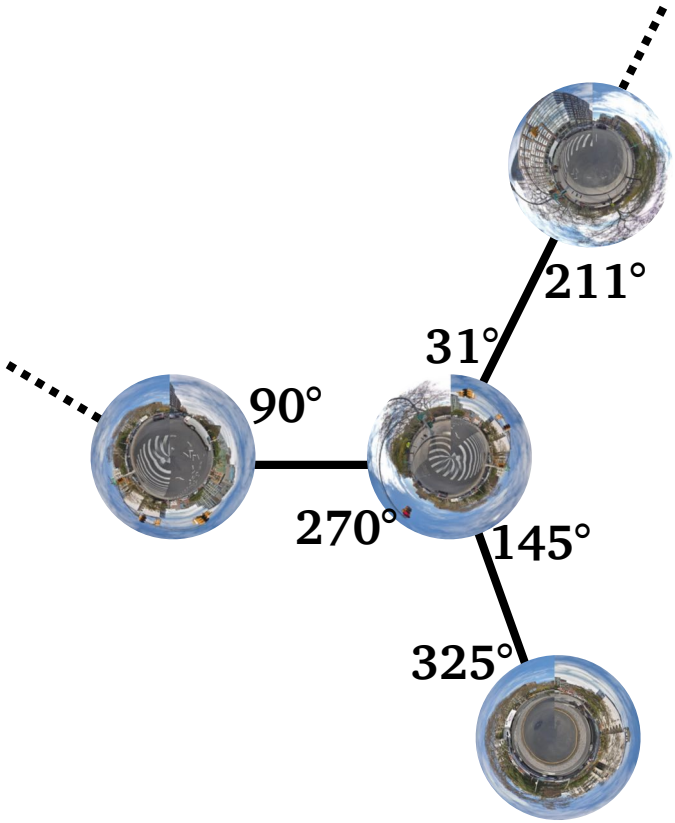
Street View Panoramas

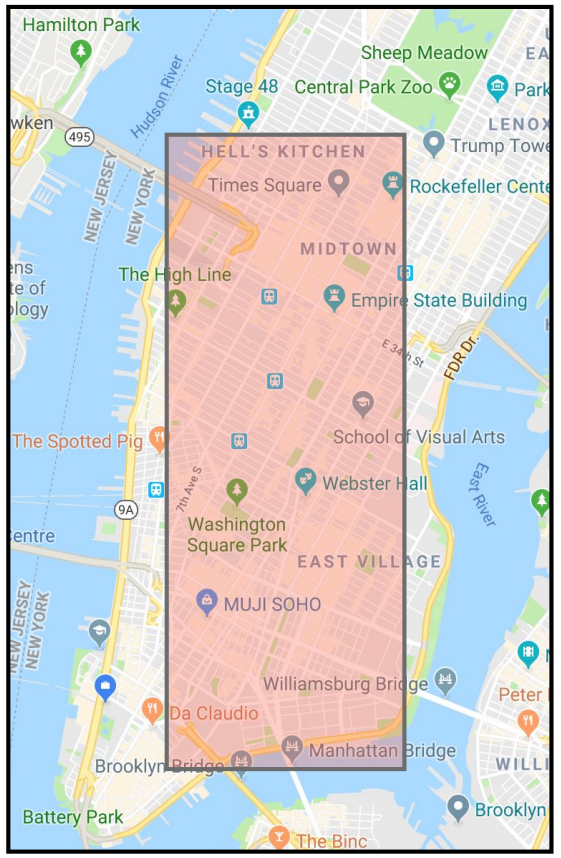
Street View Panoramas

- Object and perspective diversity
- Clutter, a lot of clutter
- Rich movement
- Still: relatively consistent point of view

The Environment

- 29,941 panoramas
- 61,319 edges
- 122,638 states for discrete navigation
- Images are part of StreetLearn





https://sites.google.com/view/streetlearn

Task-focused Navigation

- There is a lot to describe in rich environments
- Without a clear aim, it leads to overly verbose and descriptive language
- But this is not how we refer to objects or give instructions — language is generated with intent
- Focus navigation on a clear task, and incentivize instructors with a game-like dynamics

Task-focused Navigation

- Writing task: instruct to follow a path and describe the location of an object they hide
- The focused task makes the instruction more natural for the writer
- Guide workers not to count intersections and not to use text and store names
- What do we hide?

Example

Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right. Touchdown is on the back of the dinosaur.

Task-focused Navigation

- This formulation allows for multiple tasks:
  - **Navigation** only: given instruction and a starting point, navigate to the goal position
  - **Spatial description resolution (SDR)** only: given a sentence and a panorama, find Touchdown
  - **The complete task**: navigate first, and then find Touchdown

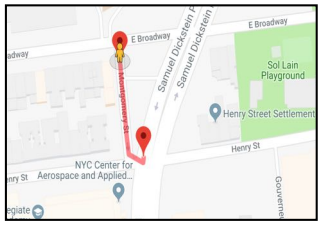
Data Collection

- A sequence of four separate tasks on MTurk
  - Writing, propagation, validation, and segmentation
  - Leader instruction writing is tied immediately to measurable execution by a follower
- Workers use a customized Street View environment

Task I: Writing

Place Touchdown

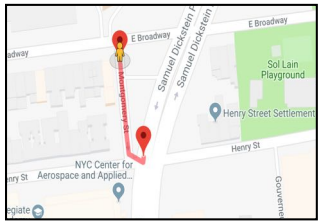
Can't Place Touchdown



Task I: Writing

Place Touchdown

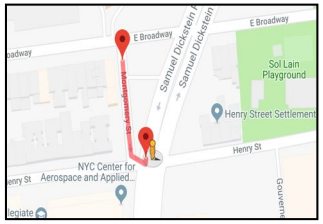
Can't Place Touchdown



Turn so that the trees are to your left. At the first intersection, turn left and stop.

Place Touchdown

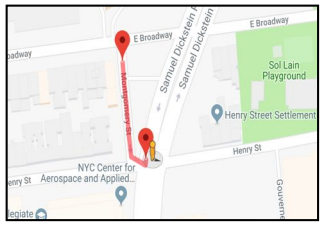
Can't Place Touchdown



Task I: Writing

Place Touchdown

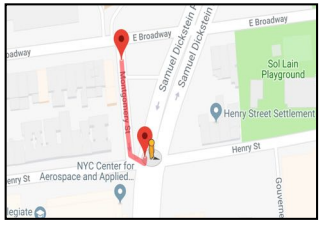
Can't Place Touchdown



Turn so that the trees are to your left. At the first intersection, turn left and stop. **Touchdown is on top of the blue mailbox on the right hand corner.**

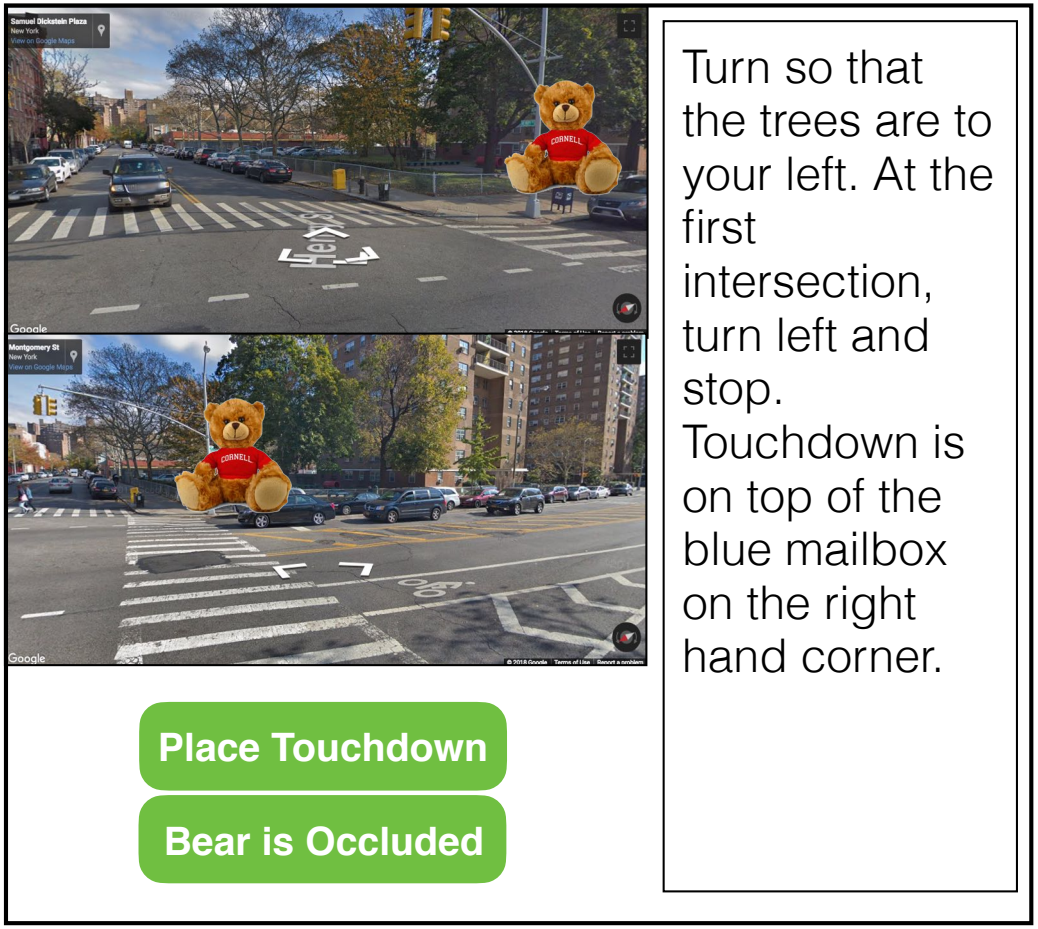
Place Touchdown

Can't Place Touchdown



Task II: Propagation

- Touchdown position may be visible from multiple panoramas
- We propagate the location to neighboring panoramas



Task III: Validation

- Validate instruction by finding Touchdown
- Easy to verify
- Give bonuses to original writer and follower if successful

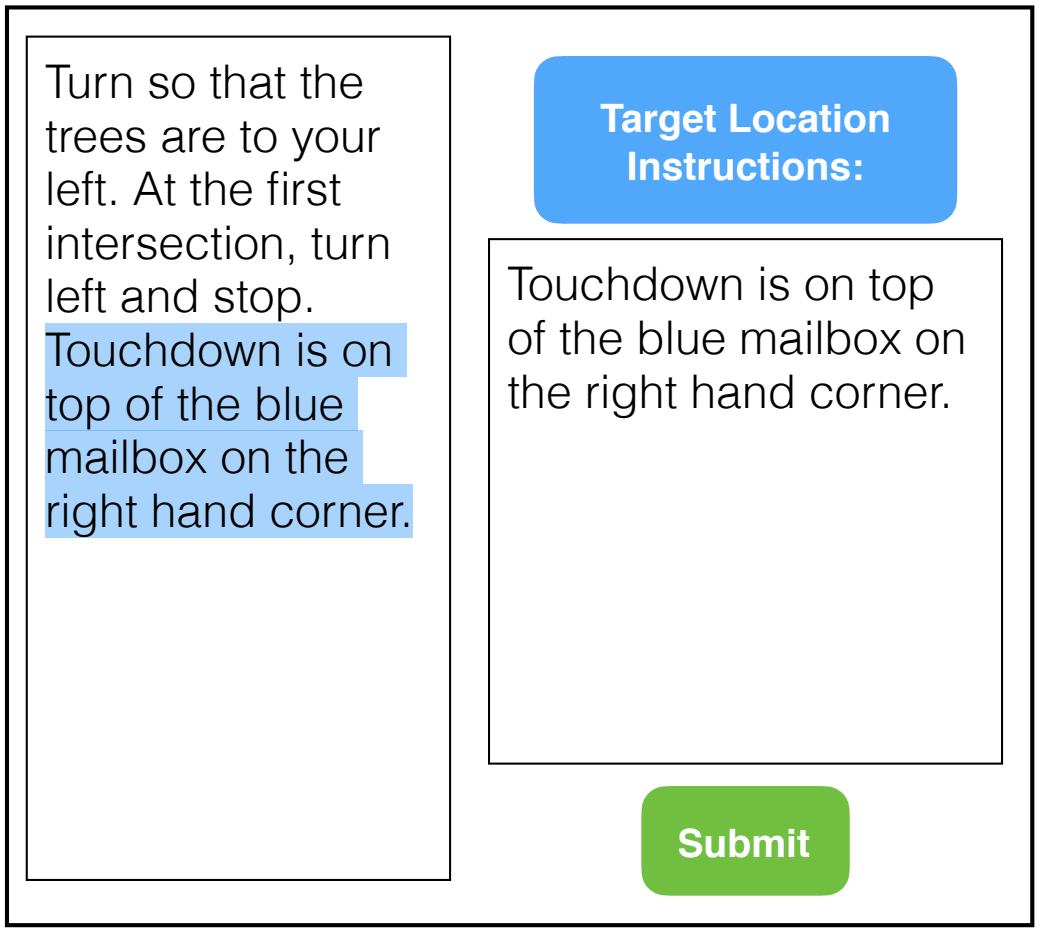
Turn so that the trees are to your left. At the first intersection, turn left and stop. Touchdown is on top of the blue mailbox on the right hand corner.

You Found Touchdown!

Remaining Attempts: 2

Task IV: Task Segmentation

- Segment the text to the two tasks: navigation and SDR
- Segments may overlap



What Did We Get?

- Over 200 people wrote and validated instructions
- Collected 9,326 examples, split to 6,526/1,391/1,409 for train/dev/test

Analysis

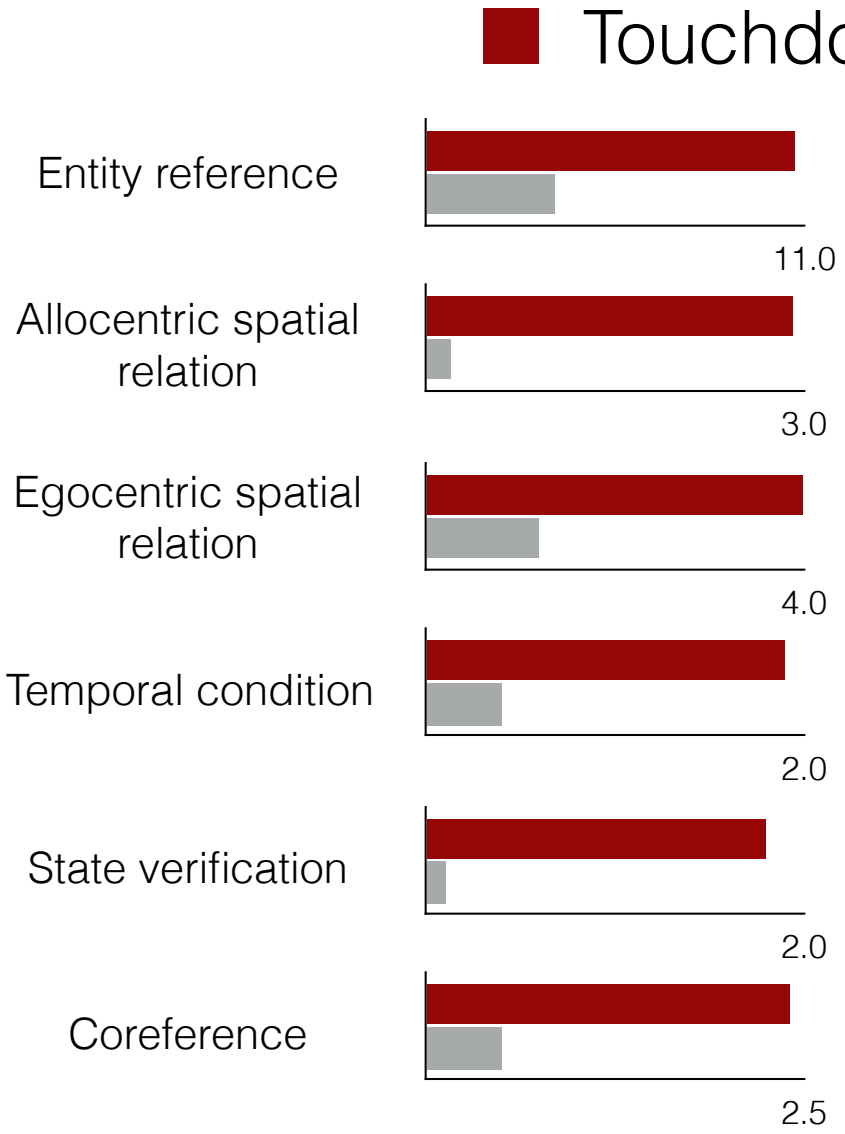
- Average length is 108 tokens on average
  - 89.6 for navigation, compared to 29.3 in R2R
  - 29.8 for SDR, compared to 8.5 in Google RefExp and 4.4 in ReferItGame
- Relatively large vocabulary size of 5,625, compared 3,156 for R2R
- Paths are on average 35.2 panoramas, compared to 6 in R2R

[R2R: Anderson et al. 2018; Google RefExp: Mao et al. 2016; ReferItGame: Kazemzadeh et al. 2014]

Language Analysis

- Sampled 25 examples from Touchdown and R2R
- Analyzed for 11 semantic categories
- Report the mean number of instances per example (more analysis in the paper)
- First used this type of analysis in NLVR, adopted in recent datasets (NLVR2, RxR, Refer360)

Linguistic Analysis



*…You'll pass **three trashcans** on your left …*

*… There is a fire hydrant, the bear is **on top** …*

*… up ahead there is some flag poles **on your right hand side** …*

*… Follow the road **until you see** a school on your right …*

*… **You should see** a small bridge ahead …*

*… a brownish colored brick building with a black fence around **it** …*

Spatial Description Resolution

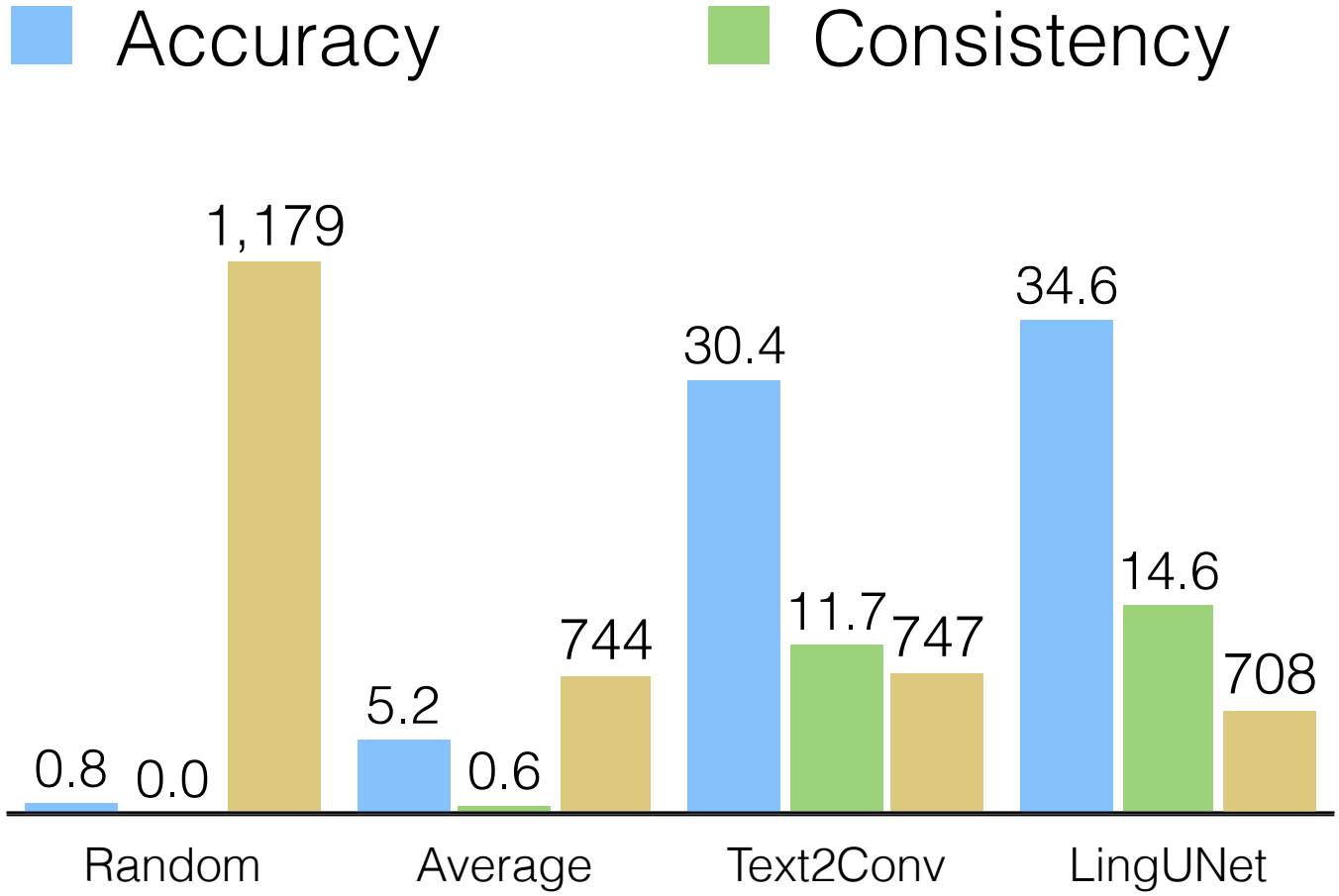
There is also a dinosaur mural to the right.
Touchdown is on the back of the dinosaur.

Where is Touchdown?

SDR Evaluation

- **Accuracy:** predicting the position close enough to the gold position (threshold: 80px)
- **Consistency:** consider a unique SDR as correct only if solved for all propagated panoramas
- **Mean distance error:** the distance of the predicted position from the gold position

Test Results



A lot of room left for improvement

Accuracy, consistency, and distance are not to scale

Example: LingUNet

If you turn right you'll see a narrow staircase and a door next to a fence. Touchdown is at the top of the staircase ✓

Example: LingUNet

a black doorway with red brick to the right of it, and green brick to the left of it. it has a light just above the doorway, and on that light is where you find Touchdown

Navigation

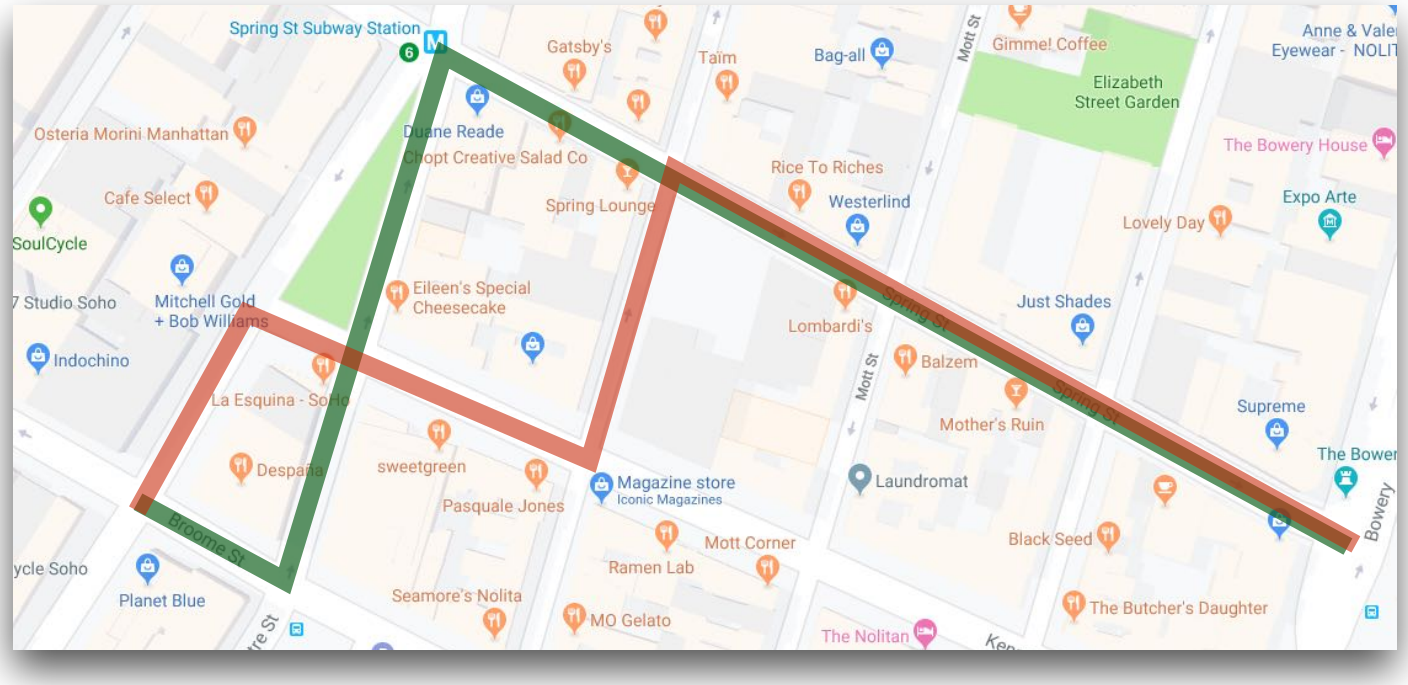
Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right.

Navigation Evaluation

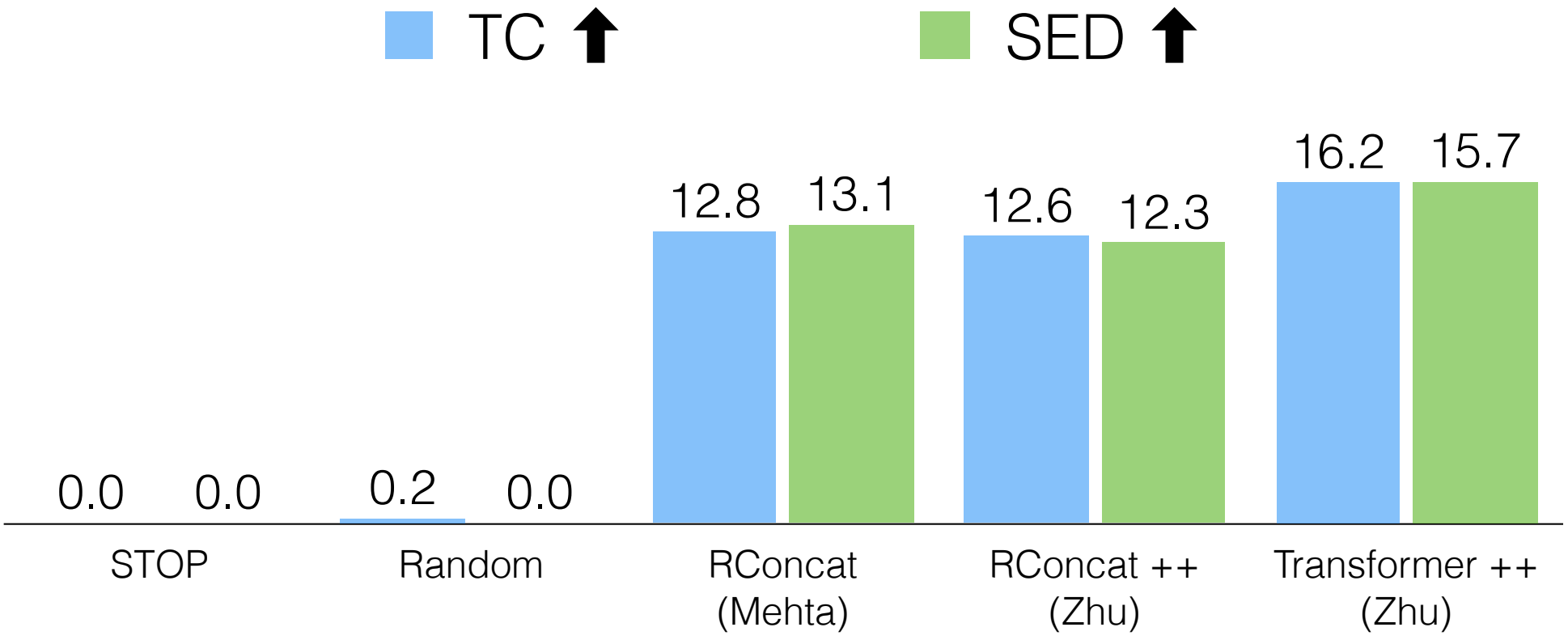
- Accuracy: stopping at the annotated goal panorama, or to one of the propagated panoramas
- Success-weighted by edit distance (SED)

Success weighted by Edit Distance (SED)

- Measure edit distance between reference and prediction
- Weight success by distance
- The closer the agent is to the correct execution, success is considered better



Test Results



Some improvements from
Transformers + pre-training + augmentation

[Mehta et al. 2020; Zhu et al. 2021]

Getting Touchdown

- Panoramas are part of StreetLearn: https://sites.google.com/view/streetlearn
- Language and demonstration data: https://github.com/lil-lab/touchdown

Related Tasks

SDR

- Refer360: A Referring Expression Recognition Dataset in 360 Images [Cirik et al. 2020]
  Slightly different SDR setup indoor panoramas

Navigation

- Talk the Walk: Navigating New York City through Grounded Dialogue [de Vries et al. 2018]
  Small set of panoramas, dialogue

- Learning To Follow Directions in Street View [Hermann et al. 2019]
  Google Maps synthetic instructions

- Learning to Navigate in Cities Without a Map [Mirowski et al. 2018]
  Goal navigation, without language

- DeepNav: Learning to Navigate Large Cities [Brahmbhatt and Hays 2017]
  Landmark navigation, without language

Forward to Realistic VLN

Three aspects of the problem, three case studies:

- Real-life observations: urban navigation
- Real-life control: drone instruction following
- Sim2real: from 3D scans to physical buildings

Real-life Control: Drone Instruction Following

- Navigation between landmarks
- Agent: quadcopter drone
- Context: pose and RGB camera image
- Today's focus: model design

Task

after the blue bale take a right towards the small white bush before the white bush take a right and head towards the right side of the banana

Mapping Instructions to Control

- The drone maintains a **configuration** of target velocities

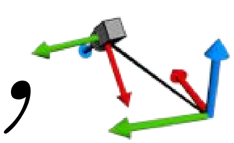
Linear forward velocity — Angular yaw rate

$$(v, \omega)$$

- Each action updates the configuration or stops

- Goal: learn a mapping from inputs to configuration updates

$f($ *after the blue bale take a right towards the small white bush before the white bush …* ,

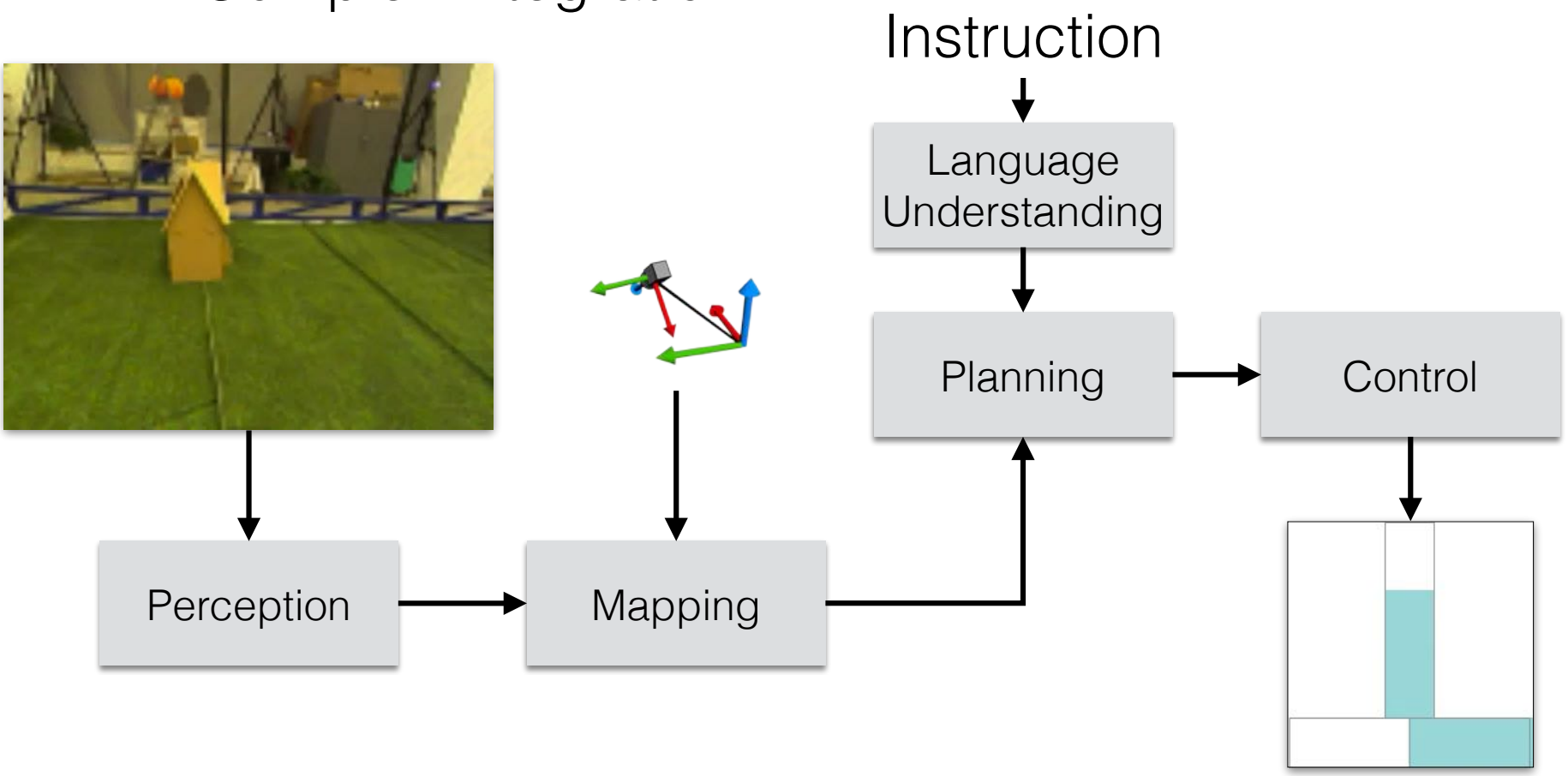


,

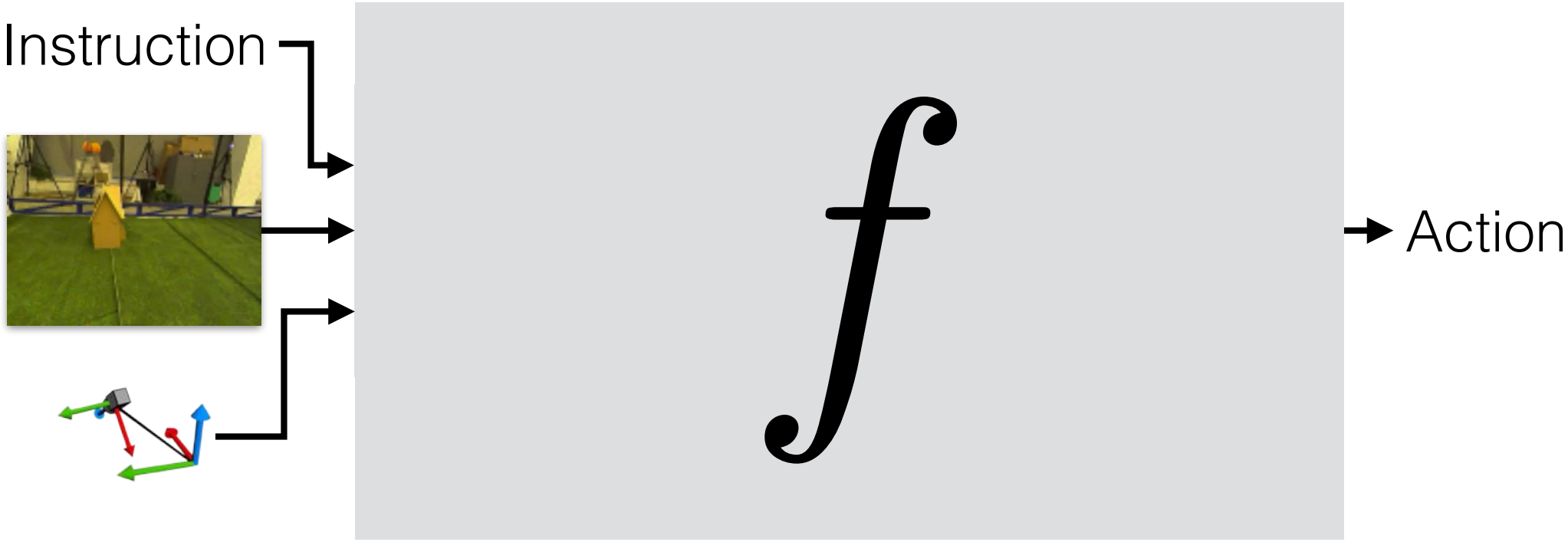
$) = STOP$

Modular Approach

- Build/train separate components
- Symbolic meaning representation
- Complex integration

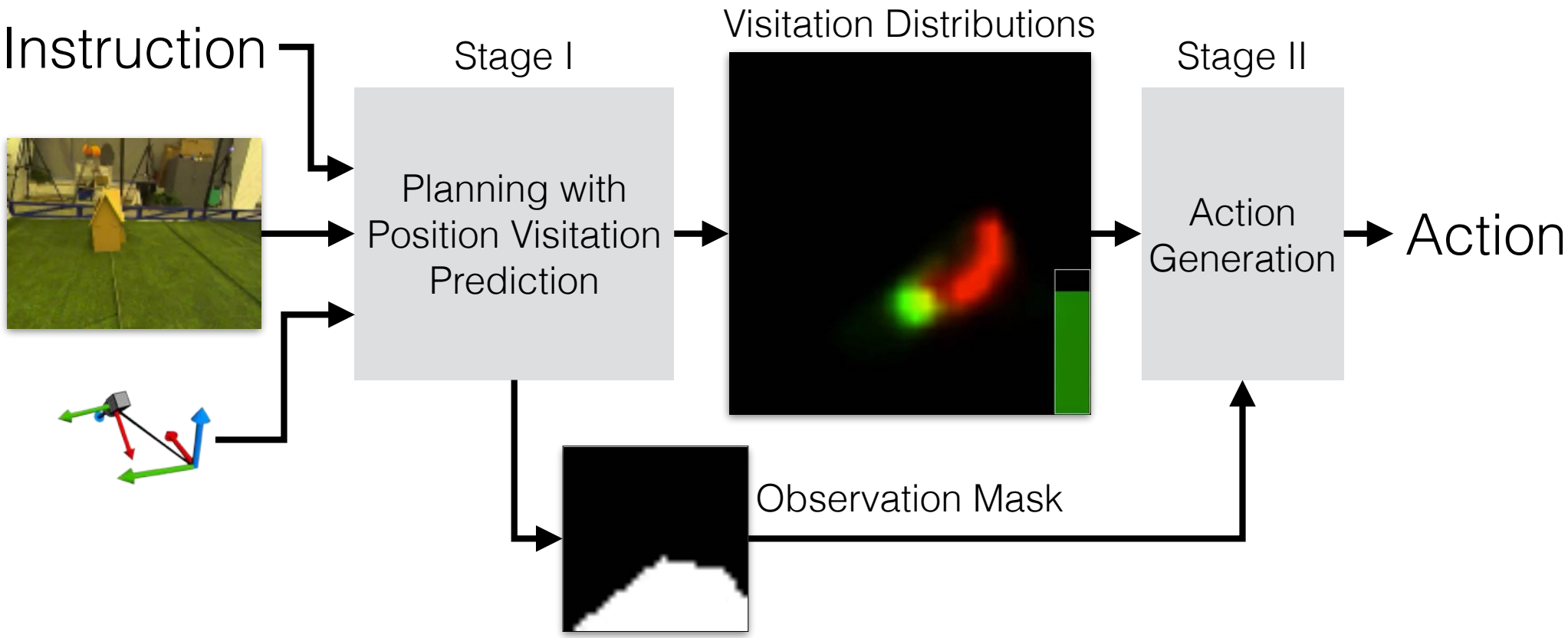


Single-model Approach



How to think of modularity and interpretability when packing everything in a single model?

Single-model Approach



1. Predict states likely to visit and track accumulated observability

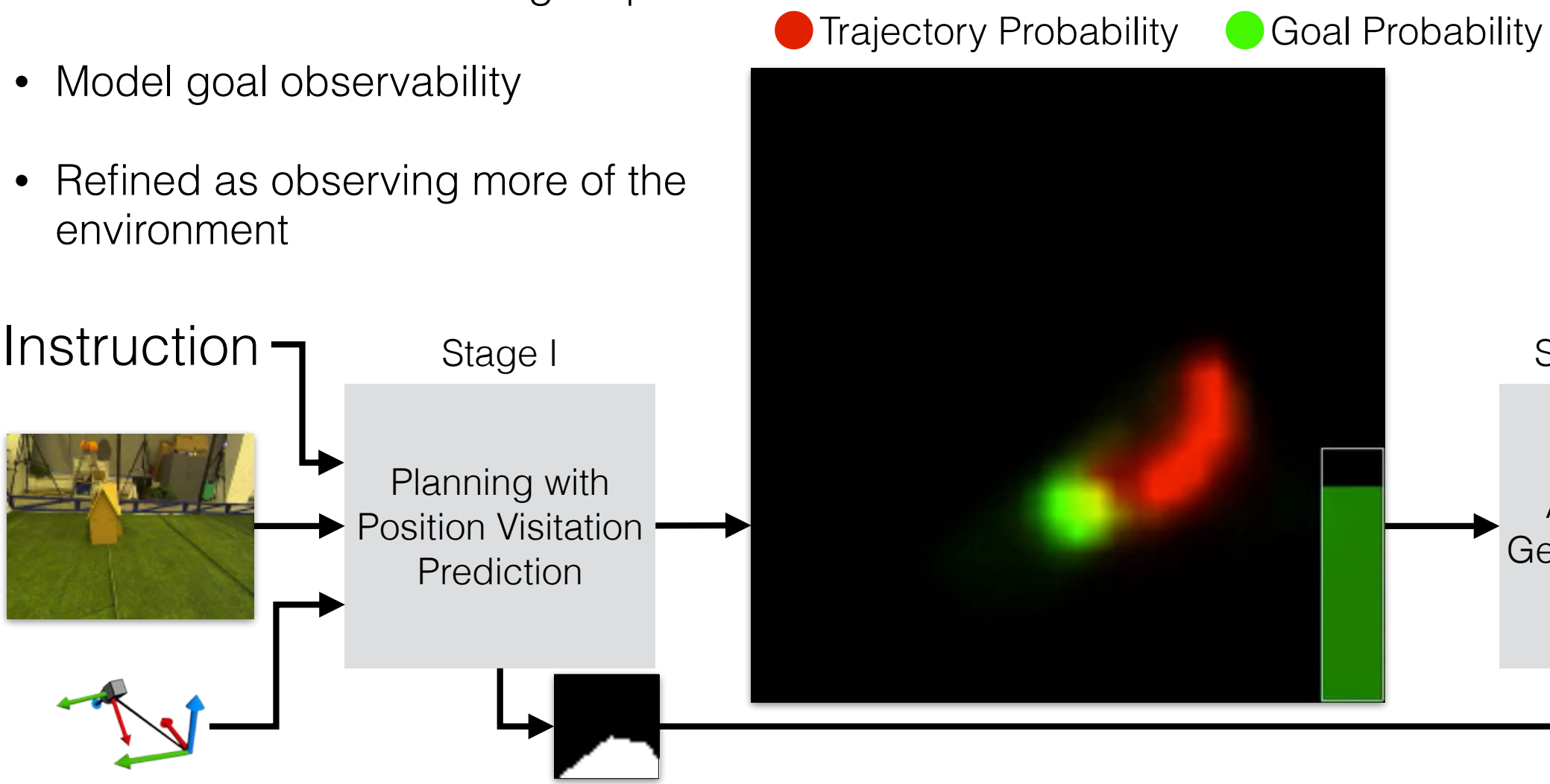
2. Generate actions to visit high-probability states and explore

Visitation Distribution

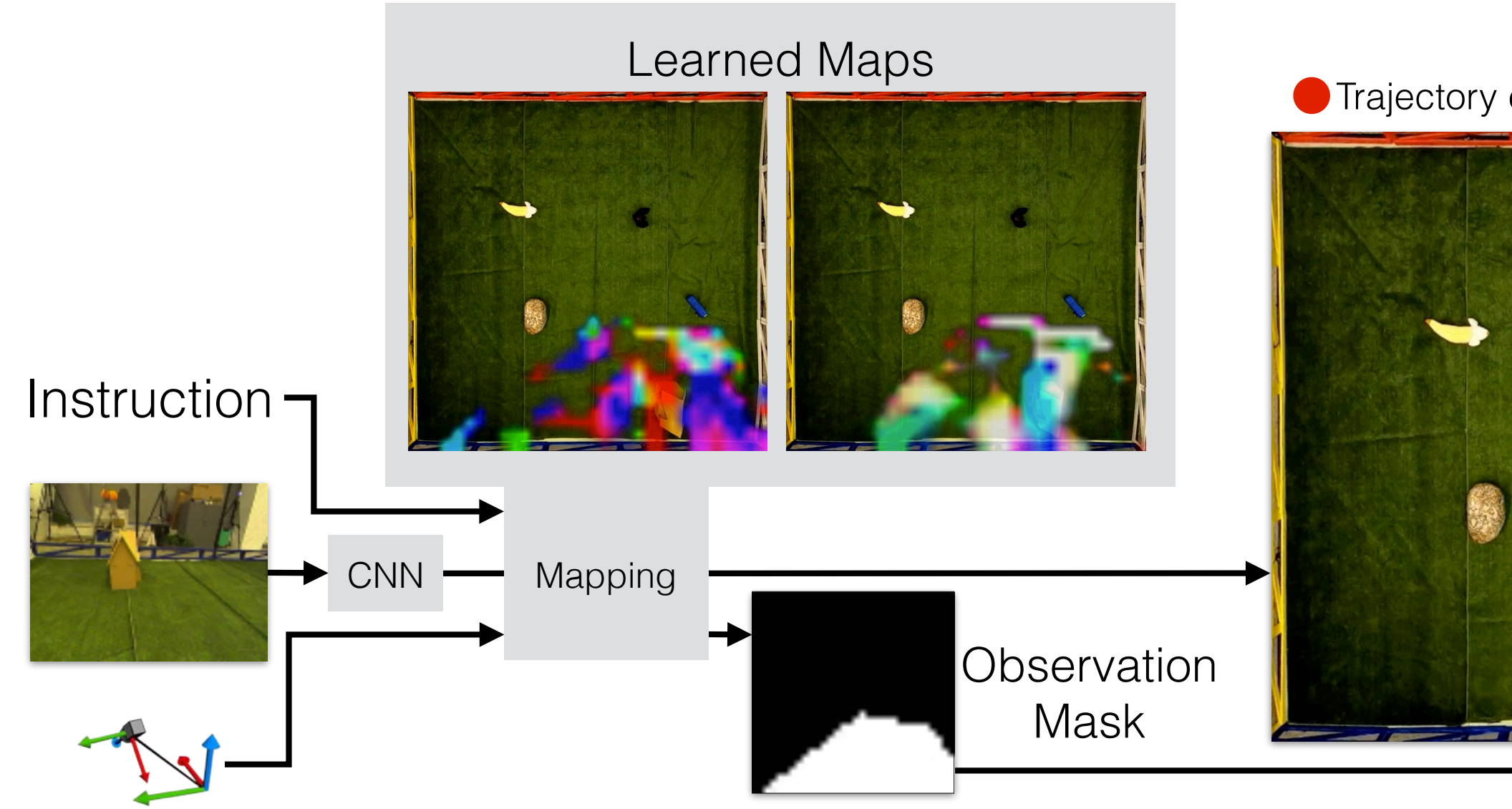
- The state-visitation distribution $d(s; \pi, s_0)$ is the probability of visiting state s following policy π from start state s_0
- Predicting $d(s; \pi^*, s_0)$ for an expert policy π^* tells us the states to visit to complete the task
- We compute two distributions: **trajectory-visitation** and **goal-visitation**

Visitation Distribution for Navigation

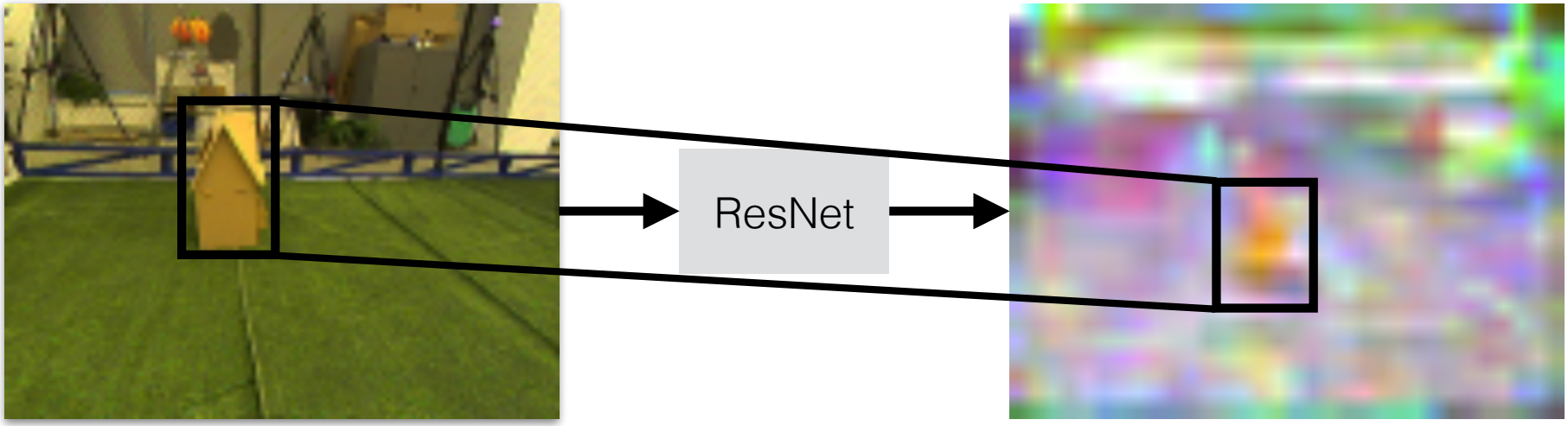
- Distributions reflect the agent plan
- Model goal observability
- Refined as observing more of the environment



Stage I: Planning with Position Visitation Prediction



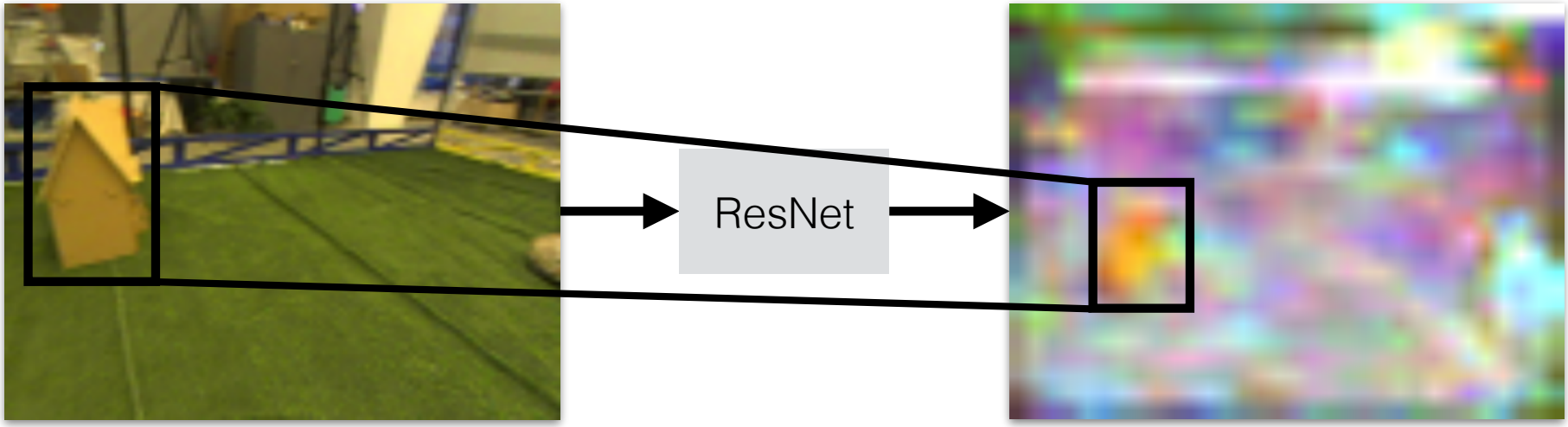
Differentiable Mapping
Step 1: Feature Extraction



- Extract features with a ResNet
- Recover a low resolution semantic view

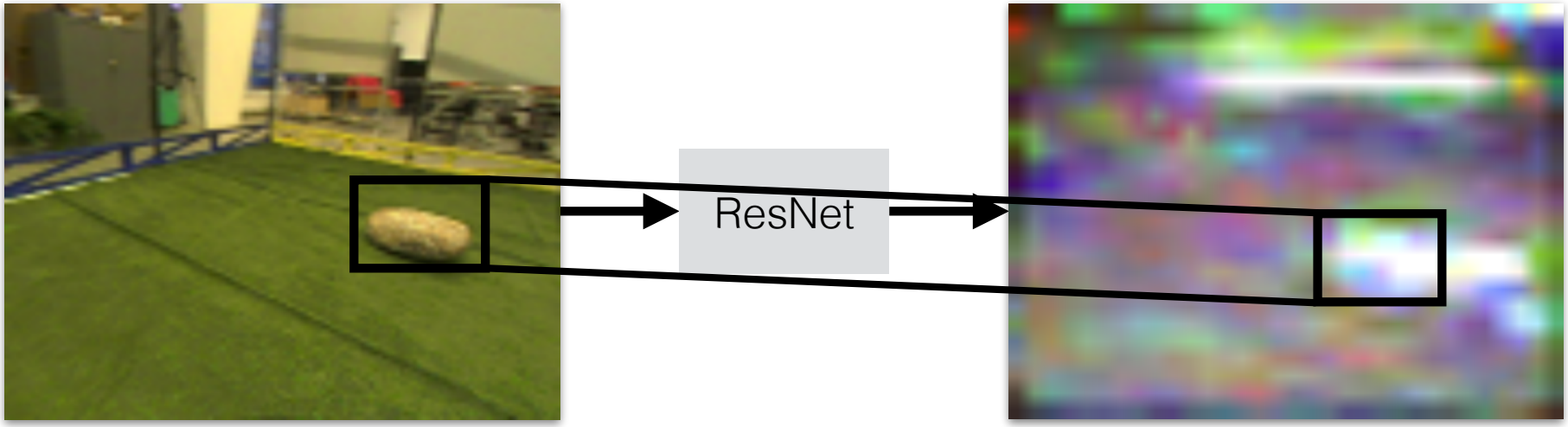
Differentiable Mapping

Step 1: Feature Extraction



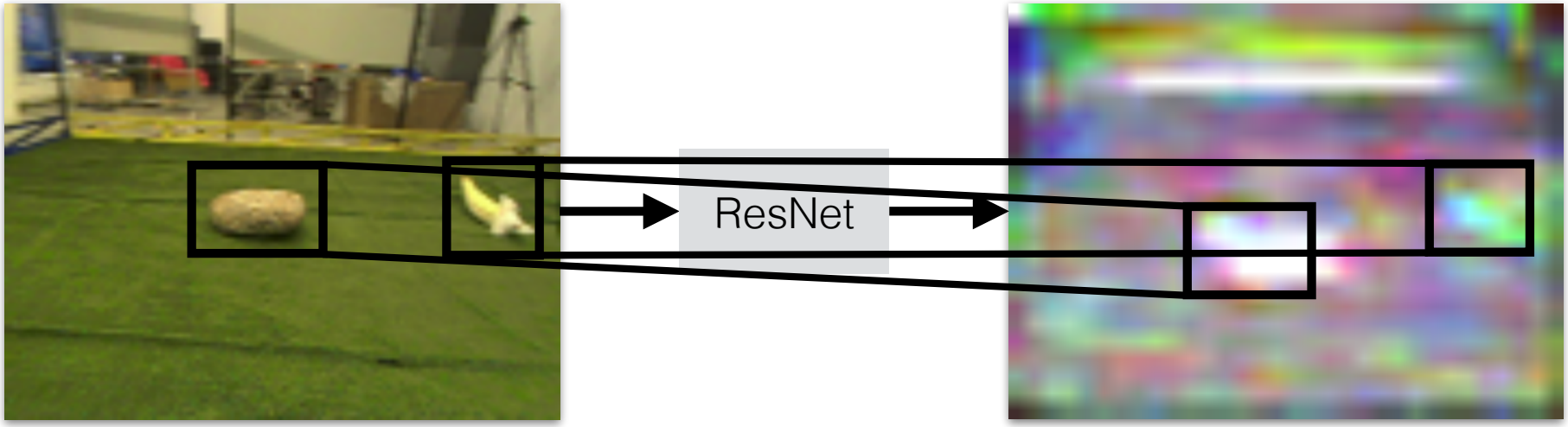
- Extract features with a ResNet
- Recover a low resolution semantic view

Differentiable Mapping
Step 1: Feature Extraction



- Extract features with a ResNet
- Recover a low resolution semantic view

Differentiable Mapping
Step 1: Feature Extraction

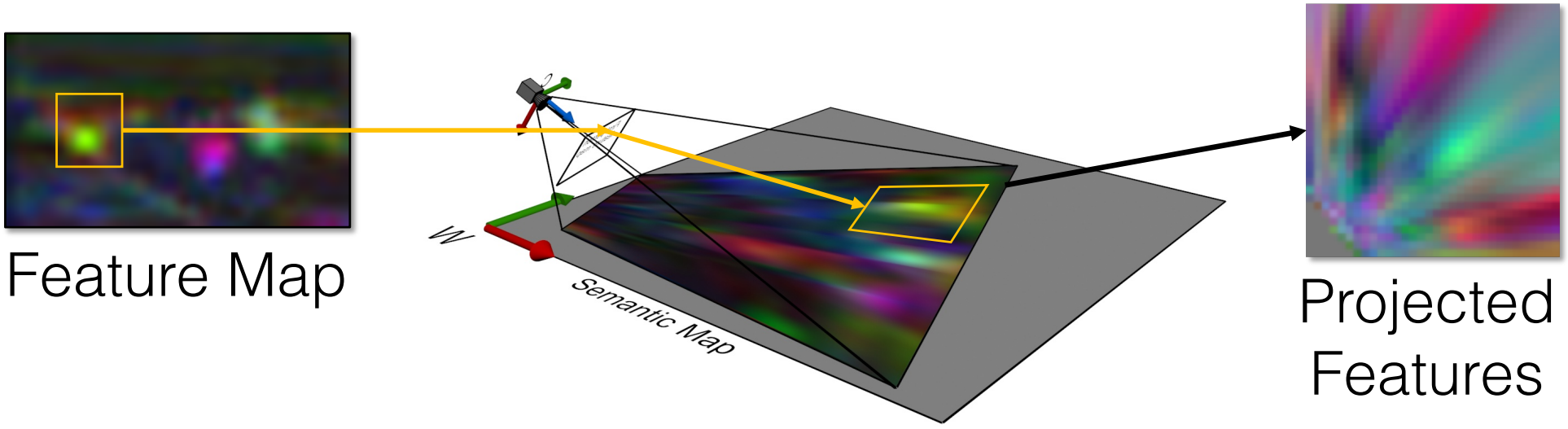


- Extract features with a ResNet
- Recover a low resolution semantic view

Differentiable Mapping

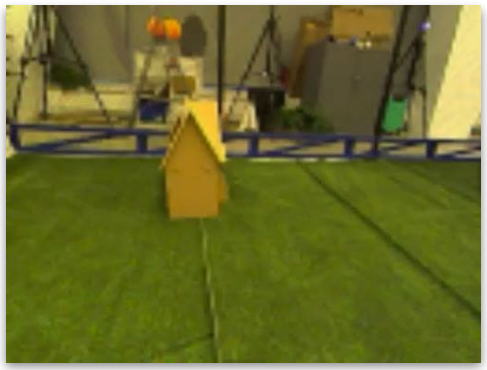
Step 2: Projection

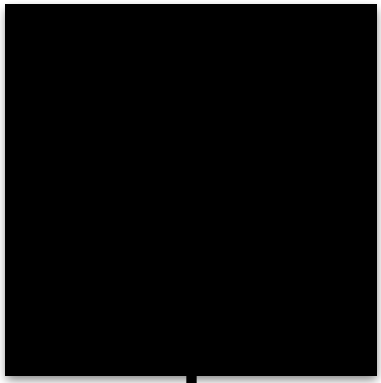
- Deterministic projection from camera image plane to environment ground with pinhole camera model
- Transform from first-person to third-person



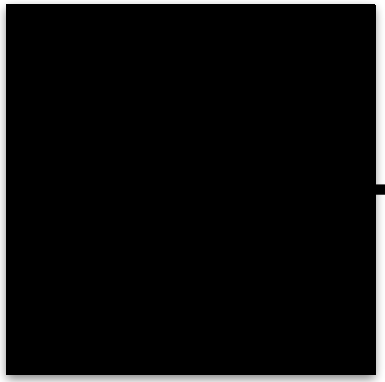
Differentiable Mapping

Step 3: Accumulation



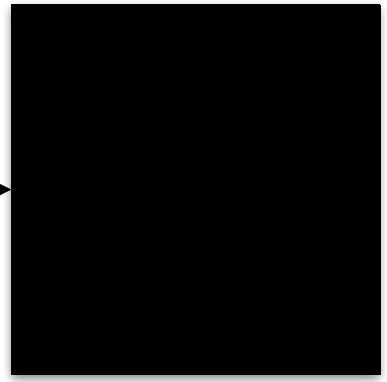


Semantic Map
(time t-1)



Σ

Leaky
Integrator

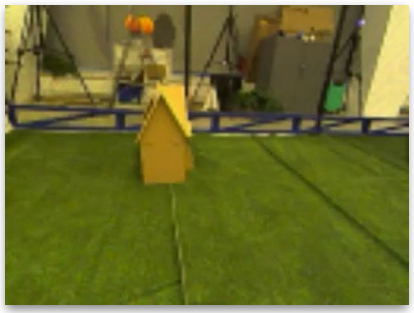


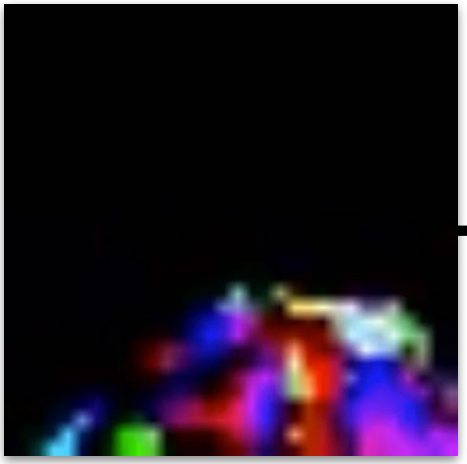
Projected Features
(time t)

Semantic Map
(time t)

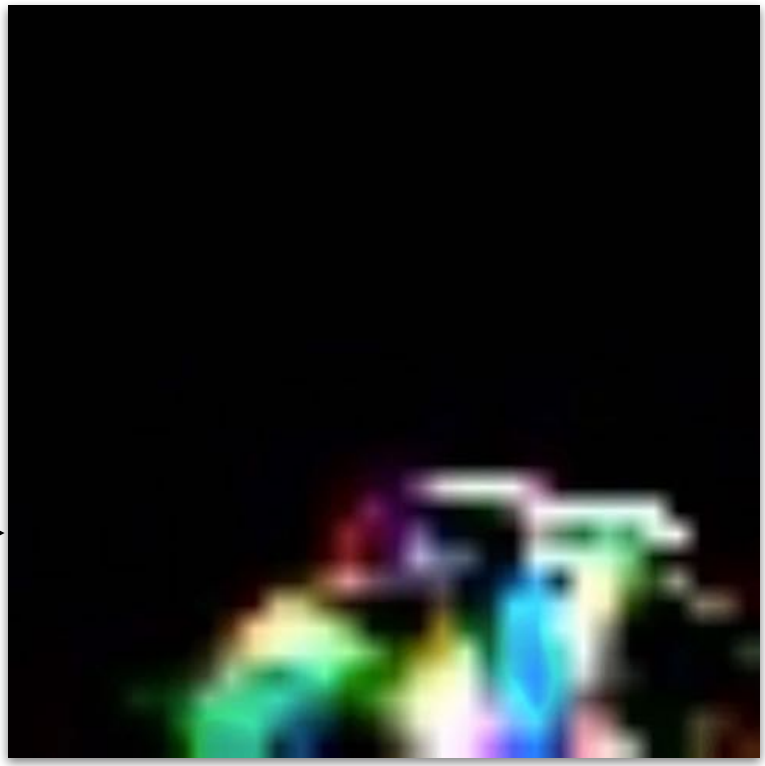
Differentiable Mapping

Step 4: Grounding





1x1 Filter

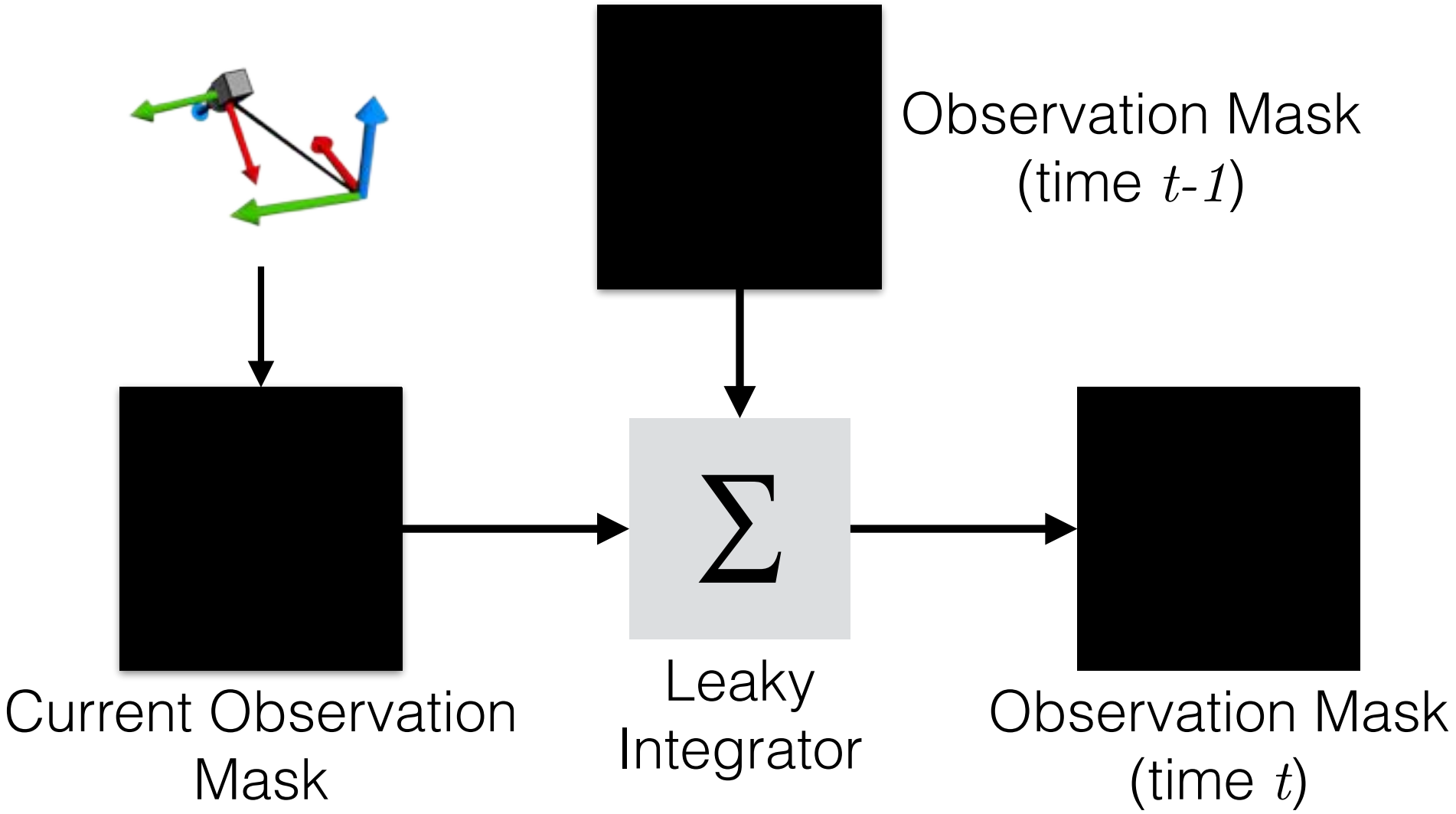


Semantic Map

Grounding Map

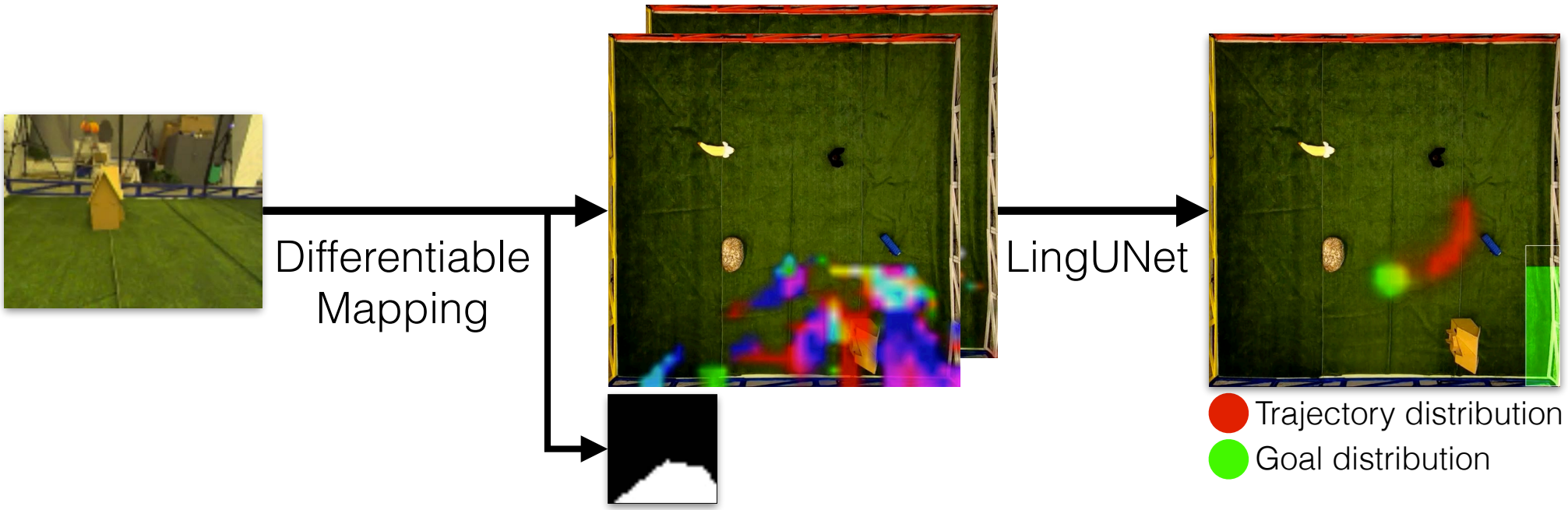
after the blue bale take a right towards the small white bush before the white bush …

Observability Mask



Stage I: Planning with Position Visitation Prediction

- ✓ Extract visual features and construct maps
- • Compute visitation distributions over the maps

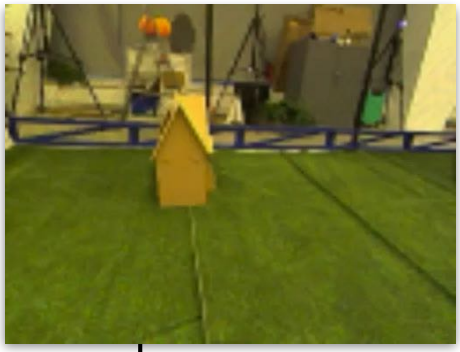


Predicting Visitation Distributions

- We compute two distributions: **trajectory-visitation** and **goal-visitation**
- Cast distribution prediction as image generation

LingUNet

- Image-to-image encoder-decoder
- Visual reasoning at multiple image scales
- Conditioned on language input at all levels of reasoning using text-based convolutions

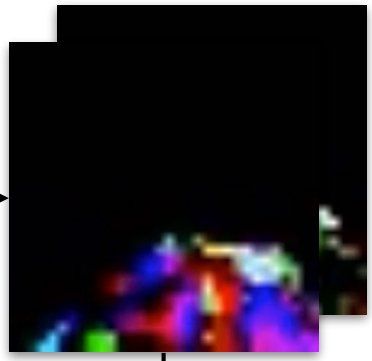


LingUNet

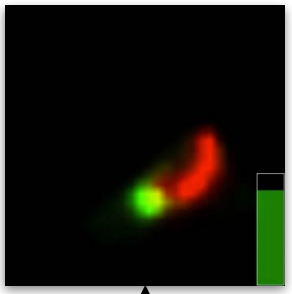
Semantic Maps

Instruction

Visitation Distributions



RNN



SoftMax

Text Kernels

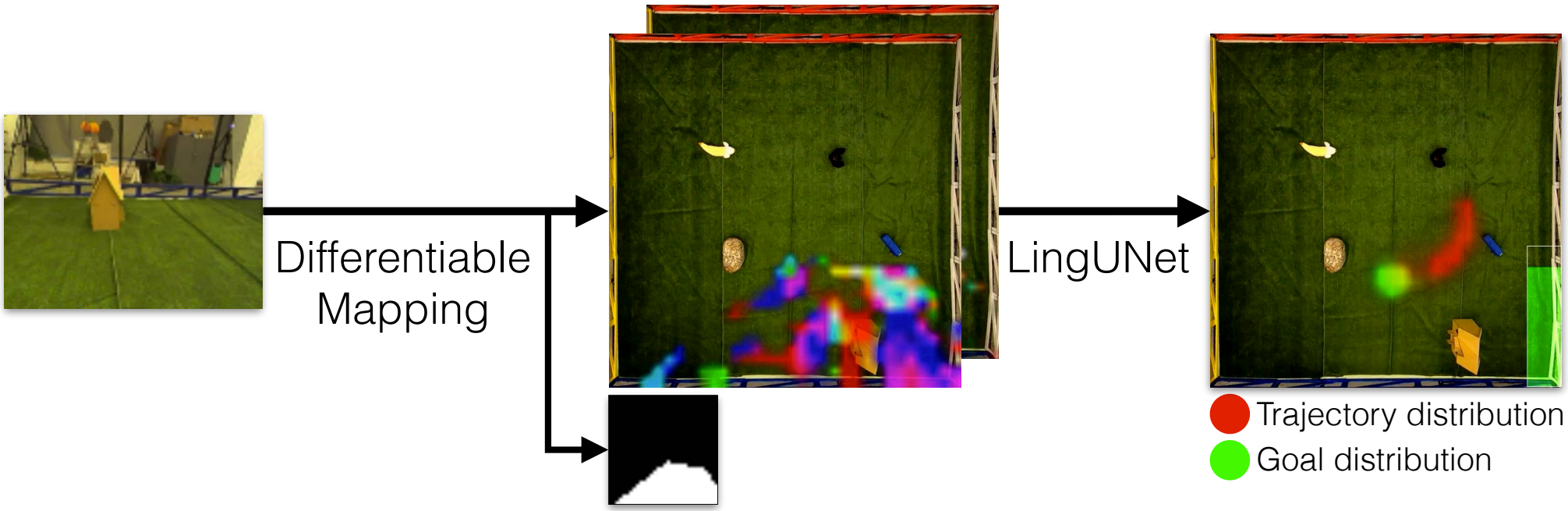
Convolutions

Deconvolutions

Text Convolutions

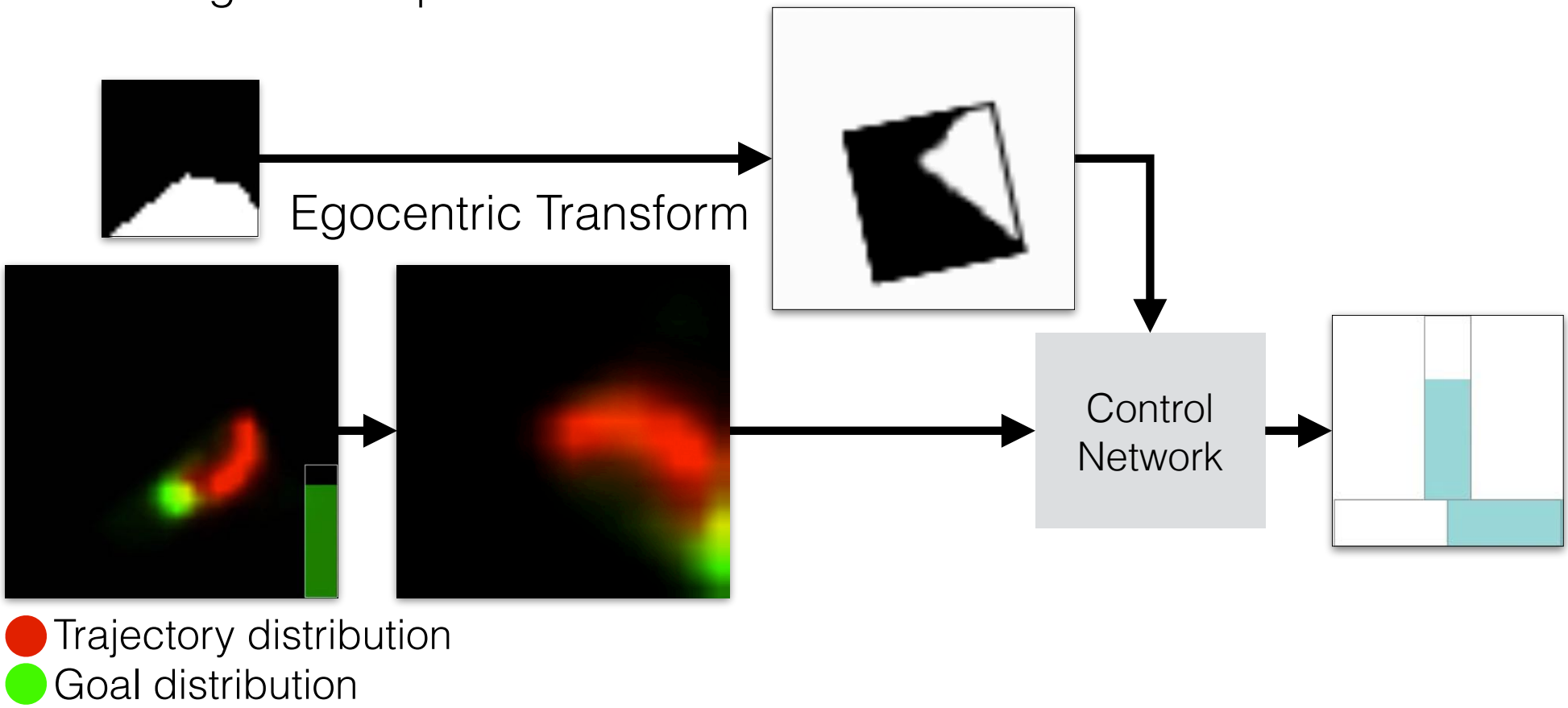
Stage I: Planning with Position Visitation Prediction

- ✓ Extract visual features and construct maps
- ✓ Compute visitation distributions over the maps



Stage II: Action Generation

- Relatively simple control problem without language
- Transform and crop to agent perspective and generate configuration update



Learning vs. Engineering

Learned 👨‍🎓

- Visual features (ResNet)
- Text representation (RNN)
- Image generation (LingUNet)
- Control (control network)

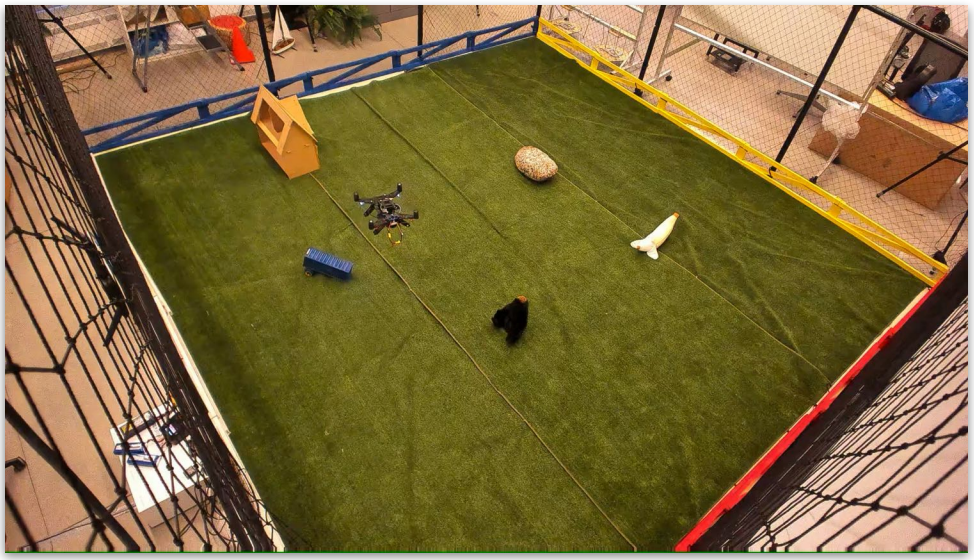
Engineered 👷

- Feature projections (pinhole camera model)
- Map accumulation (leaky integrator)
- Egocentric transformation (matrix rotations)

Complete network remains fully differentiable

Simulation-Reality Joint Learning

Go between the mushroom and flower chair the tree all the way up to the phone booth



after the blue bale take a right towards the small white bush before the white bush take a right and head towards the right side of the banana

Training Data

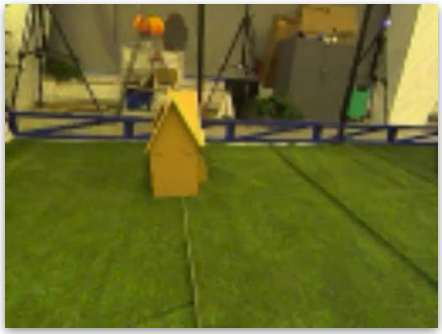
Simulator

Demonstrations and simulator

Go between the mushroom and flower chair the tree all the way up to the phone booth

Physical Environment

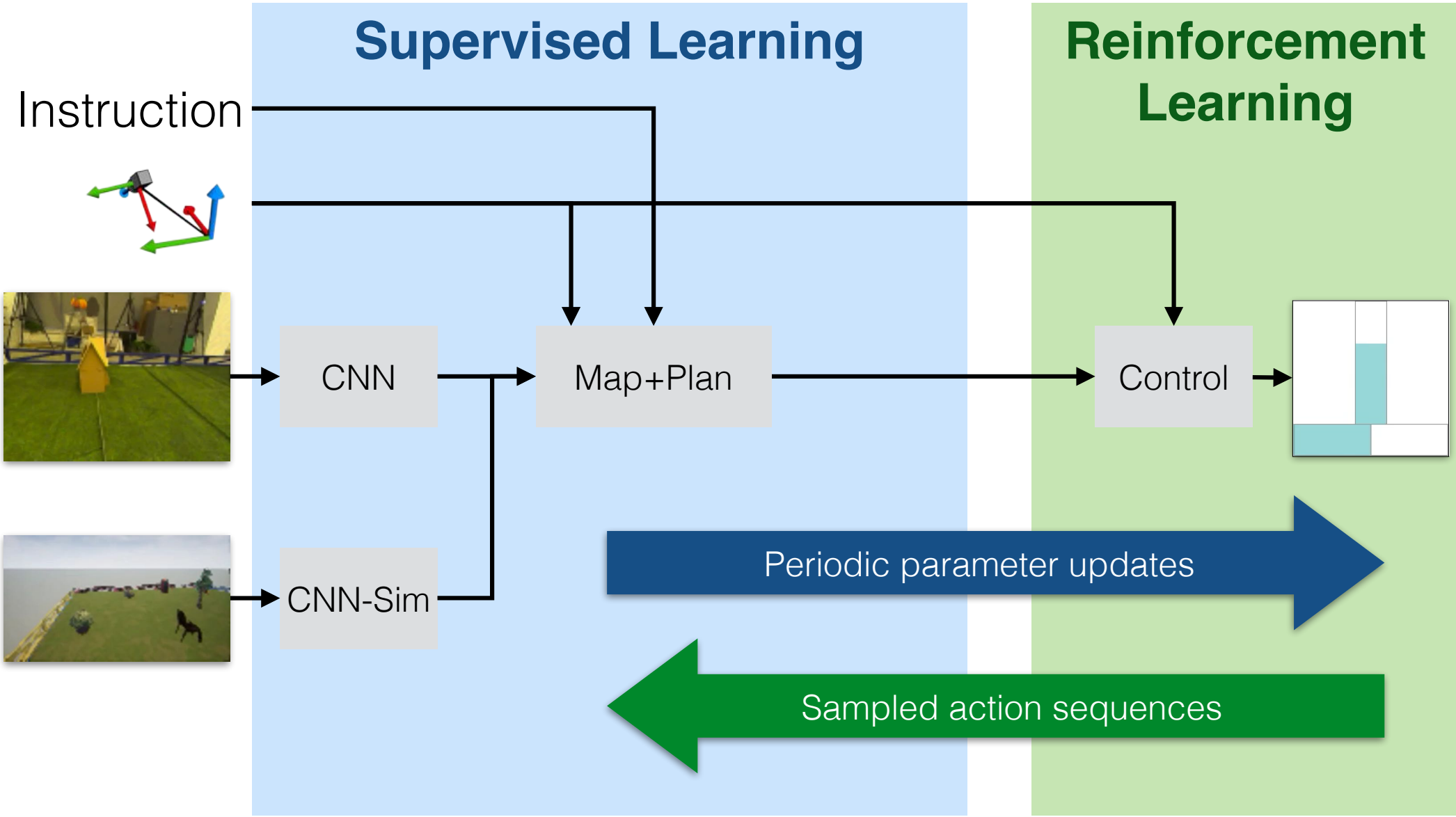
Demonstrations only



after the blue bale take a right towards the small white bush before the white bush take a right and head towards the right side of the banana

SuReAL

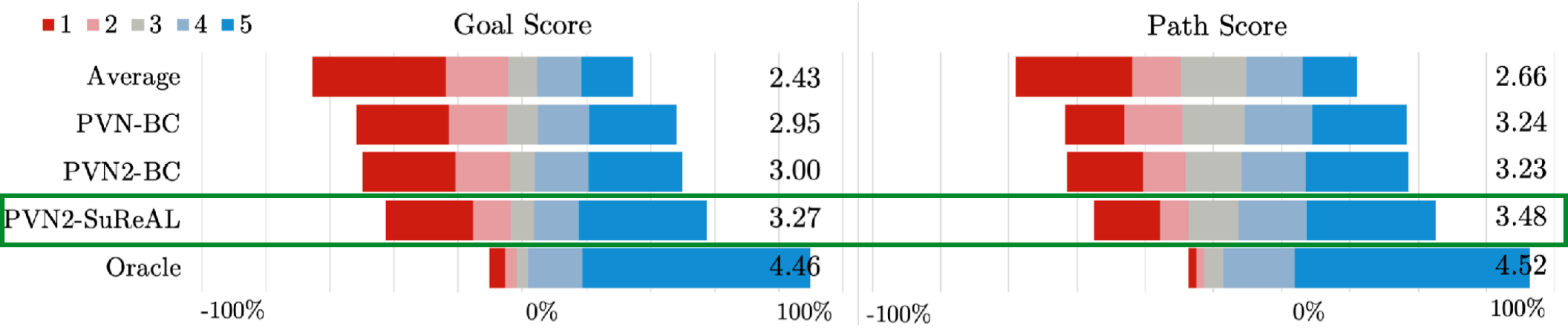
Supervised and Reinforcement Asynchronous Learning



Experimental Setup

- Intel Aero quadcopter
- Vicon motion capture for pose estimate
- Simulation with Microsoft AirSim
- Drone cage is 4.7x4.7m
- Roughly 1.5% of training data in physical environment (402 vs. 23k examples)

Human Evaluation



- Score path and goal on a 5-point Likert scale for 73 examples
- Our model receives five-point path scores 37.8% of the time, 24.8% improvement over PVN2-BC
- Improvements over PVN2-BC illustrates the benefit of SuReAL and the exploration reward

Cool Example

once near the rear of the gorilla turn right and head towards the rock stopping once near it

Forward to Realistic VLN

Three aspects of the problem, three case studies:

- Real-life observations: urban navigation
- Real-life control: drone instruction following
- Sim2real: from 3D scans to physical buildings