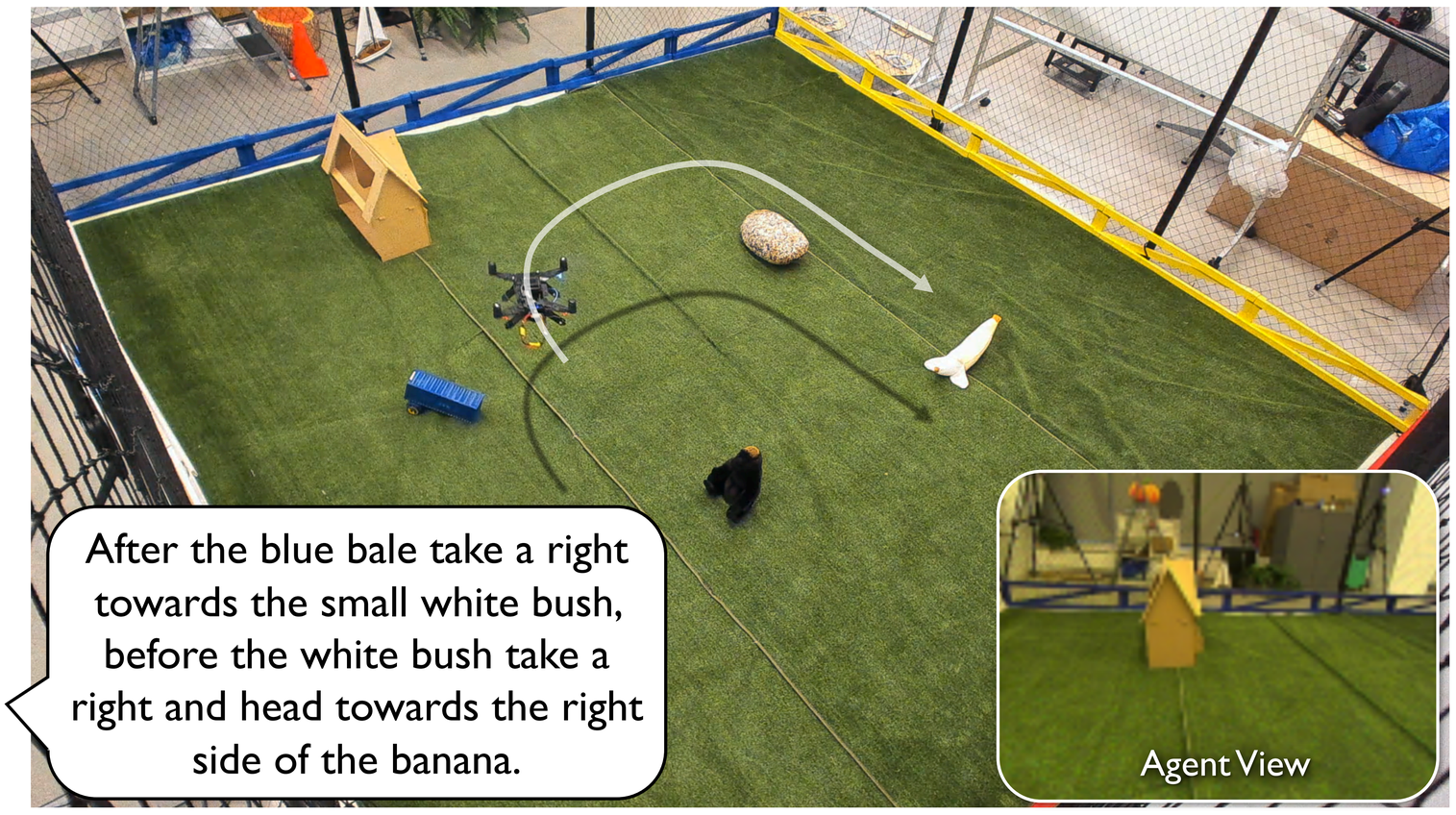
Learning to Map Natural Language Instructions to Physical Quadcopter Control using Simulated Flight

Valts Blukis, Yannick Terme,
Eyvind Niklasson, Ross A. Knepper,
Yoav Artzi

https://github.com/clic-lab/drif



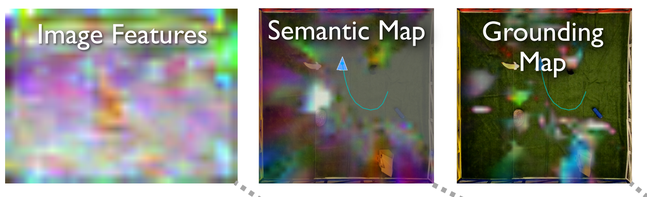
Task: Follow natural language navigation instructions on a physical quadcopter, assuming access only to first-person RGB images and pose estimates.

Challenges: Language understanding, grounding, perception, spatial reasoning, exploration and control.

Key Contributions:

- First demonstration of direct mapping of natural language and first-person observations to continuous robot control without manual representation design
- SuReAL algorithm (Supervised and Reinforcement Asynchronous Learning)
- Language-directed exploration by reducing P(goal unobserved)

Two-Stage Model (Position Visitation Network v2)

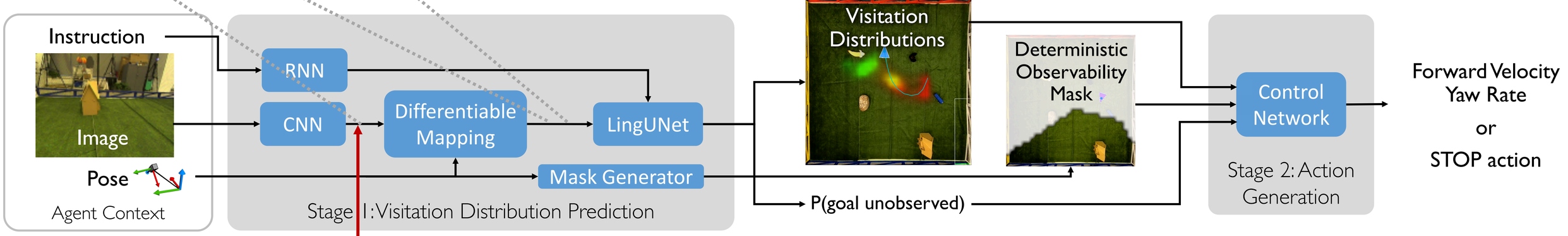


Stage 1 outputs a pair of 2D probability distributions over environment locations.

- Stop position distribution predicts where the goal is, or if the goal is not observed yet
- Position visitation distribution predicts which observed positions in the environment the agent should visit, or if it should visit positions that are not yet observed

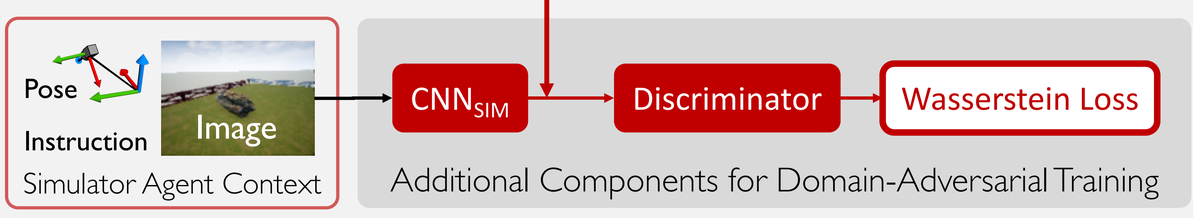
Stage 2 generates actions to:

1. Visit high-probability positions
2. Stop at a likely goal location
3. Explore the environment



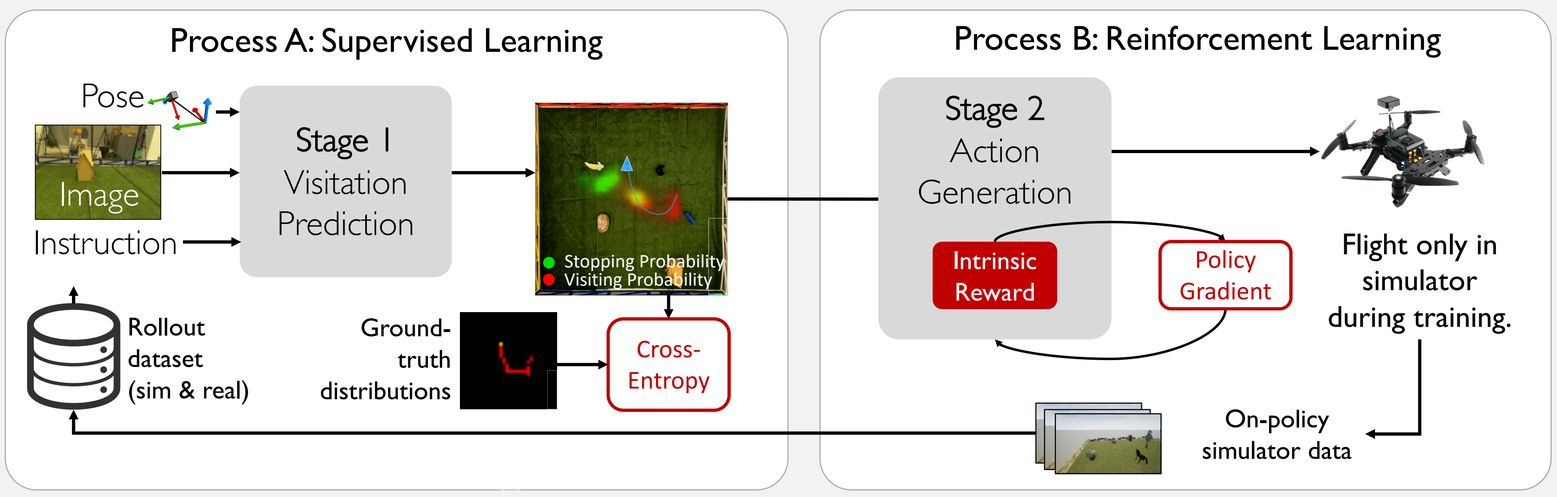
Joint Sim-to-Real Training with SuReAL

Adversarial Loss for Joint Sim and Real Training



- Extend PVNv2 model with a separate CNN feature extractor for simulation data
- Train discriminator network to distinguish features from CNN and CNN$_{SIM}$
- During training, alternate between simulator and real-world rollouts

SuReAL – Supervised and Reinforcement Asynchronous Learning



SuReAL: Simultaneously train Stage 1 with supervised learning and Stage 2 with reinforcement learning. Continuously add on-policy rollouts from RL to the data used to train Stage 1.

Advantages of SuReAL:

- More sample efficient than end-to-end reinforcement learning
- More robust to errors than end-to-end behavior cloning (supervised learning)
- Unlike behavior cloning, does not require action oracle, but requires only a visitation distribution oracle

Intrinsic Reward for Language-Directed Exploration with Partial Observability

Exploration Reward + **Trajectory Reward** + **Stopping Reward** + **Step Reward**

Predict actions to reduce P(goal unseen) and avoid stopping if P(goal unseen) is high

Fly through or near high probability positions according to predicted distribution.

Output the STOP action at or near a likely stopping position.

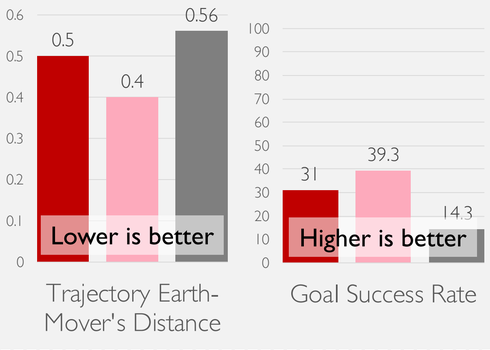
Negative per-step penalty to encourage efficiency.

Evaluation on Unseen Environments and Instructions

Automated Evaluation

Goal Success Rate:
How often did the agent stop within 47cm of the human demonstrated goal position.

Trajectory Earth-Mover's Distance:
Cost for morphing the agent trajectory to align with the human demonstration.



Human Evaluation (Mturk 5-point Likert-scale scores of agent behavior)

Goal score: How well the agent reached the correct goal. 5/5 points 40% of the time.

Path score: How well the agent followed the correct path. 5/5 points 38% of the time.