Are We There Yet? Learning to Localize in Embodied Instruction Following

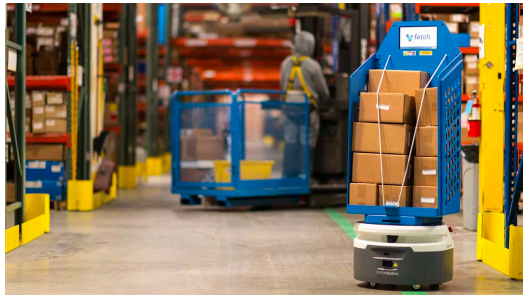
Shane Storks*^, Qiaozi Gao*, Govind Thattai*, & Gokhan Tur*

Hybrid AI @ AAAI 2021

*Amazon Alexa AI

^University of Michigan

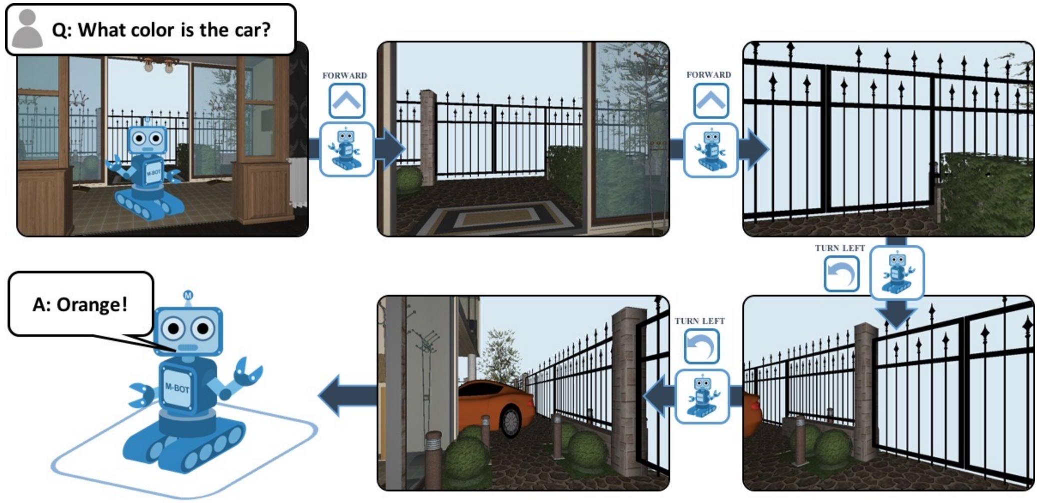
Concept Teaching



Motivation

- Mobile robots are being widely adopted for completing various pre-programmed and demonstrated tasks
- Embodied task learning: How can we teach a robot how to complete a new task using language?
  - Requires navigation and object manipulation in a physical space
  - Requires grounding language to visual inputs and primitive actions
  - Combines language, vision, and robotics
- How can we best harness the rich features in the environment, agent capabilities to guide navigation?

Related Work



https://embodiedqa.org/

- Language, vision, and robotics
  - **Embodied question answering (Das et al., 2018)**
  - Remote object grounding (Qi et al., 2020)
  - Robotic motion planning (Xia et al., 2020)
  - Vision-and-language navigation (Anderson et al., 2018)
  - Embodied task learning (Shridhar et al., 2019)
    - Action Learning From Realistic Environments and Directives (ALFRED)

Das, A. et al. (2018). Embodied Question Answering. In CVPR 2018.

Qi, Y. et al. (2020). REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In CVPR 2020.

Xia, F. et al. (2020). Interactive Gibson Benchmark: A Benchmark for Interactive Navigation in Cluttered Environments. In IEEE Robotics and Automation Letters 5(2): 713-720.

Anderson, P. et al. (2018). Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real envirnoments. In CVPR 2018.

Shridhar, M., et al. (2019). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In CVPR 2020.

Related Work

- Language, vision, and robotics
  - Embodied question answering (Das et al., 2018)
  - **Remote object grounding (Qi et al., 2020)**
  - Robotic motion planning (Xia et al., 2020)
  - Vision-and-language navigation (Anderson et al., 2018)
  - Embodied task learning (Shridhar et al., 2019)
    - Action Learning From Realistic Environments and Directives (ALFRED)

https://yuankaiqi.github.io/REVERIE_Challenge/challenge.html

Das, A. et al. (2018). Embodied Question Answering. In CVPR 2018.

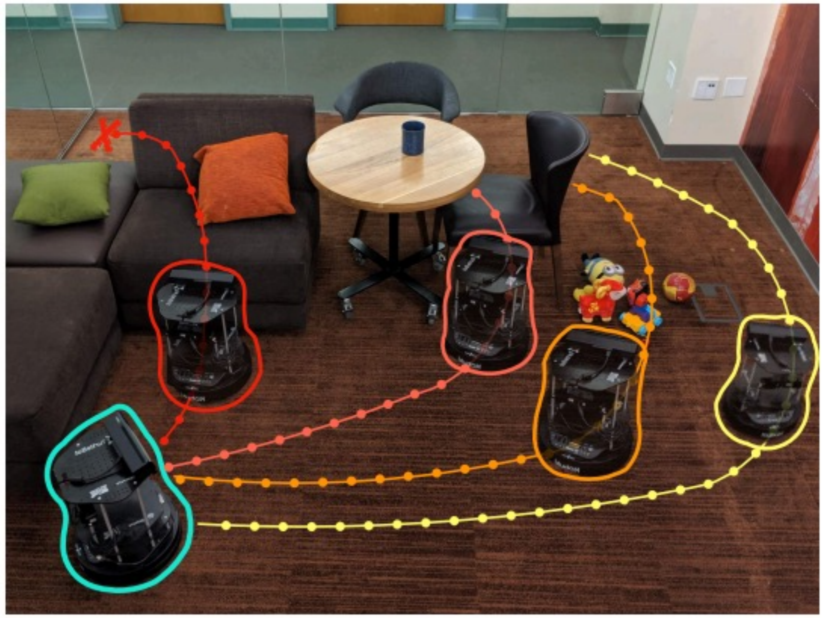
Qi, Y. et al. (2020). REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In CVPR 2020.

Xia, F. et al. (2020). Interactive Gibson Benchmark: A Benchmark for Interactive Navigation in Cluttered Environments. In IEEE Robotics and Automation Letters 5(2): 713-720.

Anderson, P. et al. (2018). Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real envirnoments. In CVPR 2018.

Shridhar, M., et al. (2019). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In CVPR 2020.

Related Work



https://arxiv.org/pdf/1910.14442.pdf

- Language, vision, and robotics
  - Embodied question answering (Das et al., 2018)
  - Remote object grounding (Qi et al., 2020)
  - **Robotic motion planning (Xia et al., 2020)**
  - Vision-and-language navigation (Anderson et al., 2018)
  - Embodied task learning (Shridhar et al., 2019)
    - Action Learning From Realistic Environments and Directives (ALFRED)

Das, A. et al. (2018). Embodied Question Answering. In CVPR 2018.

Qi, Y. et al. (2020). REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In CVPR 2020.

Xia, F. et al. (2020). Interactive Gibson Benchmark: A Benchmark for Interactive Navigation in Cluttered Environments. In IEEE Robotics and Automation Letters 5(2): 713-720.

Anderson, P. et al. (2018). Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real envirnoments. In CVPR 2018.

Shridhar, M., et al. (2019). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In CVPR 2020.

Related Work

- Language, vision, and robotics
  - Embodied question answering (Das et al., 2018)
  - Remote object grounding (Qi et al., 2020)
  - Robotic motion planning (Xia et al., 2020)
  - **Vision-and-language navigation (Anderson et al., 2018)**
  - Embodied task learning (Shridhar et al., 2020)
    - Action Learning From Realistic Environments and Directives (ALFRED)

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

https://bringmeaspoon.org/

Das, A. et al. (2018). Embodied Question Answering. In CVPR 2018.

Qi, Y. et al. (2020). REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In CVPR 2020.

Xia, F. et al. (2020). Interactive Gibson Benchmark: A Benchmark for Interactive Navigation in Cluttered Environments. In IEEE Robotics and Automation Letters 5(2): 713-720.

Anderson, P. et al. (2018). Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real envirnoments. In CVPR 2018.

Shridhar, M., et al. (2020). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In CVPR 2020.

Related Work

- Language, vision, and robotics
  - Embodied question answering (Das et al., 2018)
  - Remote object grounding (Qi et al., 2020)
  - Robotic motion planning (Xia et al., 2020)
  - Vision-and-language navigation (Anderson et al., 2018)
  - **Embodied task learning (Shridhar et al., 2019)**
    - Action Learning From Realistic Environments and Directives (ALFRED)

Das, A. et al. (2018). Embodied Question Answering. In CVPR 2018.

Qi, Y. et al. (2020). REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In CVPR 2020.

Xia, F. et al. (2020). Interactive Gibson Benchmark: A Benchmark for Interactive Navigation in Cluttered Environments. In IEEE Robotics and Automation Letters 5(2): 713-720.

Anderson, P. et al. (2018). Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real envirnoments. In CVPR 2018.

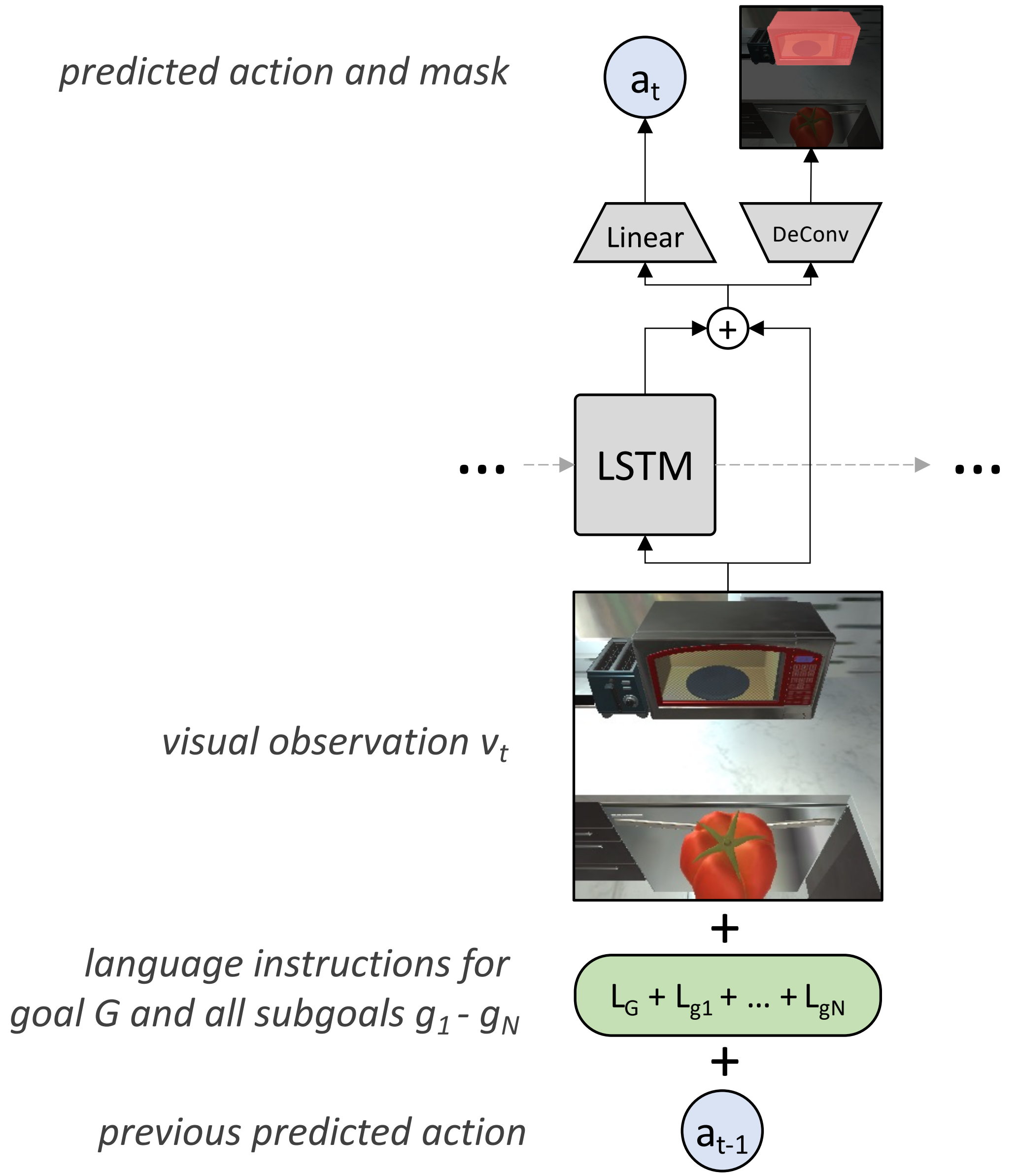
Shridhar, M., et al. (2019). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In CVPR 2020.

An instance of ALFRED consists of 3 units:

1. **Goal** G
2. **Subgoals** $g_1, g_2, \ldots, g_N$
   - Navigation
   - Object manipulation
     - Pick up object
     - Put down object
     - Clean object
     - …
3. **Actions** $a_1, a_2, \ldots, a_T$

Seq2Seq Baseline

- The baseline model uses task inputs at each timestep to predict a primitive action
  - And (if applicable) a mask over the current visual observation to indicate the object to interact with
  - (not pictured) language instructions are reweighted by an attention mechanism at every timestep



Evaluation Details

- Three granularities of inference and evaluation:
  - Goal-based
    - Can the agent achieve the goal G?
  - Subgoal-based
    - In isolation, can the agent achieve a single subgoal?
  - Action-based
    - How close is the predicted sequence of actions to the ground truth?
- Can evaluate in rooms seen during training, or rooms unseen in training
  - Validation seen and unseen partitions
- **ALFRED baseline:** 3.6% goal success rate in seen rooms, 0.4% goal success rate in unseen rooms ☹

Project Contributions

1. Granular training with ALFRED subgoals
2. Augmented navigation
    a. Full coverage of object segmentation masks
    b. Panoramic visual observations
3. Integrated object detection
4. Enabled spatial tracking in the model

Project Contributions

1. Granular training with ALFRED subgoals
2. Augmented navigation
    a. Full coverage of object segmentation masks
    b. Panoramic visual observations
3. Integrated object detection
4. Enabled spatial tracking in the model

Key Limitations

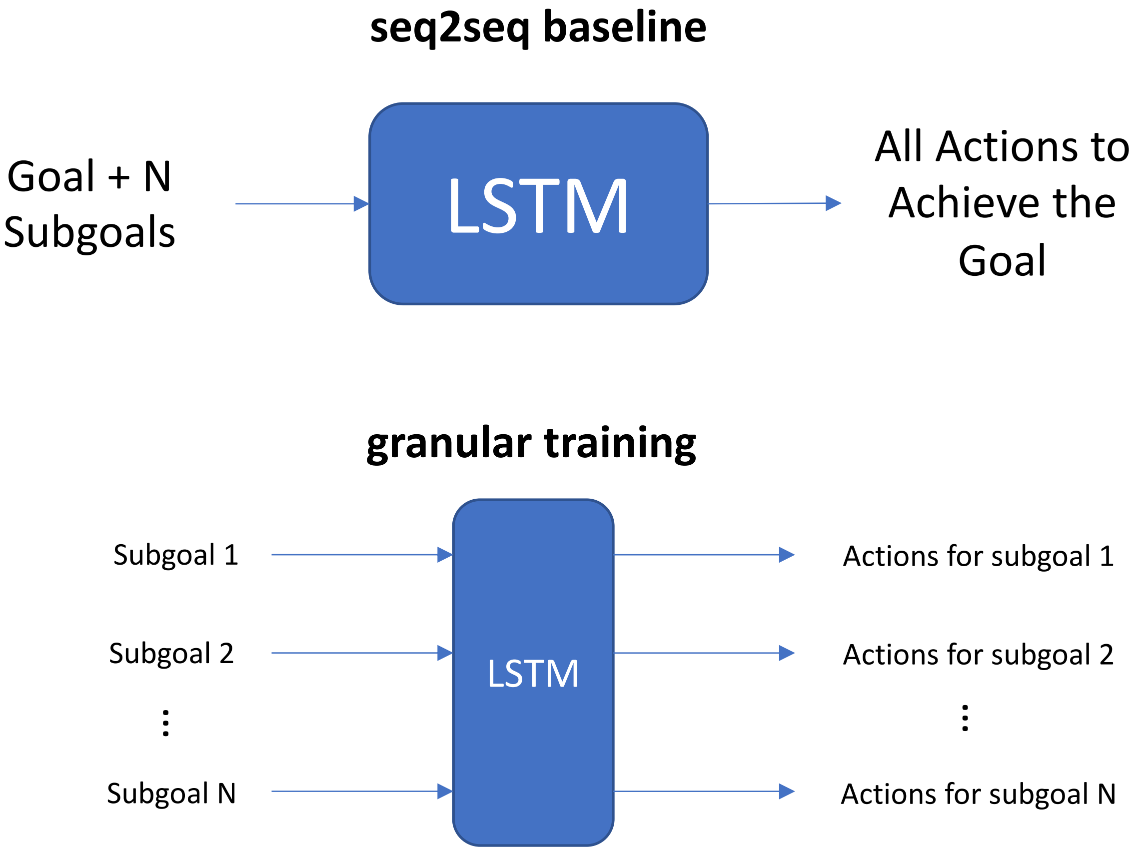
Sequence of actions is long. The model must predict a long sequence of actions from a long sequence of text.

“Rinse off a mug and place it in the coffee maker.”

- ”Walk to the coffee maker on the right.”
  - MoveAhead
  - MoveAhead
  - MoveAhead
  - MoveAhead
  - RotateRight
- “Pick up the dirty mug from the coffee maker.”
  - PickupObject(mug)
- “Turn and walk to the sink.
  - RotateLeft
  - RotateLeft
  - MoveAhead
  - MoveAhead
- “Wash the mug in the sink.”
  - PutObject(mug, sink)
  - ToggleObjectOn
  - ToggleObjectOff
  - PickupObject(mug)
- “Pick up the mug and go back to the coffee maker.”
  - RotateLeft
  - RotateLeft
  - MoveAhead
  - MoveAhead
- “Put the clean mug in the coffee maker.”
  - PutObject(mug, coffee maker)
  - STOP

Contribution 1: Granular Training

- **Solution:** break the problem down into subgoal completion



Model	Val. Seen Action F1 (%)	Avg. Subgoal Success Rate (%)
ALFRED Baseline	84.5	25.8
Granular Training	91.6	32.2

Project Contributions

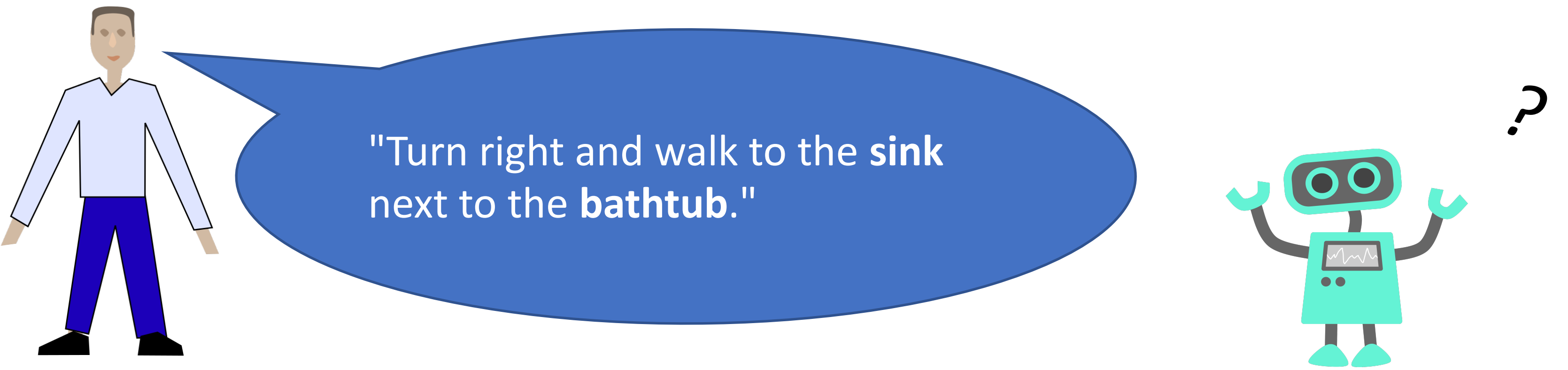
1. Granular training with ALFRED subgoals
2. Augmented navigation
    a. Full coverage of object segmentation masks
    b. Panoramic visual observations
3. Integrated object detection
4. Enabled spatial tracking in the model

Key Limitations

Navigation performance is a bottleneck for overall performance.

Success rate on navigation subgoals is low relative to some other subgoal types. Why?

a) *The agent is not explicitly trained to ground language during navigation.*
b) *The agent doesn't learn to explore.*

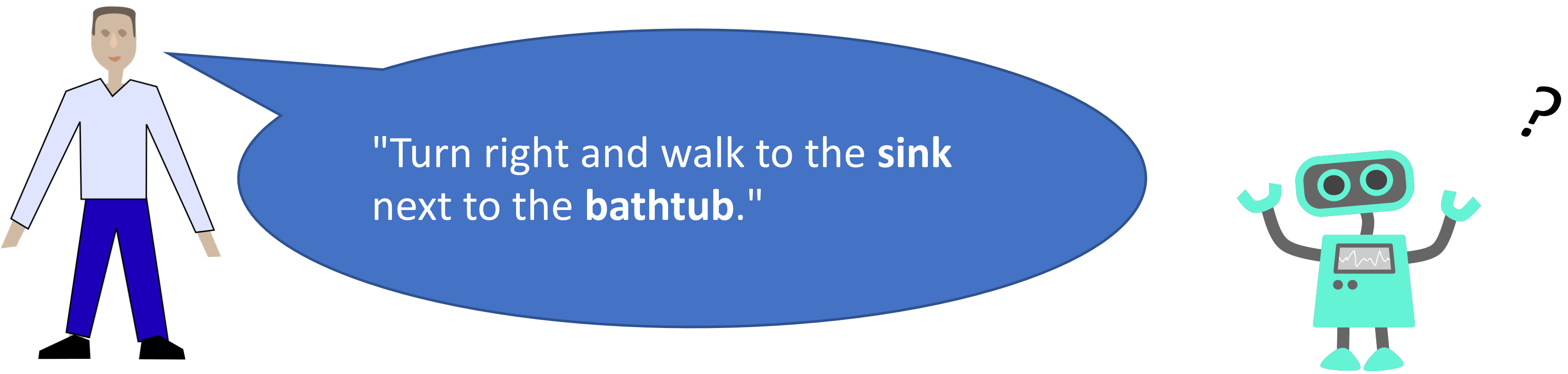


Key Limitations

Navigation performance is a bottleneck for overall performance.

Success rate on navigation subgoals is low relative to some other subgoal types. Why?

a) *The agent is not explicitly trained to ground language during navigation.*
b) *The agent doesn't learn to explore.*

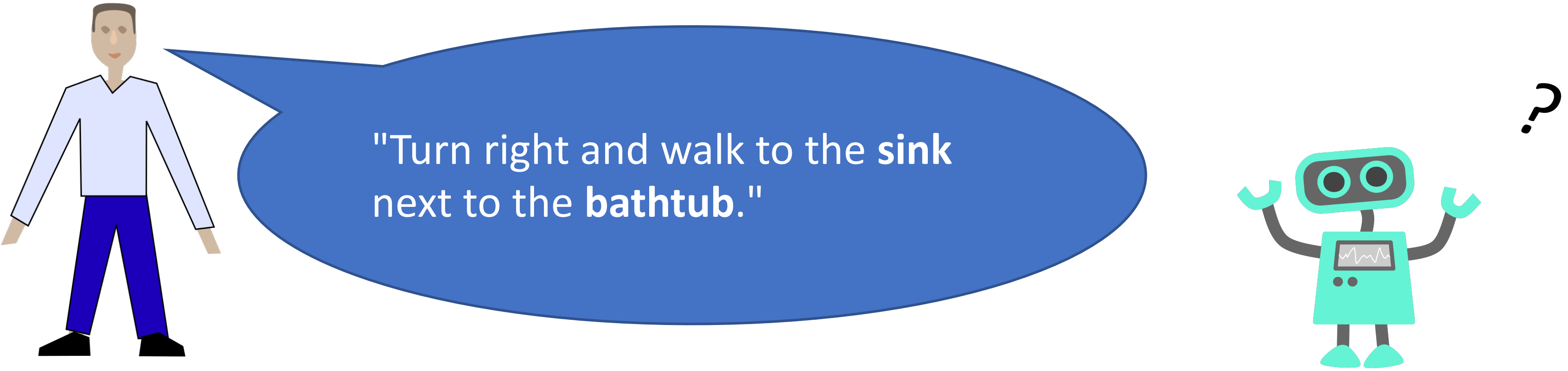


Key Limitations

Navigation performance is a bottleneck for overall performance.

Success rate on navigation subgoals is low relative to some other subgoal types. Why?

a) *The agent is not explicitly trained to ground language during navigation.*
b) *The agent doesn't learn to explore.*

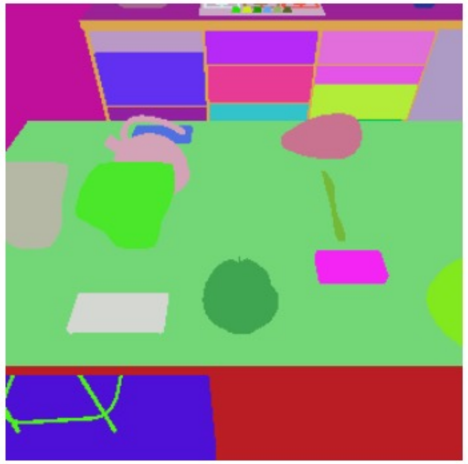


Project Contributions

1. Granular training with ALFRED subgoals
2. Augmented navigation
    a. Full coverage of object segmentation masks
    b. Panoramic visual observations
3. Integrated object detection
4. Enabled spatial tracking in the model

Additional Masks

- Base dataset only includes segmentation masks for objects that the agent must manipulate
- Collect masks for every visible object at every timestep

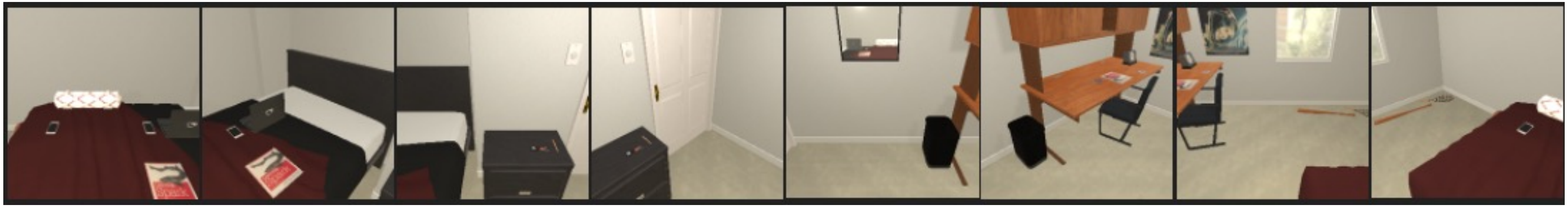


Project Contributions

1. Granular training with ALFRED subgoals
2. Augmented navigation
    a. Full coverage of object segmentation masks
    b. Panoramic visual observations
3. Integrated object detection
4. Enabled spatial tracking in the model

Panoramic Image Observations

- Performance gains have come in vision-and-language navigation (VLN) from using panoramic visual inputs
  - Fried et al. (2018). Speaker-Follower Models for Vision-and-Language Navigation.
- Training: we collect images at 8 view angles for every timestep of navigation
  - Built-in exploratory behavior
- Inference: force the agent to ”look around” 360 degrees before taking each step during navigation
  - At a cost of extra predicted actions

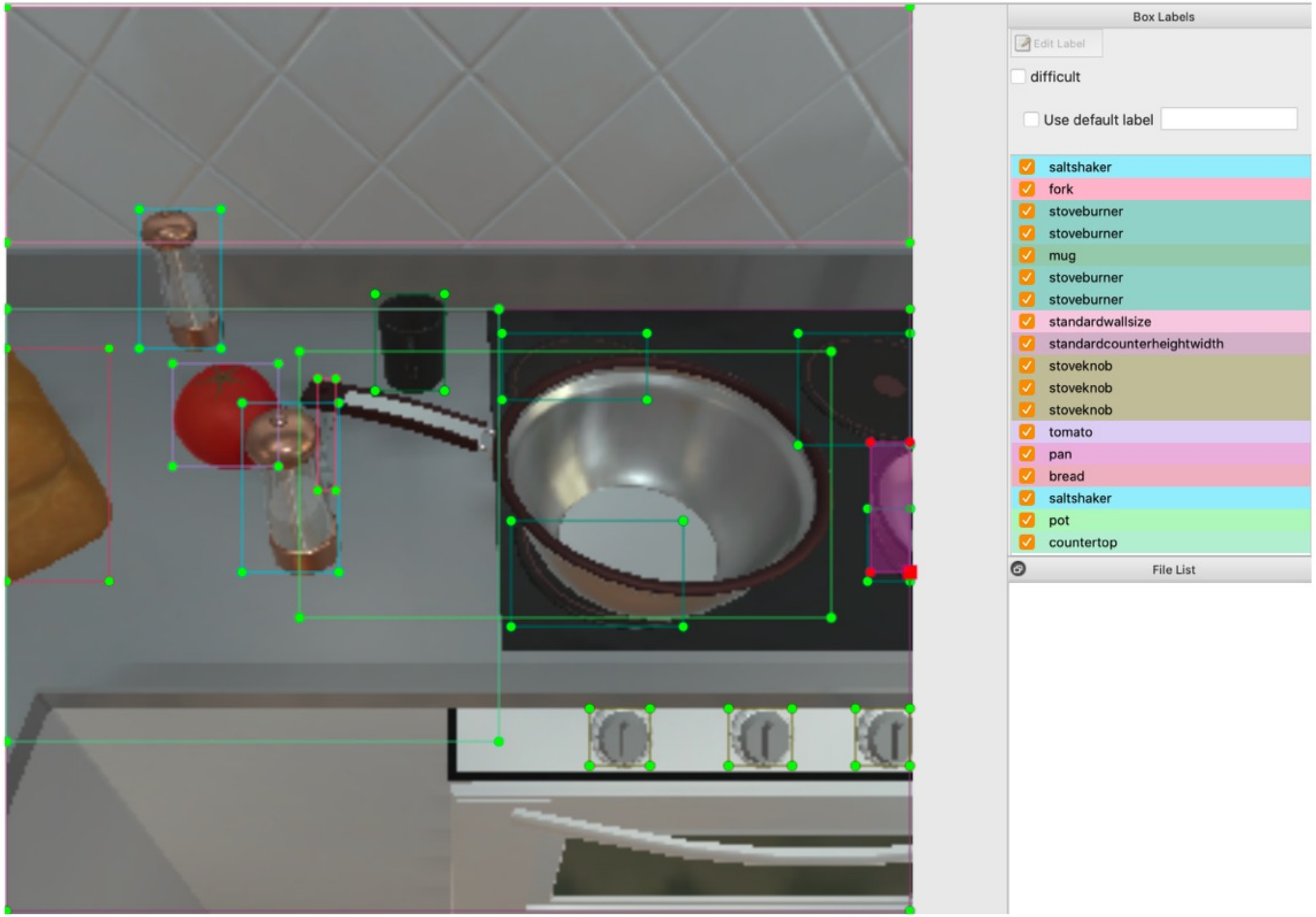


Project Contributions

1. Granular training with ALFRED subgoals
2. Augmented navigation
   a. Full coverage of object segmentation masks
   b. Panoramic visual observations
3. Integrated object detection
4. Enabled spatial tracking in the model

Introducing Object Detection

- Using newly generated masks, train an object detection model
  - [Bochkovski, A. et al. (2018). YOLOv4: Optimal Speed and Accuracy of Object Detection.]
- If we add this to the pipeline, agent can explicitly identify any object it sees
  - (even in panoramic observations)

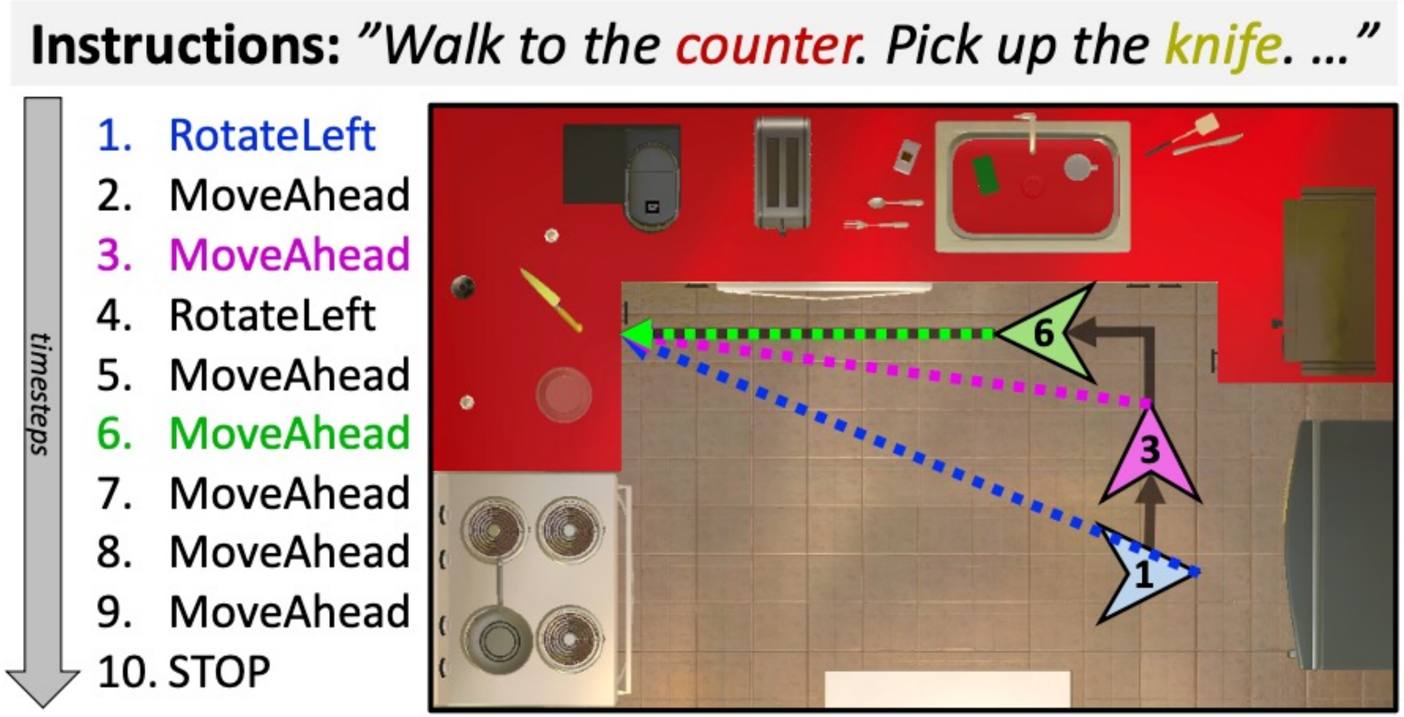


Project Contributions

1. Granular training with ALFRED subgoals
2. Augmented navigation
   a. Full coverage of object segmentation masks
   b. Panoramic visual observations
3. Integrated object detection
4. Enabled spatial tracking in the model

Oracle Angle Tracking

- In the granular trained model, the agent loses the ability to look ahead in the instructions.
  - When navigating to *the counter*, the agent doesn't know that it will need *a knife* during the next subgoal
- During navigation, enable the agent to track the relative location of the precise navigation goal
  - Angle to the goal location

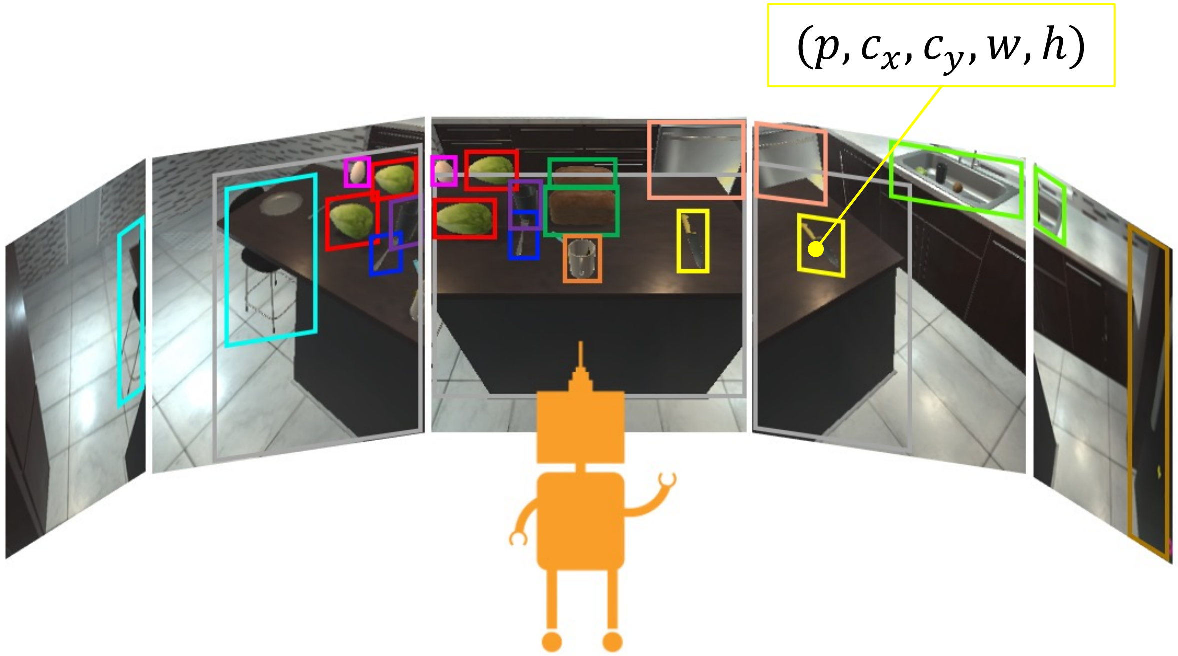


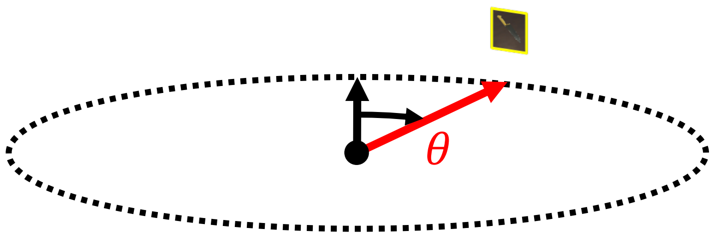
Model	Navigation Subgoal Success Rate (%)	Goal Condition Success Rate (%)
ALFRED Baseline	31.0	1.6
Granular Training	30.0	1.3
Granular Training + Oracle Angle Tracking	**67.8**	**2.8**

Predicting the Angle

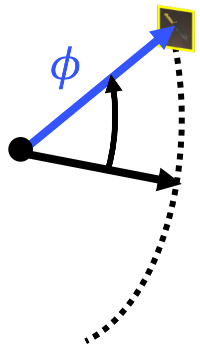
- How can we predict this angle to achieve such high performance fairly?
- We combine all the work so far into a *localizer* module:
  - Inputs at each timestep:
    - Panoramic bounding box information (coordinates and labels)
    - Current and next subgoal language instructions
  - Output:
    - Angle d_t to goal (sine and cosine)

Projecting Bounding Boxes to 3D Space





horizontal angle θ

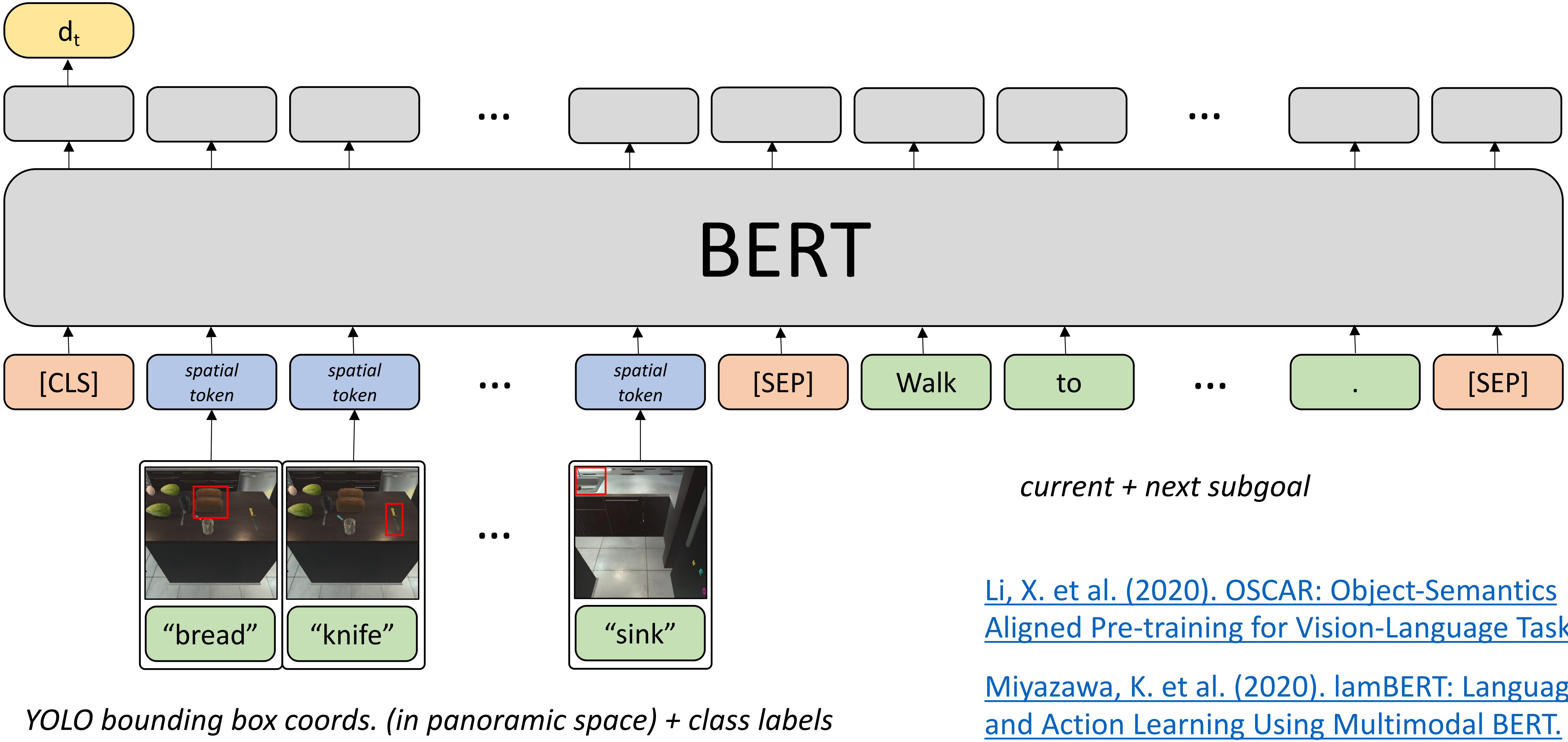


vertical angle ϕ

$$\theta = \tan^{-1}\left[2(c_x - 0.5)\tan\frac{F_x}{2}\right] + 45p \qquad \phi = \tan^{-1}\left[2(0.5 - c_y)\tan\frac{F_y}{2}\right] + \delta$$

$$(\sin\theta, \cos\theta, \sin\phi, w, h)$$

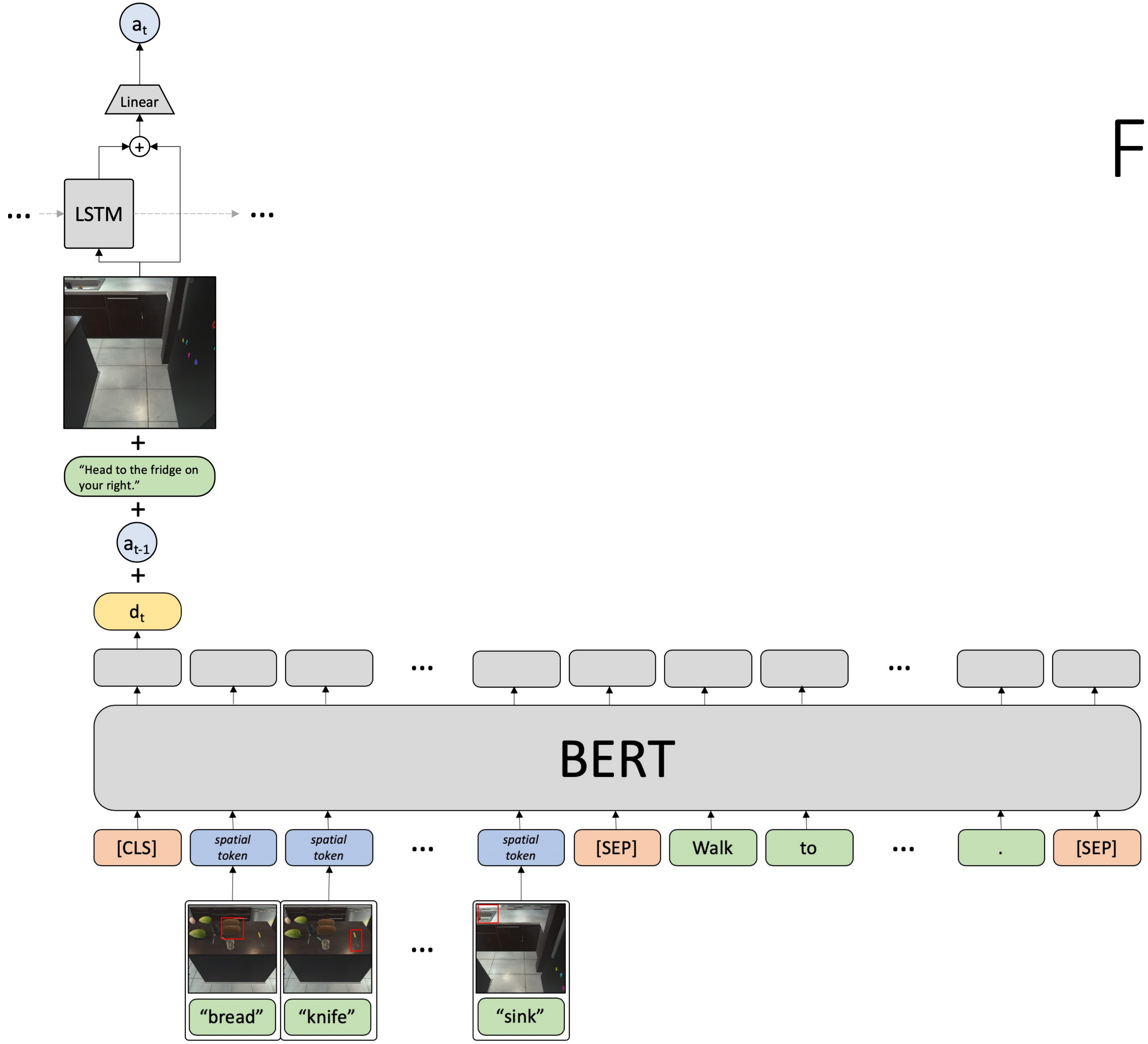
Transformer-based Angular Prediction



[Li, X. et al. (2020). OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks.]

[Miyazawa, K. et al. (2020). lamBERT: Language and Action Learning Using Multimodal BERT.]

Full Structure



(all results on stack & place task type)

Model	Action F1 (%)		Navigation Subgoal Success Rate (%)		Goal Condition Success Rate (%)	
	Val. Seen	*Val. Unseen*	*Val. Seen*	*Val. Unseen*	*Val. Seen*	*Val. Unseen*
Baseline	84.5	75.6	**31.0**	27.5	**1.6**	0.0
Granular Training	91.6	85.3	30.0	26.5	1.3	0.0
Granular Training + Oracle Goal Angle	<u>93.9</u>	86.9	<u>67.8</u>	<u>35.4</u>	<u>2.8</u>	0.0
Granular Training + BERT-Based Localizer	**93.8**	**88.7**	25.4	**28.8**	1.4	0.0

best non-oracle result <u>best overall result</u>

Summary

1. Granular training with subgoals improved performance of action prediction
2. Augmented inputs combined with object detection gave the agent new capabilities during navigation
    1. ”Looking around”
    2. Identifying objects explicitly
3. Used capabilities to enable spatial tracking in the model and improve action prediction, navigation performance

Questions?

Thank you!