Vision-and-Dialog Navigation

Jesse Thomason
University of Washington

Daniel Gordon
Yonatan Bisk
Michael Murray
Maya Cakmak
Luke Zettlemoyer
Bringing Robots from Industrial to Human Spaces
Bringing Robots from Industrial to Human Spaces
Bringing Robots from Industrial to Human Spaces

Natural Language → Navigation → Additional Safety → Robust Perception → Human

Industrial

...
Bringing Robots from Industrial to Human Spaces

Natural Language

Navigation

Additional Safety

Robust Perception
1) In Human Spaces, We Use Natural Language.
1) In Human Spaces, We Use Natural Language.
2) Human Spaces Are *Dynamic* and can be *Unseen*
2) Human Spaces Are *Dynamic* and can be *Unseen*
2) Human Spaces Are *Dynamic* and can be *Unseen*
Outline
Outline

- Language grounding in visual environments
Outline

- Language grounding in visual environments
  - For navigation
Outline

- Language grounding in visual environments
  - For navigation
  - Unimodal bias [Thomason et al., NAACL’19]
Outline

- Language grounding in visual environments
  - For navigation
  - Unimodal bias [Thomason et al., NAACL’19]
- Vision-and-Dialog Navigation [Thomason et al., in sub]
Outline

- Language grounding in visual environments
  - For navigation
  - Unimodal bias [Thomason et al., NAACL’19]
- Vision-and-Dialog Navigation [Thomason et al., in sub]
  - New dataset - CVDN
Outline

- Language grounding in visual environments
  - For navigation
  - Unimodal bias [Thomason et al., NAACL’19]
- Vision-and-Dialog Navigation [Thomason et al., in sub]
  - New dataset - CVDN
  - Navigation from dialog history
Outline

- Language grounding in visual environments
  - For navigation
  - Unimodal bias [Thomason et al., NAACL’19]
- Vision-and-Dialog Navigation [Thomason et al., in sub]
  - New dataset - CVDN
  - Navigation from dialog history
- Next steps
Connecting Language and Vision
Connecting Language and Vision

- Common paradigm:
Connecting Language and Vision

- Common paradigm:
  - **Inputs:**
    - Language tokens (e.g., question)
Connecting Language and Vision

- Common paradigm:
  - Inputs:
    - Language tokens (e.g., question)
    - Visual context (e.g., photograph)
Connecting Language and Vision

- Common paradigm:
  - **Inputs:**
    - Language tokens (e.g., question)
    - Visual context (e.g., photograph)
- Visual contexts differ in quality across datasets.
Connecting Language and Vision

● Common paradigm:
  ○ **Inputs:**
    ■ Language tokens (e.g., question)
    ■ Visual context (e.g., photograph)
● Visual contexts differ in quality across datasets.
● Output can be a single classification or a sequence.
Visual Fidelity

CLEVR
[Johnson et al., CVPR’17]

Visual Context

Static    Dynamic

Photorealistic

Rendered
Visual Context

CLEVR
[Johnson et al., CVPR’17]

VQA
[Antol et al., CVPR’15]
Outline

● Language grounding in visual environments
  ○ For navigation
  ○ Unimodal bias [Thomason et al., NAACL’19]
● Vision-and-Dialog Navigation [Thomason et al., in sub]
  ○ New dataset - CVDN
  ○ Navigation from dialog history
● Next steps
Vision-and-Language Navigation

Exit the bathroom. Turn left and exit the room using the door on the left. Wait there.
“Turn around and exit the library, head down the…”

- Low-level instructions for moving through the environment.
“Turn around and exit the library, head down the…”

- Low-level instructions for moving through the environment.
Sequence-to-Sequence Model

- Encode the language tokens.
- Decode a sequence of actions to take in the environment.
- At every timestep, receive a new visual observation.
Sequence-to-Sequence Model

- Encode the language tokens.
- Decode a sequence of actions to take in the environment.
- At every timestep, receive a new visual observation.

Learned, token-level Language Embedding
Sequence-to-Sequence Model

- Encode the language tokens.
- Decode a sequence of actions to take in the environment.
- At every timestep, receive a new visual observation.

[Anderson et al., CVPR'18]
Sequence-to-Sequence Model

- Encode the language tokens.
- Decode a sequence of actions to take in the environment.
- At every timestep, receive a new visual observation.

[Anderson et al., CVPR’18]
Sequence-to-Sequence Model

- Encode the language tokens.
- Decode a sequence of actions to take in the environment.
- At every timestep, receive a new visual observation.
Sequence-to-Sequence Model

- Encode the language sequence of observations.
- Decode a sequence of actions, one in the environment.
- At every timestep, receive a new visual observation.

We will build on this model.
Sequence-to-Sequence Model

- Train to predict the action a shortest-path planner would take from the current state.
Sequence-to-Sequence Model

- Train to predict the action a shortest-path planner would take from the current state.
Sequence-to-Sequence Model

- Train to predict the action a shortest-path planner would take from the current state.
Language Grounding for QA and Navigation

Input:
- Instruction + Frame
  - Room-2-Room [Anderson et al., CVPR’18]
- Question + Frame
  - Embodied QA [Das et al., CVPR’18]
- Question + Frame
  - Interactive QA [Gordon et al., CVPR’18]

Output:
- Navigation Actions
- Navigation Actions + Answer Action
- Navigation Actions + Answer Action
Language Grounding for QA and Navigation

Input:
- Instruction + Frame
  - Room-2-Room
    [Anderson et al., CVPR’18]
- Question + Frame
  - Embodied QA
    [Das et al., CVPR’18]
- Question + Frame
  - Interactive QA
    [Gordon et al., CVPR’18]

Output:
- Navigation Actions
- Navigation Actions + Answer Action
- Navigation Actions + Answer Action
Outline

- Language grounding in visual environments
  - For navigation
    - Unimodal bias [Thomason et al., NAACL’19]
- Vision-and-Dialog Navigation [Thomason et al., in sub]
  - New dataset - CVDN
  - Navigation from dialog history
- Next steps
Inputs and Outputs

Language: “Walk past the bar and turn right.”

Actions: Forward, turn Left & Right, tilt Up & Down, End
Inputs and Outputs

Language: “Walk past the bar and turn right.”

Actions: Forward, turn Left & Right, tilt Up & Down, End
Inputs and Outputs

Language: “Walk past the bar and turn right.”

Actions: Forward, turn Left & Right, tilt Up & Down, End
Unimodal Ablation - Vision Only

Language:

Actions: Forward, turn Left & Right, tilt Up & Down, End
Unimodal Ablation - Language Only

Language: “Walk past the bar and turn right.”

Actions: Forward, turn Left & Right, tilt Up & Down, End
Unimodal Ablation - Action Only

**Language:**

- $t_1$
- $t_2$
- $t_3$
- $t_4$

**Actions:** Forward, turn Left & Right, tilt Up & Down, End
Unimodal Model Ablations
Unimodal Ablation - Action Only

Language:

Actions: Forward, turn Left & Right, tilt Up & Down, End
Unimodal Model Ablations
Unimodal Model Ablations

Room-2-Room
[Anderson et al., CVPR’18]

Beats baseline Vision-, language-, and action-only models.
Unimodal Model Ablations

Room-2-Room
[Anderson et al., CVPR’18]

Embodied QA
[Das et al., CVPR’18]

Beats baseline

Vision-, language-, and action-only models.

Vision-, language-, and action-only models.
Unimodal Model Ablations

Room-2-Room
[Anderson et al., CVPR’18]

Embodied QA
[Das et al., CVPR’18]

Interactive QA
[Gordon et al., CVPR’18]

Beats baseline

Vision-, language-, and action-only models.
Vision-, language-, and action-only models.
Vision-only model.
Unimodal Model Ablations

Room-2-Room [Anderson et al., CVPR’18]

Embodyed QA [Das et al., CVPR’18]

Interactive QA [Gordon et al., CVPR’18]

Beats baseline Vision-, language-, and action-only models.

Beats initial model! Language-only model.
Unimodal Model Ablations

Room-2-Room
[Anderson et al., CVPR’18]

Embodied QA
[Das et al., CVPR’18]

Interactive QA
[Gordon et al., CVPR’18]

Beats baseline
Vision-, language-, and action-only models.

Vision-, language-, and action-only models.

Vision-only model.

Beats initial model!
Language-only model.

Vision-only model.
Unimodal Model Ablations

- Low-level Language
- Embodied QA
  [Das et al., CVPR’18]
- Interactive QA
  [Gordon et al., CVPR’18]

Beats baseline
Vision-, language-, and action-only models.

Beats initial model!
Language-only model.

Vision-only model.

Vision-only model.

(none)
Unimodal Model Ablations

Photorealistic Environments
[Anderson et al., ICLR’18]

Low-level Language

Vision-, language-, and action-only models.

Embodied QA
[Das et al., CVPR’18]

Vision-, language-, and action-only models.

Interactive QA
[Gordon et al., CVPR’18]

Vision-only model.

Beats baseline

Language-only model.

Beats initial model!
Unimodal Model Ablations

Photorealistic Environments

Low-level Language

Simple Visual Environments

Interactive QA
[Anderson et al., CVPR'18]

[Das et al., CVPR'18]

Vision-, language- and action-only models.

Vision-, language- and action-only models.

Vision-only model.

Language-only model.

Vision-only model.

(none)
Unimodal Model Ablations

Photorealistic Environments
[Anderson et al., NeurIPS’18]

High-level Language
[Das et al., ICLR’19]

Interactive QA
[Gordon et al., CVPR’18]

Beats baseline
Vision-, language- and action-only models.

Beats initial model!
Language-only model.

Vision-only model.

Simple Visual Environments
(none)
Unimodal Model Ablations

- Photorealistic Environments
- High-level Language
- Well-balanced
  - Low-level Language
  - Simple Visual Environments

- Beats baseline
  - Vision-, language-, and action-only models.
- Beats initial model!
  - Language-only model.
  - Vision-only model.
  - (none)
Lessons
Lessons

- Unimodal baselines expose dataset bias.
Lessons

- Unimodal baselines expose dataset bias.
  - More appropriate than non-learning baselines.
Lessons

- Unimodal baselines expose dataset bias.
  - More appropriate than non-learning baselines.
- Rich visual context prevents visual overfitting.
Lessons

- Unimodal baselines expose dataset bias.
  - More appropriate than non-learning baselines.
- Rich visual context prevents visual overfitting.
- *Underspecified* language context prevents language overfitting.
Lessons

- Unimodal baselines expose dataset bias.
  - More appropriate than non-learning baselines.
- Rich visual context prevents visual overfitting.
- *Underspecified* language context prevents language overfitting.
  - Also, low-level instructions are somewhat unnatural.
Lessons

● Unimodal baselines expose dataset bias.
  ○ More appropriate than non-learning baselines.
● Rich visual context prevents visual overfitting.
● *Underspecified* language context prevents language overfitting.
  ○ Also, low-level instructions are somewhat unnatural.
● Why not both?
Outline

- Language grounding in visual environments
  - For navigation
  - Unimodal bias [Thomason et al., NAACL’19]
- Vision-and-Dialog Navigation [Thomason et al., in sub]
  - New dataset - CVDN
  - Navigation from dialog history
- Next steps
“Turn around and exit the library, head down the…”
“Turn around and exit the library, head down the…”

- *Room-to-Room* uses low-level language.
“Turn around and exit the library, head down the…”

- *Room-to-Room* uses low-level language.
- 24% → 80% of human performance since ‘18.
“Go to the room with a plant.”

- This instruction is *underspecified*.
“Go to the room with a plant.”

- This instruction is *underspecified*.
- This instruction is *ambiguous*. 
“Go to the room with a plant.”

- This instruction is *underspecified*.
- This instruction is *ambiguous*.
What if we could just... ask?

- “Should I continue into the living room or go right towards the kitchen?”
Visual Context

CLEVR
[Johnson et al., CVPR’17]

Static

Instruction Following
[Chen and Mooney, AAAI’11]

Dynamic

VQA
[Antol et al., CVPR’15]

Photorealistic

Room-to-Room
[Anderson et al., CVPR’18]
A **cylinder** is next to a **yellow object**.

Q1: What shape is the object?
A1: Sphere

Q2: And material?
A2: Metal

Q3: What about that cylinder?
A3: Rubber

---

**Visual Dialog**

Q: What is the gender of the one in the white shirt?
A: She is a woman

Q: What is she doing?
A: Playing a Wii game

Q: Is that a man to her right?
A: No, it's a woman
Language Source

Guidance
Semantic

Abstraction
Visual

Templates
Humans
Talk the Walk
[de Vries et al., arXiv’18]
Vision-and-Dialog Navigation

- Human-human dialogs
- Both participants get an egocentric scene view.

Talk the Walk
[de Vries et al., arXiv’18]
Outline

- Language grounding in visual environments
  - For navigation
  - Unimodal bias [Thomason et al., NAACL’19]
- Vision-and-Dialog Navigation [Thomason et al., in sub]
  - New dataset
    - Cooperative Vision-and-Dialog Navigation (CVDN)
    - 2k human-human dialogs
  - Navigation from dialog history
- Next steps
Outline

- Language grounding in visual environments
  - For navigation
  - Unimodal [Gu et al., CVPR’19]
- Vision-and-Dialog Navigation [Gu et al., in sub]
  - New dataset
    - Cooperative Vision-and-Dialog Navigation (CVDN)
    - 2k human-human dialogs
  - Navigation from dialog history
- Next steps
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat*.  

-- this target object is present in at least two rooms, but only one is correct.
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat.*

-- this target object is present in at least two rooms, but only one is correct.
Visible to both Navigator and Oracle

Hint: The goal room contains a *mat*.

-- this target object is present in at least two rooms, but only one is correct.

Into the hall or the office?
Visible to both **Navigator** and **Oracle**

*Hint: The goal room contains a *mat.**  

--- this target object is present in at least two rooms, but only one is correct.

**Visible Only to the **Oracle**  

--- A shortest-path planner’s next steps; up to 5 navigation nodes in the direction of the goal.
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat*.

-- this target object is present in at least two rooms, but only one is correct.

**Visible Only to the Oracle**

Left into the hall. Follow it to a living room.

-- A shortest-path planner’s next steps; up to 5 navigation nodes in the direction of the goal.
Visible to both **Navigator** and **Oracle**

Hint: The goal room contains a *mat*.

-- this target object is present in at least two rooms, but only one is correct.

**Visible Only to the Oracle**

Into the hall or the office?

Left into the hall. Follow it to a living room.

-- A shortest-path planner’s next steps; up to 5 navigation nodes in the direction of the goal.
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat*.

-- this target object is present in at least two rooms, but only one is correct.

**Visible Only to the **Oracle**

--- A shortest-path planner’s next steps; up to 5 navigation nodes in the direction of the goal.

**Into the hall or the office?**

**Left into the hall. Follow it to a living room.**

**Should I go upstairs?**
ON AVERAGE, PATHS IN VDN ARE OVER THREE TIMES LONGER THAN R2R PATHS.
Dialog Leads to Long Paths and Rich Language

![Graph showing the distribution of path lengths for Human and Planner paths per dialog. The x-axis represents Path Length, and the y-axis represents Proportional Frequency. The graph shows that Human paths tend to be longer and have a higher frequency of occurrence compared to Planner paths.]

34
Dialog Leads to Long Paths and Rich Language

- Path Length Average:
  - Human (25.0); Planner (17.4)
  - Room-to-Room (6.0)
Dialog Leads to Long Paths and Rich Language

- Path Length Average:
  - Human (25.0); Planner (17.4)
  - Room-to-Room (6.0)
Dialog Leads to Long Paths and Rich Language

- Path Length Average:
  - Human (25.0); Planner (17.4)
  - Room-to-Room (6.0)

- Average total words:
  - CVDN (82)
  - Room-to-Room (29)
Dialog Leads to Long Paths and Rich Language

![Bar chart showing the proportional frequency of number of utterances per dialog. The x-axis represents the number of utterances per dialog ranging from 0 to 26, and the y-axis represents the proportional frequency ranging from 0.00 to 0.25. The chart shows a decreasing trend as the number of utterances increases.]
# Dialog Leads to Long Paths and Rich Language

<table>
<thead>
<tr>
<th>Number of Utterances per Dialog</th>
<th>Proportional Frequency</th>
<th>Dialog 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00</td>
<td>3 steps</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>4 steps</td>
</tr>
<tr>
<td>6</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.20</td>
<td>7 steps</td>
</tr>
<tr>
<td>10</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.10</td>
<td>6 steps</td>
</tr>
<tr>
<td>14</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

**The goal room contains a rug.**

**Navigator:** Should I go to the left or right?

**Oracle:** Go left and turn right after the bathroom.

**Navigator:** Do I need to go in the room with the run or keep on going right?

**Oracle:** Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.

**Navigator:** Should I go into the kitchen or to the right?

**Oracle:** Turn toward the front door and go up the stairs you see on the right.

**Navigator:** Do I go left or right?

**Oracle:** Go along the railing to the right. Stop at the room with a brown chair.
Dialog Leads to Long Paths and Rich Language

The goal room contains a rug.

**Navigator:** Should I go to the left or right?

**Oracle:** Go left and turn right after the bathroom.

**Navigator:** Do I need to go in the room with the run or keep on going right?

**Oracle:** Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.

**Navigator:** Should I go into the kitchen or to the right?

**Oracle:** Turn toward the front door and go up the stairs you see on the right.

**Navigator:** Do I go left or right?

**Oracle:** Go along the railing to the right. Stop at the room with a brown chair.
Outline

- Language grounding in visual environments
  - For navigation
  - Unimodal bias [Thomason et al., NAACL’19]
- Vision-and-Dialog Navigation [Thomason et al., in sub]
  - New dataset - CVDN
  - Navigation from dialog history
    - 7k dialog-based navigation inputs
- Next steps
Navigation from Dialog History

The goal room contains a rug.

**Navigator:** Should I go to the left or right?

**Oracle:** Go left and turn right after the bathroom.

**Navigator:** Do I need to go in the room with the run or keep on going right?

**Oracle:** Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.

**Navigator:** Should I go into the kitchen or to the right?

**Oracle:** Turn toward the front door and go up the stairs you see on the right.

**Navigator:** Do I go left or right?

**Oracle:** Go along the railing to the right. Stop at the room with a brown chair.
Navigation from Dialog History

The goal room contains a rug.

**Navigator:** Should I go to the left or right?

**Oracle:** Go left and turn right after the bathroom.

**Navigator:** Do I need to go in the room with the run or keep on going right?

**Oracle:** Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.
# Navigation from Dialog History

<table>
<thead>
<tr>
<th>The goal room contains a rug.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Navigator</strong>: Should I go to the left or right?</td>
</tr>
<tr>
<td><strong>Oracle</strong>: Go left and turn right after the bathroom.</td>
</tr>
<tr>
<td><strong>Navigator</strong>: Do I need to go in the room with the run or keep on going right?</td>
</tr>
<tr>
<td><strong>Oracle</strong>: Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.</td>
</tr>
<tr>
<td><strong>Navigator</strong>: Should I go into the kitchen or to the right?</td>
</tr>
<tr>
<td><strong>Oracle</strong>: Turn toward the front door and go up the stairs you see on the right.</td>
</tr>
<tr>
<td><strong>Navigator</strong>: Do I go left or right?</td>
</tr>
<tr>
<td><strong>Oracle</strong>: Go along the railing to the right. Stop at the room with a brown chair.</td>
</tr>
</tbody>
</table>
Navigation from Dialog History

The goal room contains a rug.

**Navigator:** Should I go to the left or right?

**Oracle:** Go left and turn right after the bathroom.

**Navigator:** Do I need to go in the room with the run or keep on going right?

**Oracle:** Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.

**Navigator:** Should I go into the kitchen or to the right?

**Oracle:** Turn toward the front door and go up the stairs you see on the right.

**Navigator:** Do I go left or right?

**Oracle:** Go along the railing to the right. Stop at the room with a brown chair.
Navigation from Dialog History

The goal room contains a rug.

**Navigator:** Should I go to the left or right?

**Oracle:** Go left and turn right after the bathroom.

**Navigator:** Do I need to go in the room with the run or keep on going right?

**Oracle:** Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.

**Navigator:** Should I go into the kitchen or to the right?

**Oracle:** Turn toward the front door and go up the stairs you see on the right.

**Navigator:** Do I go left or right?

**Oracle:** Go along the railing to the right. Stop at the room with a brown chair.
Navigation from Dialog History

The goal room contains a rug.

**Navigator:** Should I go to the left or right?

**Oracle:** Go left and turn right after the bathroom.

**Navigator:** Do I need to go into the room with the run or keep on going right?

**Oracle:** Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.

**Navigator:** Should I go into the kitchen or to the right?

**Oracle:** Turn toward the front door and go up the stairs you see on the right.

**Navigator:** Do I go left or right?

**Oracle:** Go along the railing to the right. Stop at the room with a brown chair.

- **Input:** History so far + visual frame per timestep.
Navigation from Dialog History

The goal room contains a rug.

Navigator: Should I go to the left or right?

Oracle: Go left and turn right after the bathroom.

Navigator: Do I need to go in the room with the run or keep on going right?

Oracle: Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.

Navigator: Should I go into the kitchen or to the right?

Oracle: Turn toward the front door and go up the stairs you see on the right.

Navigator: Do I go left or right?

Oracle: Go along the railing to the right. Stop at the room with a brown chair.

- **Input**: History so far + visual frame per timestep.
- **Output**: Navigation action per timestep.
Navigation from Dialog History

**The goal room contains a rug.**

3 steps

**Navigator:** Should I go to the left or right?

**Oracle:** Go left and turn right after the bathroom.

4 steps

**Navigator:** Do I need to go in the room with the run or keep on going right?

**Oracle:** Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.

**Navigator:** Should I go into the kitchen or to the right?

**Oracle:** Turn toward the front door and go up the stairs you see on the right.

**Navigator:** Do I go left or right?

**Oracle:** Go along the railing to the right. Stop at the room with a brown chair.

- **Input:** History so far + visual frame per timestep.
- **Output:** Navigation action per timestep.
- **Goal:** Get closer to the target object room.
Navigation from Dialog History

The goal room contains a rug.

**Navigator:** Should I go to the left or right?

**Oracle:** Go left and turn right after the bathroom.

**Navigator:** Do I need to go in the room with the run or keep on going right?

**Oracle:** Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.

**Navigator:** Should I go into the kitchen or to the right?

**Oracle:** Turn toward the front door and go up the stairs you see on the right.

**Navigator:** Do I go left or right?

**Oracle:** Go along the railing to the right. Stop at the room with a brown chair.

- **Input:** History so far + visual frame per timestep.
- **Output:** Navigation action per timestep.
- **Goal:** Get closer to the target object room.
- **2k dialogs → 7k histories.**
Navigation from Dialog History

The goal room contains a rug.

Navigator: Should I go to the left or right?

Oracle: Go left and turn right after the bathroom.

Navigator: Do I need to go in the room with the run or keep on going right?

Oracle: Turn right and take the tiny hallway on the right. You will ascend the stairs you find on the right.

Navigator: Should I go into the kitchen or to the right?

Oracle: Turn toward the front door and go up the stairs you see on the right.

Navigator: Do I go left or right?

Oracle: Go along the railing to the right. Stop at the room with a brown chair.

- Input: History so far + visual frame per timestep.
- Output: Navigation action per timestep.
- Goal: Get closer to the target object room.
- 2k dialogs → 7k histories.
Initial, Sequence-to-Sequence Model
Initial, Sequence-to-Sequence Model
Initial, Sequence-to-Sequence Model

*Navigator:* Should I turn left down the hallway ahead?
*Oracle:* ya
Initial, Sequence-to-Sequence Model

**Navigator:** Should I turn left down the hallway ahead?
**Oracle:** ya
Initial, Sequence-to-Sequence Model

*Oracle*: Through the lobby. So go through the door next to the green towel. Go to the left door next to the **two yellow lights**. Walk straight to the end of the hallway and stop

...  

*Navigator*: Are these the **yellow lights** you were talking about?
Initial, Sequence-to-Sequence Model

All Previous Questions and Answers

Last Question

<NAV> Should I go upstairs?
<ORA> Yeah, head up the stairs.

Last Answer

<NAV> into the hall or the office?
<ORA> Left into the hall. Follow it to a living room.

Target

\[ \hat{a}_0, \hat{a}_1, \ldots, \hat{a}_n \]
Evaluation - Validation (best perf over 200 epochs)

- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.
Initial, Sequence-to-Sequence Model

All Previous Questions and Answers

<NAV> into the hall or the office ? <ORA> Left into the hall . Follow it to a living room .

Last Question

<NAV> Should I go upstairs ? <ORA> Yeah , head up the stairs .

Last Answer

Target

\( \hat{a}_0 \rightarrow \hat{a}_1 \rightarrow \hat{a}_n \)
Evaluation - Validation (best perf over 200 epochs)

- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.
Evaluation - Validation (best perf over 200 epochs)

- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.

The bar chart shows the goal progress (in meters) for different dialog history encoded conditions: target object, + last answer, + last question, and + all. The best performance is noted with a score of 5.71.
Evaluation - Validation (best perf over 200 epochs)

- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.

![Bar Chart](chart.png)

<table>
<thead>
<tr>
<th>Dialog History Encoded</th>
<th>Goal Progress (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>target object</td>
<td>5.71</td>
</tr>
<tr>
<td>+ last answer</td>
<td>6.04</td>
</tr>
<tr>
<td>+ last question</td>
<td></td>
</tr>
<tr>
<td>+ all</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation - Validation (best perf over 200 epochs)

- Seen Environments:
  - Novel dialogs.
  - Houses seen at training time.

<table>
<thead>
<tr>
<th>Dialog History Encoded</th>
<th>Goal Progress (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>target object</td>
<td>5.71</td>
</tr>
<tr>
<td>+ last answer</td>
<td>6.04</td>
</tr>
<tr>
<td>+ last question</td>
<td>6.16</td>
</tr>
<tr>
<td>+ all</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation - Validation (best perf over 200 epochs)

- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.

<table>
<thead>
<tr>
<th>Dialog History Encoded</th>
<th>Goal Progress (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>target object</td>
<td>5.71</td>
</tr>
<tr>
<td>+ last answer</td>
<td>6.04</td>
</tr>
<tr>
<td>+ last question</td>
<td>6.16</td>
</tr>
<tr>
<td>+ all</td>
<td>5.92</td>
</tr>
</tbody>
</table>
Evaluation - Validation (best perf over 200 epochs)

**Goal Progress (m)**
- Target object: 5.1
- + last answer: 6.04
- + last question: 6.16
- + all: 5.92

- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.

*Adding Question-Answer Context Helps in Seen Environments.*
Evaluation - Test (epoch of best Val Unseen)

- Unseen Envs:
  - Novel dialogs.
  - Novel houses not seen during training.
Evaluation - Test (epoch of best Val Unseen)

- **Unseen Envs:**
  - Novel dialogs.
  - Novel houses not seen during training.

![Graph showing goal progress](image)
Evaluation - Test (epoch of best Val Unseen)

- Unseen Envs:
  - Novel dialogs.
  - Novel houses not seen during training.
Evaluation - Test (epoch of best Val Unseen)

- **Unseen Envs:**
  - Novel dialogs.
  - Novel houses not seen during training.
Evaluation - Test (epoch of best Val Unseen)

- Unseen Envs:
  - Novel dialogs.
  - Novel houses not seen during training.
Evaluation - Test (epoch of best Val Unseen)

- Unseen Envs:
  - Novel dialogs.
  - Novel houses not seen during training.

Adding Dialog History Helps in Unseen Environments.
Evaluation - Unimodal Baselines

![Graph showing Goal Progress (m) for different models: Action-only, Vis-only, Lang-only, Lang+Vis, and Best Possible. The graph compares Validation (Seen) and Test (Unseen) results. The Lang+Vis model has the highest Goal Progress (5.92 m) for Validation and 2.35 m for Test (Unseen).]
Evaluation - Unimodal Baselines

Goal Progress (m)

- Action-only
- Vis-only
- Lang-only
- Lang + Vis

Validation (Seen) vs Test (Unseen)

Best Possible

5.92
2.35
Evaluation - Unimodal Baselines

![Graph showing goal progress for different models: Action-only, Vis-only, Lang-only, Lang+Vis, and Best Possible. The graph compares validation (seen) and test (unseen) performance. The Lang+Vis model achieves the highest goal progress of 5.92 meters in the validation set and 2.35 meters in the test set.](image)
Evaluation - Unimodal Baselines

Action-only
Evaluation - Unimodal Baselines

Action-only

Vis-only

Goal Progress (m)

- Validation (Seen)
- Test (Unseen)

<table>
<thead>
<tr>
<th>Model</th>
<th>Action-only</th>
<th>Vis-only</th>
<th>Lang-only</th>
<th>Lang+Vis</th>
<th>Best Possible</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.91</td>
<td>0.52</td>
<td>1.74</td>
<td>5.92</td>
<td>2.35</td>
</tr>
</tbody>
</table>
Evaluation - Unimodal Baselines

Action-only

Vis-only

Goal Progress (m)

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation (Seen)</th>
<th>Test (Unseen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action-only</td>
<td>0.91</td>
<td>0.52</td>
</tr>
<tr>
<td>Vis-only</td>
<td>5.72</td>
<td>1.74</td>
</tr>
<tr>
<td>Lang-only</td>
<td>5.92</td>
<td>2.35</td>
</tr>
<tr>
<td>Lang+Vis</td>
<td>5.92</td>
<td>2.35</td>
</tr>
<tr>
<td>Best Possible</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation - Unimodal Baselines

Action-only

Vis-only

![Diagram showing goal progress comparison for different models: Action-only, Vis-only, Lang-only, Lang+Vis, and Best Possible. The chart shows the goal progress in meters for validation (seen) and test (unseen) data.]
Evaluation - Unimodal Baselines

Action-only

Vis-only

Lang-only

[Diagram showing unimodal baselines]

Goal Progress (m)

Validation (Seen) vs. Test (Unseen)

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation (Seen)</th>
<th>Test (Unseen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action-only</td>
<td>0.91</td>
<td>0.52</td>
</tr>
<tr>
<td>Vis-only</td>
<td>5.72</td>
<td>1.74</td>
</tr>
<tr>
<td>Lang-only</td>
<td>1.58</td>
<td>1.40</td>
</tr>
<tr>
<td>Lang+Vis</td>
<td>5.92</td>
<td>2.35</td>
</tr>
</tbody>
</table>

Best Possible
Evaluation - Unimodal Baselines

Initial Model Uses Multimodal Input in Unseen Environments.
Evaluation - Unimodal Baselines

Action-only

Vis-only

Lang-only

Goal Progress (m)

Validation (Seen)  Test (Unseen)

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action-only</td>
<td>0.91</td>
<td>0.52</td>
</tr>
<tr>
<td>Vis-only</td>
<td>5.72</td>
<td>1.74</td>
</tr>
<tr>
<td>Lang-only</td>
<td>1.58</td>
<td>1.40</td>
</tr>
<tr>
<td>Lang+Vis</td>
<td>5.92</td>
<td>2.35</td>
</tr>
</tbody>
</table>

Best Possible
Evaluation - Unimodal Baselines

Action-only

Vis-only

Lang-only

Goal Progress (m)

Validation (Seen) vs Test (Unseen)

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation</th>
<th>Test (Unseen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action-only</td>
<td>0.91</td>
<td>0.52</td>
</tr>
<tr>
<td>Vis-only</td>
<td>1.74</td>
<td>1.40</td>
</tr>
<tr>
<td>Lang-only</td>
<td>1.58</td>
<td>1.40</td>
</tr>
<tr>
<td>Lang+Vis</td>
<td>2.35</td>
<td>9.76</td>
</tr>
</tbody>
</table>

Best Possible: 9.52
Evaluation - Unimodal Baselines

Action-only

Vis-only

Lang-only

Lots of Headroom for Future Models

Validation (Seen)  Test (Unseen)

0.91  0.52  1.40  5.92  2.35  9.52  9.76

Action-only  Vis-only  Lang-only  Lang+Vis  Best Possible

Model
Evaluation Lessons
Evaluation Lessons

- Dialog history:
Evaluation Lessons

- Dialog history:
  - Longer context leads to better performance.
  - Particularly helpful in unseen environments.
Evaluation Lessons

- **Dialog history:**
  - Longer context leads to better performance.
  - Particularly helpful in unseen environments.
- **Unimodal baselines:**
  - Initial model makes use of multimodal information.
Evaluation Lessons

- **Dialog history:**
  - Longer context leads to better performance.
  - Particularly helpful in unseen environments.

- **Unimodal baselines:**
  - Initial model makes use of multimodal information
Evaluation Lessons

- Dialog history:
  - Longer context leads to better performance.
  - Particularly helpful in unseen environments.

- Unimodal baselines:
  - Initial model makes use of multimodal information
  - Multimodal most helpful in unseen environments.
Evaluation Lessons

- Dialog history:
  - Longer context leads to better performance.
  - Particularly helpful in unseen environments.
- Unimodal baselines:
  - Initial model makes use of multimodal information
  - Multimodal most helpful in unseen environments.
- Headroom remains for more nuanced models.
Outline

- Language grounding in visual environments
  - For navigation
  - Unimodal bias [Thomason et al., NAACL’19]
- Vision-and-Dialog Navigation [Thomason et al., in sub]
  - New dataset - CVDN
  - Navigation from dialog history
- Next steps
Incorporating Navigation History

<NAV> into the hall or the office ? <ORA> Left into the hall . Follow it to a living room .

<NAV> Should I go upstairs ? <ORA> Yeah , head up the stairs . <TAR> mat <EOS>

\[ \hat{a}_0 \rightarrow \hat{a}_1 \rightarrow \ldots \rightarrow \hat{a}_n \]
Incorporating Navigation History

*Oracle: You were there briefly but left. There is a turntable behind you a bit. Enter the bedroom next to it.*
Incorporating Navigation History

<NAV> into the hall or the office ? <ORA> Left into the hall . Follow it to a living room .

<NAV> Should I go upstairs ? <ORA> Yeah , head up the stairs .

RN RN RN
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat*.

Into the hall or the office?

Left into the hall. Follow it to a living room.

Should I go upstairs?

Visible Only to the **Oracle**
Visible to both **Navigator** and **Oracle**

Hint: The goal room contains a *mat*.

- **Into the hall or the office?**
- **Left into the hall. Follow it to a living room.**
- **Should I go upstairs?**

Visible Only to the **Oracle**

**Navigation**
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat*.

**Into the hall or the office?**

**Left into the hall. Follow it to a living room.**

**Should I go upstairs?**

Visible Only to the **Oracle**
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat*.

- Into the hall or the office?
- Left into the hall. Follow it to a living room.
- Should I go upstairs?

**Question Generation**
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat*.

Into the hall or the office?

Left into the hall. Follow it to a living room.

Should I go upstairs?

**Visible Only to the Oracle**

**Question Answering**
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a **mat**.

- Into the hall or the office?
- Left into the hall. Follow it to a living room.
- Should I go upstairs?
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat*.

Into the hall or the office?

**Left into the hall. Follow it to a living room.**

Should I go upstairs?

Visible Only to the **Oracle**
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat*.

---

**Navigation**

**Question Generation**

---

Into the hall or the office?

---

Left into the hall. Follow it to a living room.

---

Should I go upstairs?
Visible to both **Navigator** and **Oracle**

**Hint:** The goal room contains a *mat*.

**Navigation**

**Question Generation**

**Visible Only to the **Oracle**

**Question Answering**

- **Into the hall or the office?**
- **Left into the hall. Follow it to a living room.**
- **Should I go upstairs?**
Hint: The goal room contains a *mat*.

- Environment exploration

- Navigation

- Question Generation

- Question Answering
Hint: The goal room contains a *mat*.

- Environment exploration
Hint: The goal room contains a *mat*.

- Environment exploration
- Self-play
- Environment exploration
- Self-play
- “Language” evolution

Hint: The goal room contains a *mat*. 
• Environment exploration
• Self-play
• "Language" evolution
• Reinforcement Learning

Hint: The goal room contains a *mat.*
Bringing Robots from Industrial to Human Spaces
Bringing Robots from Industrial to Human Spaces
Bringing Robots from Industrial to Human Spaces

Industrial

Natural Language

Navigation

Additional Safety

Robust Perception

Human
Bringing Robots from Industrial to Human Spaces

Natural Language → Navigation

Additional Safety → Robust Perception

Industrial → Human

...
Takeaways
Takeaways

- For vision-and-language navigation:
  - Unimodal ablations expose dataset bias.
Takeaways

- For vision-and-language navigation:
  - Unimodal ablations expose dataset bias.
Takeaways

- For vision-and-language navigation:
  - Unimodal ablations expose dataset bias.
- For vision-and-dialog navigation:
Takeaways

* For vision-and-language navigation:
  - Unimodal ablations expose dataset bias.

* For vision-and-dialog navigation:
  - Cooperative dialog facilitates a mix of high- and low-level language.
Takeaways

- For vision-and-language navigation:
  - Unimodal ablations expose dataset bias.

- For vision-and-dialog navigation:
  - Cooperative dialog facilitates a mix of high- and low-level language.
  - Dialog context helps agents infer better navigation actions.
Vision-and-Dialog Navigation
Vision-and-Dialog Navigation

- Jesse Thomason, Daniel Gordon, and Yonatan Bisk.
Vision-and-Dialog Navigation


Vision-and-Dialog Navigation


- Cooperative Vision-and-Dialog Navigation dataset
Vision-and-Dialog Navigation

- Cooperative Vision-and-Dialog Navigation dataset
  - Data+code: https://github.com/mmurray/cvdn/
Vision-and-Dialog Navigation


- Cooperative Vision-and-Dialog Navigation dataset
  - Data+code: https://github.com/mmurray/cvdn/
  - Live demo: https://cvdn.dev/
Vision-and-Dialog Navigation

- Cooperative Vision-and-Dialog Navigation dataset
  - Data+code: https://github.com/mmurray/cvdn/
  - Live demo: https://cvdn.dev/
- Navigation from Dialog History task
Vision-and-Dialog Navigation

- Jesse Thomason, Daniel Gordon, and Yonatan Bisk.  
  “Shifting the Baseline: Single Modality Performance on Visual Navigation and QA”,  
  NAACL’19 [https://arxiv.org/abs/1811.00613]


- Cooperative Vision-and-Dialog Navigation dataset  
  ○ Data+code: https://github.com/mmurray/cvdn/  
  ○ Live demo: https://cvdn.dev/

- Navigation from Dialog History task  
  ○ Leaderboard coming soon!
Future Work: More Expressive Simulator

- High-level instructions.
  - With optional, accompanying low-level.
- Navigation
  + manipulation
  + interaction.
- No dialog to support clarifications.

“Put a slice of bread in the microwave.”