Vision-and-Dialog Navigation

Jesse Thomason
University of Washington

UWNLP

Daniel Gordon

Yonatan Bisk

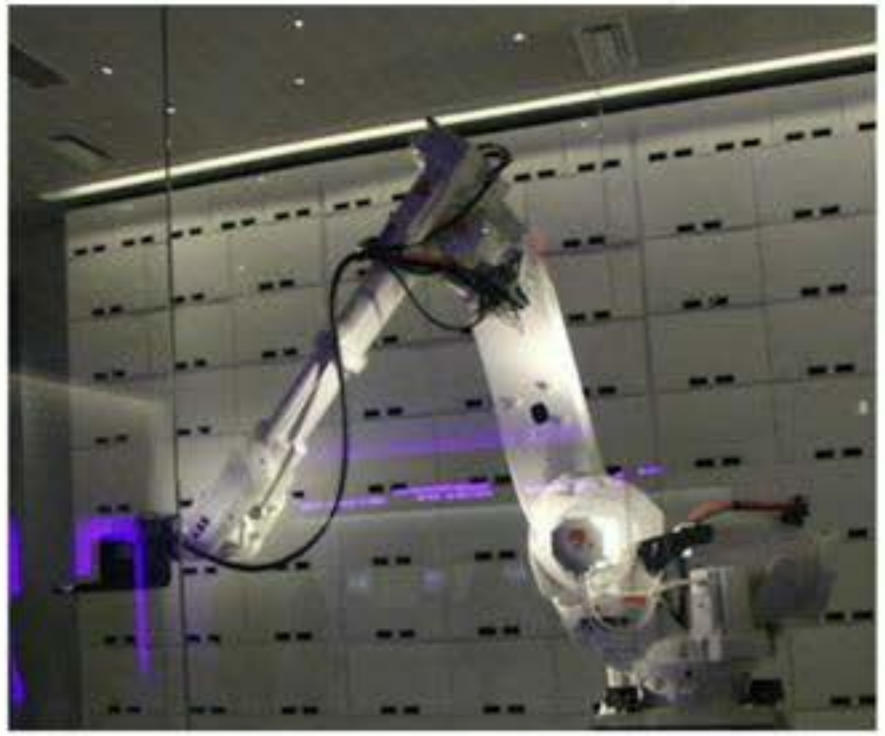
Michael Murray

Maya Cakmak

Luke Zettlemoyer

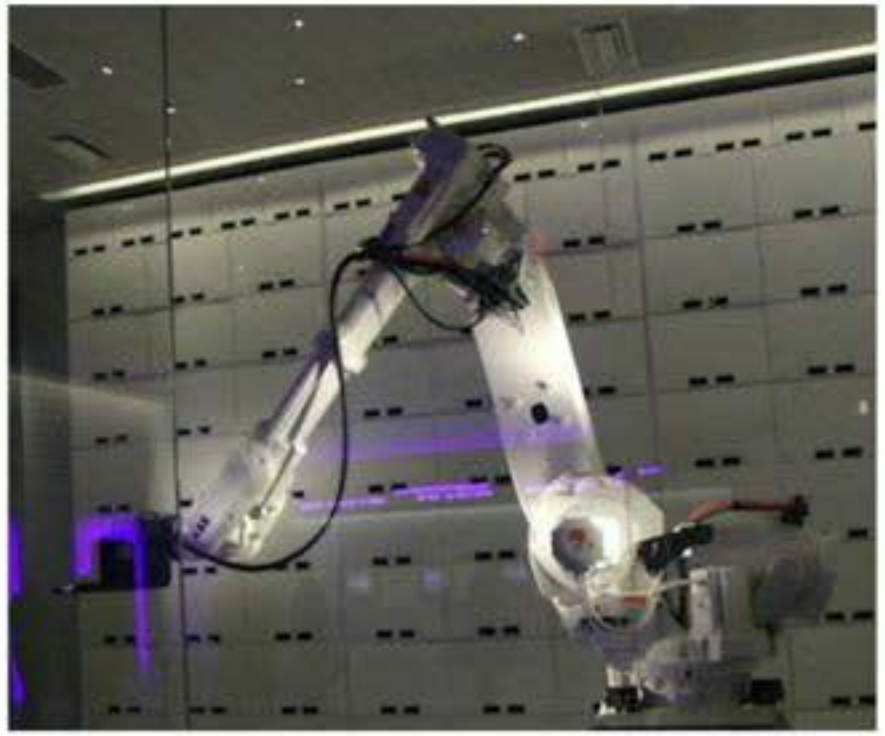
Bringing Robots from Industrial to Human Spaces

Industrial



Bringing Robots from Industrial to Human Spaces

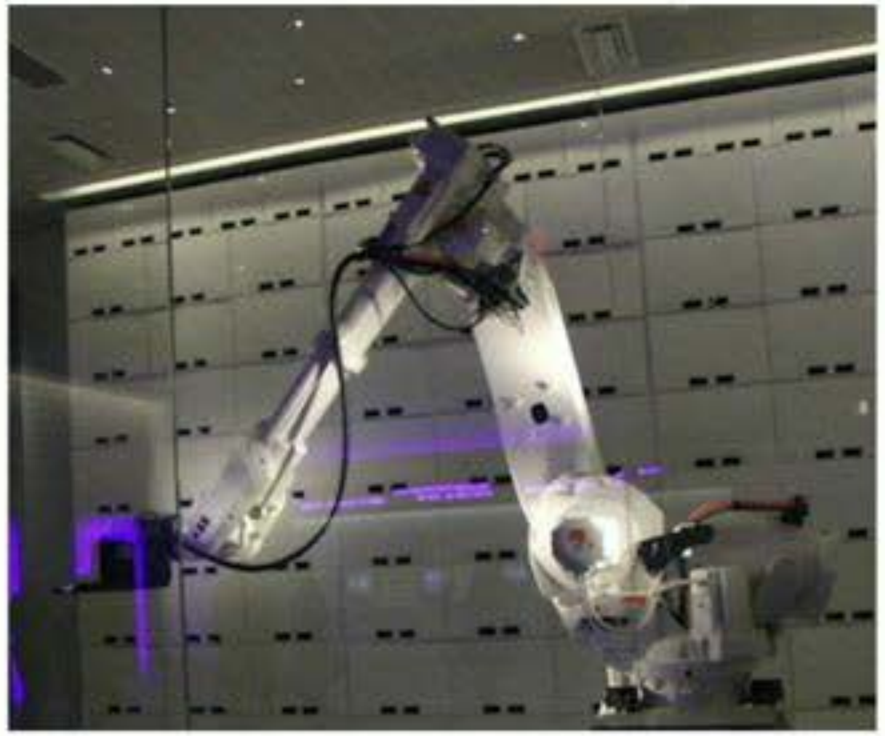
Industrial



Human

Bringing Robots from Industrial to Human Spaces

Industrial



Natural Language →

Navigation →

Additional Safety →

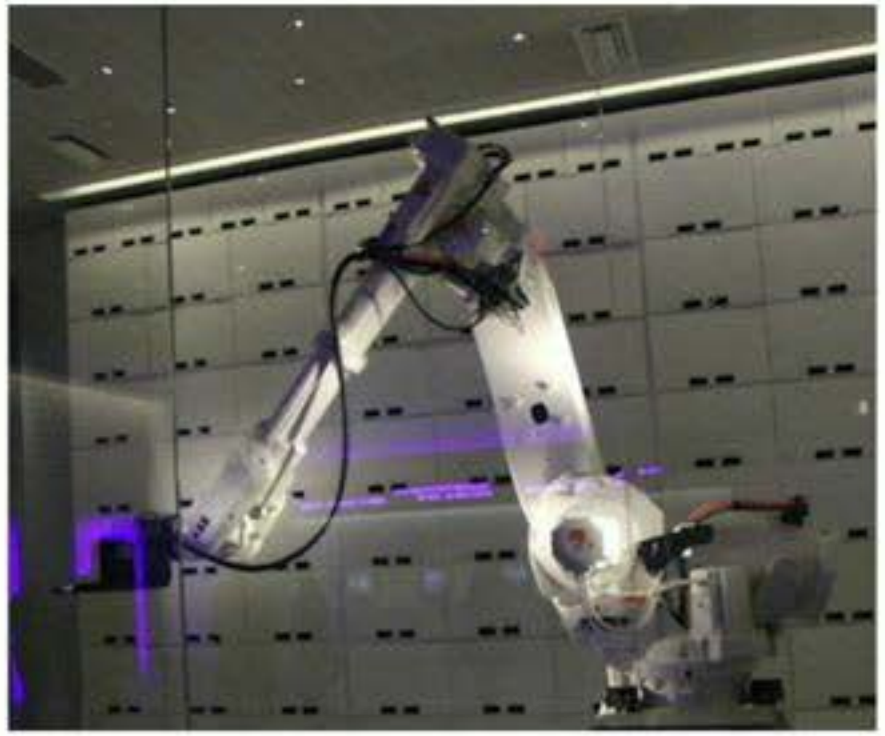
Robust Perception →

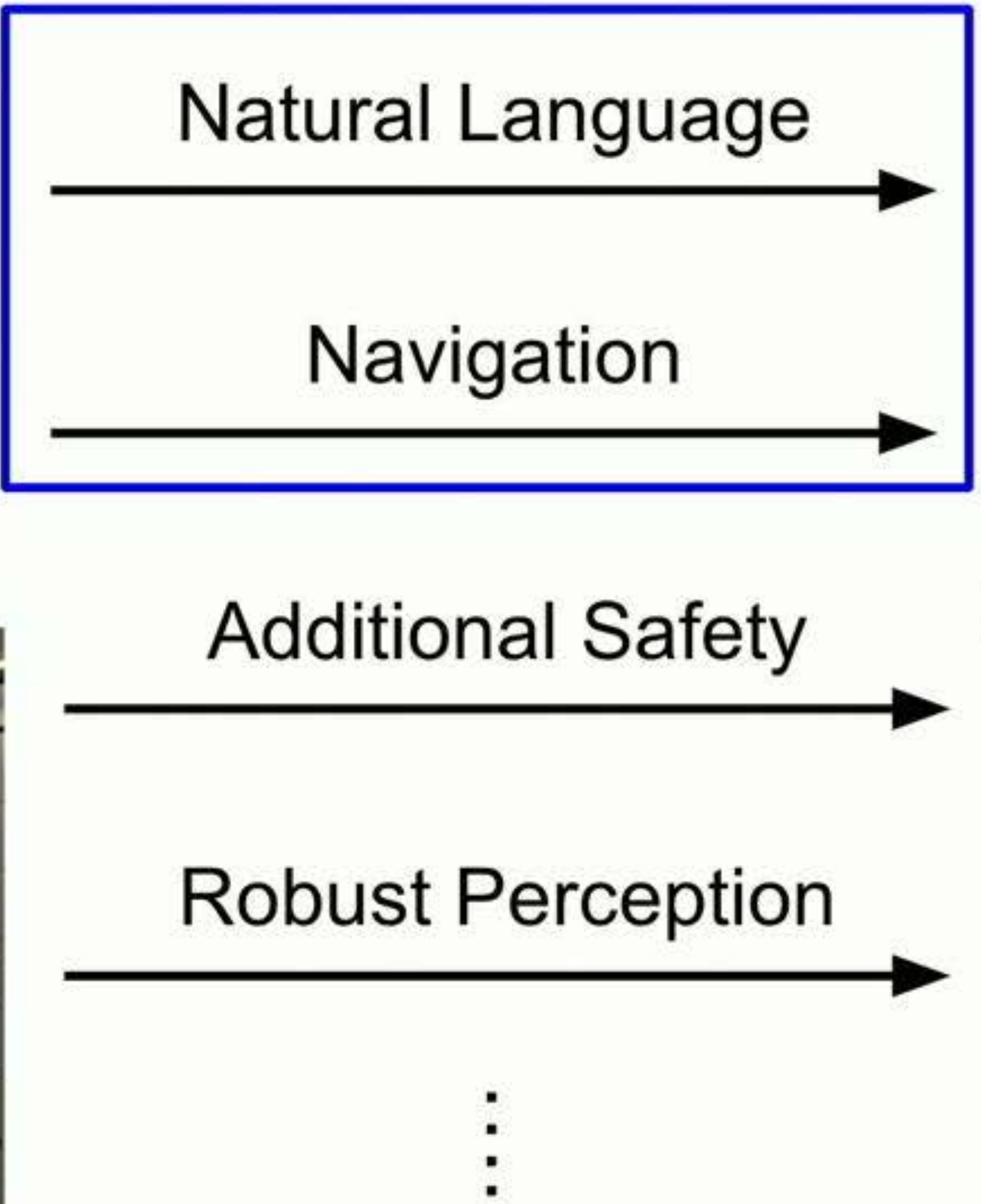
⋮

Human

Bringing Robots from Industrial to Human Spaces

Industrial





Natural Language

Navigation

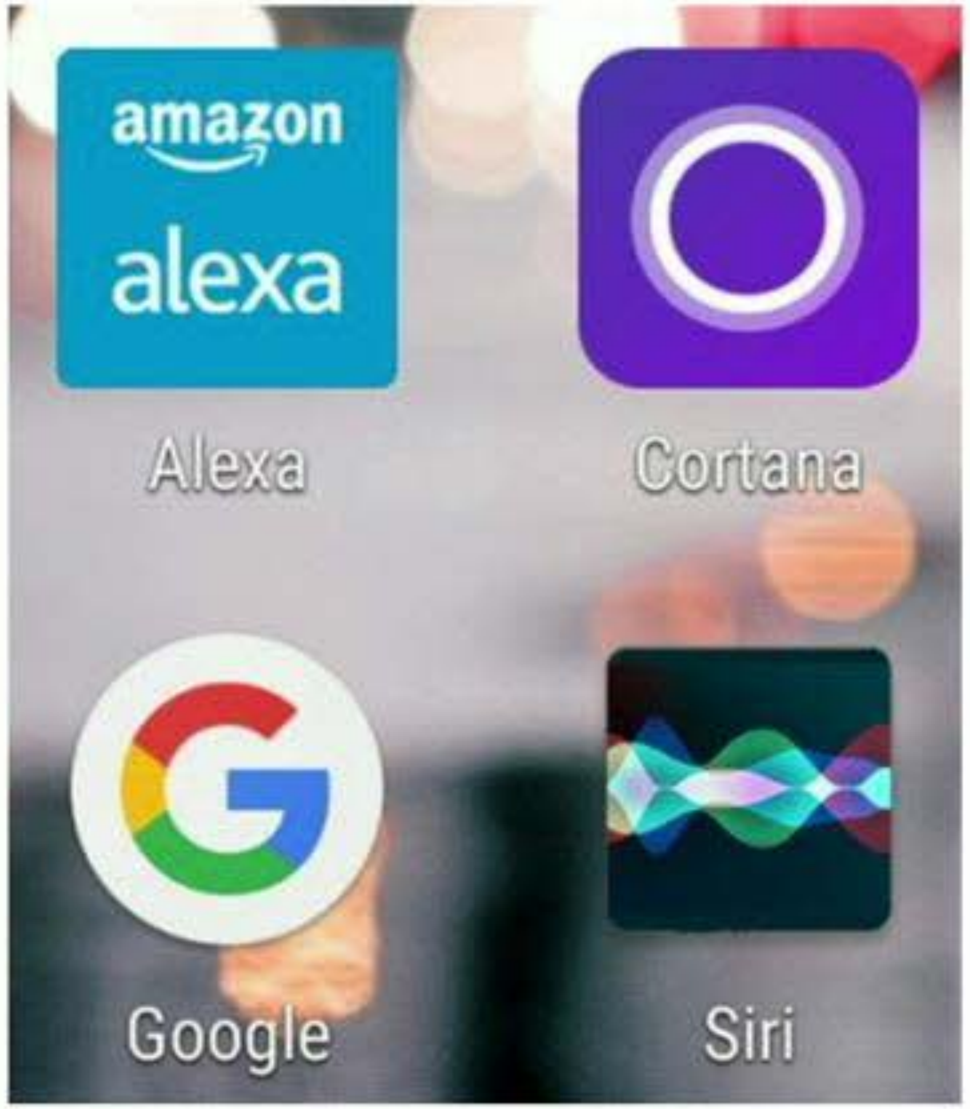
Additional Safety

Robust Perception

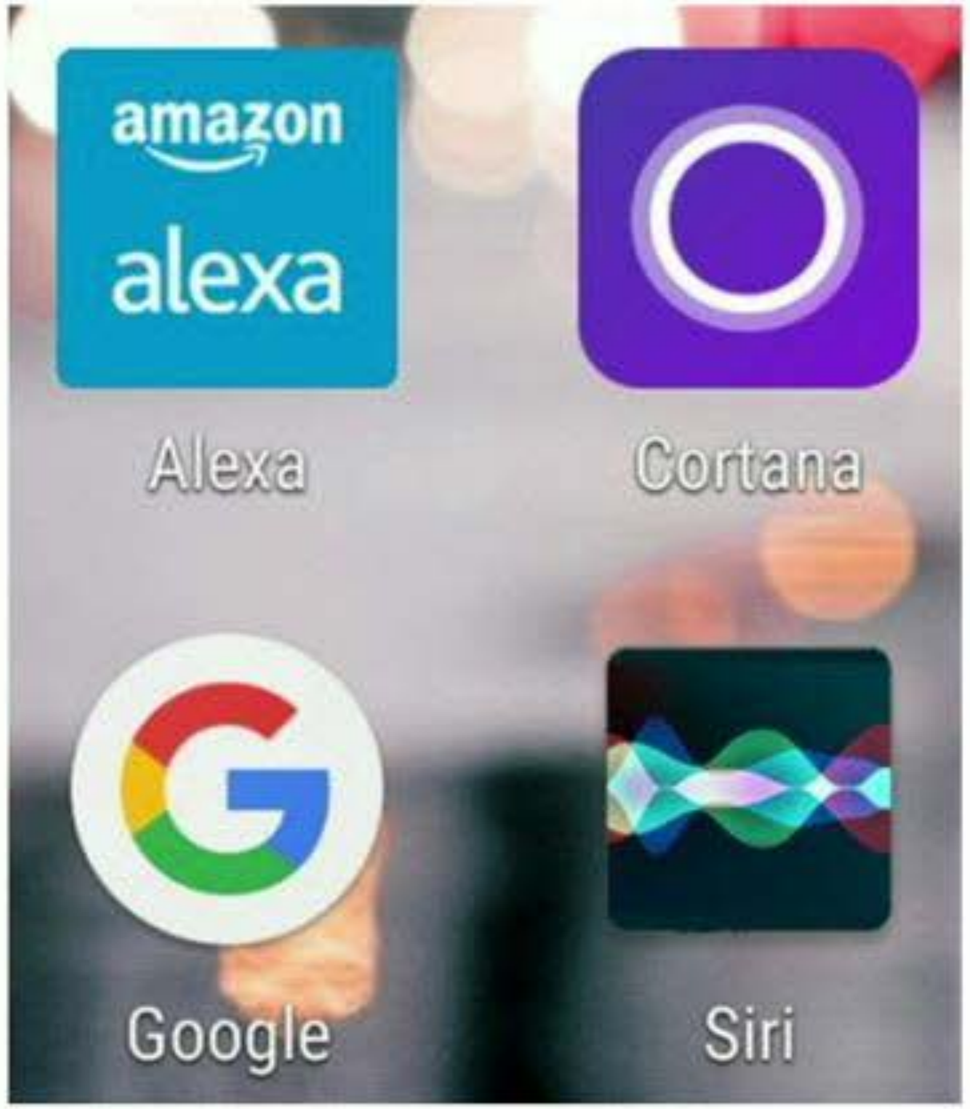
⋮

Human

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

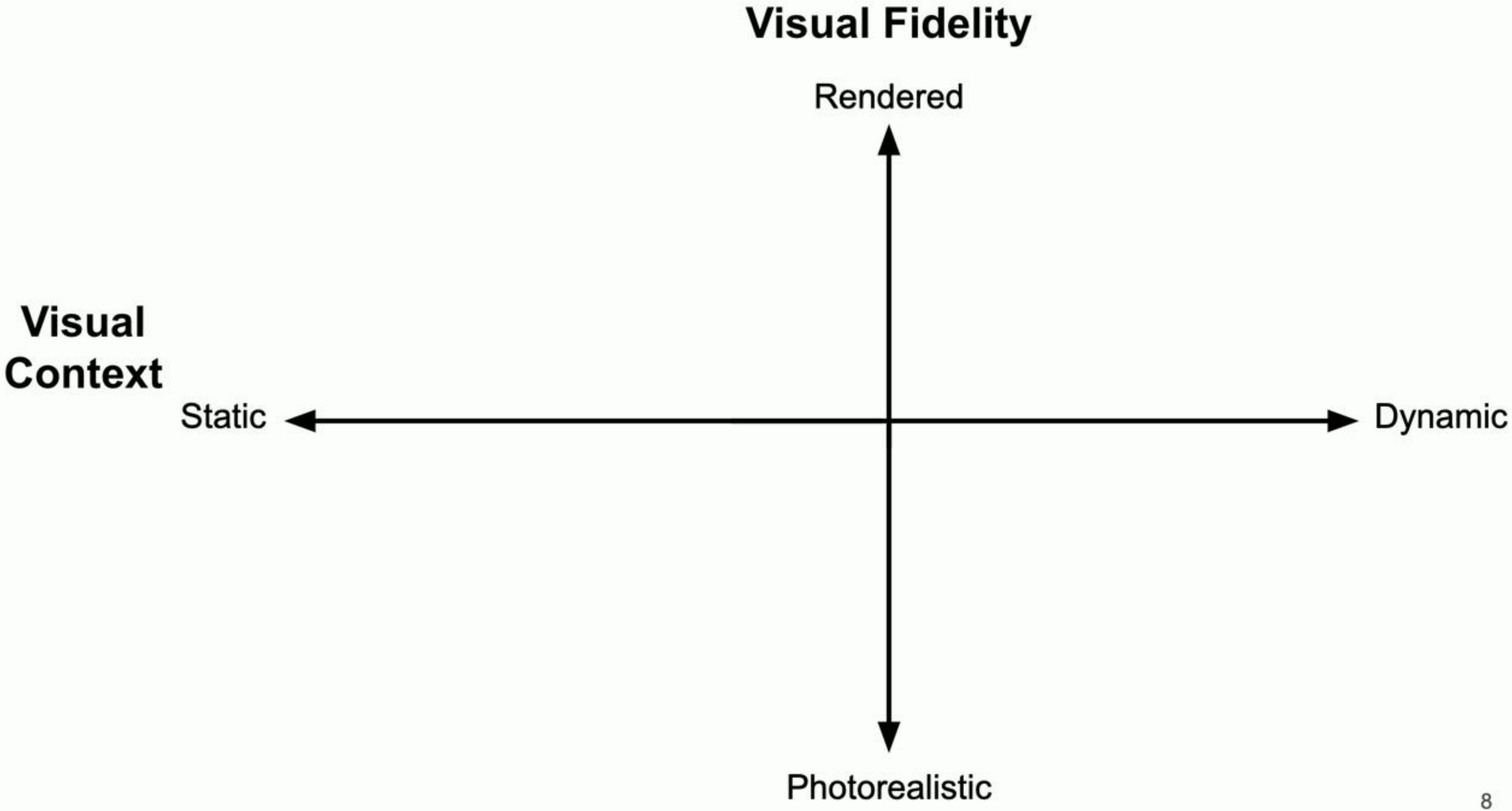
- Common paradigm:
  - **Inputs**:
    - Language tokens (e.g., question)
    - Visual context (e.g., photograph)

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

Rendered

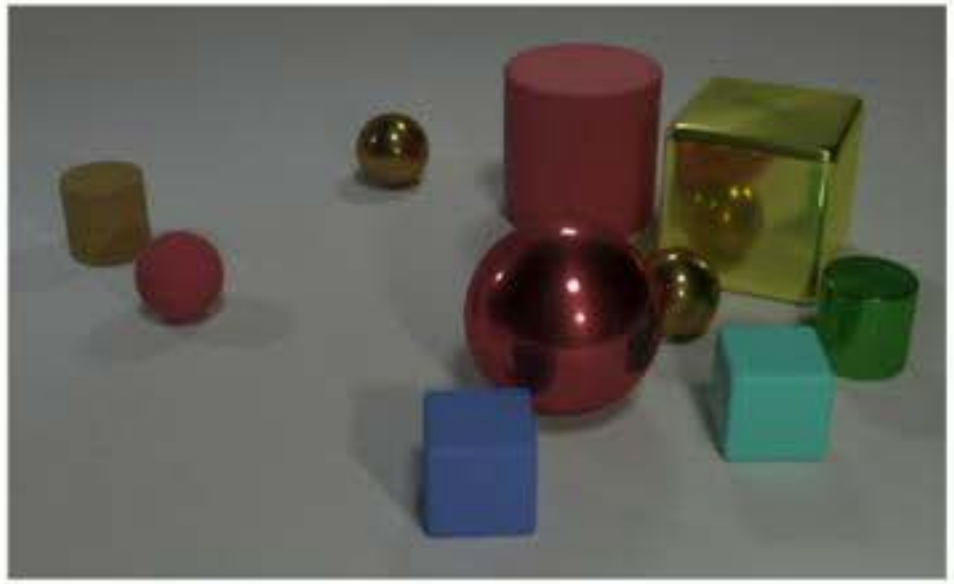
Photorealistic

Visual Context

Static

Dynamic

Visual Fidelity



CLEVR
[Johnson et al., CVPR'17]

Rendered

Visual Context

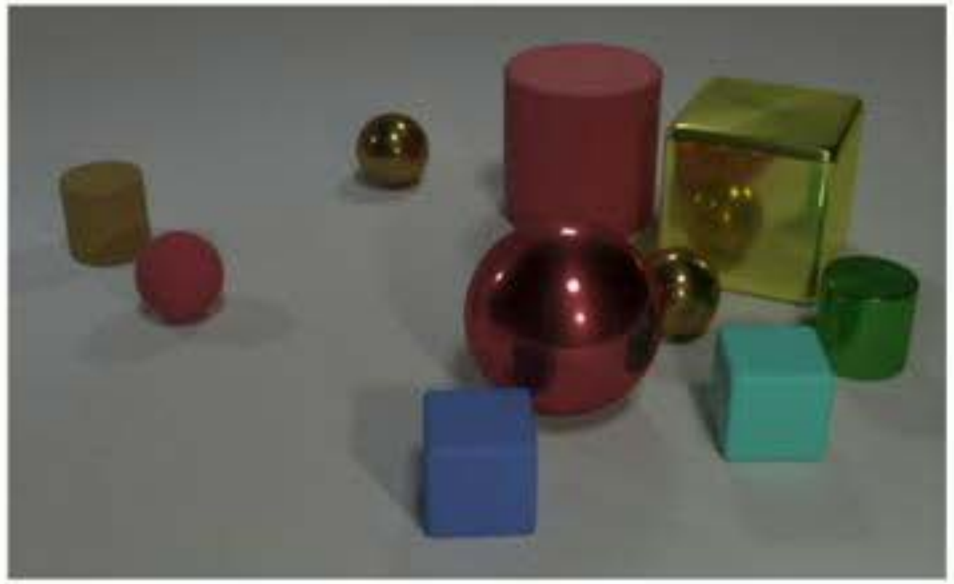
Static

Dynamic

Photorealistic

Visual Fidelity

Rendered



CLEVR
[Johnson et al., CVPR'17]

Visual Context

Static

Dynamic

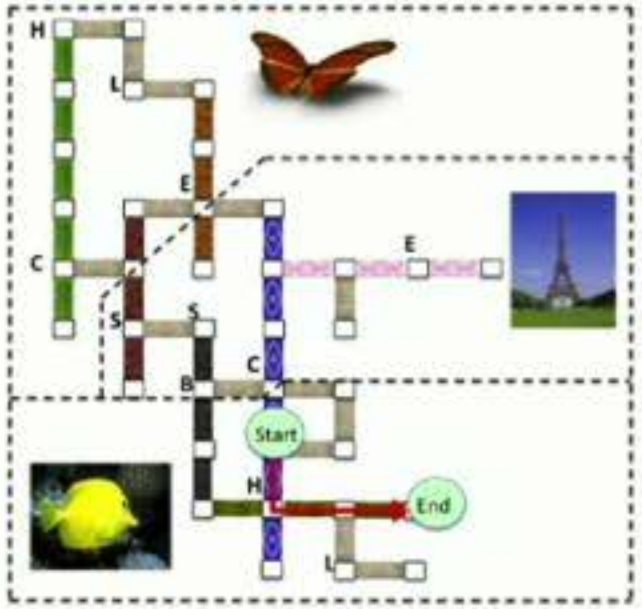
VQA
[Antol et al., CVPR'15]

Photorealistic

Visual Fidelity

Rendered

CLEVR
[Johnson et al., CVPR’17]

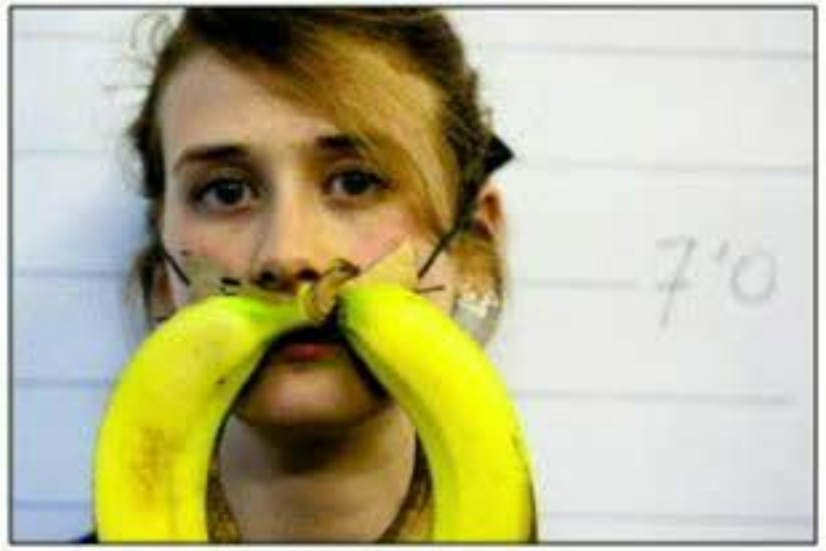


Instruction Following
[Chen and Mooney, AAAI’11]

Visual Context

Static

Dynamic

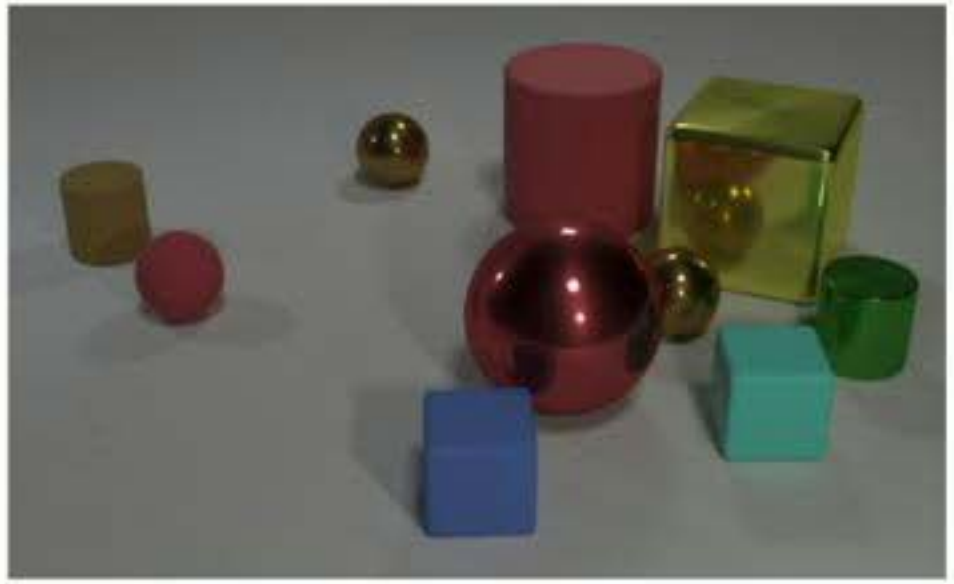


VQA
[Antol et al., CVPR’15]

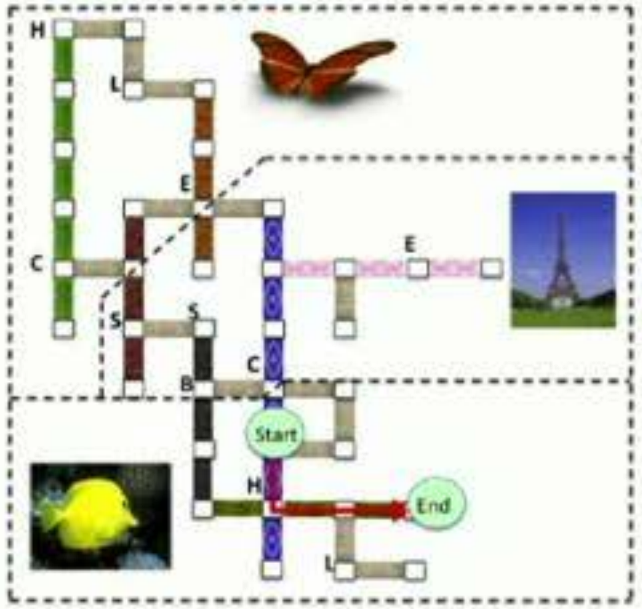
Photorealistic

Visual Fidelity

Rendered



CLEVR
[Johnson et al., CVPR’17]



Instruction Following
[Chen and Mooney, AAAI’11]

Visual Context

Static

Dynamic

VQA
[Antol et al., CVPR’15]

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

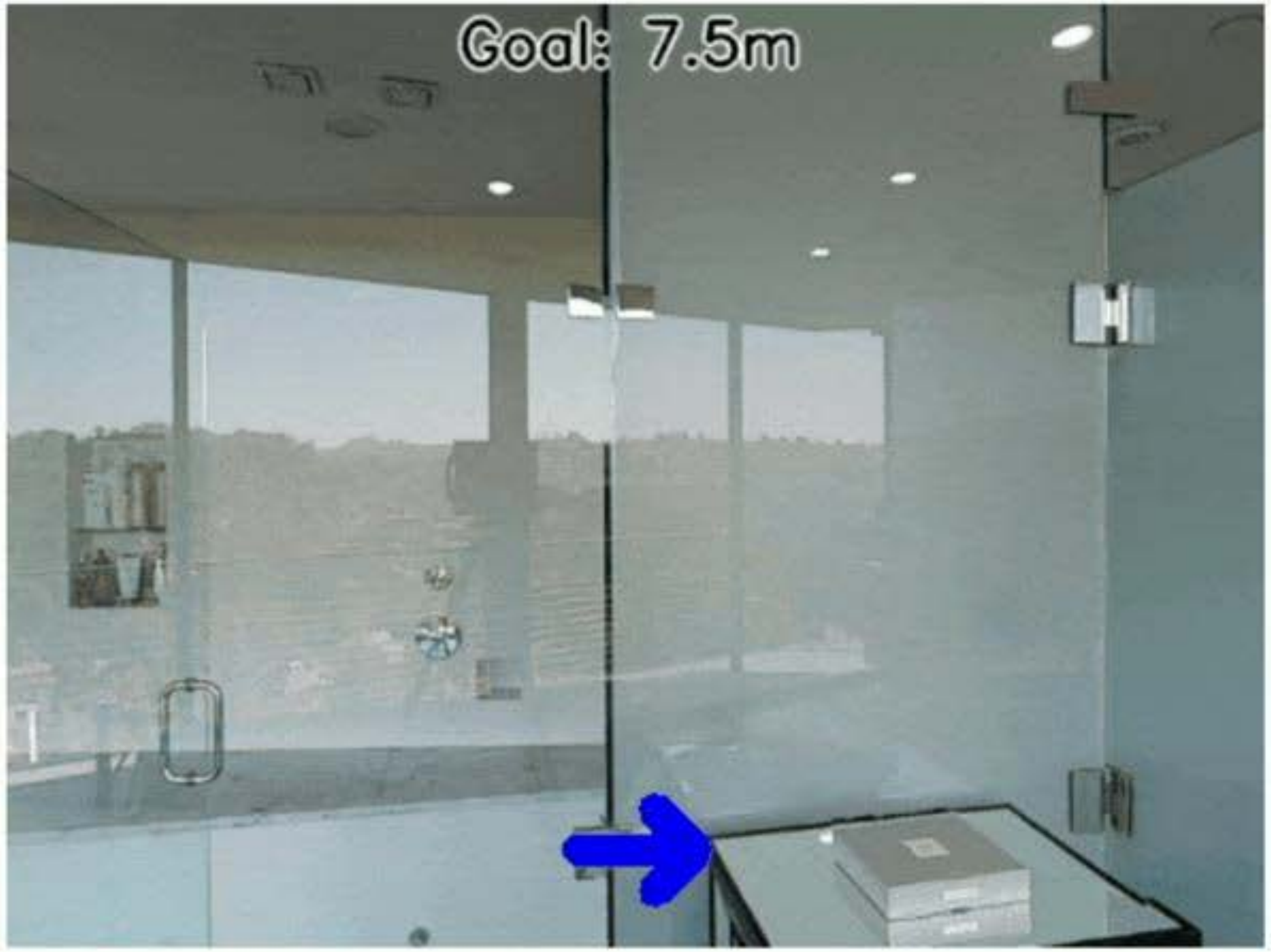
Photorealistic

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

[Anderson et al., CVPR’18]

Vision-and-Language Navigation

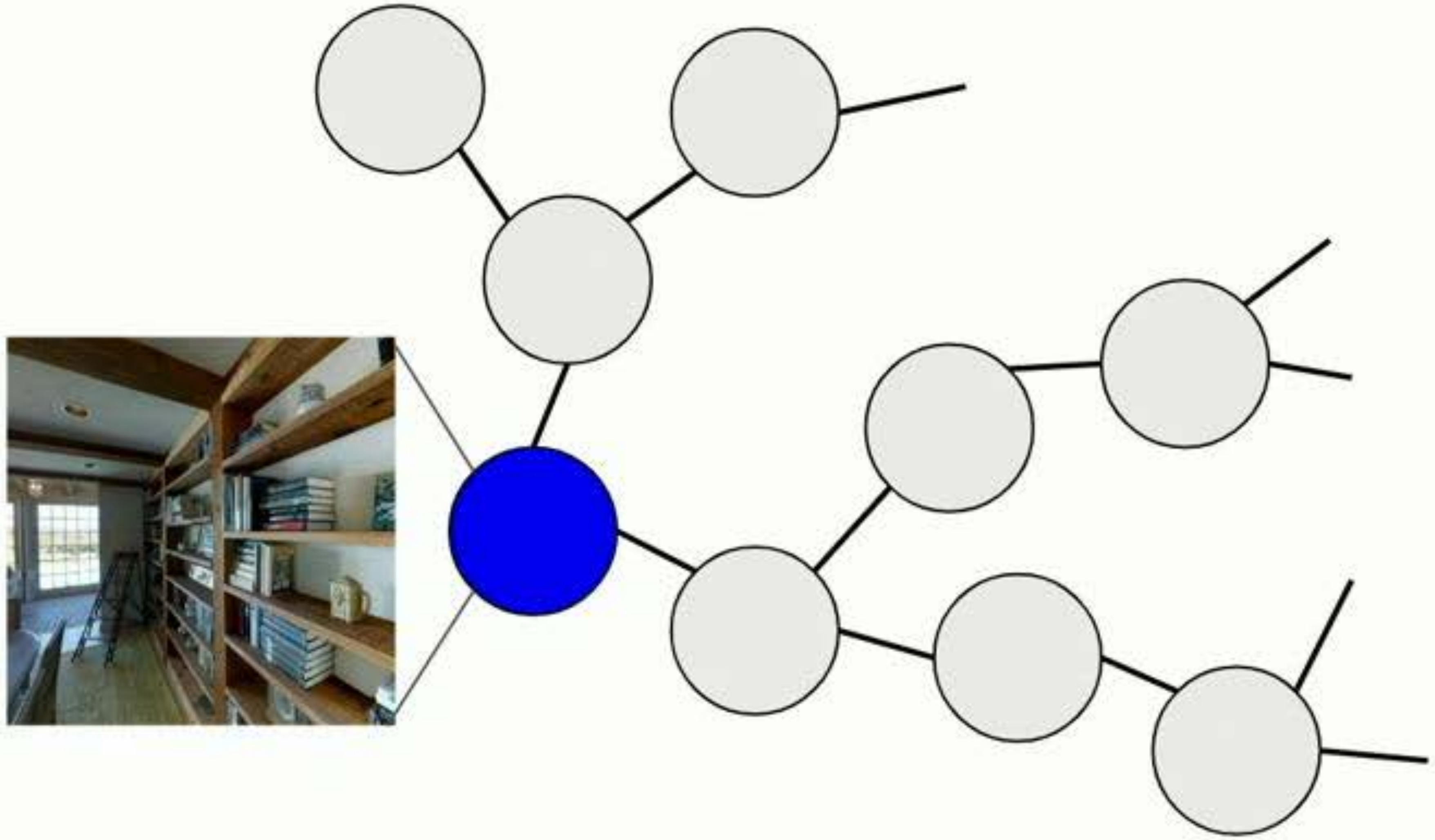


Exit the bathroom. Turn left and exit the room using the door on the left. Wait there.

[Anderson et al., CVPR’18]

“Turn around and exit the library, head down the…”

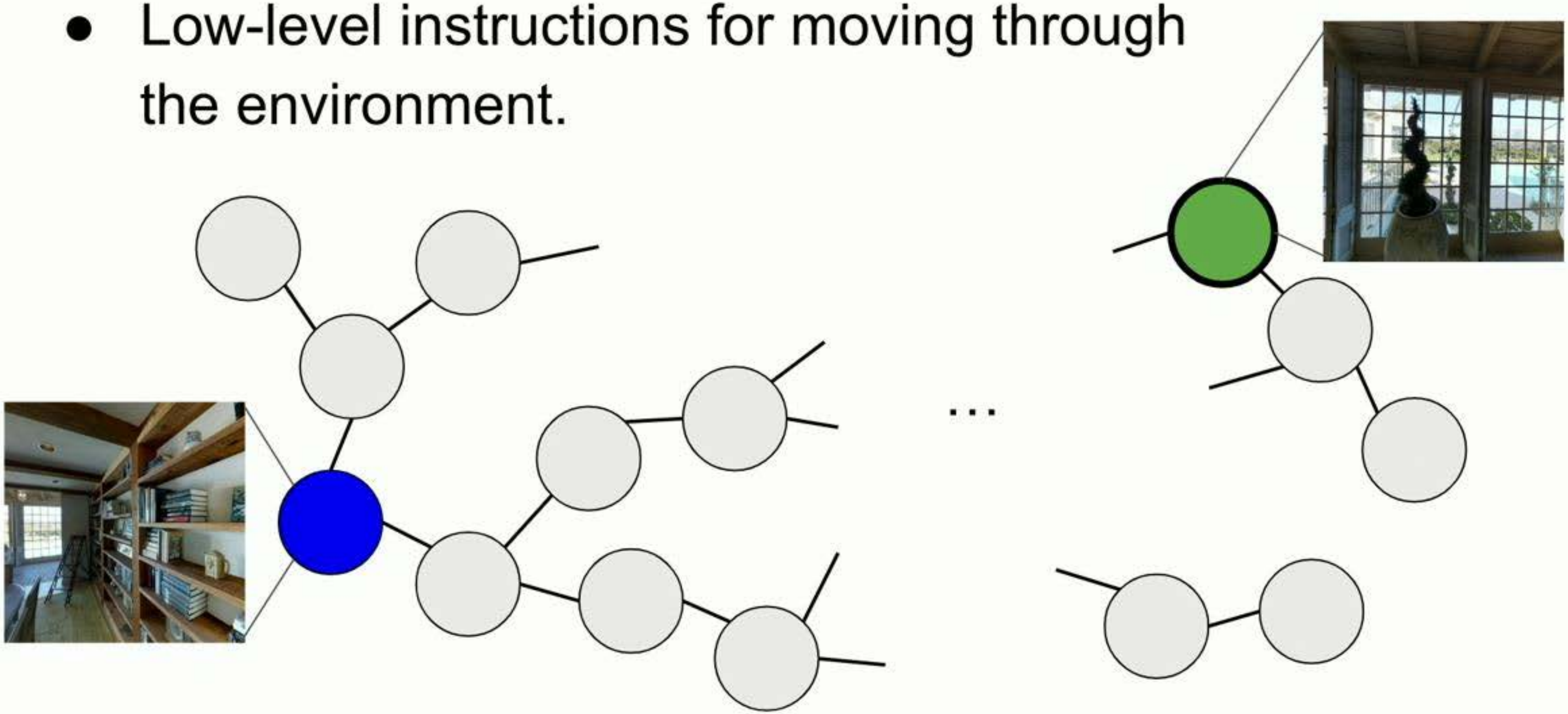
- Low-level instructions for moving through the environment.



[Anderson et al., CVPR'18]

“Turn around and exit the library, head down the…”

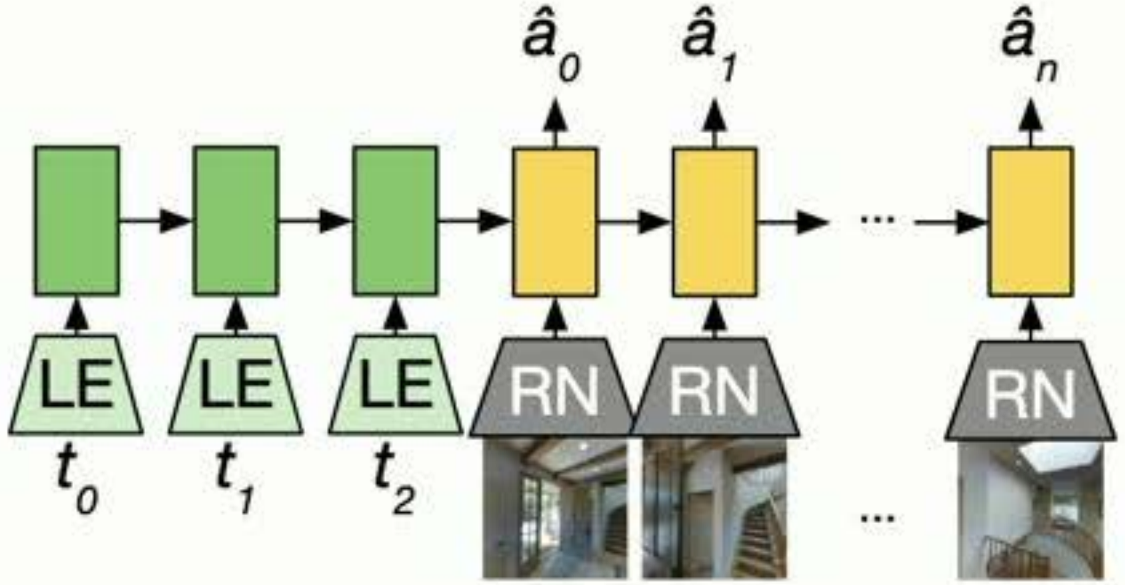
- Low-level instructions for moving through the environment.



[Anderson et al., CVPR'18]

Sequence-to-Sequence Model

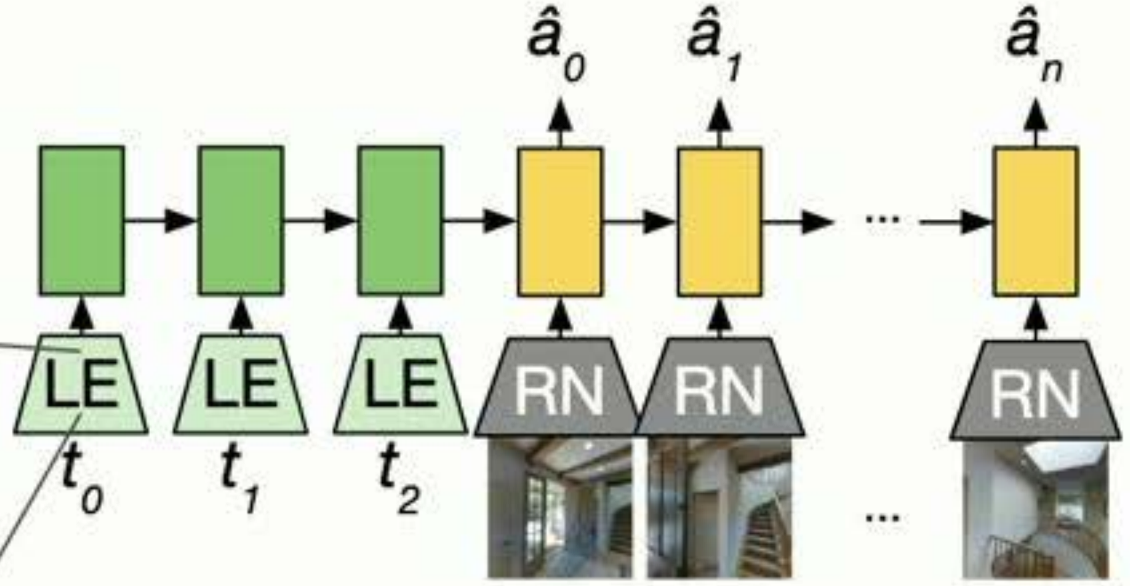
- Encode the language tokens.
- Decode a sequence of actions to take in the environment.
- At every timestep, receive a new visual observation.



[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.

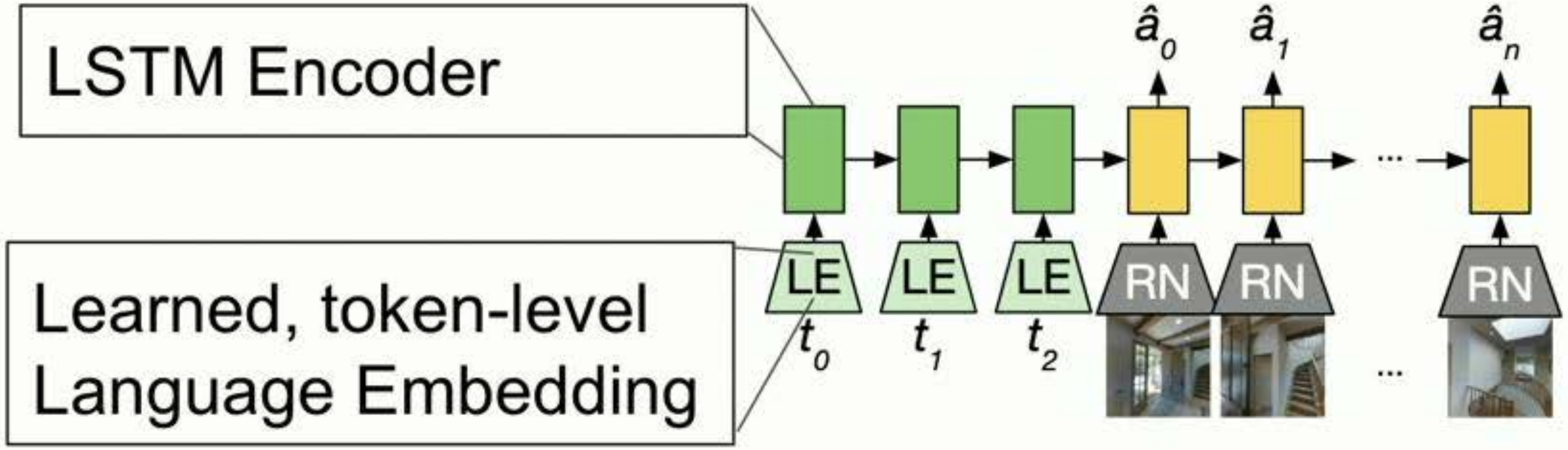


Learned, token-level Language Embedding

[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.



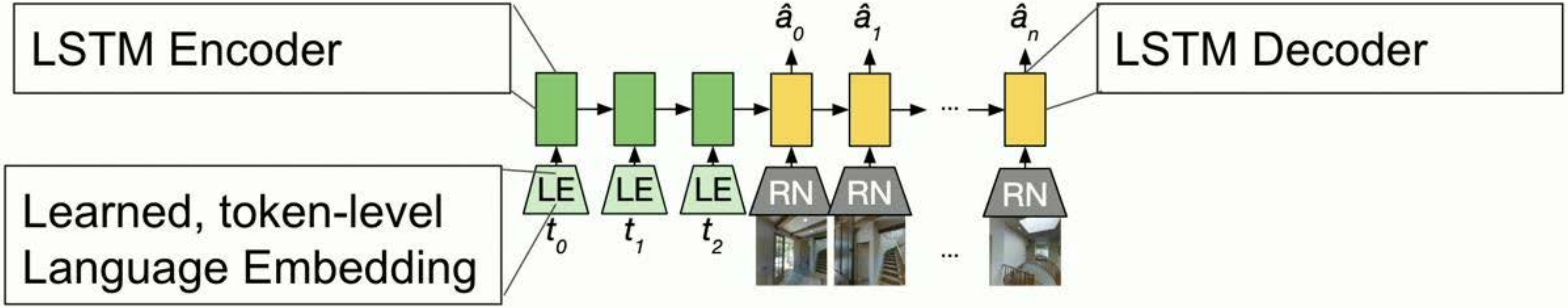
LSTM Encoder

Learned, token-level Language Embedding

[Anderson et al., CVPR'18]

Sequence-to-Sequence Model

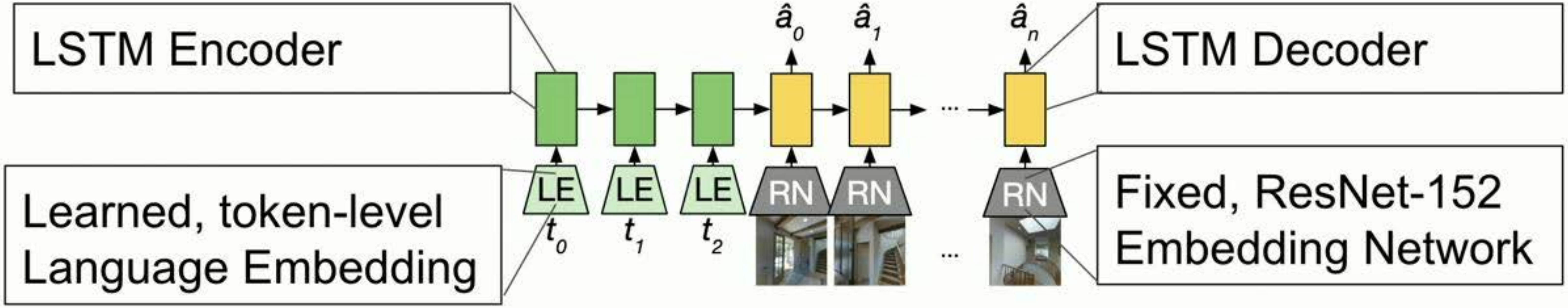
- Encode the language tokens.
- Decode a sequence of actions to take in the environment.
- At every timestep, receive a new visual observation.



[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.

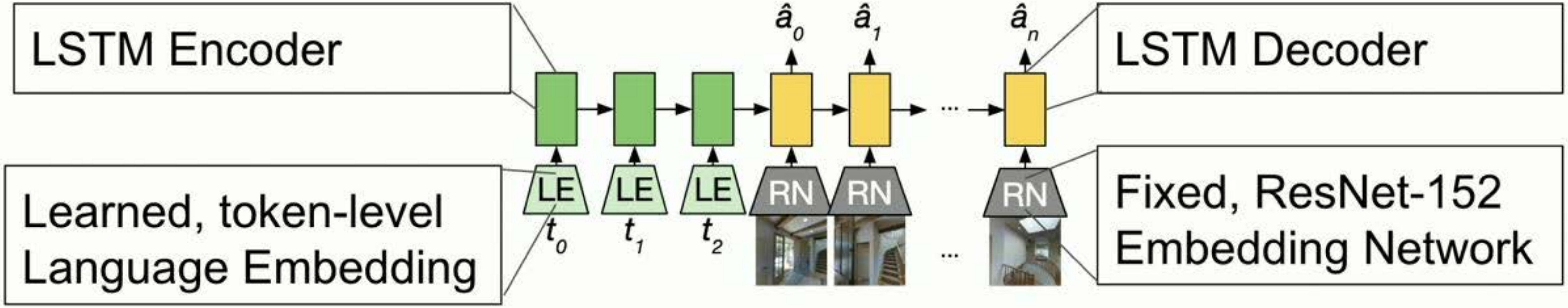


[Anderson et al., CVPR'18]

Sequence-to-Sequence Model

- Encode the lang
- Decode a seque e in the environment.
- At every timestep, receive a new visual observation.

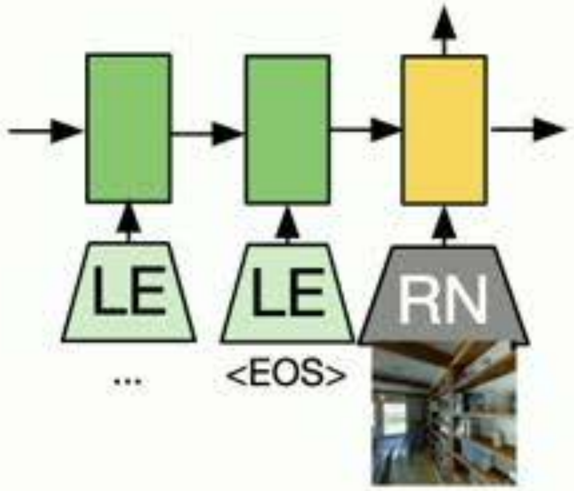
We will build on this model.

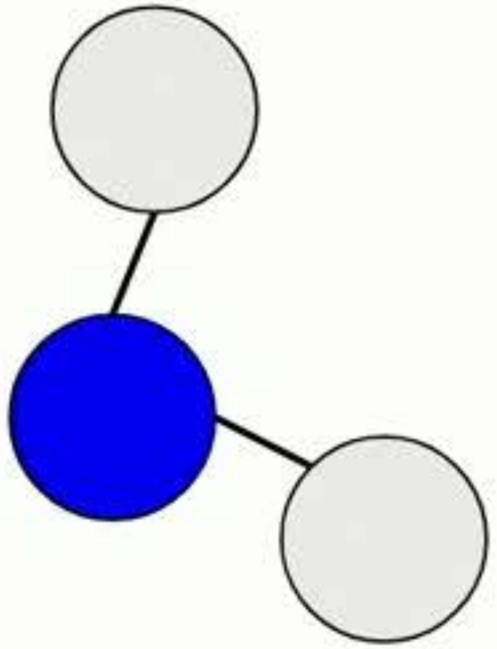


[Anderson et al., CVPR'18]

Sequence-to-Sequence Model

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

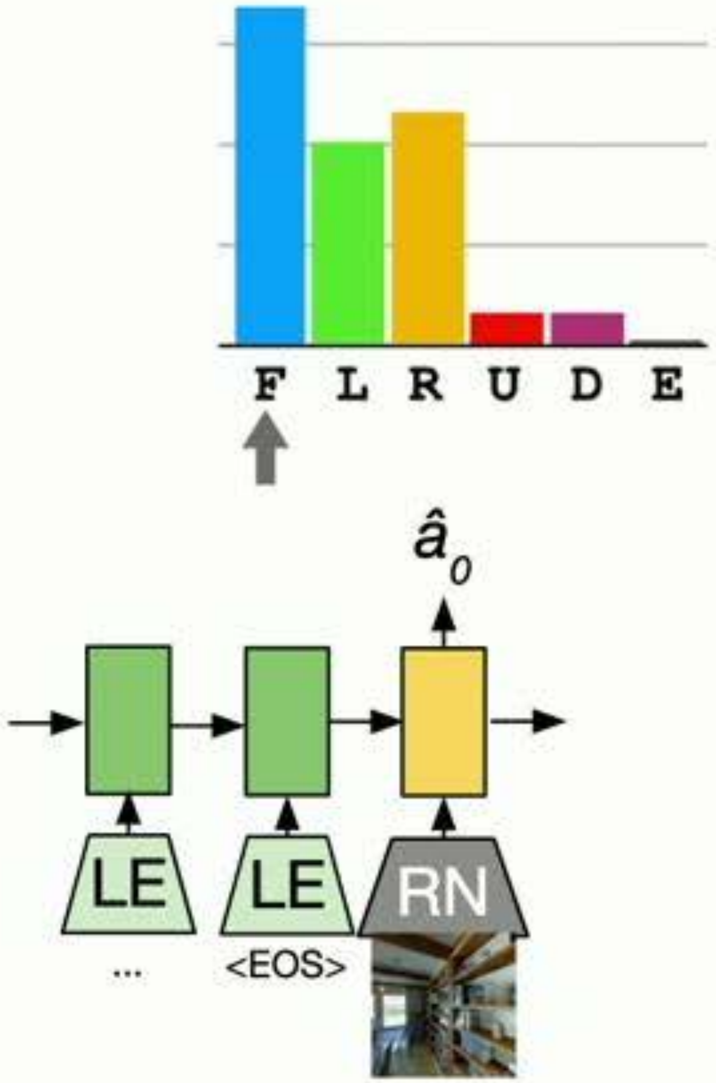


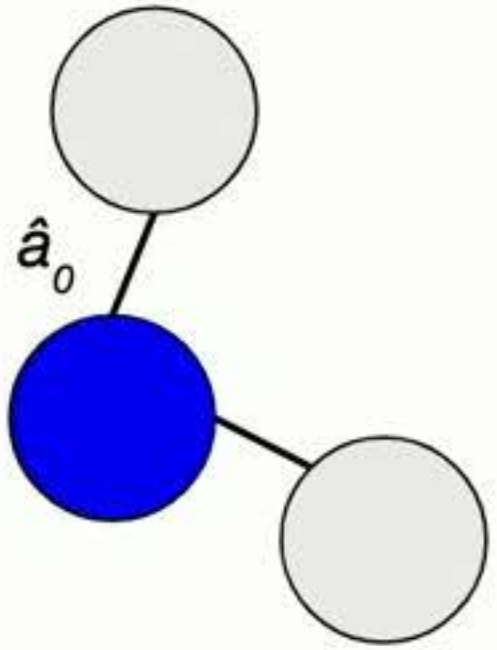


[Anderson et al., CVPR'18]

Sequence-to-Sequence Model

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

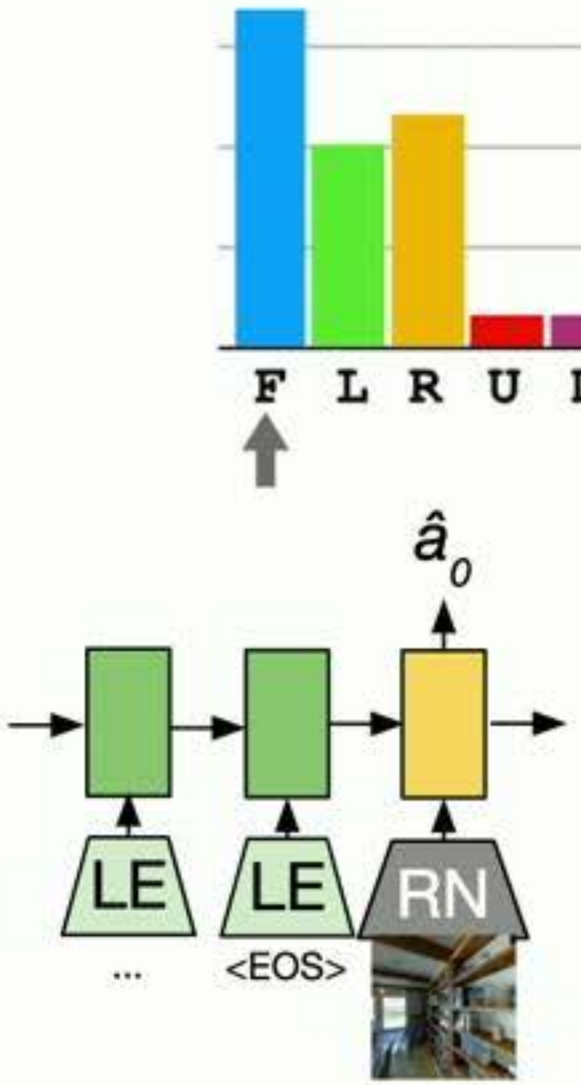


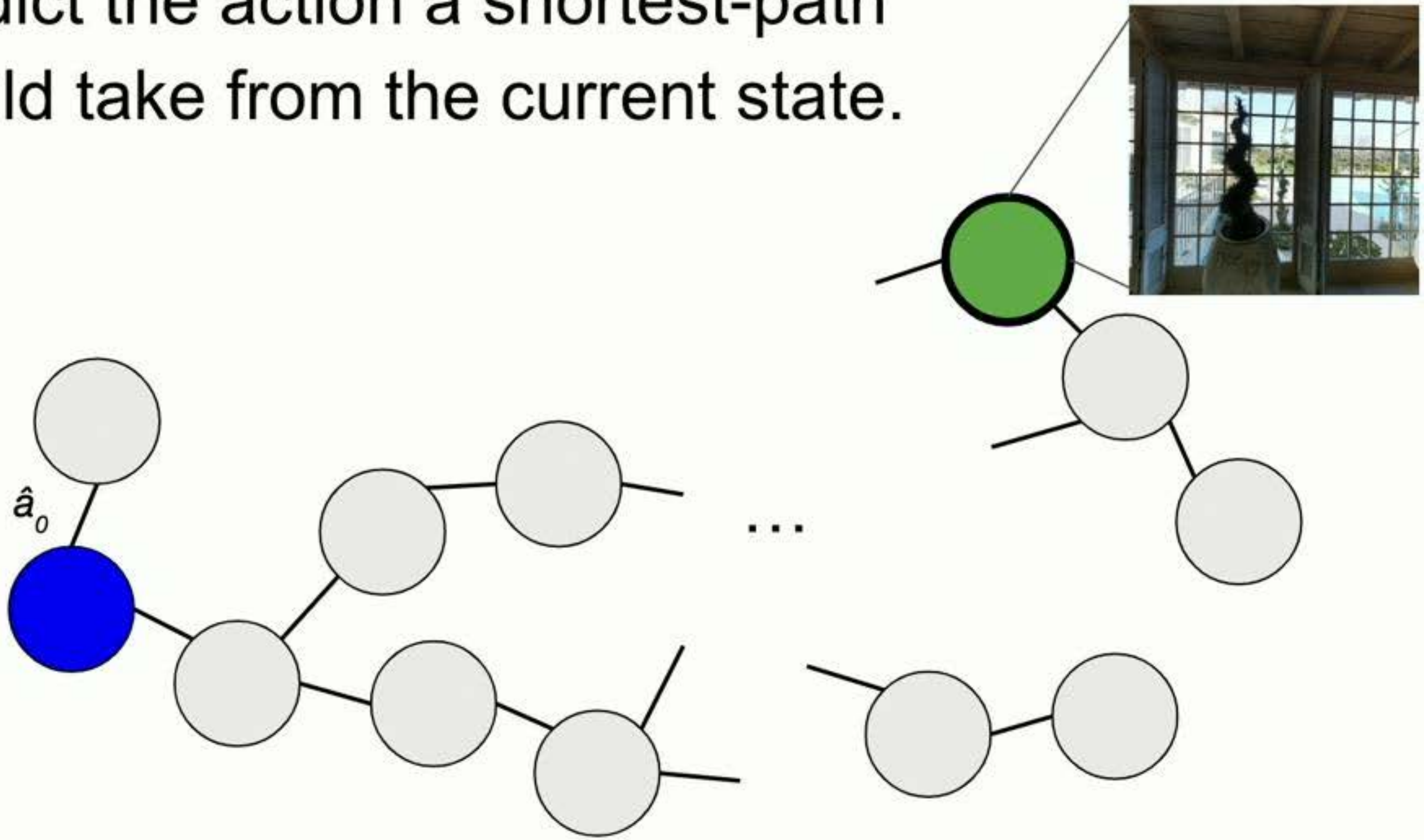


[Anderson et al., CVPR'18]

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**	**Question + Frame**	**Question + Frame**
	Room-2-Room [Anderson et al., CVPR'18]	Embodied QA [Das et al., CVPR'18]	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

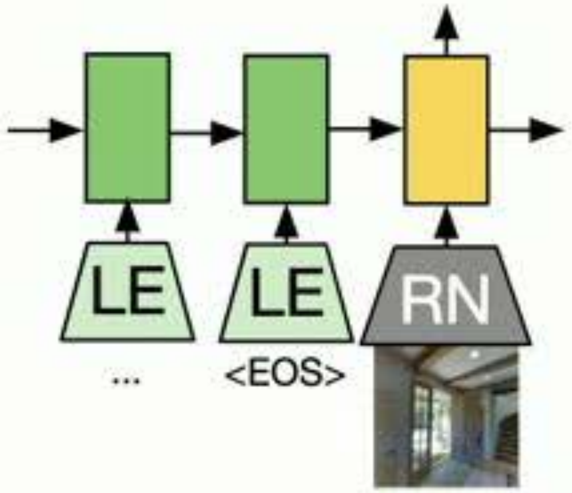
Navigation Actions + Answer Action

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."*



t₁

t₂

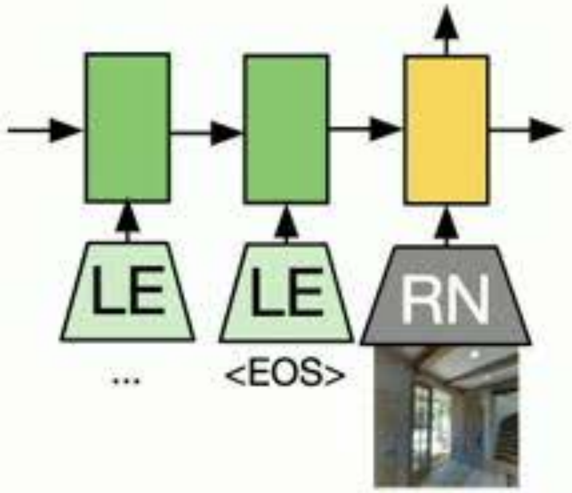
t₃

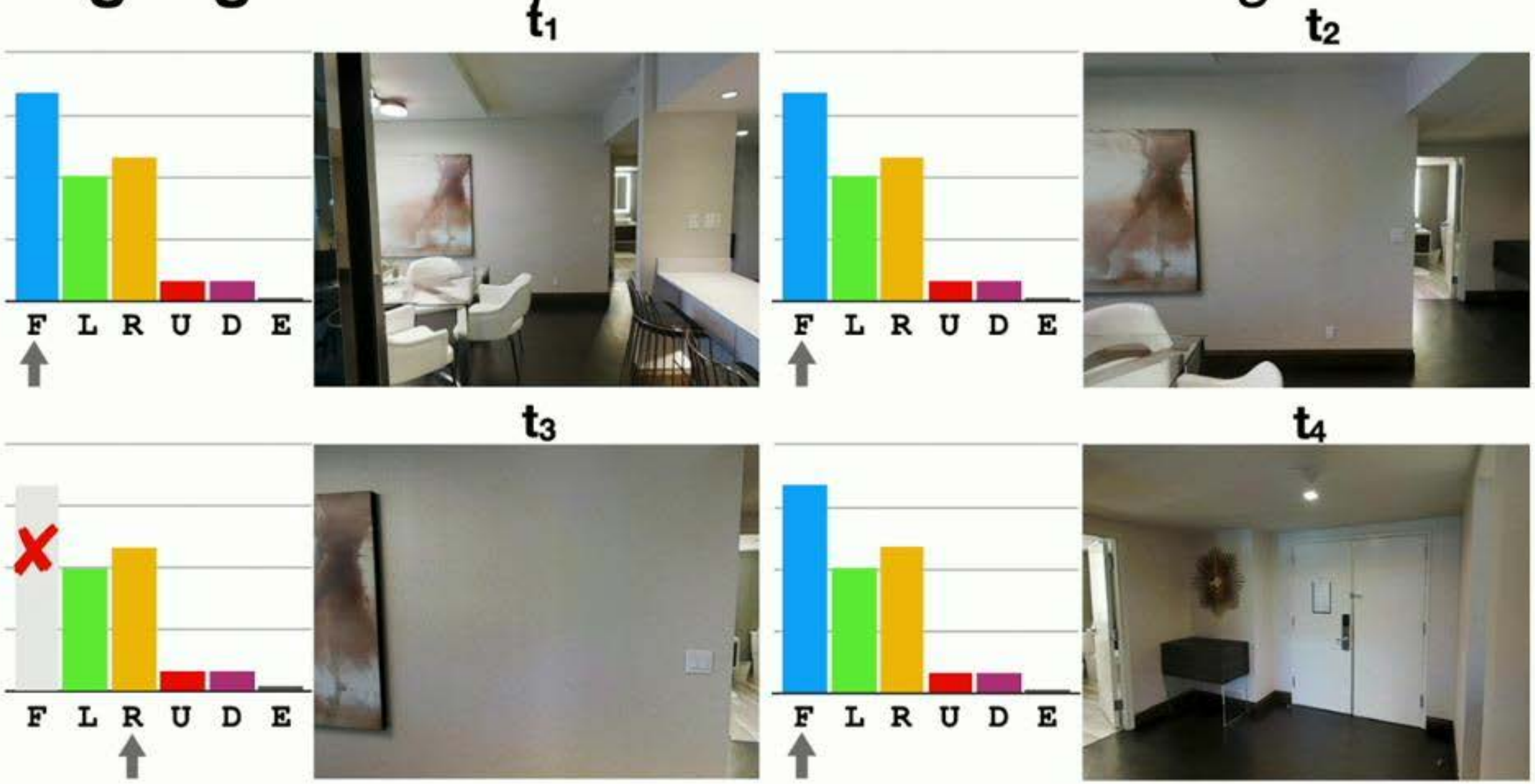
t₄

Actions: **F**orward, turn **L**eft & **R**ight, tilt **U**p & **D**own, **E**nd

Inputs and Outputs

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

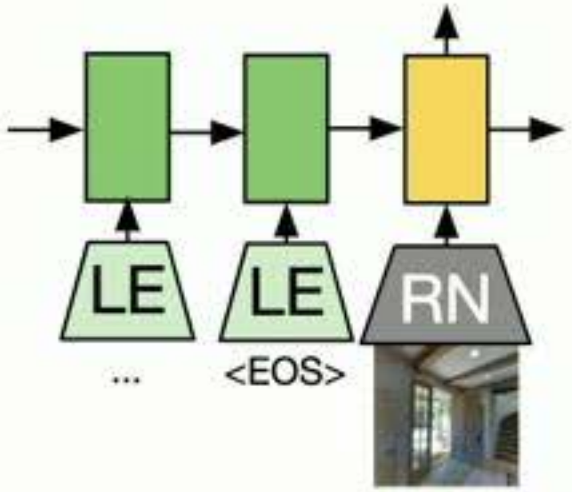


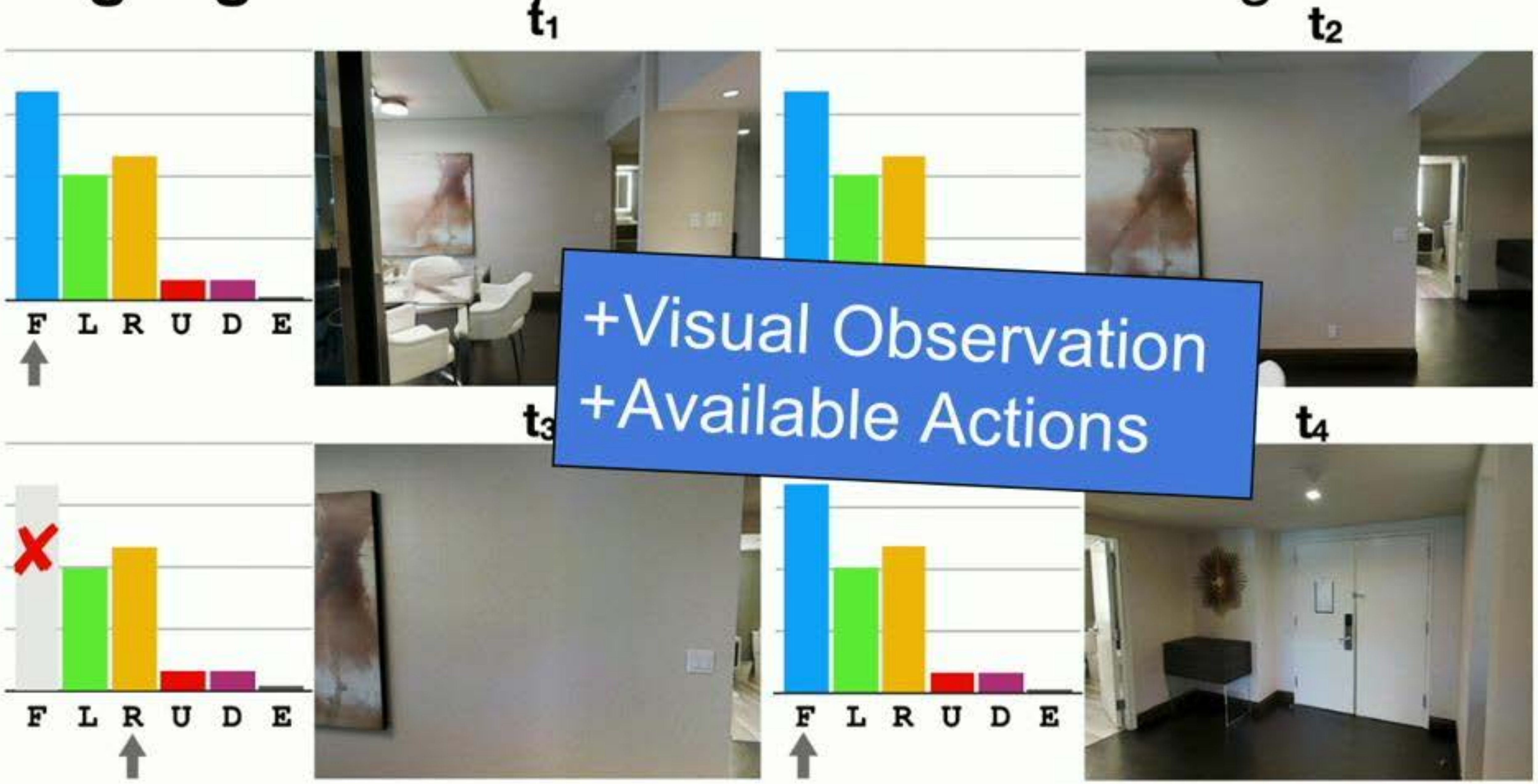


Actions: **F**orward, turn **L**eft & **R**ight, tilt **U**p & **D**own, **E**nd

Inputs and Outputs

Language: *"Walk past the bar and turn right."*



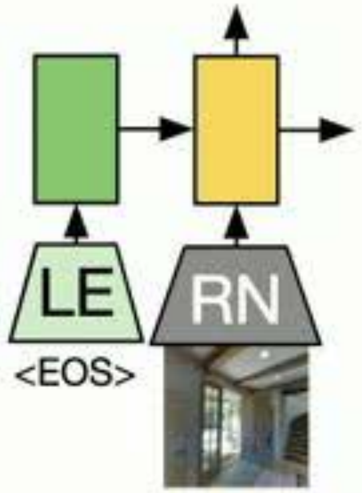


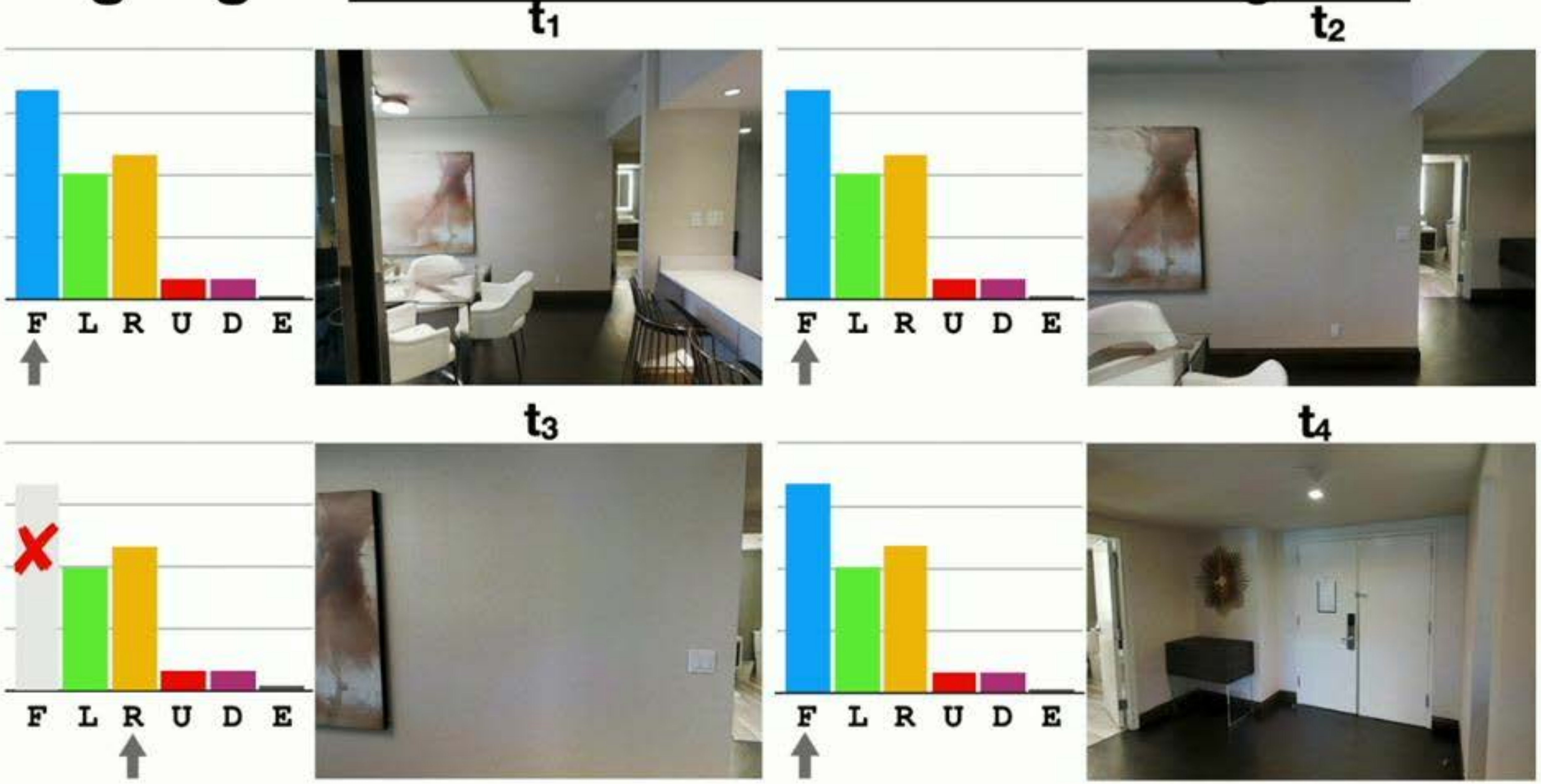
+Visual Observation
+Available Actions

Actions: **F**orward, turn **L**eft & **R**ight, tilt **U**p & **D**own, **E**nd

Unimodal Ablation - *Vision Only*

Language: [redacted]

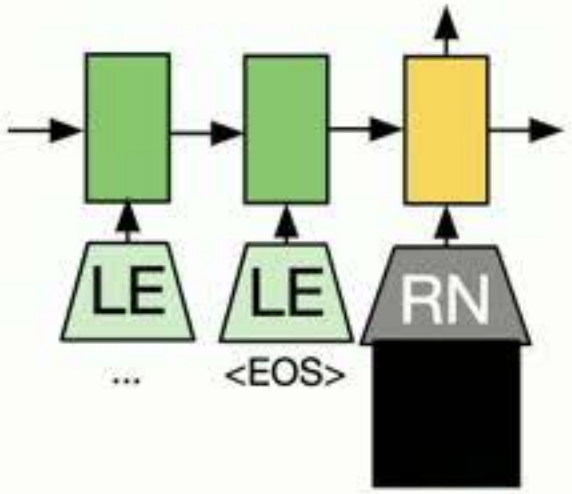


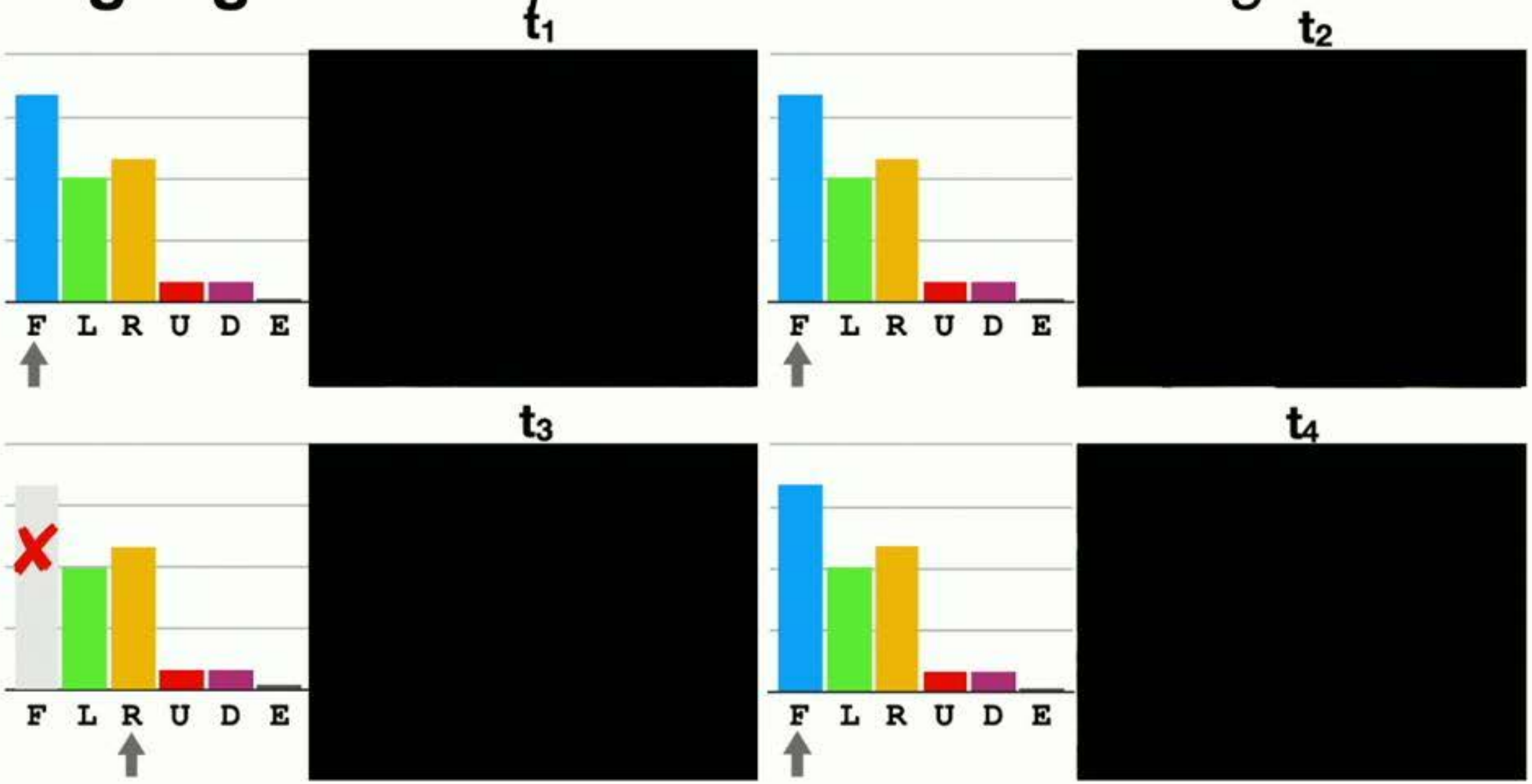


Actions: **F**orward, turn **L**eft & **R**ight, tilt **U**p & **D**own, **E**nd

Unimodal Ablation - *Language Only*

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

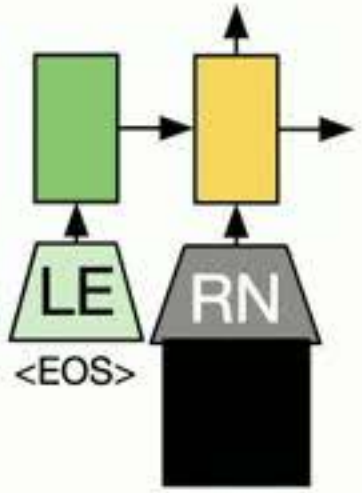




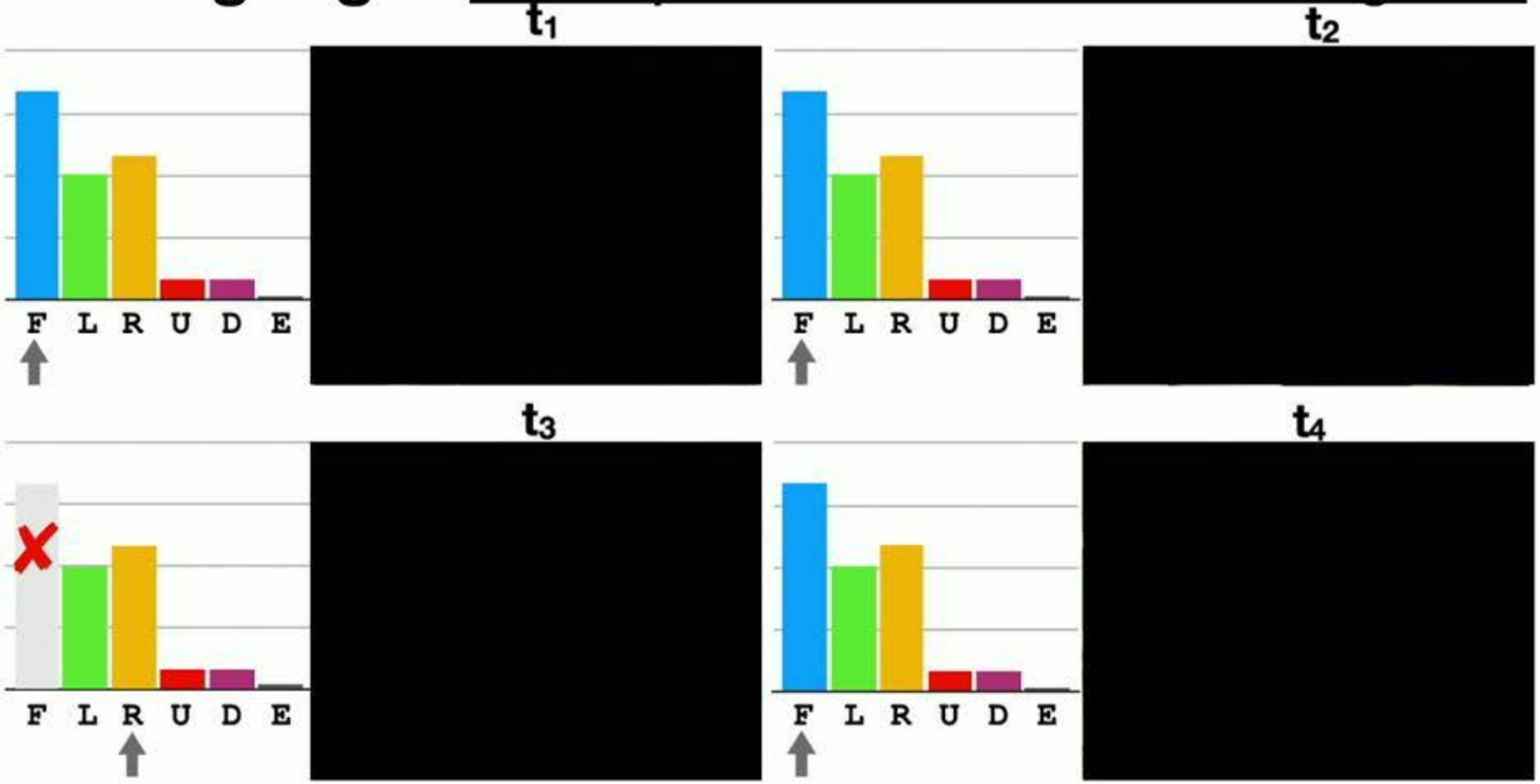
Actions: **F**orward, turn **L**eft & **R**ight, tilt **U**p & **D**own, **E**nd

Unimodal Ablation - *Action Only*

Language: [redacted]



[redacted]



[redacted]

[redacted]

[redacted]

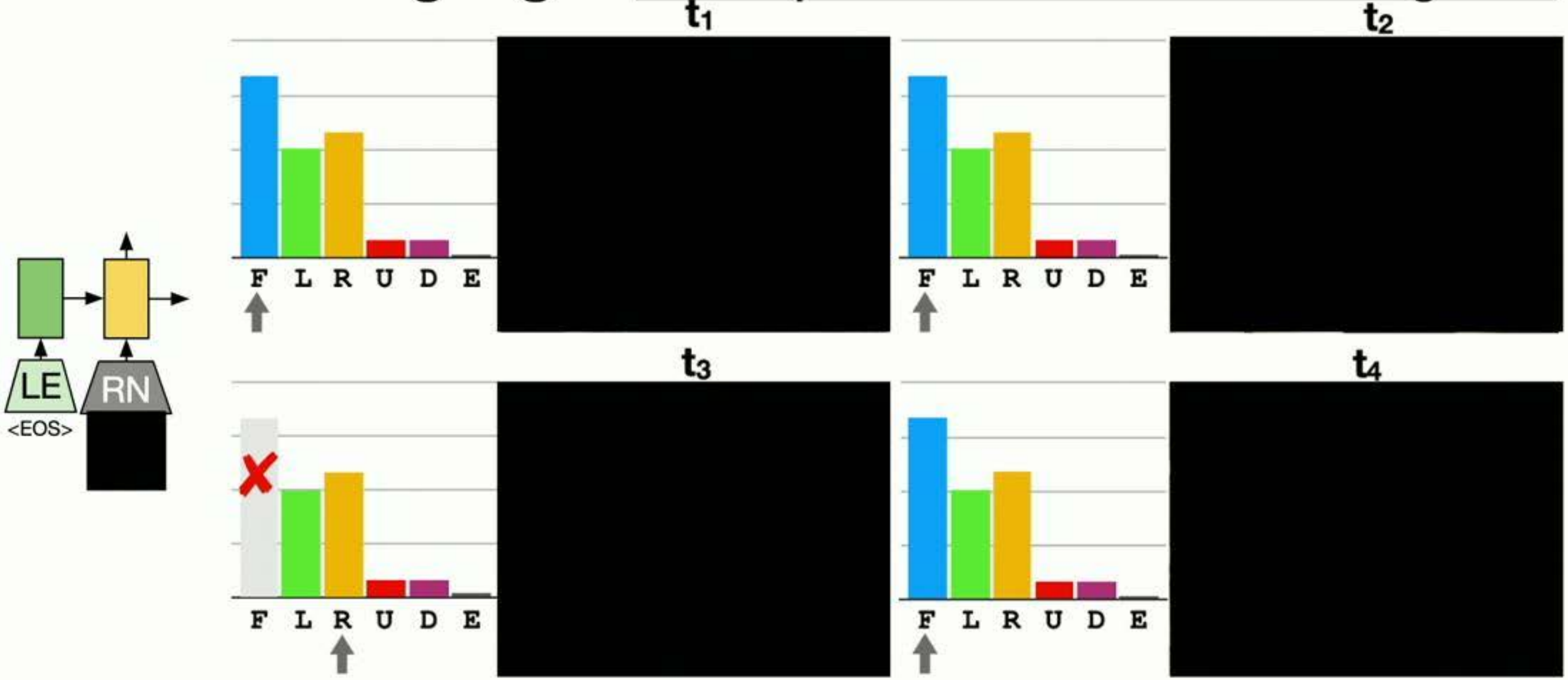
[redacted]

Actions: **F**orward, turn **L**eft & **R**ight, tilt **U**p & **D**own, **E**nd

Unimodal Model Ablations

Unimodal Ablation - *Action Only*

Language: [redacted]



[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Actions: **F**orward, turn **L**eft & **R**ight, tilt **U**p & **D**own, **E**nd

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]

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.**		

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

	m [Anders ’18]	Embodied QA [Das et al., CVPR’18]	Interactive QA [Gordon et al., CVPR’18]
Beats baseline	**Vision-, langu and action-only mo els.**	**Vision-, language-, and action-only models.**	**Vision-only model.**
Beats initial model!	**Language-only model.**	**Vision-only model.**	**(none)**

Unimodal Model Ablations

Photorealistic Environments

Low-level Language

m
[Anders ’18]

Embodied QA
[Das et al., CVPR’18]

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

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

Unimodal Model Ablations

Photorealistic Environments

Low-level Language

m
[Anders ’18]

Simple Visual Environments

[Das e

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

Beats baseline	**Vision-, langu and action-only mo els.**	**Vision-, langua and action-only models.**	**Vision-only model.**
Beats initial model!	**Language-only model.**	**Vision-only model.**	**(none)**

Unimodal Model Ablations

Photorealistic Environments

Low-level Language

m
[Anders '18]

High-level Language

Simple Visual Environments

[Das e

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

Beats baseline	**Vision-, langu and action-only models.**	**Vision-, langua and action-only models.**	**Vision-only model.**
Beats initial model!	**Language-only model.**	**Vision-only model.**	**(none)**

Unimodal Model Ablations

Photorealistic Environments

Low-level Language

m

[Anders '18]

High-level Language

Simple Visual Environments

[Das e

Well-balanced

Interacti

[Gordon et al., CV 18]

Beats baseline	**Vision-, langu and action-only mo els.**	**Vision-, langua and action-only models.**	**Vision-only model.**
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.

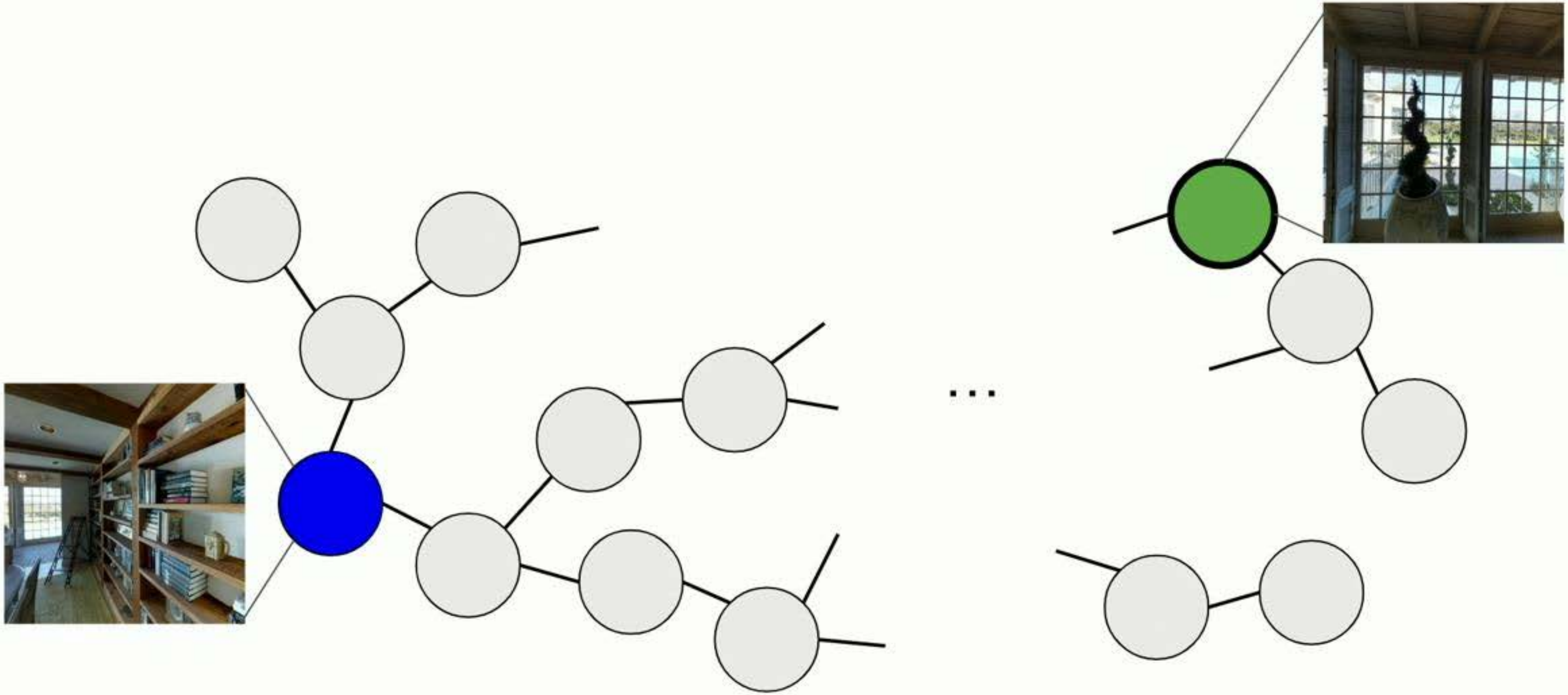
Lessons

- 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

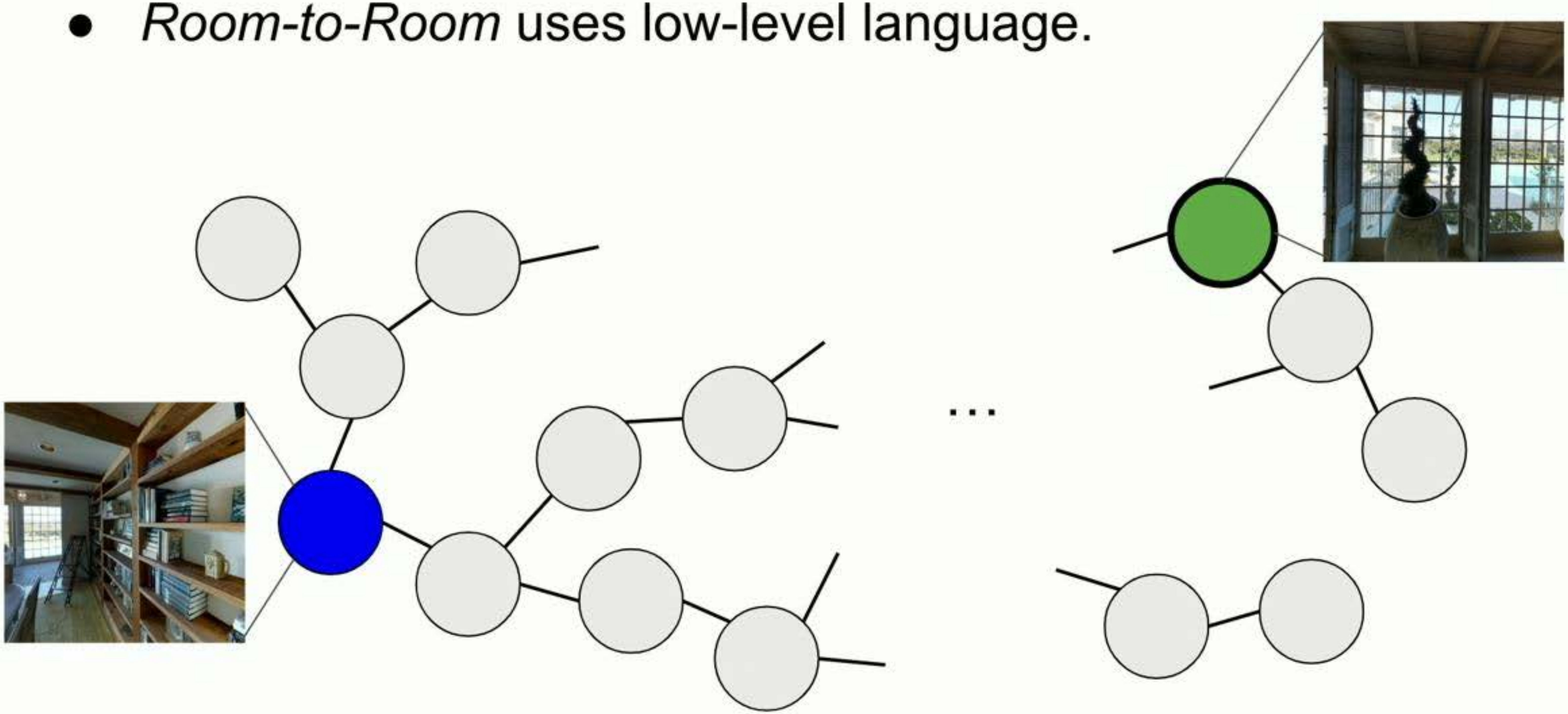
- Language grounding in visual environments
    - For navigation
    - Unimodal bias [Thomason et al., NAACL’19]
- Vision-and-Dialog Navigation [Thomason et al., *in sub*]
    - New dataset - CVDN
    - Navigation from dialog history
- Next steps

“Turn around and exit the library, head down the…”



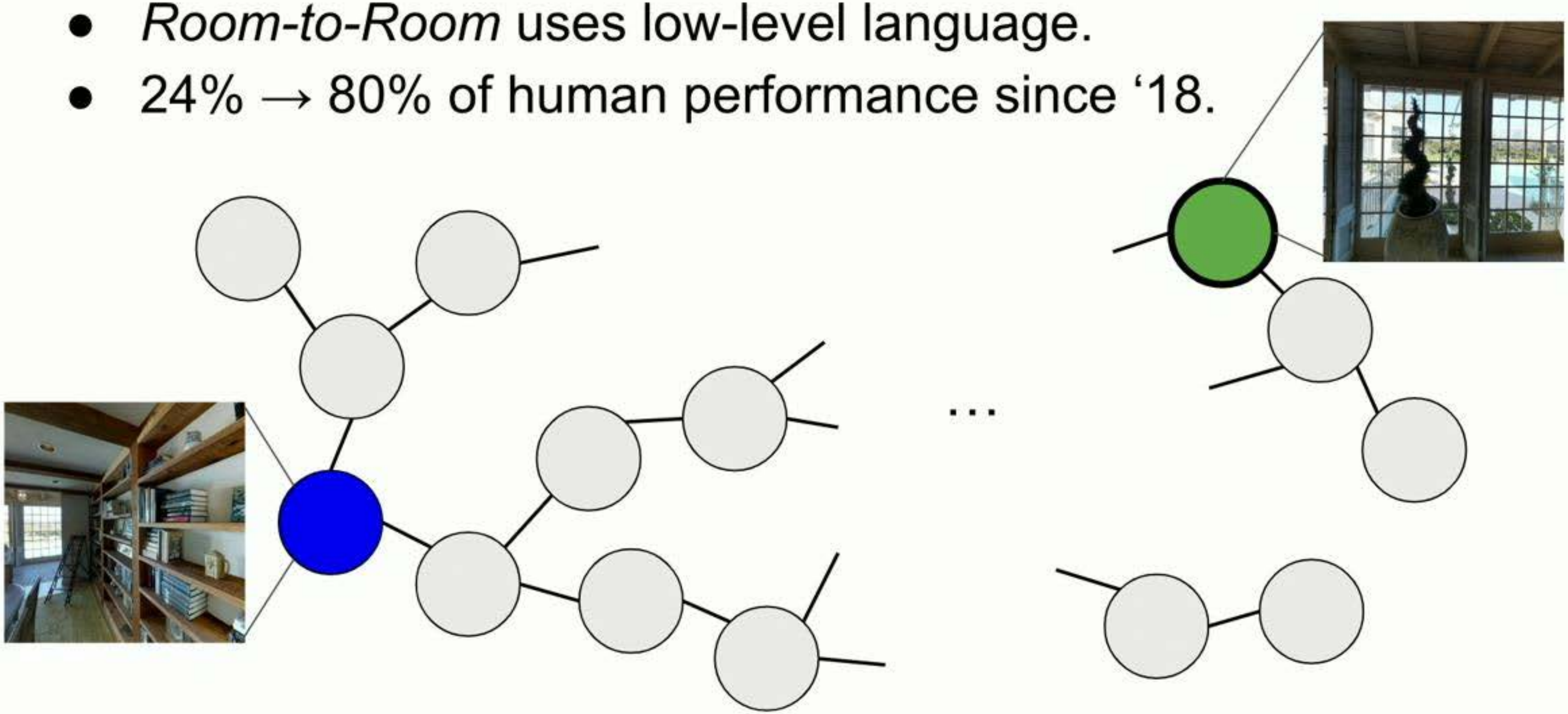
"Turn around and exit the library, head down the…"

- *Room-to-Room* uses low-level language.



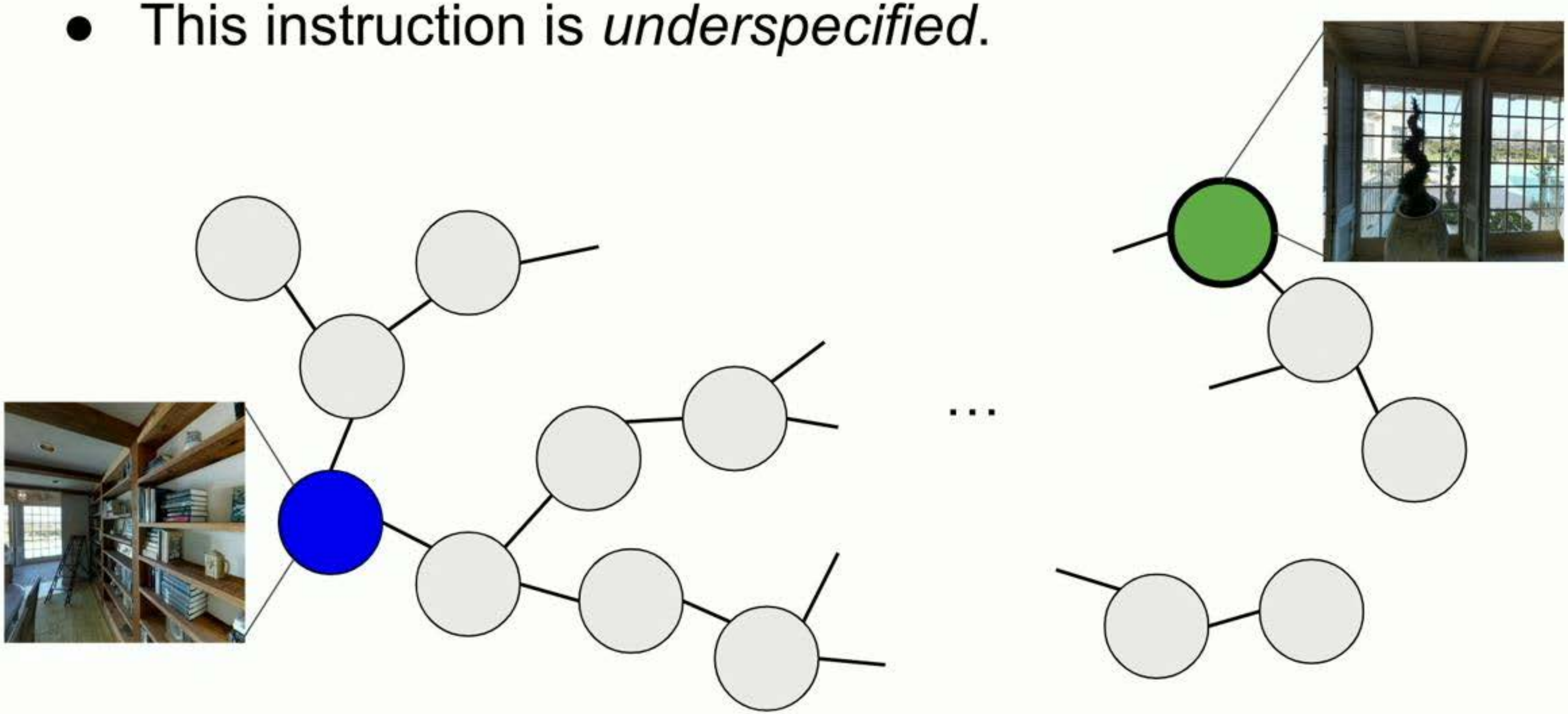
“Turn around and exit the library, head down the…”

- *Room-to-Room* uses low-level language.
- 24% → 80% of human performance since ‘18.



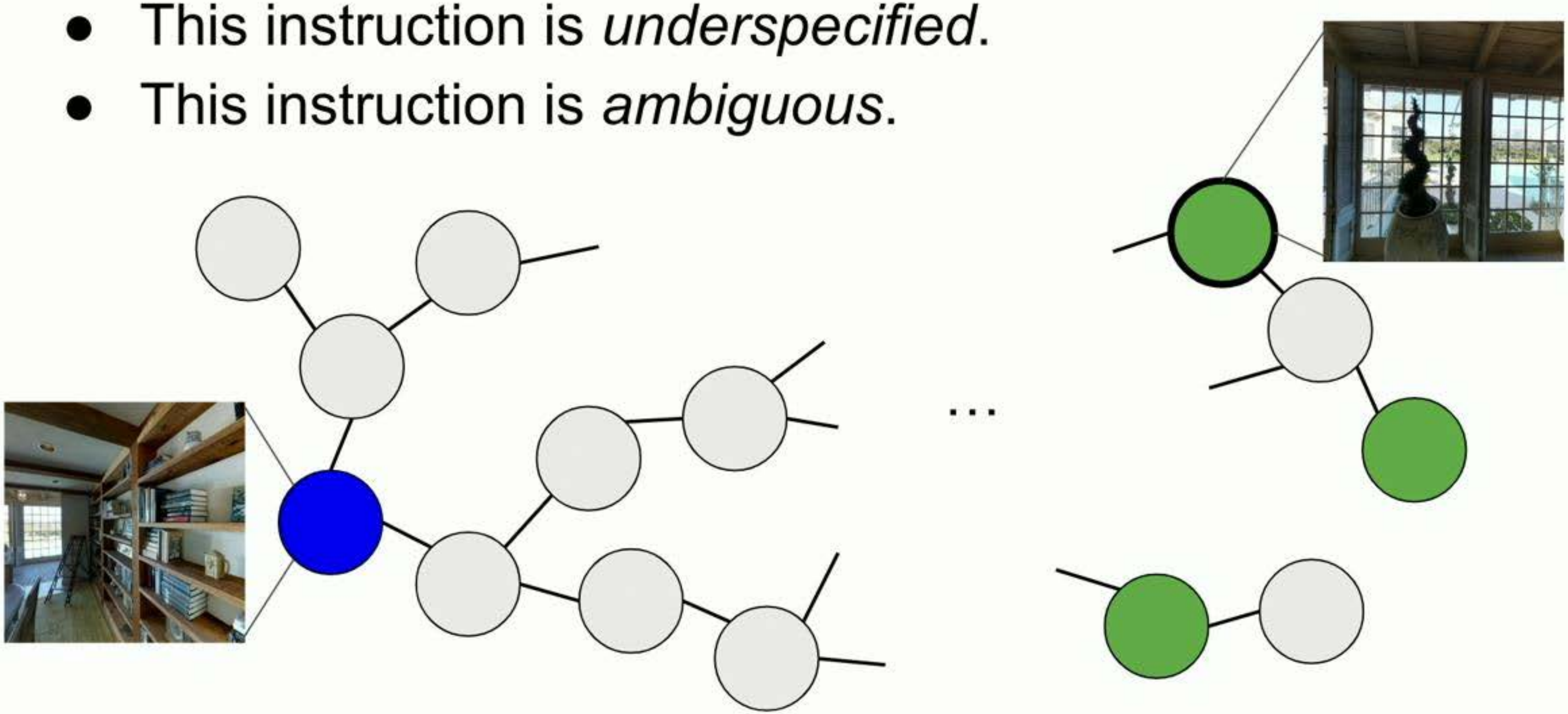
"Go to the room with a plant."

- This instruction is *underspecified*.



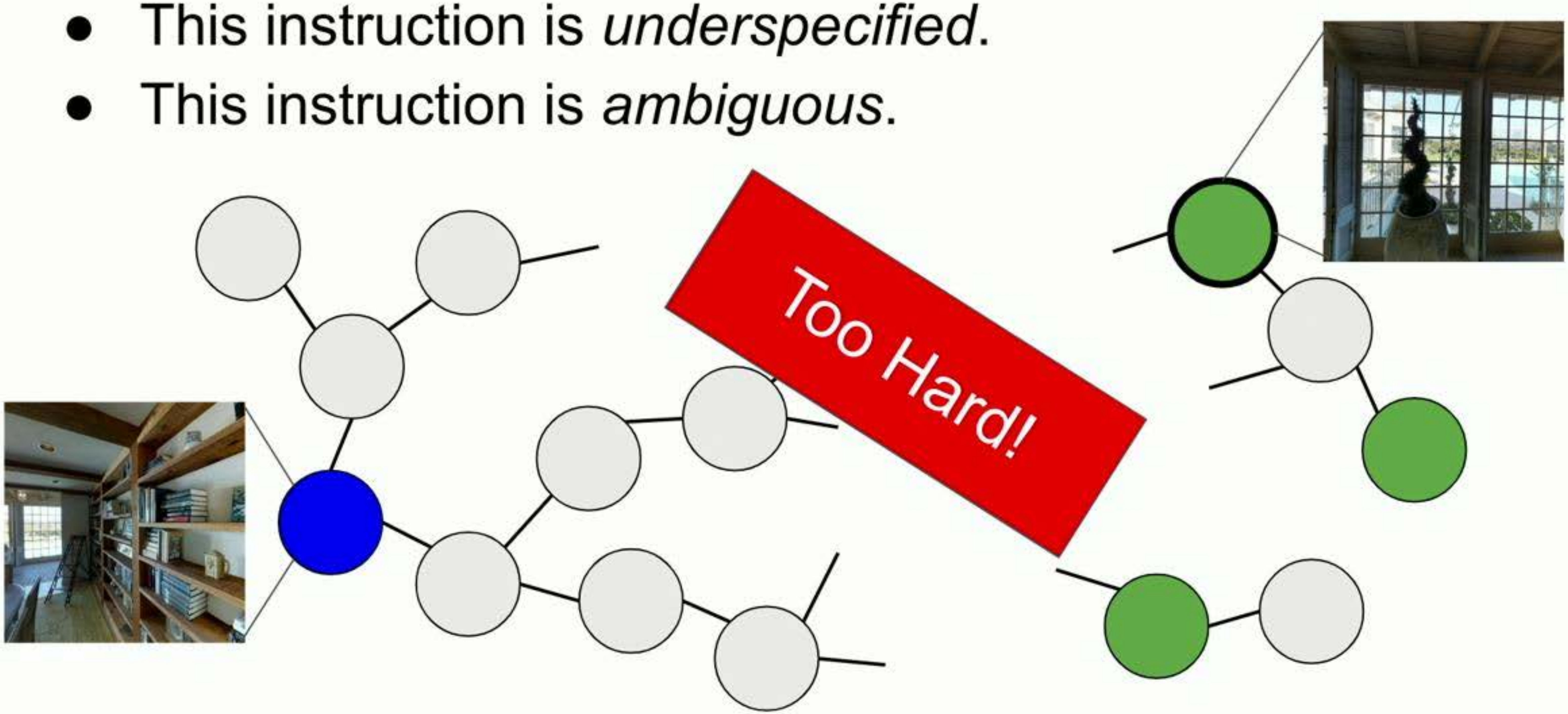
“Go to the room with a plant.”

- This instruction is *underspecified*.
- This instruction is *ambiguous*.



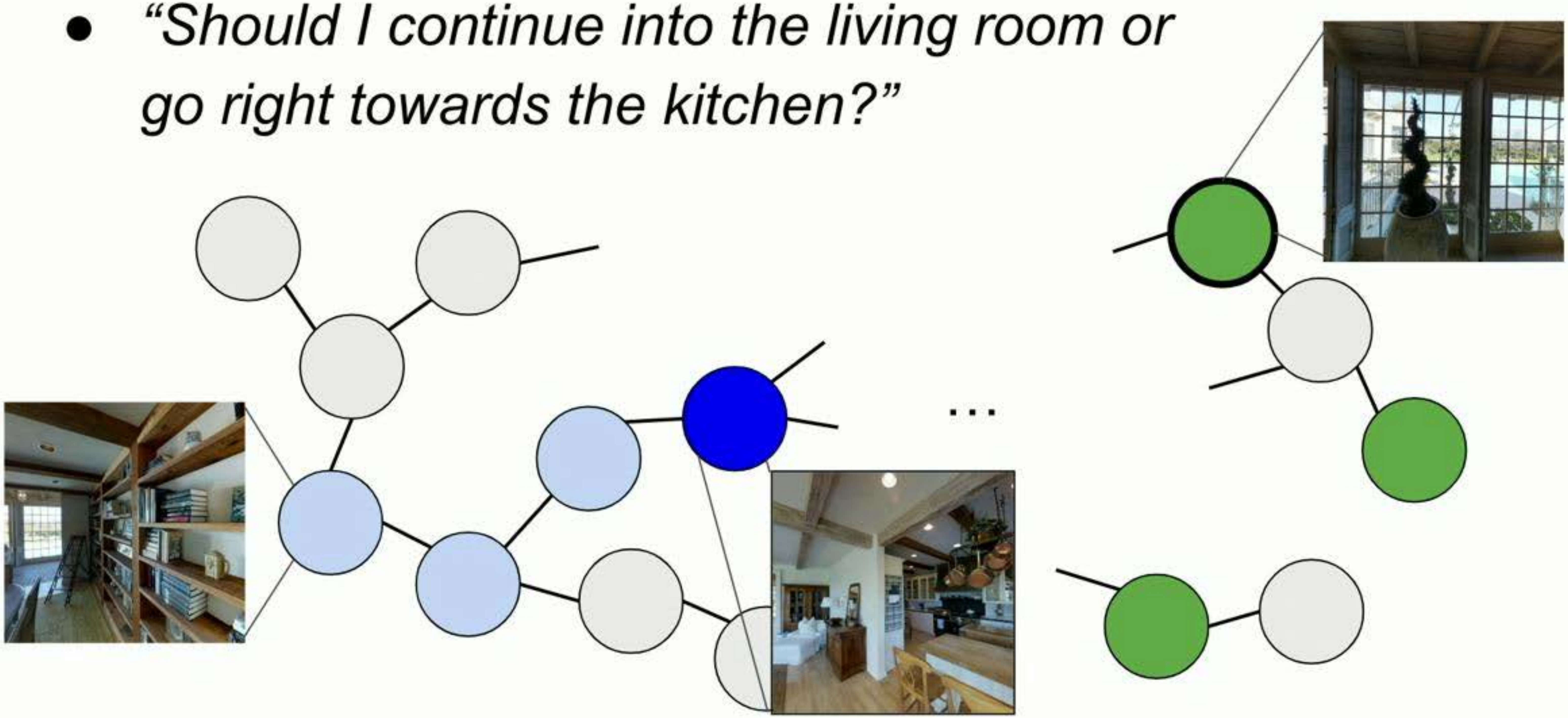
"Go to the room with a plant."

- This instruction is *underspecified*.
- This instruction is *ambiguous*.



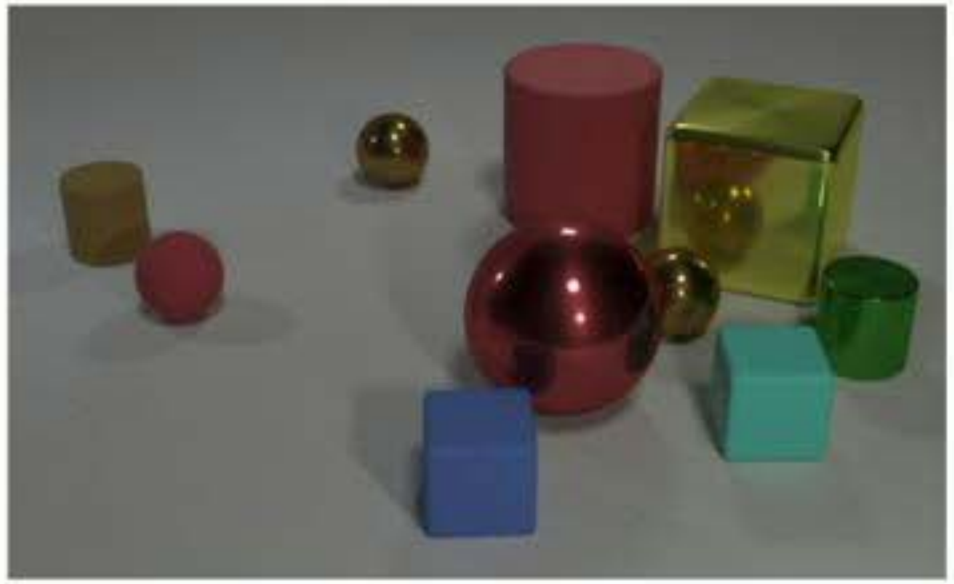
What if we could just… ask?

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

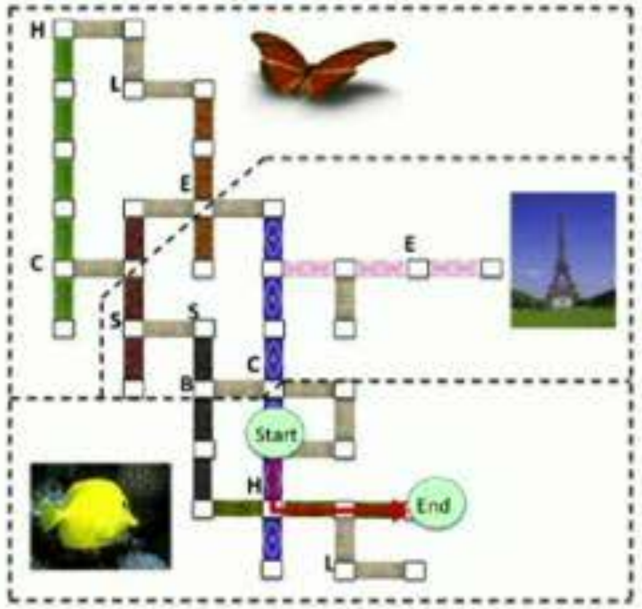


Visual Fidelity

Rendered



CLEVR
[Johnson et al., CVPR'17]



Instruction Following
[Chen and Mooney, AAAI'11]

Visual Context

Static

Dynamic

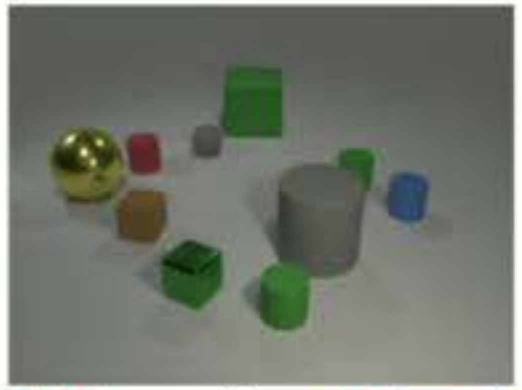
VQA
[Antol et al., CVPR'15]

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

Photorealistic

Visual Fidelity

Rendered



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

CLEVR-Dialog [Kottur et al., NAACL'19]

Visual Context

Static

Dynamic

Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

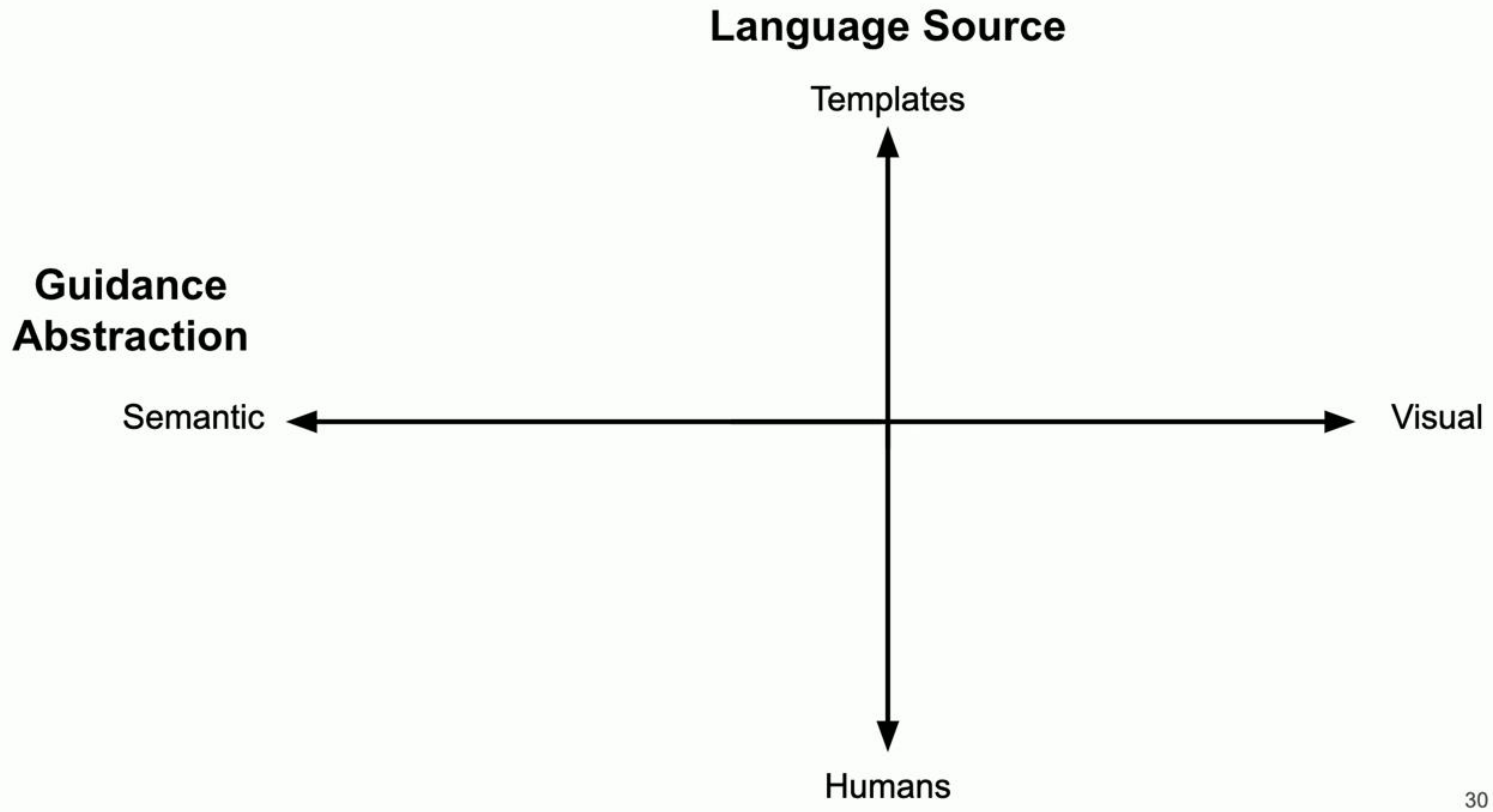
A: Playing a Wii game

Q: Is that a man to her right

A: No, it's a woman

VisDial [Das et al., CVPR'17]

Photorealistic


Language Source
Templates
Guidance
Abstraction
Semantic
Visual
Humans

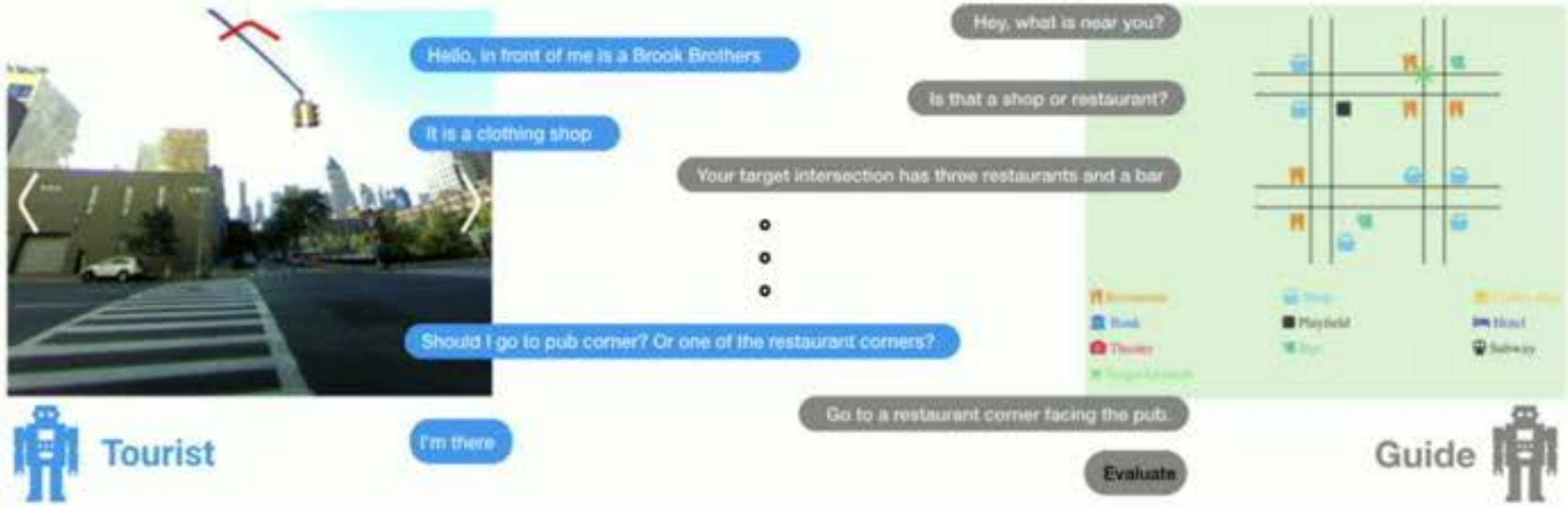
Language Source

Templates

Guidance Abstraction

Semantic

Visual



Hey, what is near you?

Hello, in front of me is a Brook Brothers

Is that a shop or restaurant?

It is a clothing shop

Your target intersection has three restaurants and a bar

Should I go to pub corner? Or one of the restaurant corners?

Go to a restaurant corner facing the pub.

I'm there

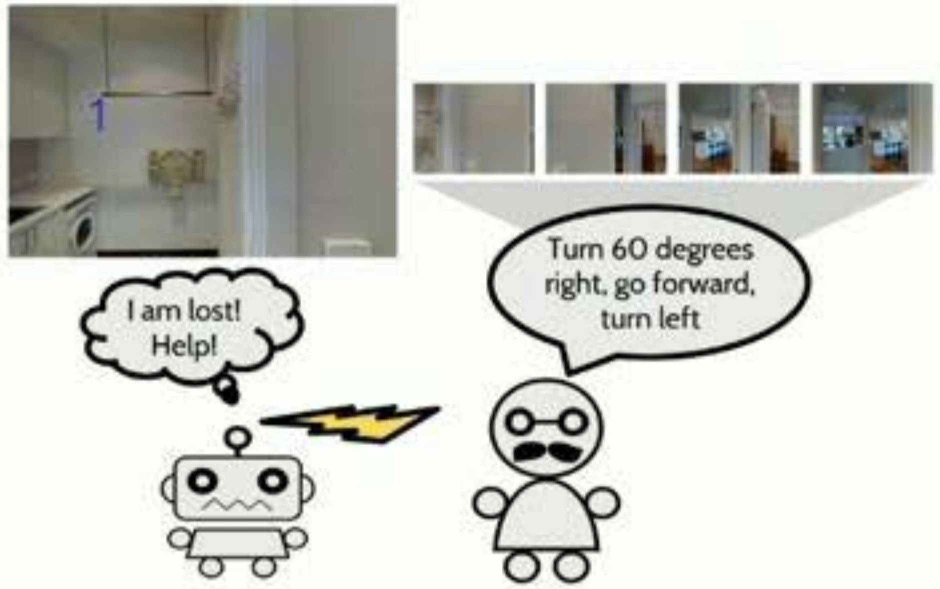
Evaluate

Humans

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

Language Source

Templates

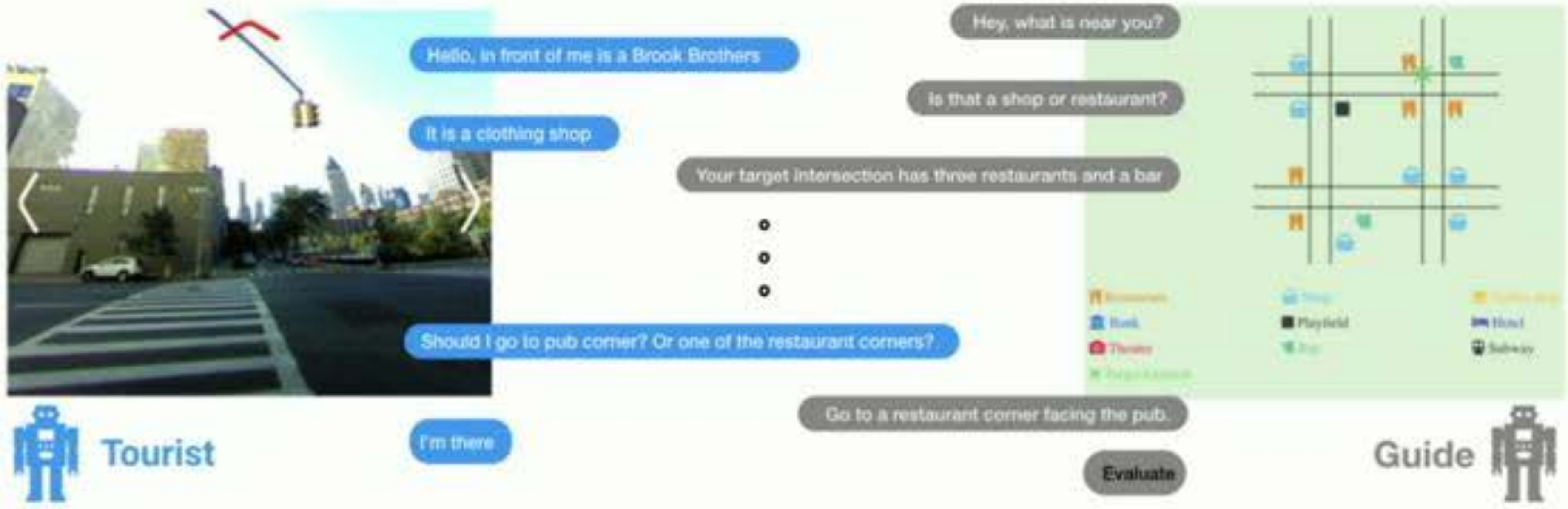


VLNA [Nguyen et al., CVPR'19]

Guidance Abstraction

Semantic

Visual



Hey, what is near you?

Hello, in front of me is a Brook Brothers

Is that a shop or restaurant?

It is a clothing shop

Your target intersection has three restaurants and a bar

Should I go to pub corner? Or one of the restaurant corners?

Go to a restaurant corner facing the pub.

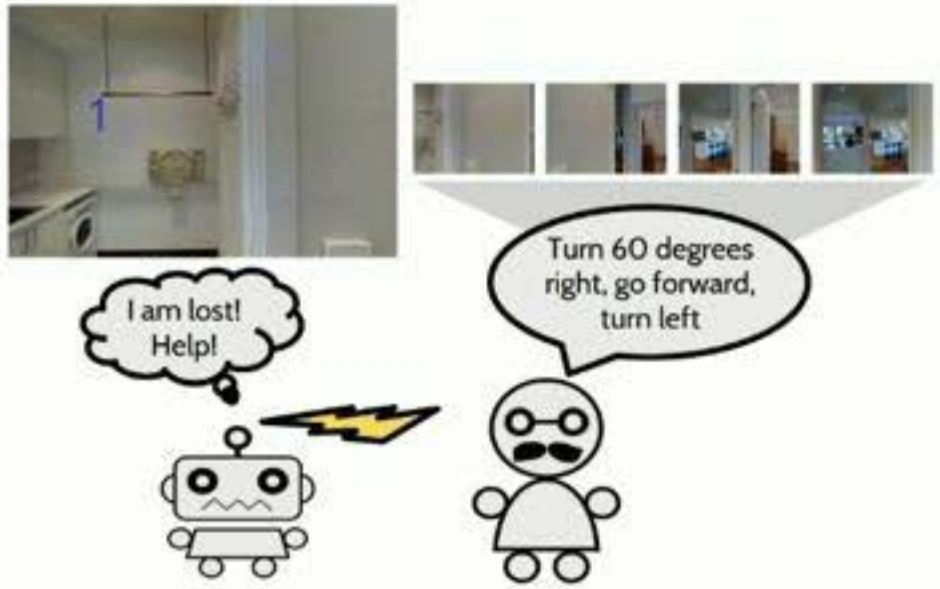
I'm there

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

Humans

Language Source

Templates

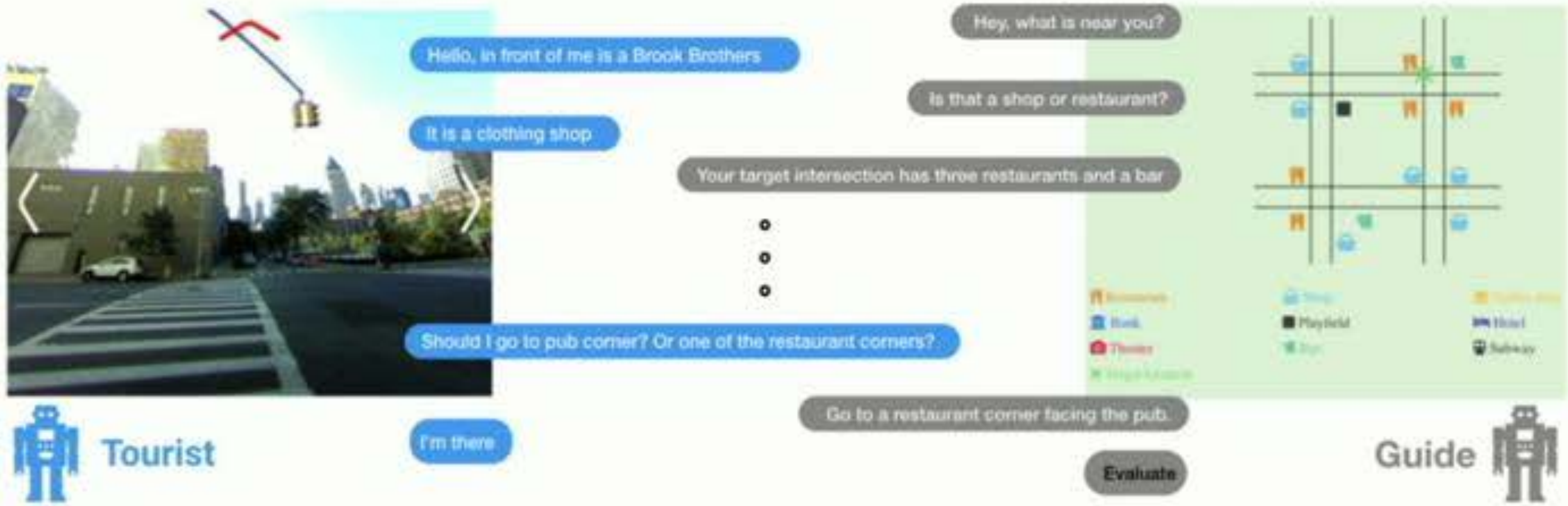


VLNA [Nguyen et al., CVPR'19]

Guidance Abstraction

Semantic

Visual



Hey, what is near you?

Hello, in front of me is a Brook Brothers

Is that a shop or restaurant?

It is a clothing shop

Your target intersection has three restaurants and a bar

Should I go to pub corner? Or one of the restaurant corners?

Go to a restaurant corner facing the pub.

I'm there

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

Humans

Vision-and-Dialog Navigation

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

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 navig
  - Unimodal L’19]
- Vision-and-D et al., *in sub*]
  - New data
    - Cooperative Vision-and-Dialog Navigation (CVDN)
    - 2k human-human dialogs
  - Navigation from dialog history
- Next steps

Initial model adapted from R2R does not exhibit unimodal overfitting.

Visible to both Navigator and Oracle

Hint: The goal room contains a *mat*.

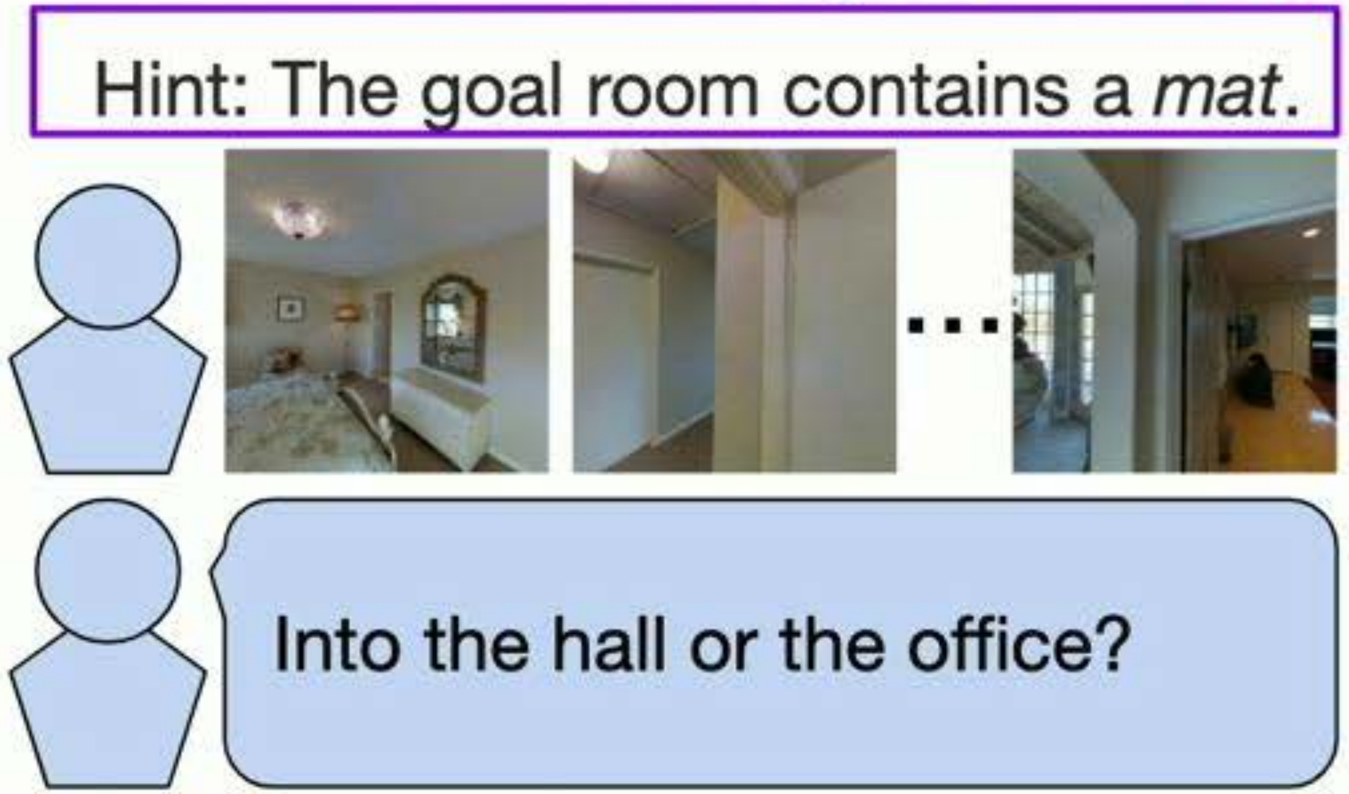
-- this target object is present in at least two rooms, but only one is correct.

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*.

Into the hall or the office?

-- 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 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.

Into the hall or the office?

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

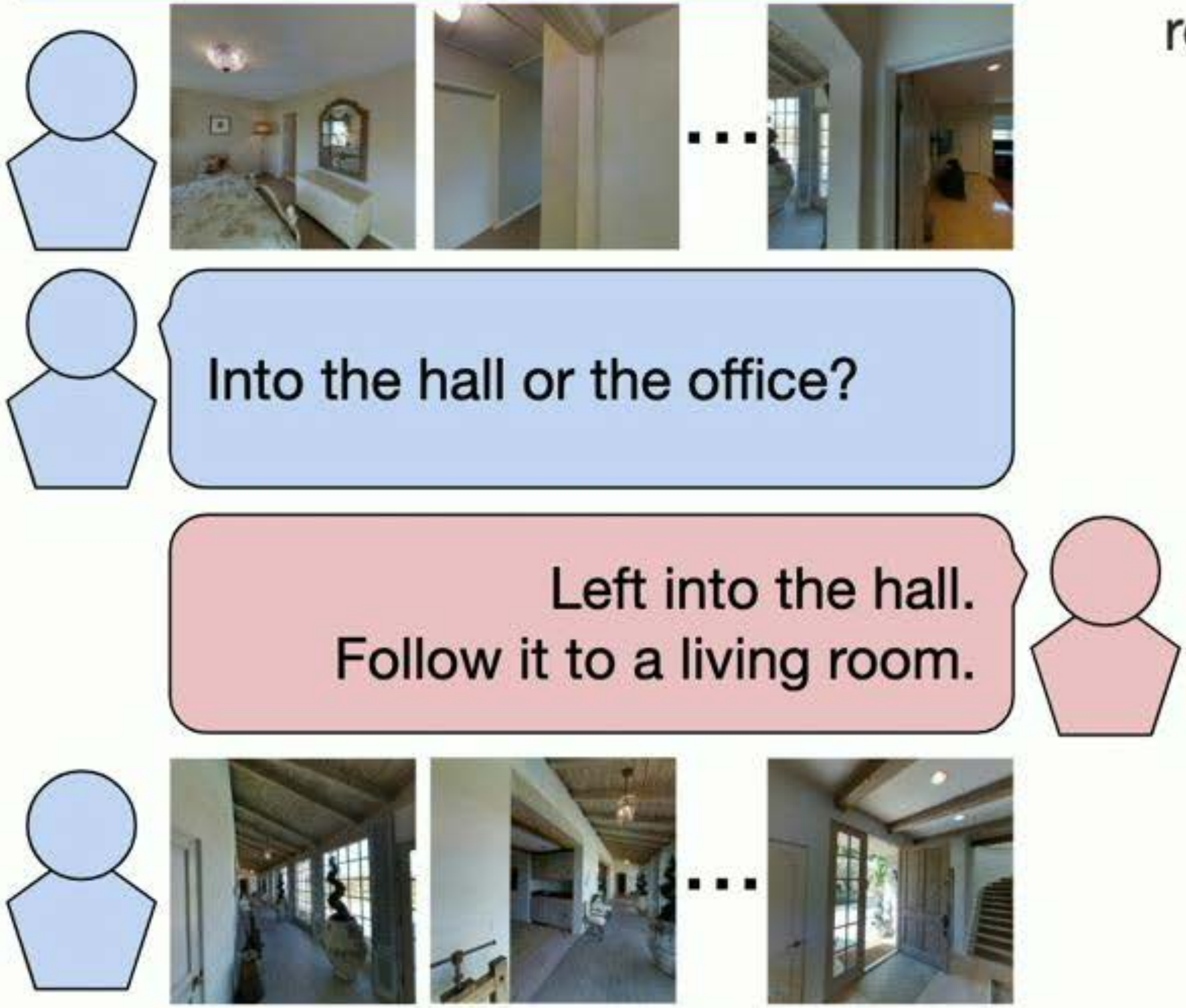
Visible Only to the Oracle

-- 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.



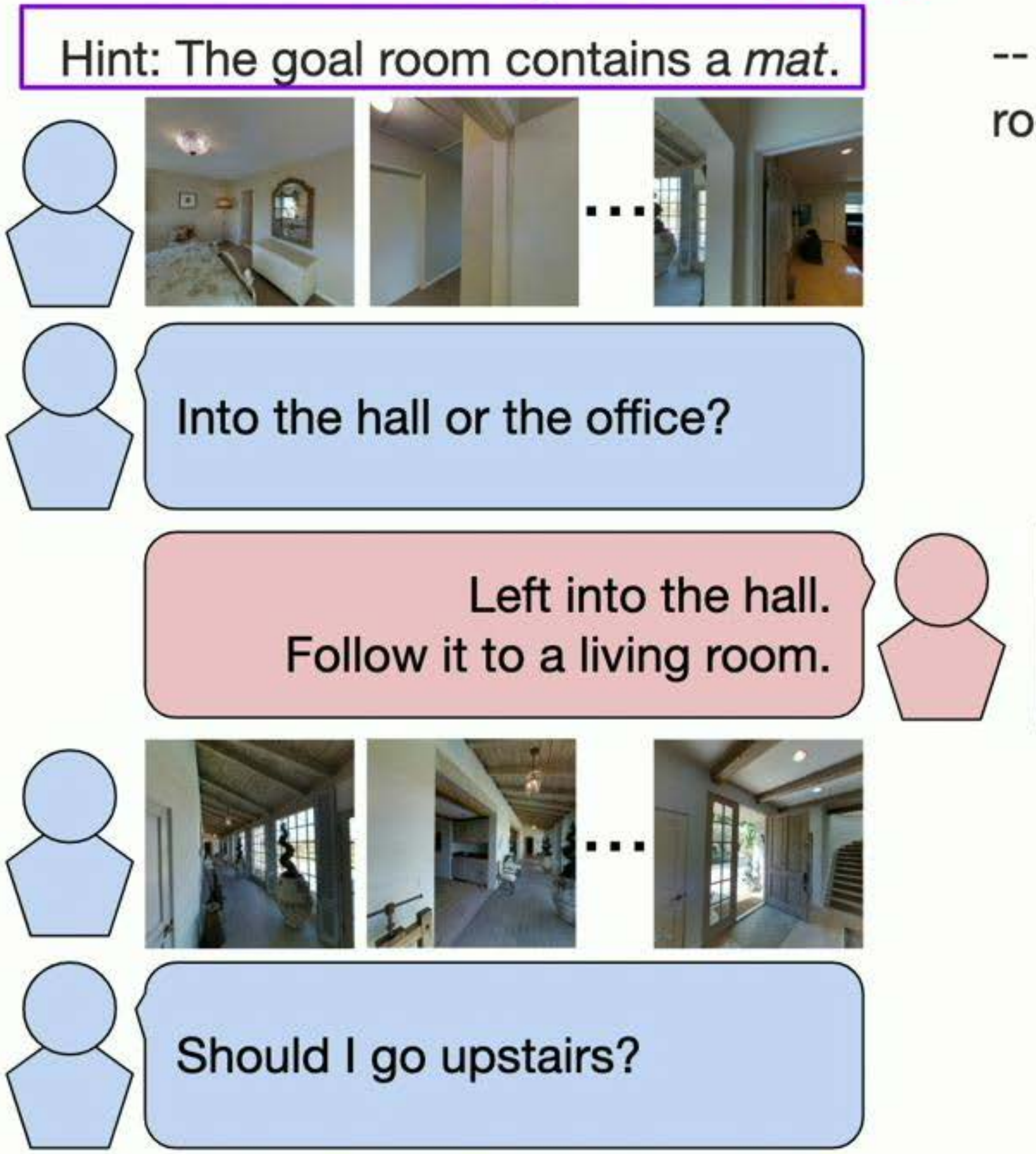
Into the hall or the office?

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

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*.

Into the hall or the office?

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

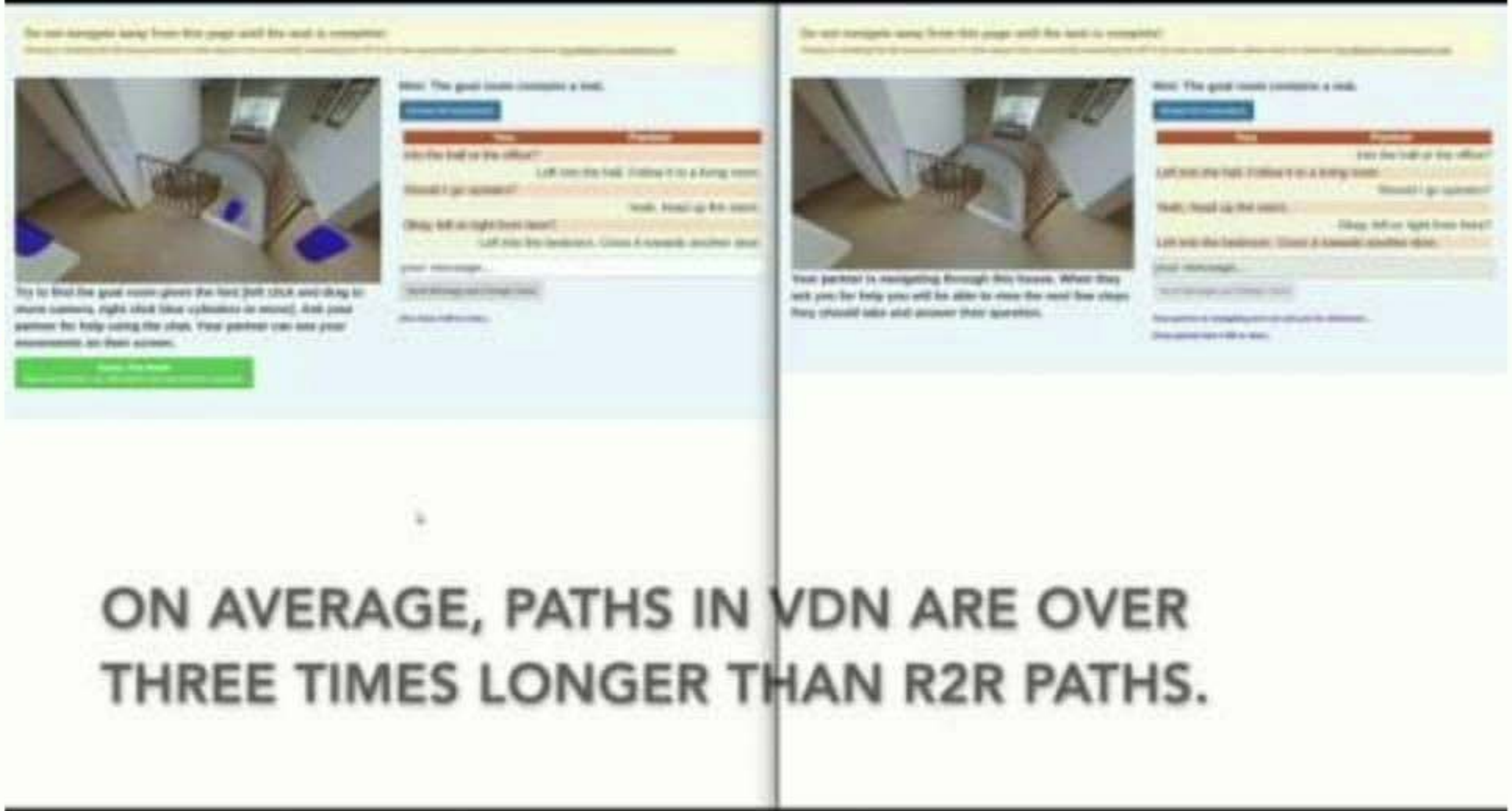
Should I go upstairs?

-- 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.

Navigator View



Do not navigate away from this page until the task is complete!

Hint: The goal room contains a bed.

You	Partner
Into the hall or the office?	
	Left into the hall. Follow it to a living room
Should I go upstairs?	
	Yeah, head up the stairs
Okay, left or right from here?	
	Left into the bedroom. Cross it towards another door

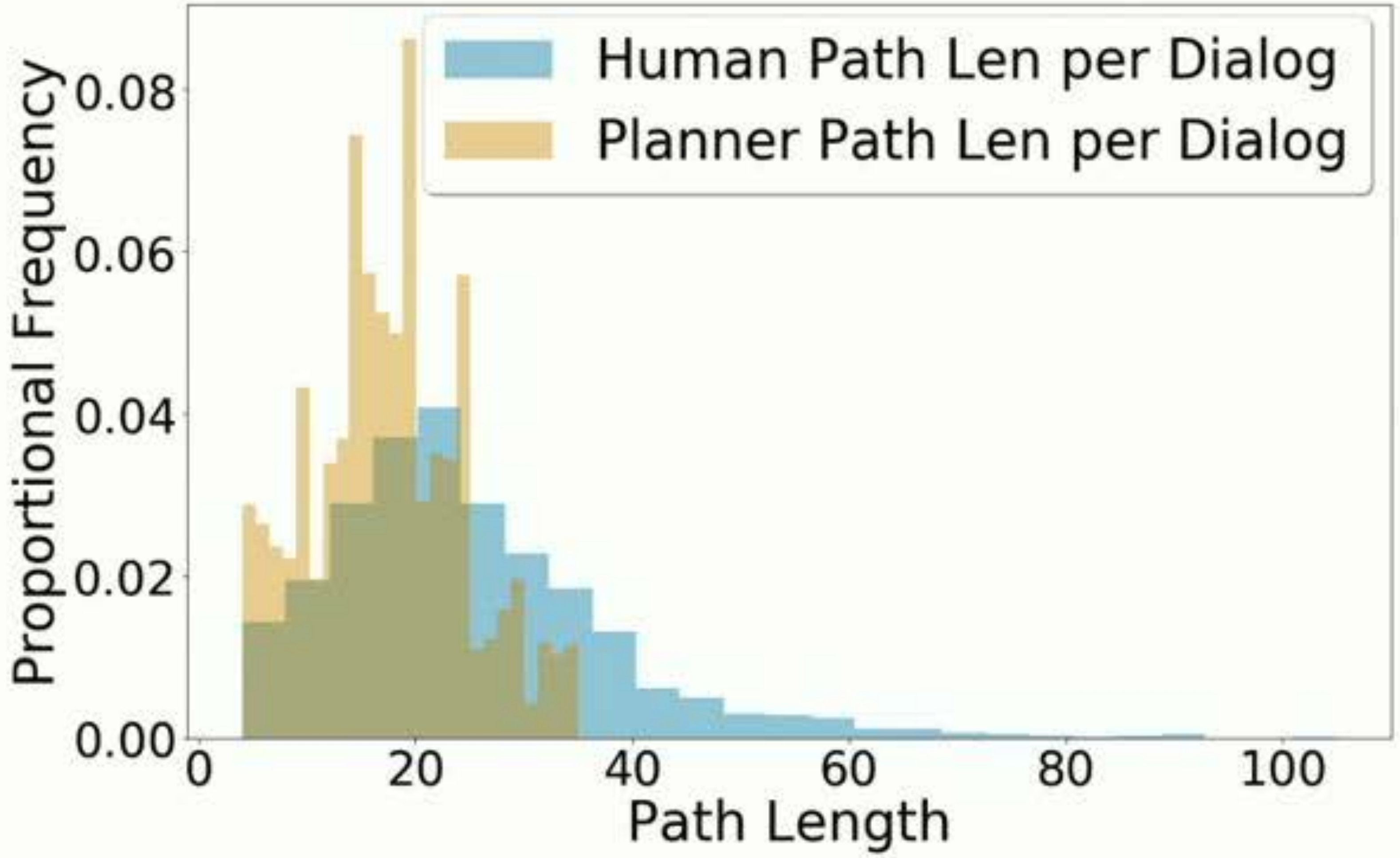
Try to find the goal room given the hint (left click and drag to move camera, right click blue cylinders to move). Ask your partner for help using the chat. Your partner can see your movements on their screen.

Your partner is navigating through this house. When they ask you for help you will be able to view the next few steps they should take and answer their question.

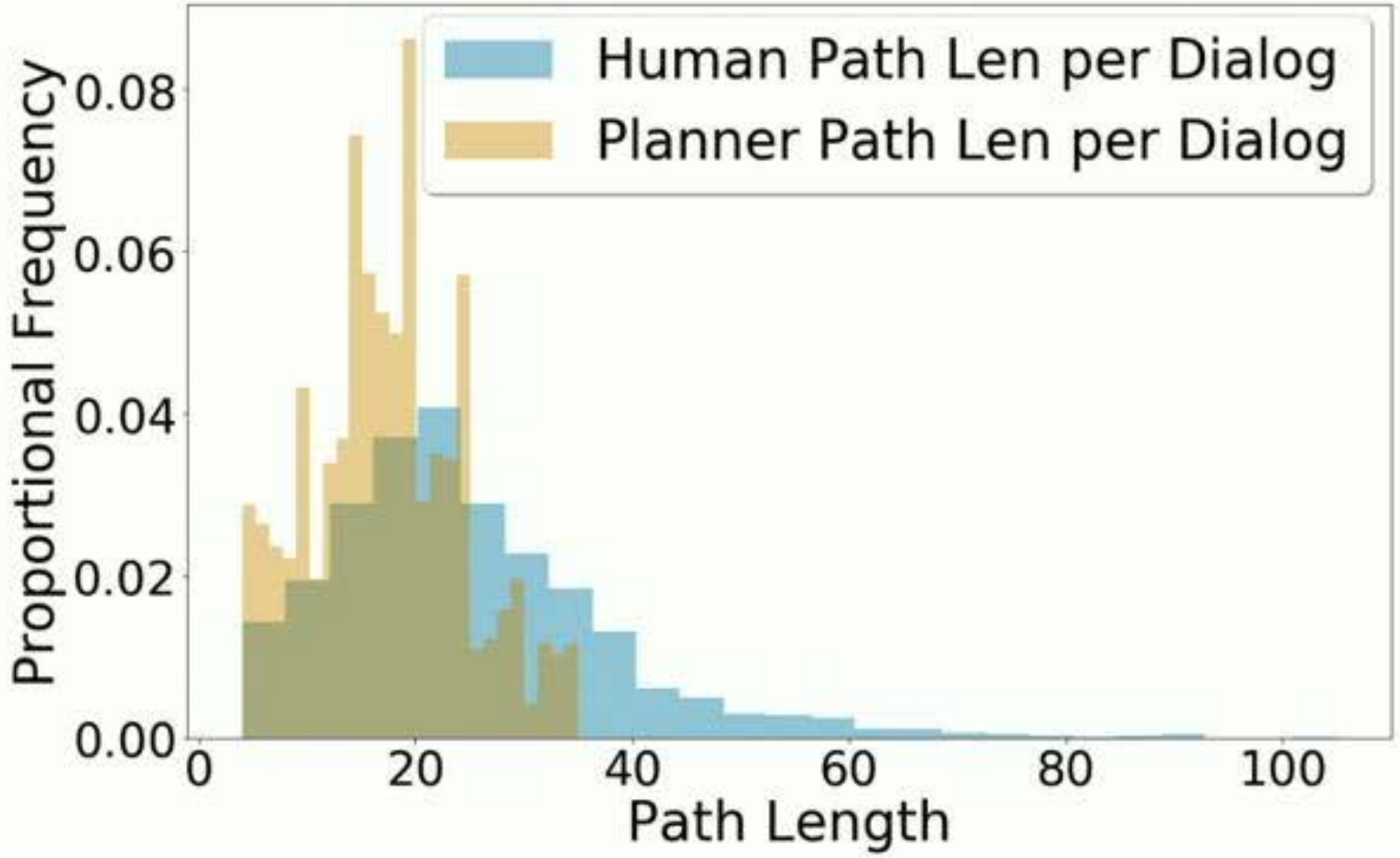
ON AVERAGE, PATHS IN VDN ARE OVER THREE TIMES LONGER THAN R2R PATHS.

Oracle View

Dialog Leads to Long Paths and Rich Language

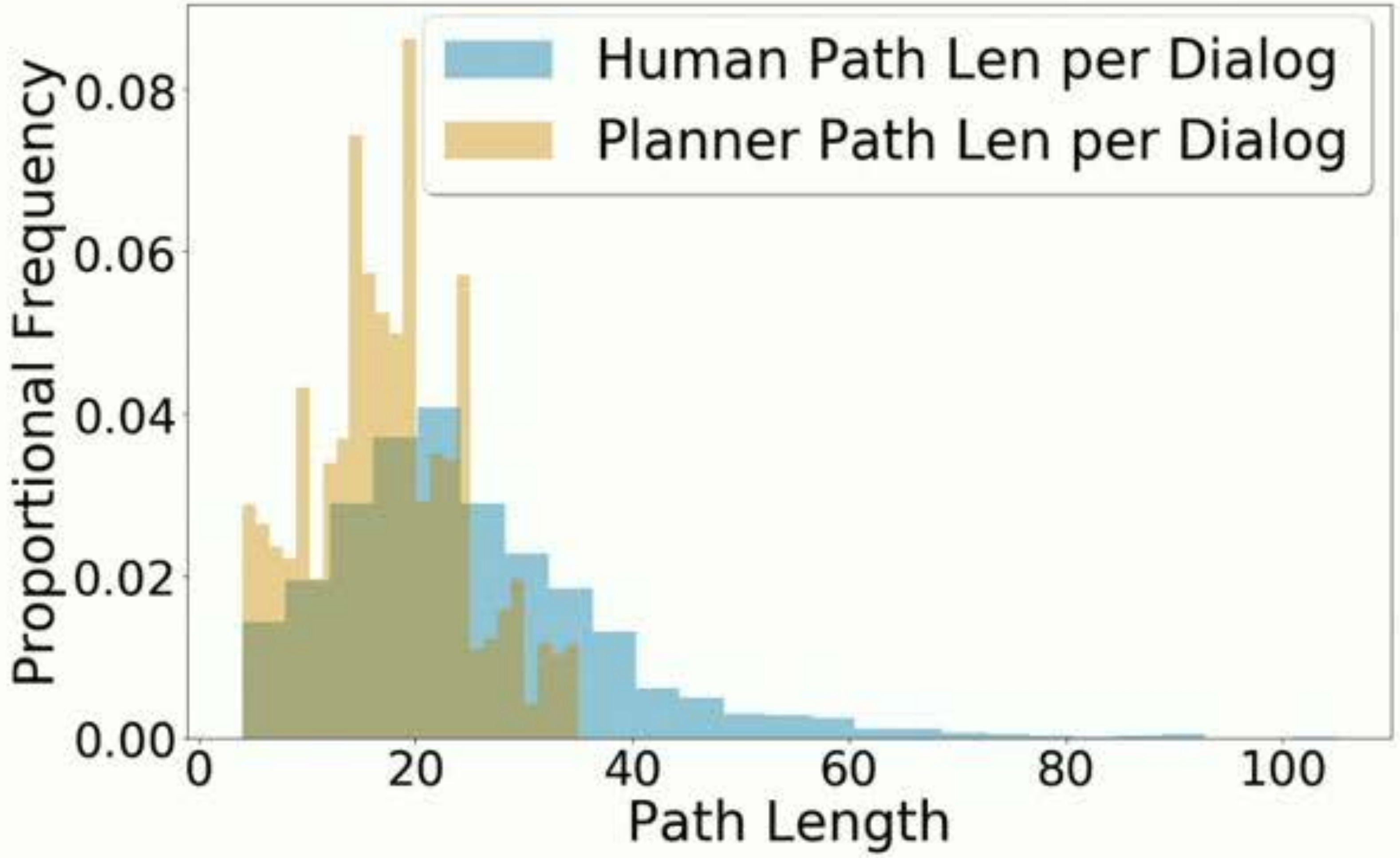


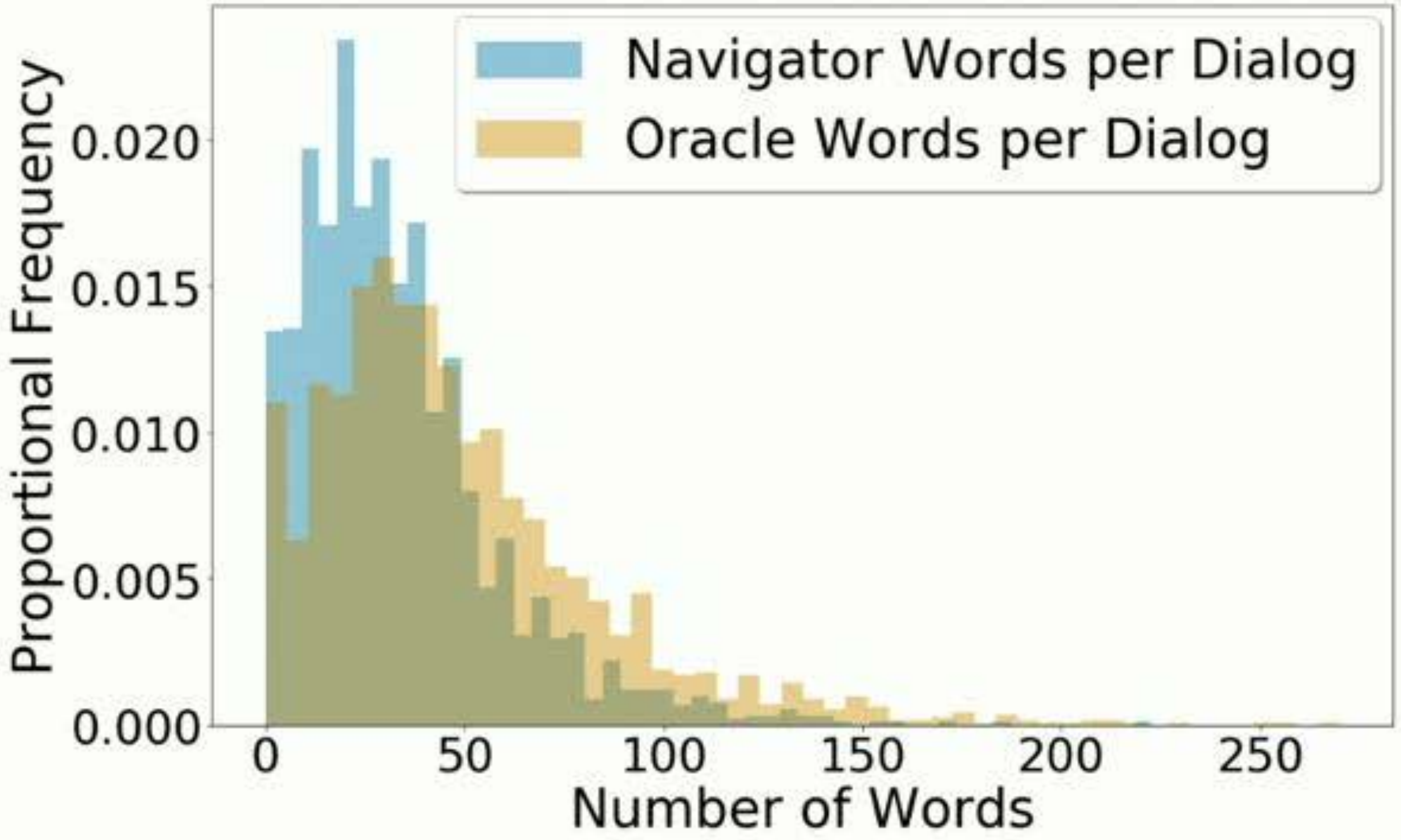
Dialog Leads to Long Paths and Rich Language



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

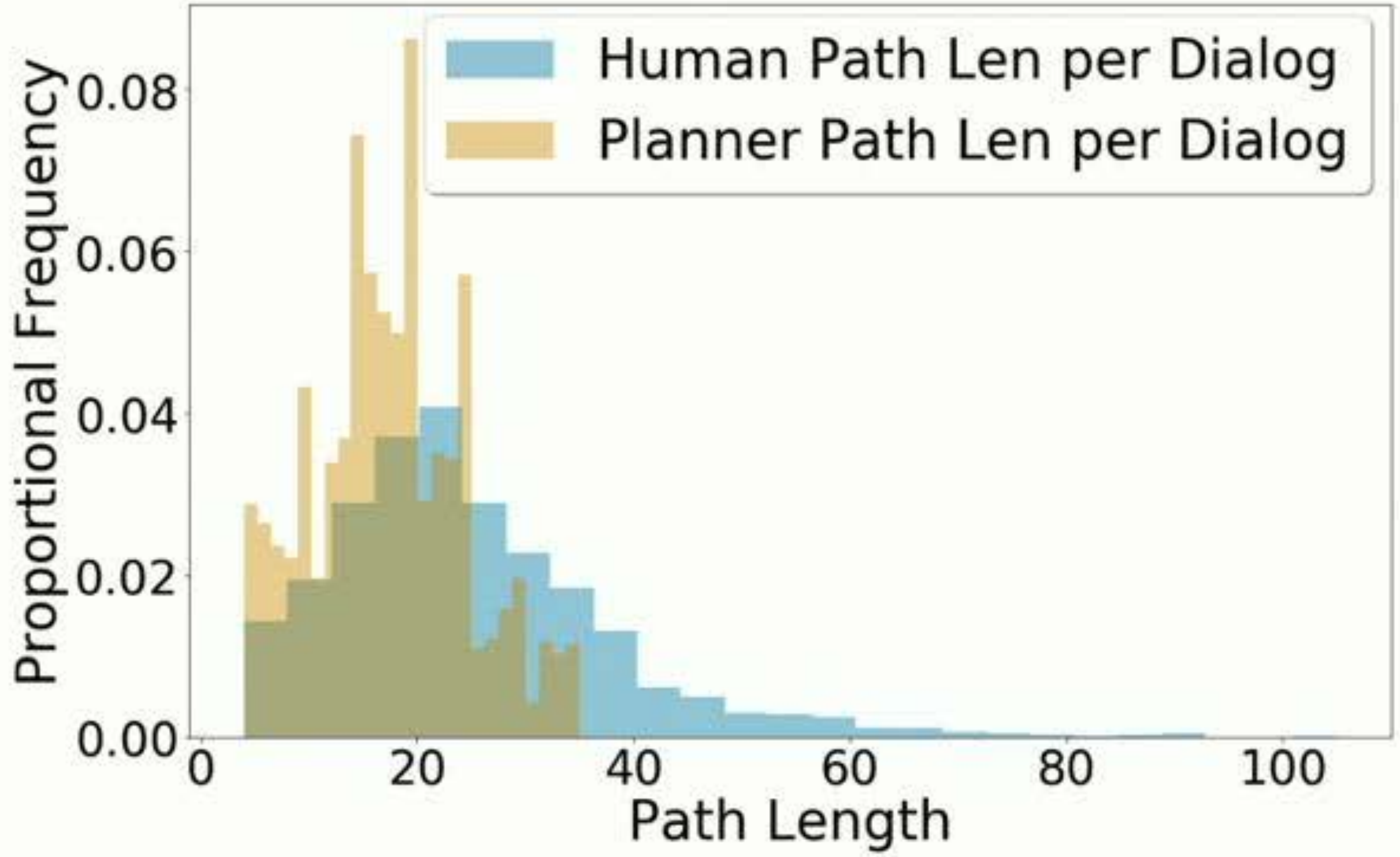
Dialog Leads to Long Paths and Rich Language

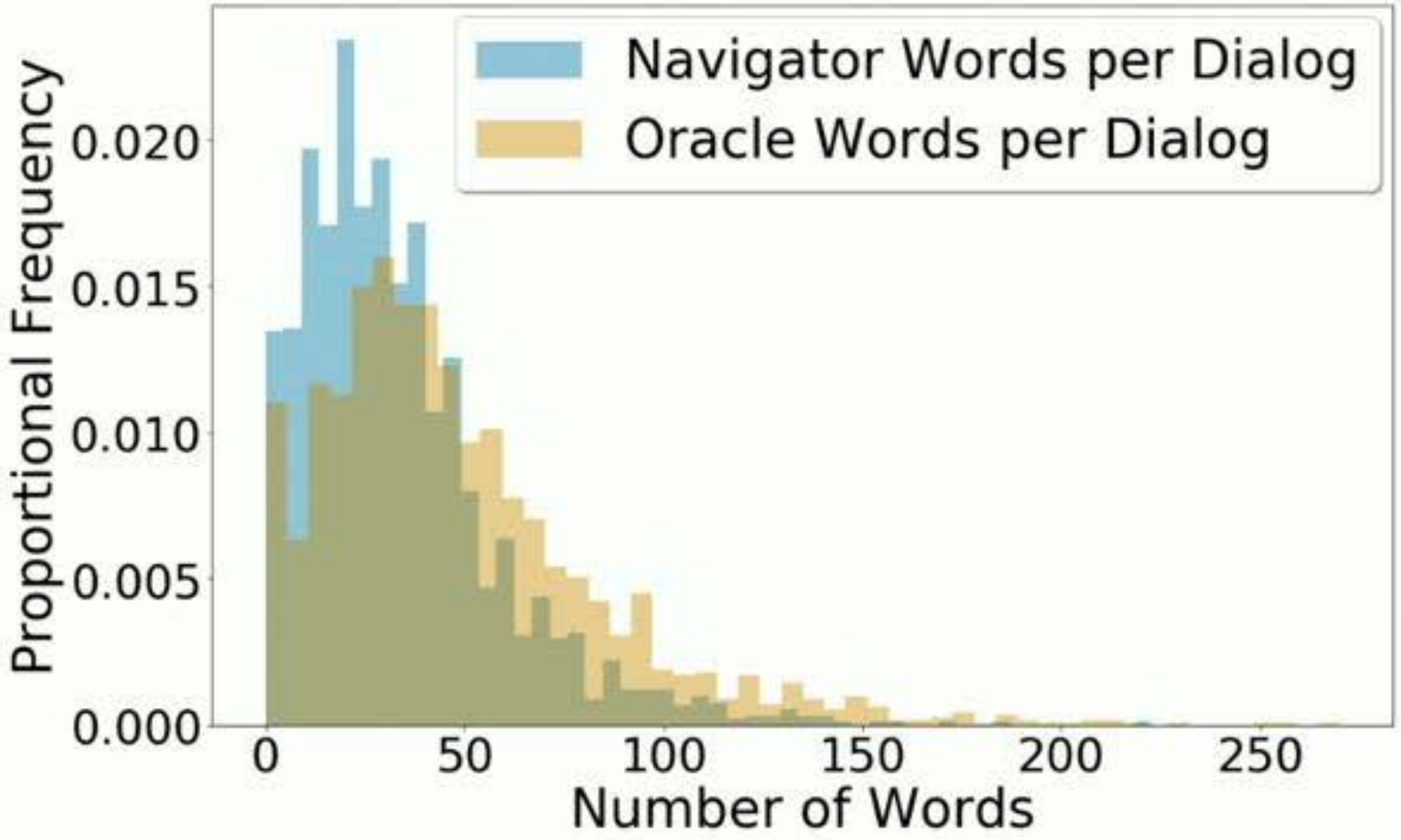




- Path Length Average:
  - Human (25.0); Planner (17.4)
  - Room-to-Room (6.0)

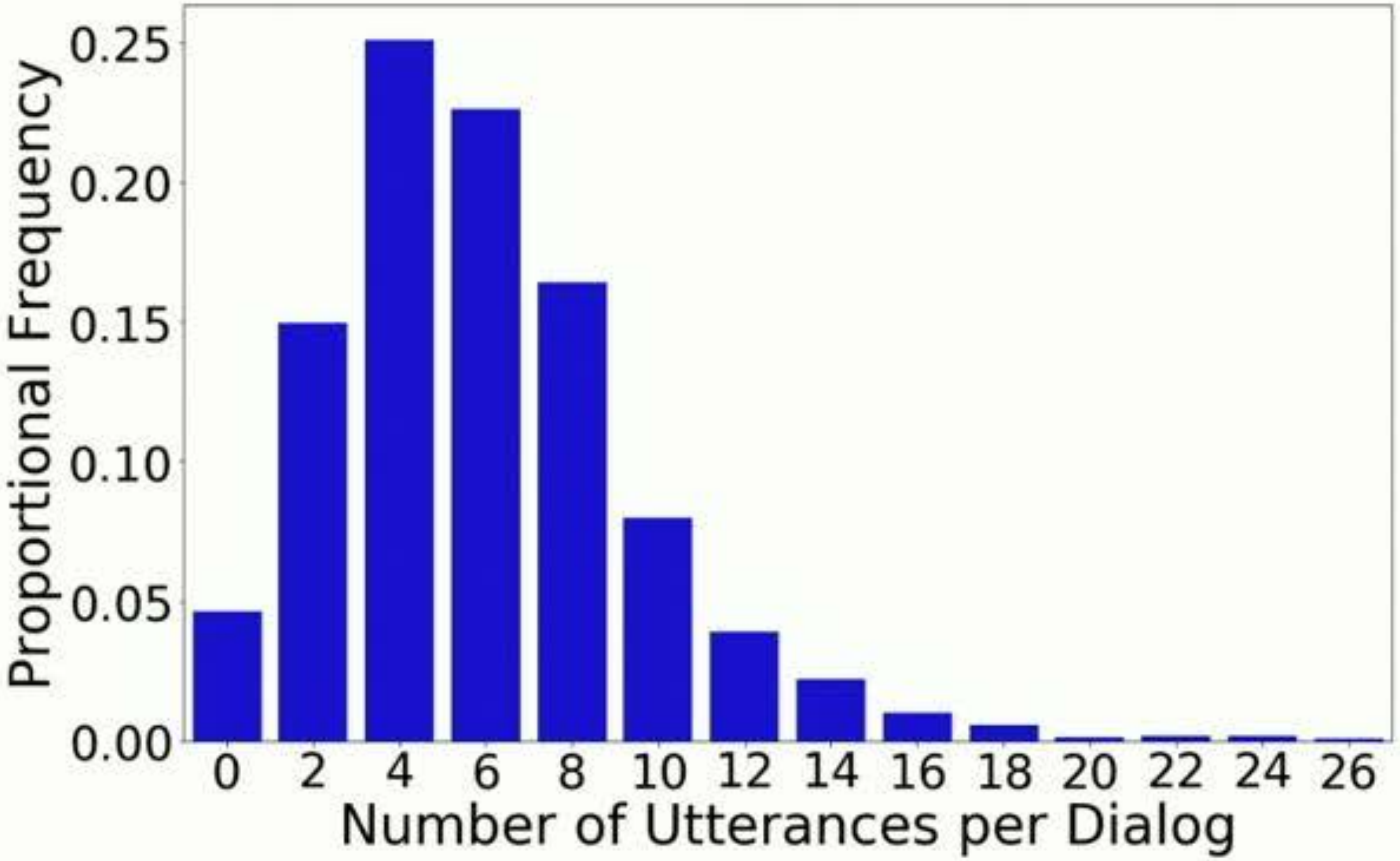
Dialog Leads to Long Paths and Rich Language



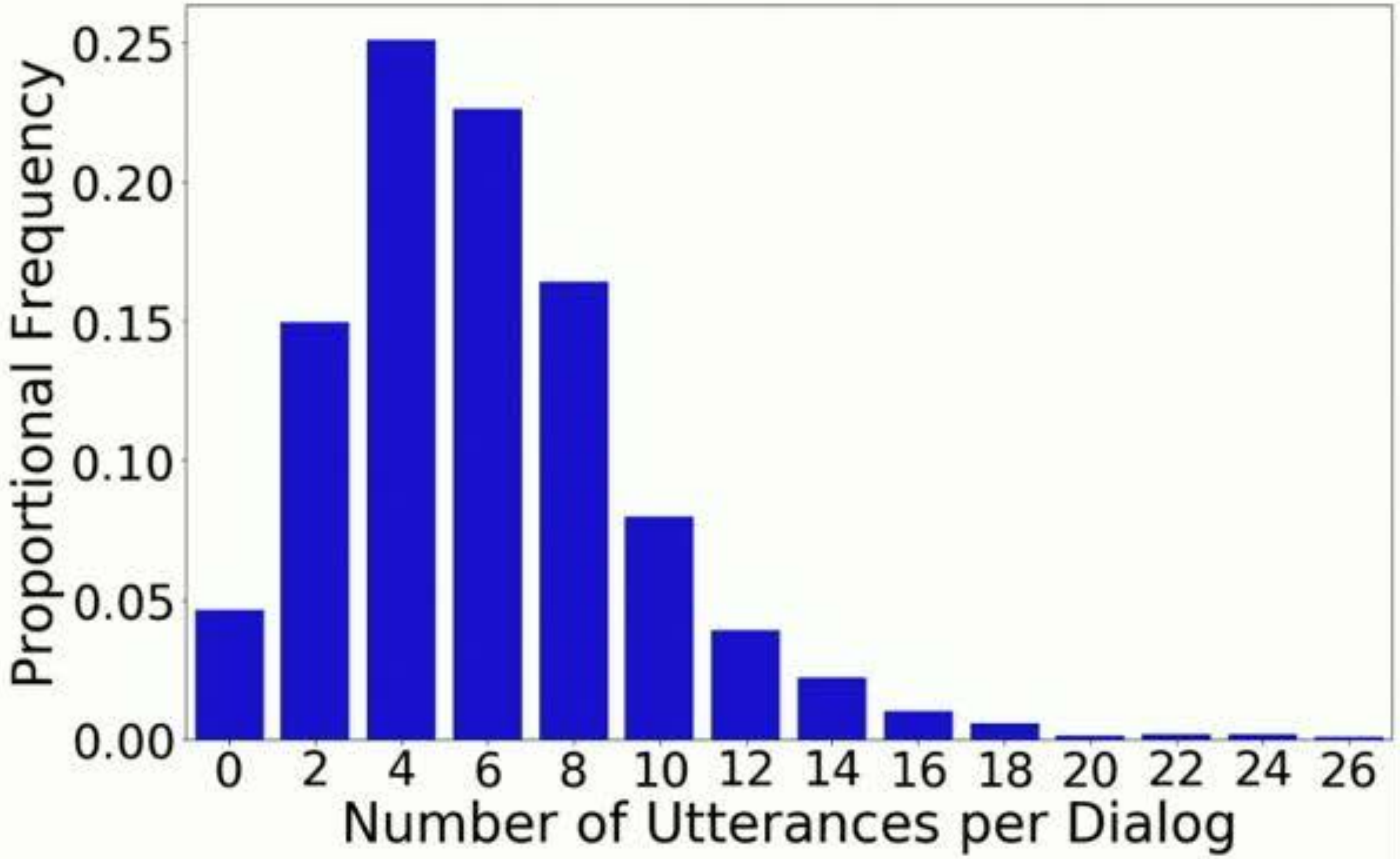


- Path Length Average:
    - Human (25.0); Planner (17.4)
    - Room-to-Room (6.0)
- Average total words:
    - CVDN (82)
    - Room-to-Room (29)

Dialog Leads to Long Paths and Rich Language

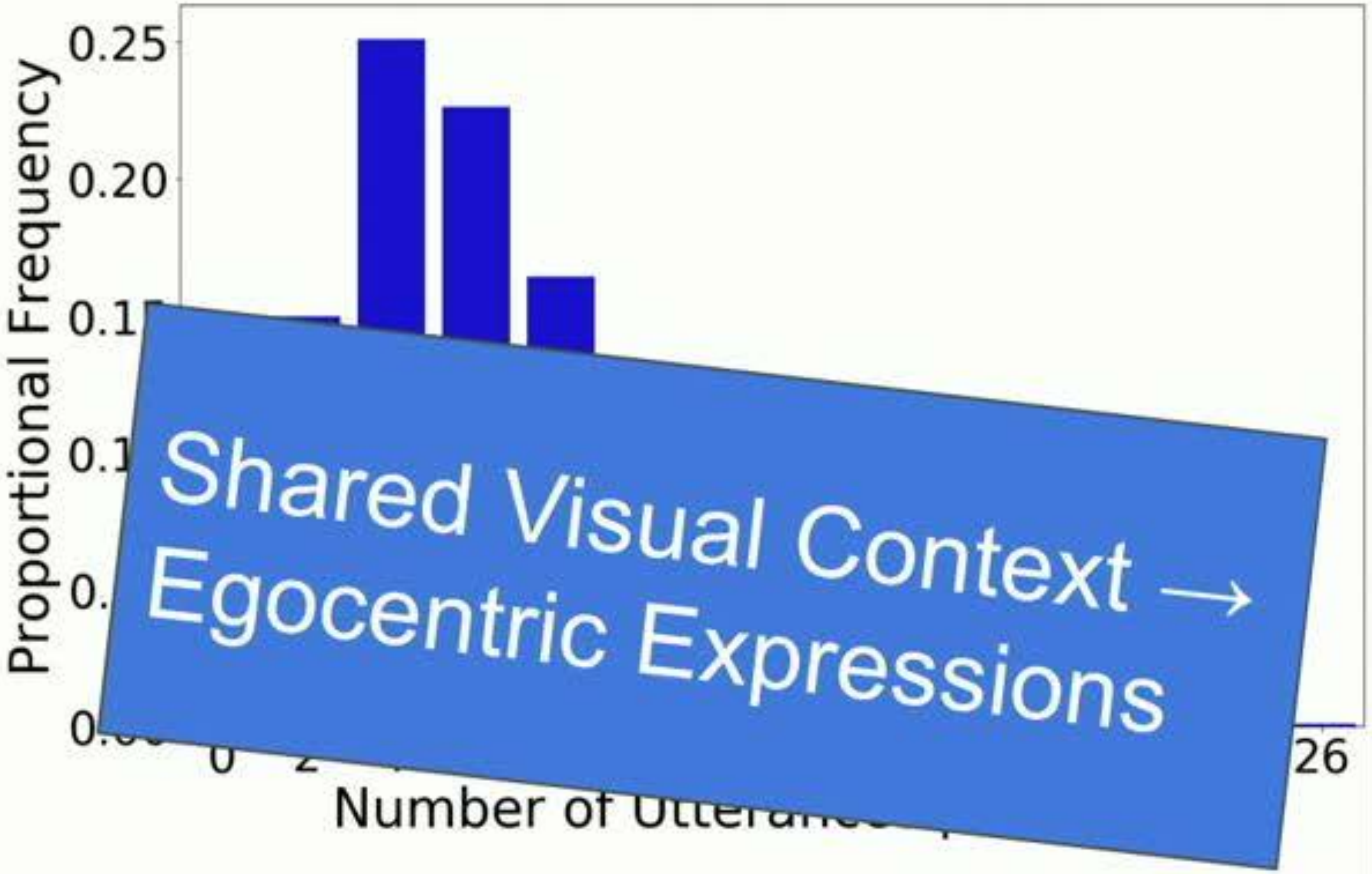


Dialog Leads to Long Paths and Rich Language



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.	➡ 7 steps
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.	➡ 6 steps
Navigator: Do I go left or right?	
Oracle: Go along the railing to the right. Stop at the room with a brown chair.	➡ 4s 🛑

Dialog Leads to Long Paths and Rich Language



Shared Visual Context → Egocentric Expressions

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.	➡ 7 steps
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.	➡ 6 steps
Navigator: Do I go left or right?	
Oracle: Go along the railing to the right. Stop at the room with a brown chair.	➡ 4s 🛑

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*.	→ 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.	→ 7 steps
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.	→ 6 steps
Navigator: Do I go left or right?	
Oracle: Go along the railing to the right. Stop at the room with a brown chair.	→ 4s 🛑

Navigation from Dialog History

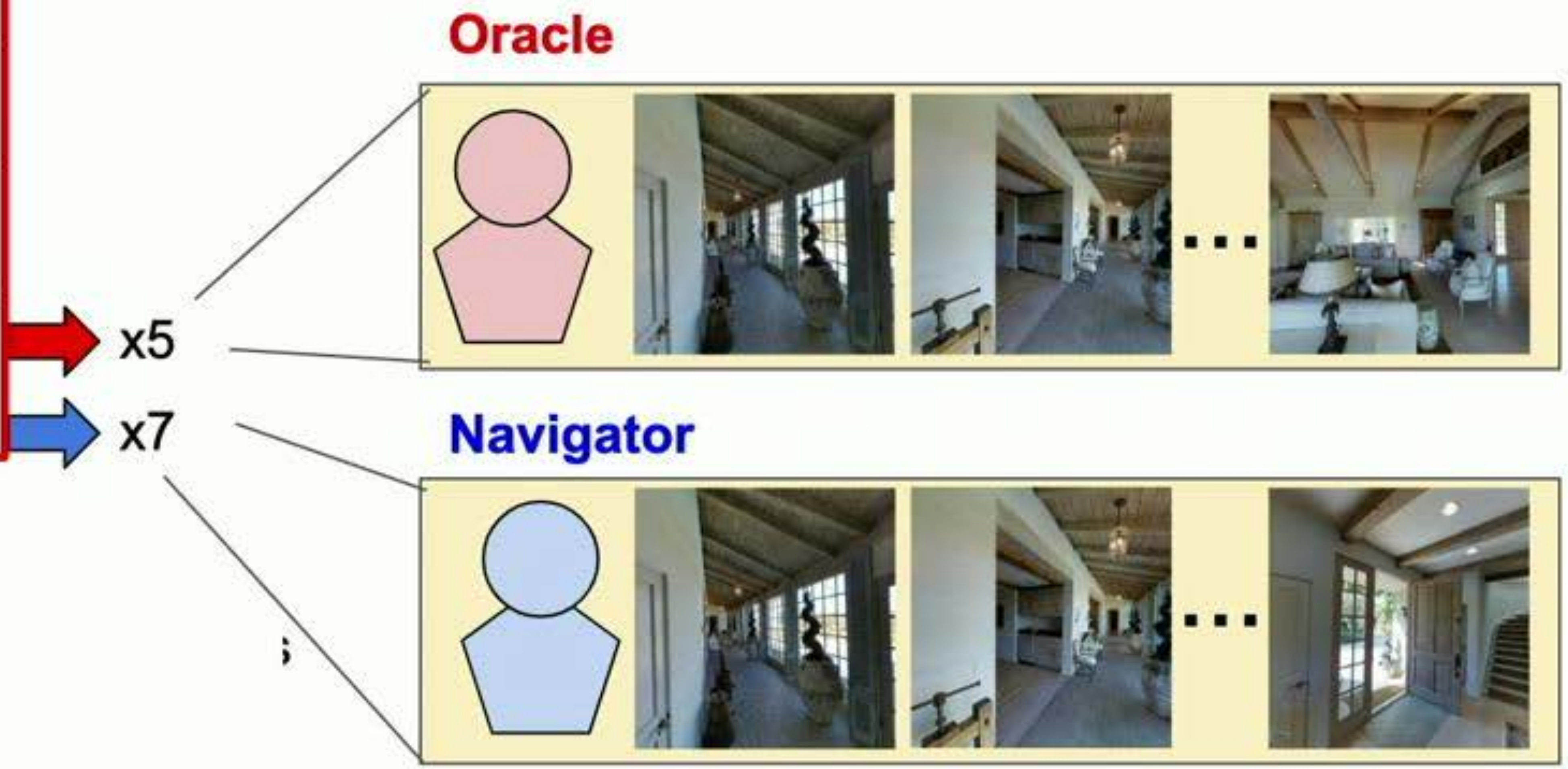
The goal room contains a *rug*.

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

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.

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*. → 3 steps

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.

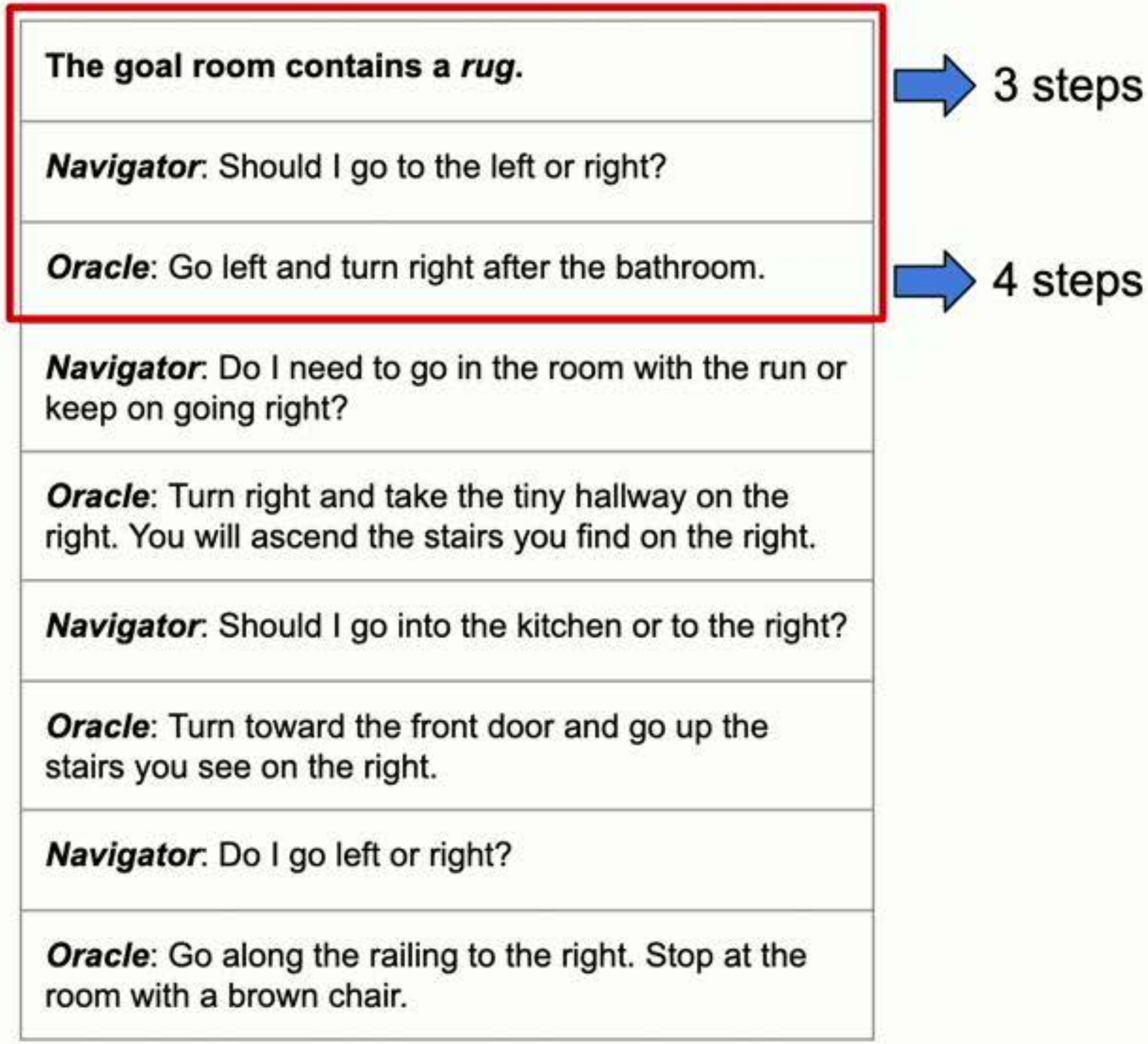
Navigator: Should I go into the kitchen or to the right?

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.

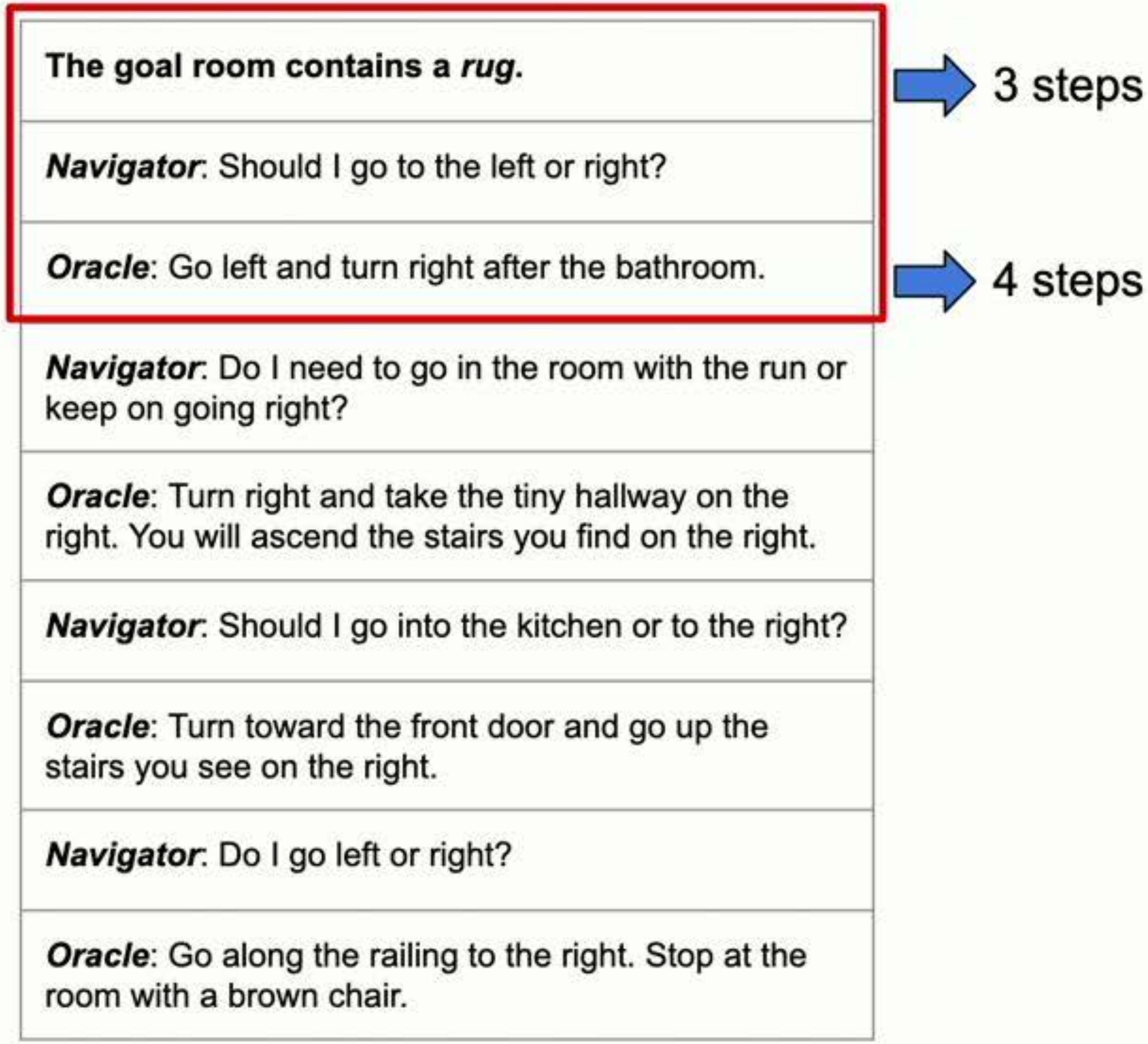
Navigator: Should I go into the kitchen or to the right?

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.

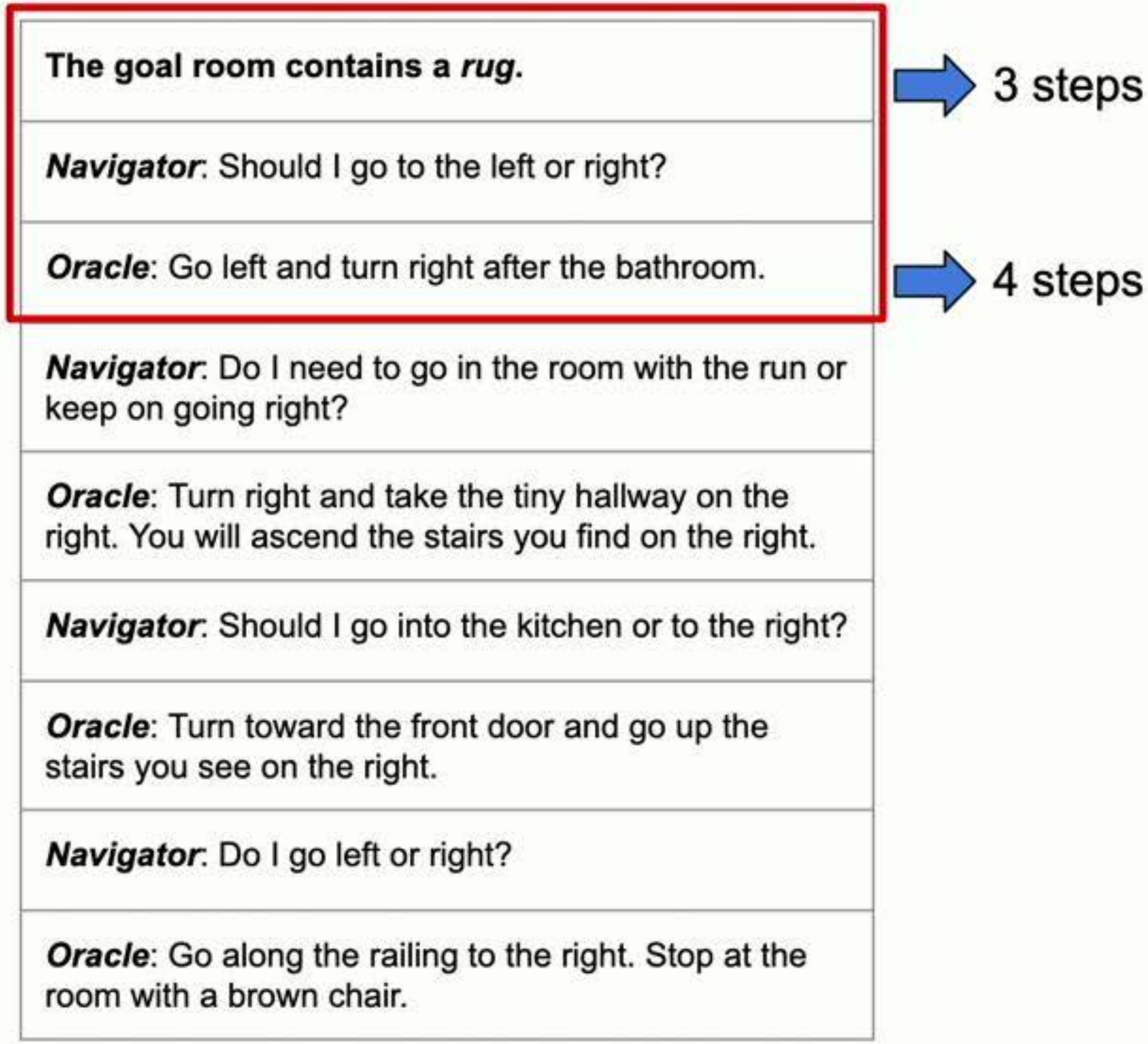
Navigation from Dialog History



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 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.

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?

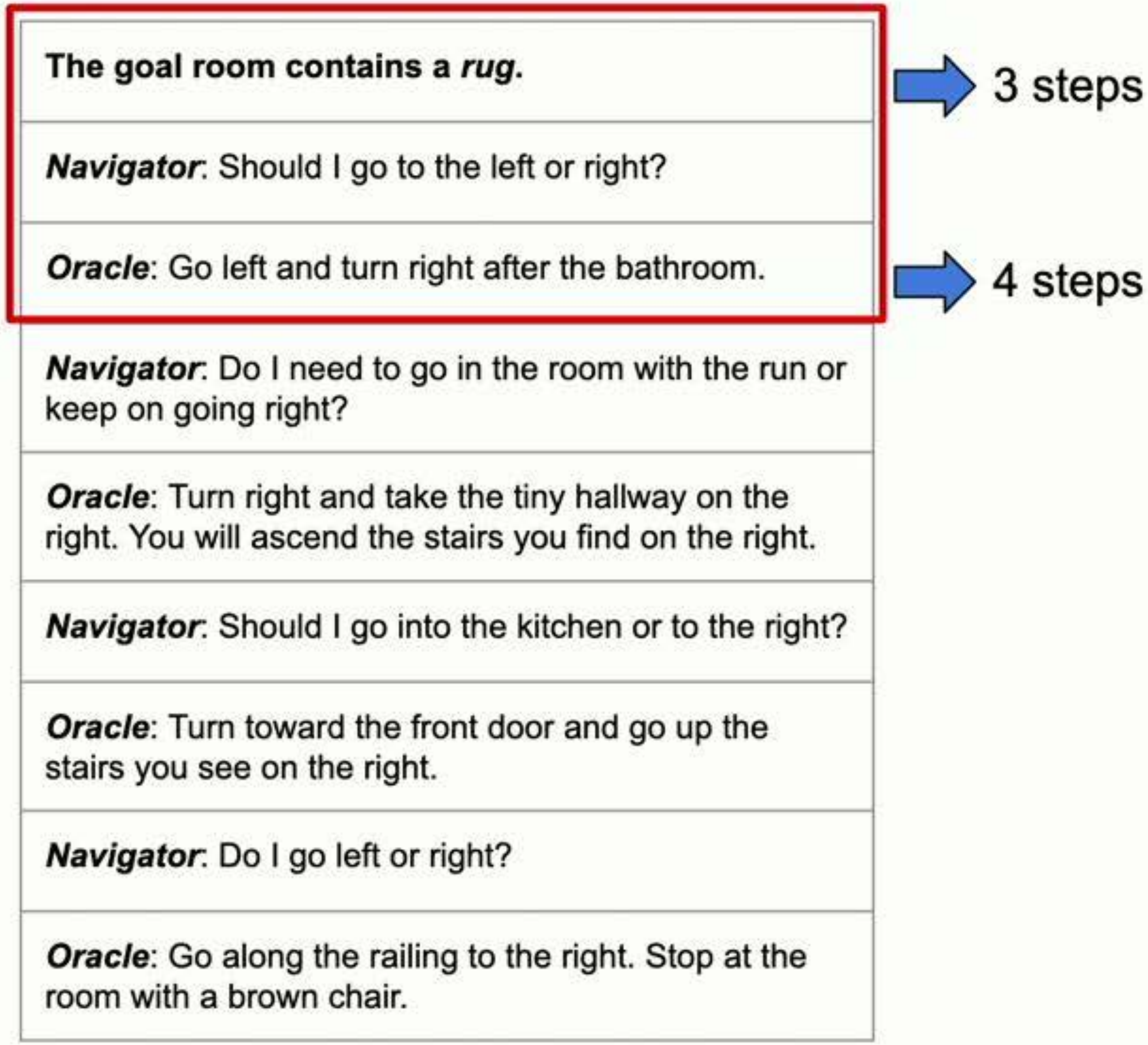
Oracle: Turn toward the front door and go up the stairs you see on the right.

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*.

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?

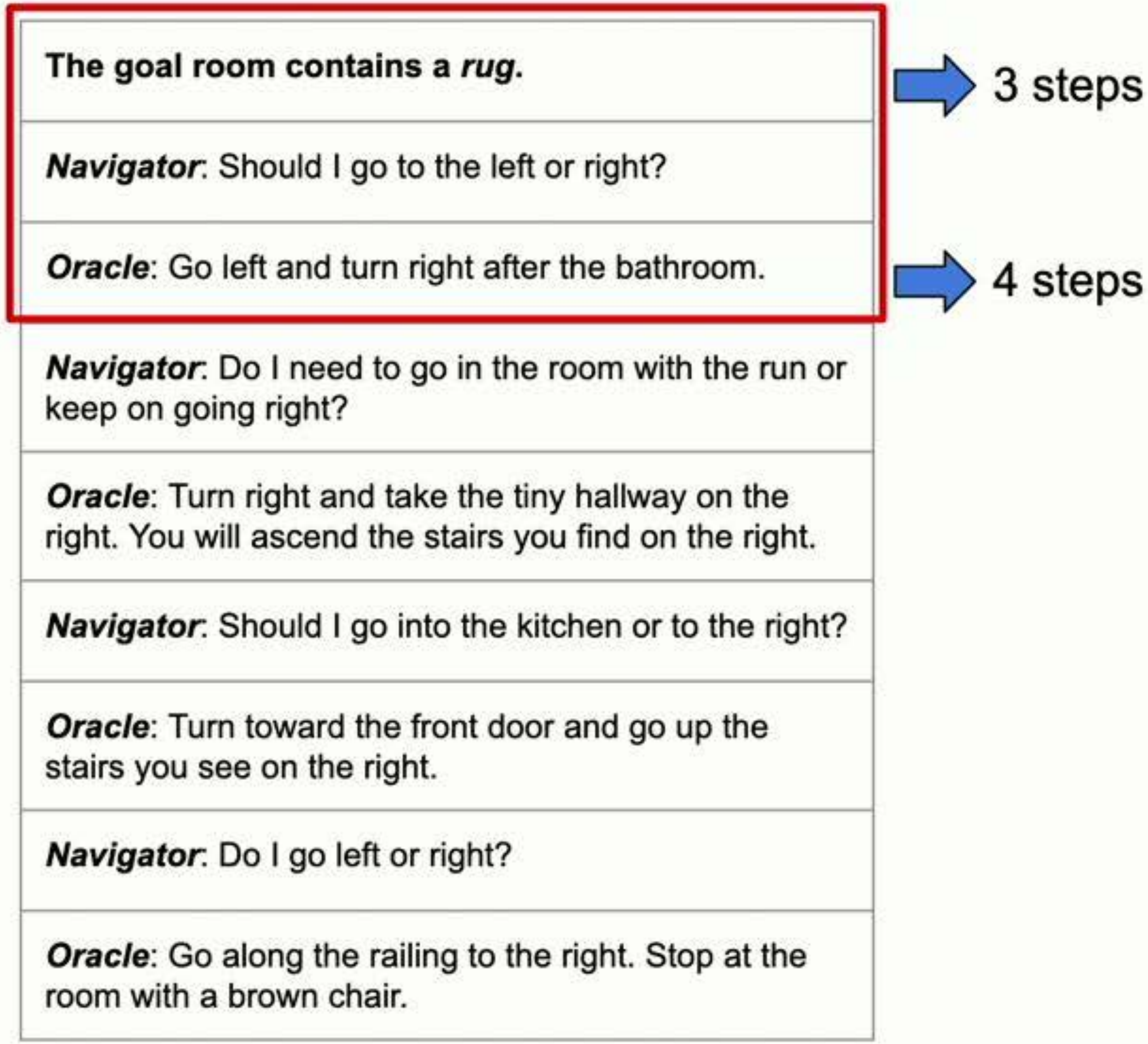
Oracle: Go along the railing to the right. Stop at the room with a brown chair.

3 steps

4 steps

- **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?

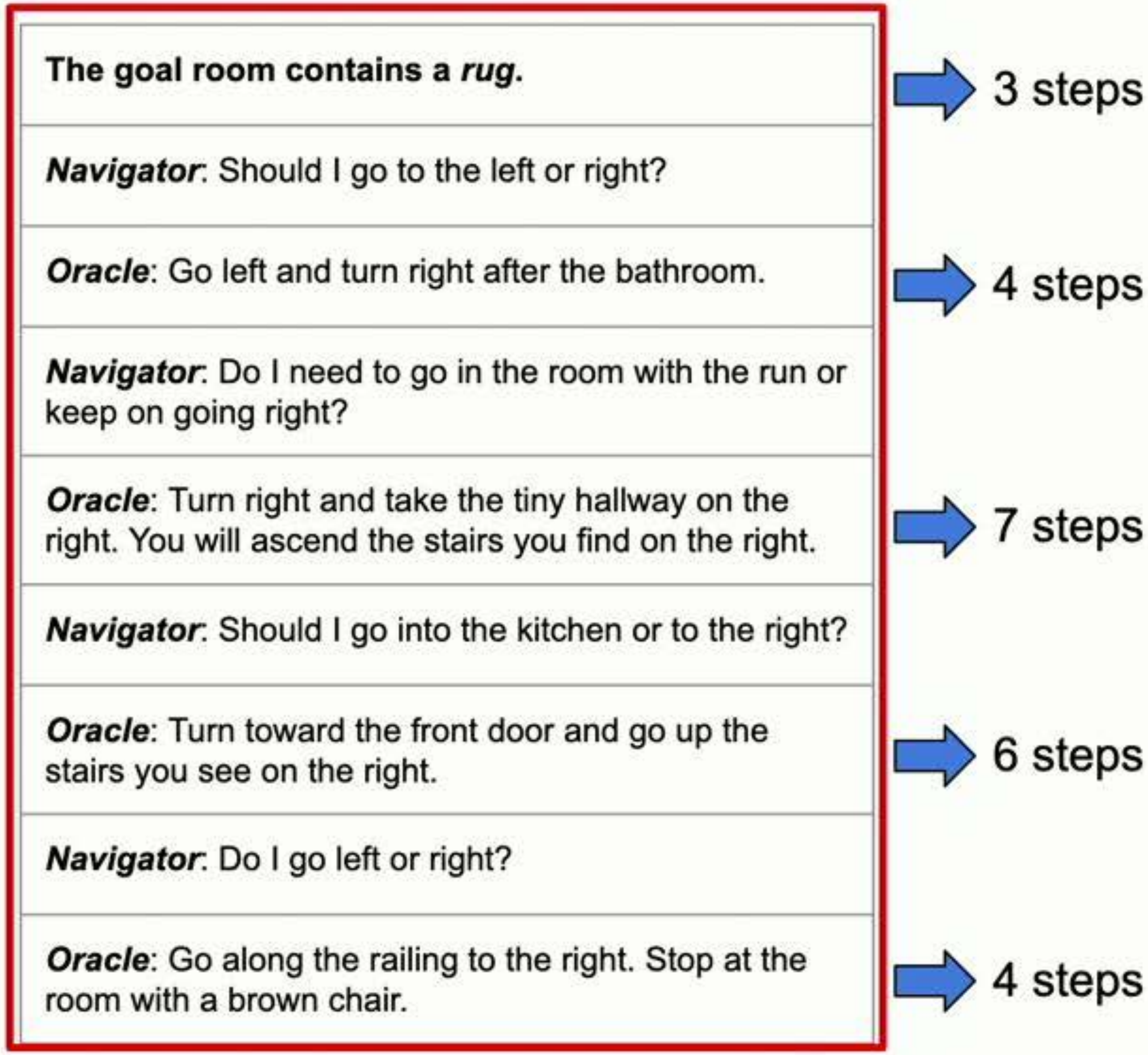
Oracle: Turn toward the front door and go up the stairs you see on the right.

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*. → 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. → 7 steps

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

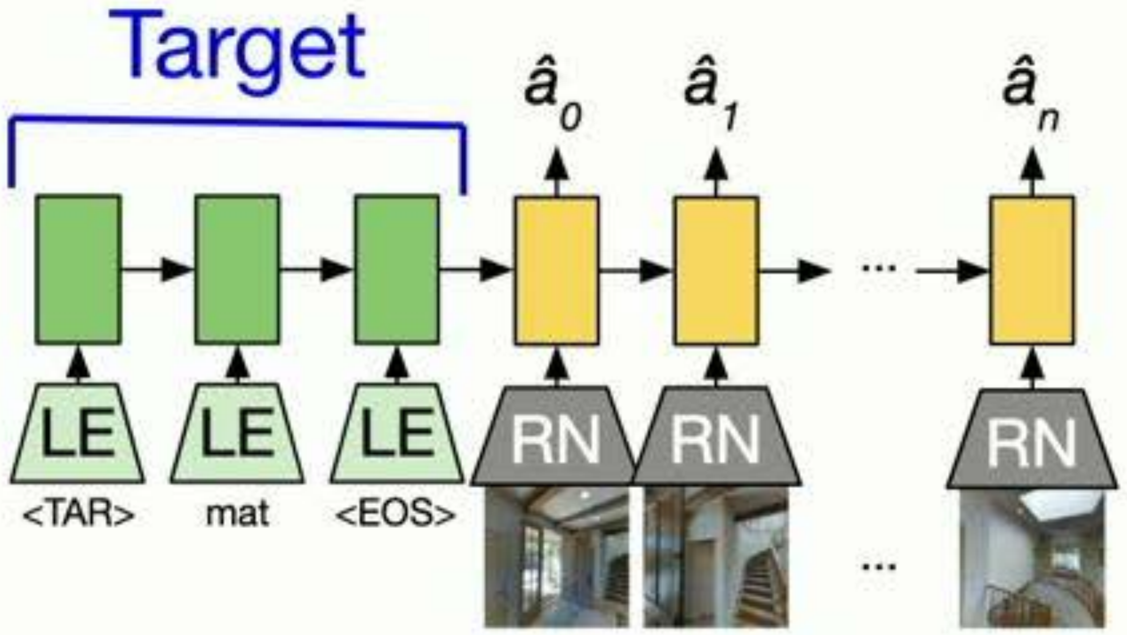
Oracle: Turn toward the front door and go up the stairs you see on the right. → 6 steps

Navigator: Do I go left or right?

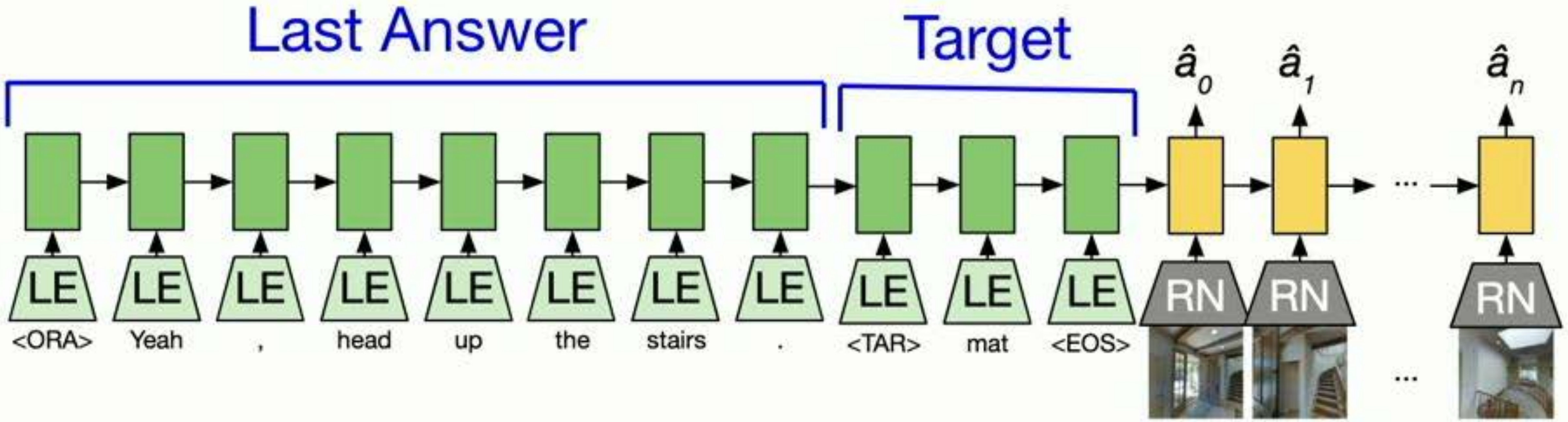
Oracle: Go along the railing to the right. Stop at the room with a brown chair. → 4 steps

- **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

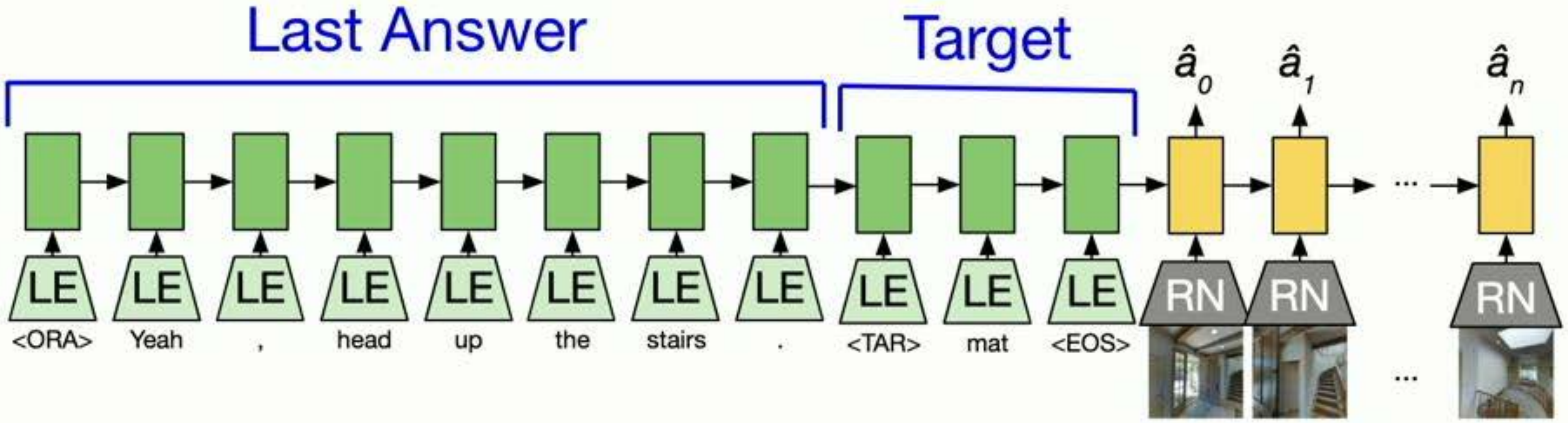


Initial, Sequence-to-Sequence Model



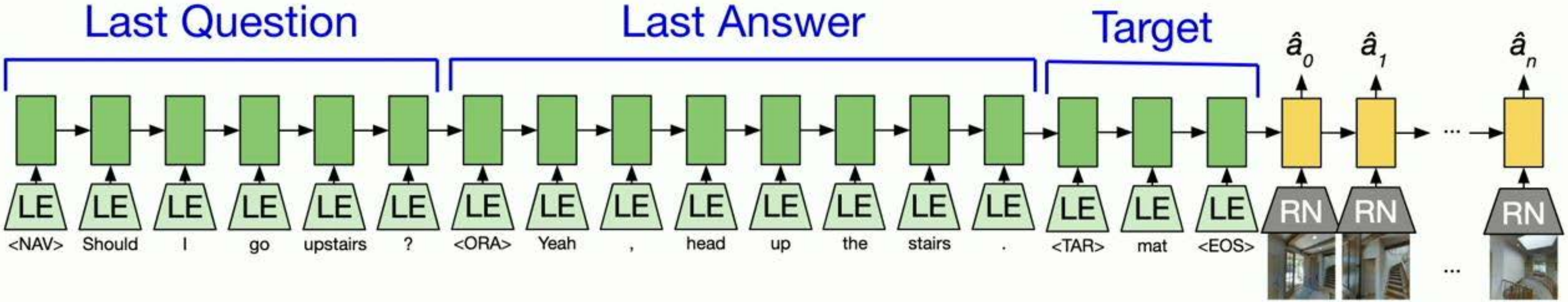
Initial, Sequence-to-Sequence Model

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

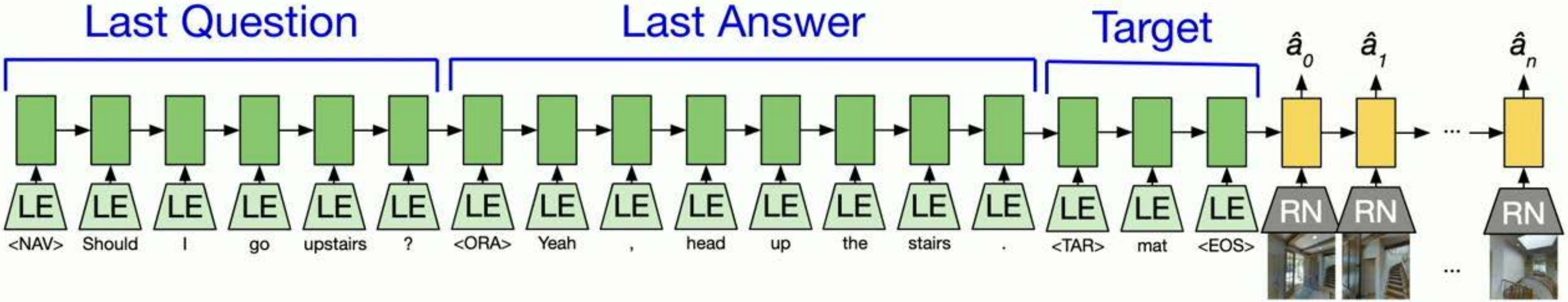


Initial, Sequence-to-Sequence Model

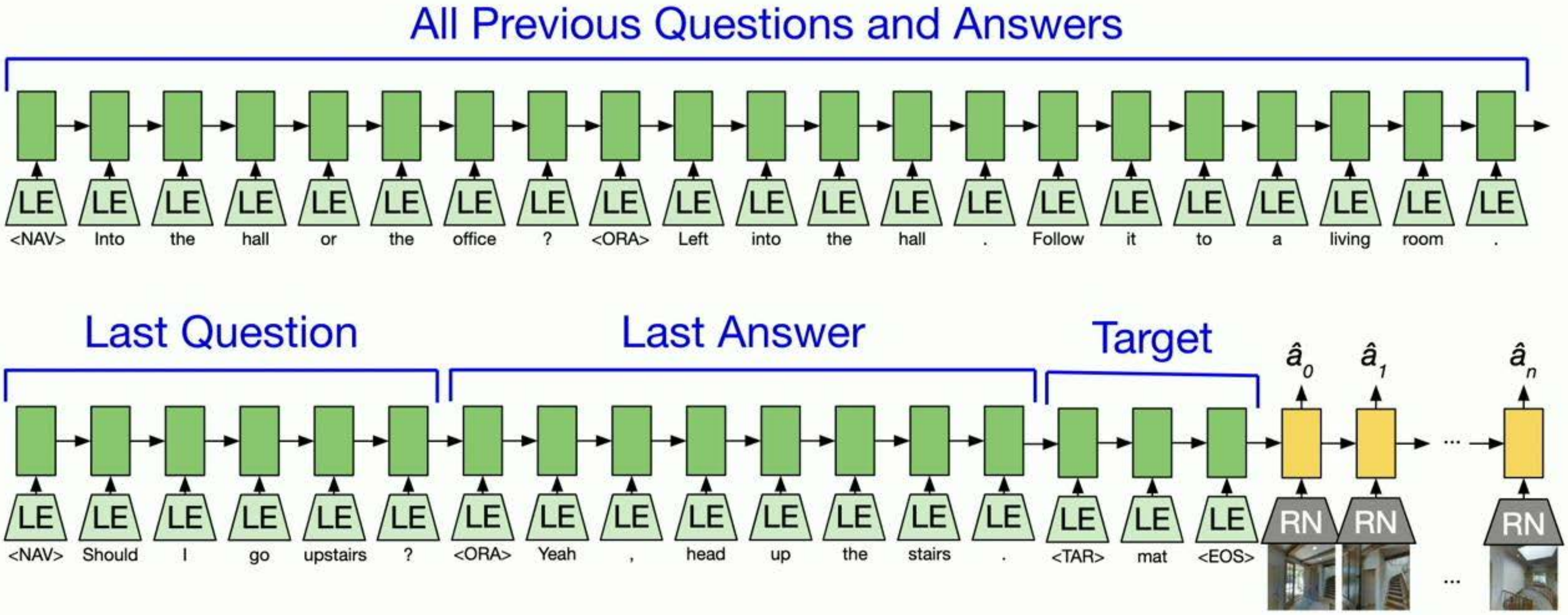
Oracle: Through the lobby. So go through the door next to the green towel. Go to the left door next to the two yellow lights. Walk straight to the end of the hallway and stop

…

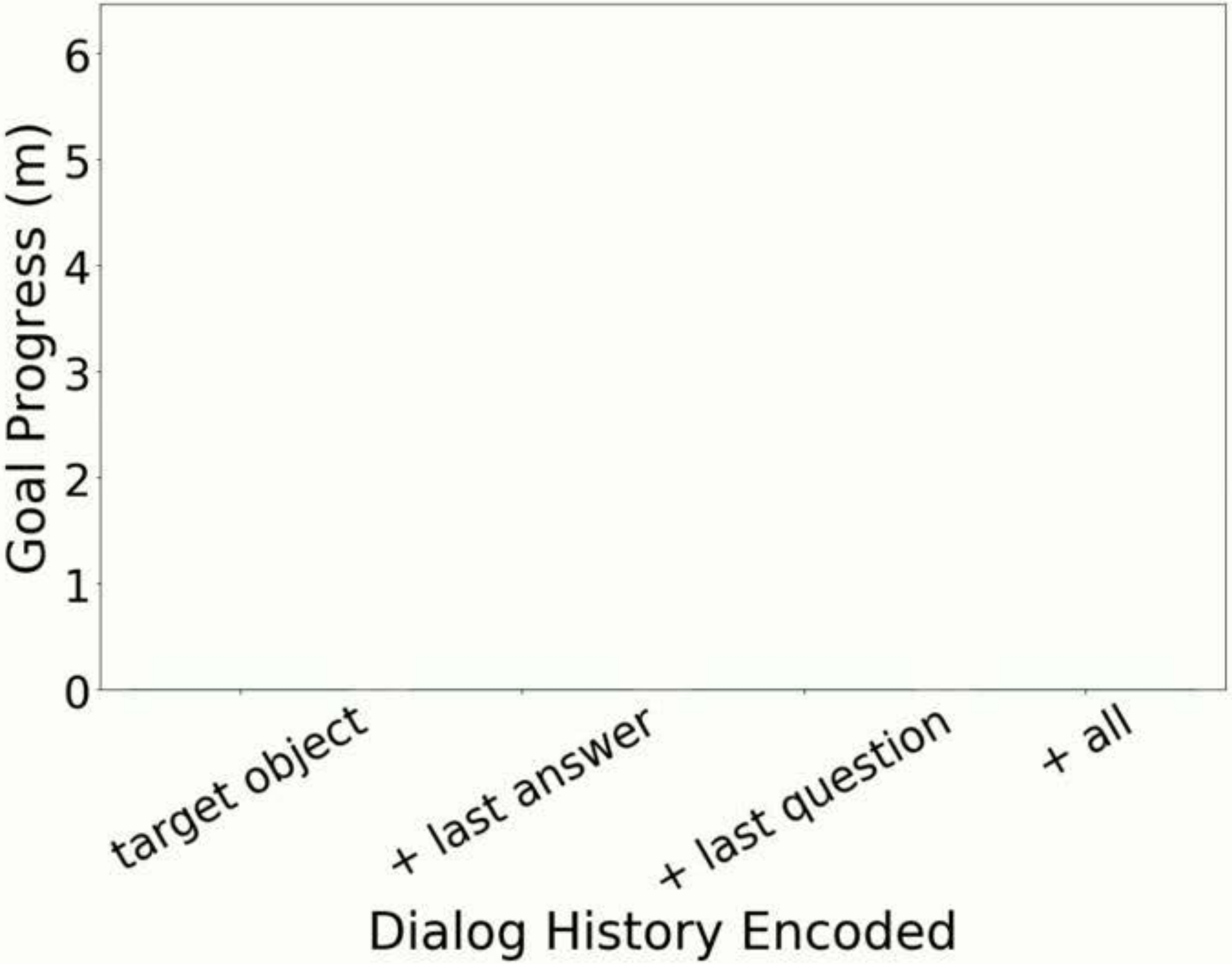
Navigator: Are these the yellow lights you were talking about?



Initial, Sequence-to-Sequence Model

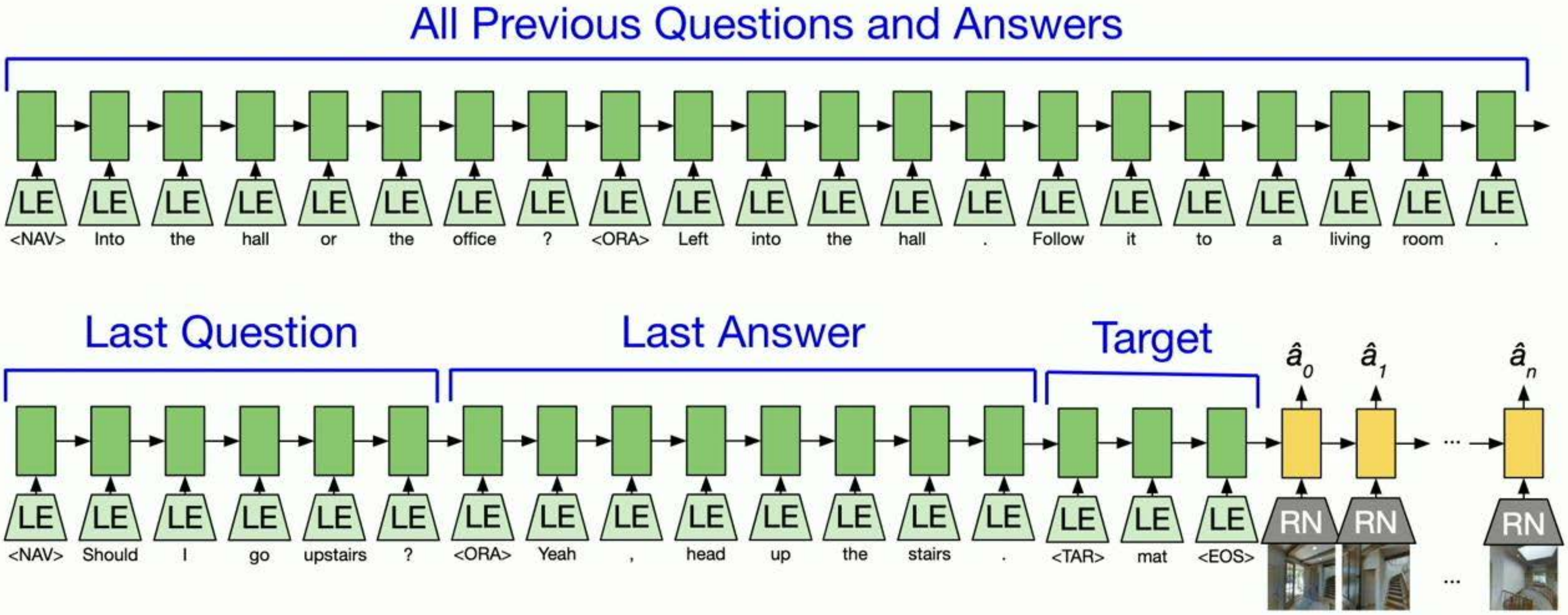

All Previous Questions and Answers
LE
<NAV> Into the hall or the office ? <ORA> Left into the hall . Follow it to a living room .
Last Question
Last Answer
Target
$\hat{a}_0$
$\hat{a}_1$
$\hat{a}_n$
...
RN
<NAV> Should I go upstairs ? <ORA> Yeah , head up the stairs . <TAR> mat <EOS>

Evaluation - Validation (best perf over 200 epochs)

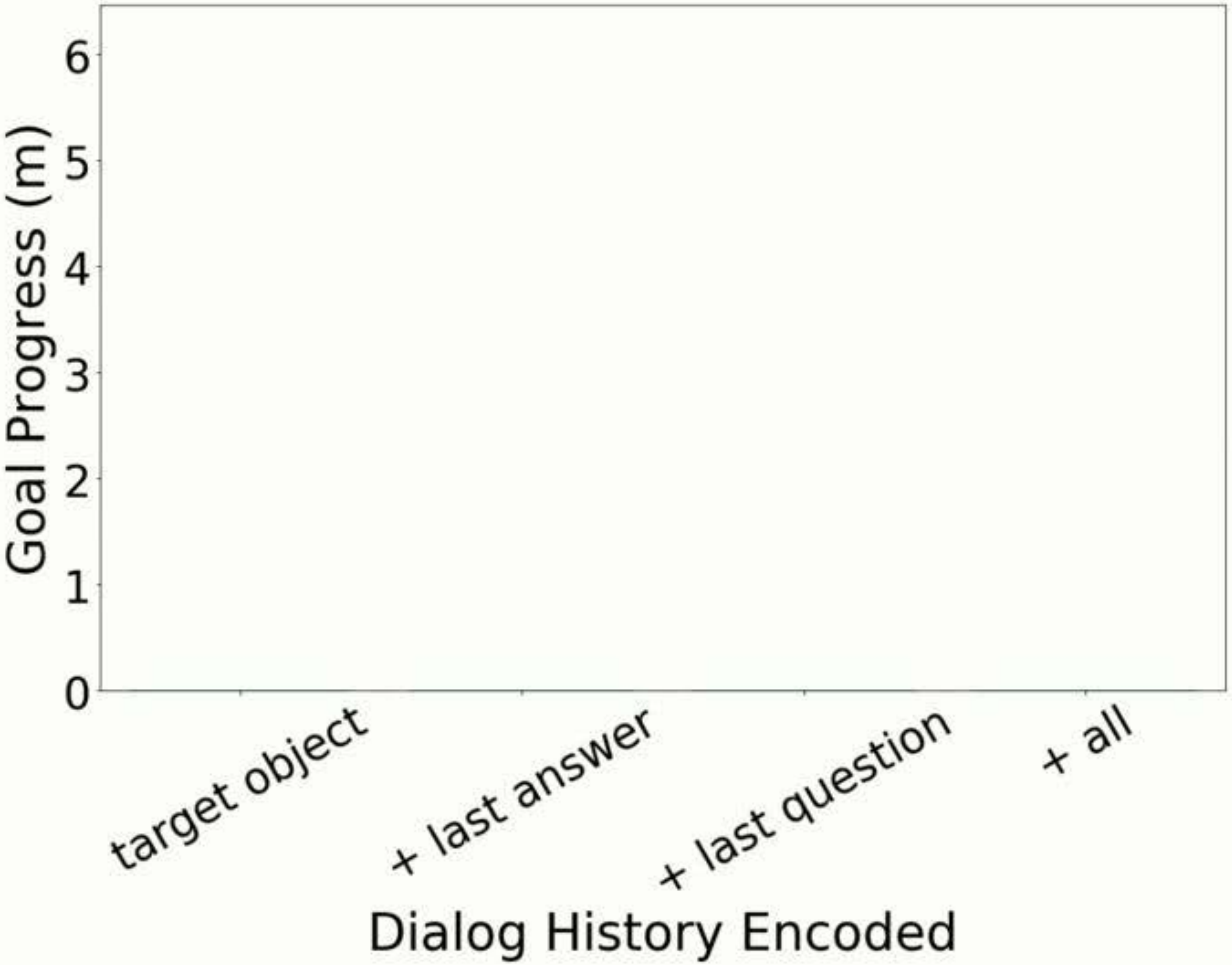


- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.

Initial, Sequence-to-Sequence Model

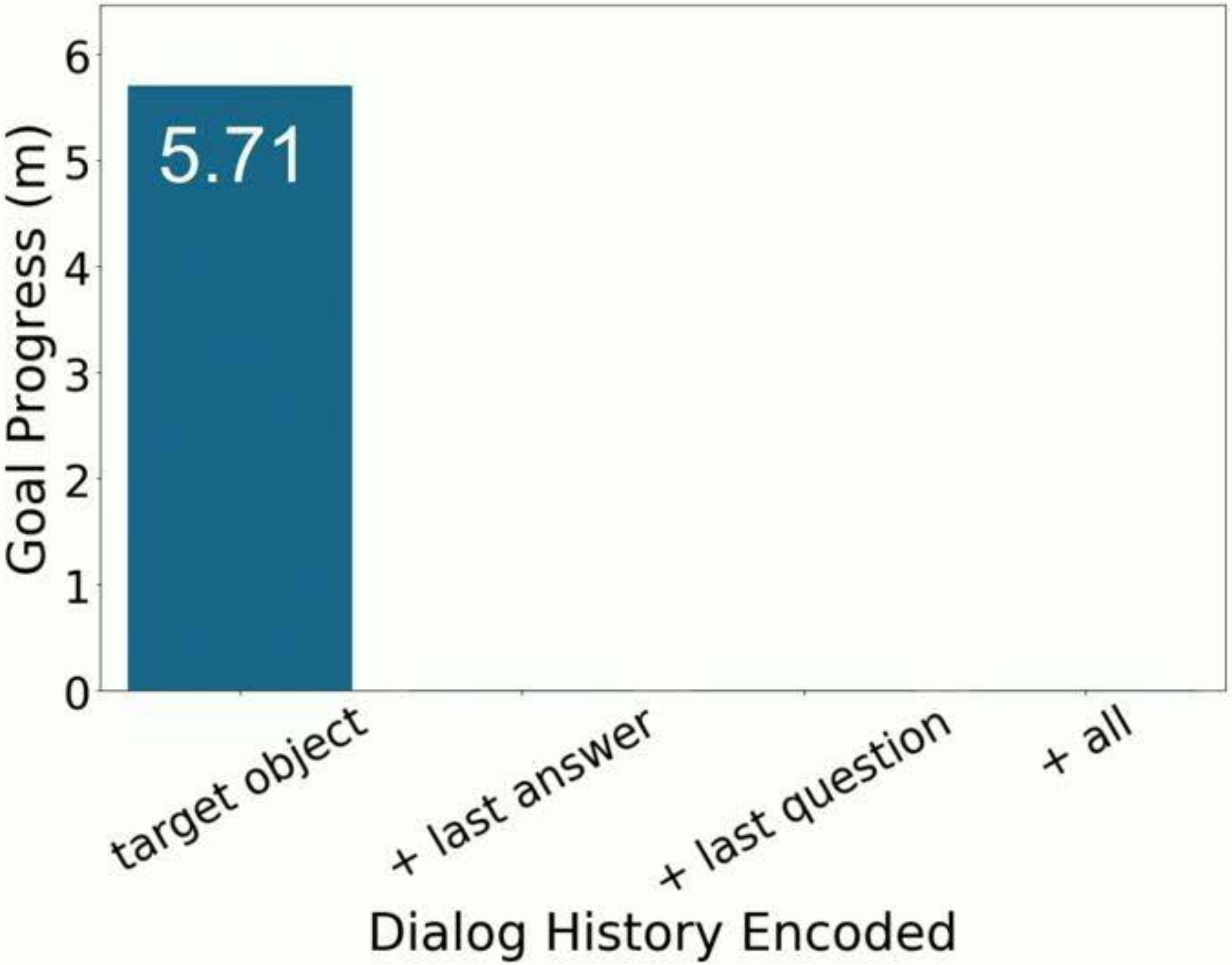


Evaluation - Validation (best perf over 200 epochs)



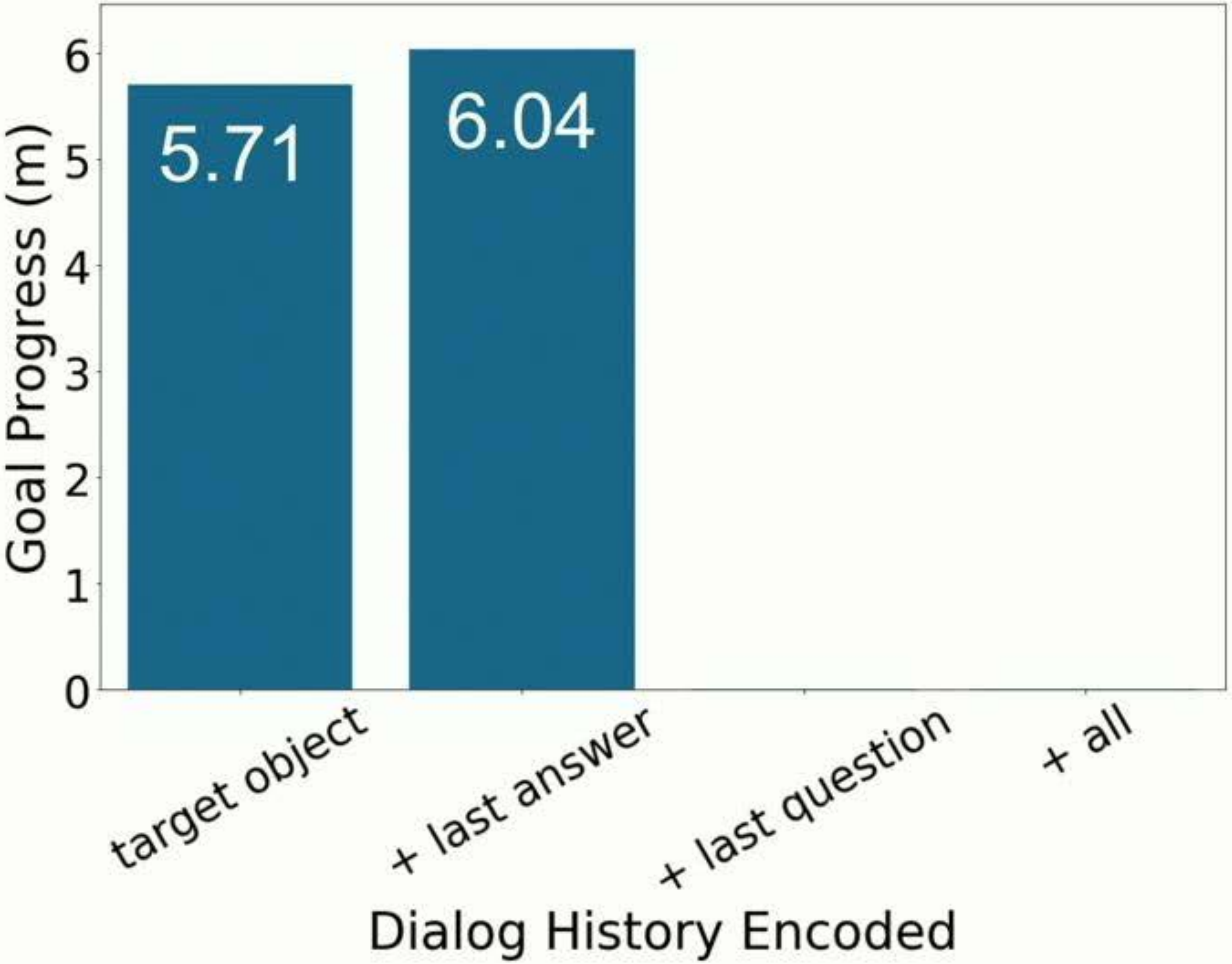
- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.

Evaluation - Validation (best perf over 200 epochs)



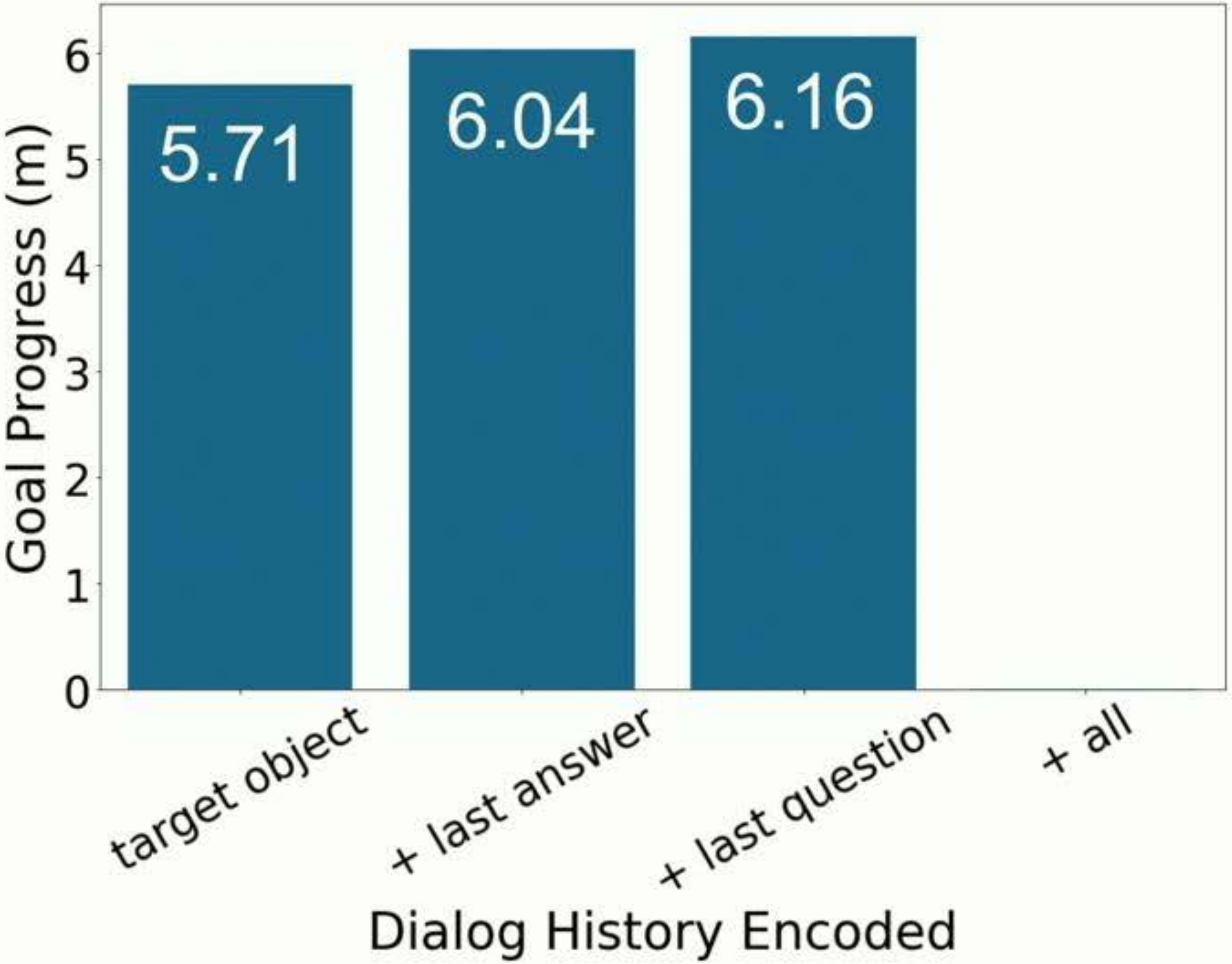
- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.

Evaluation - Validation (best perf over 200 epochs)



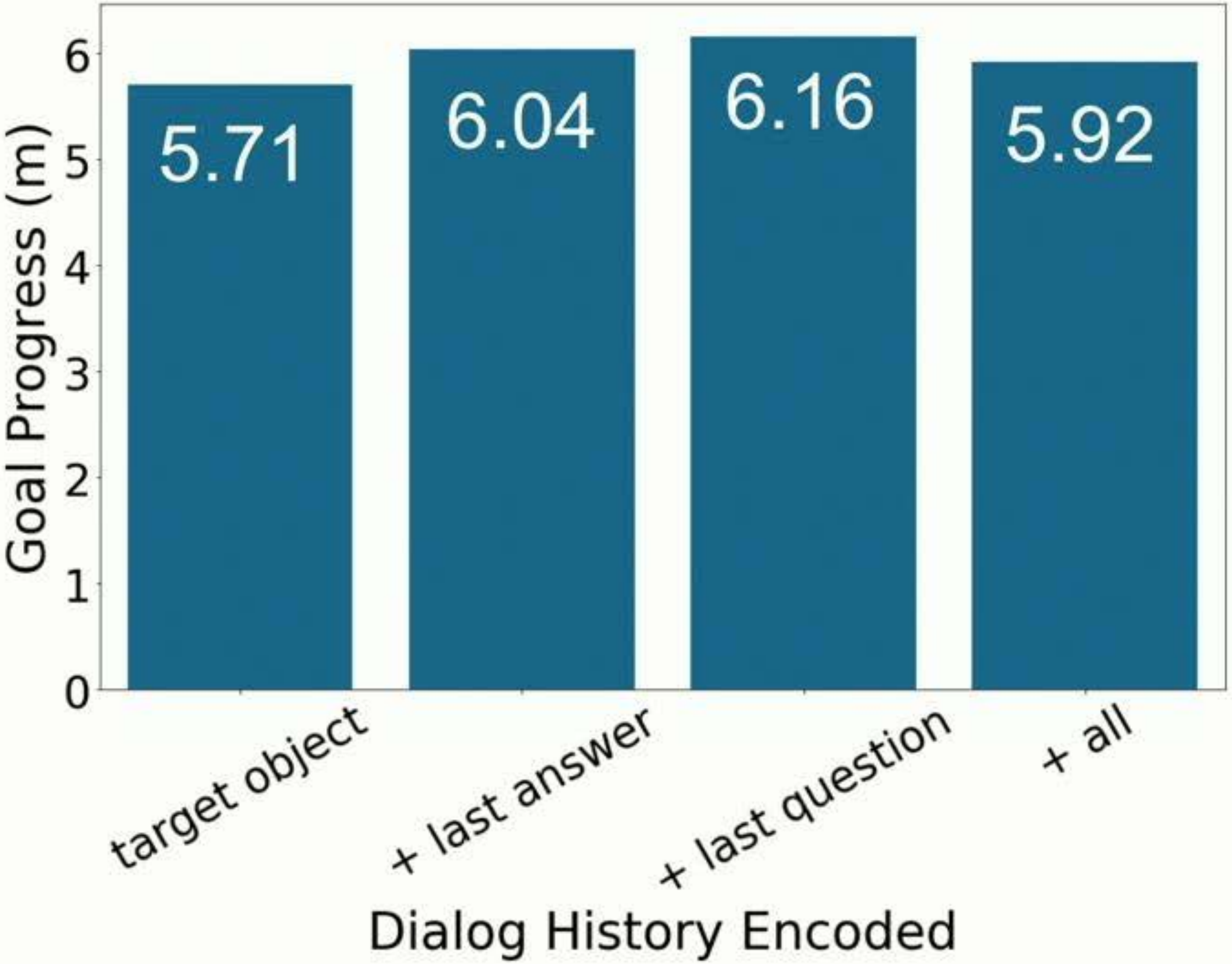
- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.

Evaluation - Validation (best perf over 200 epochs)



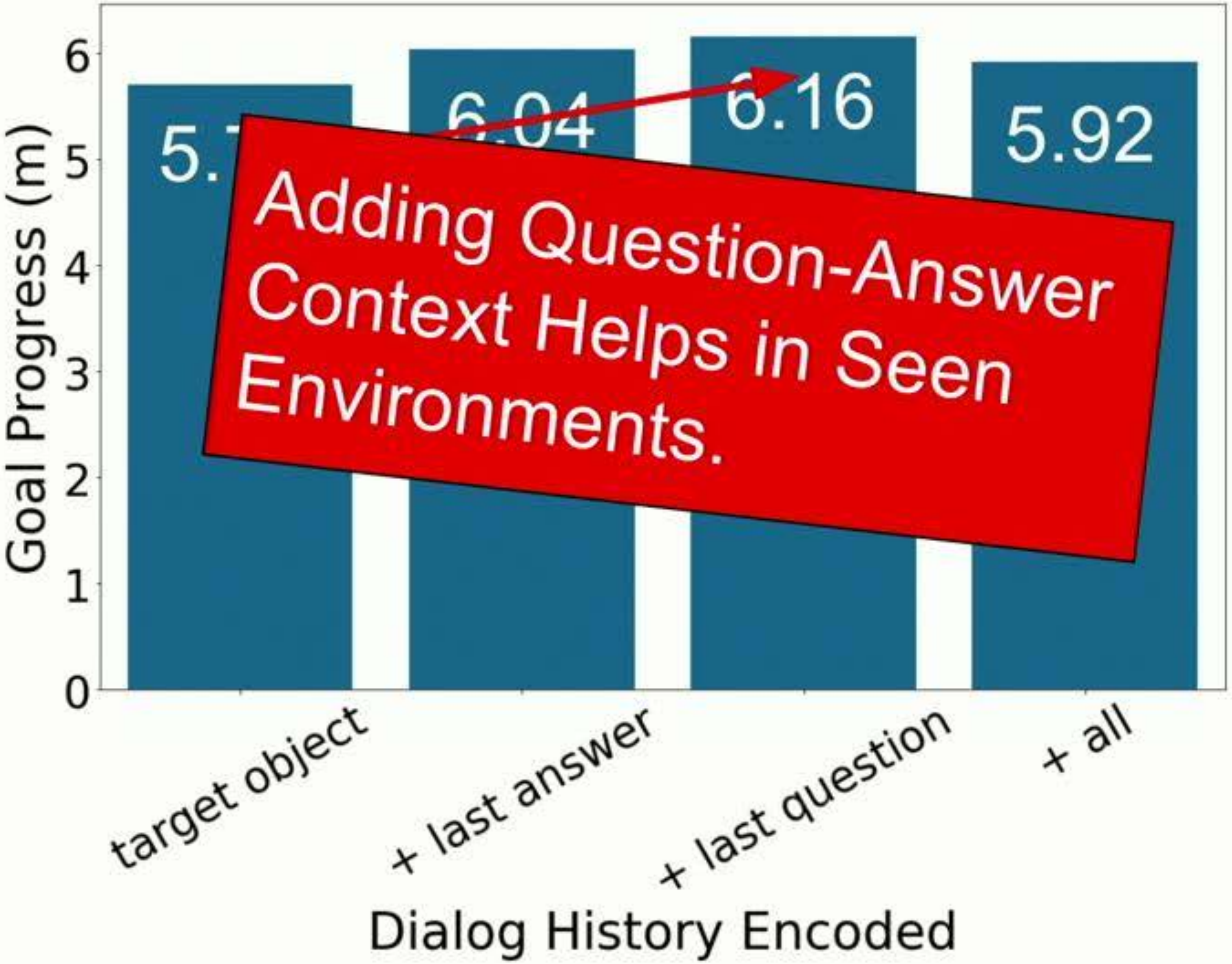
- **Seen Environments:**
  - Novel dialogs.
  - Houses seen at training time.

Evaluation - Validation (best perf over 200 epochs)



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

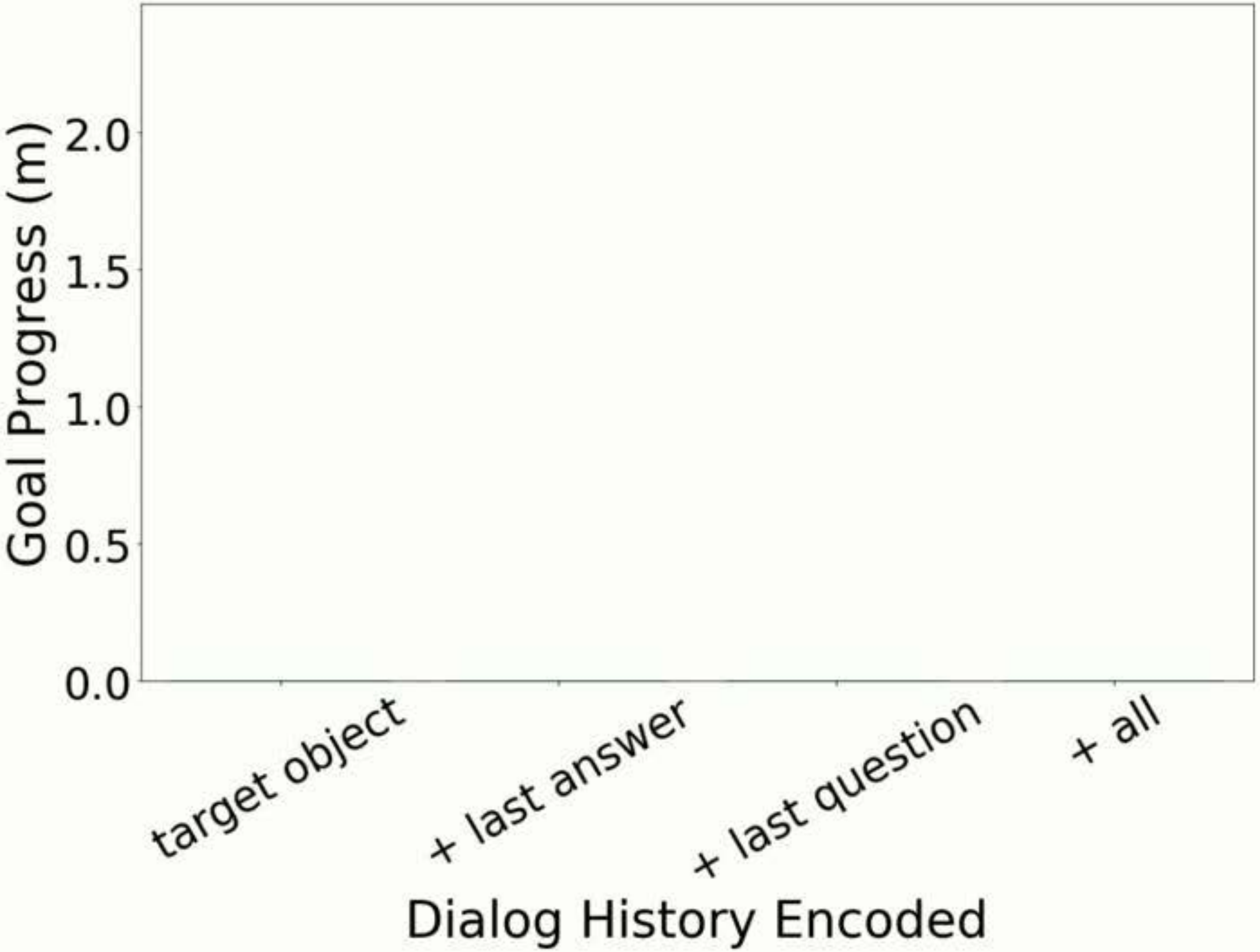
Evaluation - Validation (best perf over 200 epochs)



Adding Question-Answer Context Helps in Seen Environments.

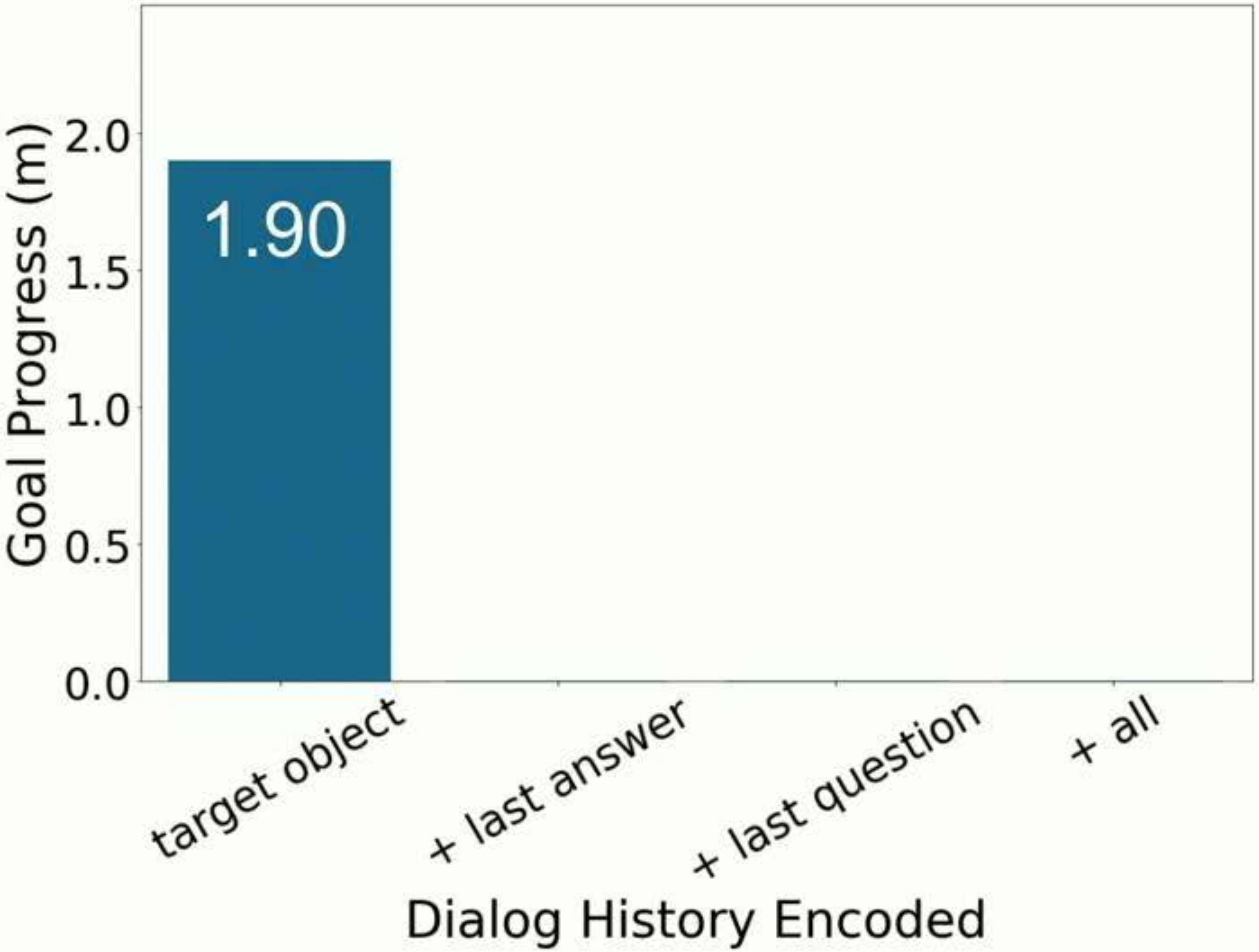
- **Seen Environments:**
  - ○ Novel dialogs.
  - ○ Houses seen at training time.

Evaluation - Test (epoch of best Val Unseen)



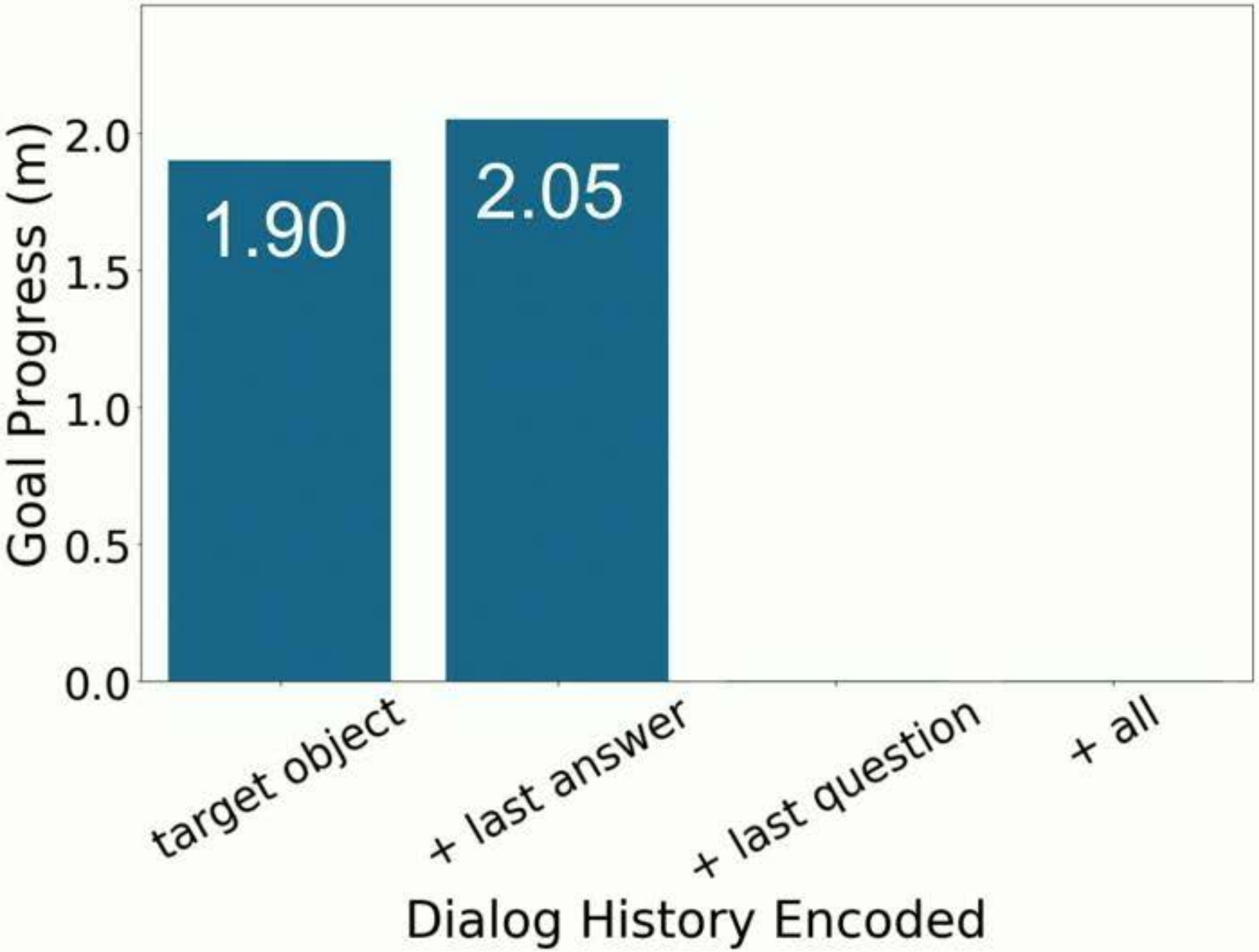
- **Unseen Envs:**
  - Novel dialogs.
  - Novel houses not seen during training.

Evaluation - Test (epoch of best Val Unseen)



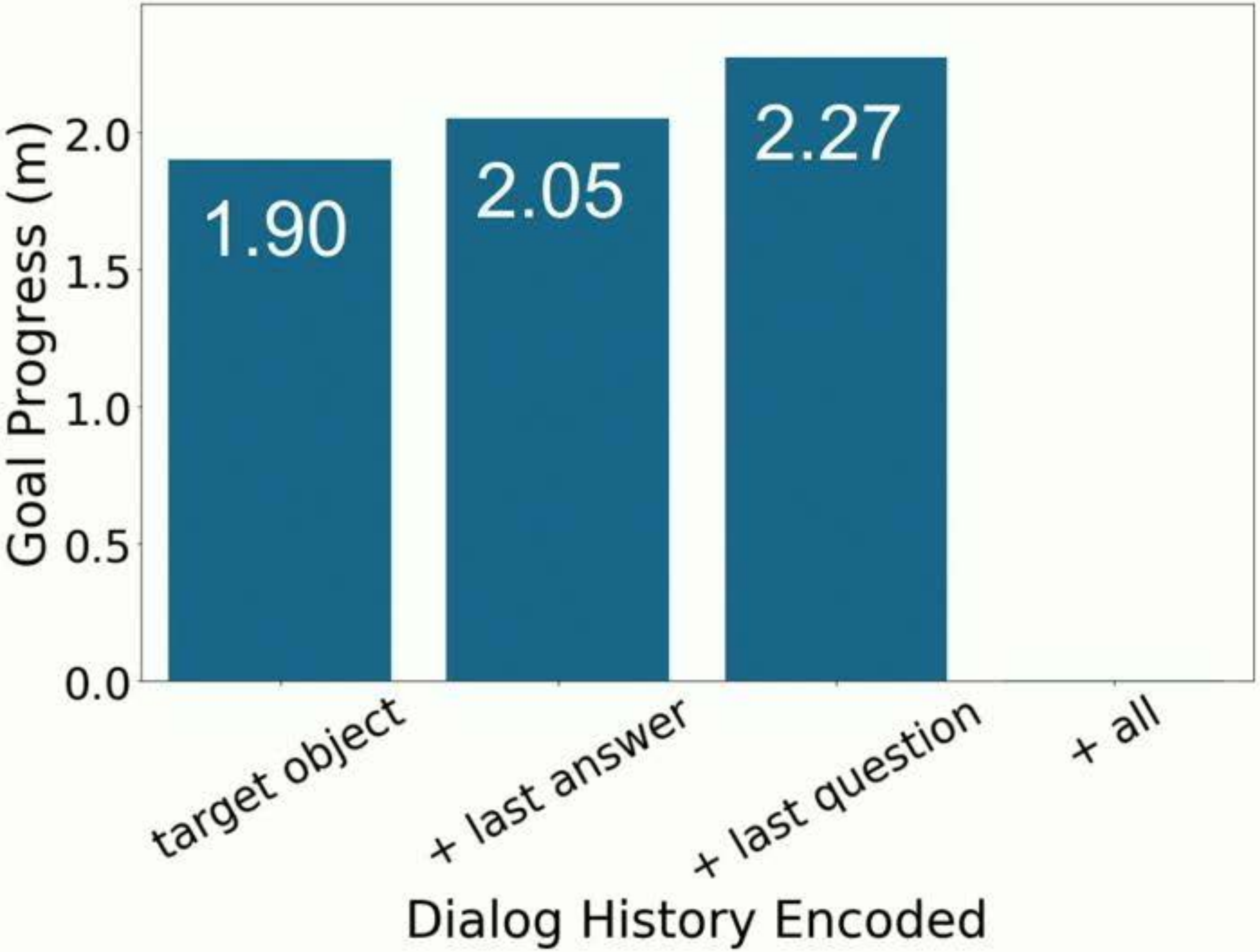
- **Unseen Envs:**
  - Novel dialogs.
  - Novel houses not seen during training.

Evaluation - Test (epoch of best Val Unseen)



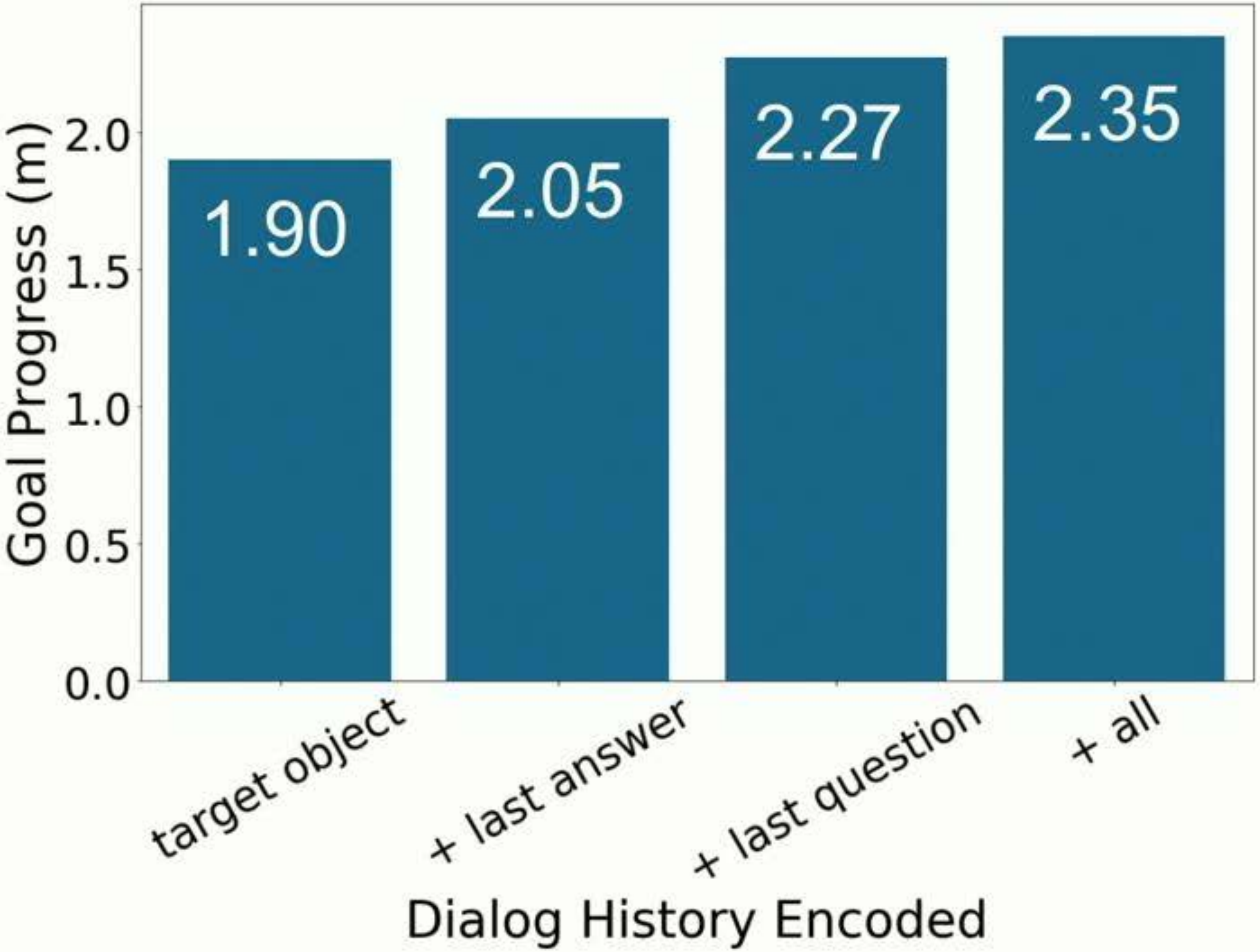
- **Unseen Envs:**
    - ○ Novel dialogs.
    - ○ Novel houses not seen during training.

Evaluation - Test (epoch of best Val Unseen)



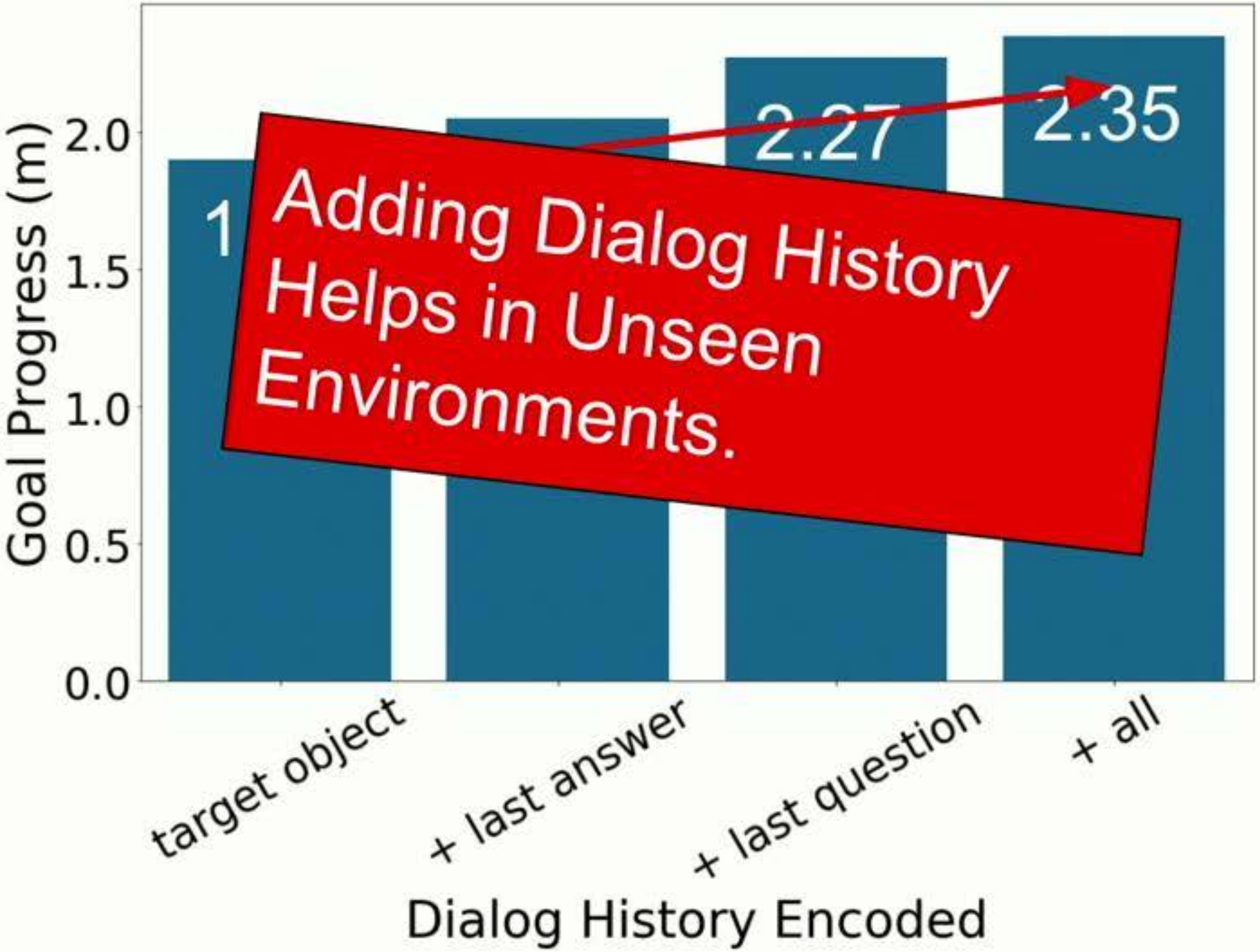
- **Unseen Envs:**
  - ○ Novel dialogs.
  - ○ Novel houses not seen during training.

Evaluation - Test (epoch of best Val Unseen)



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

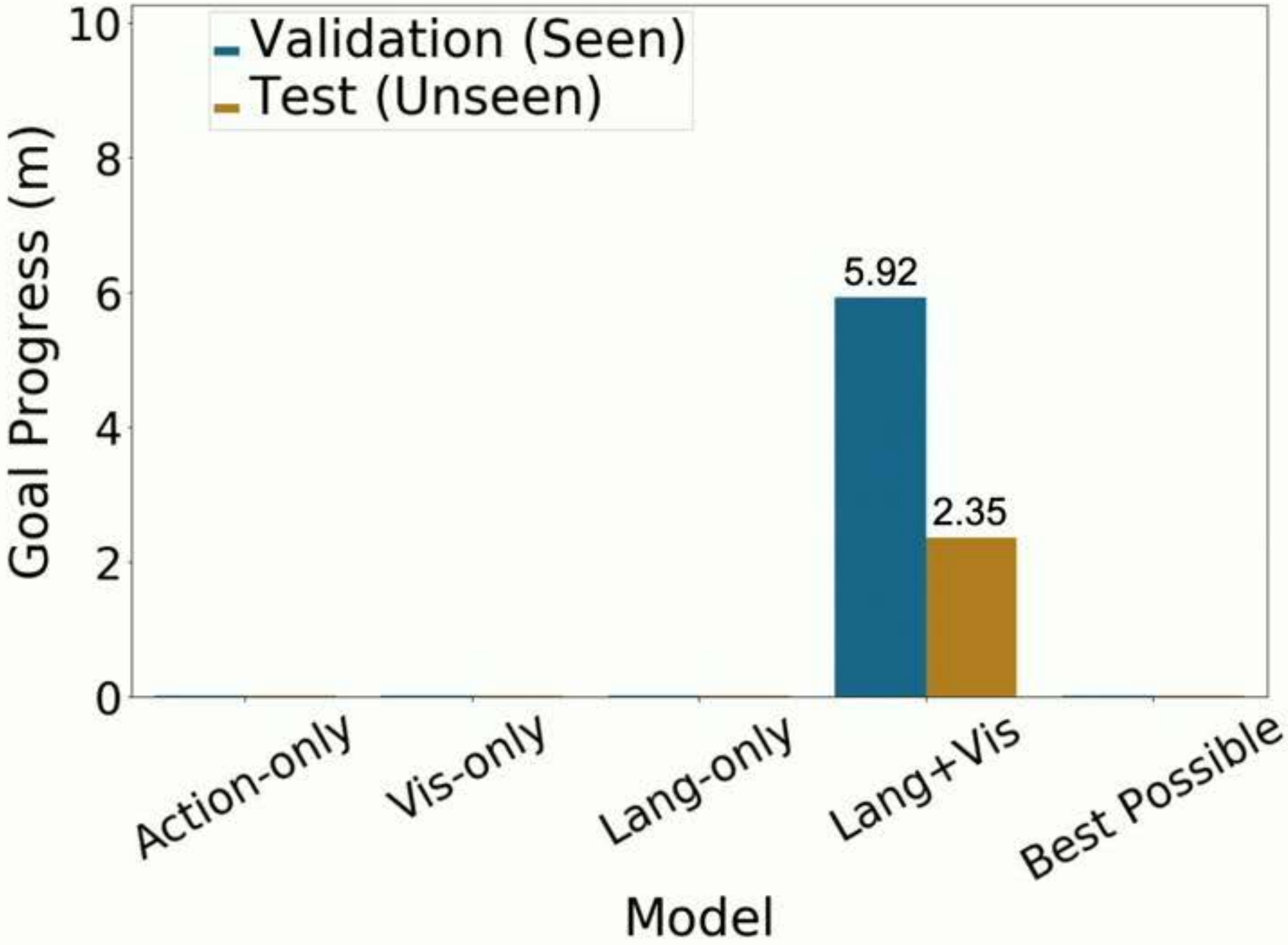
Evaluation - Test (epoch of best Val Unseen)



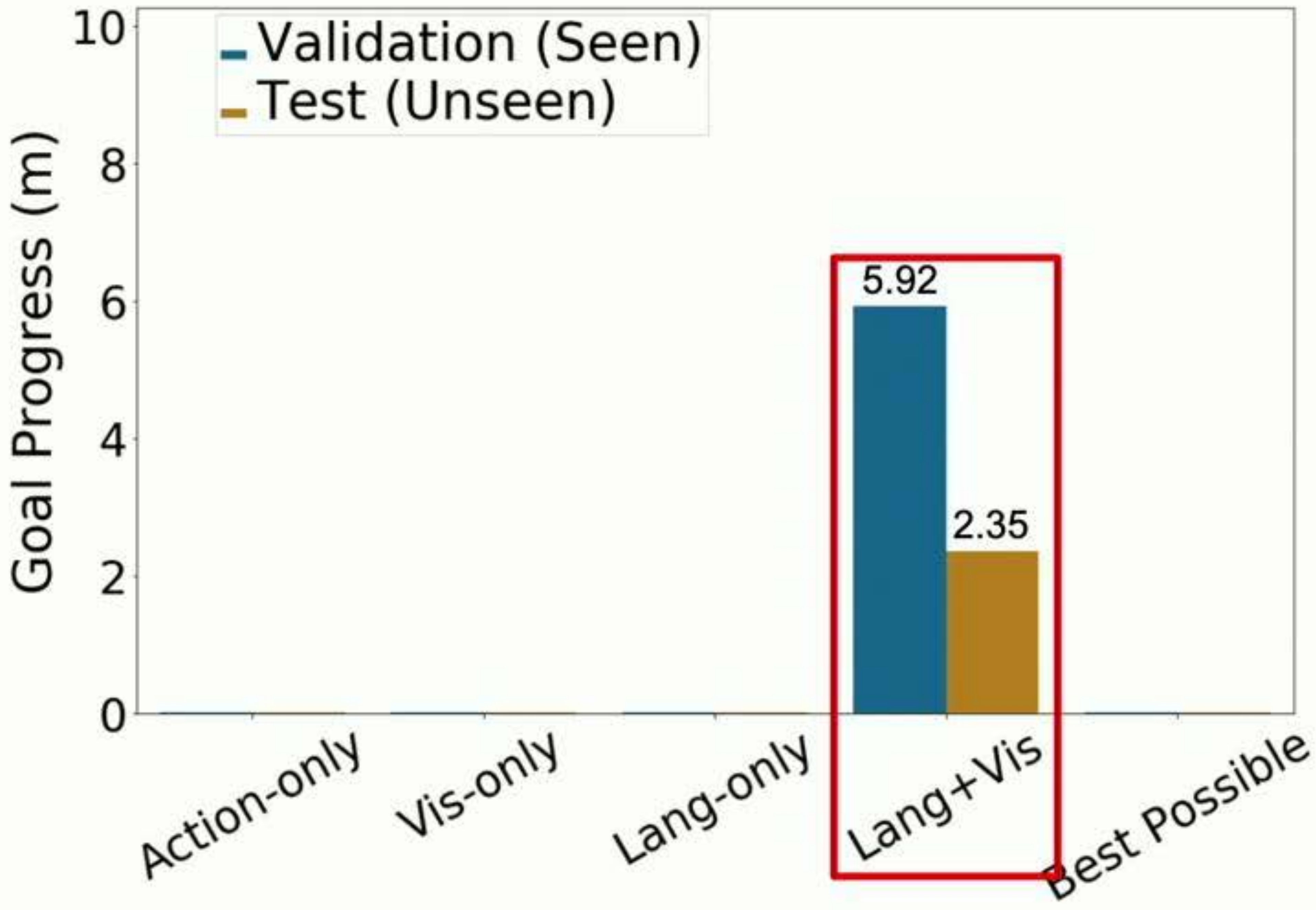
Adding Dialog History Helps in Unseen Environments.

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

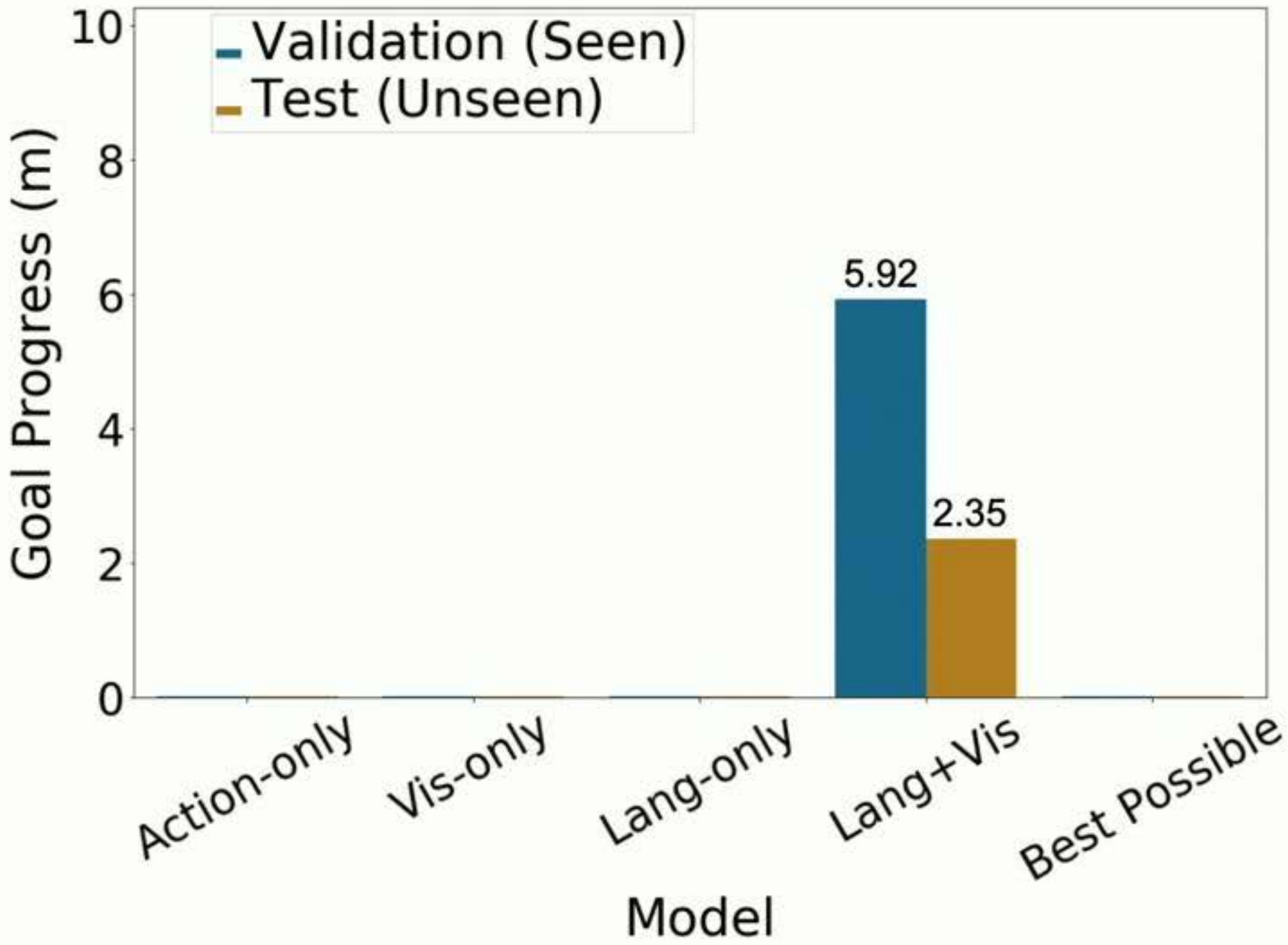
Evaluation - Unimodal Baselines



Evaluation - Unimodal Baselines

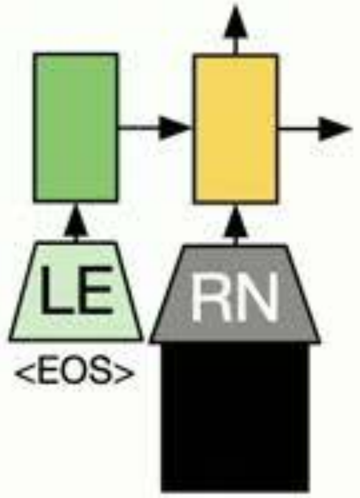


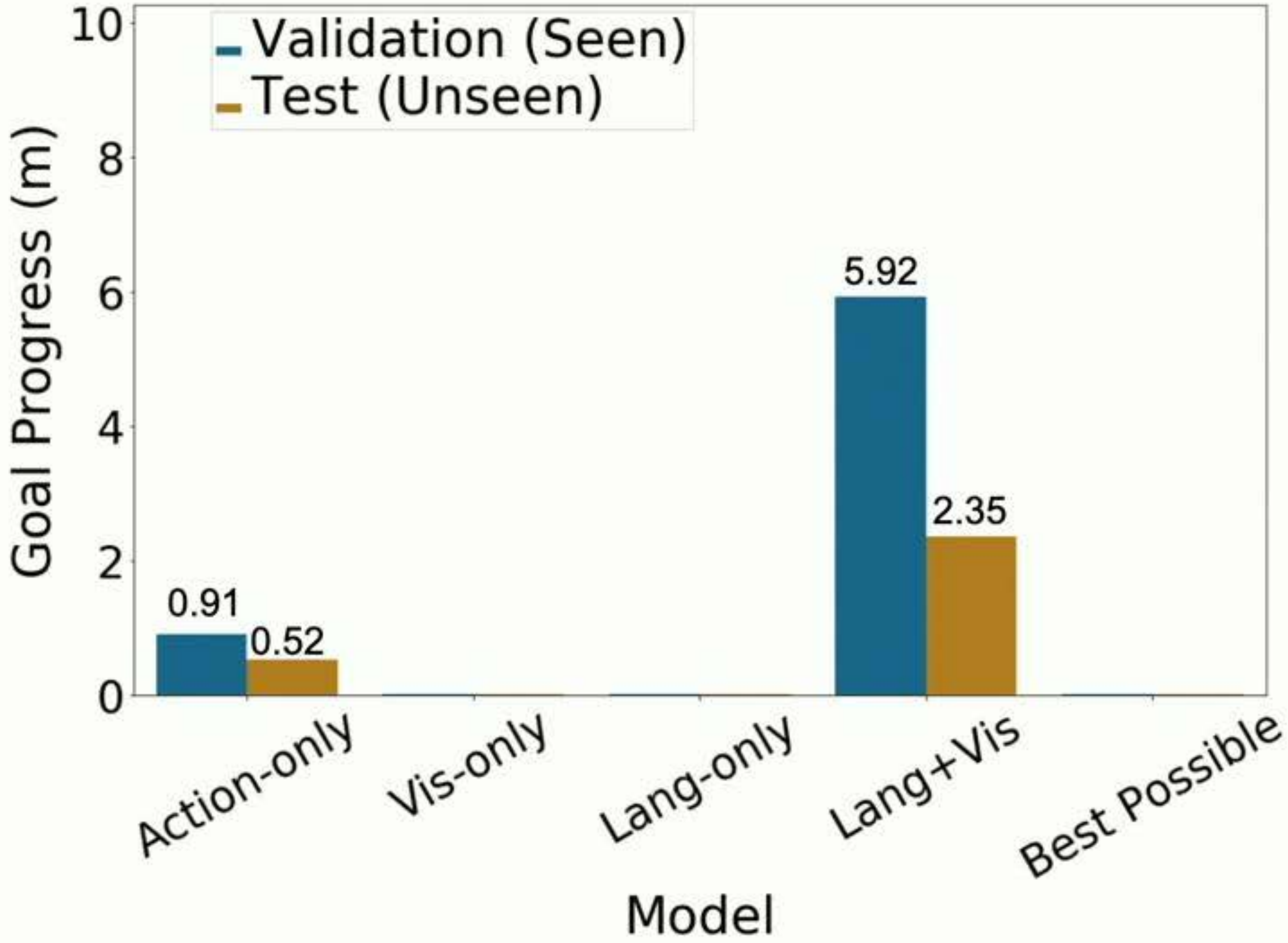
Evaluation - Unimodal Baselines



Evaluation - Unimodal Baselines

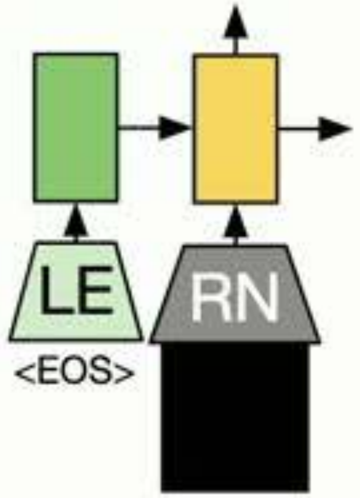
Action-only



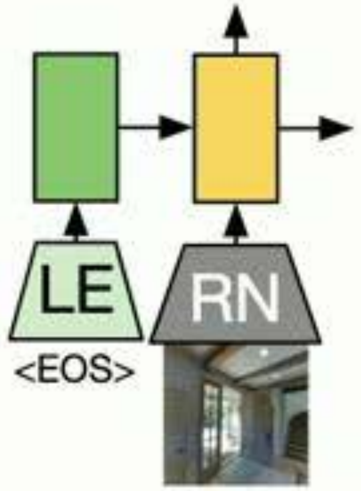


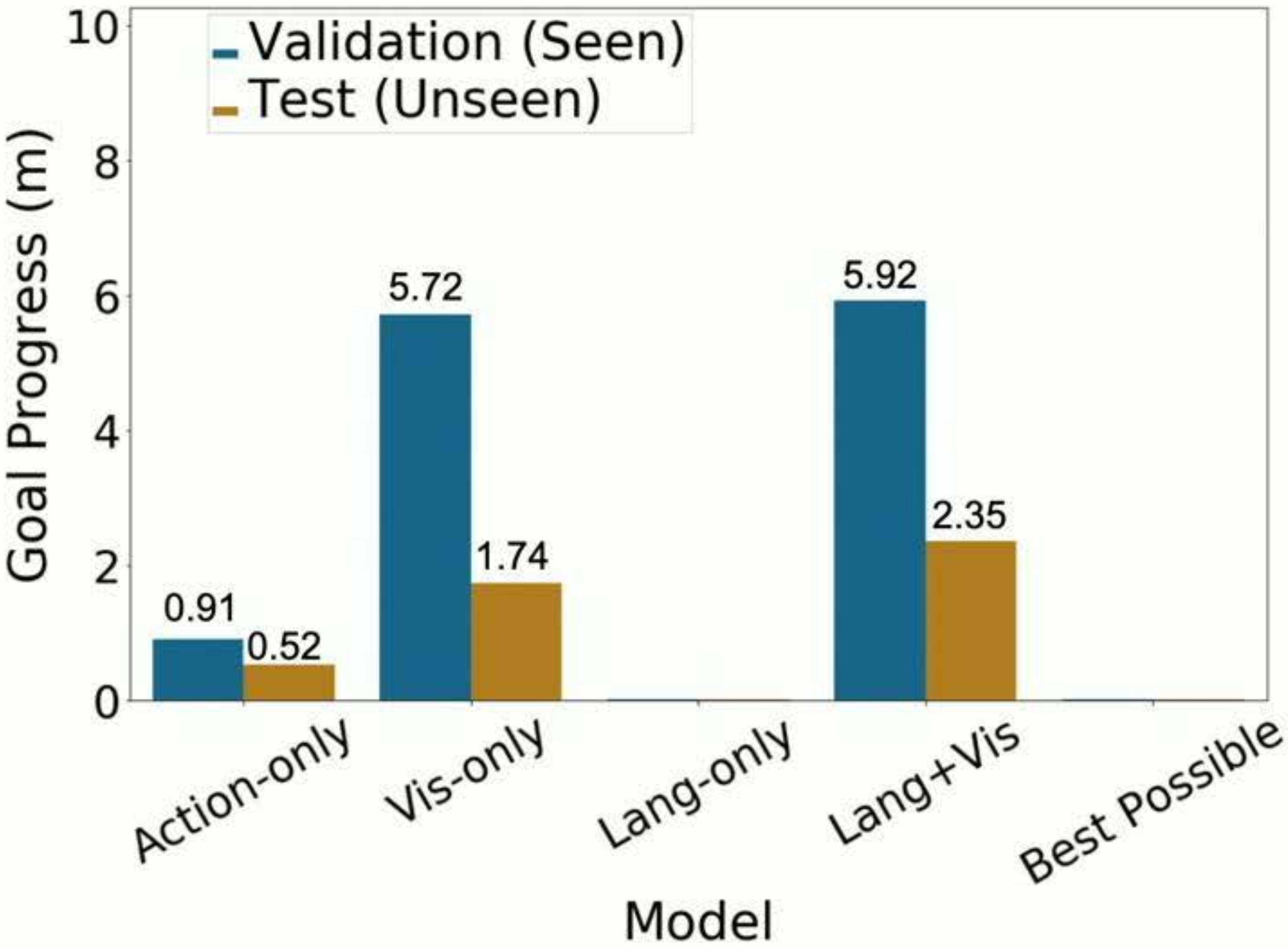
Evaluation - Unimodal Baselines

Action-only



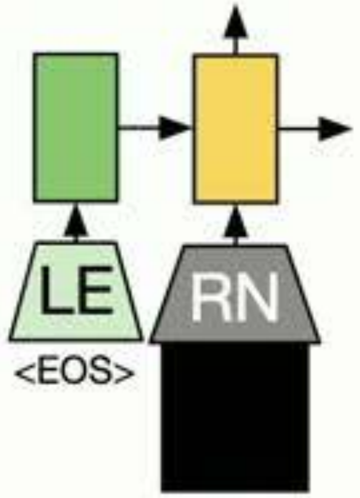
Vis-only



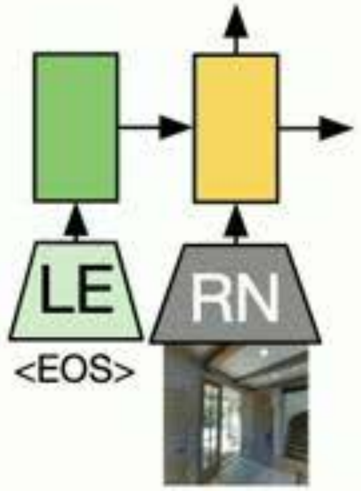


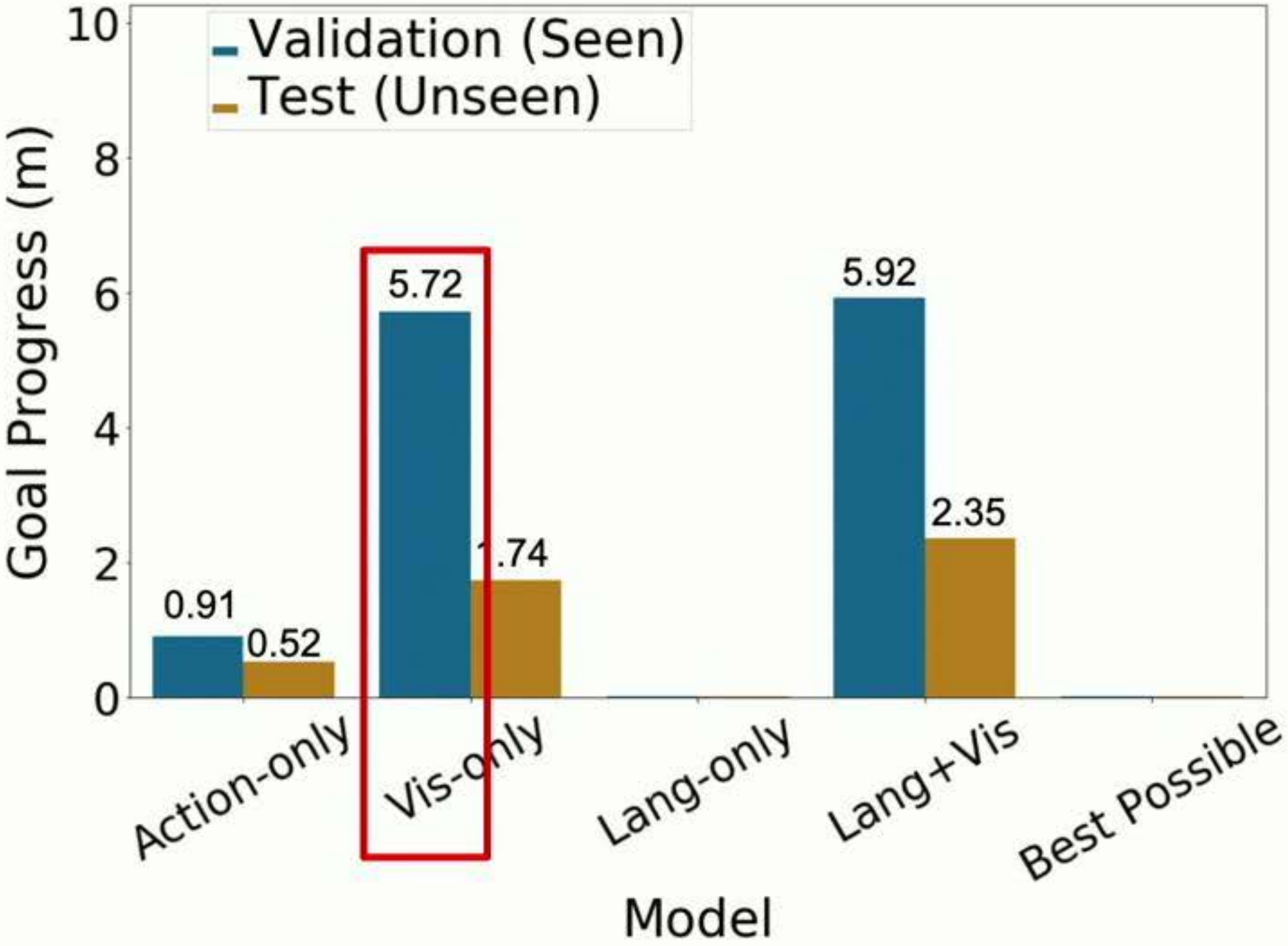
Evaluation - Unimodal Baselines

Action-only



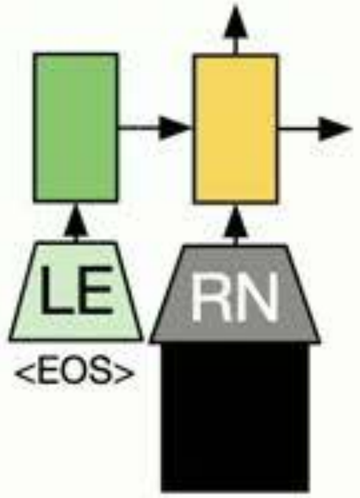
Vis-only



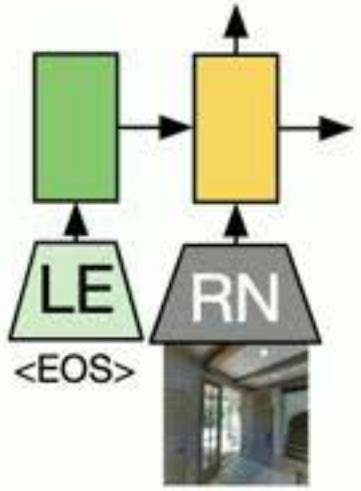


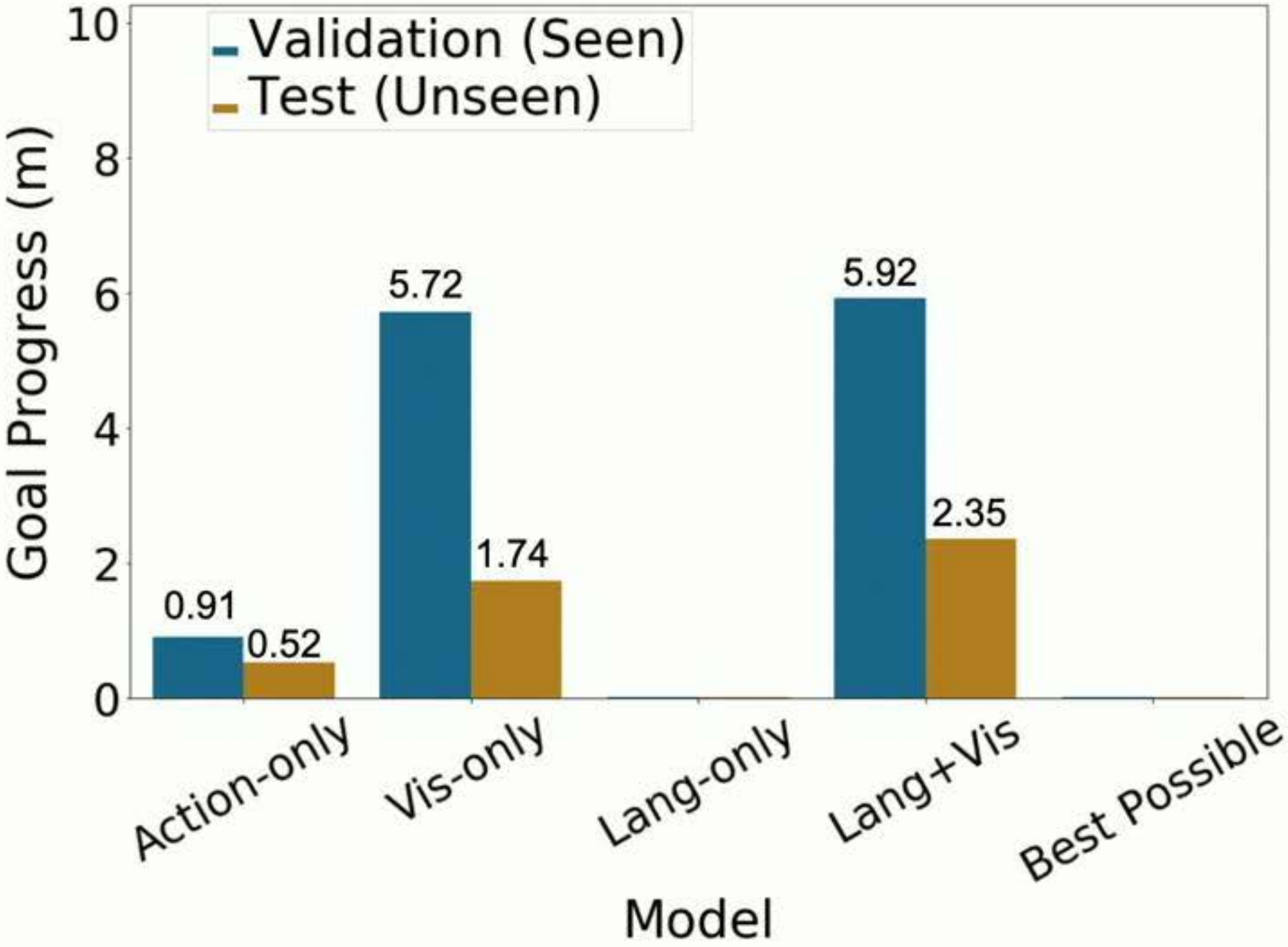
Evaluation - Unimodal Baselines

Action-only



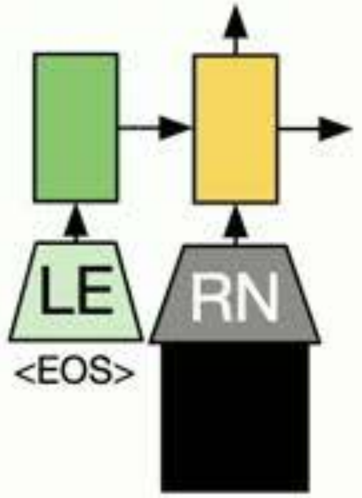
Vis-only



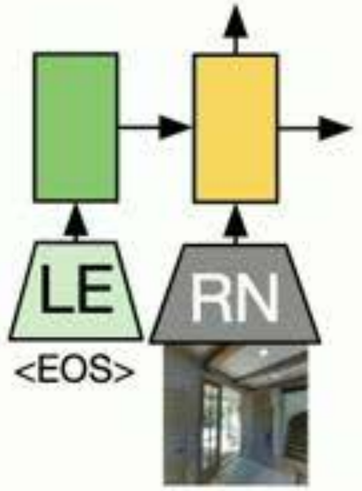


Evaluation - Unimodal Baselines

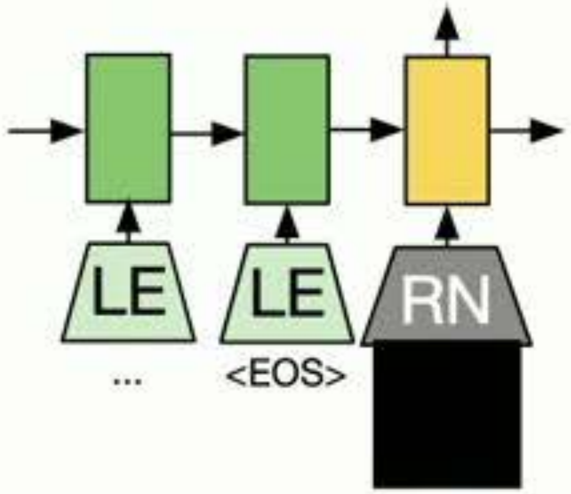
Action-only

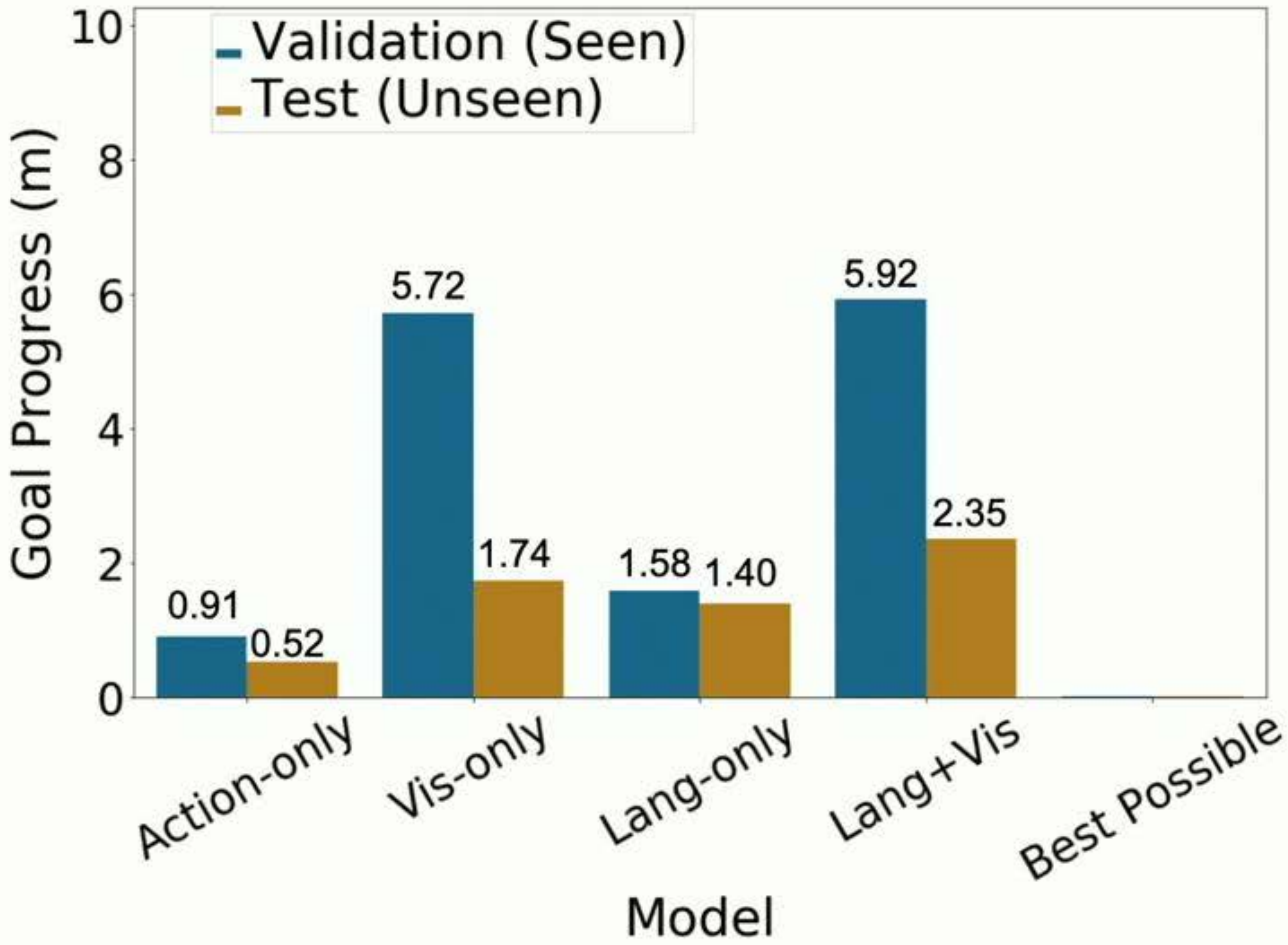


Vis-only



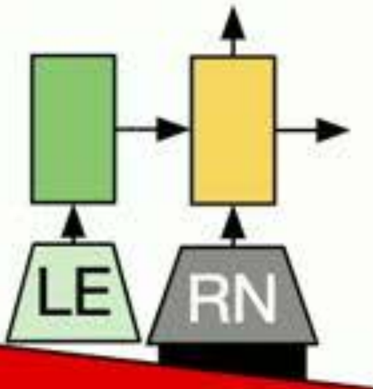
Lang-only





Evaluation - Unimodal Baselines

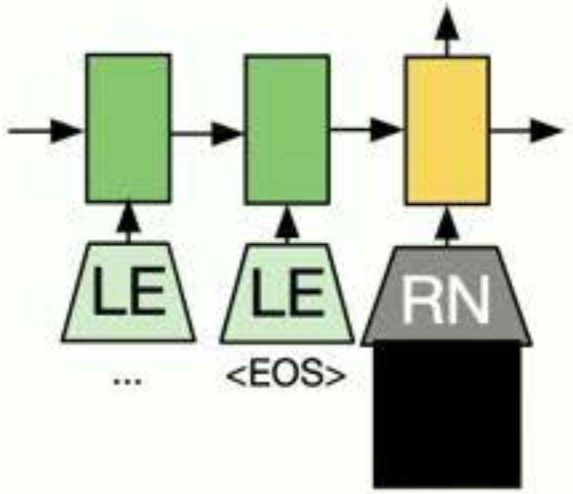
Action-only

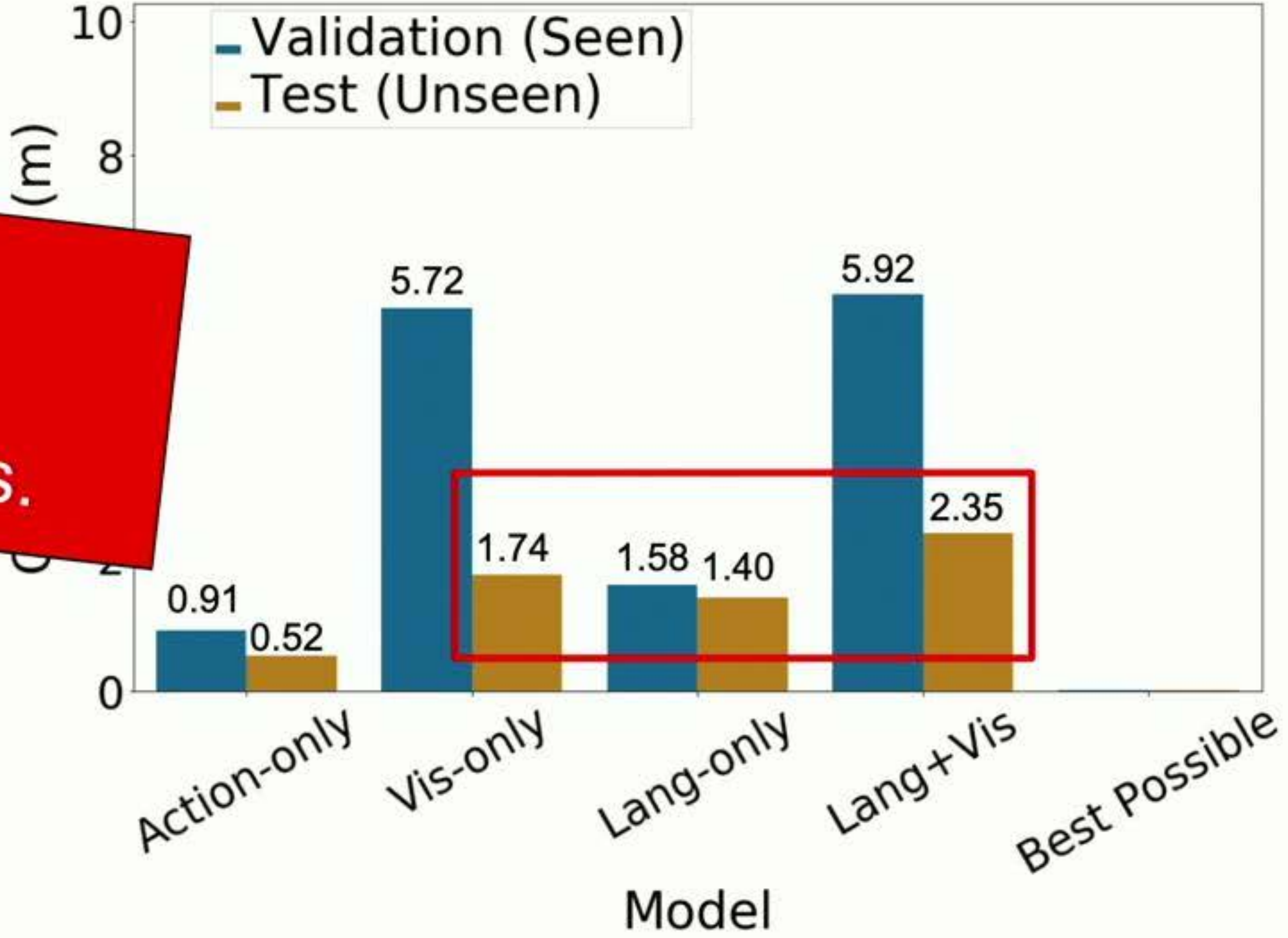


Initial Model Uses Multimodal Input in Unseen Environments.

only

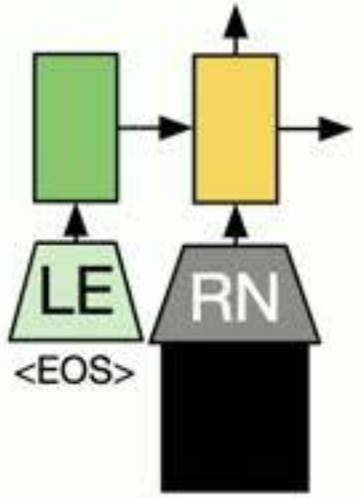
Lang-only



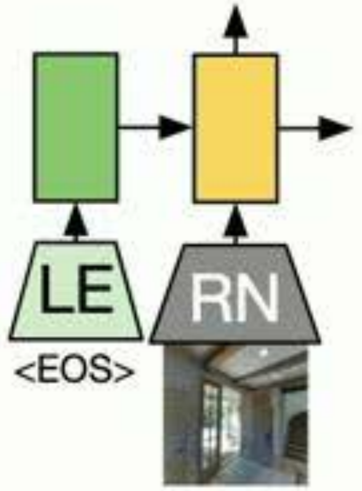


Evaluation - Unimodal Baselines

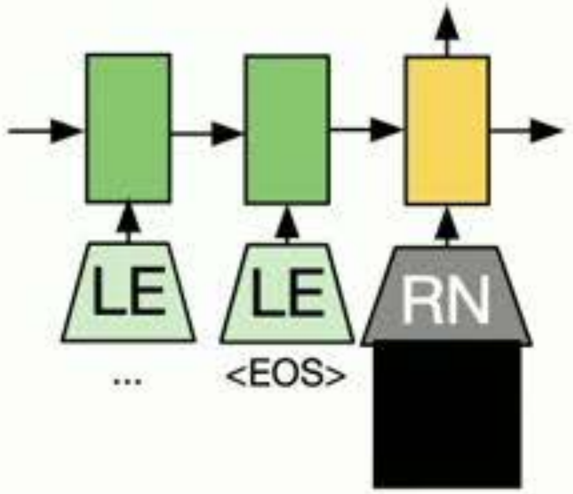
Action-only

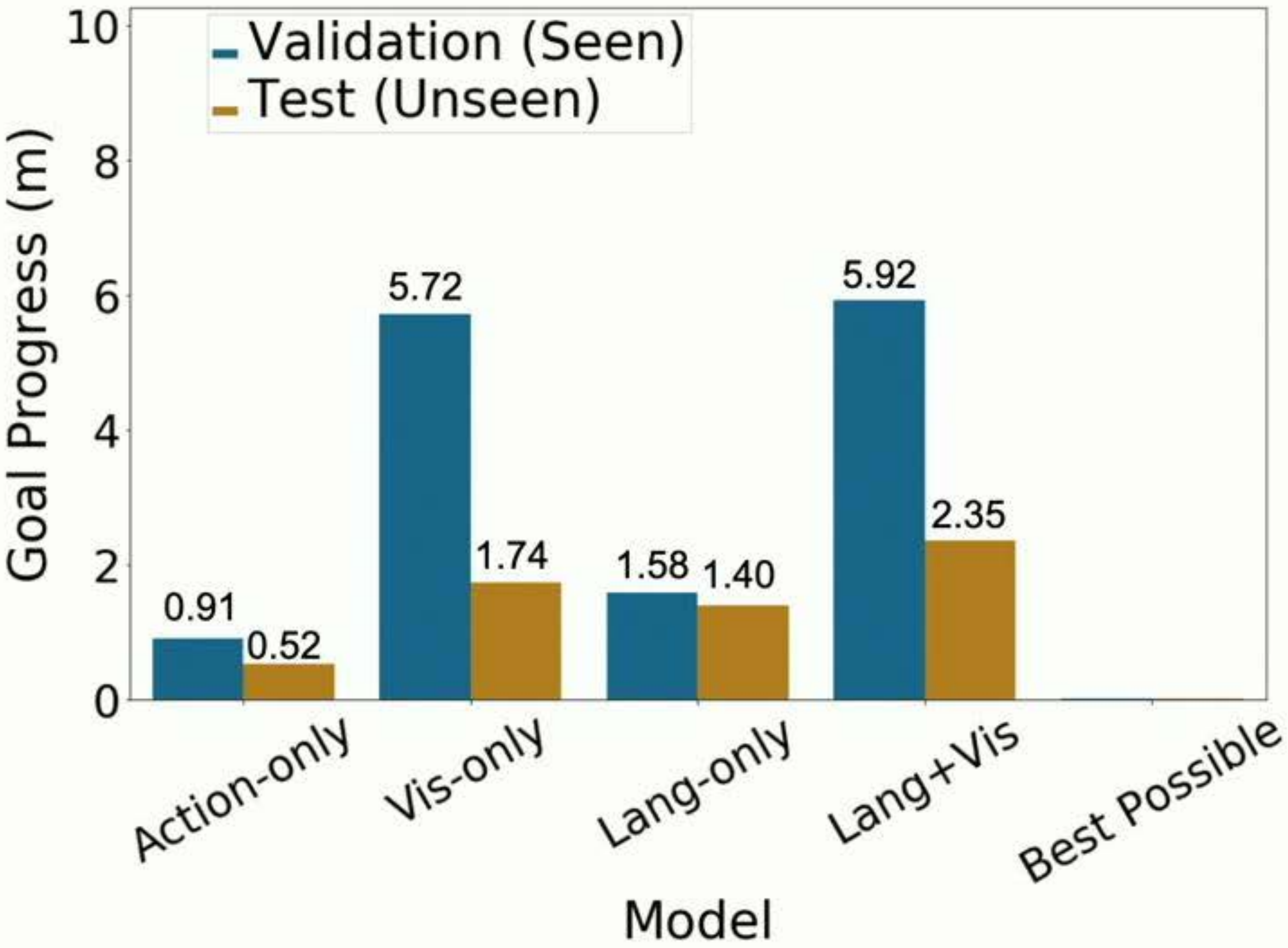


Vis-only



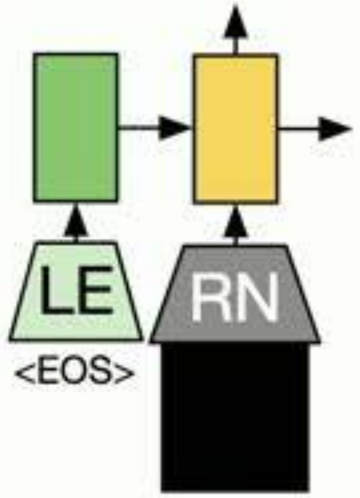
Lang-only



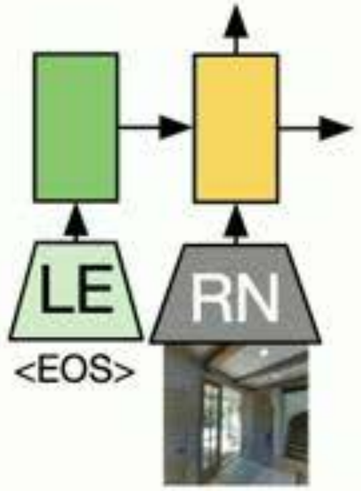


Evaluation - Unimodal Baselines

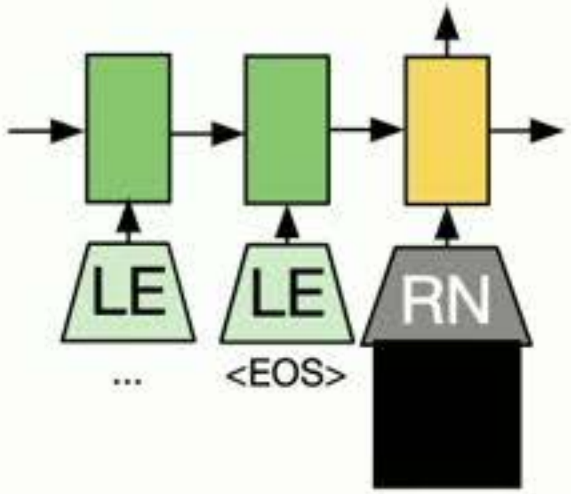
Action-only

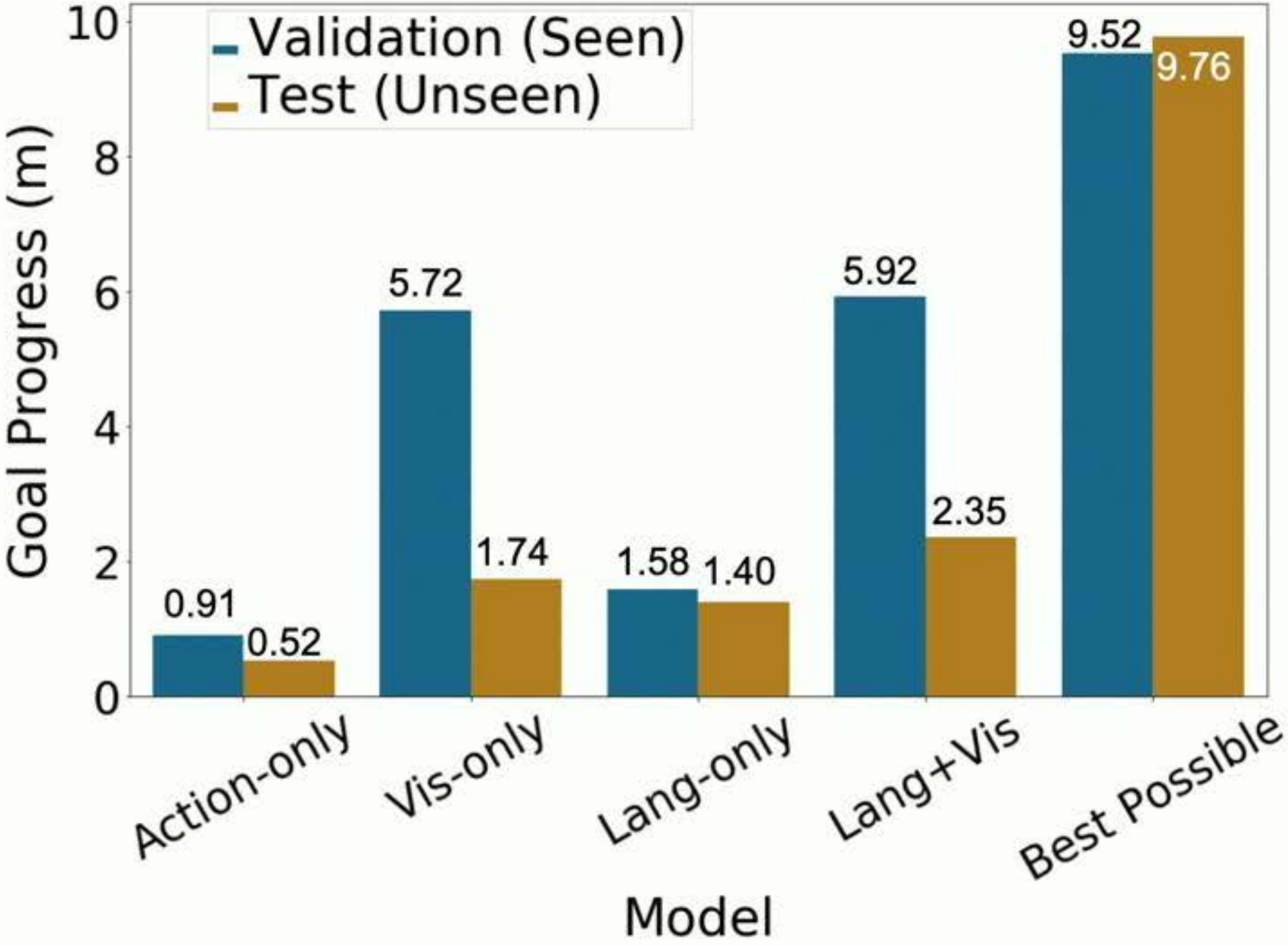


Vis-only



Lang-only



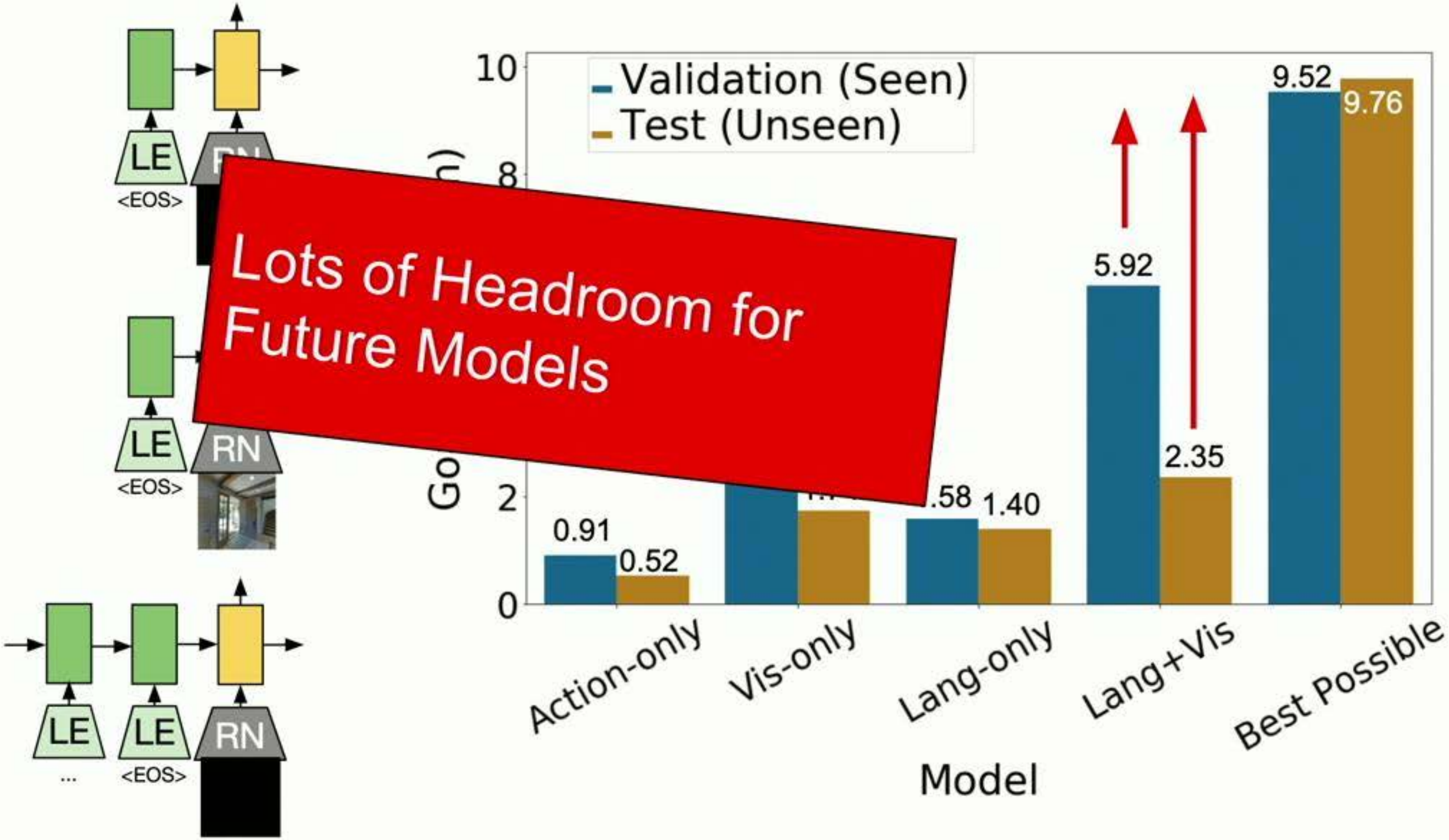


Evaluation - Unimodal Baselines

Action-only

Vis-only

Lang-only



Lots of Headroom for Future Models

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.

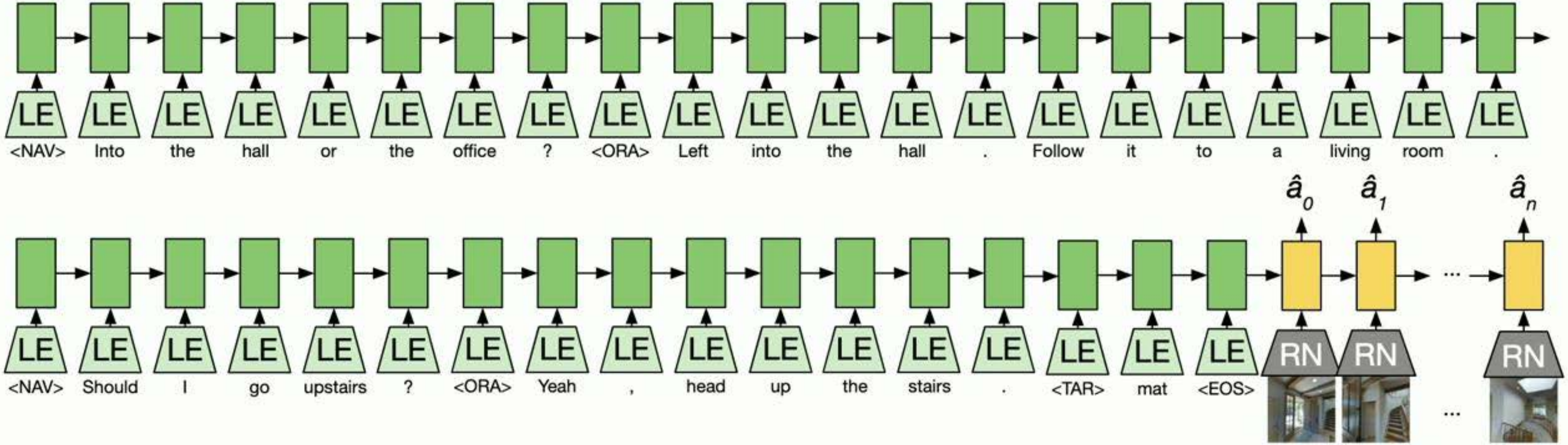
Evaluation Lessons

- 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

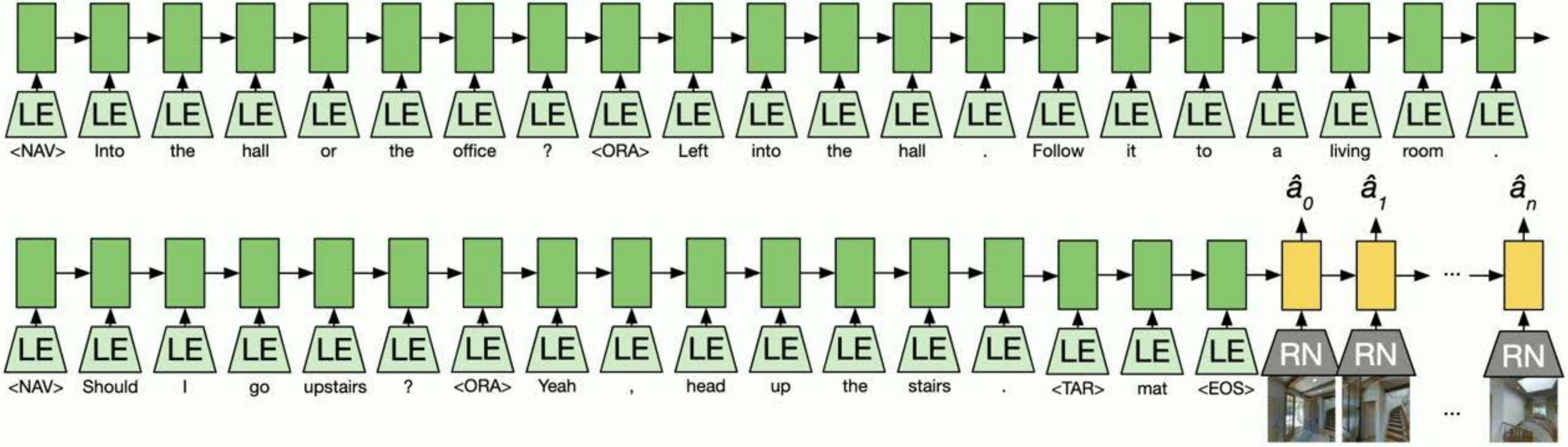
- Language grounding in visual environments
  - For navigation
  - Unimodal bias [Thomason et al., NAACL'19]
- Vision-and-Dialog Navigation [Thomason et al., *in sub*]
  - New dataset - CVDN
  - Navigation from dialog history
- Next steps

Incorporating Navigation History

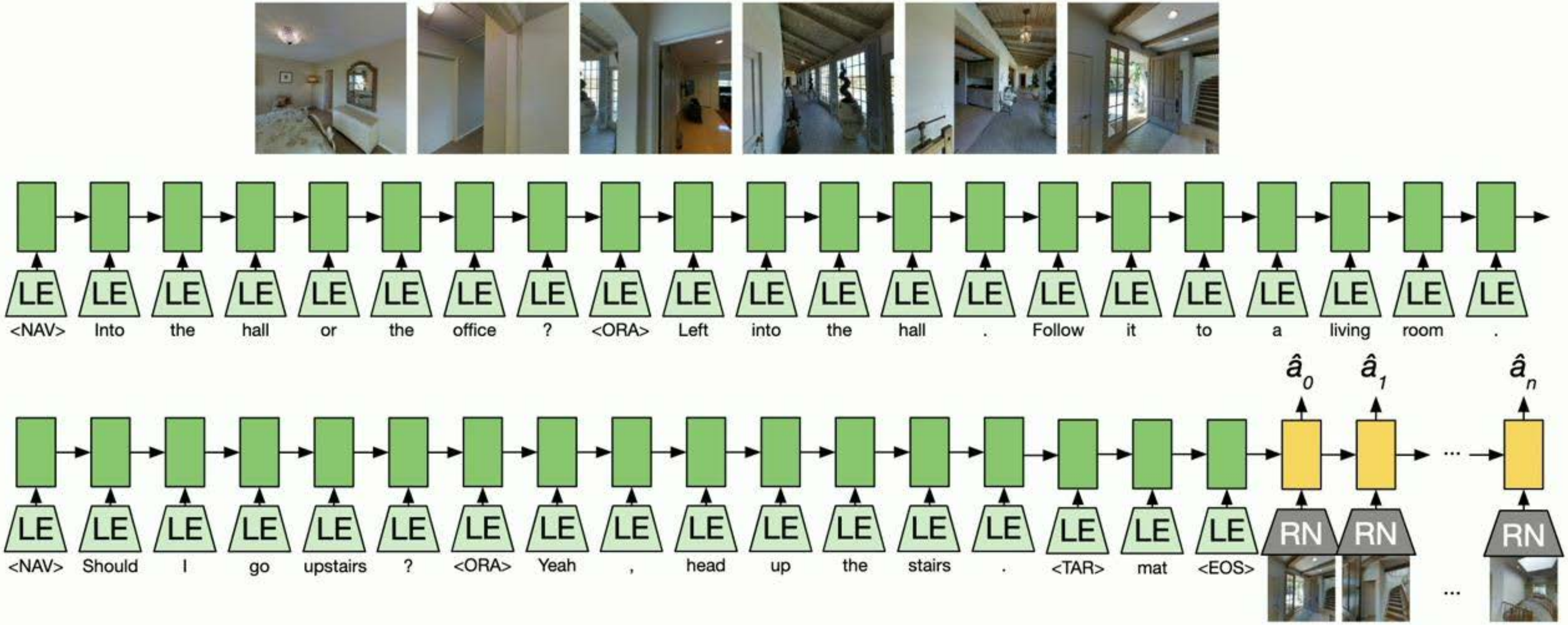


Incorporating Navigation History

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

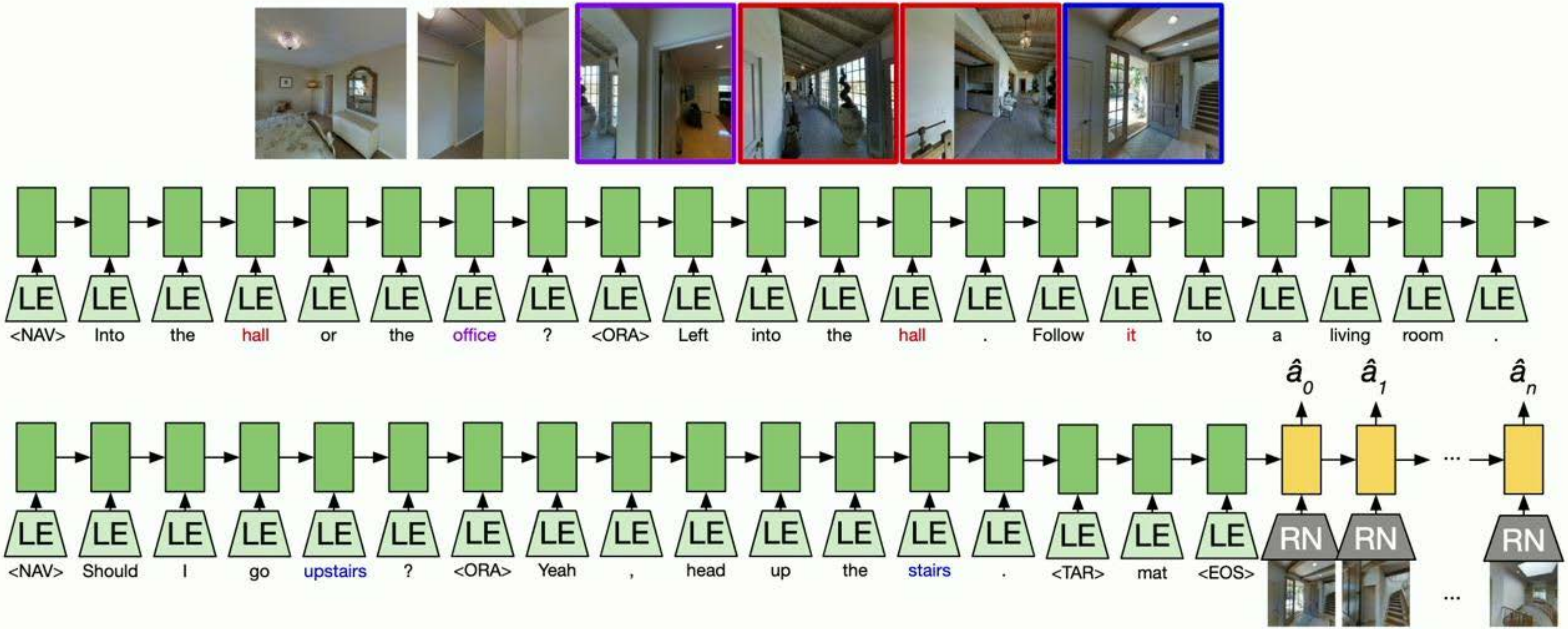


Incorporating Navigation History



LE
<NAV> Into the hall or the office ? <ORA> Left into the hall . Follow it to a living room .
$\hat{a}_0$ $\hat{a}_1$ $\hat{a}_n$
<NAV> Should I go upstairs ? <ORA> Yeah , head up the stairs . <TAR> mat <EOS>
RN
...

Incorporating Navigation History



Visible to both Navigator and Oracle

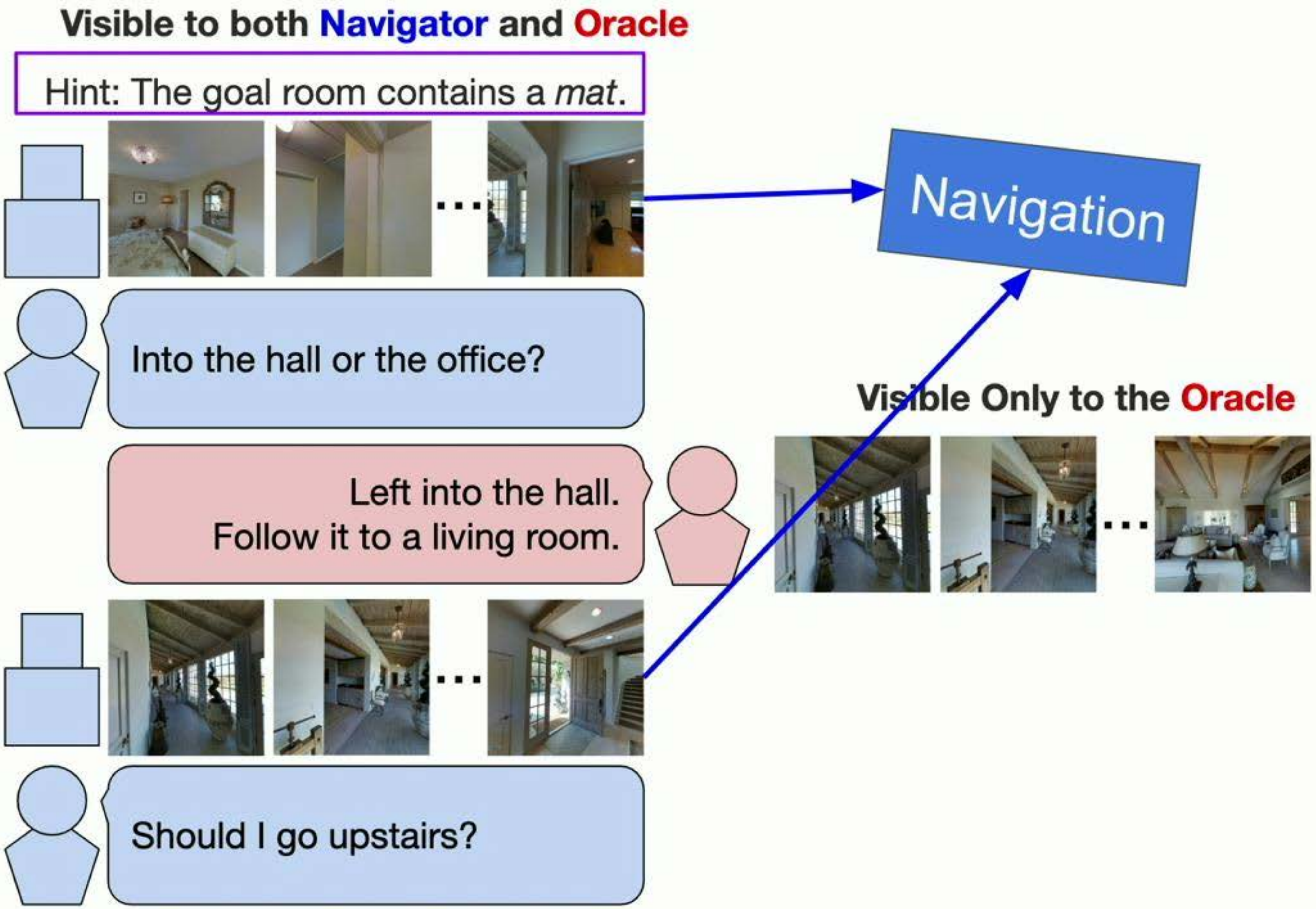
Hint: The goal room contains a *mat*.

Into the hall or the office?

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

Visible Only to the Oracle

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*.

Into the hall or the office?

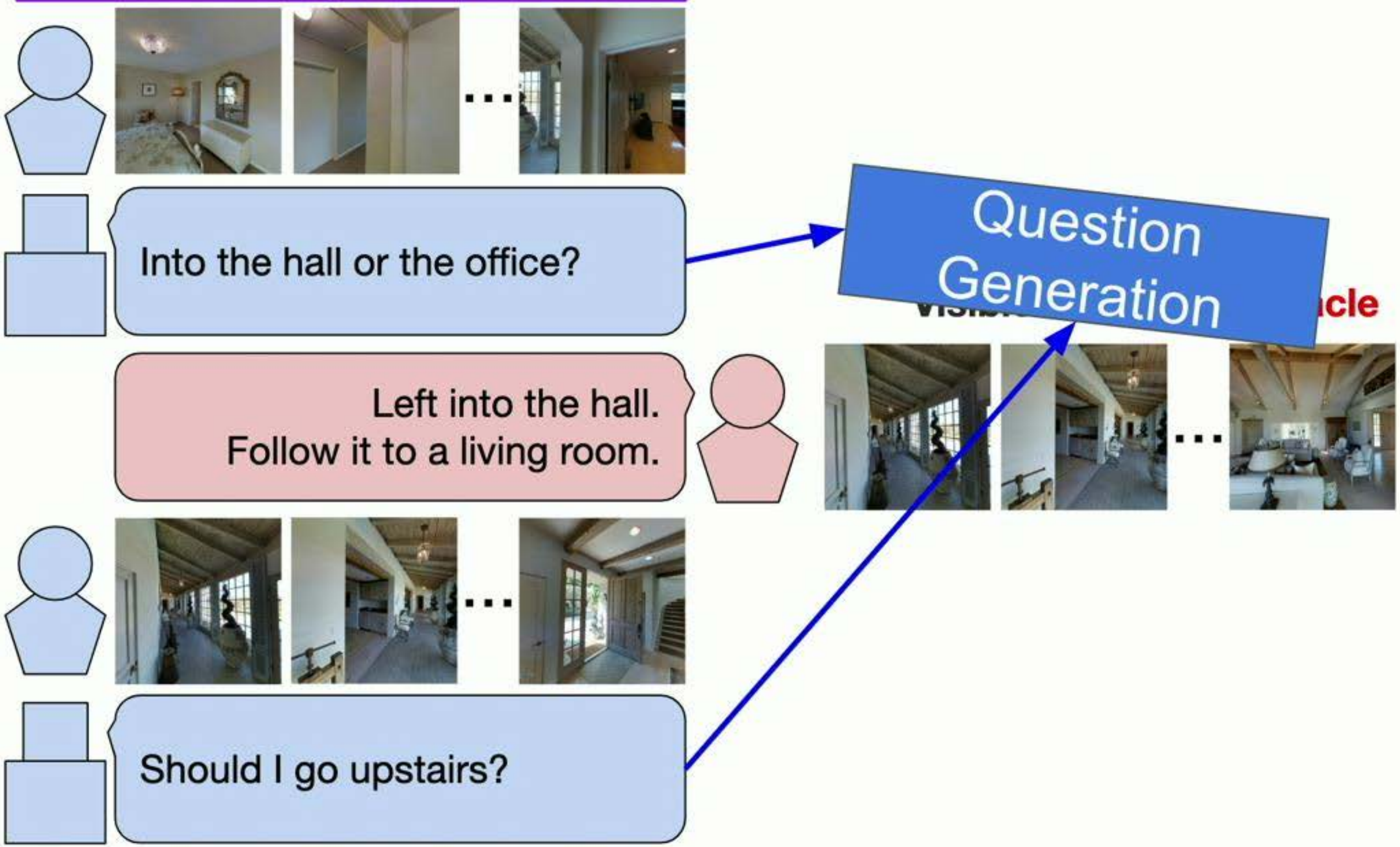
Left into the hall.
Follow it to a living room.

Visible Only to the Oracle

Should I go upstairs?

Visible to both Navigator and Oracle

Hint: The goal room contains a *mat*.



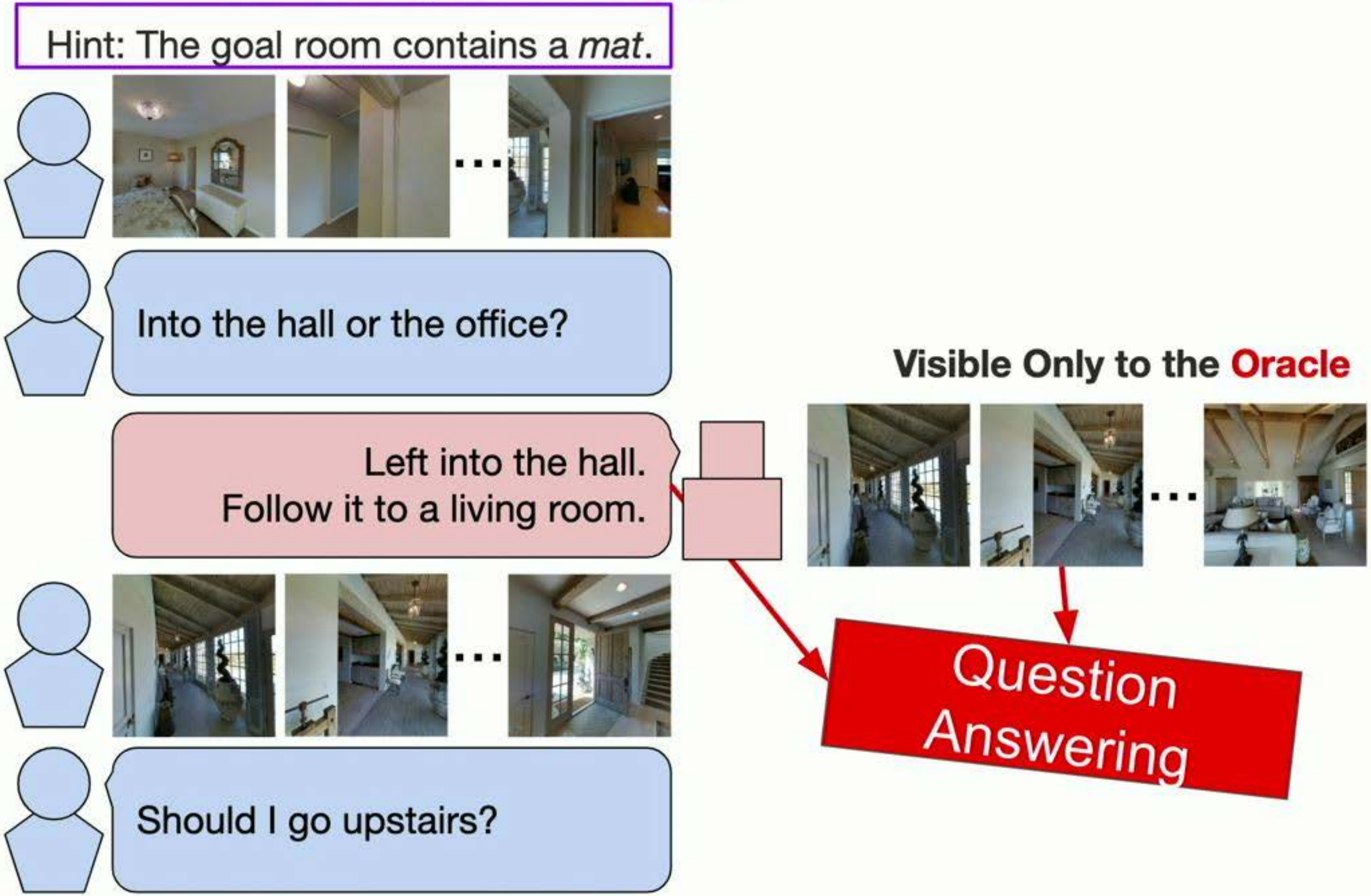
Into the hall or the office?

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

Should I go upstairs?

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.

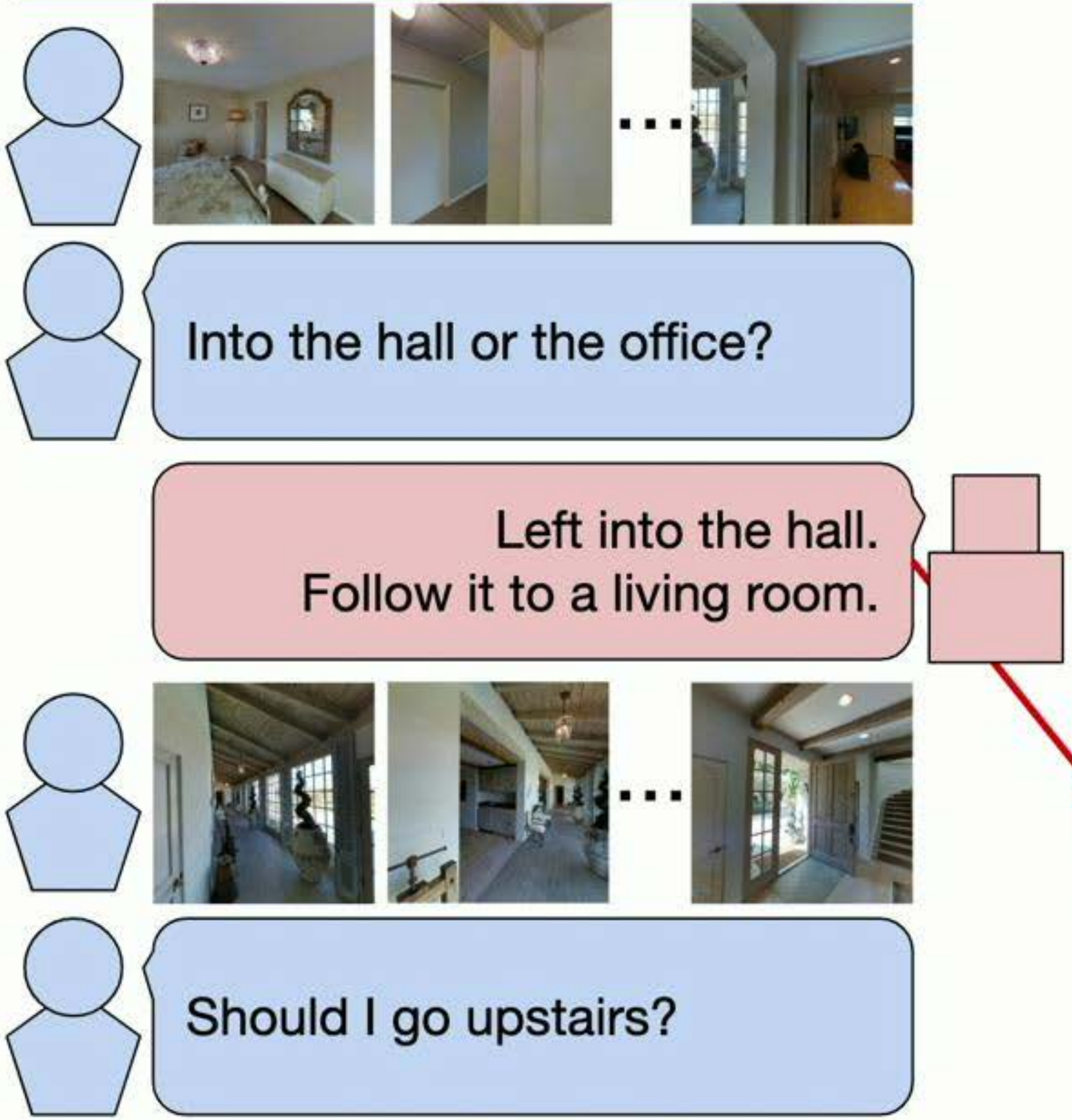
Should I go upstairs?

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?

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.

Visible Only to the Oracle

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?

Navigation

Question Generation

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?

Navigation

Question Generation

Visible Only to the Oracle

Question Answering

Hint: The goal room contains a *mat*.

Navigation

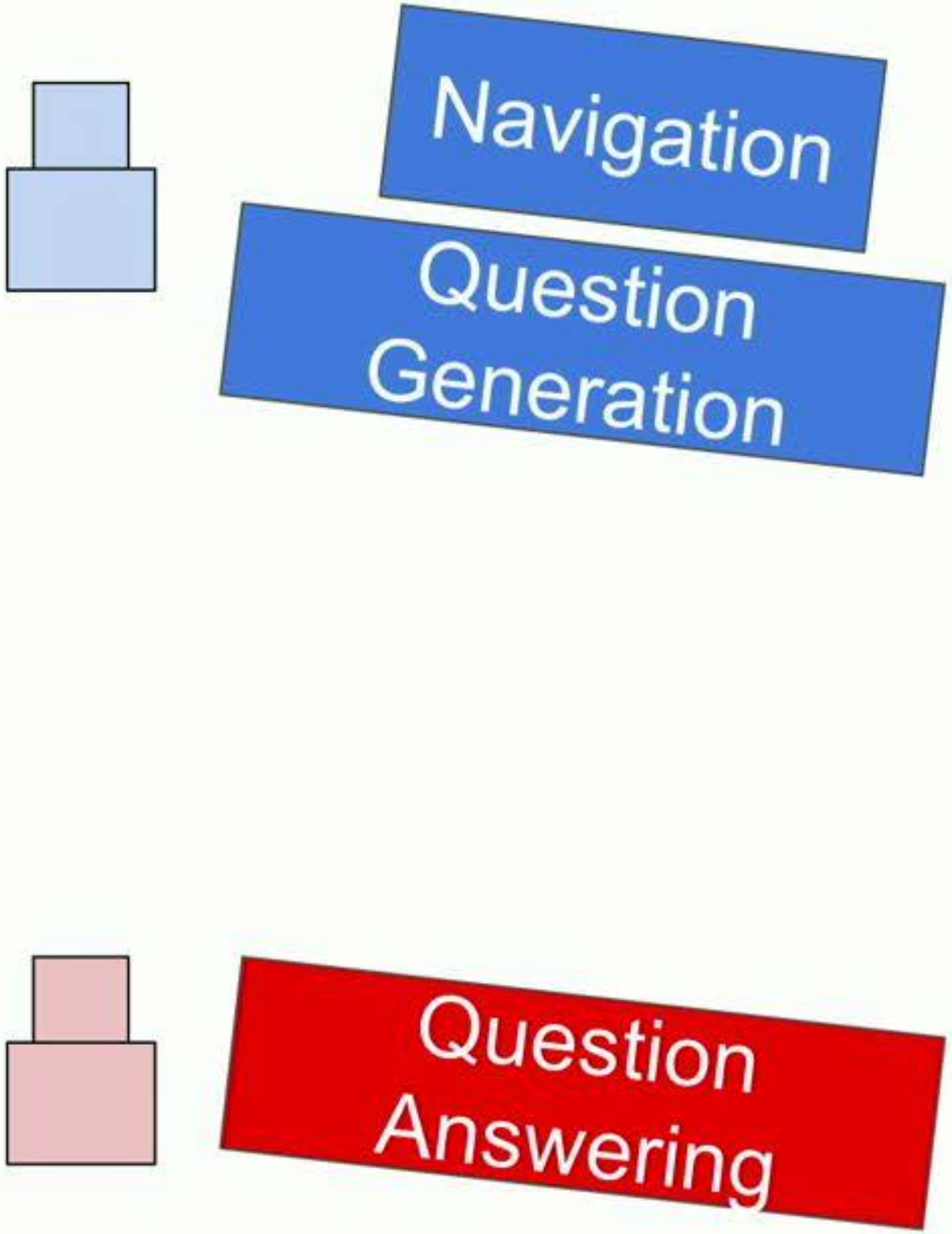
Question Generation

- Environment exploration

Question Answering

Hint: The goal room contains a *mat*.

- Environment exploration



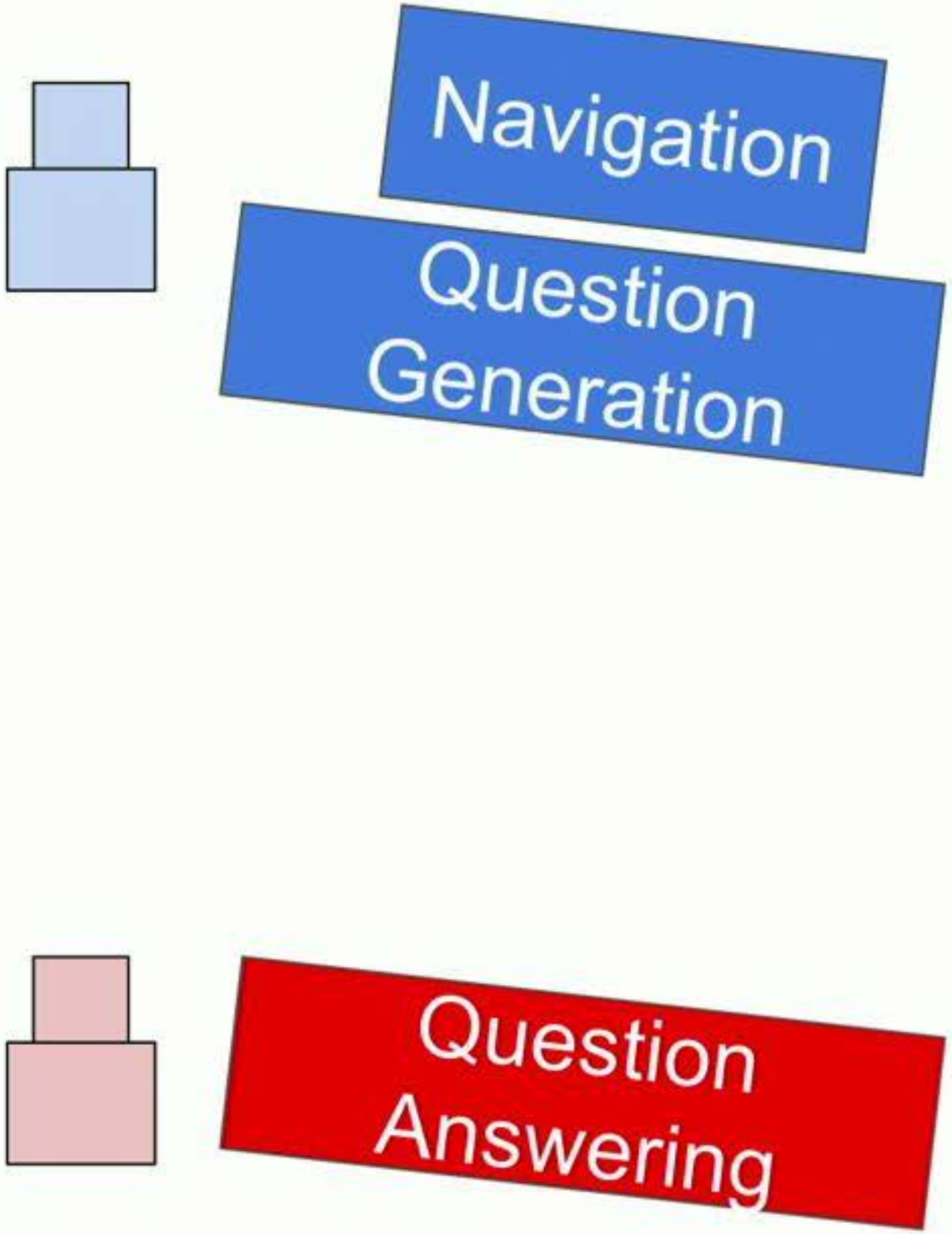
Navigation

Question Generation

Question Answering

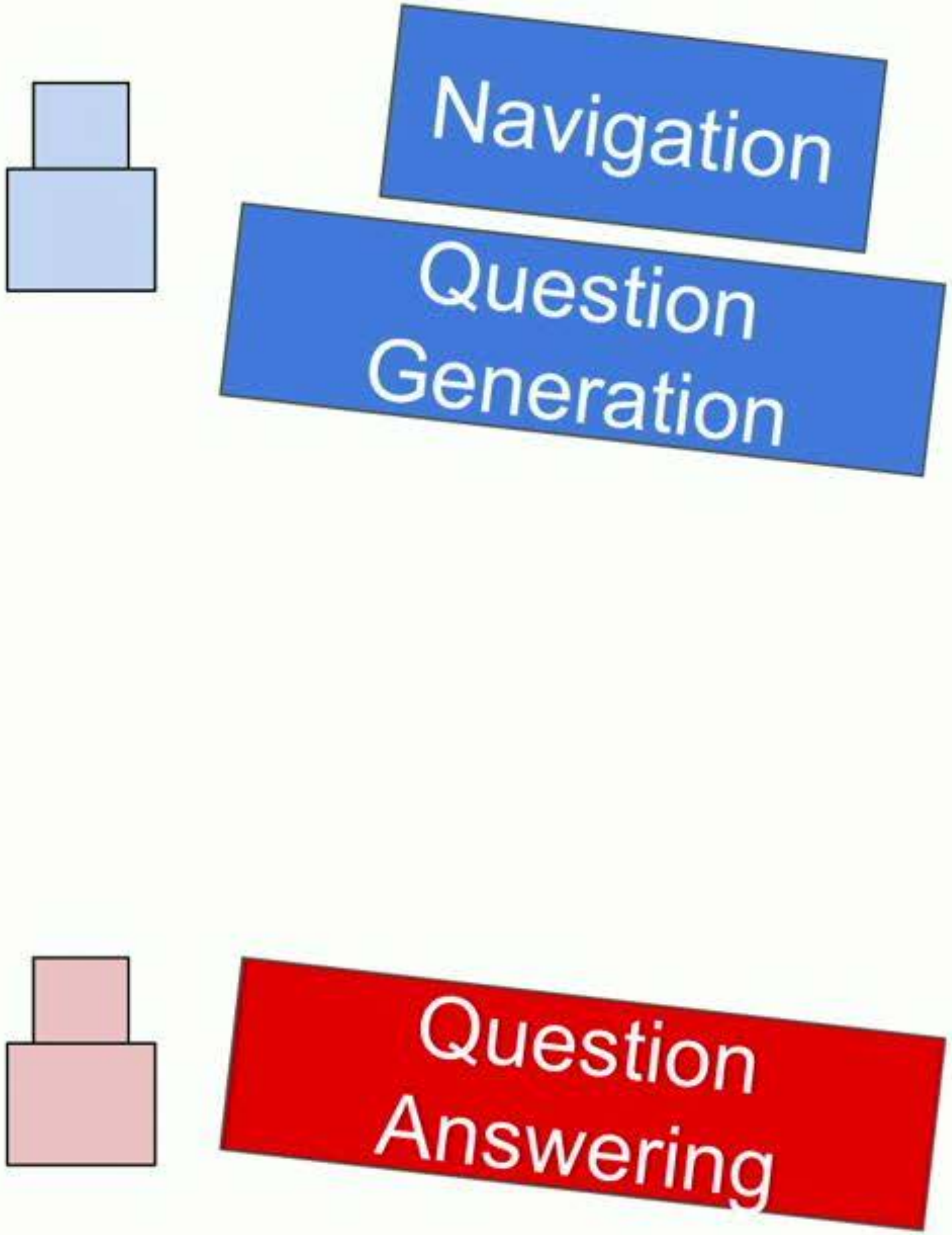
Hint: The goal room contains a *mat*.

- Environment exploration
- Self-play



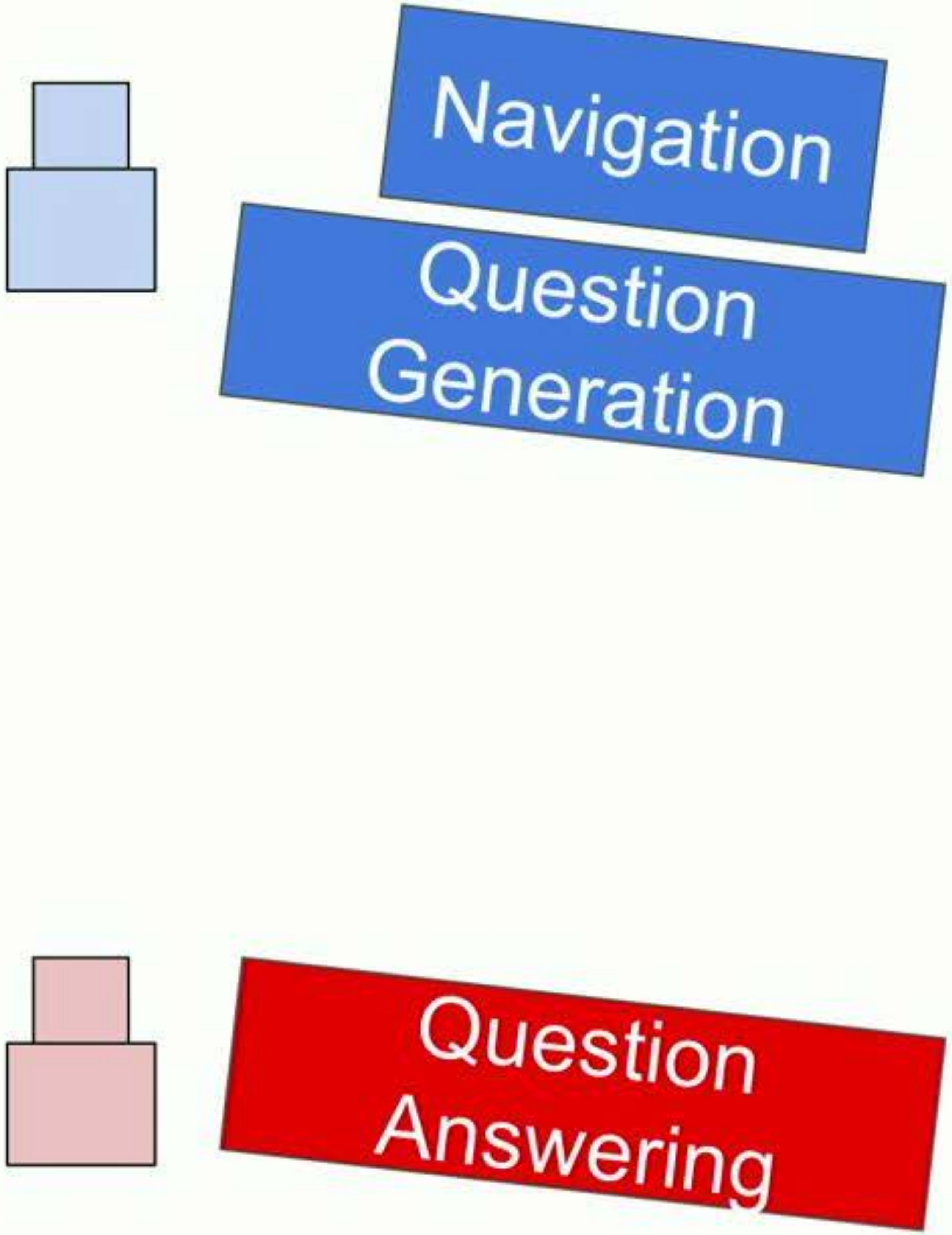
Hint: The goal room contains a *mat*.

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



Hint: The goal room contains a *mat*.

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



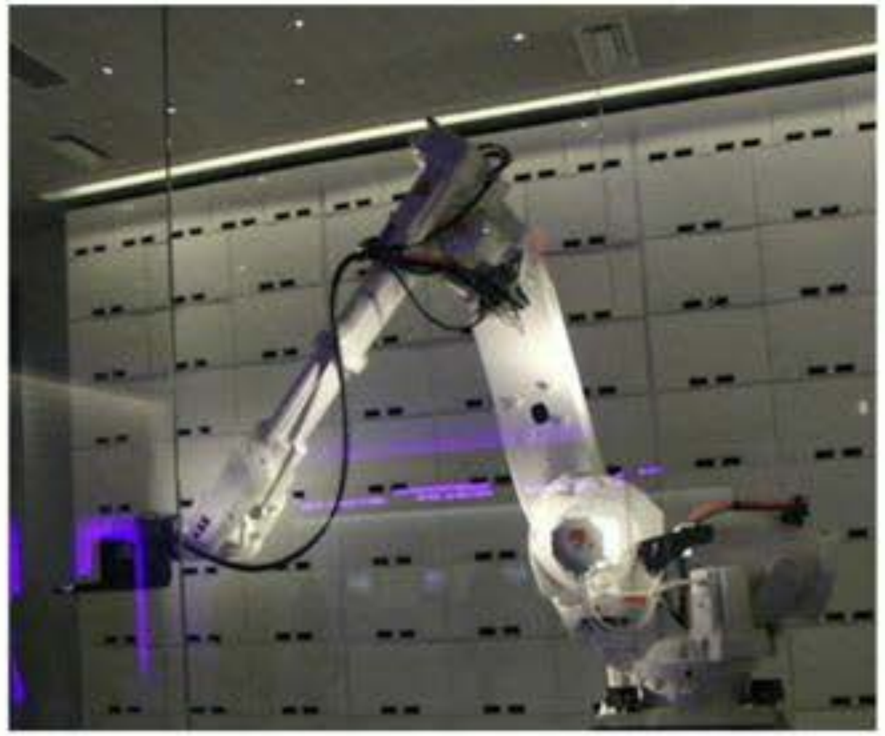
Navigation

Question Generation

Question Answering

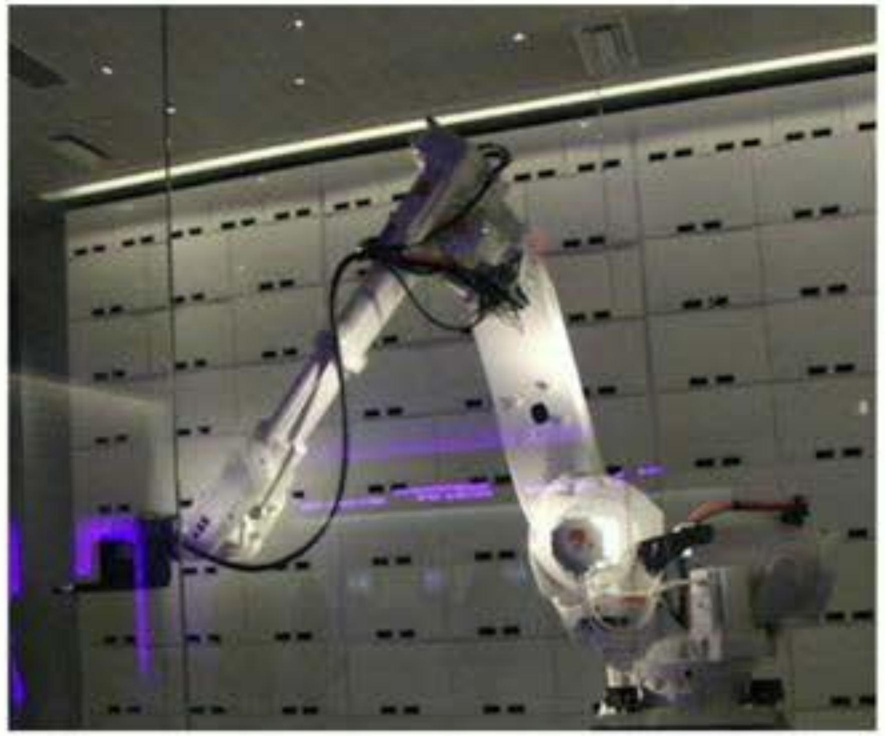
Bringing Robots from Industrial to Human Spaces

Industrial



Bringing Robots from Industrial to Human Spaces

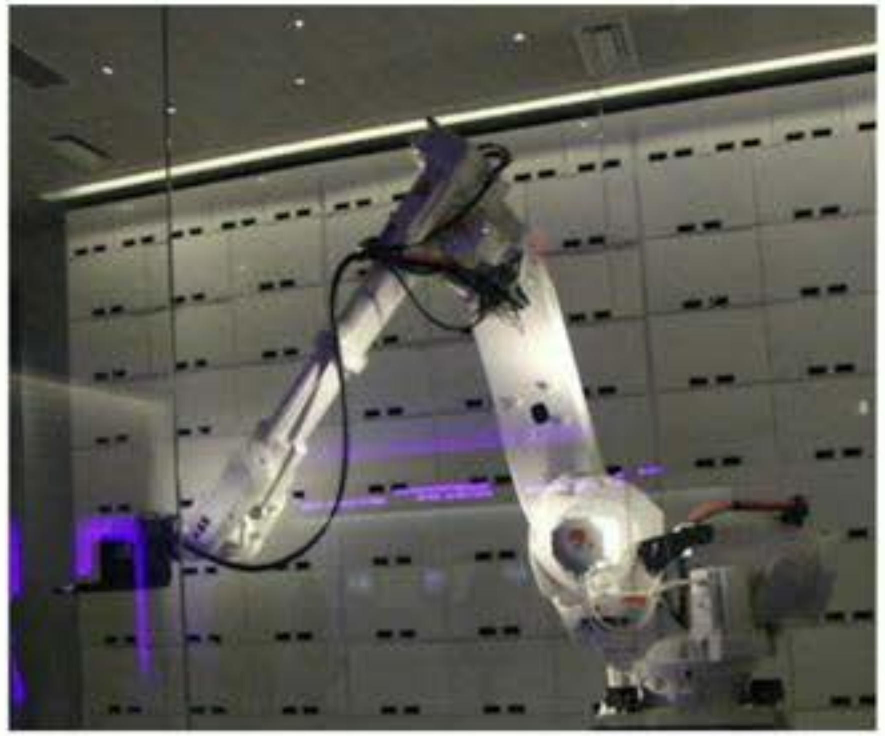
Industrial



Human

Bringing Robots from Industrial to Human Spaces

Industrial



Natural Language

Navigation

Additional Safety

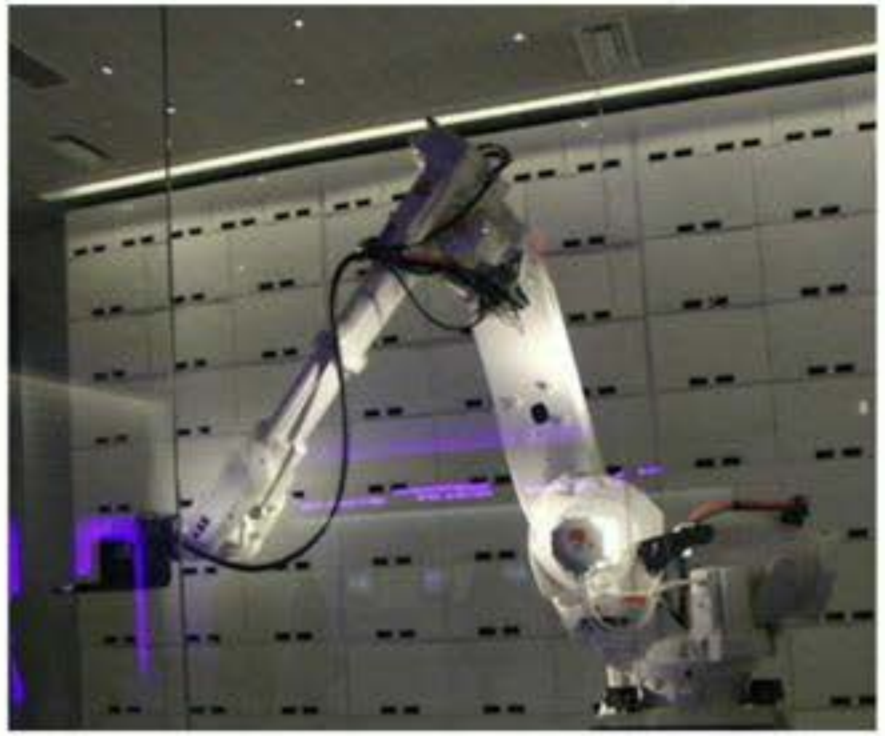
Robust Perception

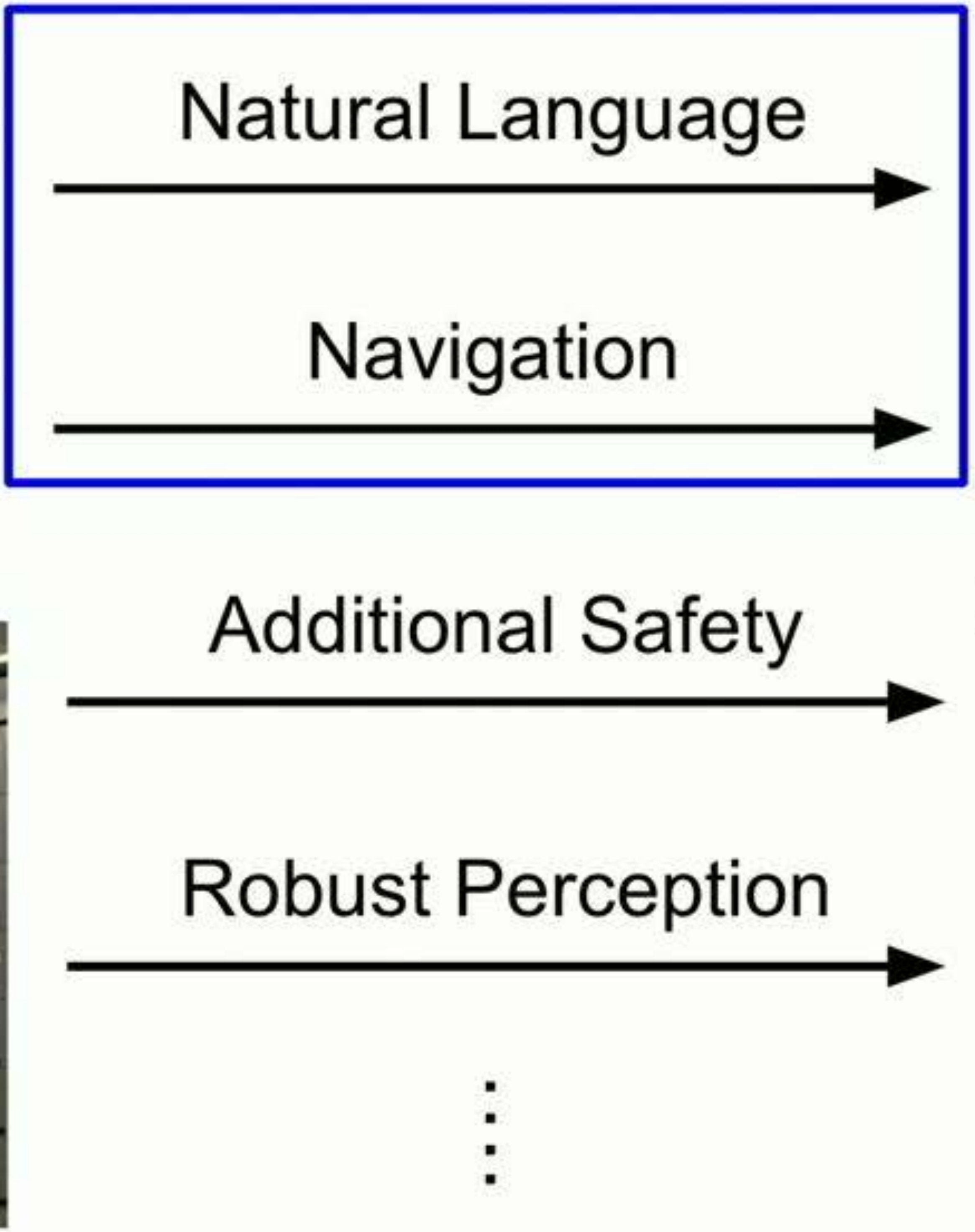
⋮

Human

Bringing Robots from Industrial to Human Spaces

Industrial





Natural Language

Navigation

Additional Safety

Robust Perception

⋮

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.
  "Shifting the Baseline: Single Modality Performance on Visual Navigation and QA", NAACL'19 [https://arxiv.org/abs/1811.00613]

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]
- Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer.
  "Vision-and-Dialog Navigation", *in submission* [https://arxiv.org/abs/1907.04957]

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]
- Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer.
  "Vision-and-Dialog Navigation", *in submission* [https://arxiv.org/abs/1907.04957]
- Cooperative Vision-and-Dialog Navigation dataset

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]
- Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer.
  "Vision-and-Dialog Navigation", *in submission* [https://arxiv.org/abs/1907.04957]
- Cooperative Vision-and-Dialog Navigation dataset
  - Data+code: https://github.com/mmurray/cvdn/

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]
- Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. "Vision-and-Dialog Navigation", *in submission* [https://arxiv.org/abs/1907.04957]
- Cooperative Vision-and-Dialog Navigation dataset
  - Data+code: https://github.com/mmurray/cvdn/
  - Live demo: https://cvdn.dev/

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]
- Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer.
  "Vision-and-Dialog Navigation", *in submission* [https://arxiv.org/abs/1907.04957]
- 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]
- Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer.
  "Vision-and-Dialog Navigation", *in submission* [https://arxiv.org/abs/1907.04957]
- 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

“Put a slice of bread in the microwave.”

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