NAV-NF: A Benchmark and Framework for Vision-Language Navigation under Infeasible (Not-Found) Instructions

針對潛在不可行指令之視覺語言導航之基準與方法設計

Presenter: Ting-Jun Wang 王廷郡

Advisor: Prof. Winston H. Hsu, 徐宏民 教授

2025/12/02

Introduction

Vision-Language Navigation

- Receive a language instruction from human, navigate through the environment using an egocentric visual perspective
  - Considered sucessful if the agent arrives within a specified distance(e.g. 3m)[1]

Enter the kitchen to find the chair

[1] Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments (CVPR 2018)

Human Errors in Object Localization

- A cognitive science study[1] shows that humans cannot locate a target object's 2D position with 100% accuracy after environmental changes (landmark shifts or changes in neighboring objects)
- The real world is more complex than experimental environments.

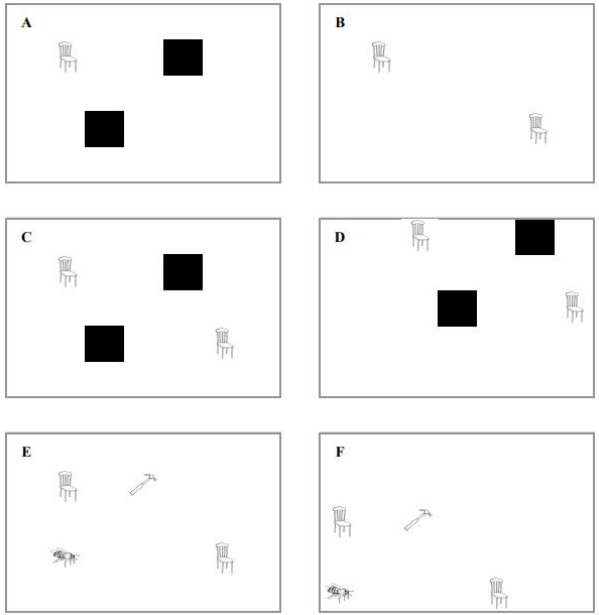


Figure 1. The design of Experiment 1. A, an encoding trial; B, fixed-nocue retrieval; C, fixed-landmark retrieval; D, shifted-landmark retrieval; E, fixed-object retrieval; F, shifted-object retrieval.

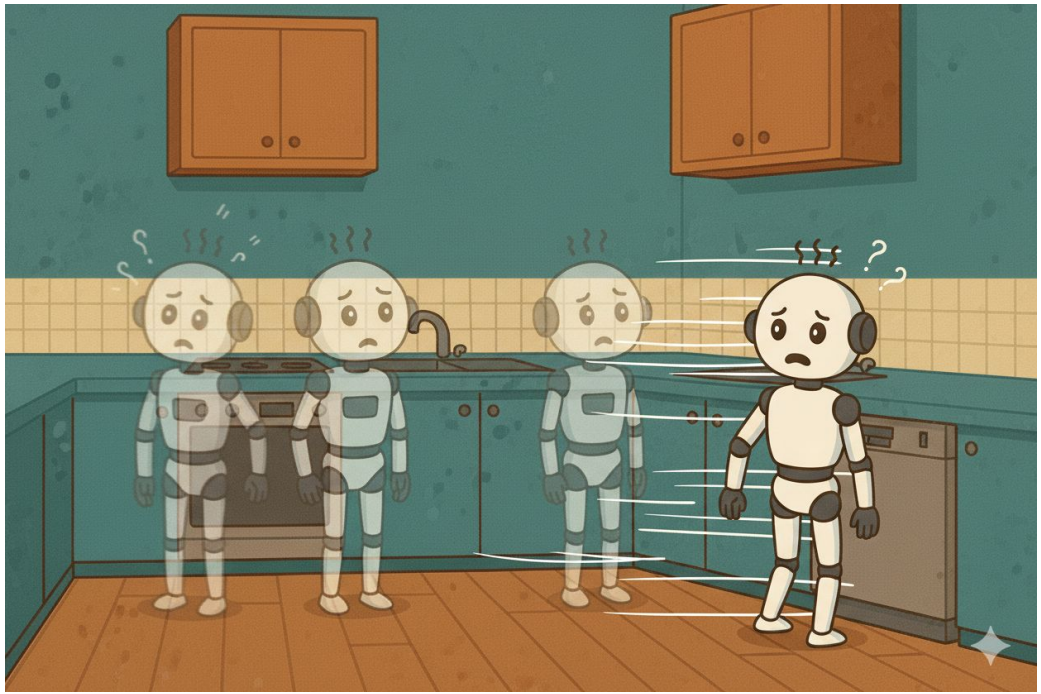
[1] A Study of Object-Location Memory. (CogSci 2002 (volume 24))

New Task: Navigation Not Found (NAV-NF)

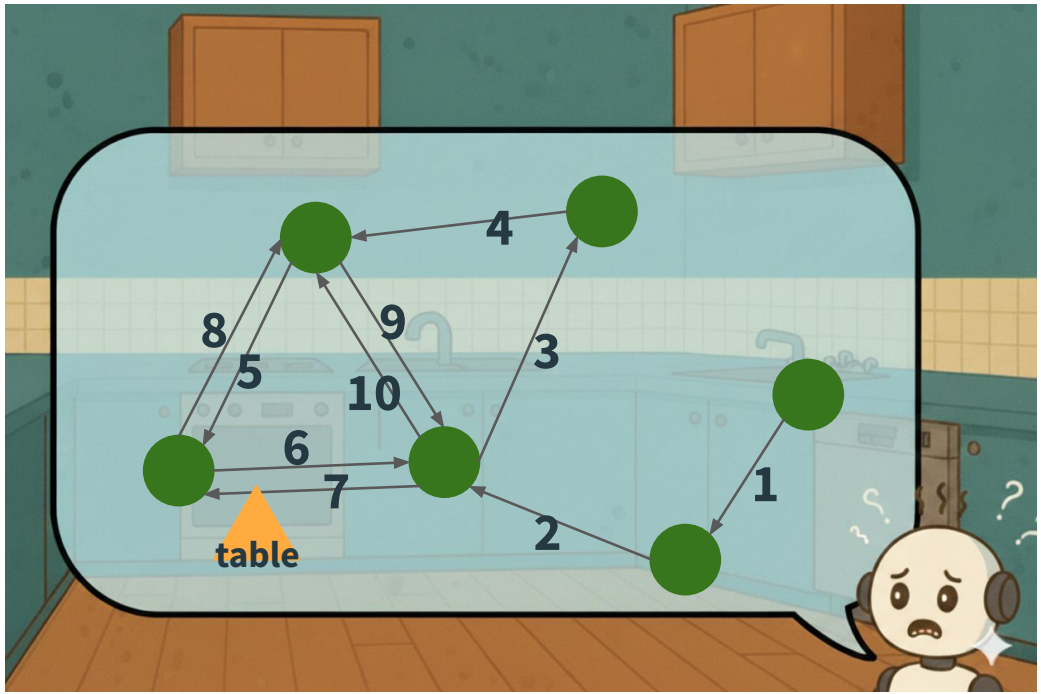
New Task: Navigation Not Found (NAV-NF)

New Task: Navigation Not Found (NAV-NF)

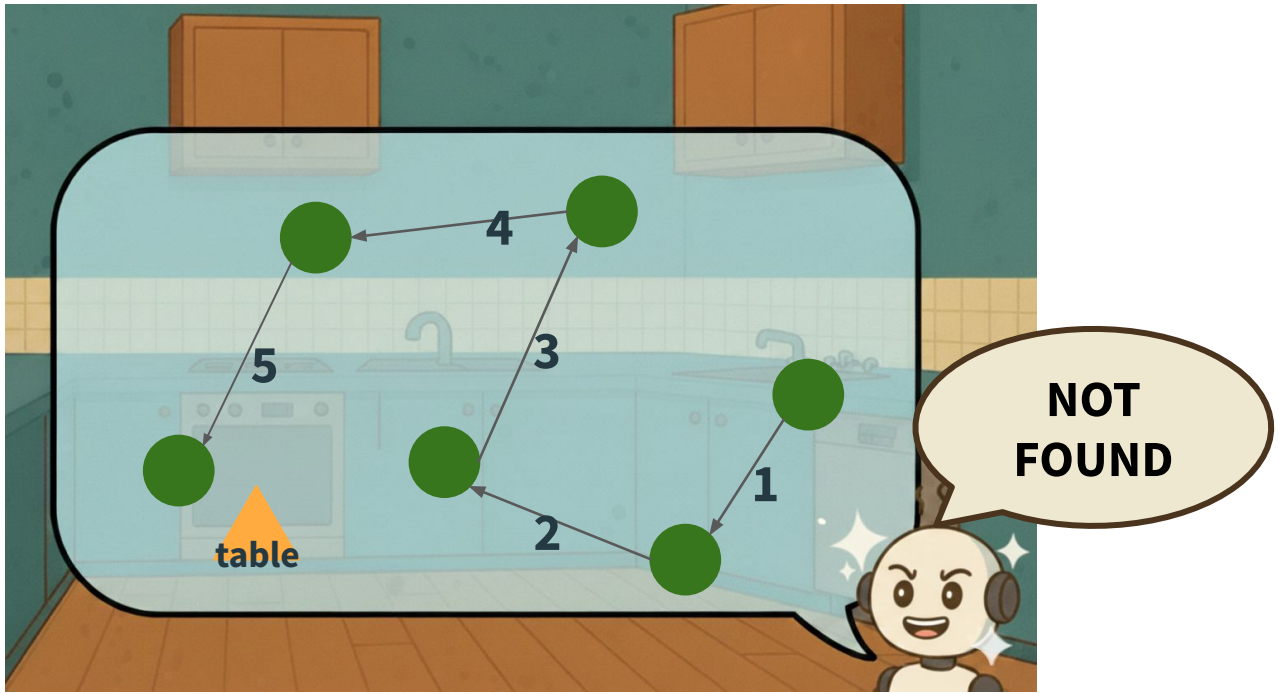
New Task: Navigation Not Found (NAV-NF)



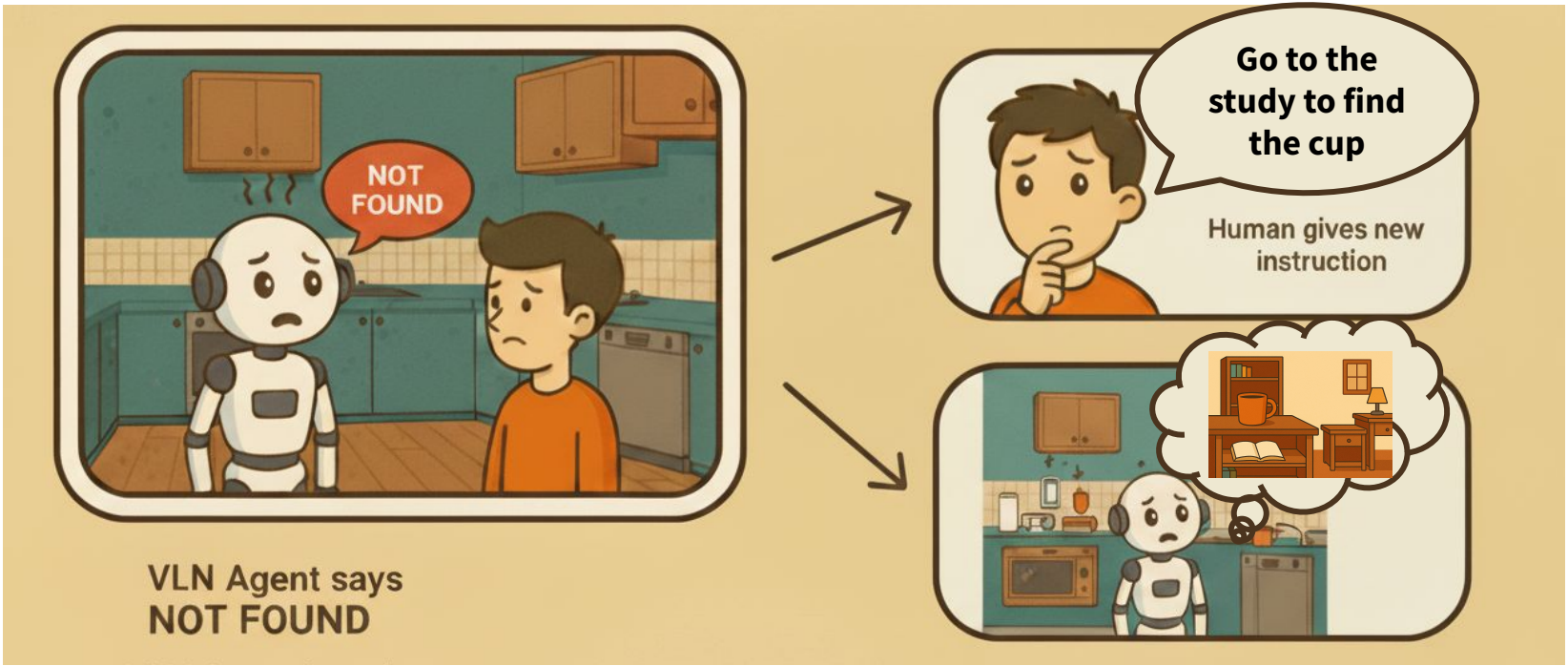
New Task: Navigation Not Found (NAV-NF)



New Task: Navigation Not Found (NAV-NF)



New Task: Navigation Not Found (NAV-NF)



Vision-Language Navigation (Datasets)

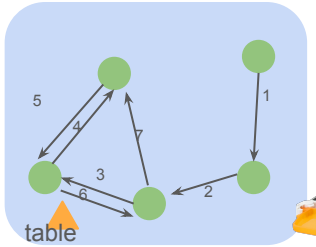
- **Room-to-Room (R2R)[1]**
  - Path following
  - Example: *Leave the bedroom. Cross the hall to the sitting room. Wait near the bed.*
- **REVERIE [2]**
  - Navigation + object grounding
  - Example: *Go to the study room and pick up the cup on the table*
  - better reflects real-world scenarios
- **More…**
  - SOON[3], CVDN[4], RxR[5]
  - More complex instructions, incorporation of novel objects, and integration of dialogue…
- While the instructions might be incorrect in the real world due to human mistakes…

[1] Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments (CVPR 2018)
[2] REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments (CVPR 2020)
[3] SOON: Scenario Oriented Object Navigation with Graph-based Exploration (CVPR 2021)
[4] Cooperative Vision-and-Dialog Navigation (CoRL 2019)
[5] Room-Across-Room}: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding (EMNLP 2020)

New Task: Navigation Not Found (NAV-NF)

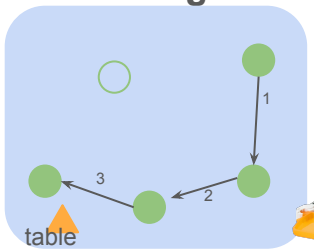
Go to the study room and
pick up the cup on the table
(Actually the cup has been moved to the kitchen)

Current VLN Agent:



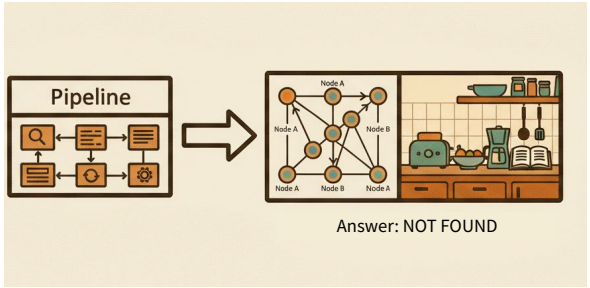
I have reached the study!
I found the table
Where is the cup? …

Ideal VLN Agent:

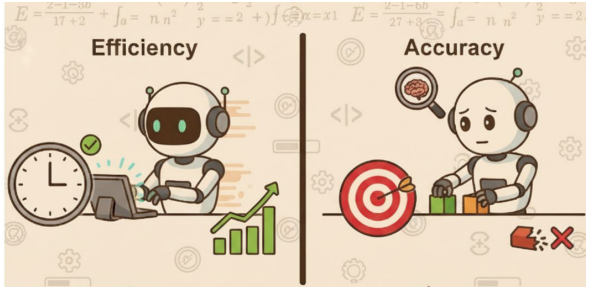


I have reached the study!
I found the table
but there is no cup
(Not Found)

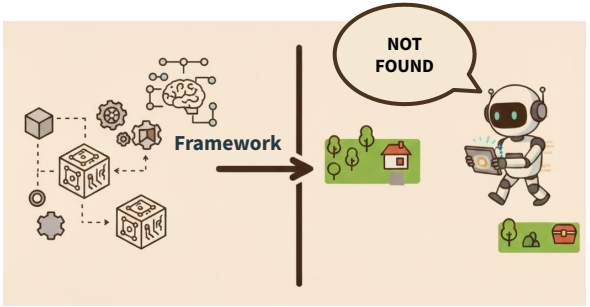
New Task: Navigation Not Found (NAV-NF)



Dataset



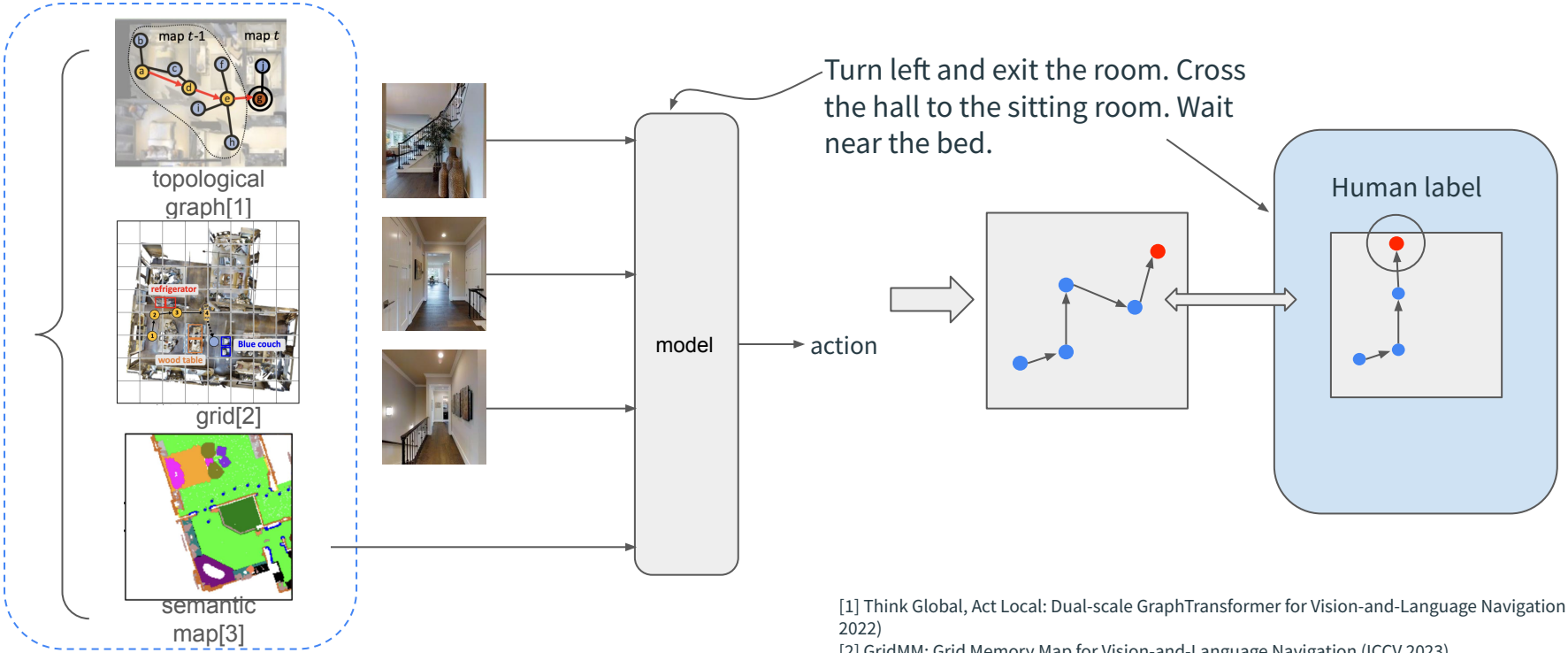
Metrics



ROAM Framework

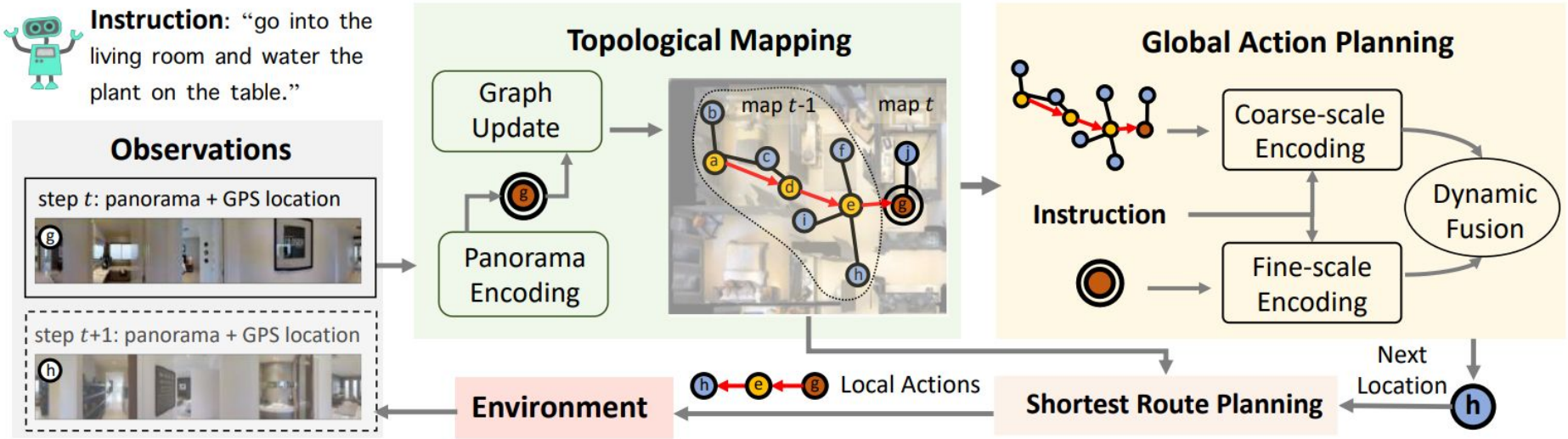
Related Work

Supervised Methods



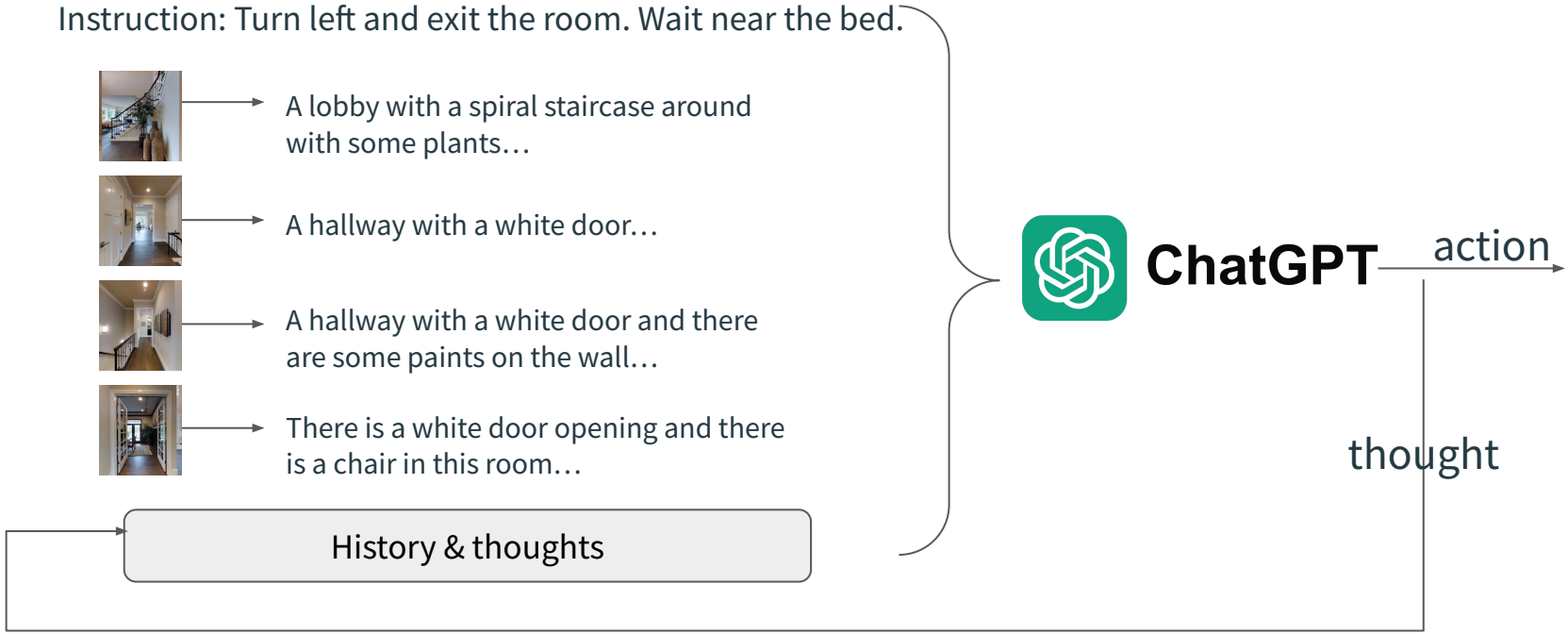
[1] Think Global, Act Local: Dual-scale GraphTransformer for Vision-and-Language Navigation (CVPR 2022)
[2] GridMM: Grid Memory Map for Vision-and-Language Navigation (ICCV 2023)
[3] Cross-modal Map Learning for Vision and Language Navigation (CVPR 2022)

Supervised Method: DUET



[1] Think Global, Act Local: Dual-scale GraphTransformer for Vision-and-Language Navigation (CVPR 2022)

Unsupervised Methods

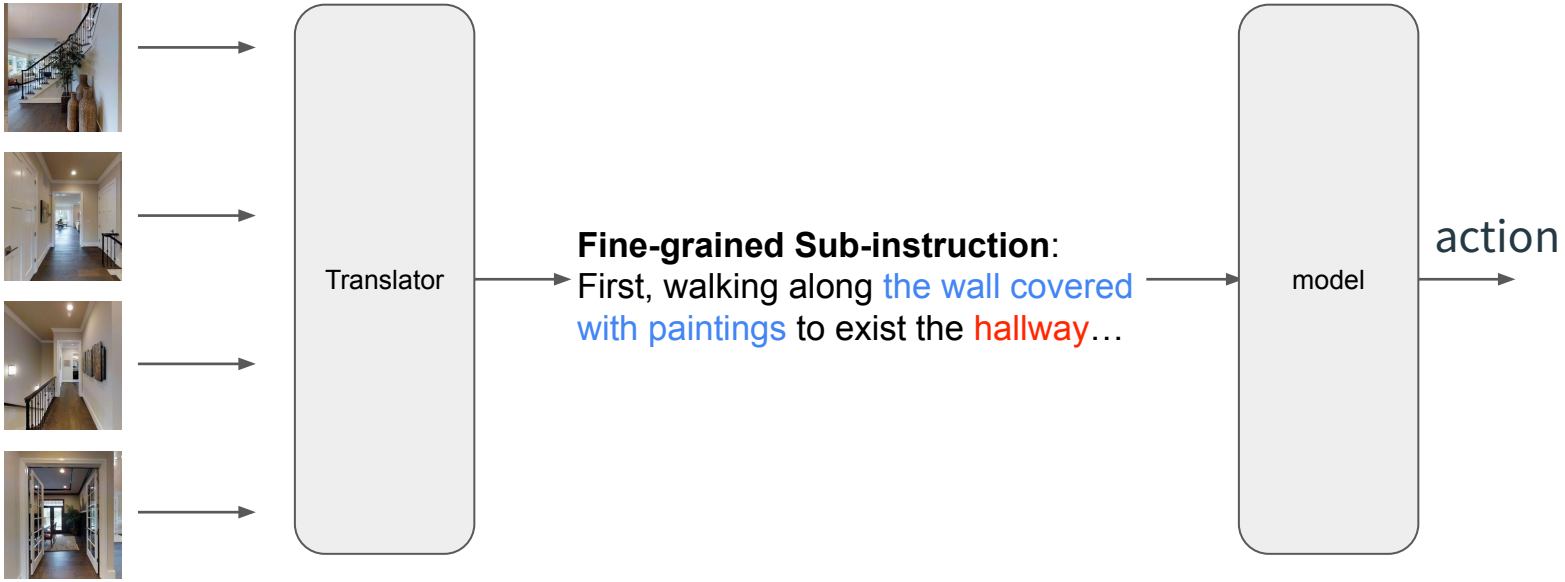


[1] NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models (AAAI 2024)

Comparison

	Supervised	Unsupervised
Visual Utilization	✓	
Performance	✓	
Annotation Cost		✓
Generalization		✓
Interpretability		✓

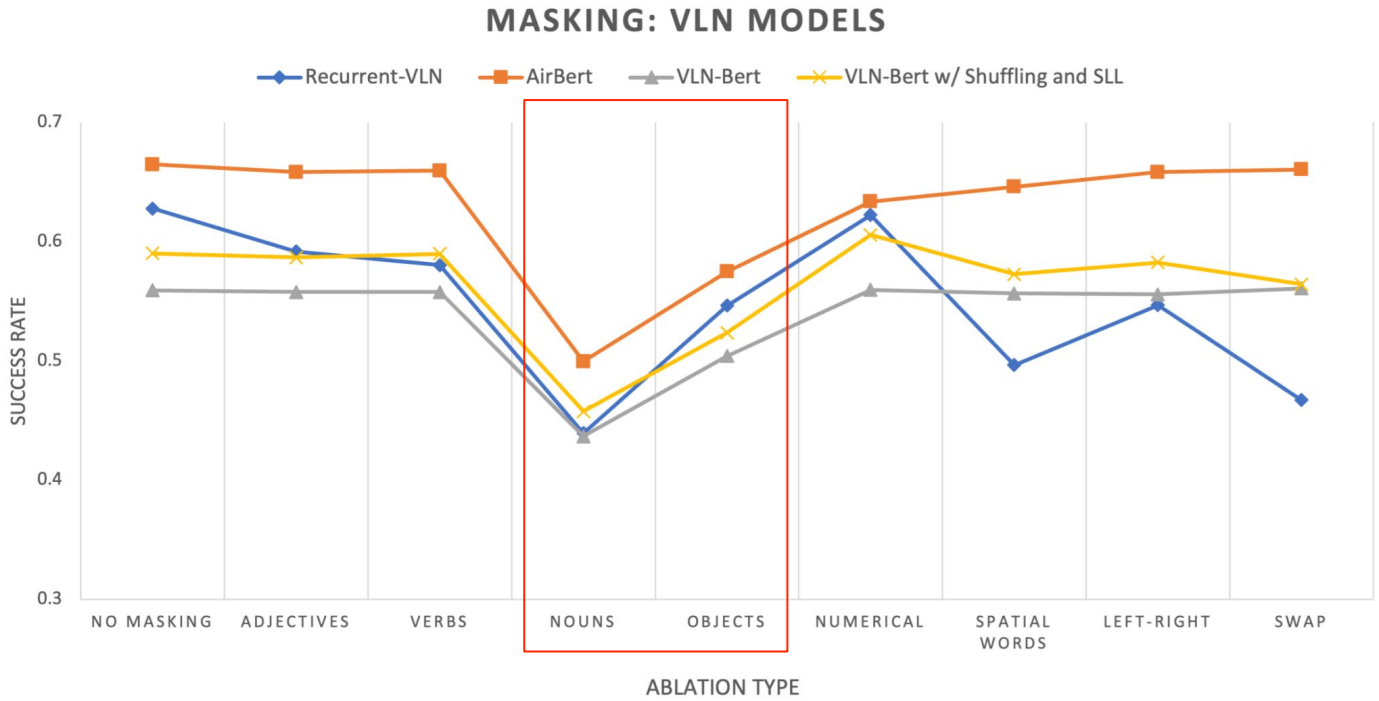
Different Strategy: Rewrite Instructions



Instruction:
Exist the hallway and enter the bedroom. Wait near the bed.

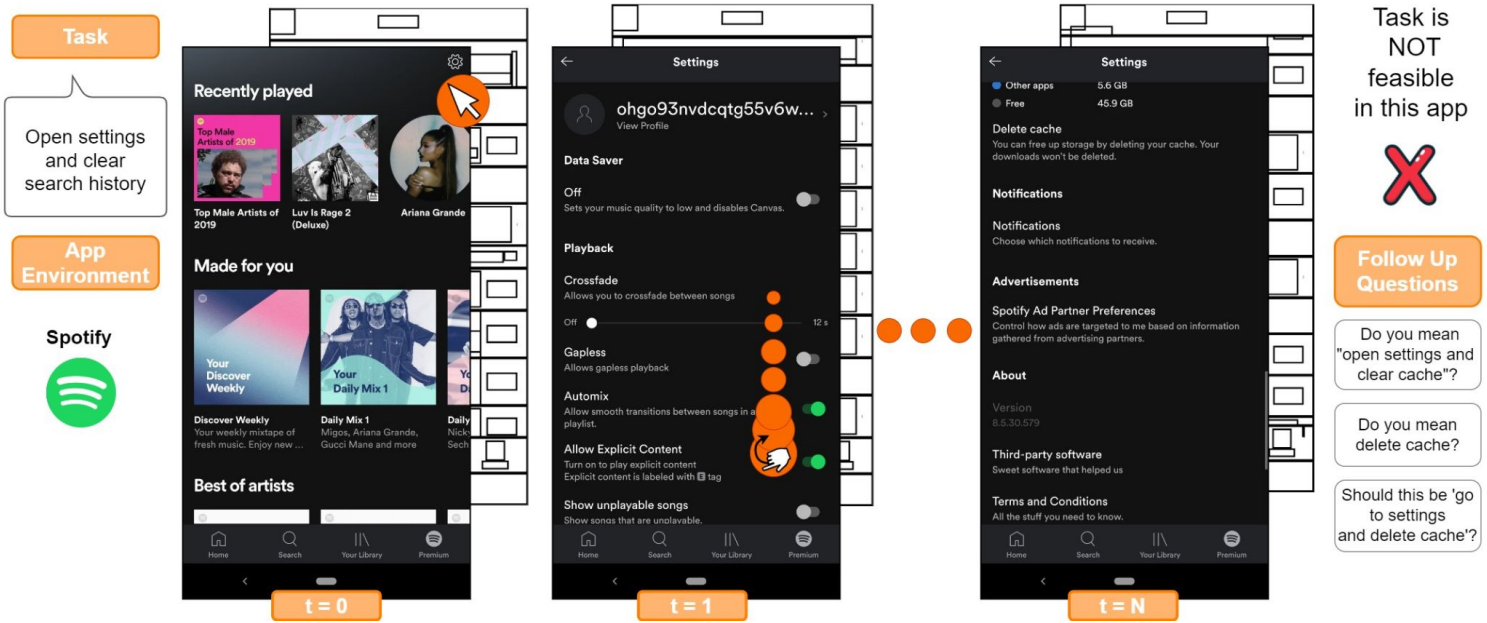
[1] VLN-Trans: Translator for the Vision and Language Navigation Agent (ACL 2023)

Instruction Errors in VLN



[1] Which way is 'right'?: Uncovering limitations of Vision-and-Language Navigation Models (AAMAS 2023)

Infeasible Instructions in VLN



[1] A Dataset for Interactive Vision-Language Navigation with Unknown Command Feasibility (ECCV 2022)

NAV-NF as a Hybrid Search-Decision Problem

Input

Search Problem

Definition
"Find solutions" to reach the goal state

Example
Find a path from s to t

In our task
(Method) Find a path to reach the target object

Decision Problem

Definition
Determine whether a "solution exists" (Yes/No)

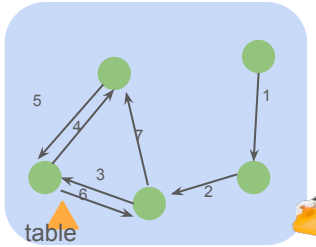
Example
Is there a path from s to t?

In our task
(Result) whether the target object exist?

New Task: Navigation Not Found (NAV-NF)

Go to the study room and pick up the cup on the table
(Actually the cup has been moved to the kitchen)

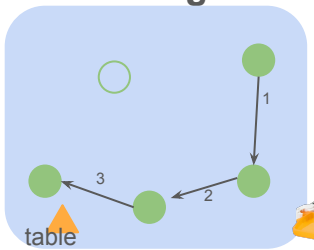
Current VLN Agent:



I have reached the study!
I found the table
Where is the cup? …

Search Problem

Ideal VLN Agent:



I have reached the study!
I found the table
but there is no cup
(Not Found)

Decision Problem

Summary of Related Work

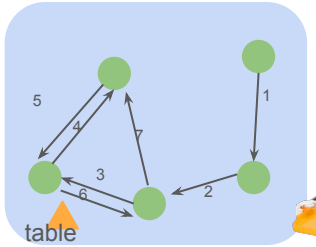
- **Dataset:** Most existing datasets assume instructions are correct
  - Ours → not always correct in real-world
- **Methods**: Supervised & unsupervised training on existing datasets cannot handle NAV-NF
  - Ours → Multi-stage with both supervised and unsupervised advantages
- **Infeasible Instruction Research**: Noun is the key part of instruction
  - Ours → focuses on cases with incorrect nouns in a real-world environment

Dataset Generation Pipeline

New Task: Navigation Not Found (NAV-NF)

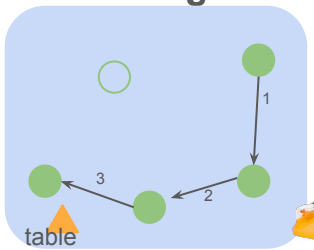
Go to the study room and
pick up the cup on the table
(Actually the cup has been moved to the kitchen)

Current VLN Agent:



I have reached the study!
I found the table
Where is the cup? …

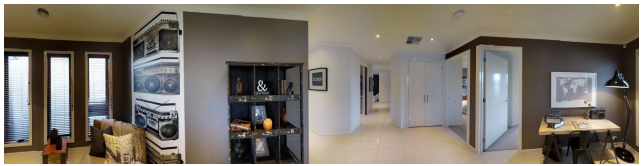
Ideal VLN Agent:



I have reached the study!
I found the table
but there is no cup
(Not Found)

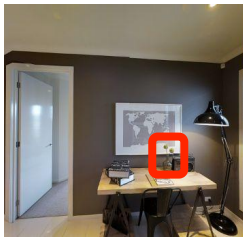
Strategy Options

Go to the study room and
pick up the cup on the table



Feasible Instruction Pair (Original Pair)

1



remove cup

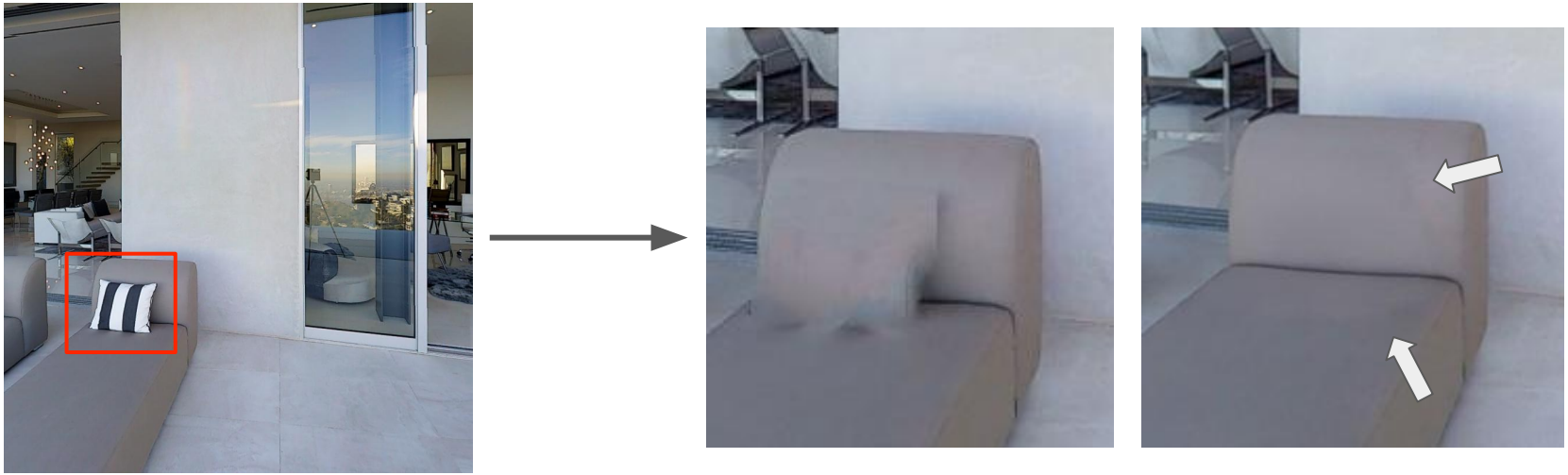
2

Go to the study room and
pick up the cup on the table

Go to the study room and
turn off the lamp on the table.

Failure Cases: Image Inpainting

1



Hard to automatically detect failures

Our Strategy: change the target object

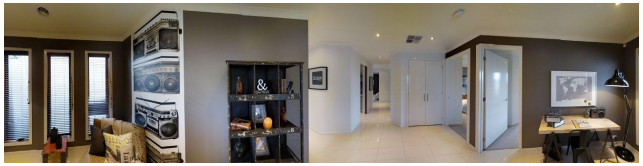
2

Go to the study room and pick up the cup on the table

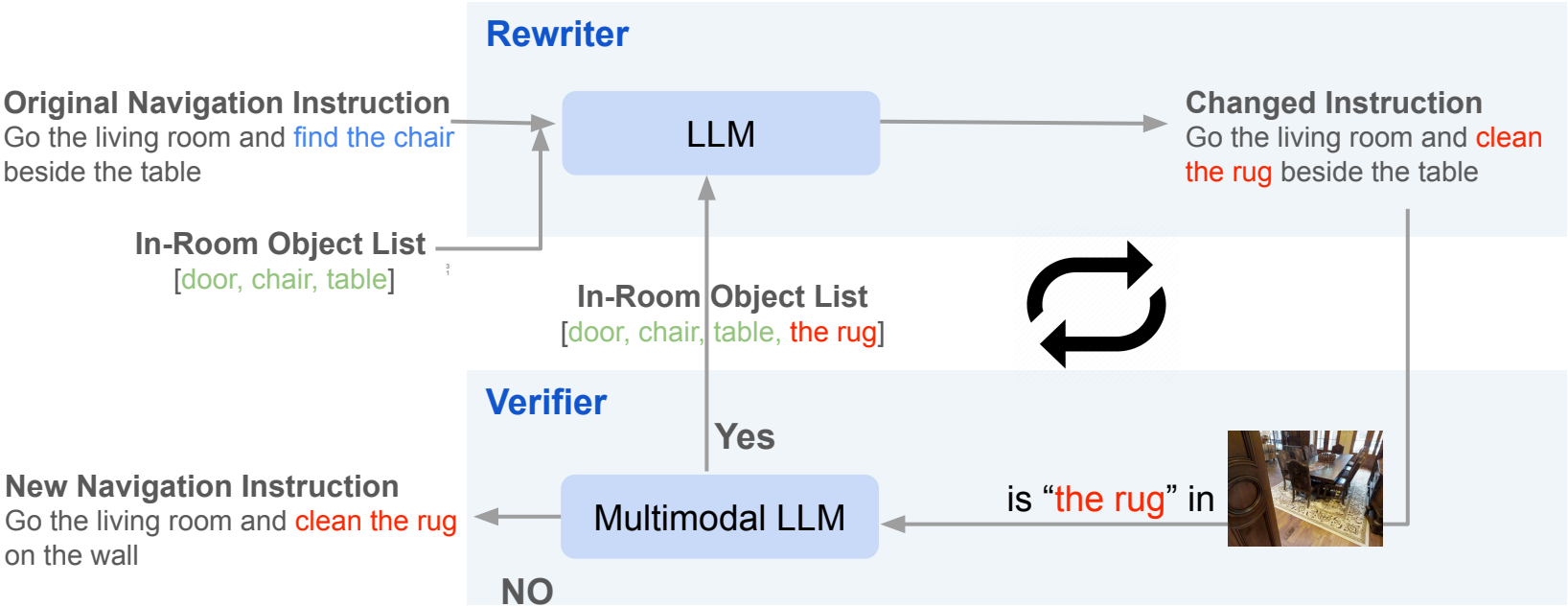
Go to the study room and turn off the lamp on the table.

&

Verify that the lamp is not in



Infeasible Instruction Generation

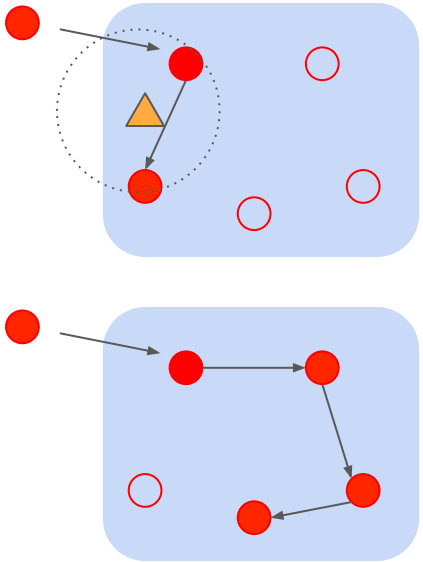


Path Generation for Infeasible Instructions

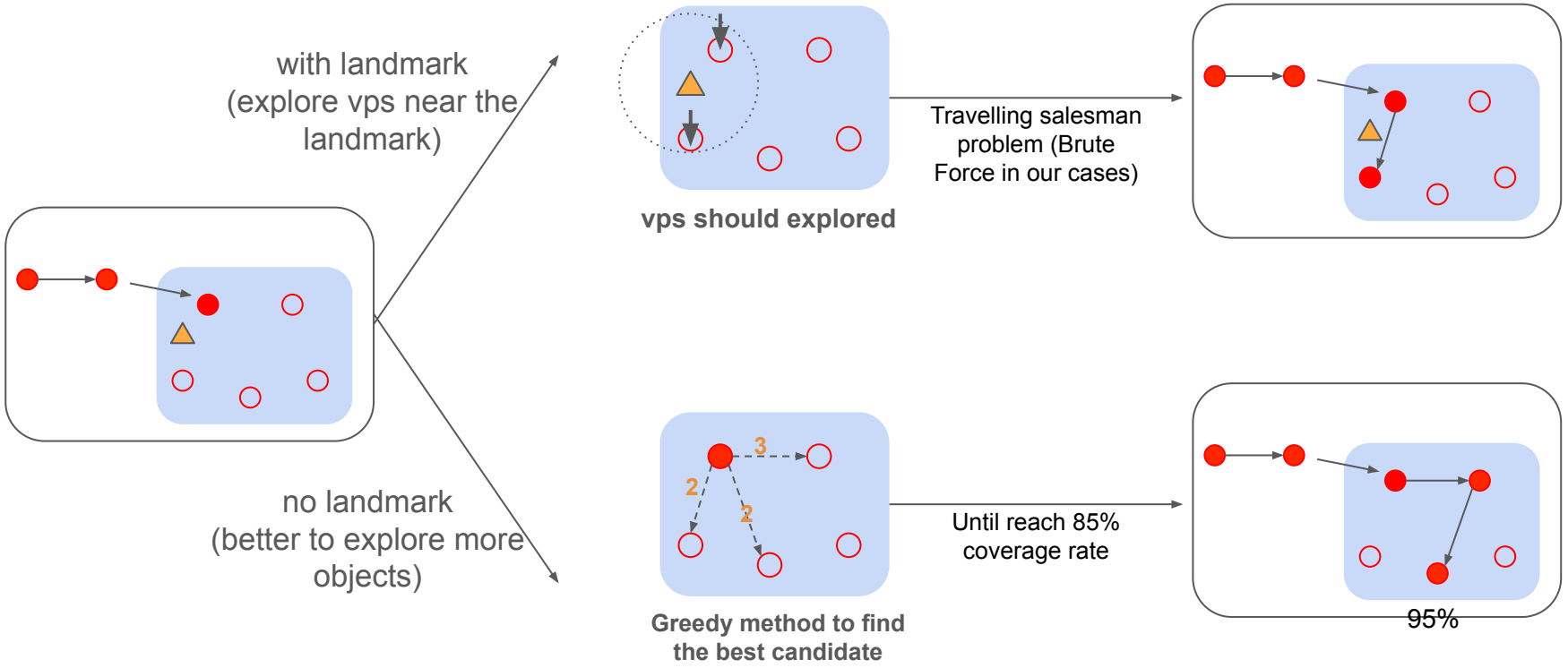
Go to the lounge room and pick up the picture above the lamp

(landmark)

Go to the lounge area and clean the top picture



Path Generation for Infeasible Instructions



Dataset Overview

- **Dataset Size**: 4,724 training instructions, 55 scenes (~45% of original REVERIE)
  - Removed instructions incompatible with NAV-NF
- **Feasibility Split**: 50% feasible (from REVERIE), 50% infeasible (generated by our pipeline)
- **Objects Covered**: 1,072 categories
- **Quality Check**: Human evaluation → <2% errors

Metrics

Evaluation Metrics

- **Success Rate (SR)**
  - **Reach SR**: Agent reaches the correct room
  - **Reach&Found SR**: Agent reaches room & correctly reports Found/Not Found
- **Success weighted by Path Length (SPL)**
  - **Reach SPL**: Efficiency in reaching the correct room (vs. ground-truth path)
  - **Reach&Found SPL**: Efficiency & thoroughness in exploring the target room
- **Coverage Rate**: % of objects in target room explored by agent

Evaluation Metric - Reach SPL

- To align with our task definition, we use "oracle room success" & “the shortest path to the target room”.

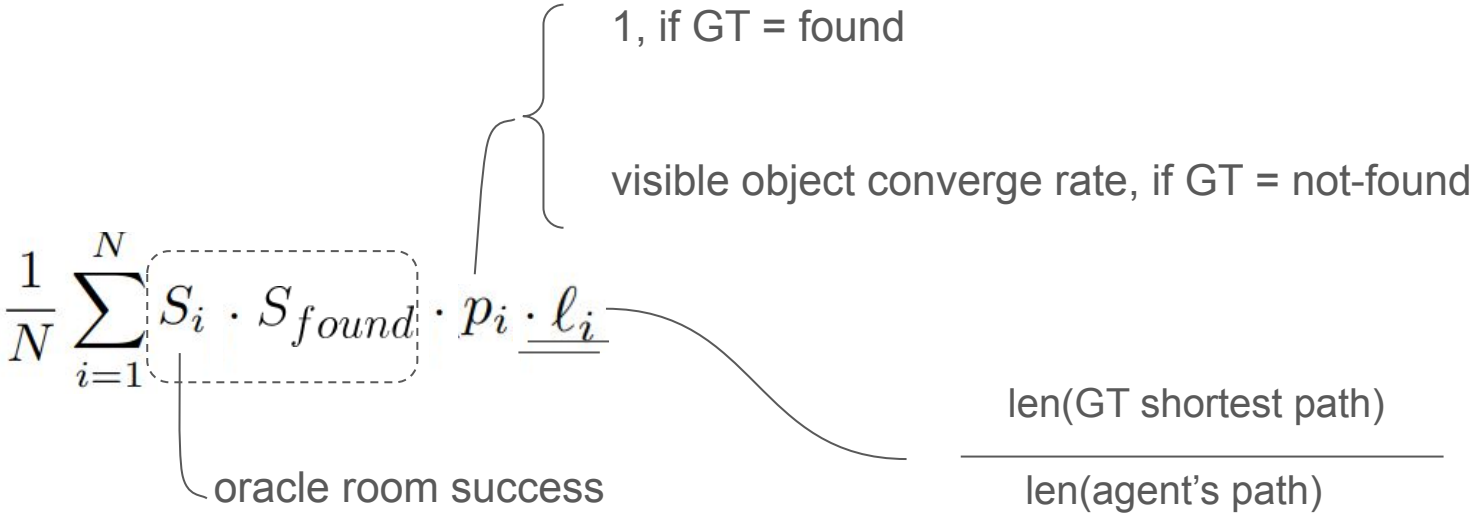
$$\frac{1}{N}\sum_{i=1}^{N} S_i \frac{\ell_i}{\max(p_i, \ell_i)}.$$

GT Path Length

Agent’s Path Length

Evaluation Metric - Reach & Found SPL

- To align with our task definition, we use "oracle room success".



Issue: Unbalanced Prediction & Early Stopping

- We found that some models prefer to predict NOT FOUND to cut corners to get a higher Reach & Found SPL.
  - Unbalanced prediction
  - Early Stopping

Always predicting NOT_FOUND
→ Gains score from GT = NOT_FOUND cases
→ longer GT path (make agent paths relatively shorter)
→ higher SPL

$$\frac{1}{N}\sum_{i=1}^{N} S_i \cdot S_{found} \cdot p_i \cdot \frac{\ell_i}{\max(p_i, \ell_i)}.$$

GT Path Length

Agent's Path Length

Fixed Reach & Found SPL metric

- Issue 1: Unbalanced Prediction
  - F1 Score
- Issue 2: Early Stopping
  - Penalty

$$\mathrm{F1}_{\text{found}} \cdot \frac{1}{N} \sum_{i=1}^{N} S_i \cdot p_i \cdot \frac{\ell_i}{\max(p_i, \ell_i)} \cdot \min(1, \frac{p_i}{\ell_i})$$

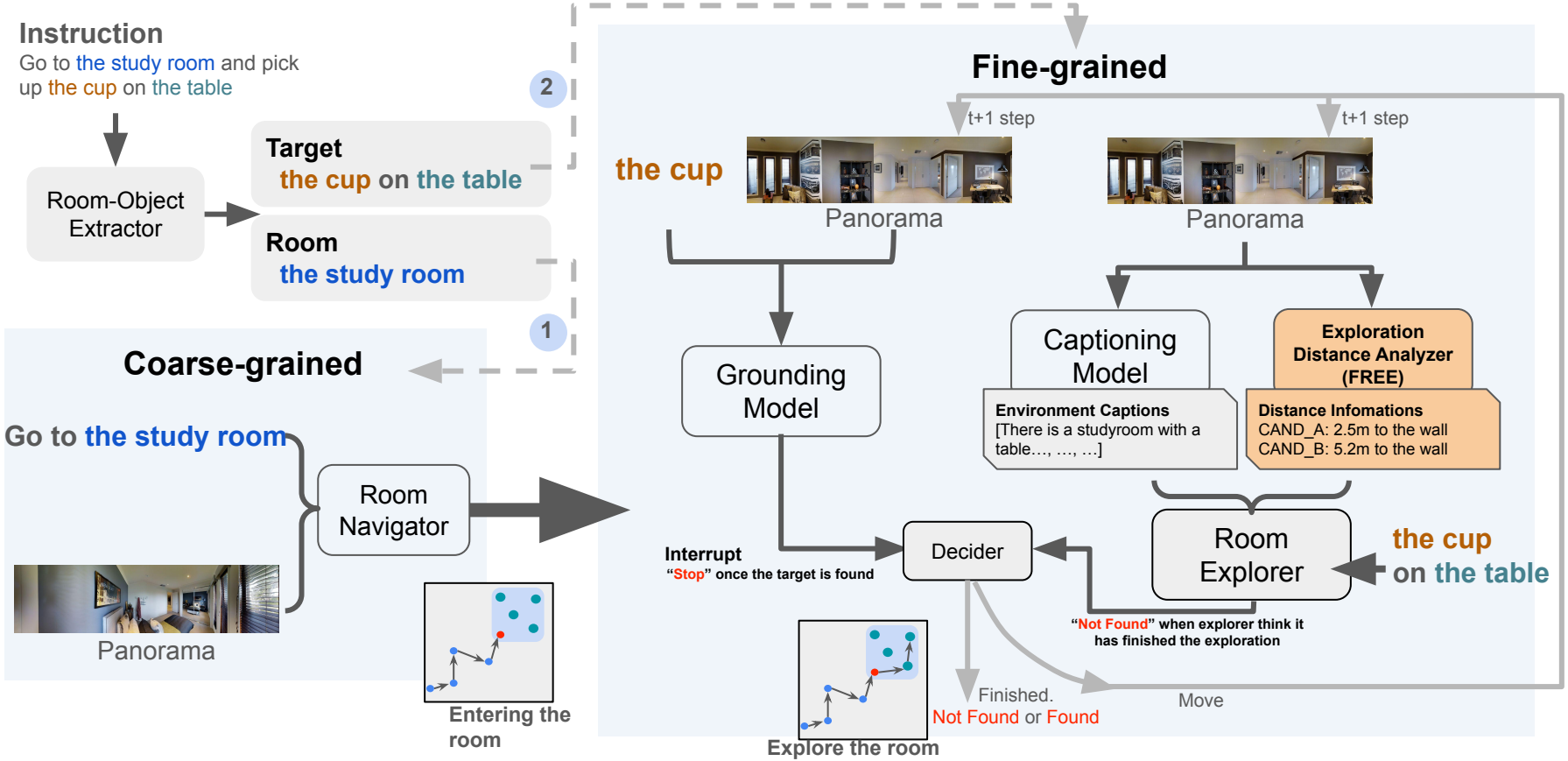
GT Path Length

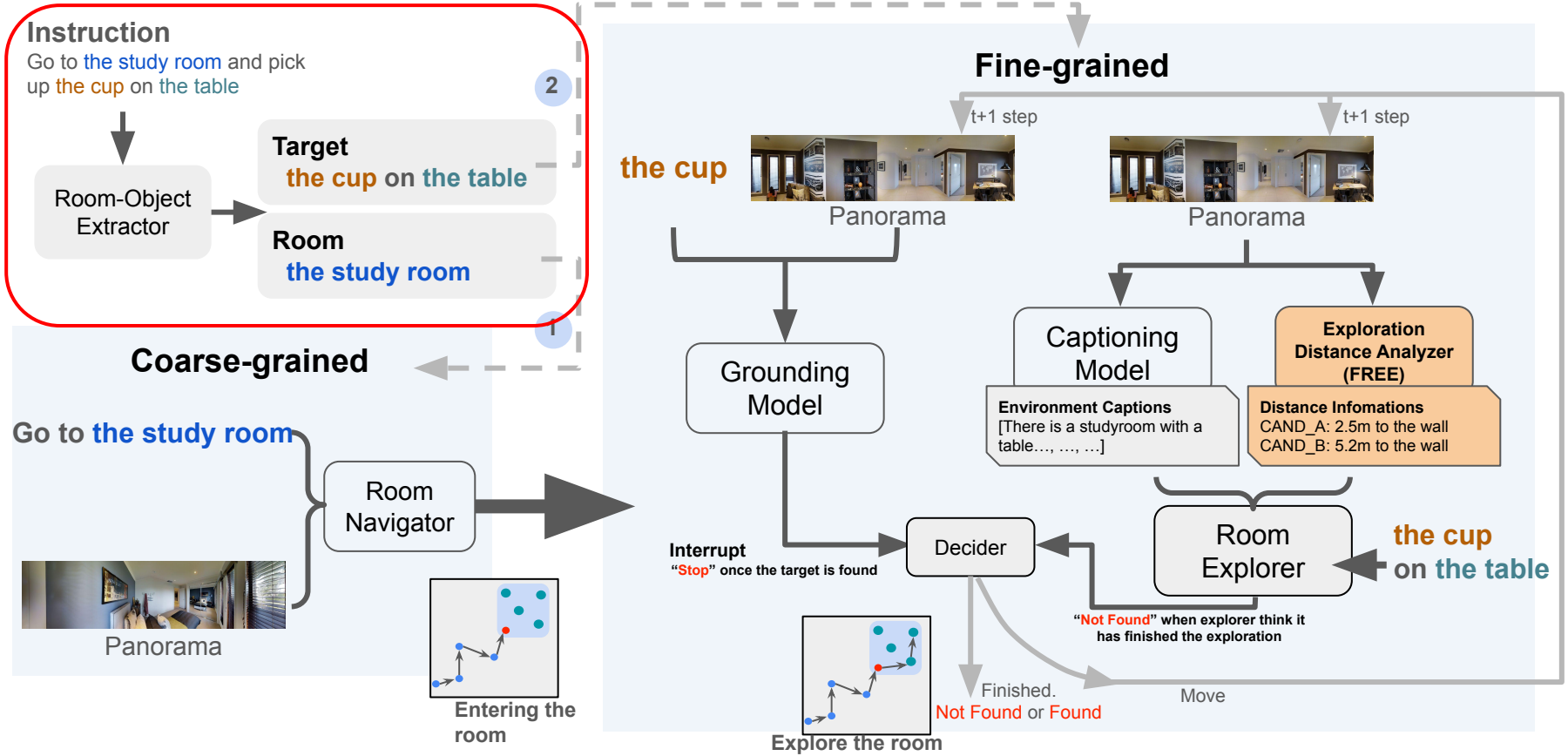
Agent's Path Length

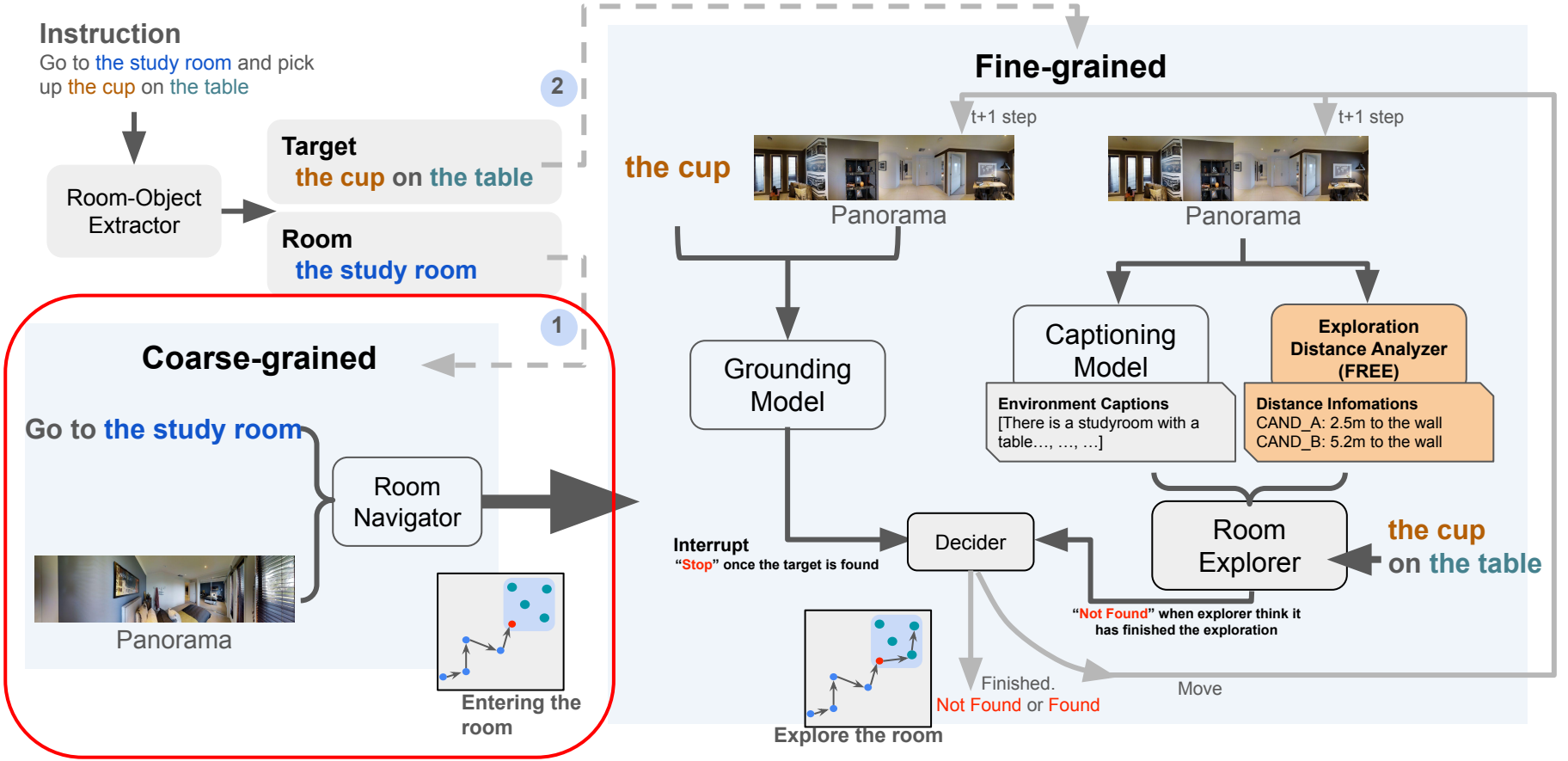
Method

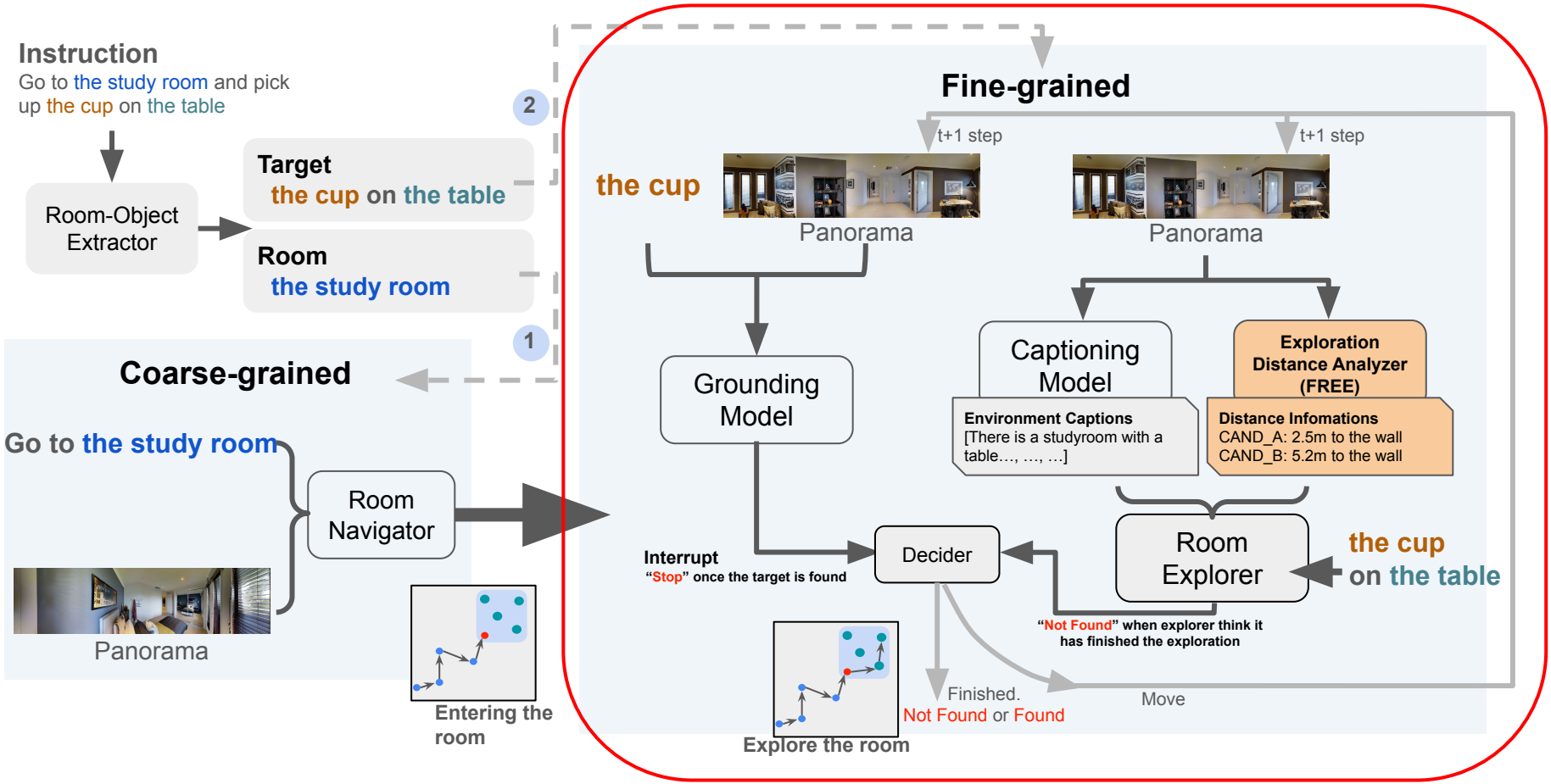
Comparison

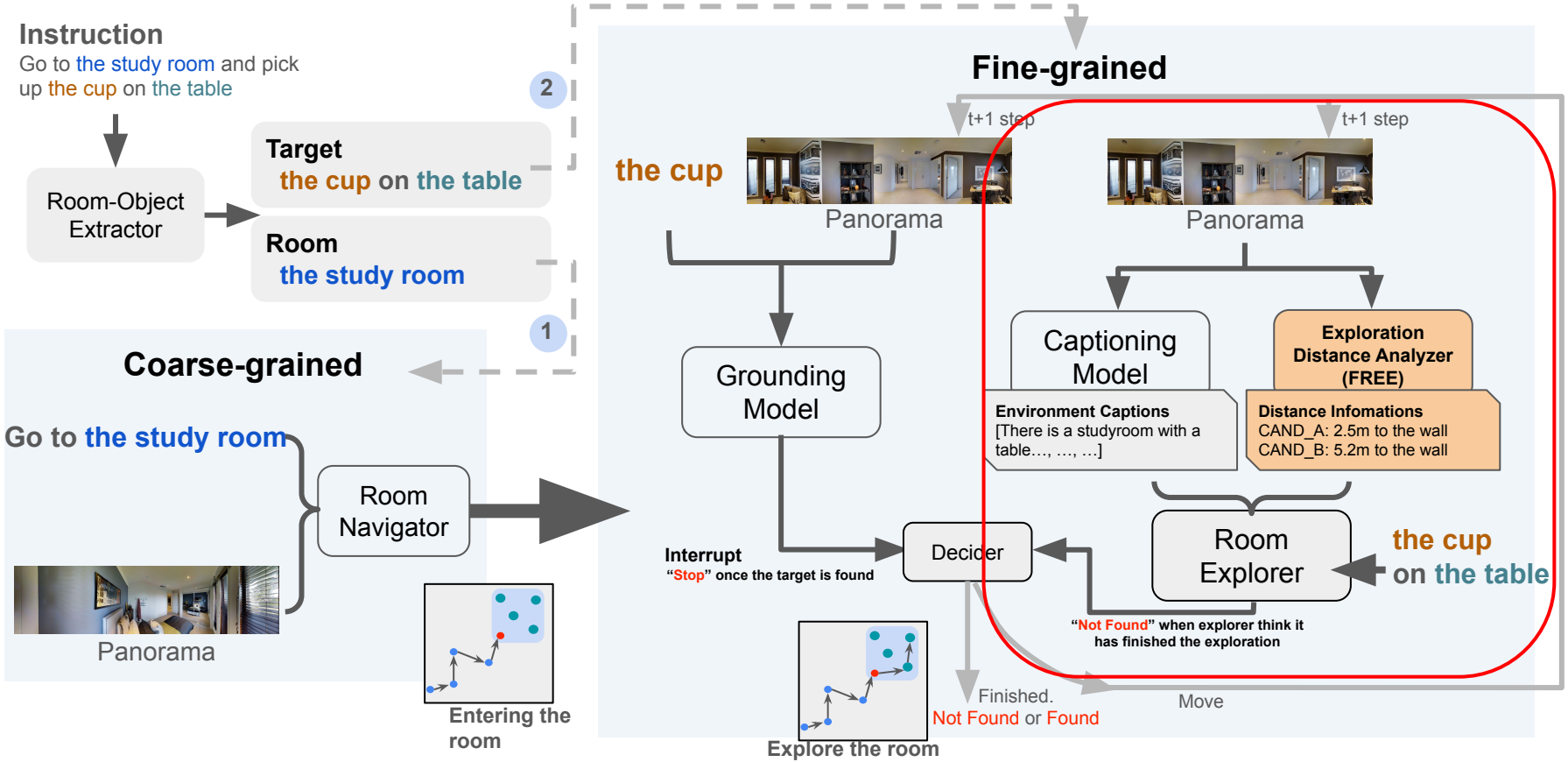
	Supervised	Unsupervised
Visual Utilization	✓	
Performance	✓	
Annotation Cost		✓
Generalization		✓
Interpretability		✓

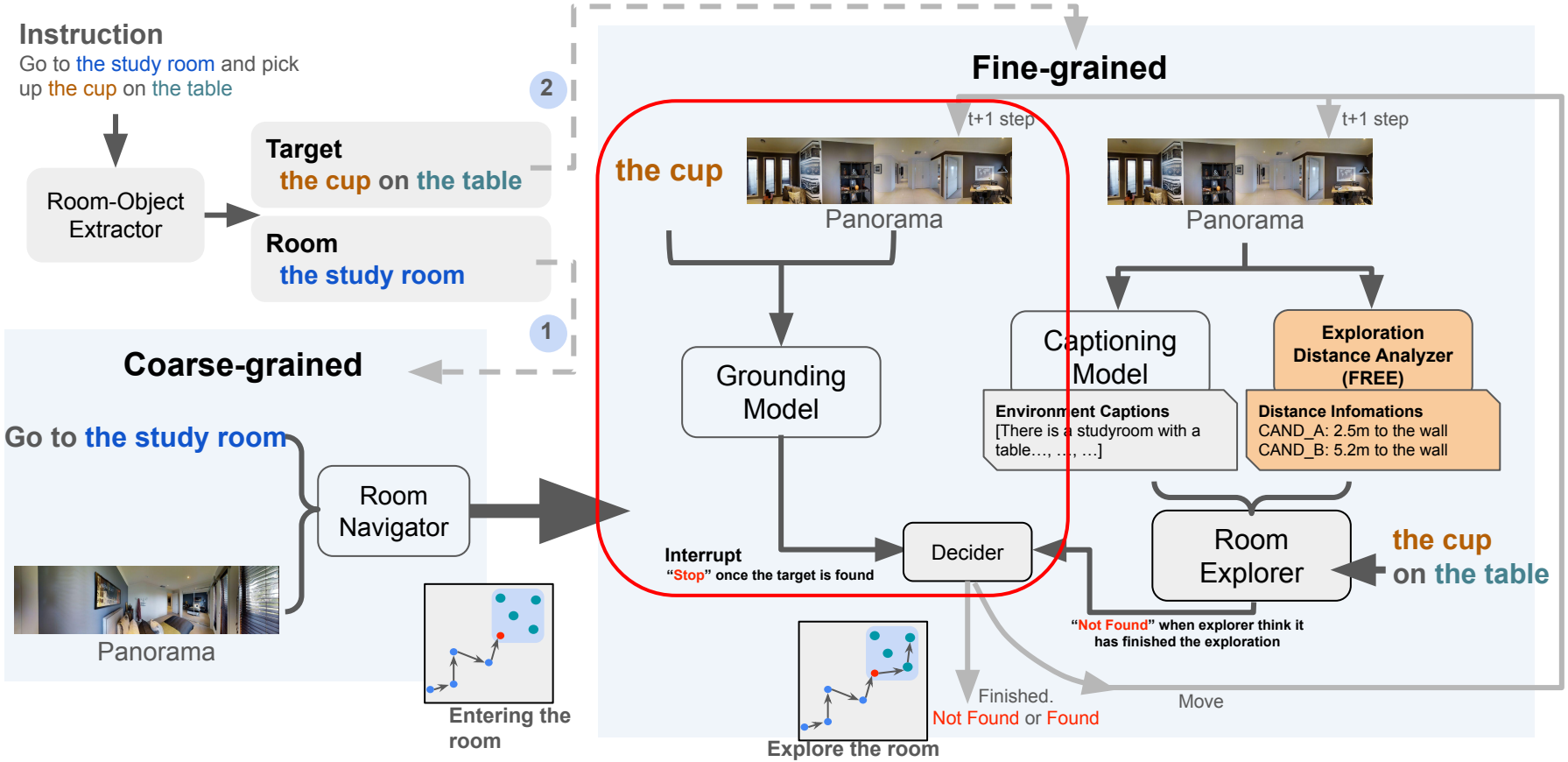

Instruction
Go to the study room and pick up the cup on the table
Room-Object Extractor
Target
the cup on the table
Room
the study room
2
1
Coarse-grained
Go to the study room
Room Navigator
Panorama
Entering the room
Fine-grained
the cup
t+1 step
Panorama
t+1 step
Panorama
Grounding Model
Captioning Model
Environment Captions
[There is a studyroom with a table…, …, …]
Exploration Distance Analyzer (FREE)
Distance Infomations
CAND_A: 2.5m to the wall
CAND_B: 5.2m to the wall
Room Explorer
the cup on the table
Interrupt
"Stop" once the target is found
Decider
"Not Found" when explorer think it has finished the exploration
Finished.
Not Found or Found
Move
Explore the room

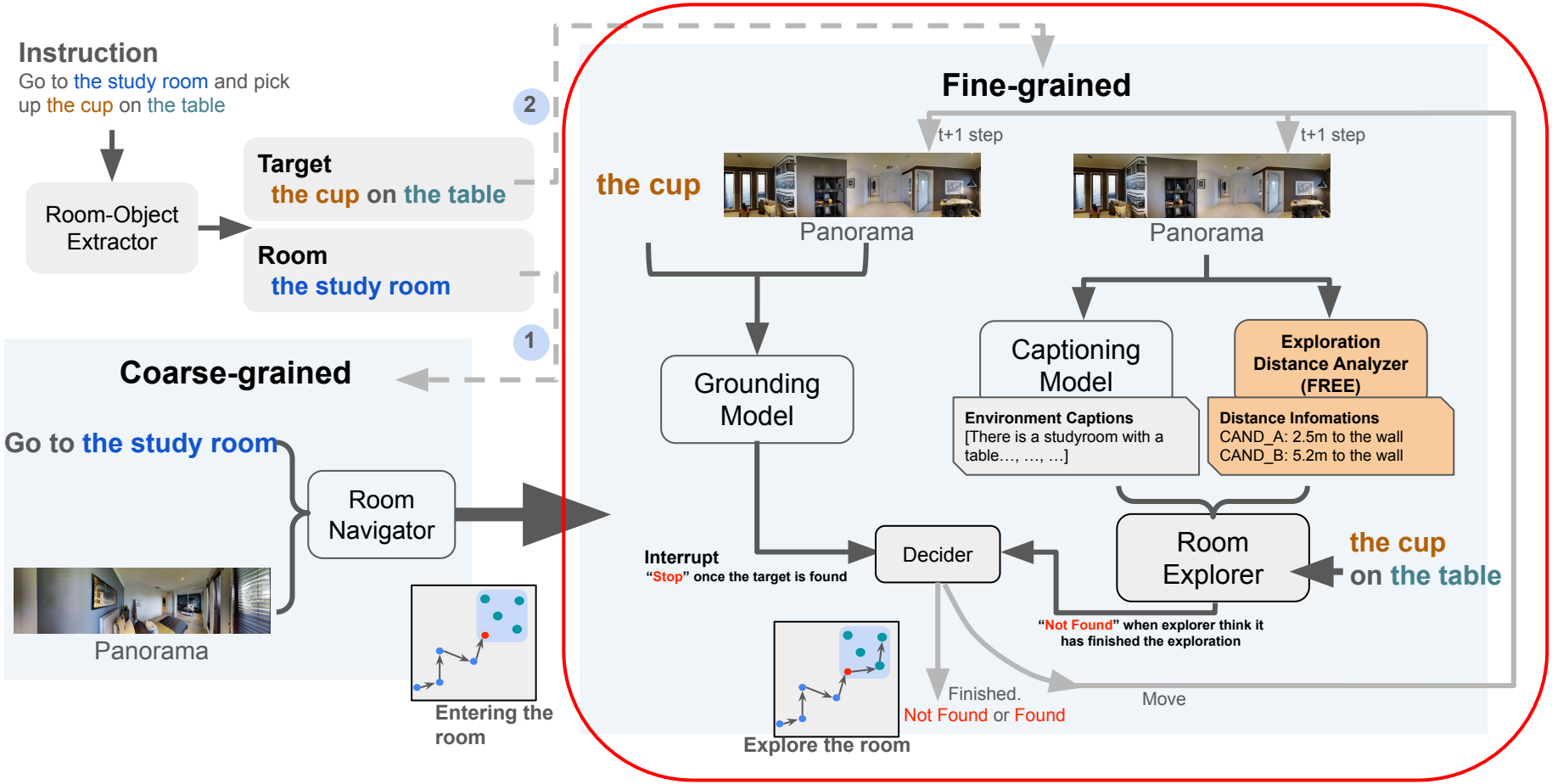

Instruction
Go to the study room and pick up the cup on the table
2
Room-Object Extractor
Target
the cup on the table
Room
the study room
1
Coarse-grained
Go to the study room
Room Navigator
Panorama
Entering the room
Fine-grained
t+1 step
t+1 step
the cup
Panorama
Panorama
Grounding Model
Captioning Model
Environment Captions
[There is a studyroom with a table…, …, …]
Exploration Distance Analyzer (FREE)
Distance Infomations
CAND_A: 2.5m to the wall
CAND_B: 5.2m to the wall
Interrupt
"Stop" once the target is found
Decider
Room Explorer
the cup on the table
"Not Found" when explorer think it has finished the exploration
Finished.
Not Found or Found
Move
Explore the room


Instruction
Go to the study room and pick up the cup on the table
Room-Object Extractor
Target
the cup on the table
Room
the study room
2
1
Coarse-grained
Go to the study room
Room Navigator
Panorama
Entering the room
Fine-grained
t+1 step
t+1 step
the cup
Panorama
Panorama
Grounding Model
Captioning Model
Environment Captions
[There is a studyroom with a table…, …, …]
Exploration Distance Analyzer (FREE)
Distance Infomations
CAND_A: 2.5m to the wall
CAND_B: 5.2m to the wall
Interrupt
"Stop" once the target is found
Decider
Room Explorer
the cup on the table
"Not Found" when explorer think it has finished the exploration
Finished.
Not Found or Found
Move
Explore the room


Instruction
Go to the study room and pick up the cup on the table
Room-Object Extractor
Target
the cup on the table
Room
the study room
2
1
Coarse-grained
Go to the study room
Room Navigator
Panorama
Entering the room
Fine-grained
t+1 step
t+1 step
the cup
Panorama
Panorama
Grounding Model
Captioning Model
Environment Captions
[There is a studyroom with a table…, …, …]
Exploration Distance Analyzer (FREE)
Distance Infomations
CAND_A: 2.5m to the wall
CAND_B: 5.2m to the wall
Interrupt
"Stop" once the target is found
Decider
Room Explorer
the cup on the table
"Not Found" when explorer think it has finished the exploration
Finished.
Not Found or Found
Move
Explore the room

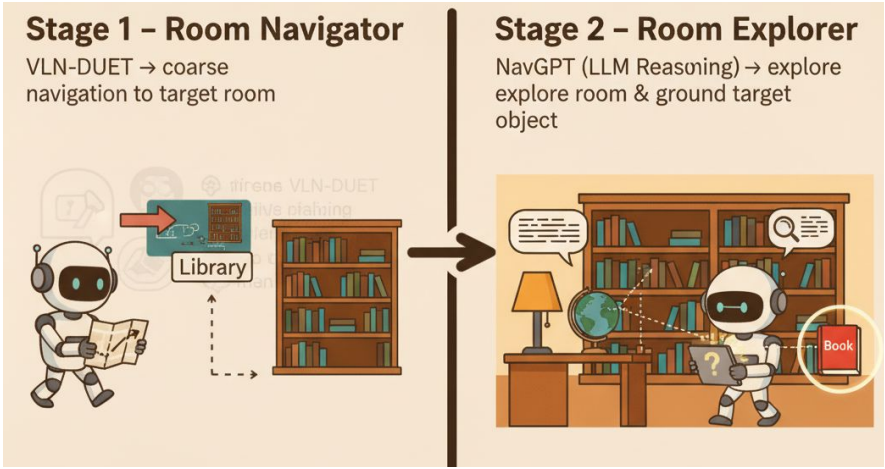

Instruction
Go to the study room and pick up the cup on the table
Room-Object Extractor
Target
the cup on the table
Room
the study room
2
1
Coarse-grained
Go to the study room
Room Navigator
Panorama
Entering the room
Fine-grained
the cup
t+1 step
Panorama
t+1 step
Panorama
Grounding Model
Captioning Model
Exploration Distance Analyzer (FREE)
Environment Captions
[There is a studyroom with a table…, …, …]
Distance Infomations
CAND_A: 2.5m to the wall
CAND_B: 5.2m to the wall
Interrupt
"Stop" once the target is found
Decider
Room Explorer
the cup on the table
"Not Found" when explorer think it has finished the exploration
Finished.
Not Found or Found
Move
Explore the room


Instruction
Go to the study room and pick up the cup on the table
Room-Object Extractor
Target
the cup on the table
Room
the study room
2
1
Coarse-grained
Go to the study room
Room Navigator
Panorama
Entering the room
Fine-grained
the cup
t+1 step
Panorama
t+1 step
Panorama
Grounding Model
Captioning Model
Environment Captions
[There is a studyroom with a table…, …, …]
Exploration Distance Analyzer (FREE)
Distance Infomations
CAND_A: 2.5m to the wall
CAND_B: 5.2m to the wall
Interrupt
"Stop" once the target is found
Decider
Room Explorer
the cup on the table
"Not Found" when explorer think it has finished the exploration
Finished.
Not Found or Found
Move
Explore the room

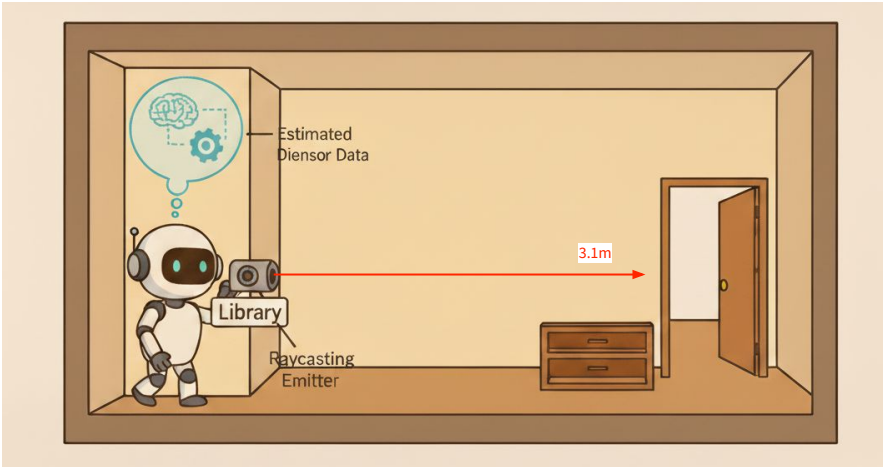

Instruction
Go to the study room and pick up the cup on the table
Room-Object Extractor
Target
the cup on the table
Room
the study room
2
1
Coarse-grained
Go to the study room
Room Navigator
Panorama
Entering the room
Fine-grained
the cup
t+1 step
Panorama
t+1 step
Panorama
Grounding Model
Captioning Model
Environment Captions
[There is a studyroom with a table…, …, …]
Exploration Distance Analyzer (FREE)
Distance Infomations
CAND_A: 2.5m to the wall
CAND_B: 5.2m to the wall
Interrupt
"Stop" once the target is found
Decider
Room Explorer
the cup on the table
"Not Found" when explorer think it has finished the exploration
Finished.
Not Found or Found
Move
Explore the room

Our Design

Coarse-to-Fine Design



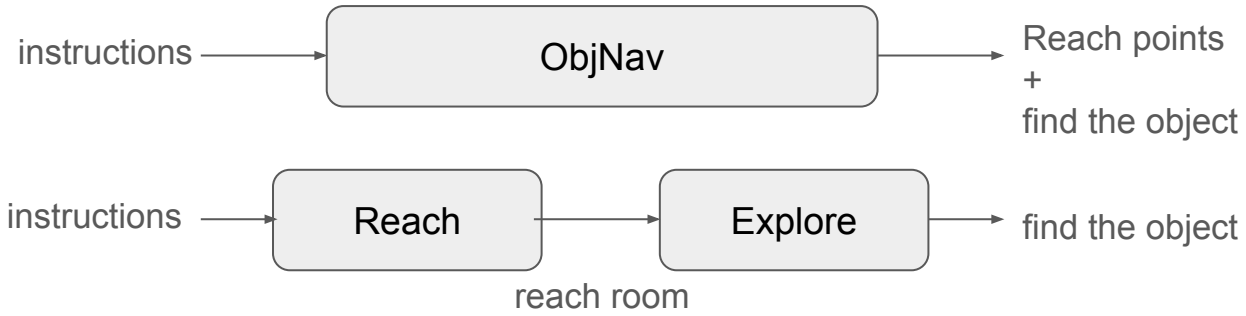
FREE module (Free-space Raycasting Estimation Engine)



Our Design

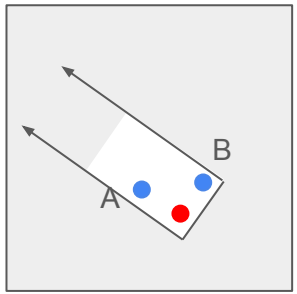
Coarse-to-Fine Design

- We directly add an additional action, “Not Found,” to the existing model, which slightly decreases its ability to navigate to the target room, reducing the Reach SR from 59.4% to 47%.
- Our navigation module ensures that the agent has at least a 57% chance of reaching the target room.



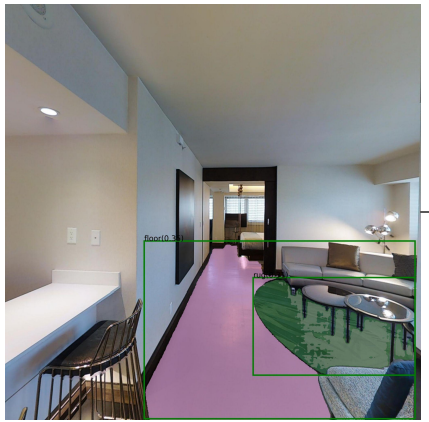
Free-space Raycasting Estimation Engine (FREE)

- Limitation of NavGPT
  - Relies on captions & object info
  - **Loses spatial distance** info during captioning
- Human Search Behavior Insight
  - Use common sense → move toward likely object locations
  - Prefer **open areas** over narrow paths (avoid backtracking, improve efficiency)



A: There is a dining table…… (5m)
B: A mirror on the wall…… (1.4m)
Instruction: find the fourth chair in the living room

Free-space Raycasting Estimation Engine (FREE)

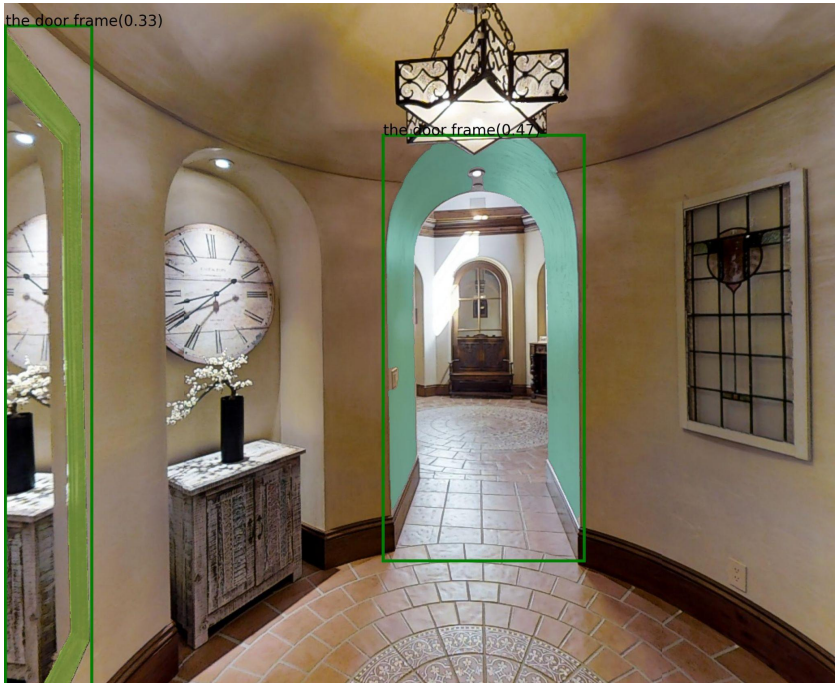


SAM output

camera projection

A: 5 meters
B: 2 meters

Free-space Raycasting Estimation Engine (FREE)

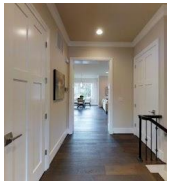


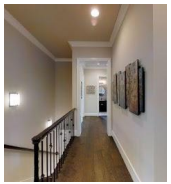
Experiments & Discussion

Baseline

- **DUET**[1]: Supervised method.
- **NavGPT**[2]: Unsupervised method.
- **Gemini**: Unsupervised method(Multimodal LLM).

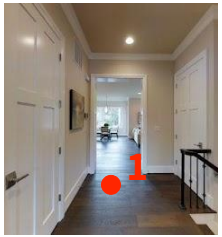
Baseline - Gemini

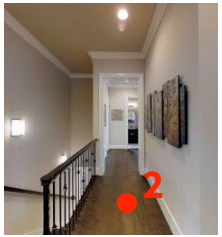


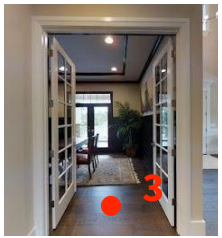


Candidate's WCS coordinates

(1) (2.63, 1.25, 1.02)
(2) (2.14, 1.12, 1.032)
(3) (2.23, 1.8, 1.031)







Gemini

action

[1] Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V (arxiv)

Comparison with Baseline

- **ROAM leads** in most metrics, esp. Reach SR / SPL
  - Coarse-to-Fine design: room-level labels → lower cost, higher accuracy
  - **Reach&Found SR**: clear gain by exploring within correct room
- DUET: (short paths, low coverage)
- LLM baselines: weak on spatial/visual understanding → poor results

	Methods	Coverage	len (reach)	len (explore)	Reach SR	Reach&Found SR	Reach SPL	Reach&Found SPL
supervised	DUET	69.5%	13.1	7.8	53.8%	33.8%	0.370	0.042
unsupervised	NavGPT	59.1%	10.6	2.66	8.2%	5.4%	0.070	0.010
unsupervised	Gemini-2.0-flash-baseline	80.9%	10.0	14.2	39.4%	22.0%	0.289	0.015
hybrid	ROAM(Ours)-GPT-3.5	**82.1%**	11.2	9.9	58.6%	37.6%	**0.441**	**0.061**
hybrid	ROAM(Ours)-GPT-4o	**82.8%**	12.8	22.0	**62.6%**	**41.4%**	**0.454**	0.056

Ablation Study

- (1) → (2): Two-stage design ↑ Reach SR / SPL, (enabling exploration in the correct room)
- (2) → (3): FREE ↑ SPL (LLM gains spatial awareness for more efficient exploration)

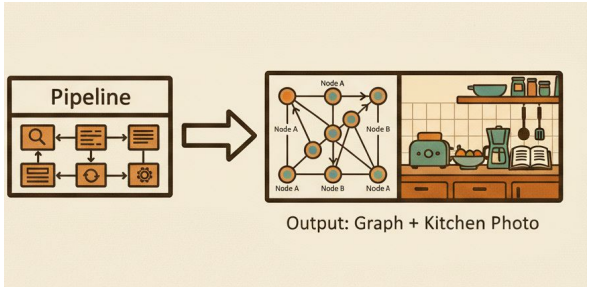
#	Two-Staged Pipeline	FREE Module	Coverage	Reach SR	Reach&Found	Reach SPL	Reach&Found SPL
(1)	✗	✗	75.7%	7.8%	4.4%	0.067	0.008
(2)	✓	✗	79.2%	59.4%	37.2%	0.442	0.056
(3)	✓	✓	**82.1%**	58.6%	37.6%	0.441	**0.061**

ROAM in Original REVERIE

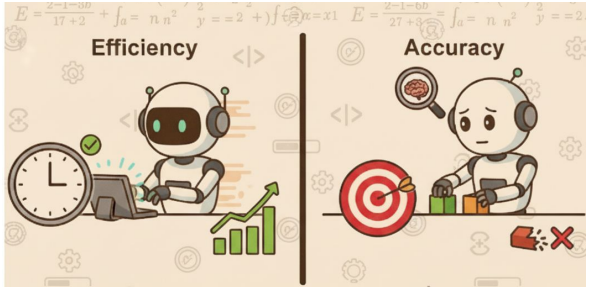
#	Two-Staged Pipeline	FREE Module	Reach SR	Reach SPL	SR	SPL
(1)	✗	✗	8.1%	0.066	9.2%	0.068
(2)	✓	✗	62.0%	0.481	42.0%	0.237
(3)	✓	✓	64.0%	0.484	45.2%	0.251

Conclusion

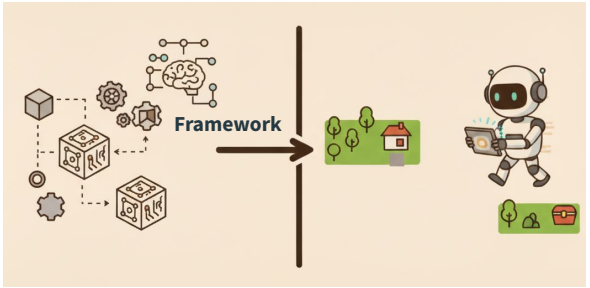
New Task: Navigation Not Found (NAV-NF)



Dataset



Metrics



ROAM Framework

Thanks for listening