
NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models

Gengze Zhou¹ Yicong Hong² Qi Wu¹

¹The University of Adelaide ²The Australian National University

{gengze.zhou, qi.wu01}@adelaide.edu.au yicong.hong@anu.edu.au

- Introduction
- Method
- Experiment

- Introduction
- Method
- Experiment

Introduction

- A trend of integrating LLMs into embodied robotics:
 - The **development** of techniques for **processing textual information** provides **an abundant source of natural language training data** for learning **interdisciplinary and generalizable** knowledge.
 - By accessing unlimited language data, significant **emergent abilities** are observed when **scaling up the model**, resulting in a remarkable **enhancement** in the **reasoning capabilities** when solving problems across **wide domains**.
- However, the **reasoning ability of LLMs in navigation** is still **under-explored**, i.e. , can LLMs **understand the interactive world, the actions, and consequences in text form**, and use all the information to solve a navigation task?

- Introduction
- Method
- Experiment

VLN Problem Formulation

- Given a natural language instruction W , composed of a series of words: $\{w_1, w_2, w_3, \dots, w_n\}$
- At every step s_t , the agent interprets the current location via the simulator to obtain an observation O . This observation comprises N alternative viewpoints, representing the egocentric perspectives of agents in varying orientations.
- Each unique viewpoint is denoted as $o_i (i \leq N)$, with its associated angle direction represented as $a_i (i \leq N)$. The observation can thus be defined as $\mathcal{O}_t \triangleq [\langle o_1, a_1 \rangle, \langle o_2, a_2 \rangle, \dots, \langle o_N, a_N \rangle]$
- Select from the M navigable viewpoints: $\mathcal{O}_t^C \triangleq [\langle o_1^C, a_1^C \rangle, \langle o_2^C, a_2^C \rangle, \dots, \langle o_M^C, a_M^C \rangle]$

- NavGPT is a system that interacts with environments, language guidance, and navigation history to perform action prediction.
- The navigation history of observation O , LLM reasoning R and action A triplets for the previous t steps:

$$\mathcal{H}_{<t+1} \triangleq [\langle \mathcal{O}_1, \mathcal{R}_1, \mathcal{A}_1 \rangle, \langle \mathcal{O}_2, \mathcal{R}_2, \mathcal{A}_2 \rangle, \dots, \langle \mathcal{O}_t, \mathcal{R}_t, \mathcal{A}_t \rangle]$$

- NavGPT needs to synergize the visual perception from VFMs F , language instruction W , history H and navigation system principle P with the help of prompt manager M , define as follow:

$$\langle \mathcal{R}_{t+1}, \mathcal{A}_{t+1} \rangle = LLM(\mathcal{M}(\mathcal{P}), \mathcal{M}(\mathcal{W}), \mathcal{M}(\mathcal{F}(\mathcal{O}_t)), \mathcal{M}(\mathcal{H}_{<t+1}))$$

- **Navigation System Principle P .** The Navigation System Principle formulates the behavior of LLM as a VLN agent. It clearly defines the VLN task and the basic reasoning format and rules for NavGPT at each navigation step.
- **Visual Foundation Models F .** NavGPT as an LLM agent requires visual perception and expression ability from VFMs to translate the current environment's visual observation into natural language description.

- The navigation history of observation O , LLM reasoning R and action A triplets for the previous t steps:

$$\mathcal{H}_{<t+1} \triangleq [\langle \mathcal{O}_1, \mathcal{R}_1, \mathcal{A}_1 \rangle, \langle \mathcal{O}_2, \mathcal{R}_2, \mathcal{A}_2 \rangle, \dots, \langle \mathcal{O}_t, \mathcal{R}_t, \mathcal{A}_t \rangle]$$

- NavGPT needs to synergize the visual perception from VFMs F , language instruction W , history H and navigation system principle P with the help of prompt manager M , define as follow:

$$\langle \mathcal{R}_{t+1}, \mathcal{A}_{t+1} \rangle = LLM(\mathcal{M}(\mathcal{P}), \mathcal{M}(\mathcal{W}), \mathcal{M}(\mathcal{F}(\mathcal{O}_t)), \mathcal{M}(\mathcal{H}_{<t+1}))$$

- **Navigation History $\mathcal{H}_{<t+1}$.** The navigation history is essential for NavGPT to evaluate the progress of the completion of the instruction, to update the current state, and make the following decisions.
- **Prompt Manager M .** The key to using LLM as a VLN agent is to convert all the above content into a natural language that LLM can understand.

Navigation System Principle

- Defines the VLN task and the basic reasoning format and rules for NavGPT at each navigation step.

You are an intelligent embodied agent that follows an instruction to navigate in an indoor environment. Your task is to move among the static viewpoints (positions) of a pre-defined graph of the environment, and try to reach the target viewpoint as described by the given instruction with the least steps.

At the beginning of the navigation, you will be given an instruction of a trajectory which describes all observations and the action you should take at each step.

During navigation, at each step, you will be at a specific viewpoint and receive the history of previous steps you have taken (containing your "Thought", "Action", "Action Input" and "Observation" after the "Begin!" sign) and the observation of current viewpoint (including scene descriptions, objects, and navigable directions/distances within 3 meters).

Orientations range from -180 to 180 degrees: "0" signifies forward, "right 90" rightward, "right (or left) 180" backward, and "left 90" leftward.

You make actions by selecting navigable viewpoints to reach the destination. You are encouraged to explore the environment while avoiding revisiting viewpoints by comparing current navigable and previously visited IDs in previous "Action Input". The ultimate goal is to stop within 3 meters of the destination in the instruction. If destination visible but the target object is not detected within 3 meters, move closer.

Navigation System Principle

At each step, you should consider:

(1) According to Current Viewpoint observation and History, have you reached the destination?

If yes you should stop, output the 'Final Answer: Finished!' to stop.

If not you should continue:

(2) Consider where you are on the trajectory and what should be the next viewpoint to navigate according to the instruction.

Use the action_maker tool, input the next navigable viewpoint ID to move to that location.

Show your reasoning in the Thought section.

Here are the descriptions of the action_maker tool:

Can be used to move to next adjacent viewpoint.

The input to this tool should be a viewpoint ID string of the next viewpoint you wish to visit.

For example:

Action: action_maker

Action Input: "4a153b13a3f6424784cb8e5dabbb3a2c".

Every viewpoint has a unique viewpoint ID. You are very strict to the viewpoint ID and will never fabricate nonexistent IDs.

Navigation System Principle

Starting below, you should follow this format:

Instruction: an instruction of a trajectory which describes all observations and the actions should be taken

Initial Observation: the initial observation of the environment

Thought: you should always think about what to do next and why

Action: the action to take, must be one of the tools [action_maker]

Action Input: "Viewpoint ID"

Observation: the result of the action

... (this Thought/Action/Action Input/Observation can repeat N times)

Thought: I have reached the destination, I can stop.

Final Answer: Finished!

Begin!

Instruction: {instruction}

Initial Observation: {init_observation}

Thought: I should start navigation according to the instruction,

Visual Perceptron for NavGPT

- Take visual signals as a foreign language and handle the visual input using different visual foundation models to translate them into natural language, shown in figure 2.

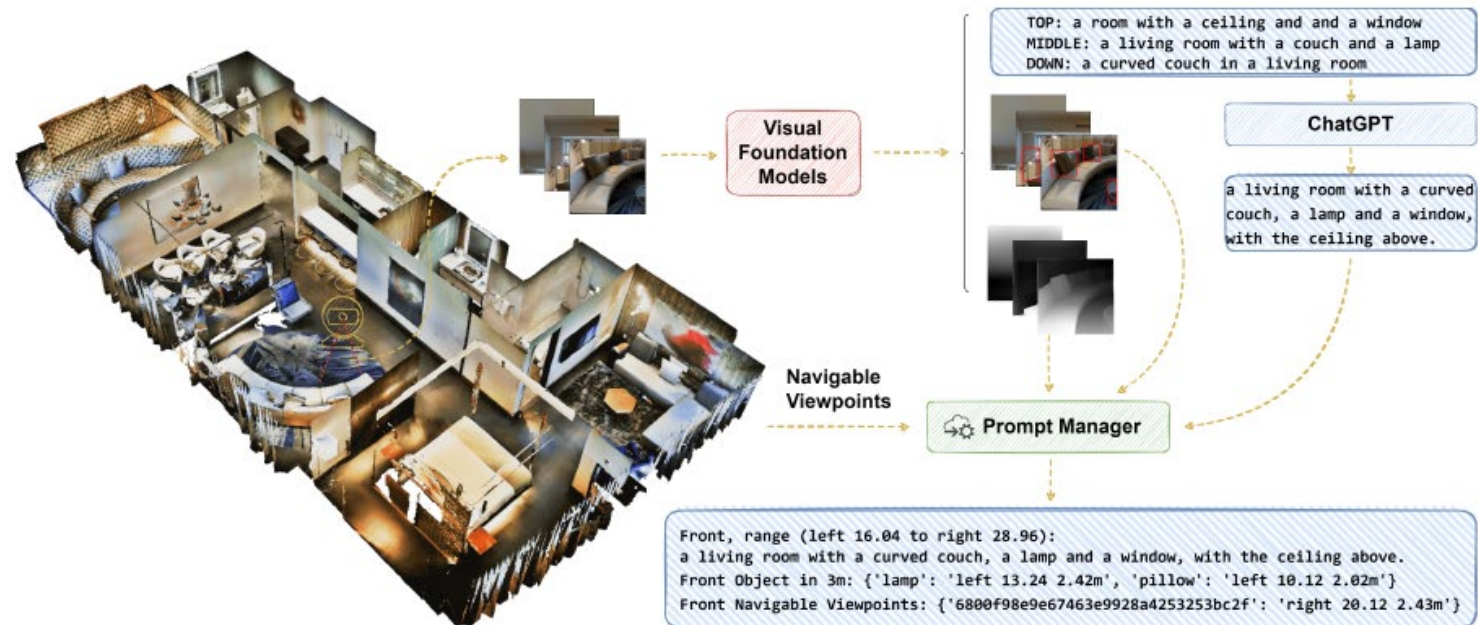
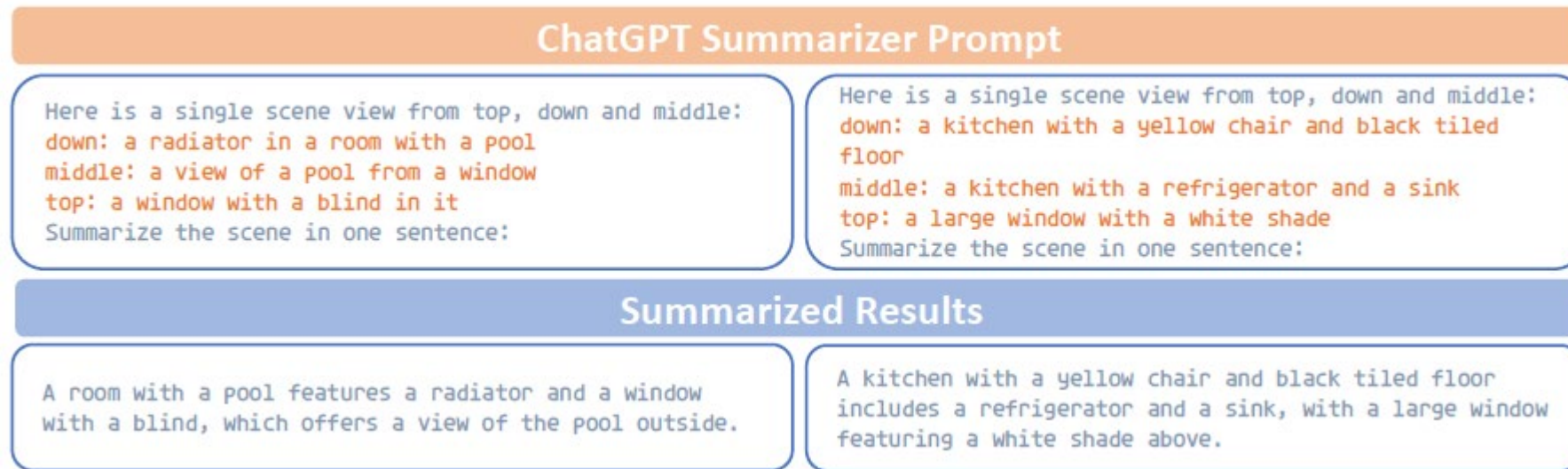


Figure 2: The process of forming natural language description from visual input. We used 8 directions to represent a viewpoint and show the process of forming the descriptions for one of the directions.

- Prompt BILP-2 to generate a decent language description of each view with a detailed depiction of the shapes and color of objects and the scenes they are in.

Convert Visual Perception to Language Description(Appendix)

- The visual perception process for each direction includes two steps, including using BILP-2 to caption the three images(down, middle and top), then, summarizing the descriptions by the GPT-3.5 summarizer.
- **BILP-2 Prompt.** The navigation history is essential for NavGPT to evaluate the progress of the completion of the instruction, to update the current state, and make the following decisions.
 - Selected "This is a scene of " as the preferred prompt for BILP-2 to generate descriptions for each image.



- **GPT-3.5 Summarizer Prompt.** Adopt a GPT-3.5 summarizer to summarize them into one sentence following the template:
 - "Here is a single scene view from top, down and middle:\n{description}\nSummarize the scene in one sentence:",

Convert Visual Perception to Language Description(Appendix)

- **Observation description examples.** The prompt manager will take the current heading of the agent as the “front” direction, and **calculate** the **relative angle** between the agent’s **current heading and the detected objects** as well as **the navigable viewpoints**, concatenating the descriptions from each direction clockwise.

Observation for a viewpoint

```
Front, range (left 23.50 to right 21.50):  
'The scene features a brick wall with a fireplace, displaying a mix of brown, white, and yellow colors.'  
Front Objects in 3m: None  
Front Navigable Viewpoints: None  
Front Right, range (right 21.50 to right 66.50):  
'A brick wall with a doorway is illuminated by a light shining down on it.'  
Front Right Objects in 3m: None  
Front Right Navigable Viewpoints: None  
Right, range (right 66.50 to right 111.50):  
'A staircase with a bottle in it leads up to a brick archway with a staircase inside, all surrounded by a brick wall with a white circle on it.'  
Right Objects in 3m: None  
Right Navigable Viewpoints: {'1d337fde52e84923871db95009731c41': 'right 86.26, 2.03m'}  
Rear Right, range (right 111.50 to right 156.50):  
'A brick wall with a light shining through it and a brick pillar stands beneath a brick archway with a surrounding brick wall.'  
Rear Right Objects in 3m: None  
Rear Right Navigable Viewpoints: None
```

```
Rear, range (right 156.50 to left 158.50):  
'A small room with a fireplace, vase, and hanging light fixture surrounded by brick walls.'  
Rear Objects in 3m: {'table': 'left 165.41, 1.75m'}  
Rear Navigable Viewpoints: {'4b587c327d8040feaa14adfeaa6e84d': 'right 180.00, 1.07m'}  
Rear Left, range (left 158.50 to left 113.50):  
'A scene with a brick wall featuring vases, a picture of a man, and another brick wall from different perspectives.'  
Rear Left Objects in 3m: {'table': 'left 137.15, 1.75m'}  
Rear Left Navigable Viewpoints: None  
Left, range (left 113.50 to left 68.50):  
'The scene features a brick wall with a brown color and a white ceiling viewed from both the top and middle perspectives.'  
Left Objects in 3m: None  
Left Navigable Viewpoints: None  
Front Left, range (left 68.50 to left 23.50):  
'The scene features a brick wall with a corner, a brick floor, and a yellow and brown color scheme.'  
Front Left Objects in 3m: None  
Front Left Navigable Viewpoints: None
```

Summarized Observation in History

Scene from the viewpoint is a small room with a brick wall, fireplace, vase, picture of a man, and various colors on the walls and ceiling, as well as a staircase and archway with brick walls and pillars.

- **Auxiliary translators**, translating visual input into their own "language" like the class of objects and corresponding bounding boxes.
 - Extract the depth information of the center pixel of the object provided by the Matterport3D simulator, leaving the object within 3 meters from the current viewpoint.

History with Summarizer

Observation for a viewpoint

Front, range (left 23.50 to right 21.50):

'The scene features a brick wall with a fireplace, displaying a mix of brown, white, and yellow colors.'

Front Objects in 3m: None

Front Navigable Viewpoints: None

Front Right, range (right 21.50 to right 66.50):

'A brick wall with a doorway is illuminated by a light shining down on it.'

Front Right Objects in 3m: None

Front Right Navigable Viewpoints: None

Right, range (right 66.50 to right 111.50):

'A staircase with a bottle in it leads up to a brick archway with a staircase inside, all surrounded by a brick wall with a white circle on it.'

Right Objects in 3m: None

Right Navigable Viewpoints: {'1d337fde52e84923871db95009731c41': 'right 86.26, 2.03m'}

Rear Right, range (right 111.50 to right 156.50):

'A brick wall with a light shining through it and a brick pillar stands beneath a brick archway with a surrounding brick wall.'

Rear Right Objects in 3m: None

Rear Right Navigable Viewpoints: None

Rear, range (right 156.50 to left 158.50):

'A small room with a fireplace, vase, and hanging light fixture surrounded by brick walls.'

Rear Objects in 3m: {'table': 'left 165.41, 1.75m'}

Rear Navigable Viewpoints: {'4b587c327d8040feaa14adfeaa6e84d': 'right 180.00, 1.07m'}

Rear Left, range (left 158.50 to left 113.50):

'A scene with a brick wall featuring vases, a picture of a man, and another brick wall from different perspectives.'

Rear Left Objects in 3m: {'table': 'left 137.15, 1.75m'}

Rear Left Navigable Viewpoints: None

Left, range (left 113.50 to left 68.50):

'The scene features a brick wall with a brown color and a white ceiling viewed from both the top and middle perspectives.'

Left Objects in 3m: None

Left Navigable Viewpoints: None

Front Left, range (left 68.50 to left 23.50):

'The scene features a brick wall with a corner, a brick floor, and a yellow and brown color scheme.'

Front Left Objects in 3m: None

Front Left Navigable Viewpoints: None

Summarized Observation in History

Scene from the viewpoint is a small room with a brick wall, fireplace, vase, picture of a man, and various colors on the walls and ceiling, as well as a staircase and archway with brick walls and pillars.

- Prompt template:

"Given the description of a viewpoint. Summarize the scene from the viewpoint in one concise sentence.

\nDescription:

\n{description}

\n Summarization: The scene from the viewpoint is a".

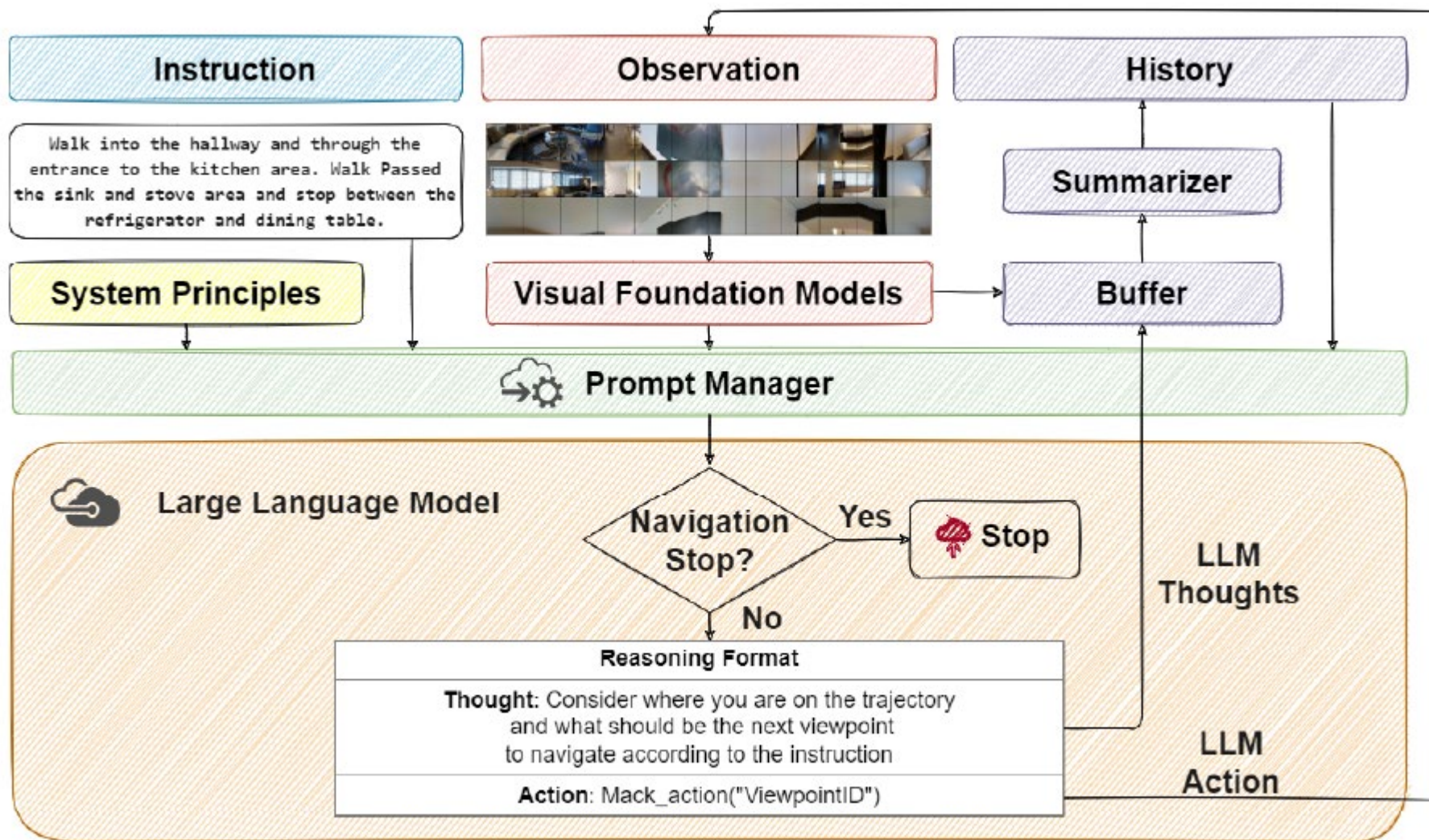
Synergizing reasoning and actions in LLMs

- In order to explicitly access and enhance the agent's comprehension of the current state during navigation, expand the agent's action space to $\tilde{\mathcal{A}} = \mathcal{A} \cup \mathcal{R}$, where $\mathcal{R} \in \mathcal{L}$ is in the entire language space \mathcal{L} , denoting the thought or reasoning trace of the agent.
- The reasoning traces R of the agent will not trigger any interaction with the external environment, therefore no observation will be returned when the agent is outputting the reasoning during each navigation step.
- Synergize the NavGPT's actions and thoughts by prompting it to make navigation decisions after outputting the reasoning trace at each step.
- Introducing the reasoning traces aims to bootstrap the LLMs in two aspects:
 - Firstly, prompting the LLMs to think before choosing an action, enables LLMs to perform complex reasoning in planning and creating strategies to follow the instructions under the new observations.
 - Secondly, including reasoning traces R in the navigation history $H_{<t}$ enhances the problem-solving ability of NavGPT.

$$\mathcal{H}_{<t+1} \triangleq [\langle \mathcal{O}_1, \mathcal{R}_1, \mathcal{A}_1 \rangle, \langle \mathcal{O}_2, \mathcal{R}_2, \mathcal{A}_2 \rangle, \dots, \langle \mathcal{O}_t, \mathcal{R}_t, \mathcal{A}_t \rangle]$$

- With the Navigation System Principle P , translated results from VFMs, and the History of Navigation $\mathcal{H}_{<t}$, the prompt manager parses and reformates them into prompts for LLMs.
- For Navigation System Principle P , NavGPT prompt manager will create a prompt to convey LLMs with the **rules**, declaring the VLN **task definition**, **defining** the simulation environment for NavGPT, and **restricting** LLMs' **behavior** in the given reasoning format.
- For perception results from VFMs F , the prompt manager **gathers** the **results from each direction** and **orders** the **language description** by taking the current orientation of NavGPT as the front, arranging the description from 8 directions into prompt by **concatenating** them **clockwise**.
- For navigation history $\mathcal{H}_{<t}$, the observation, reasoning, and actions triples $\langle \mathcal{O}_i, \mathcal{R}_i, \mathcal{A}_i \rangle$ are stored in a history buffer, to **handle the length of history**, the prompt manager utilizes GPT-3.5 to **summarize the observations from viewpoints** in the trajectory, inserting the summarized observations into the observation, reasoning, and actions triples in the prompt.

NavGPT prompt manager



NavGPT prompt manager



Visual Foundation Models

Prompt Manager

NavGPT Prompt (t = 0)

{Navigation system principles}

Begin!

Instruction: Exit the sewing room. Turn right. Go toward the glass cabinet with the dolls in it. Turn into the doorway on the left. Pass the bed and go through the next doorway on the left into the bathroom. Wait by the sink.

Init Observation:

Front, range (left 3.51 to right 41.49):

'A bathroom with a wooden door and tile floor has a doorway with a picture of a pigeon on it and a wooden archway with a light fixture above it.'

Front Objects in 3m: {'cabinet': 'right 3.31, 1.82m'}

Front Navigable Viewpoints: None

Front Right, range (right 41.49 to right 86.49):

'A view of a bathroom with a white tile floor and door, a kitchen with a wooden door and pictures, and a ceiling fan with a light in it from a top, down, and middle perspective.'

Front Right Objects in 3m: None

Front Right Navigable

Viewpoints: {'cc0df704278a4f009b2239afd9a2ec9d': 'right 61.44, 1.40m'}

Right, range (right 86.49 to right 131.49):

'A wooden door with a metal handle and a picture on it is framed by a wooden frame.'

Right Objects in 3m: None

Right Navigable Viewpoints: None

Rear Right, range (right 131.49 to right 176.49):

'A wooden door with a handle on it is shown from different angles, including a close-up and with a light shining through it.'

Rear Right Objects in 3m: {'light': 'right 167.02, 1.23m'}

Rear Right Navigable Viewpoints: None

Rear, range (right 176.49 to left 138.51):

'A wooden closet with a metal rack and hanger rack, and a small bathroom with a wooden cabinet and light.'

Rear Objects in 3m: {'light': 'left 160.33, 1.23m'}

Rear Navigable Viewpoints: None

NavGPT Response

first, I need to exit the sewing room. The only navigable viewpoint is in the front right direction, so I will move there.

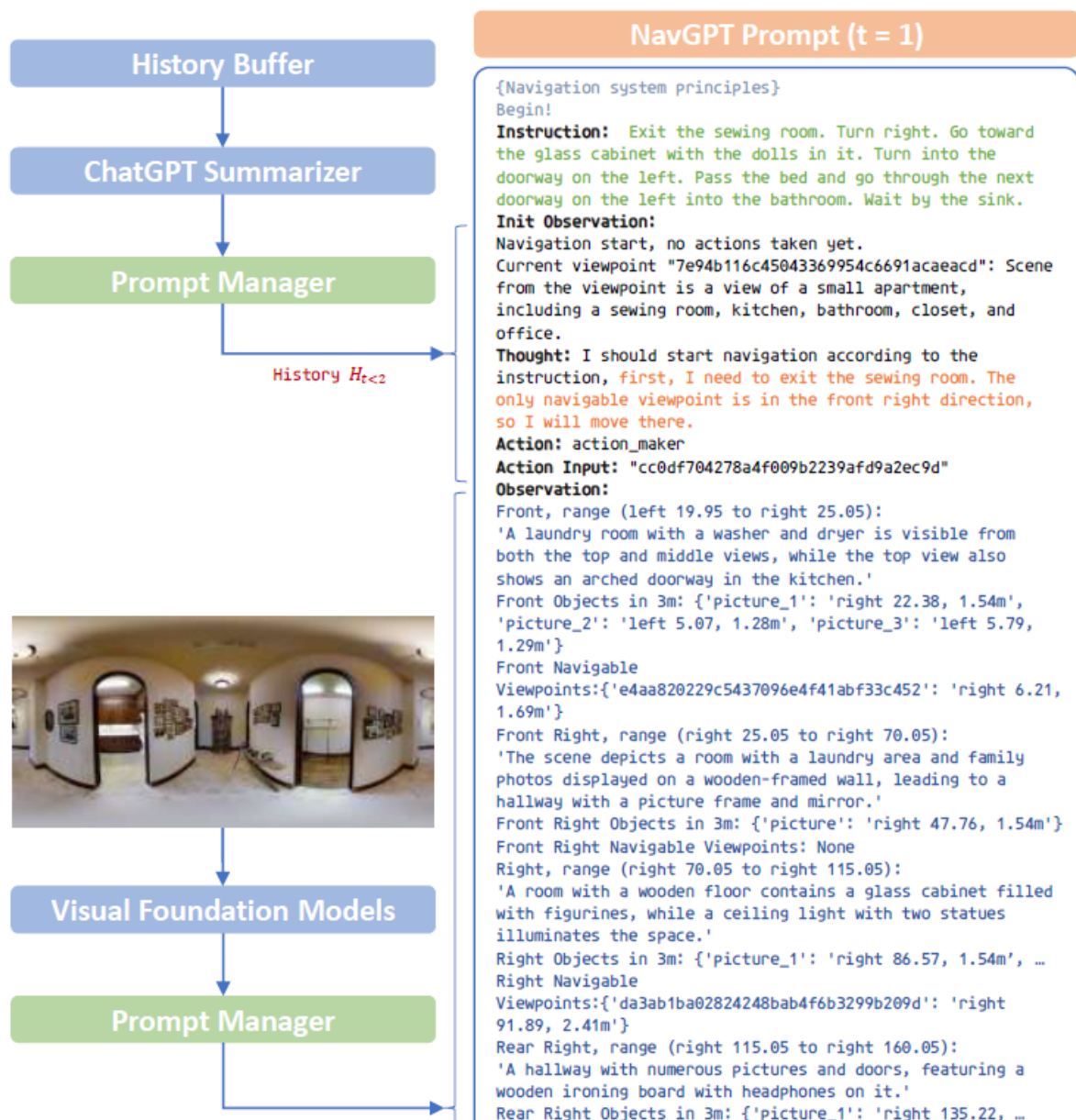
Action: action_maker

Action Input:

"cc0df704278a4f009b2239afd9a2ec9d"

Rear Left, range (left 138.51 to left 93.51):
'A view of a metal rod with a wooden floor and a metal handle, under a white ceiling with a gold and brown pattern.'
Rear Left Objects in 3m: {'light_1': 'left 109.68, 1.27m', 'light_2': 'left 124.96, 1.23m'}
Rear Left Navigable Viewpoints: None
Left, range (left 93.51 to left 48.51):
'The scene depicts a small office with a desk and chair, adjacent to a small bathroom with a sink and cabinets, and a light fixture hanging above the toilet.'
Left Objects in 3m: {'cabinet': 'left 64.14, 1.82m', 'armchair': 'left 60.39, 1.58m', 'sewing machine': 'left 69.51, 1.77m', 'light': 'left 66.11, 1.27m'}
Left Navigable Viewpoints: None
Front Left, range (left 48.51 to left 3.51):
'The scene depicts a small room with a sewing machine and chair, a small kitchen with a sink and wooden cabinet, and a closet with a light fixture and wooden cabinets.'
Front Left Objects in 3m: {'cabinet': 'left 25.28, 1.82m', 'armchair': 'left 17.36, 1.58m', 'sewing machine': 'left 17.26, 1.77m', 'light': 'left 28.54, 1.27m'}
Front Left Navigable Viewpoints: None
Thought: I should start navigation according to the instruction,

NavGPT prompt manager



NavGPT Response

I have exited the sewing room. Now, I need to turn right and go toward the glass cabinet with the dolls in it. I see a room with a glass cabinet filled with figurines to my right. I will move to the navigable viewpoint in that direction.

Action: action_maker
Action Input:
 "da3ab1ba02824248bab4f6b3299b209d"

Rear Right Navigable Viewpoints: None
 Rear, range (right 160.05 to left 154.95):
 'A closet with a clothes rack and metal rack, a mirror and a light shining through it.'
 Rear Objects in 3m: {'picture_1': 'left 159.73, 1.21m', ...
 Rear Navigable
 Viewpoints: {'7e94b116c45043369954c6691acaeacd': 'right 180.00, 1.40m'}
 Rear Left, range (left 154.95 to left 109.95):
 'The scene depicts a hallway with framed pictures and a clock leading to a room with a yellow archway and more framed pictures.'
 Rear Left Objects in 3m: {'light': 'left 112.76, 1.43m', ...
 Rear Left Navigable Viewpoints: None
 Left, range (left 109.95 to left 64.95):
 'A hallway with a circular tiled floor and a violin hanging from the ceiling, illuminated by a circular ceiling light.'
 Left Objects in 3m: {'light': 'left 87.36, 1.43m', ...
 Left Navigable
 Viewpoints: {'0942757fbc9474590d047d55e04868d': 'left 84.6...
 Front Left, range (left 64.95 to left 19.95):
 'A man is sitting in a room with a picture on the wall, while a hallway with a picture of a man and woman and a mirror with framed photos on it can be seen from different angles.'
 Front Left Objects in 3m: {'light': 'left 63.68, 1.43m', ...
 Front Left Navigable Viewpoints: None
Thought:

- Introduction
- Method
- Experiment

Qualitative Results

- Reasoning capability of GPT-4 for Language-guide Navigation

Short instruction with sub-goal planning

Instruction: Go forward past the bed and the stairs. Stop in the doorway straight ahead.

Init observation: ... A bedroom with hardwood floors, a bed, dresser, nightstand, and a picture hanging on a white ceiling ...

Thought: I need to go forward past the bed and the stairs, and stop in the doorway straight ahead. The rear navigable viewpoint seems to be in the direction of the bed, so I should move there first.

Action: action_maker("b0a6cf6e9b9043248ed1c1ed8db08436")

Observation: ... A hallway with hardwood floors leads to a bathroom with a mirror and light fixtures ...

Thought: I have moved past the bed, and now I see a hallway leading to a bathroom. I should continue forward to find the stairs and the doorway straight ahead. The front navigable viewpoint seems to be in the direction of the hallway, so I should move there next.

Action: action_maker("c53934a9642942ae9b8d9766ee1e51ac")

Observation: ... A view of a hallway with hardwood floors and a staircase...

Thought: I have found the stairs, and now I need to find the doorway straight ahead. The front navigable viewpoint seems to be in the direction of the doorway, so I should move there next.

Action: action_maker("bf10ccfbb089402d9f3e606803dadb0d")

Observation: ... A view of a room with hardwood floors and a doorway leading to a hallway with another doorway...

Thought: I have reached a viewpoint where I can see a doorway straight ahead. I should move to the front navigable viewpoint to get closer to the doorway and check if I have reached the destination.

Action: action_maker("df14f9089a6646f0bbf5054e18fb8951")

Observation: ... A hallway with hardwood floors and white doors leads to a white doorway with a white frame ...

Thought: I have reached the destination of the doorway, I should stop now.

Action: Finish!

Integrating commonsense knowledge

Thought: first I need to walk towards the sink. The sink is not visible in the current viewpoint, but there is a bathroom to the left, which might lead me closer to the sink.

Thought: I am now in the living room area, and I can see the fireplace. I should move closer to the fireplace and look for the stool and table. I will move to the viewpoint 'faed19f97550433b958958f4df869251' to get closer to the fireplace.

Comparison with Supervised Methods

Table 1: Comparison with previous methods on R2R validation unseen split.

Training Schema	Method	TL	NE↓	OSR↑	SR↑	SPL↑
Train Only	Seq2Seq [2]	8.39	7.81	28	21	-
	Speaker Follower [14]	-	6.62	45	35	-
	EnvDrop [53]	10.70	5.22	-	52	48
Pretrain + Finetune	PREVALENT [21]	10.19	4.71	-	58	53
	VLN \curvearrowright BERT [26]	12.01	3.93	69	63	57
	HAMT [6]	11.46	2.29	73	66	61
	DuET [7]	13.94	3.31	81	72	60
No Train	DuET (Init. LXMERT [52])	22.03	9.74	7	1	0
	NavGPT (Ours)	11.45	6.46	42	34	29

- The **limitations** inhibiting the performance of LLMs in solving VLN can be primarily attributed to **two factors**: the **precision of language-based depiction of visual scenes** and the **tracking capabilities regarding objects**.
 - The target object delineated in the instruction is absent.
 - Manage the length of the navigation history.

Effect of Visual Components

- **Effect of granularity in visual observation descriptions**

Table 2: The effect of granularity in visual observation descriptions.

Granularity	#	TL	NE↓	OSR↑	SR↑	SPL↑
FoV@60, 12 views	1	12.38	9.07	14.35	10.19	6.52
FoV@30, 36 views	2	12.67	8.92	15.28	13.89	9.12
FoV@45, 24 views	3	12.18	8.02	26.39	16.67	13.00

- An overly large FoV leading to generalized room descriptions.
- An extremely small FoV hindering object recognition due to limited content.
- **Effect of semantic scene understanding and depth estimation**

Table 3: The effect of semantic scene understanding and depth estimation.

Agent Observation	#	TL	NE↓	OSR↑	SR↑	SPL↑
Baseline	1	16.11	9.83	15.28	11.11	6.92
Baseline + Obj	2	11.07	8.88	23.34	15.97	11.71
Baseline + Obj + Dis	3	12.18	8.02	26.39	16.67	13.00

- Object detectors and depth estimators.
- Baseline: based on the caption results from BILP-2 and powered by GPT-3.5.