VELMA: Verbalization Embodiment of LLM Agents for Vision and Language Navigation in Street View

Raphael Schumann[1], Wanrong Zhu[2], Weixi Feng[2], Tsu-Jui Fu[2], Stefan Riezler[1 3], William Yang Wang[2]

[1] Computational Linguistics, Heidelberg University, Germany
[2] University of California, Santa Barbara
[3] IWR, Heidelberg University, Germany

Outdoor VLN Task

Egocentric Spatial Reasoning
... turn so the orange construction barrier is on your left ... ... a red truck in front of you ... ... a playground on the far right corner ahead ...
Allocentric Spatial Reasoning
... green metal pole with pink flowers on top ... ... building with columns around the windows ... ... stop in between Chase and Dunkin' Donuts ...
Temporal Reasoning
... go straight until you see Chipotle and then ... ... once you passed the underpass ... ... stop when the park on your right ends ...
Other
... proceed straight through three more intersections ... ... if you see Dory Oyster Bar, you have gone too far ...

The agent is embodied in Street View and receives navigation instructions describing the route to a target location. **The instructions are written by humans and make use of landmarks and other visual observations**. Agents need to reason about observations and past trajectory in order to predict the next action. A successfull **trajectory is, on average, 40 steps long**.

Environment

We use the **Touchdown** environment introduced by Chen et al., 2019. It is based on Google's Street View and features **29,641 panorama images** connected by a navigation graph. It covers the dense urban street network spanning lower Manhattan. **The action space is**:

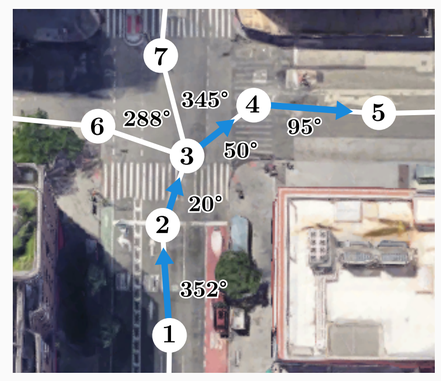
`FORWARD`, `LEFT`, `RIGHT`, `TURN_AROUND`, `STOP`

We fix a problem at intersections that prevented general agents to operate in the old environment.

Path: 1→2→3→4

Old Environment:
`FORWARD`, `FORWARD`, `FORWARD`

New Environment (our):
`FORWARD`, `FORWARD`, `RIGHT`, `FORWARD`



Agent Workflow

Prompt Sequence

Navigate to the described target location!
Action Space: forward, left, right, turn_around, stop
Navigation Instructions:
"*Go straight down the road and turn right at the next intersection. Go straight until there is a* ***Starbucks*** *on your right and turn left at the following intersection. Continue down the block and stop when* ***a mail truck*** *is on your left.*"
Action Sequence:
1. forward
2. forward

There is a 4-way intersection.

4. right
5. forward
6. forward
7. forward

There is a Starbucks on your right.

8. <***next word prediction***>
...

Landmark Extractor

Write a list of visible landmarks in the navigation instructions:
- **Starbucks**
- **a mail truck**

Landmarks

Landmark Scorer

left	slightly left	ahead	slightly right	right

Standardized CLIP Scores (Threshold: 3.5)

"picture of ***Starbucks***"				
2.85	1.29	-1.12	-2.27	**4.15**

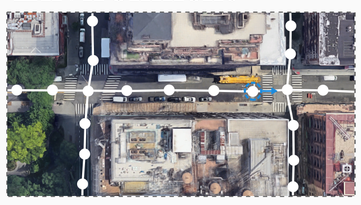
"picture of ***a mail truck***"				
2.15	-0.76	-2.20	1.87	1.98

Structured Output

{ "landmarks": { "*Starbucks*": "*right*" } }

Visible Landmarks

Environment



Action

Panorama & Heading

Number of Edges

Verbalizer

Template Based

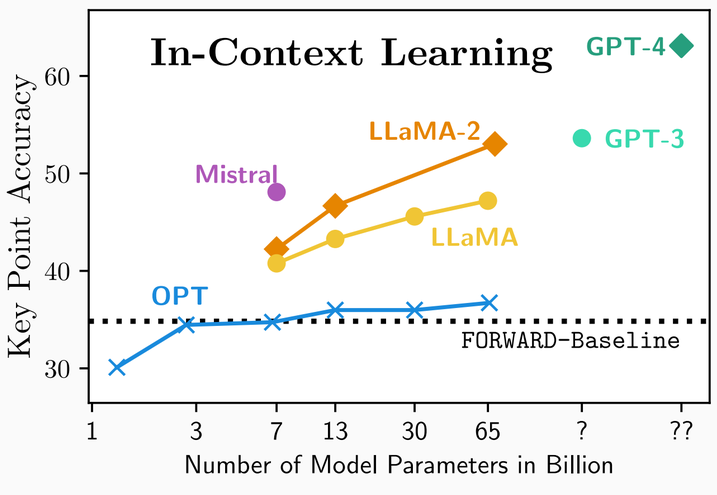
There is a ***Starbucks*** on your ***right***.

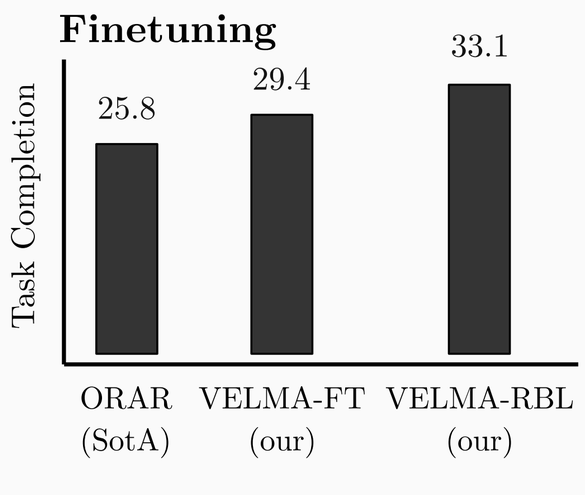
There is a *N*-way intersection

Observation

Every step the LLLM is queried to predict the next action. The prompt includes the **task description, navigation instructions, past trajectory and environment observations like intersections and landmarks**. The action is executed and the new observations are verbalized and appended to the prompt.

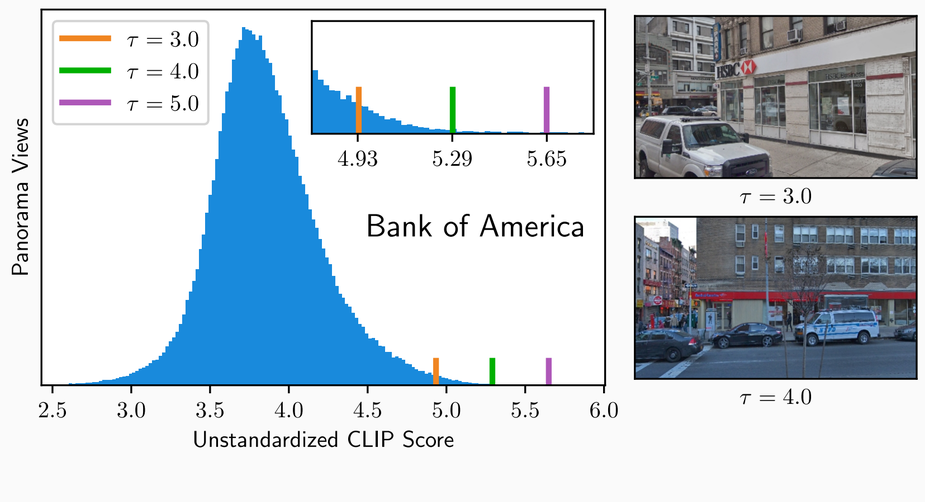
Results





We evaluate different **LLMs with two in-context examples**. GPT-3 (TC: 8.0) and GPT-4 (TC: 16.2) are the only untrained LLMs that achieve task completion rate (TC) > 2.
We finetune LLaMA-7B on the training set (6,000 instances) and beat the previous SotA model ORAR that has explicit access to the images via ResNet features. Using **response based learning (RBL)** instead of teacher forcing, we can improve TC by another 4 points. The results show that verbalization allows to leverage the reasoning of LLMs for urban VLN.

Landmark Detection



We determine the visibility of a landmark in the current panorama by the **similarity score of image embedding and text embedding**, generated by OpenCLIP. The score is standardized over scores of all images in the training area. **This allows us to set a single threshold** (set to 3.5) that generalizes well to **unseen landmarks and panoramas**.

Project Page & Demo:

https://velma.schumann.pub

Workshop on
Foundation Models for Decision Making

NEURAL INFORMATION PROCESSING SYSTEMS