NEURAL INFORMATION PROCESSING SYSTEMS

Mind's Eye of LLMs: Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models

Project Page

Wenshan Wu †, Shaoguang Mao†, Yadong Zhang†, ‡, Yan Xia†, Li Dong†, Lei Cui†, Furu Wei†
† Microsoft Research ‡ East China Normal University

38th Conference on Neural Information Processing Systems (NeurIPS 2024)

Microsoft Research

SCAN ME

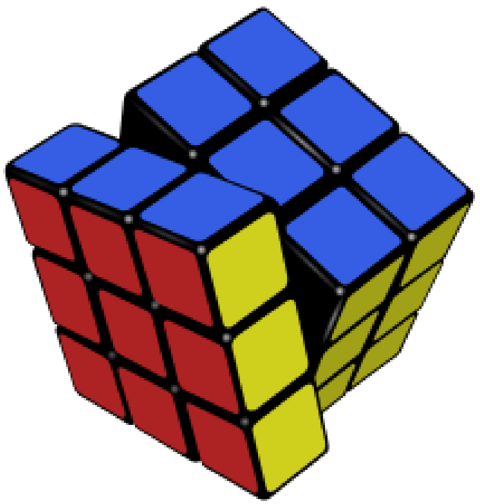
Introduction

Spatial Reasoning in Human Cognition

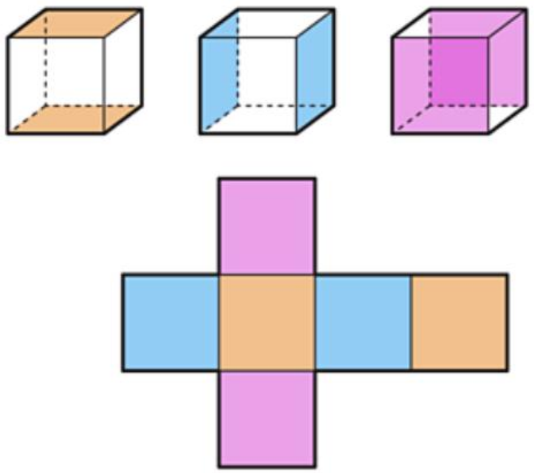
- Mental Image: abstract representations from visual perception
- Mind’s Eye: mental image manipulation

Navigation

Mental rotation



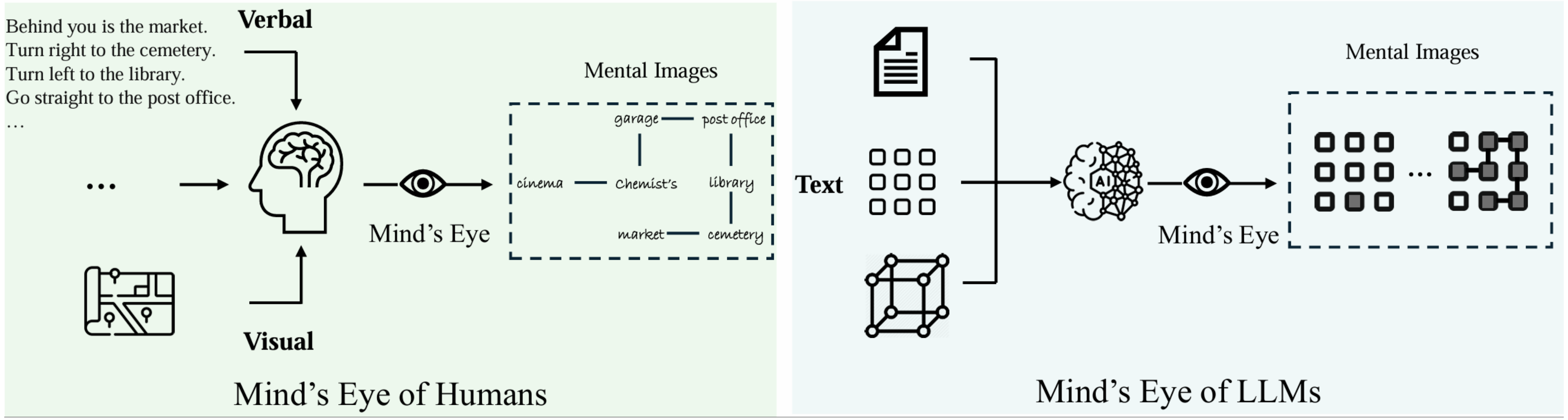
Mental paper folding



Motivation

Similar Mechanism in LLMs: Mind's Eye

- Visualize internal states
- Manipulate mental images to guide subsequent reasoning



Contribution

- We conduct quantitative and qualitative analyses on the mind’s eye of LLMs and its limitations. We also explore cues about the **origin of this generalized ability from code pre-training**.
- We develop two tasks of "visual navigation" and "visual tiling", along with corresponding synthetic datasets, emulating various sensory inputs for LLMs. These tasks are structured to support **varying levels of difficulty**.
- We propose **Visualization-of-Thought (VoT) prompting** to elicit the mind’s eye of LLMs for spatial reasoning and provide empirical evaluations on three tasks.

Spatial Reasoning Tasks

- Existing Benchmarks
  - Spatial semantics are embedded in text, spatial term focused
  - Could be solved by logic programming after converting spatial terms to logical forms through LLMs
- Ours
  - Focus on spatial awareness
    - Various aspects: spatial relationships, directions, and geometric shapes
    - Essential for action planning in the physical world.
  - Emulating various sensory inputs for LLMs
    - Natural language
    - 2D grid comprising of special text characters

Spatial Reasoning Tasks

- Natural Language Navigation [1]
  - Square map $W = \{(l_1, o_1), (l_2, o_2), \ldots, (l_n, o_n)\}$, each location associated with an object
  - Navigation instructions $I = \{i_1, i_2, \ldots, i_k\}$
  - Task: Find the object o at specific location l determined by navigation instructions

$$o \sim p(o \in W | W = \{(l_1, o_1), (l_2, o_2), \ldots, (l_n, o_n)\}, I)$$

- Visual Navigation
  - Grid map M consisting of k consecutive edges $E = \{e(s_0, s_1), e(s_1, s_2), \ldots, e(s_{k-1}, s_k)\}$
  - Route planning: generate a sequence of correct directions

$$D \sim p(\{d(s_0, s_1), d(s_1, s_2), \cdots, d(s_{k-1}, s_k)\} \mid M)$$

  - Next step prediction: given t navigation instructions, identify the direction of next step

$$d \sim p(d(s_t, s_{t+1}) \mid M, D_{t, 0<t<k})$$

- Visual Tiling
  - Rectangle R masked with k unique polyominoes $MP = \{mp_1, mp_2, \ldots, mp_k\}$
  - Two variants of each polyomino $v_{i<k} = \{v_{i1}, v_{i2}\}$, a polyomino query $q \in MP$
  - Task: identify the correct variant of q

$$v \sim p(v_q \mid R, \{mp_1, \cdots, mp_k\}, \{v_{11}, v_{12} \cdots, v_{k1}, v_{k2}\}, q)$$

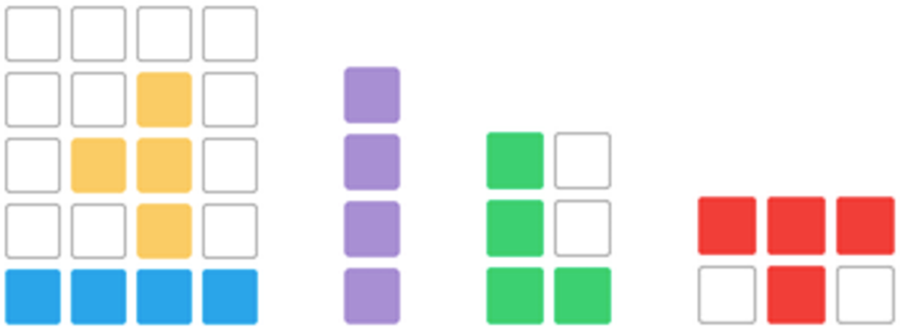
[1] Yamada, Yutaro, et al. "Evaluating spatial understanding of large language models." *arXiv preprint arXiv:2310.14540* (2023).

Spatial Reasoning Tasks

Various Levels of Difficulties

(a) k=2 (b) k=3 (c) k=4 (d) k=5 (e) k=6 (f) k=7

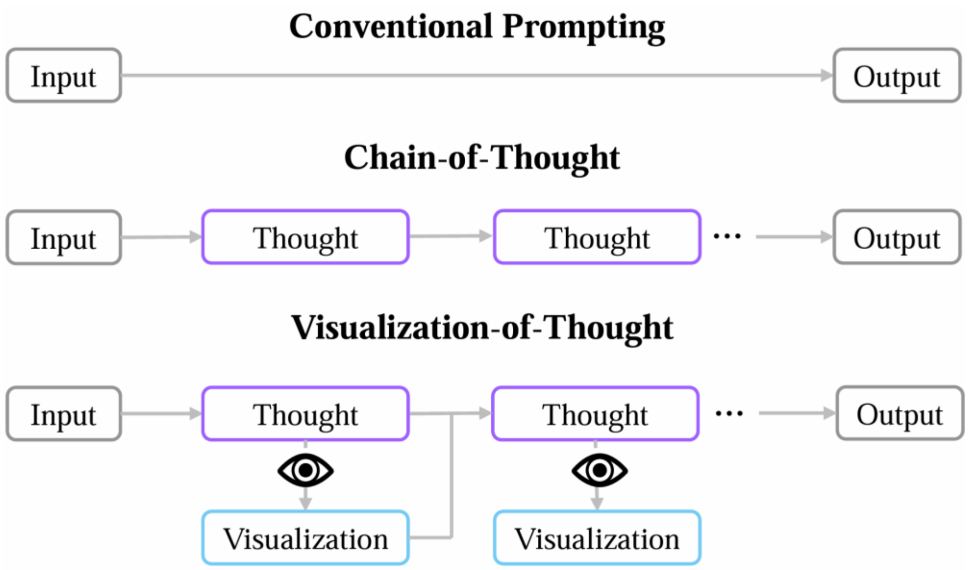
(a) Fit 2 pieces into a masked rectangle



(b) Fit 3 pieces into a masked rectangle

Visualization-of-Thought Prompting

Visualize the state after each reasoning Step

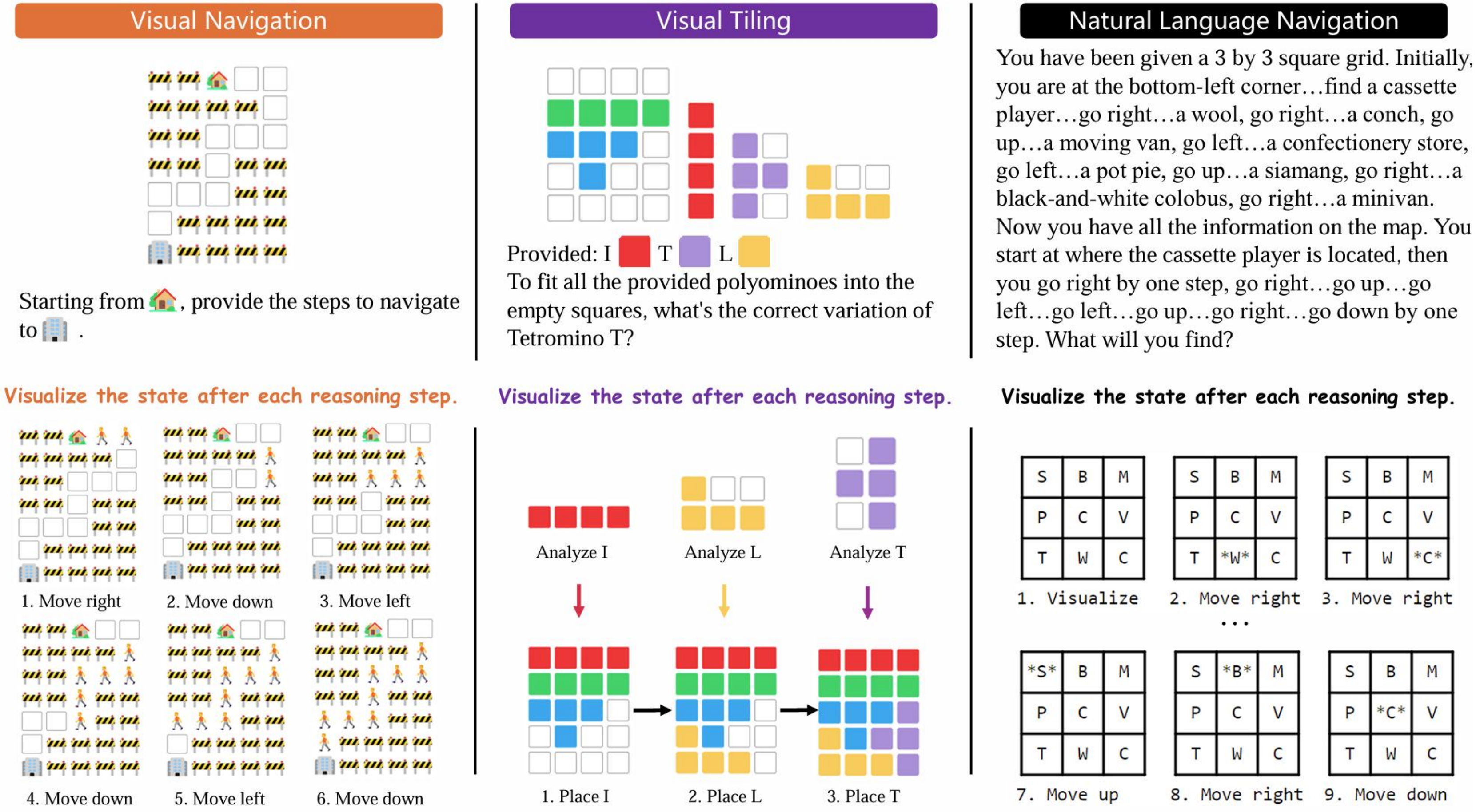


- $\boldsymbol{x}$: text sequence of input
- $\boldsymbol{v}$: visualization sequence in text form
- $\boldsymbol{z}$: language sequence of intermediate steps

$$v_i \sim p_\theta(v_i \mid prompt_{VoT}, x, z_{1 \dots i}, v_{1 \dots i-1})$$

$$z_{i+1} \sim p_\theta(z_{i+1} \mid prompt_{VoT}, x, z_{1 \dots i}, v_{1 \dots i})$$

Qualitative Results



Dataset

Natural Language Navigation

- 200 square maps of size 3x3

Visual Navigation

- 496 navigation maps and 2520 QA instances
- Map size up to 7×9 and 9×7

- ## Visual Tiling
  - 5 x 4 rectangle with 2 or 3 polyomino masked

Task	K Step						Total
	2	**3**	**4**	**5**	**6**	**7**	
Route Planning	8	16	32	64	128	248	496
Next Step Prediction	8	32	96	256	640	1488	2520

	Mask count		Total
	2	**3**	
Configuration	248	124	376
QA Instance	489	307	796

Experiments

- Settings
  - GPT-4 CoT: Let's think step by step.
  - GPT-4 w/o Viz: Don't use visualization. Let's think step by step.
  - GPT-4V CoT: Let's think step by step.
  - GPT-4 VoT: Visualize the state after each reasoning step.

Settings	Visual Navigation			Visual Tiling	Natural-Language Navigation
	Route Planning		Next Step Prediction		
	Completing Rate	Succ Rate			
GPT-4 CoT	37.02	9.48	48.61	54.15	54.00
GPT-4 w/o Viz	37.17	10.28	48.49	46.98	35.50
GPT-4V CoT	33.36	5.65	46.59	49.62	/
GPT-4 VoT	**40.77**	**14.72**	**55.28**	**63.94**	**59.00**

NEURAL INFORMATION PROCESSING SYSTEMS

Analysis

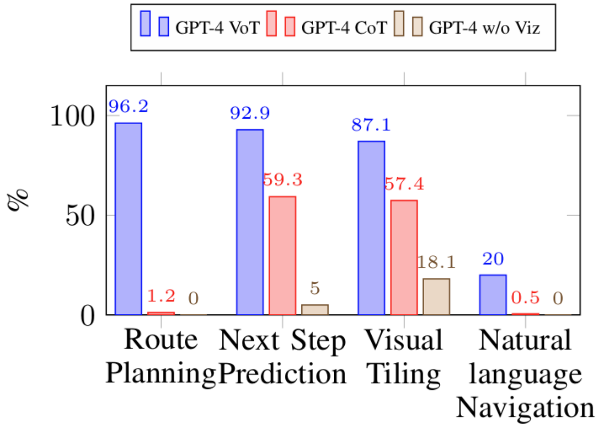
Do visual state tracking behaviors differ among prompting methods?

l_v : length of visualization sequence, l_s : number of reasoning steps

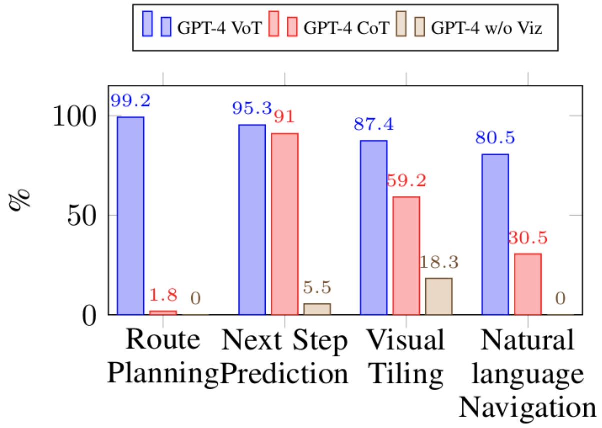
$$Complete\ tracking\ rate = \sum_i^n (l_v == l_s)/n$$

$$Patial\ tracking\ rate = \sum_i^n (l_v < 0)/n$$

- VoT markedly improves the visual tracking rate
- LLMs inherently exhibit the capability of visual state tracking in some tasks.



(a) Complete tracking rate



(b) Partial tracking rate

Analysis

How visualizations enhance final answers?

- Visualization Quality
  - Compliance: visualization satisfies requirements in 51-52% cases
  - Accuracy: visualization aligns with the corresponding state in 24%-26% cases
- Performance enhancement
  - LLMs are able to make correct decisions in 65%-77% of the cases when accurate internal state visualizations are generated

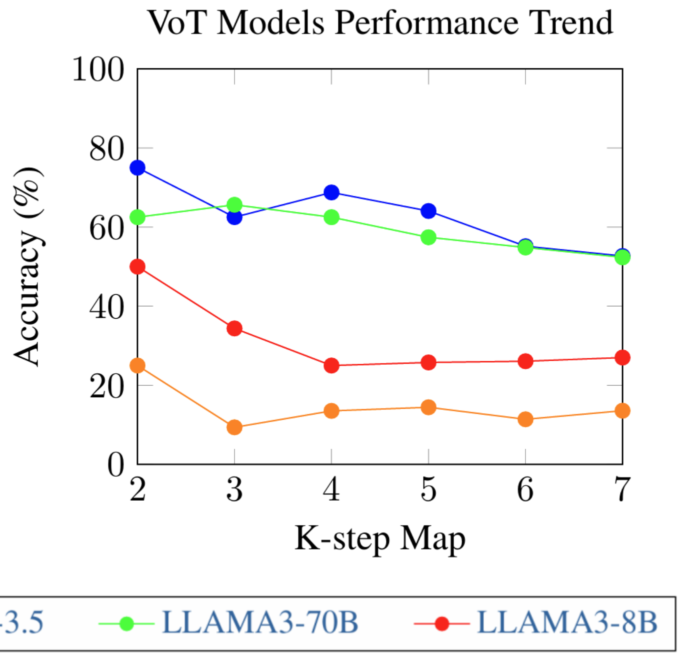
Task	Spatial Visualization		Spatial Understanding
	Compliance	Accuracy	Accuracy
Visual Navigation	51.14	**26.48**	65.16
Visual Tilling	**52.01**	24.25	**77.20**

Analysis

Can VoT benefit less powerful language models?

- VoT offers a scaling advantage when applied to more advanced models
- Less capable models tend to rely on random guessing

Settings	Visual Navigation			Visual Tiling	Natural-Language Navigation
	Route Planning		Next Step Prediction		
	Completing Rate	Succ Rate			
GPT-3.5 CoT	16.10	**2.62**	**17.42**	44.10	8.50
GPT-3.5 VoT	**19.02**	1.61	13.10	**47.99**	**9.00**
LLAMA3-8B CoT	4.65	0	**28.73**	**47.24**	**16.50**
LLAMA3-8B VoT	**4.97**	**0.2**	26.75	46.73	15.50
LLAMA3-70B CoT	19.90	2.62	49.01	**56.41**	26.00
LLAMA3-70B VoT	**30.24**	**5.85**	**54.09**	56.03	**32.50**

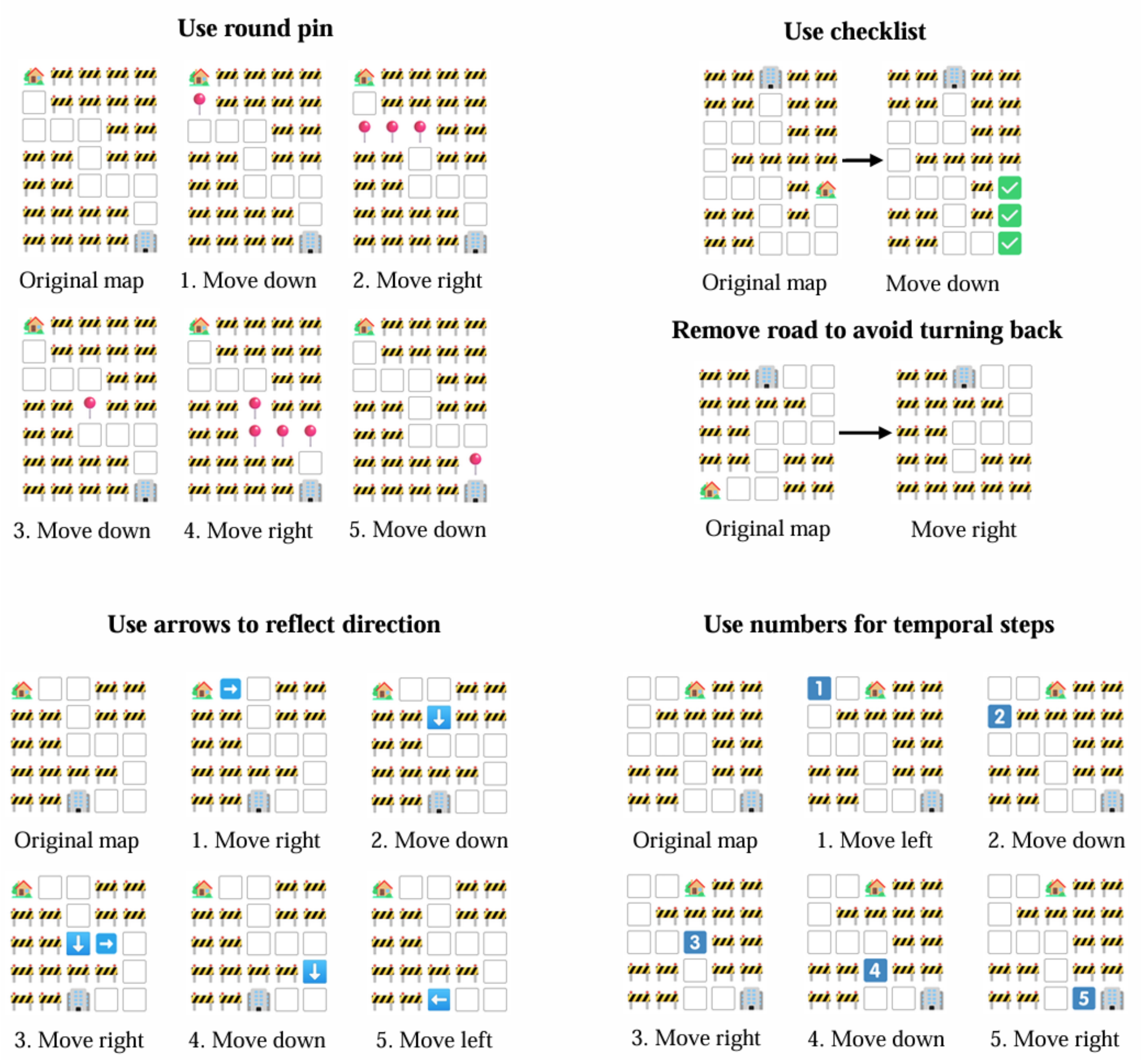


NEURAL INFORMATION PROCESSING SYSTEMS

Appendix

Mental Images for State Tracking

- Mark the path with unique symbols
- Mark path and direction with arrows
- Mark path with temporal steps
- Remove road: turning roads into obstacles



Appendix

Ascii-art in Code Comments

- Represents data structure, diagram, geometry
- Illustrates how an algorithm works or simulates an operation
  - Spatial Causality: Double-ended queue in Rust, Scrolling web pages, tree rotation present triplets of previous visual state, instruction, and updated state of instruction following.
  - Temporal Causality: Undo systems from emacs provides various temporal states of the undo system when undo operation happens in different timelines and corresponding visualizations in an interleaved manner. Each visualization reflects the temporal casuality of the system state.