Is A Picture Worth A Thousand Words? Delving Into Spatial Reasoning For Vision Language Models

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

Microsoft Research

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon (Yixuan) Li, Neel Joshi

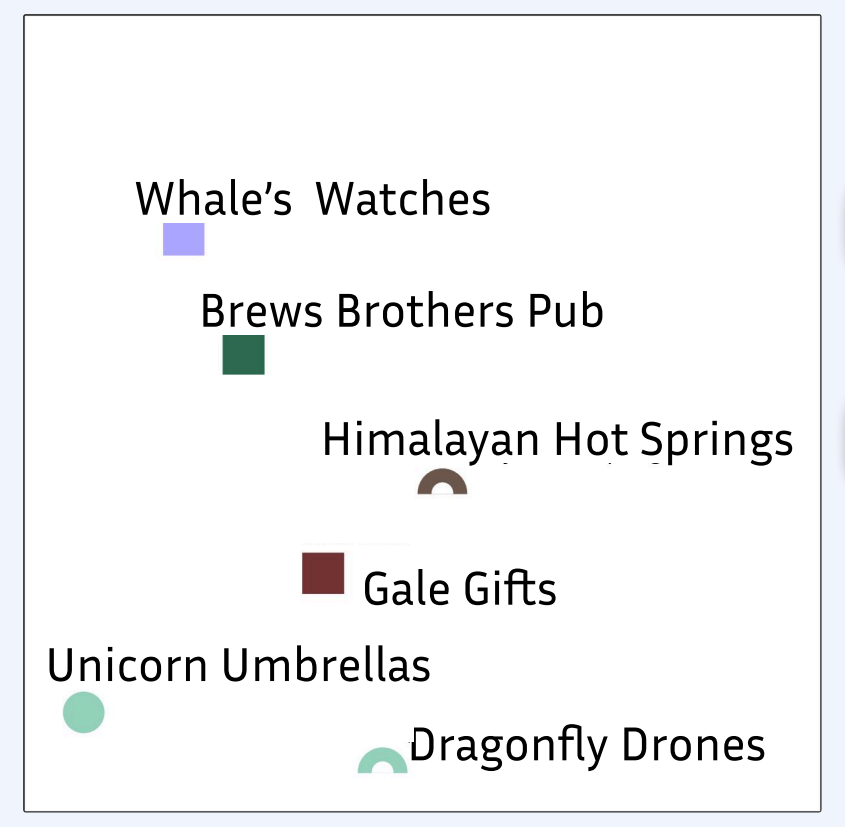
NEURAL INFORMATION PROCESSING SYSTEMS

Paper Code

SpatialEval: a new benchmark for LLMs and VLMs

- Motivation: Spatial reasoning for LLMs and VLMs are under-explored
- Scope: Spatial understanding and reasoning

Spatial-Map



TQA (Text-only): Consider a map with multiple objects:
Whale's Watches is in the map. Brews Brothers Pub is to the Southeast of Whale's Watches. Himalayan Hot Springs is to the Southeast of Whale's Watches... Gale Gifts is to the Northeast of Unicorn Umbrellas. Gale Gifts is to the Southwest of Himalayan Hot Springs.

VQA (Vision-only): <img> The figure represents a map with multiple objects. Each object is associated with a name as shown in the figure.

VTQA (Vision-text): <img> The figure represents a map with multiple objects. Each object is associated with a name as shown in the figure. The same figure can be described as follows:
Whale's Watches is in the map. Brews Brothers Pub is to the Southeast of Whale's Watches. Himalayan Hot Springs is to the Southeast of Whale's Watches.... Gale Gifts is to the Southwest of Himalayan Hot Springs.

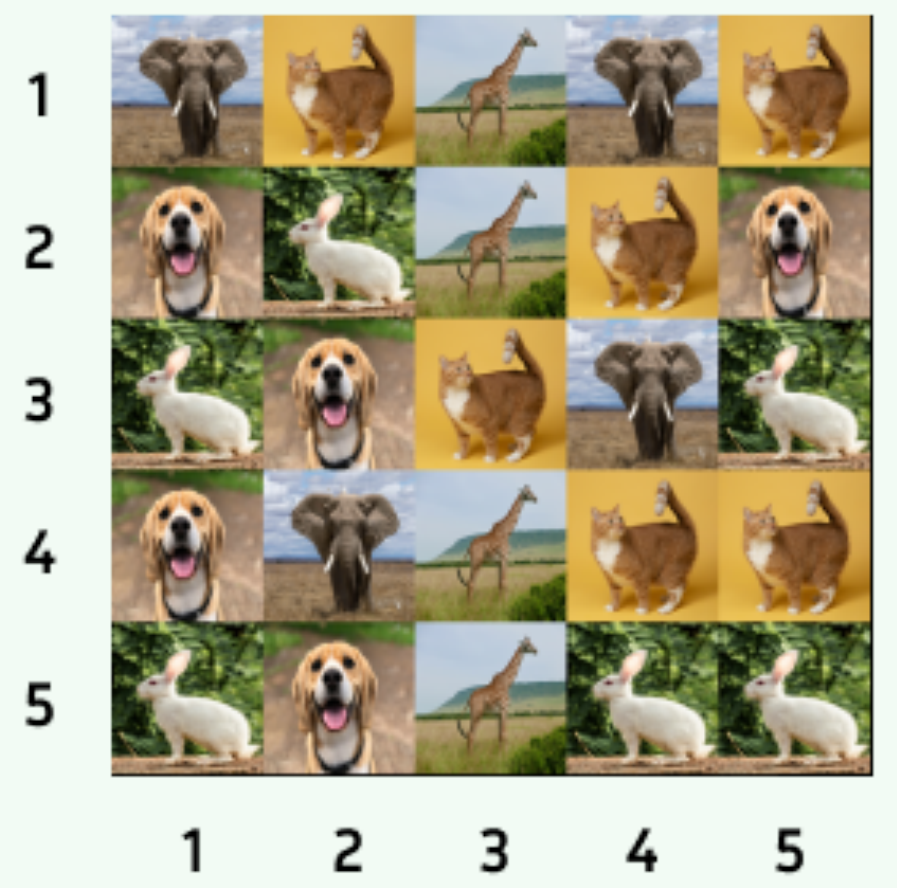
Questions

Q: In which direction is Whale's Watches relative to Dragonfly Drones?
A. Northwest B. Southwest C. Southeast D. Northeast

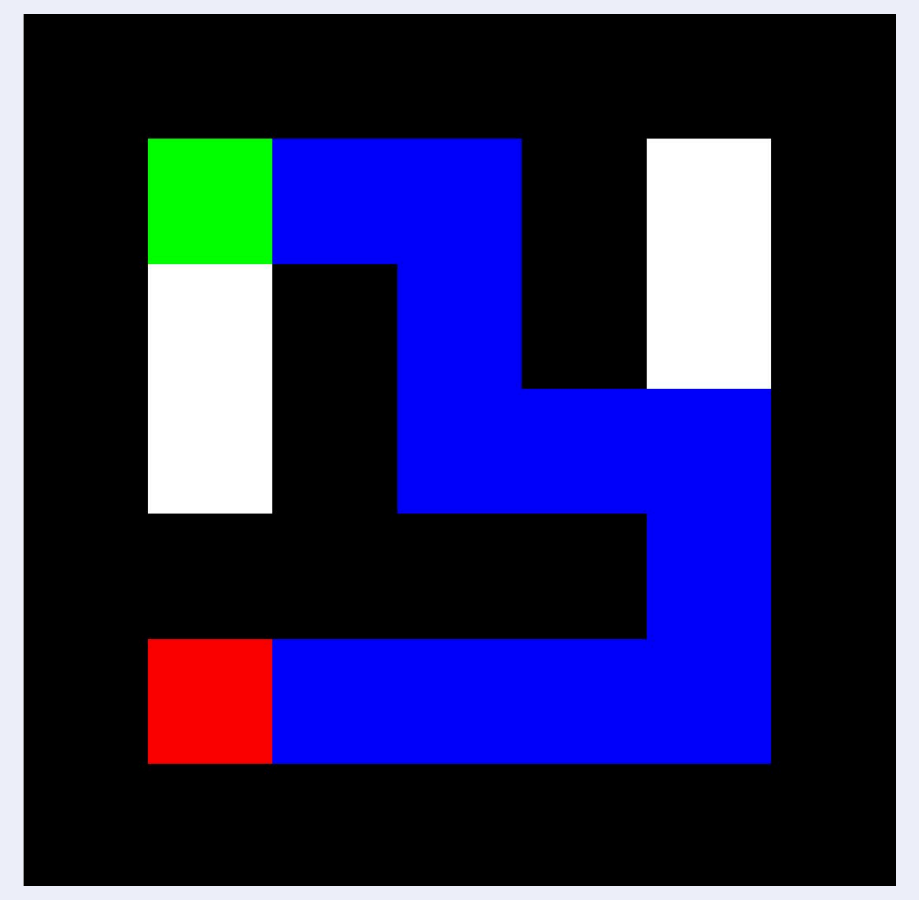
Q: Which object is in the Southwest of Gale Gifts?
A. Dragonfly Drones **B. Unicorn Umbrellas** C. Himalayan Hot Springs D. Brews Brothers Pub

Q: How many objects are in the Southwest of Himalayan Hot Springs?
A. 3 B. 4 C. 1 D. 0

Spatial-Grid



Maze-Nav



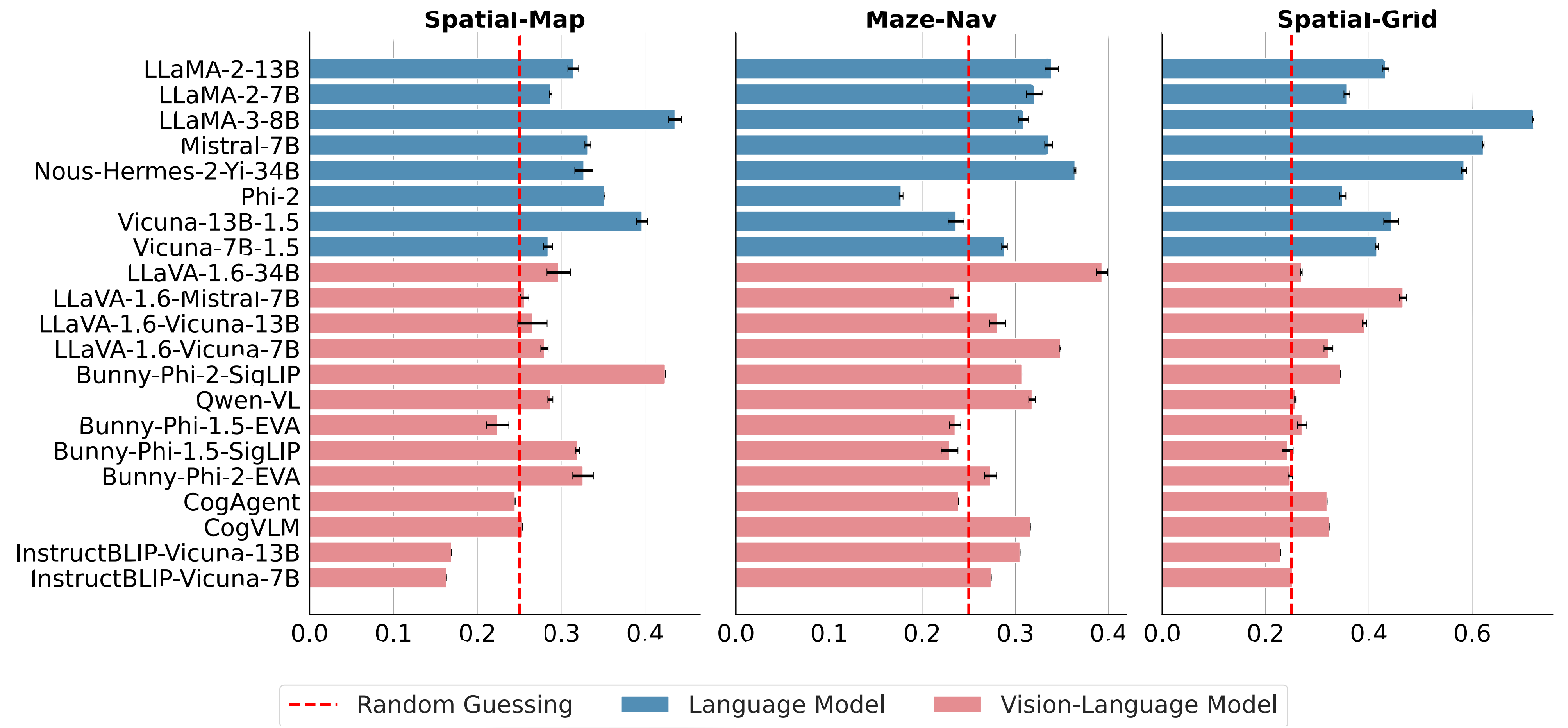
Spatial-Real

- Spatial Relationship
- Position Understanding
- Object Counting
- Navigation

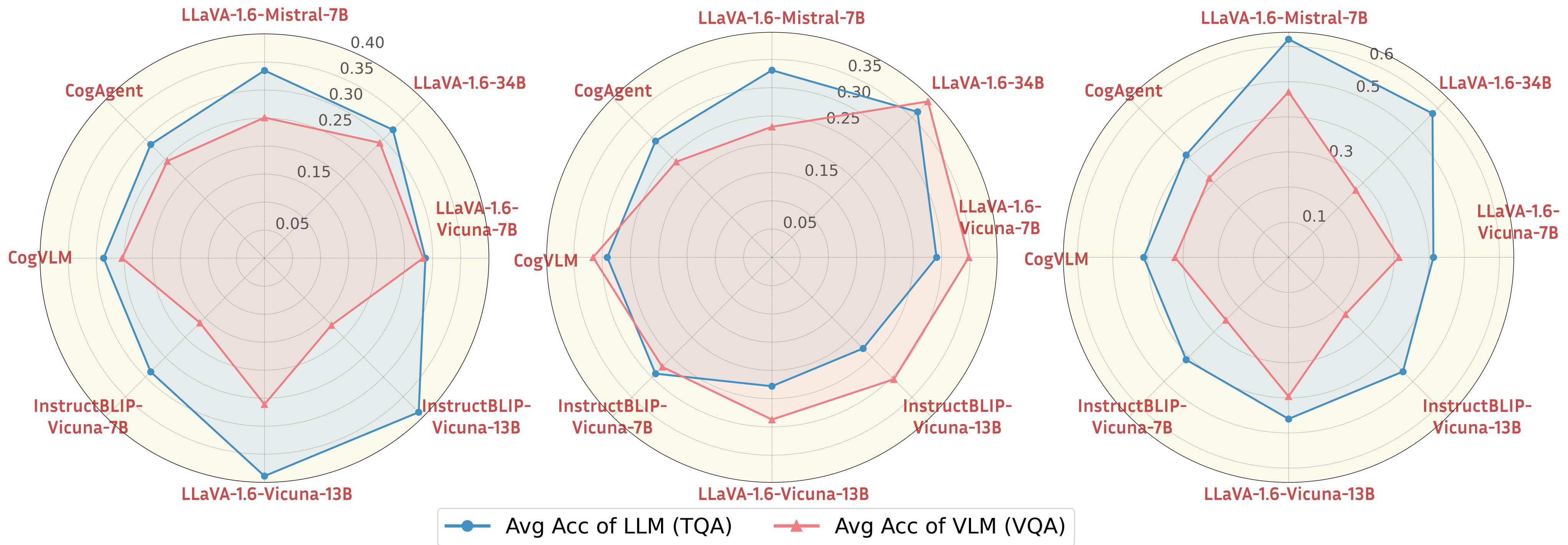
Model	Input Modality	Term	Description
LLM	Text-only	TQA (LLM)	Text-only input that includes all necessary information to answer questions without visual context.
VLM	Text-only	TQA (VLM)	Text-only input as in TQA (LLM) but applied to VLMs (*e.g.*, the LLaVA family).
VLM	Vision-only	VQA	Input only includes an image without corresponding textual description.
VLM	Vision-text	VTQA	Input includes both an image and its textual description.

Main Results

Only a few models outperform random guessing for spatial reasoning tasks

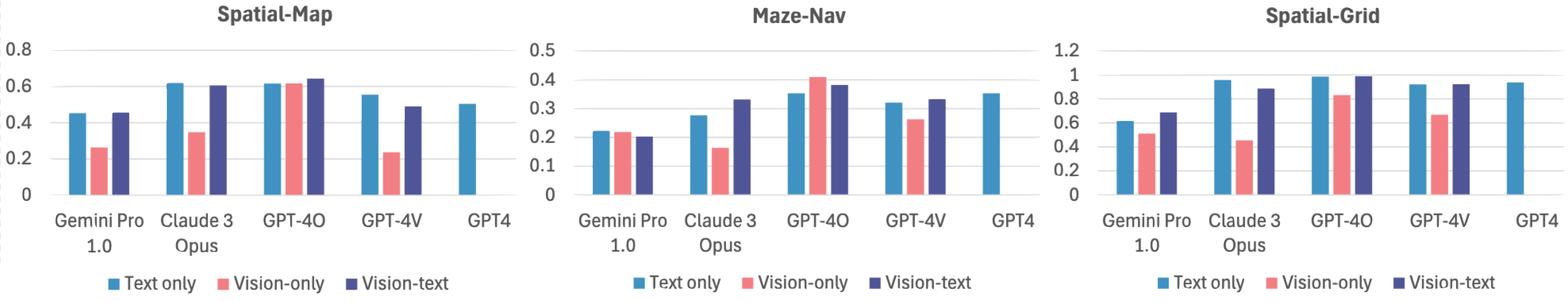


Vision information does not help with VQA? TQA (LLM) > VQA

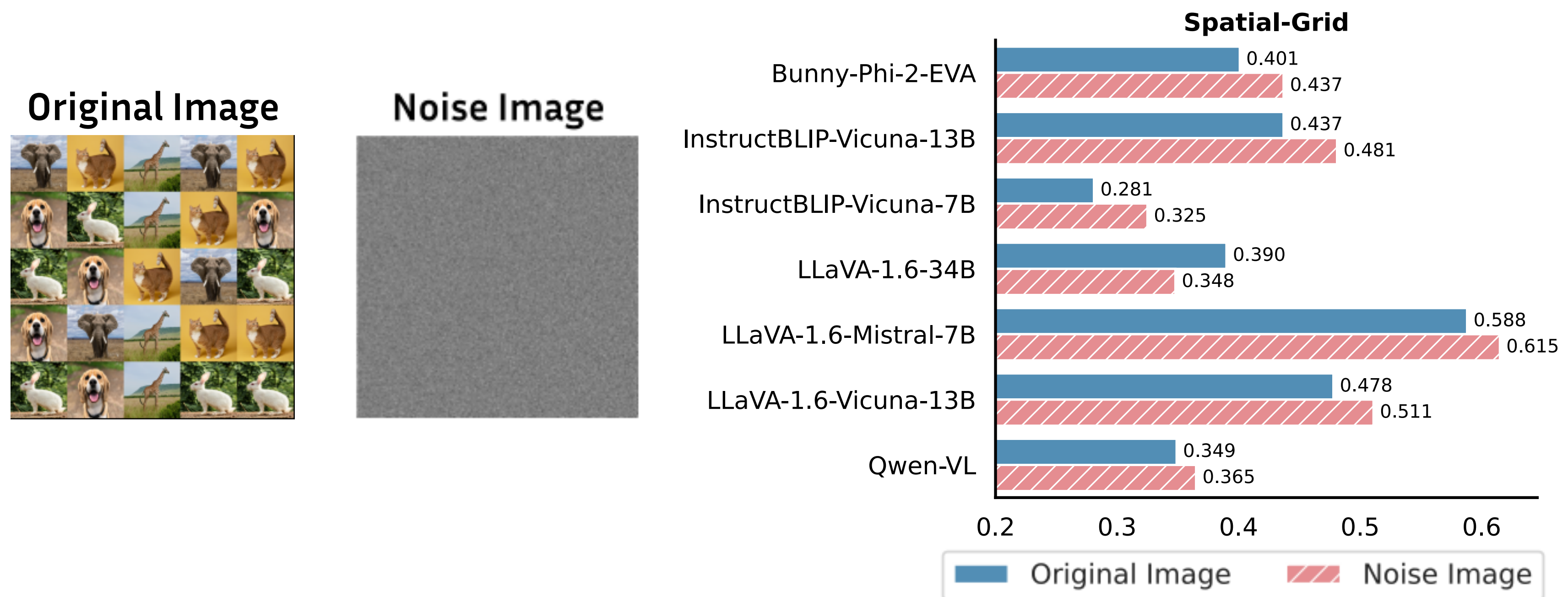


(a) TQA vs. VQA on Spatial-Map (b) TQA vs. VQA on Maze-Nav (c) TQA vs. VQA on Spatial-Grid

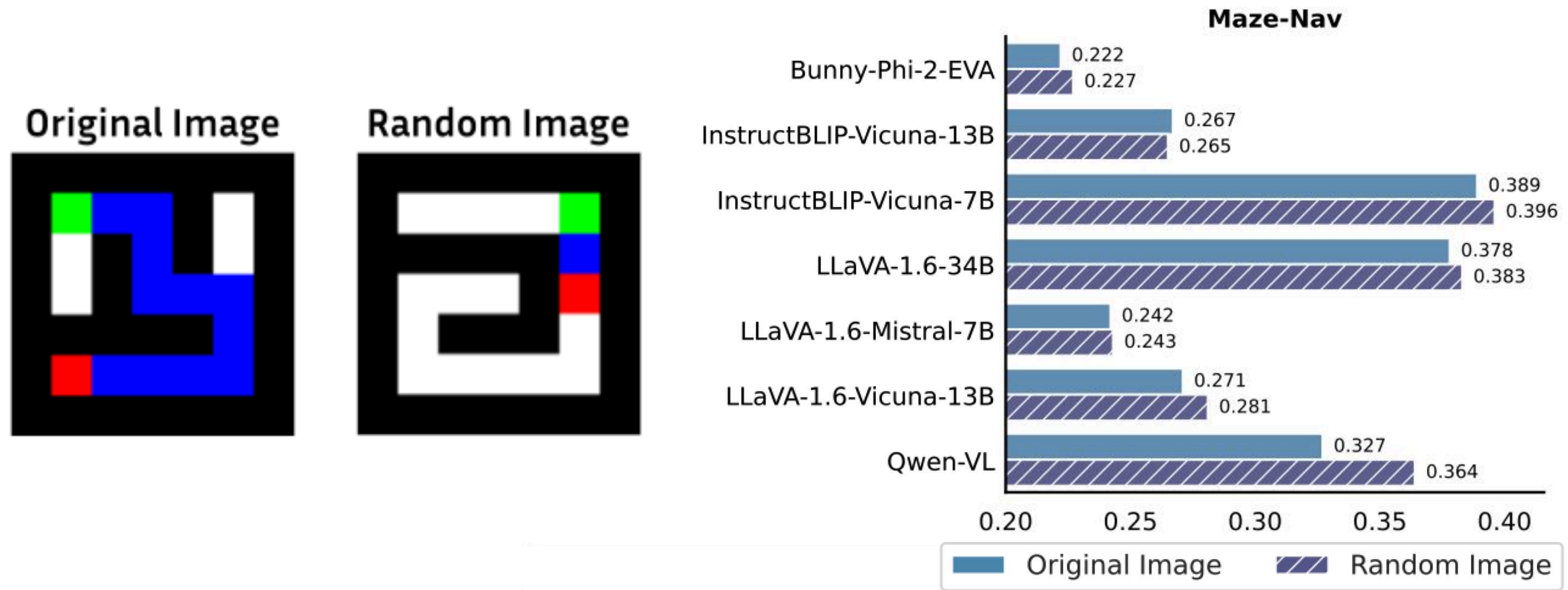
Similar Trends Hold for Proprietary Models as Open-source Models



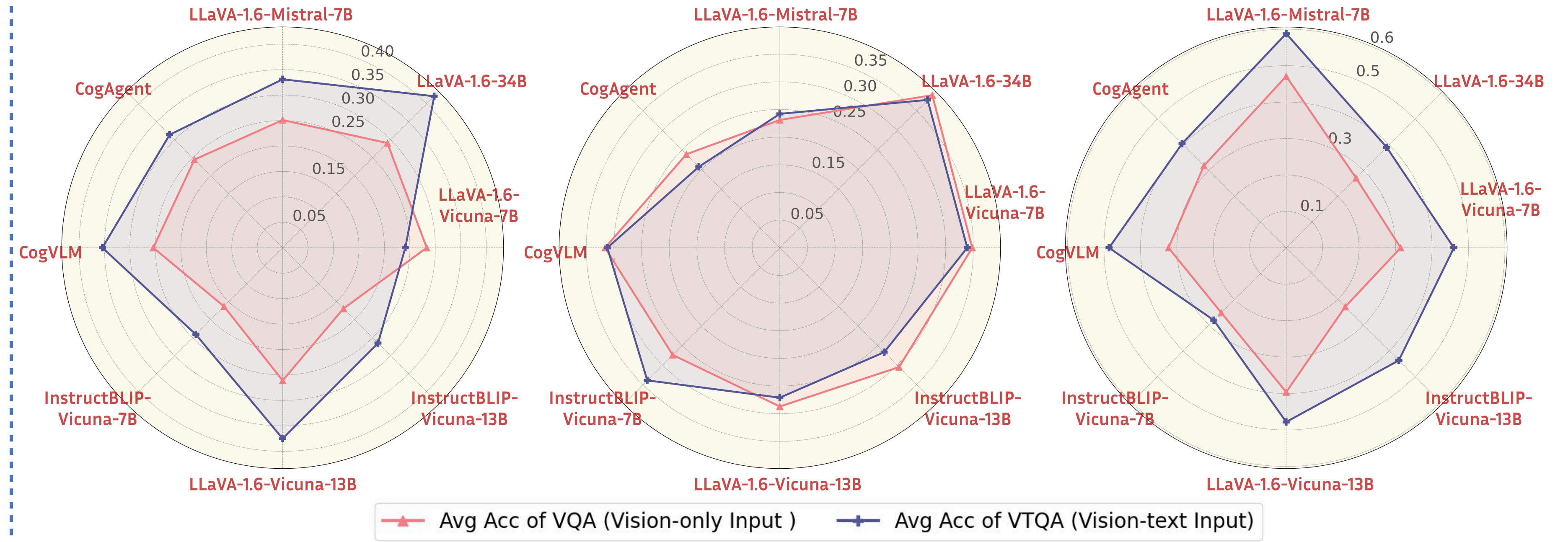
Noise Image can help VQA: Original Image vs Noise Image in VTQA



Mismatched multimodal information does not necessarily hurt



Leveraging redundancy in multimodal inputs can improve VLM performance



(a) VQA vs. VTQA on Spatial-Map (b) VQA vs. VTQA on Maze-Nav (c) VQA vs. VTQA on Spatial-Grid

Comparison	Summary of Findings
TQA (LLM) vs. VQA	VQA rarely enhances the performance compared to TQA (LLM).
VTQA vs. TQA (VLM)	VLMs exhibit improved performance in spatial reasoning tasks when the image input is absent.
VQA vs. VTQA	Given the same image input, additional textual description enhances VLM's performance.
TQA (VLM) vs. TQA (LLM)	Multimodal fine-tuning enhances LLM's spatial reasoning ability.
TQA (LLM) vs. VTQA	No definitive winner.