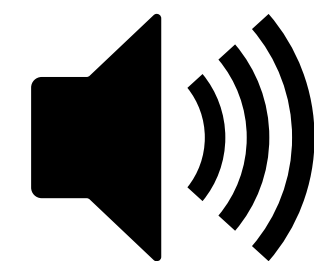


How Well Does GPT-4o Understand Vision?

Evaluating Multimodal Foundation Models on Standard Computer Vision Tasks

Rahul Ramachandran, Ali Garjani, Roman Bachmann,
Andrei Atanov*, Oğuzhan Fatih Kar*, Amir Zamir*

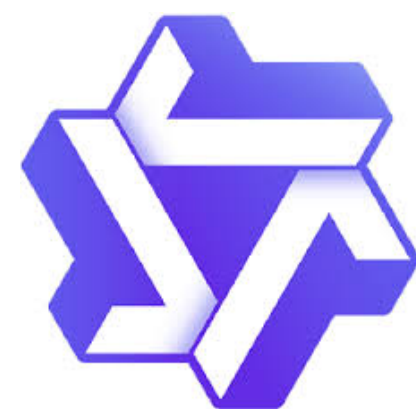
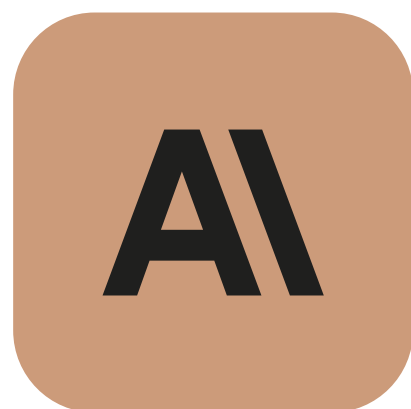
*Equal technical advising.



Please turn on sound for narration

MFMs Are Everywhere!

- MFMs have demonstrated an impressive range of capabilities across various domains



Write a haiku for this image



Dancing robots whirl,
Colors burst, music unfolds,
Joy knows no bounds here.

But How Do They Perform on Vision Tasks?

What city is this? If I want to tour around the landmark in this picture, where should I go?



New York City

This is New York City, and the prominent landmark in the image is the Washington Square Arch, which is located in the Greenwich Village neighborhood. To tour around this landmark, you should visit the nearby Washington Square Park, which is a popular spot for locals and tourists alike. The park offers a variety of activities, including walking paths, a fountain, and a statue of George Washington. Additionally, the surrounding area is rich with historical buildings, art galleries, and restaurants, making it a vibrant and culturally diverse place to explore.

LMSYS Chatbot Arena Leaderboard

[Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#) | [Kaggle Competition](#)

This is a mirror of the live leaderboard created and maintained by the LMSYS Organization. Please link to [leaderboard.lmsys.org](#) for citation purposes.

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over 1,000,000 human pairwise comparisons to rank LLMs with the Bradley-Terry model and display the model ratings in Elo-scale. You can find more details in our paper. Chatbot arena is dependent on community participation, please contribute by casting your vote!

Arena

NEW: Arena (Vision)

Full Leaderboard

Total #models: 11. Total #votes: 29,500. Last updated: 2024-07-06.

Code to recreate leaderboard tables and plots in this [notebook](#). You can contribute your vote at [chat.lmsys.org](#)!

Category

Overall

Overall Questions

#models: 11 (100%) #votes: 29,500 (100%)

Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1225	+7/-6	6002	OpenAI	Proprietary	2023/10
2	Claude 3.5 Sonnet	1210	+7/-4	8536	Anthropic	Proprietary	2024/4
3	GPT-4-Turbo-2024-04-09	1168	+5/-7	5538	OpenAI	Proprietary	2023/12
3	Gemini-1.5-Pro-API-0514	1162	+6/-6	5956	Google	Proprietary	2023/11
5	Gemini-1.5-Flash-API-0514	1083	+6/-6	5952	Google	Proprietary	2023/11
5	Claude 3 Opus	1082	+7/-7	6063	Anthropic	Proprietary	2023/8
7	Claude 3 Sonnet	1050	+6/-7	6112	Anthropic	Proprietary	2023/8

- How good are these models at standard vision benchmarks (like COCO)?
- How do these models stack up against specialized vision models?

Challenges



Provide a segmentation mask for this image



Provide bounding box coordinates for the person



... 

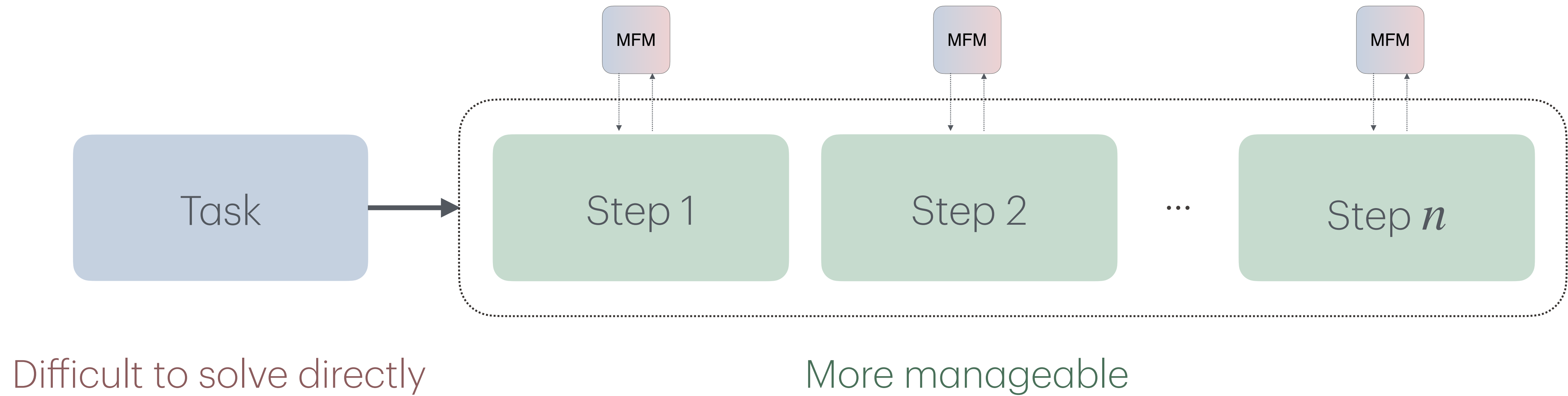


[0.3, 0.2, 0.5, 0.5]

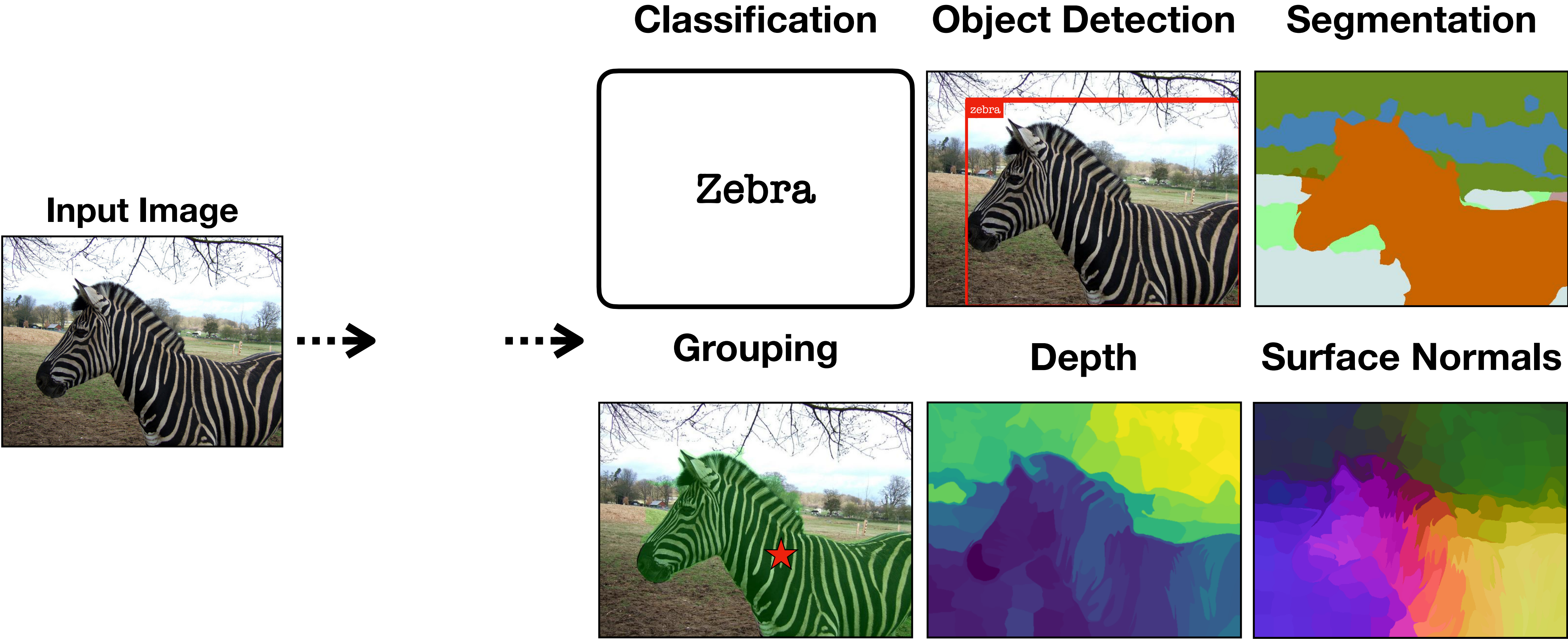


- 1. Cannot provide dense outputs
- 2. Subpar when directly prompted

Solution

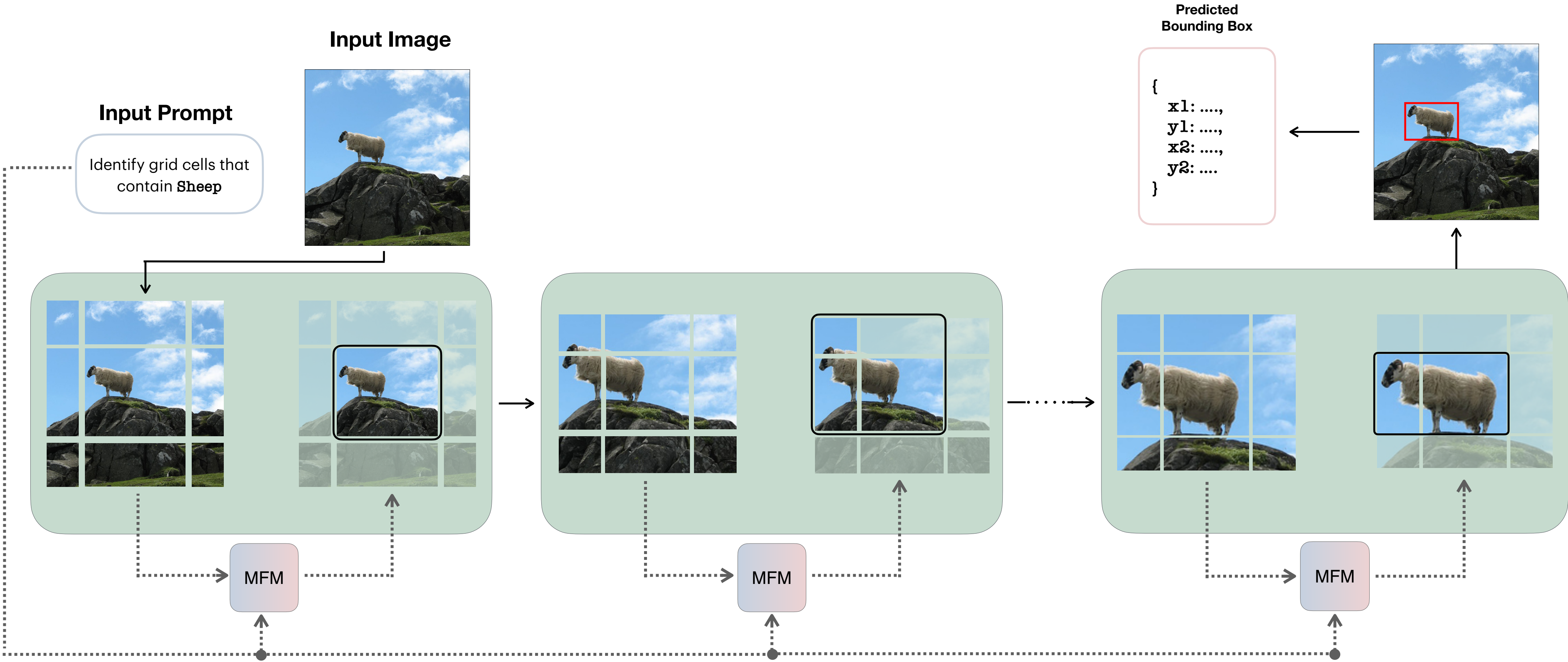


Solution

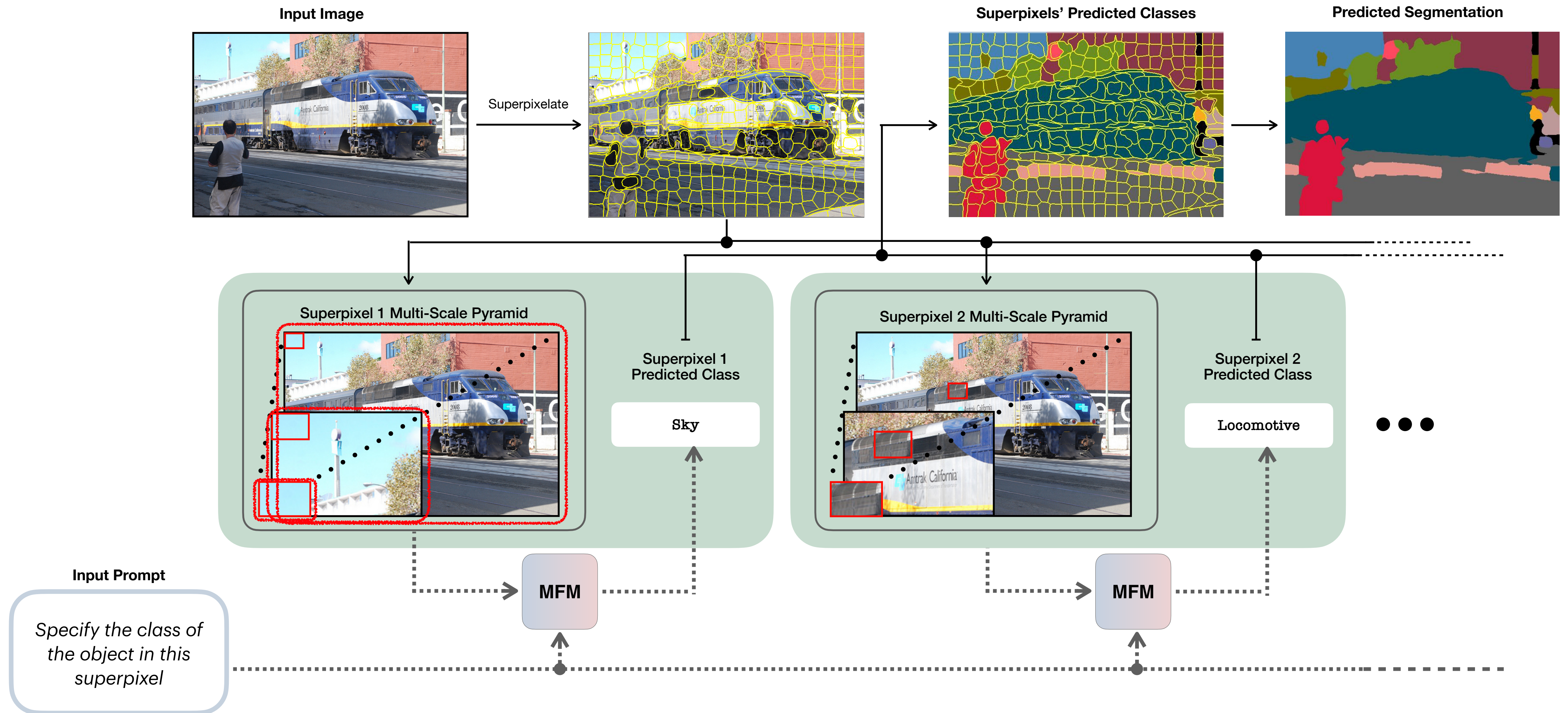


Predictions obtained using GPT-4o

Object Detection



Semantic Segmentation



Grouping



Grouping

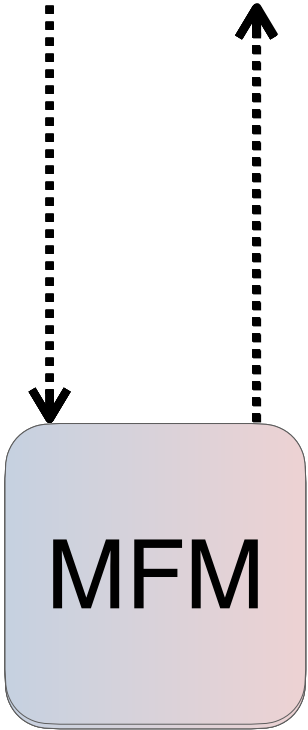
Input Image



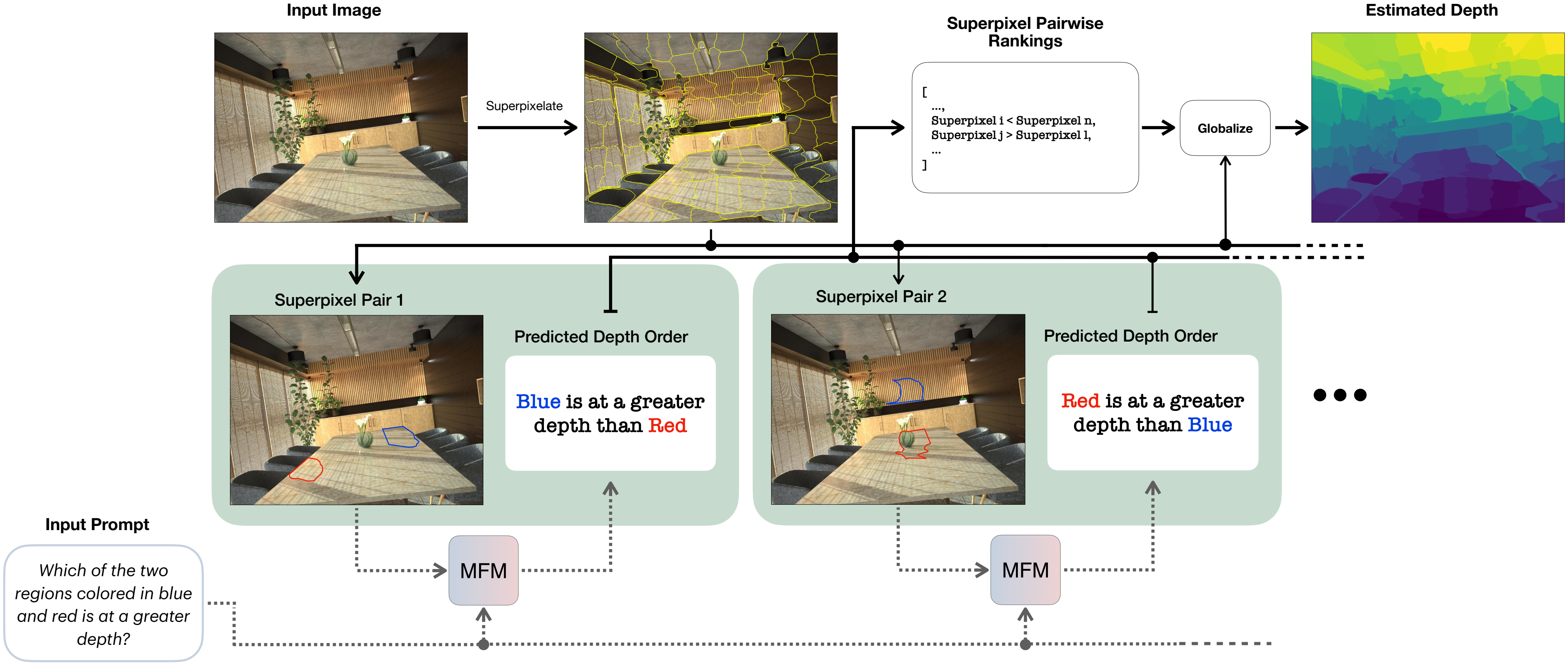
Superpixelate
→



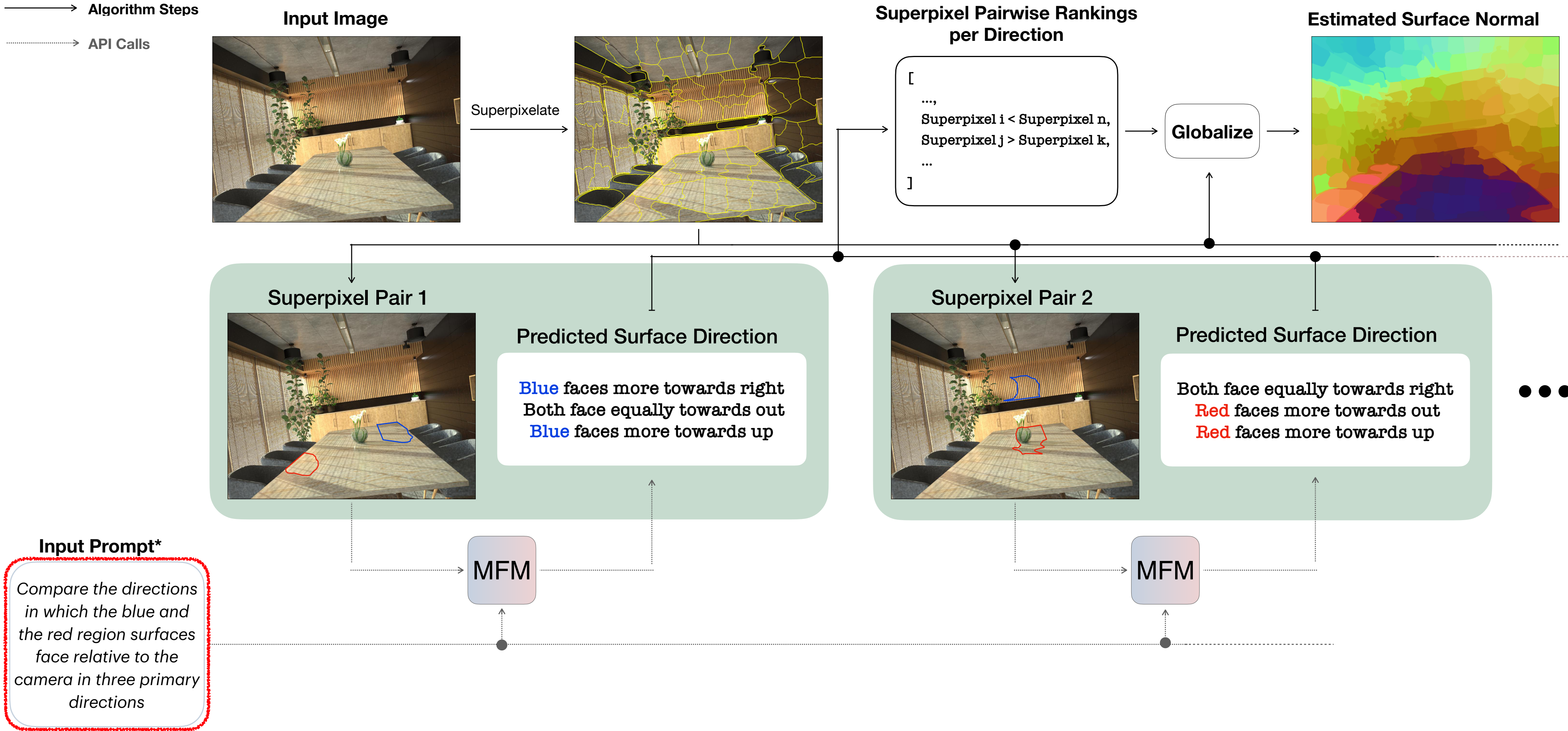
Predicted Grouping



Depth Prediction



Surface Normal Prediction



Experimental Results: Overview

Tasks

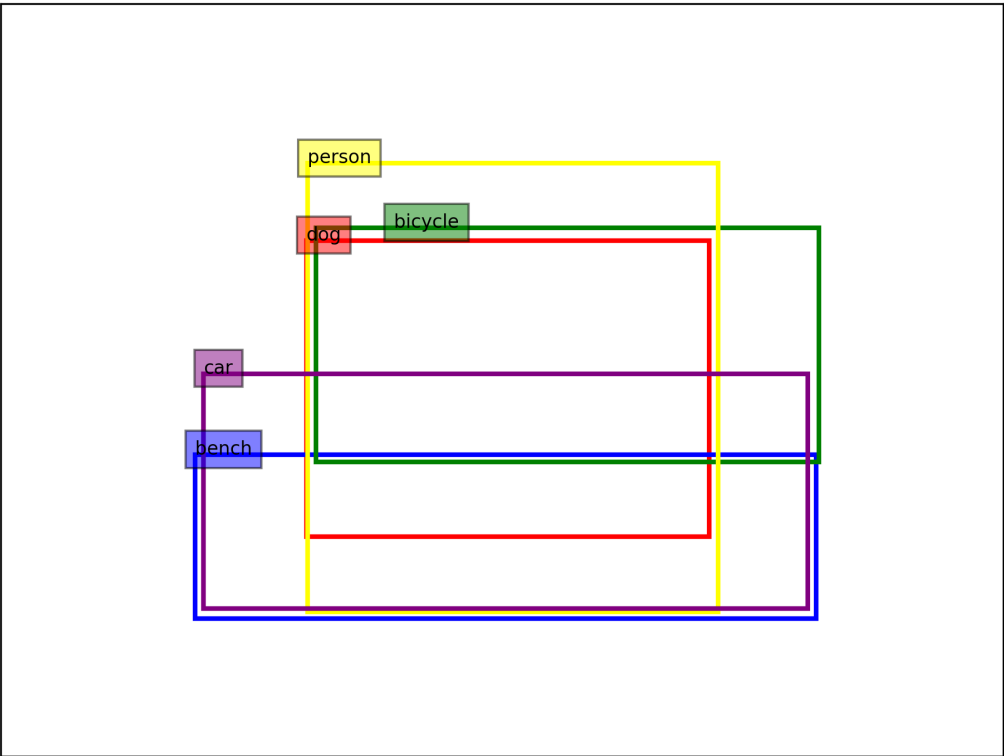
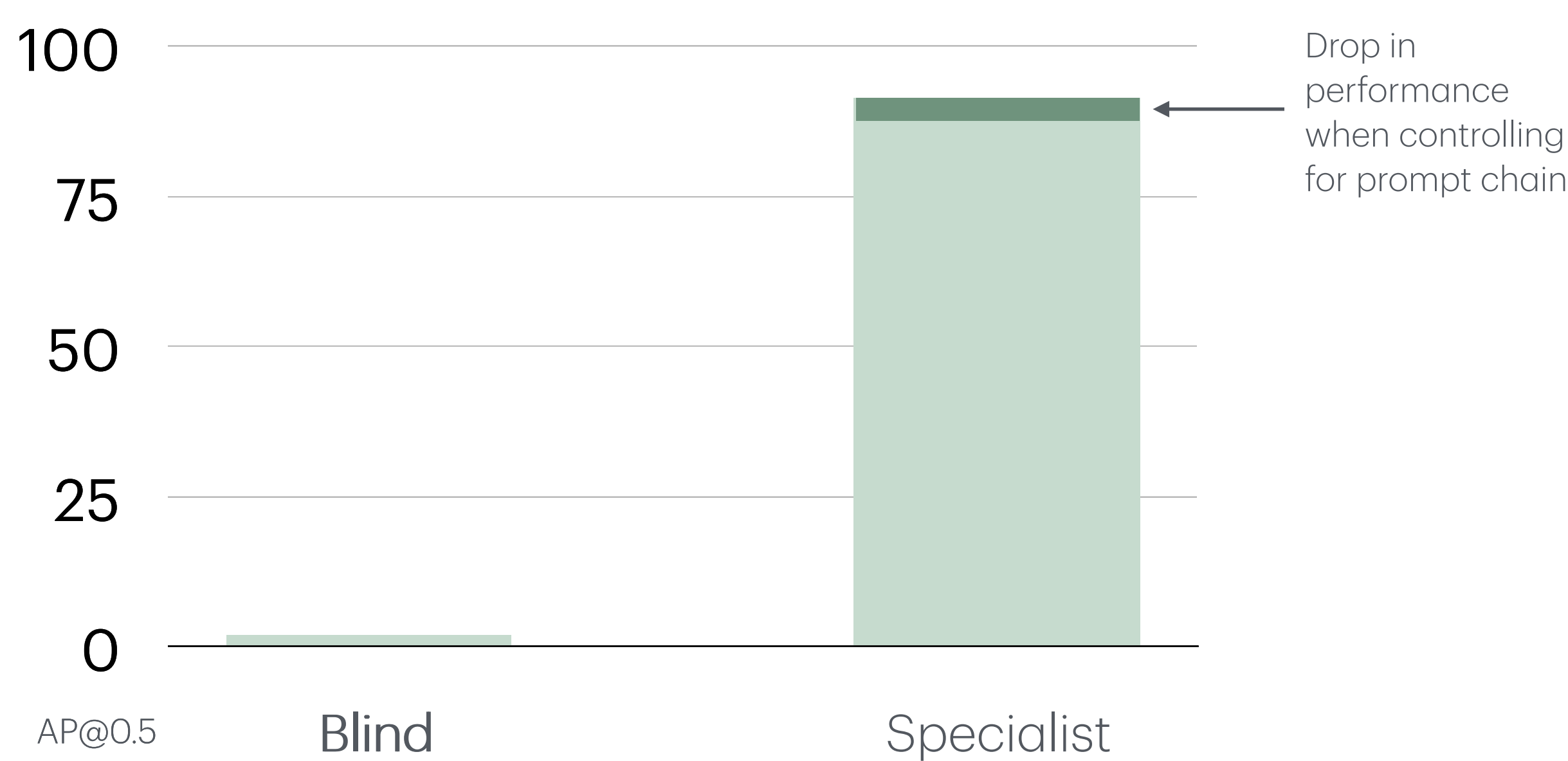
- Classification, Object Detection, Semantic Segmentation, Grouping, Depth Prediction, Surface Normal Prediction

Datasets

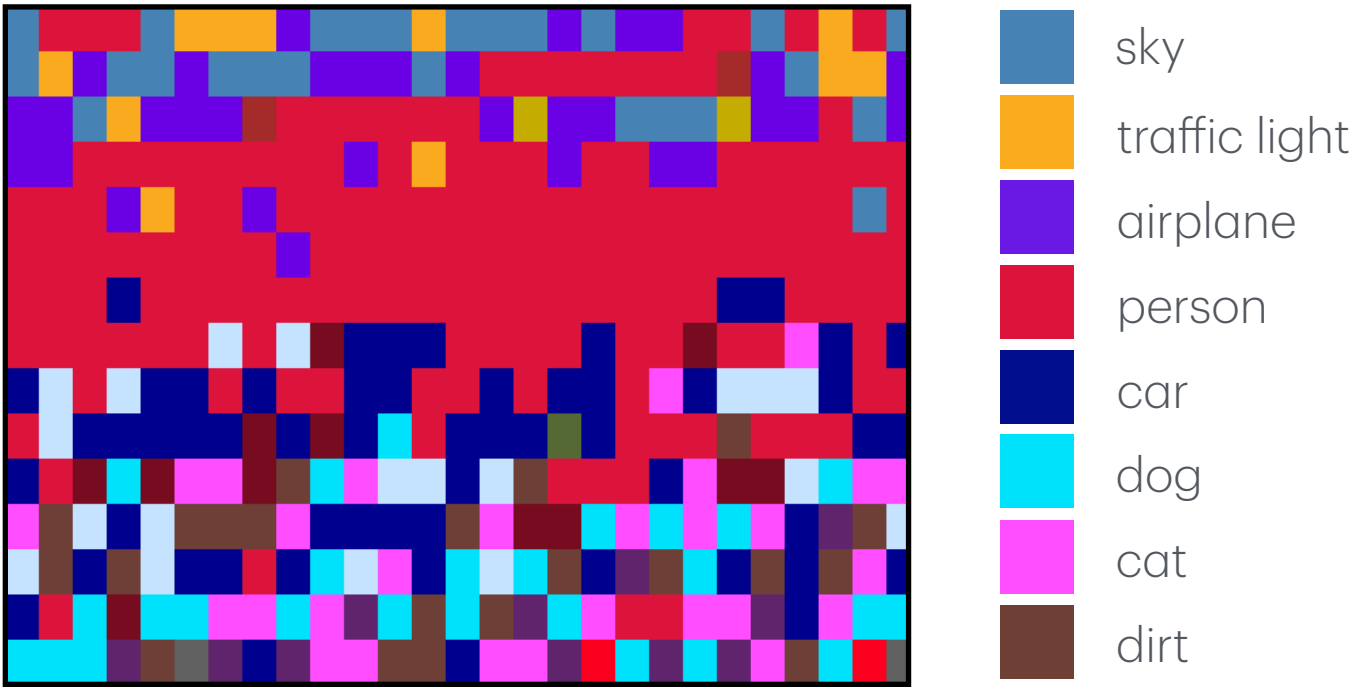
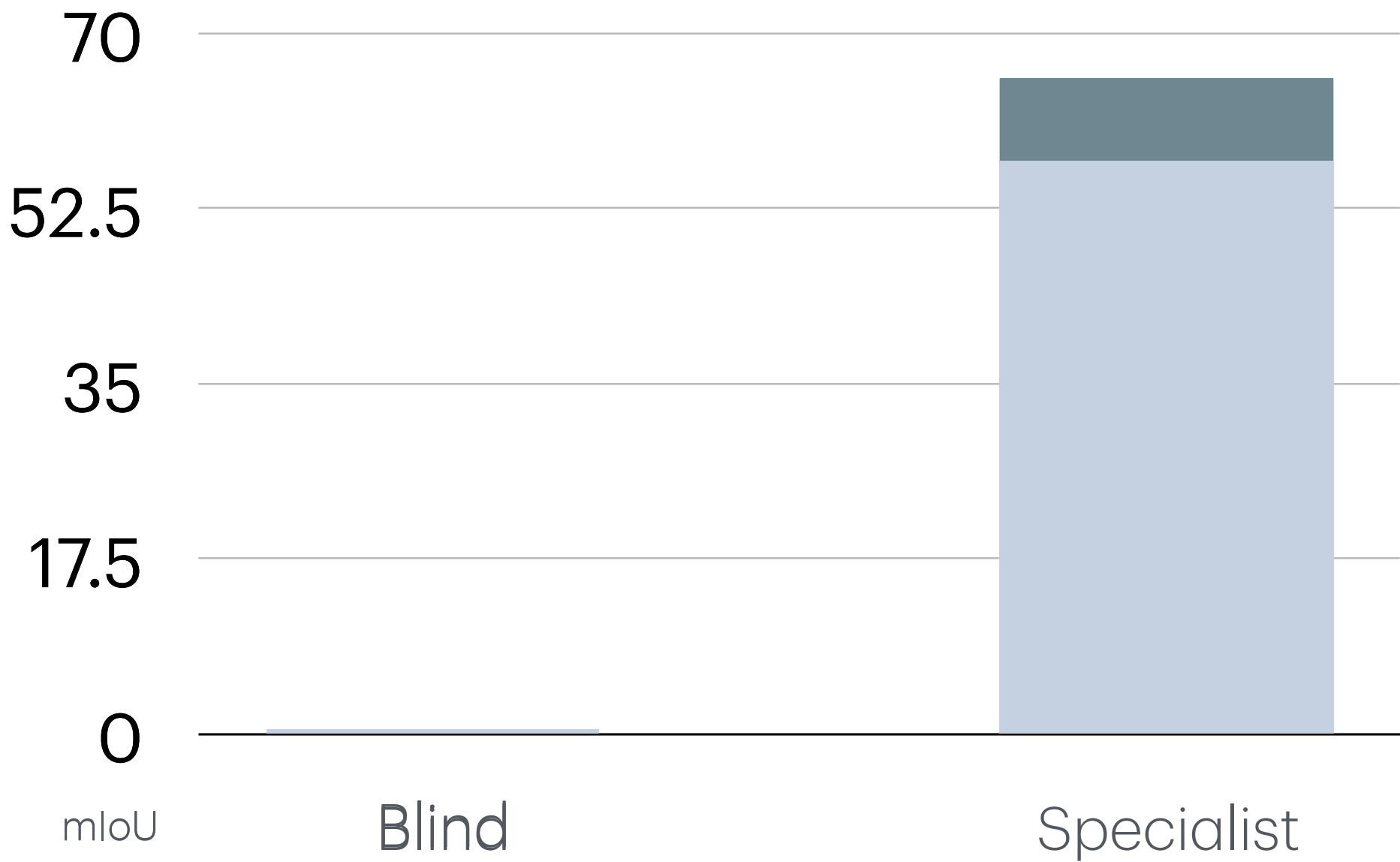
- Classification: **ImageNet** (and variants)
- Object Detection and Segmentation: **COCO**
- Depth and surface normal prediction: **Hypersim**

Experimental Results: Baselines

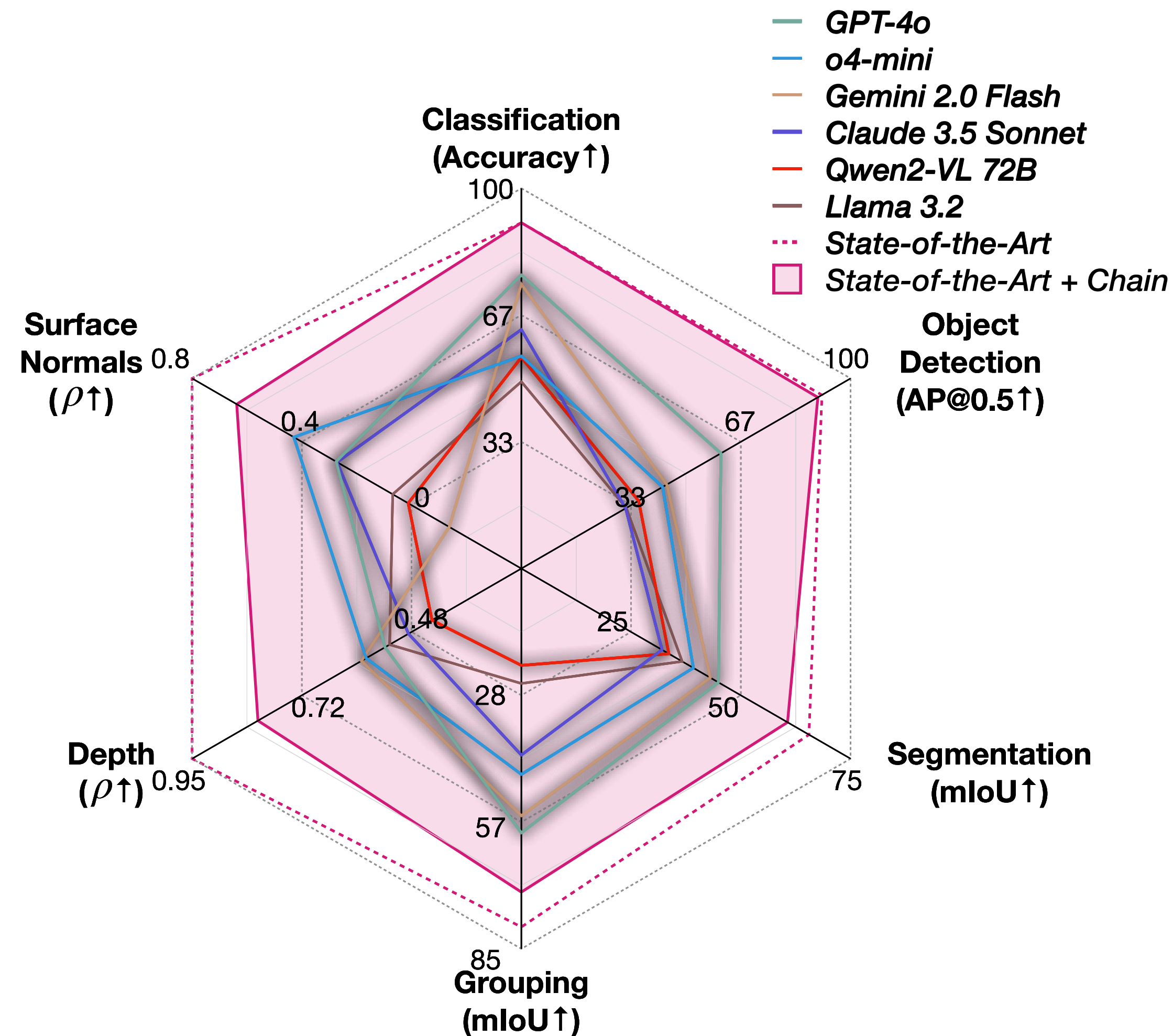
Object Detection



Segmentation



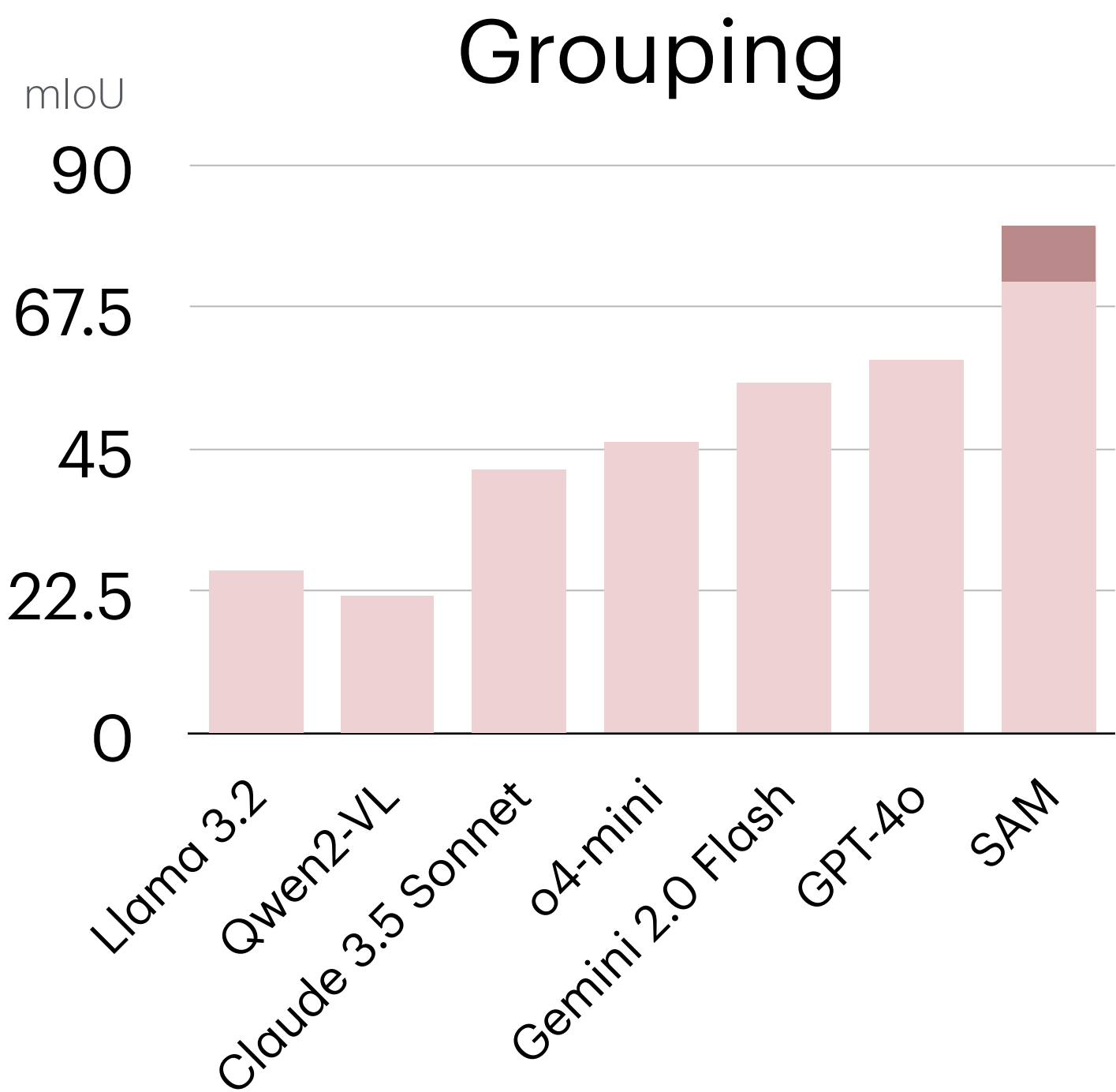
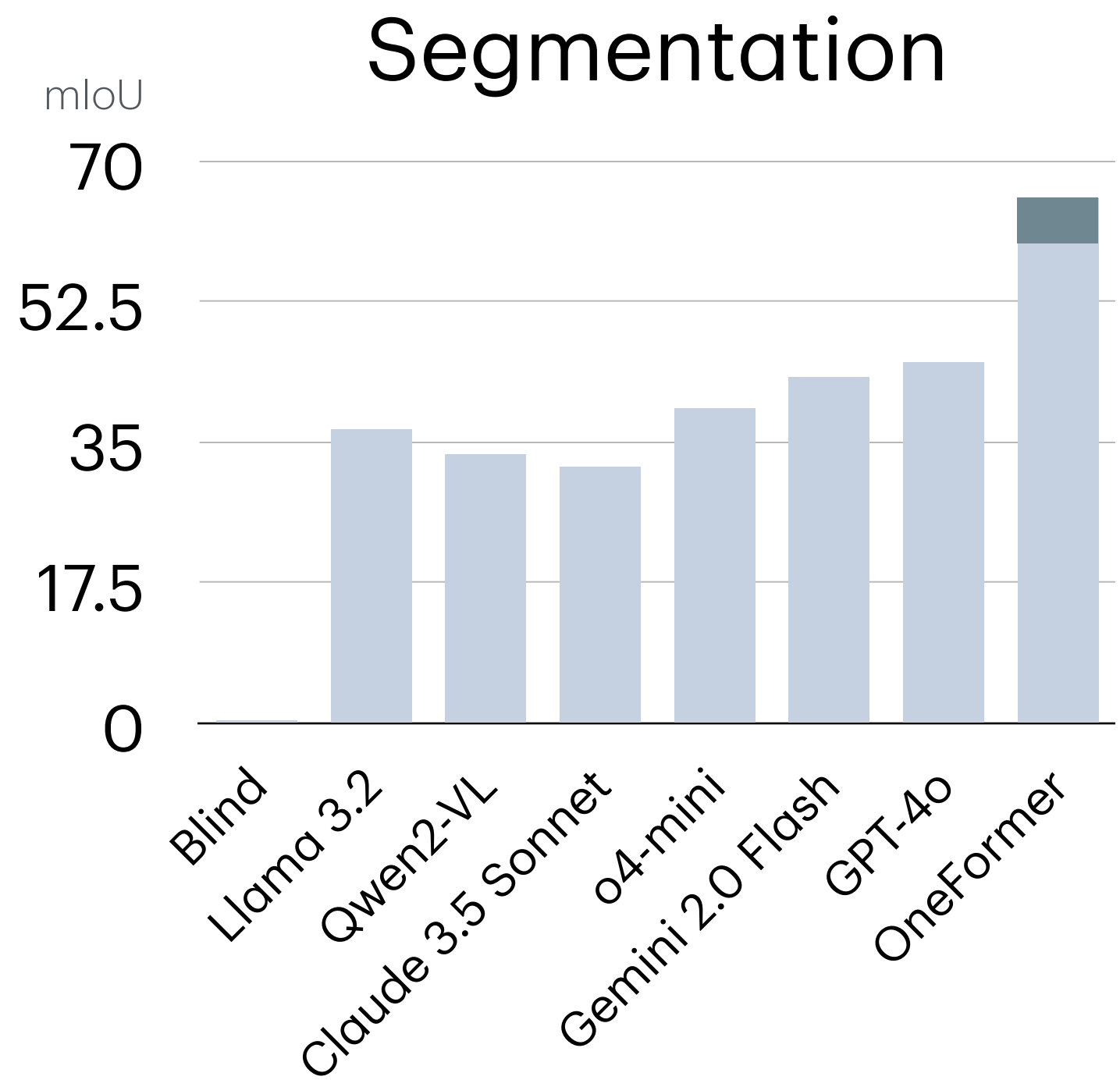
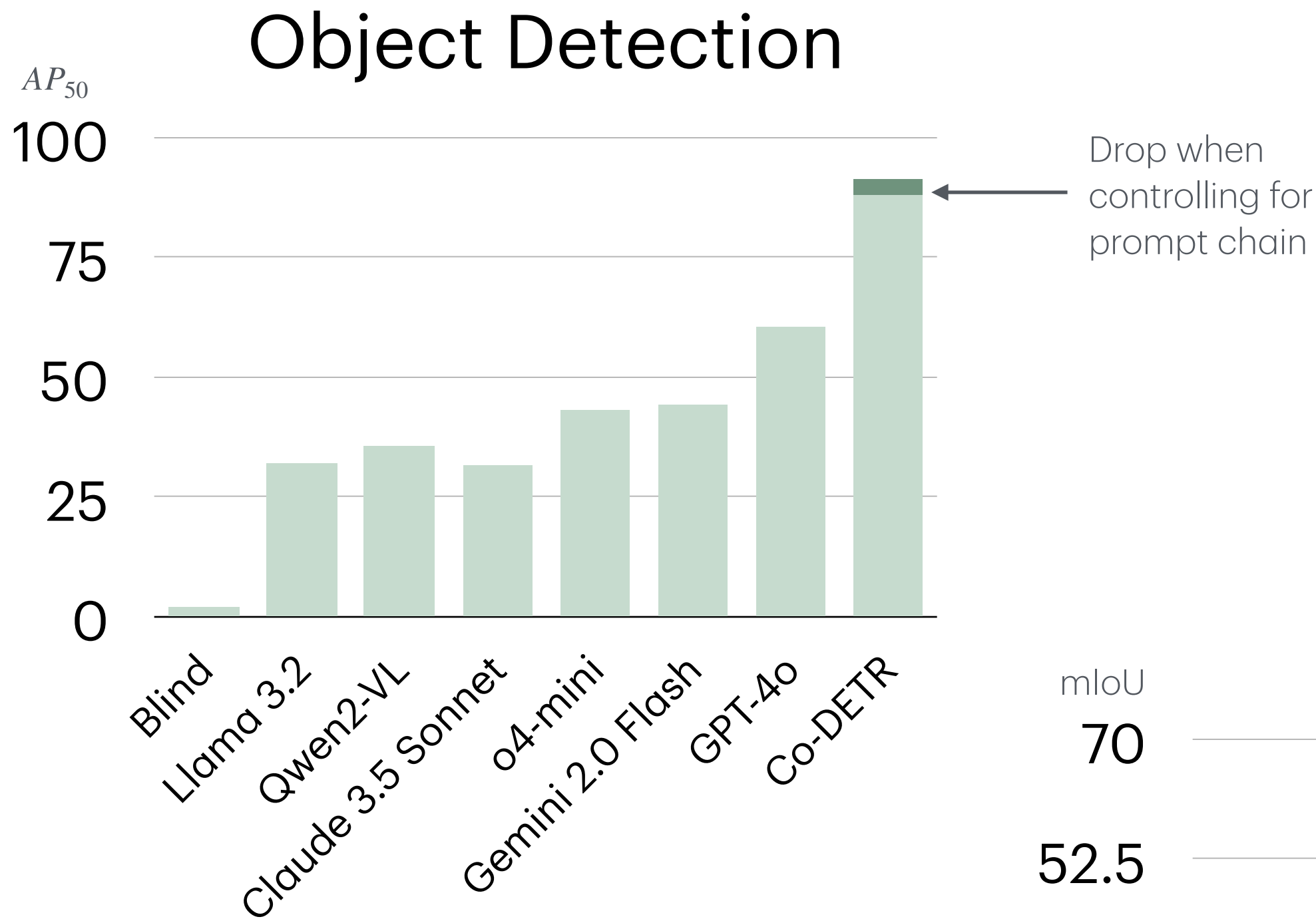
Experimental Results: Takeaways



Key Takeaways

- Not state-of-the-art but respectable generalists.
- Stronger at semantic tasks than geometric tasks.
- GPT-4o outperforms other non-reasoning models across most tasks.
- Reasoning models like o4-mini perform well at geometric tasks.

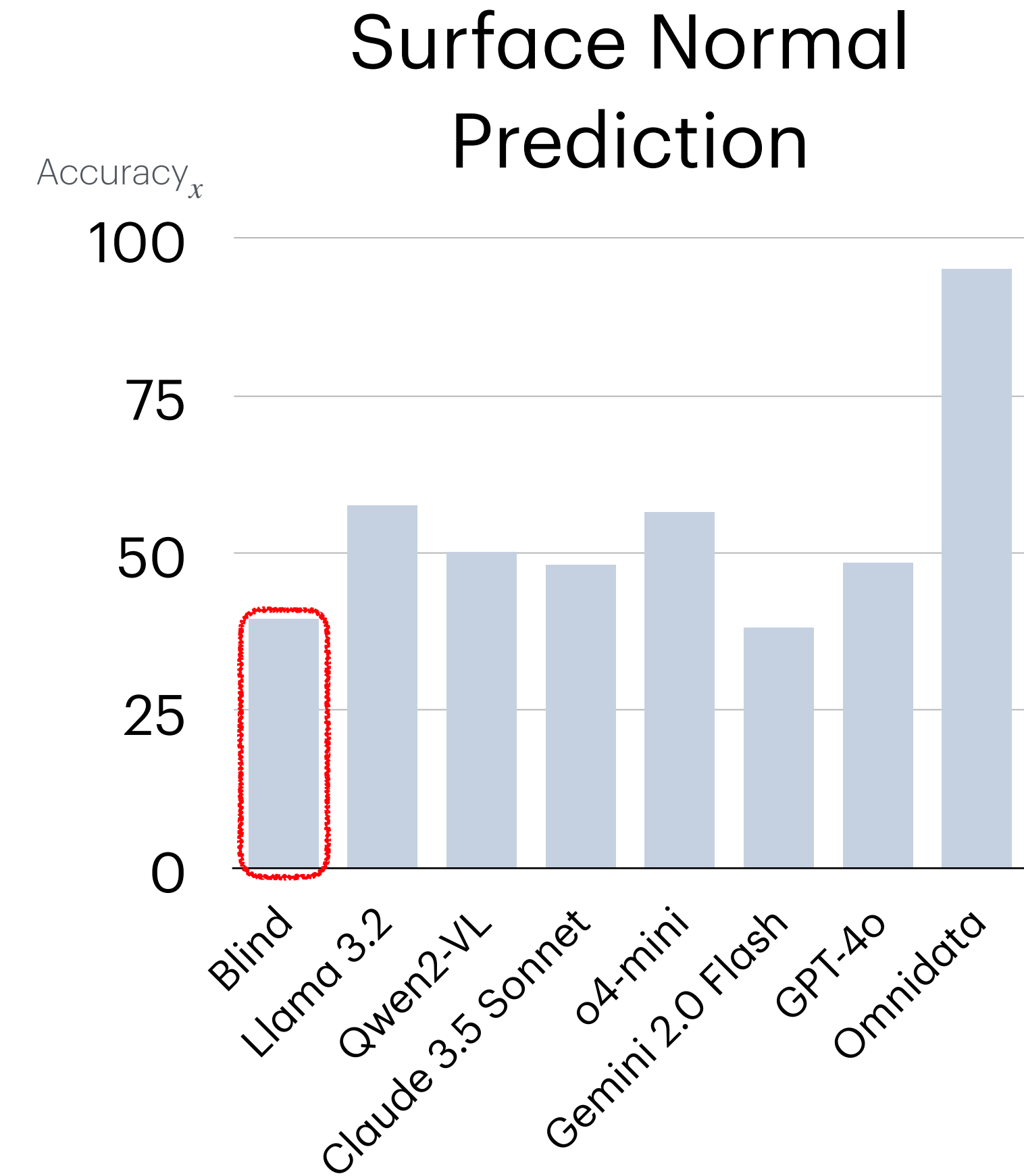
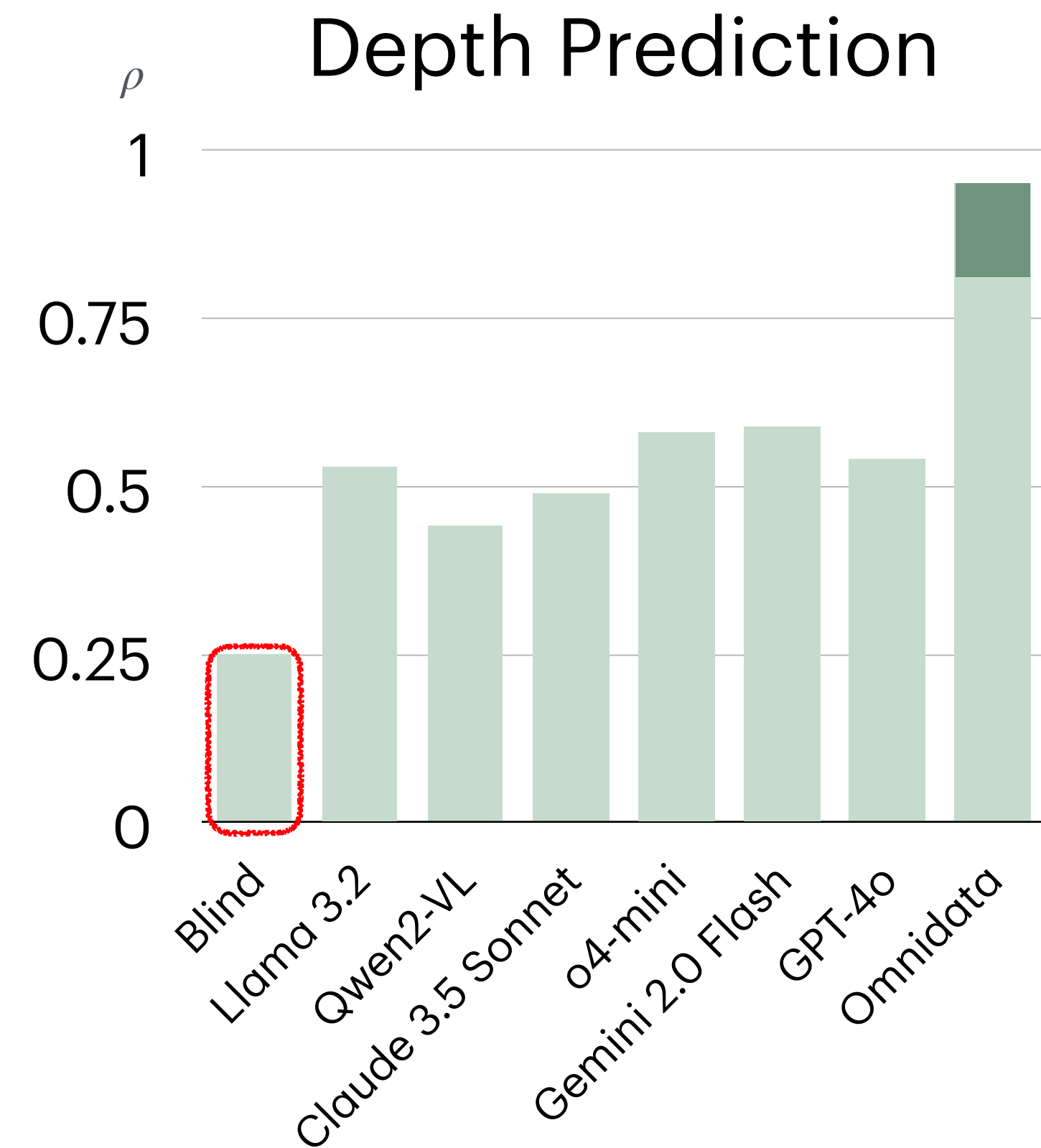
Experimental Results: Semantic Tasks



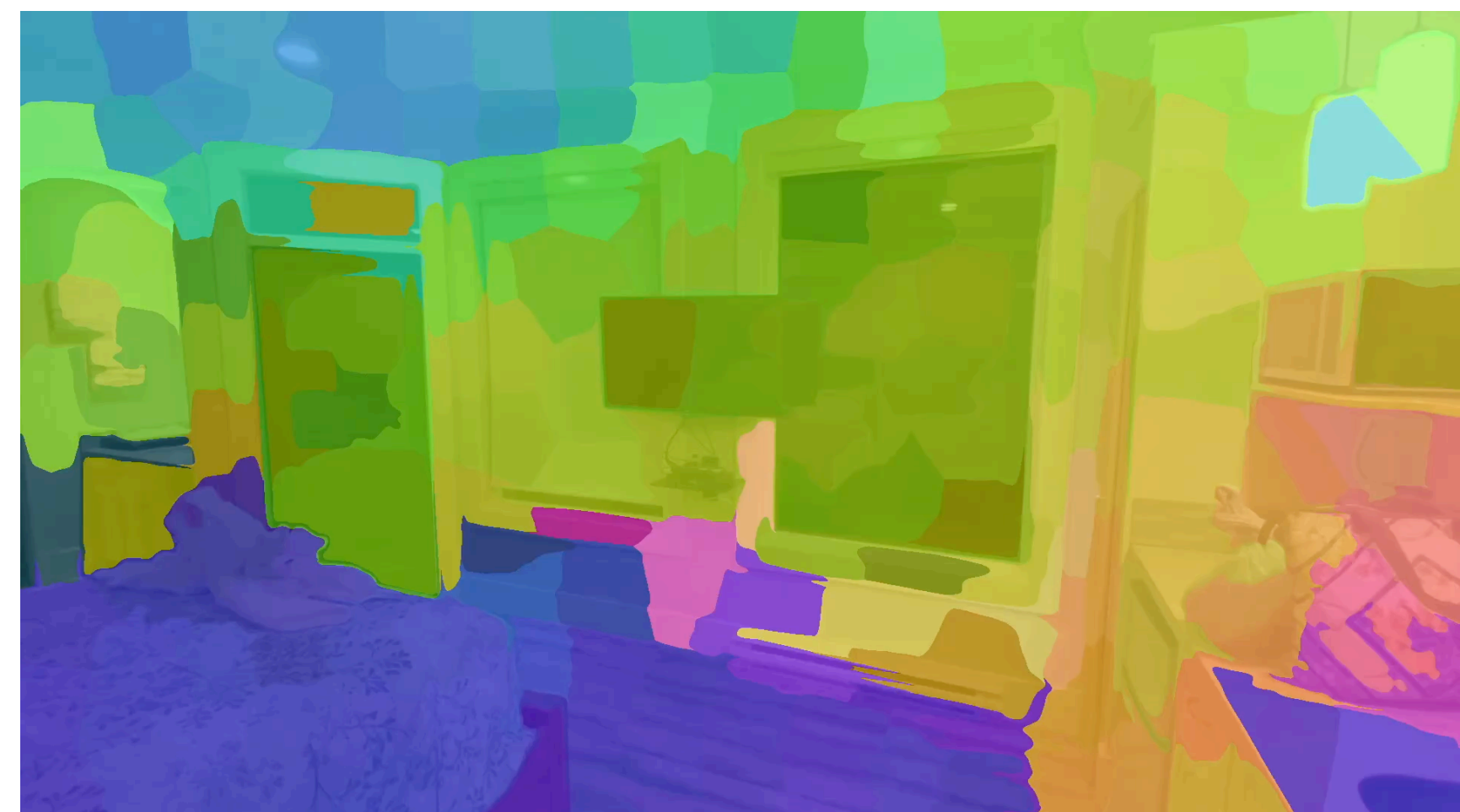
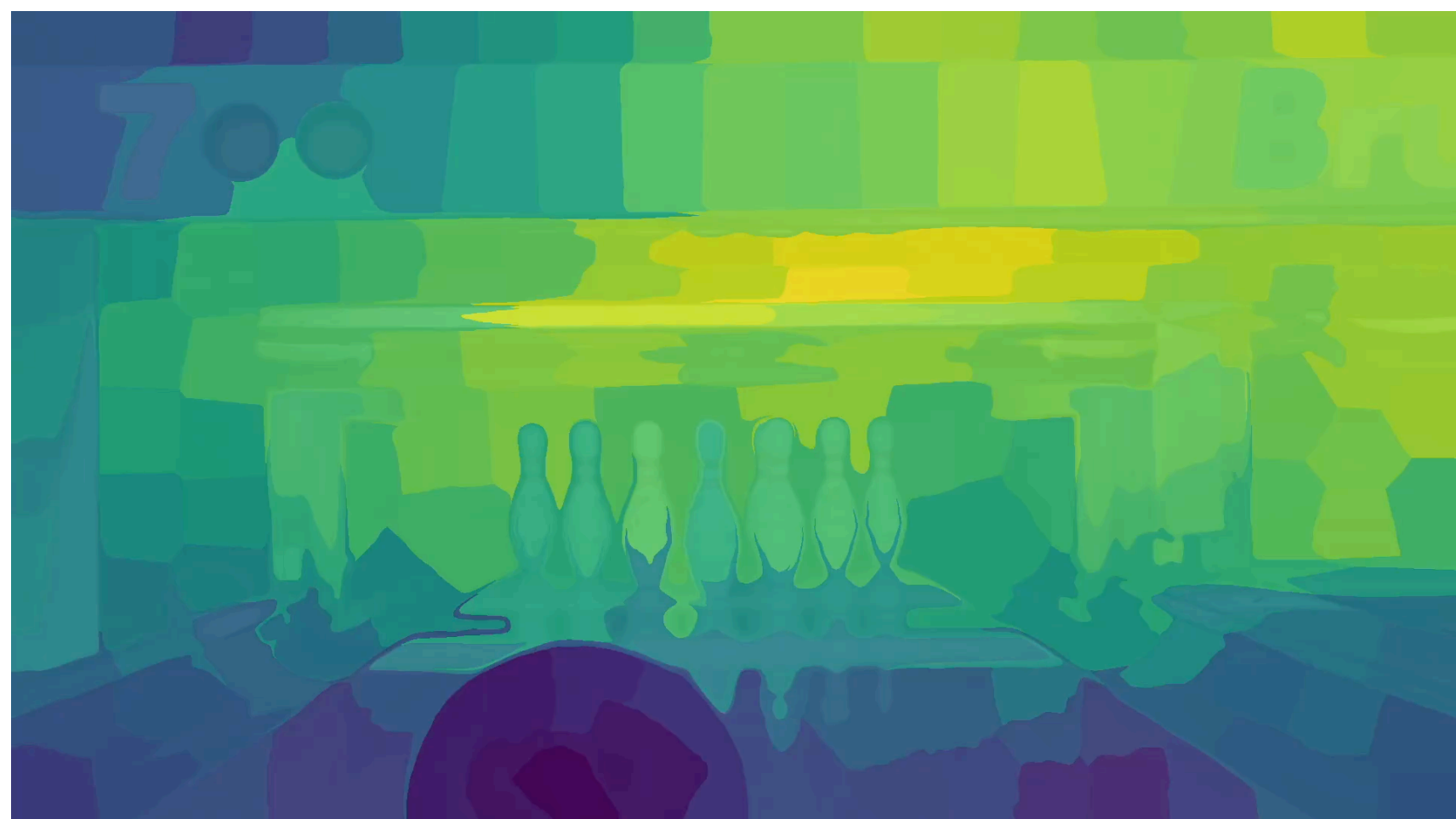
Experimental Results: Semantic Tasks



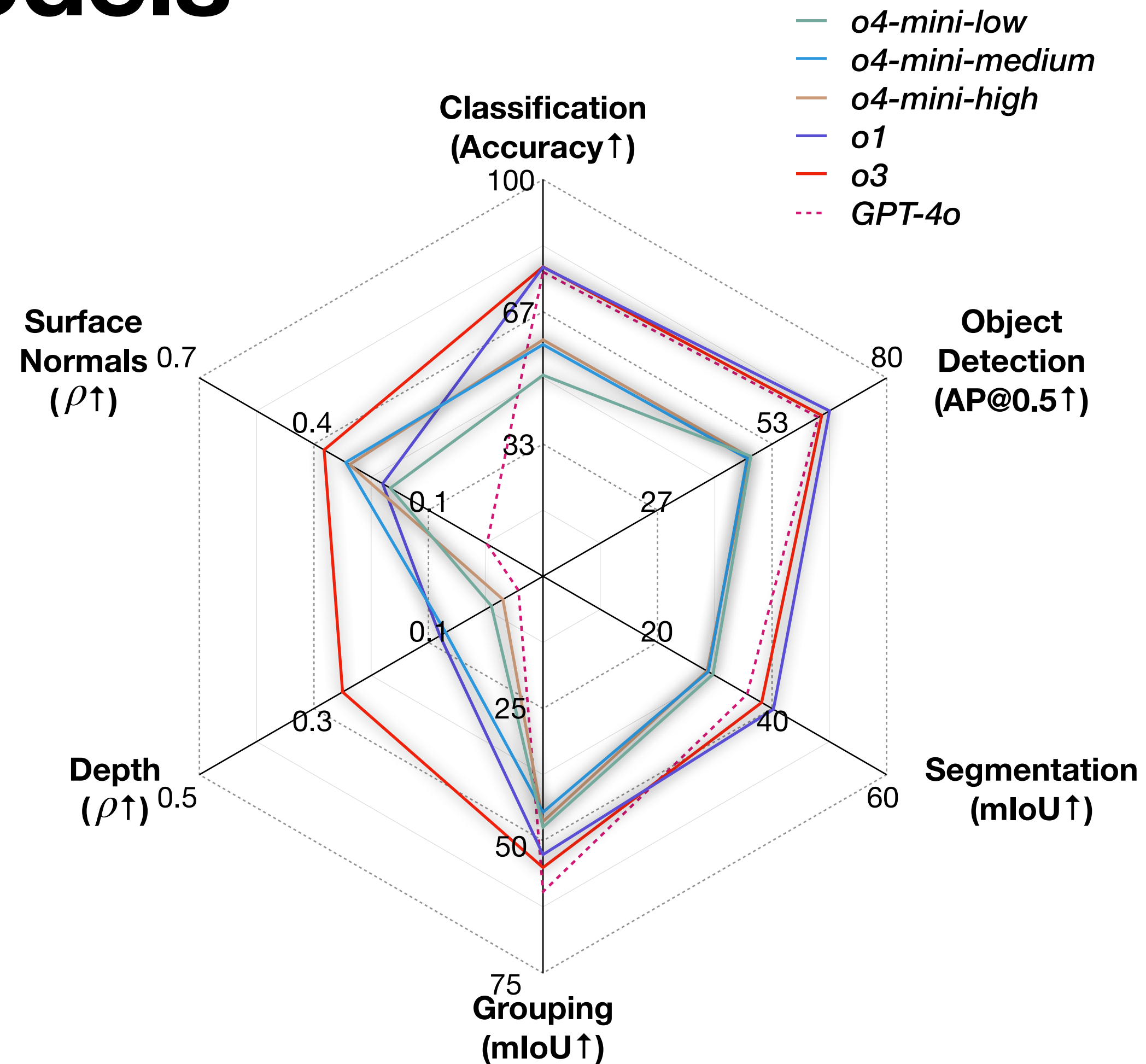
Experimental Results: Geometric Tasks



Experimental Results: Geometric Tasks



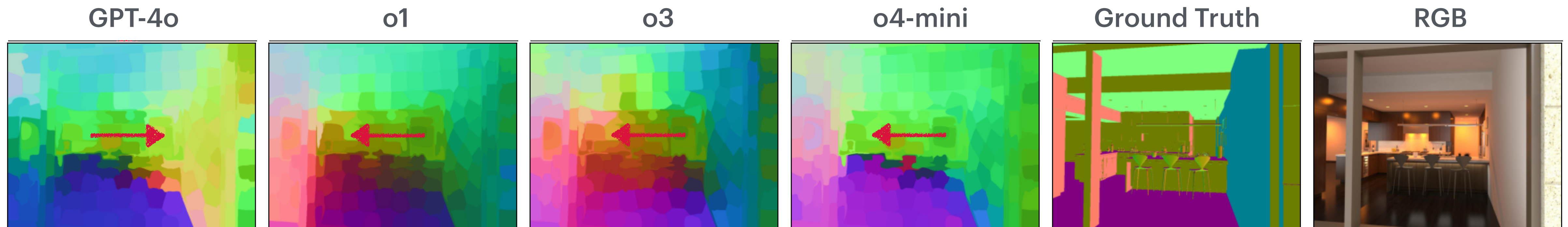
Experimental Results: Comparing Reasoning Models



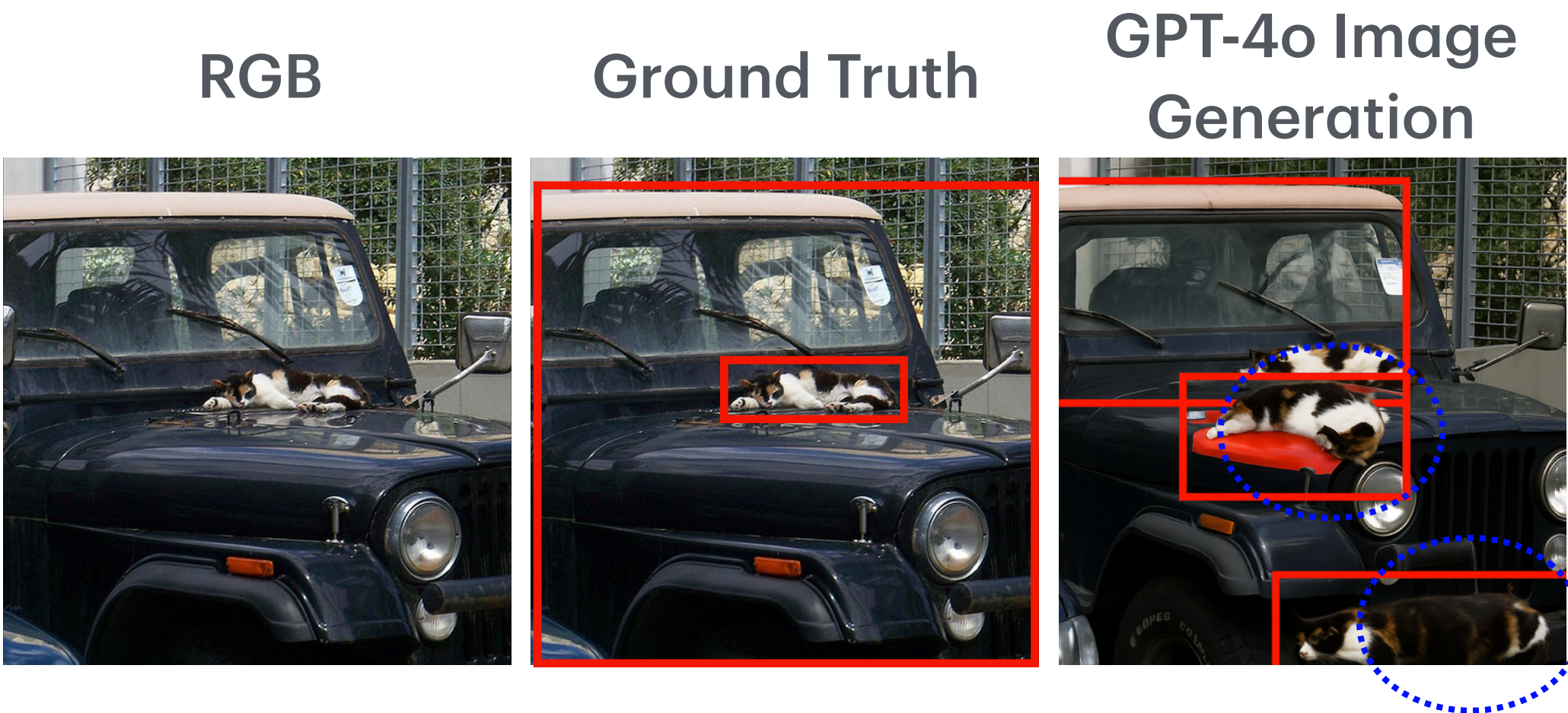
Takeaways

- o1 and o3 are slightly stronger than GPT-4o on semantic tasks.
- All reasoning models outperform GPT-4o on geometric tasks

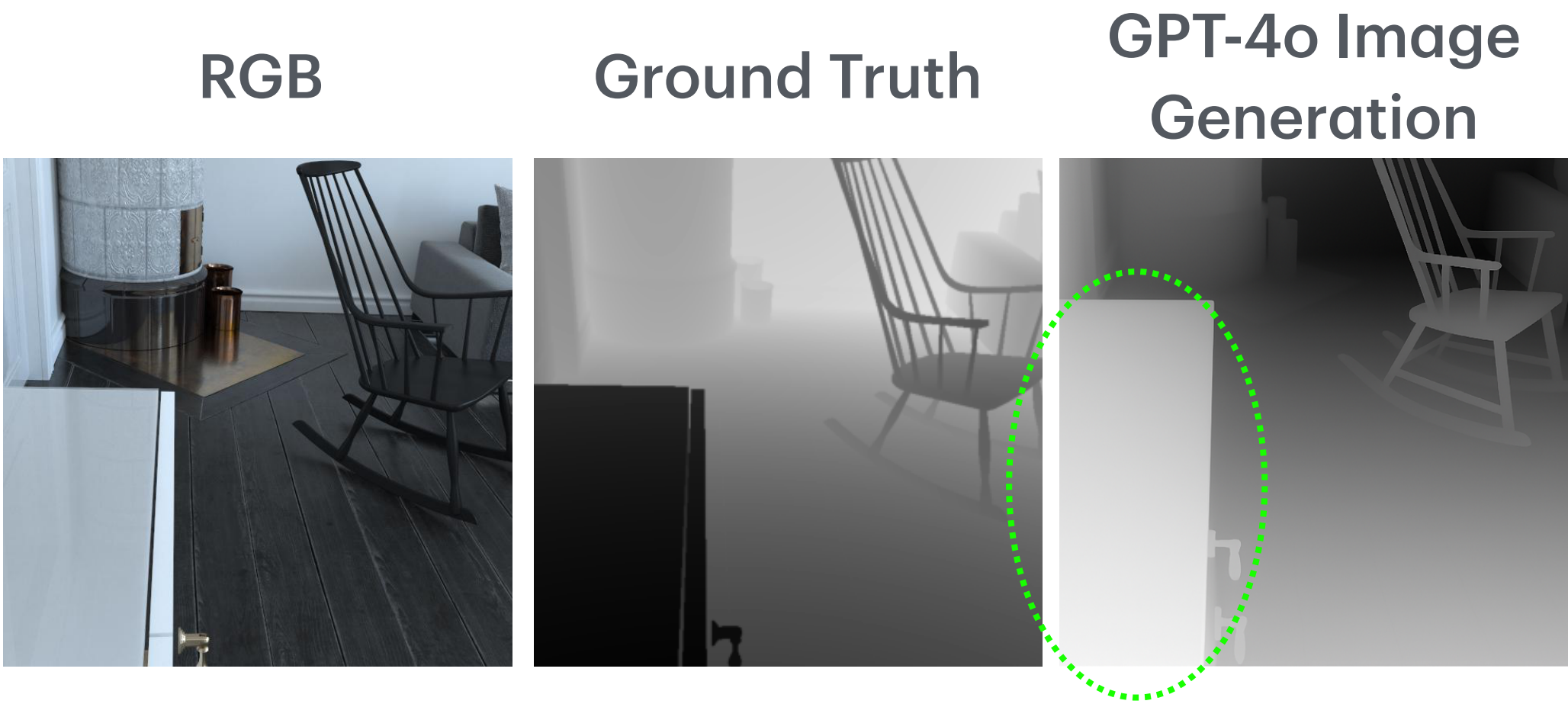
Experimental Results: Comparing Reasoning Models



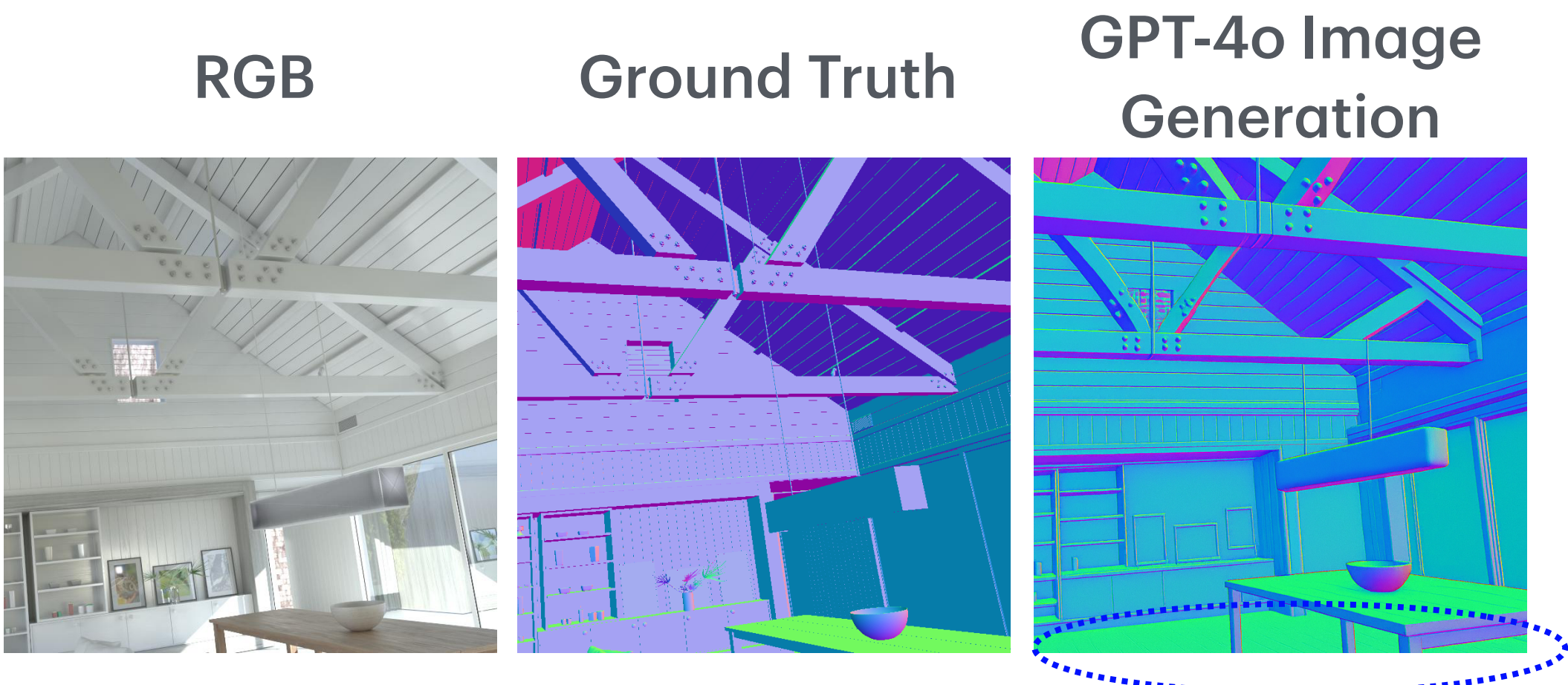
Experimental Results: GPT-4o Image Generation



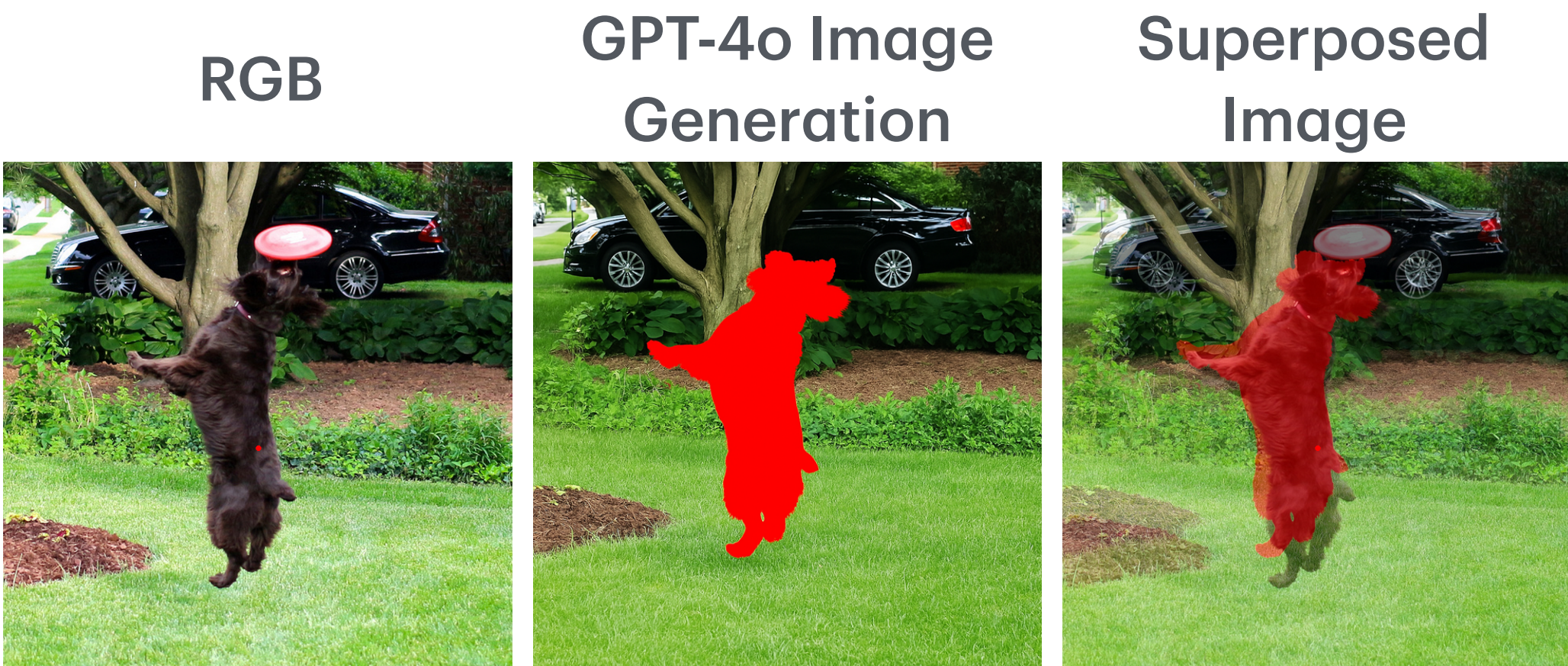
Object Detection



Depth Estimation



Depth Estimation



Grouping

How Well Does GPT-4o Understand Vision? Evaluating Multimodal Foundation Models on Standard Computer Vision Tasks

Rahul Ramachandran, Ali Garjani, Roman Bachmann,
Andrei Atanov*, Oğuzhan Fatih Kar*, Amir Zamir*

