

360° Image Perception with MLLMs: A Comprehensive Benchmark and a Training-Free Method

Huyen T. T. Tran¹, Van-Quang Nguyen¹, Farros Alferro¹, Kang-Jun Liu¹, Takayuki Okatani^{1,2}

¹Graduate School of Information Sciences, Tohoku University · ²RIKEN Center for AIP



Why 360° images break MLLMs

- Flatten a sphere → **distortion** or **fragmentation**



Equirectangular projection (ERP)

Cubemap projection (CMP)

Question: What is the relative position of **the Toys store** and **the Yorozuya store**?

- A. They are on the same side of the street but separated by a side street.
B. They are right next to each other.
C. They are positioned at a right angle to each other.
D. They are across from each other.

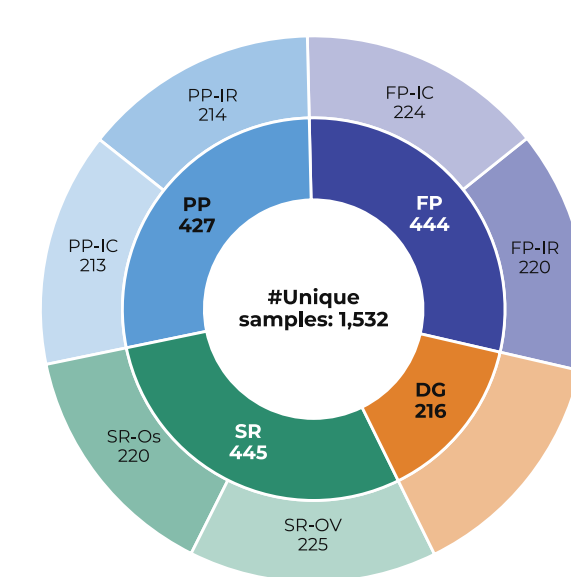
Answer:

GPT-4o: C

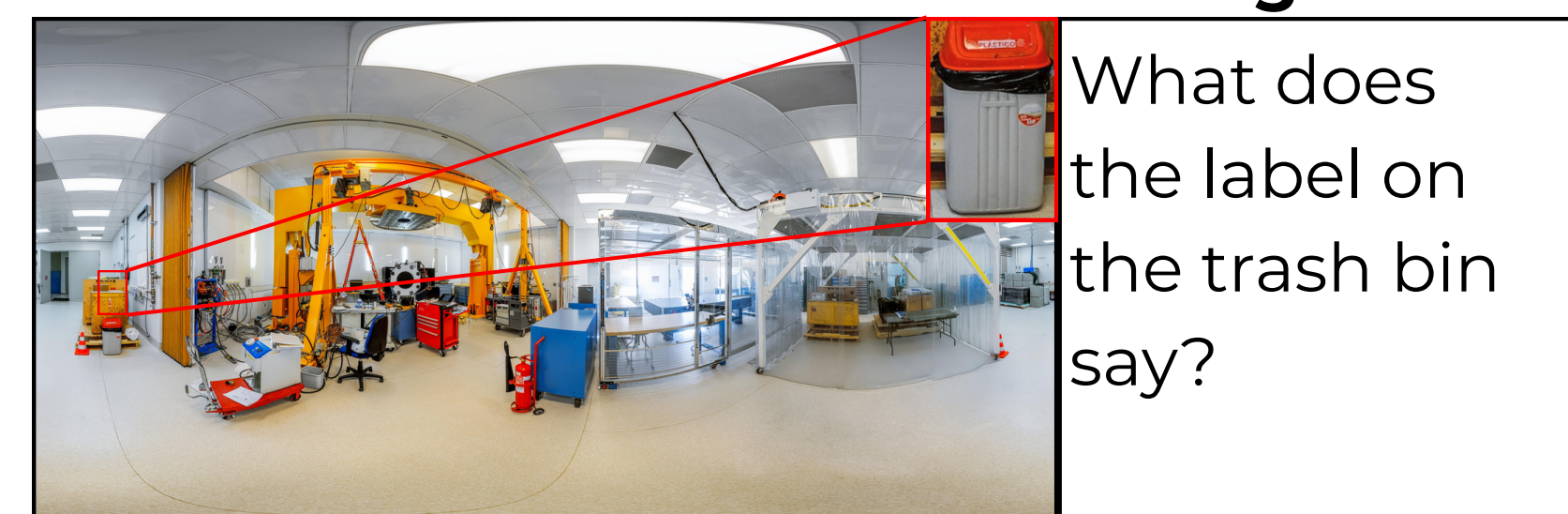
Ours: D

360Bench

- 643 7K-resolution images · 1,532 MCQ samples
- 7 subtasks · human-curated



Fine-grained Perception (FP)



Instance Recognition (IR)

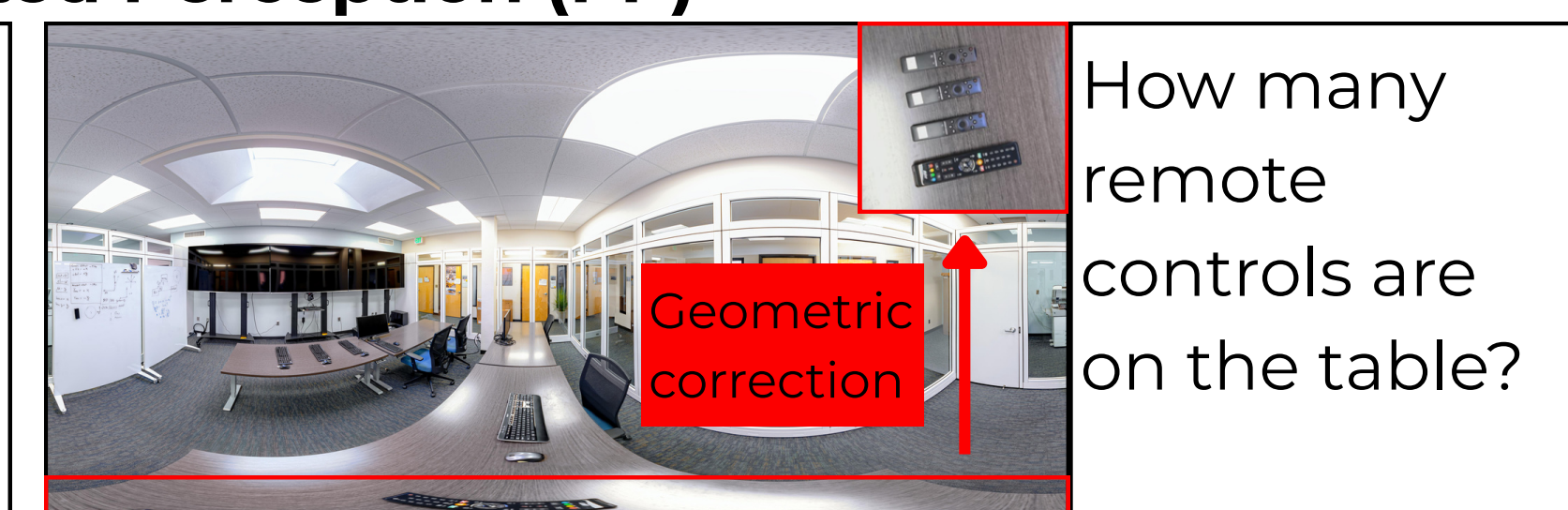


Instance Counting (IC)

Projection-distorted Perception (PP)



Instance Recognition (IR)



Instance Counting (IC)

Spatial Reasoning (SR)



between Objects (Os)



between Object and View (OV)

Direction Giving (DG)



How can I get to Quartier Ville Vieille-Léopold?

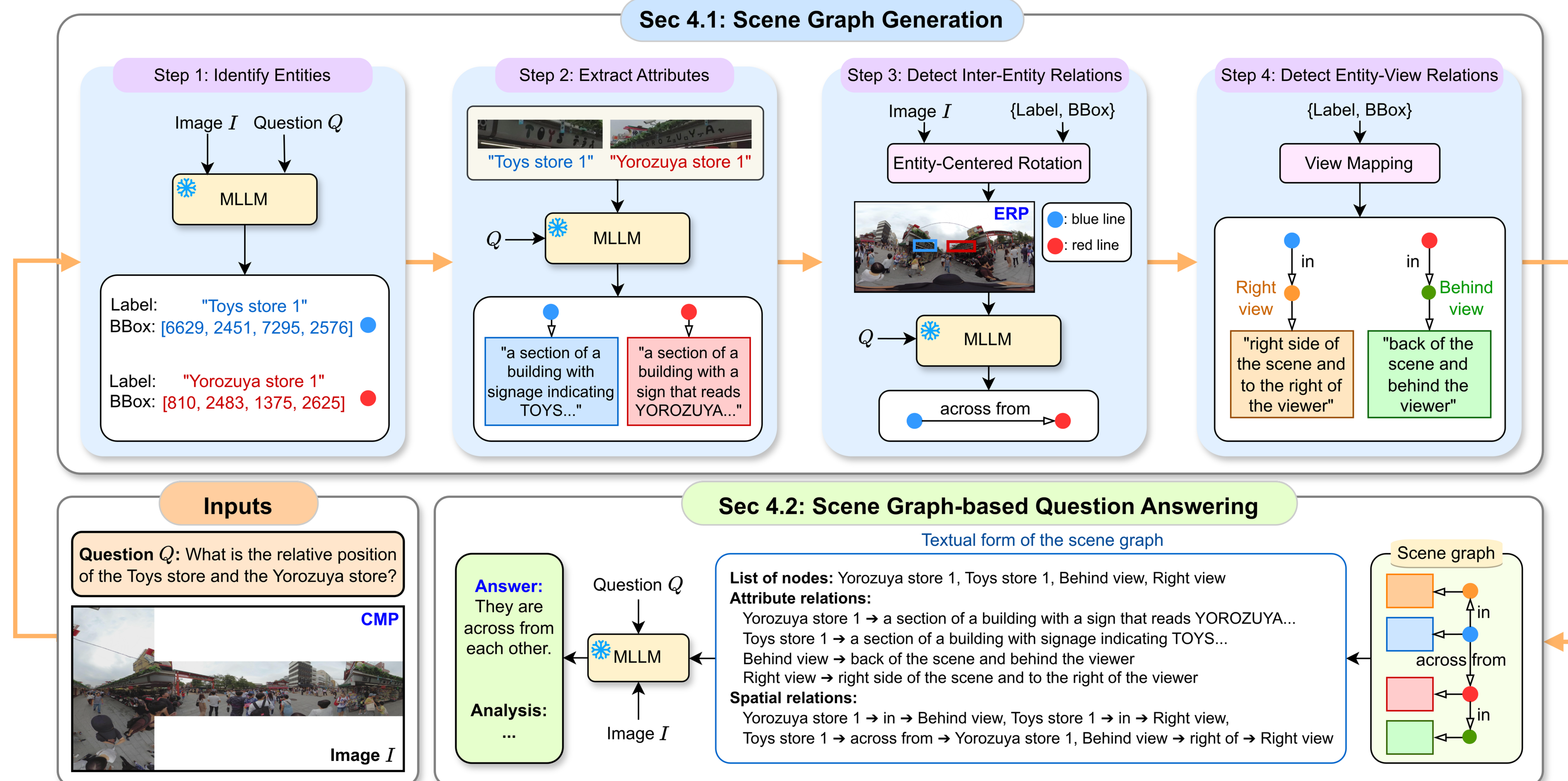
CMP: low distortion → clean object perception (PP ↑, up to **+14.1%**)

ERP: global layout preserved → coherent spatial reasoning (SR ↑, up to **+14.5%**)

Free360

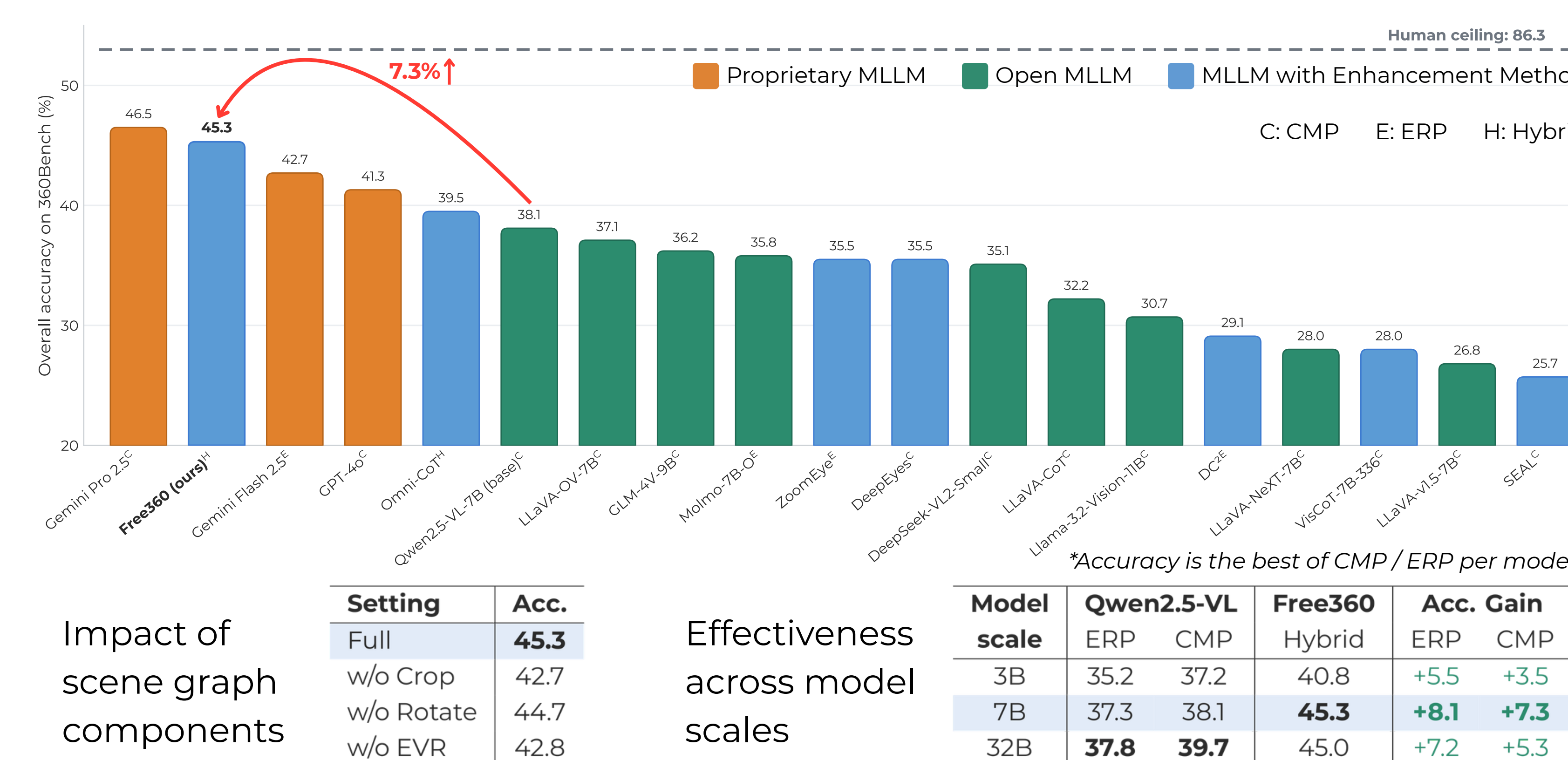
- Build a question-relevant **scene graph** on a frozen MLLM, using **CMP to see** and **ERP to reason**.

Sec 4.1: Scene Graph Generation



Evaluation

- Up to **+22.9% on subtasks** and **+7.3% overall**



Summary

Problem → MLLMs are planar-native; no benchmark measures 360° perception

360Bench → Humans 86.3% >> best model 46.5%; CMP ↔ ERP are complementary

Free360 → Training-free scene graph + 360° operations

Result → Improves every task (up to +22.9%), +7.3% overall

Check our project page →

