



SpaceLLaVA

Advance the Frontier
of Spatial Reasoning

What is Visual Spatial Reasoning?

1



VLMs can caption 



VLMs are weak on 3D spatial relationships 



Trained on internet images without distances



Can we enhance VLMs for embodied AI/robotics?

Beyond Frontier Spatial Reasoning 2



what is the distance between the man with a red hat and the shelving unit?



From the image you provided, I cannot precisely calculate the distance between the man with the red hat and the shelving unit as it would require additional context like exact measurements or scale references in the image. However, based on visual estimation, it appears he is relatively close, perhaps a few meters away from the shelving unit. If you need a more precise distance for specific purposes, consider using tools or methods that can measure distances in images with known scales or dimensions.



Understanding the 3D Scene

3



Sensors: LiDAR



Multi-View Stereo: RealSense



Monocular Depth Estimation: MiDaS

Scene Geometry

4



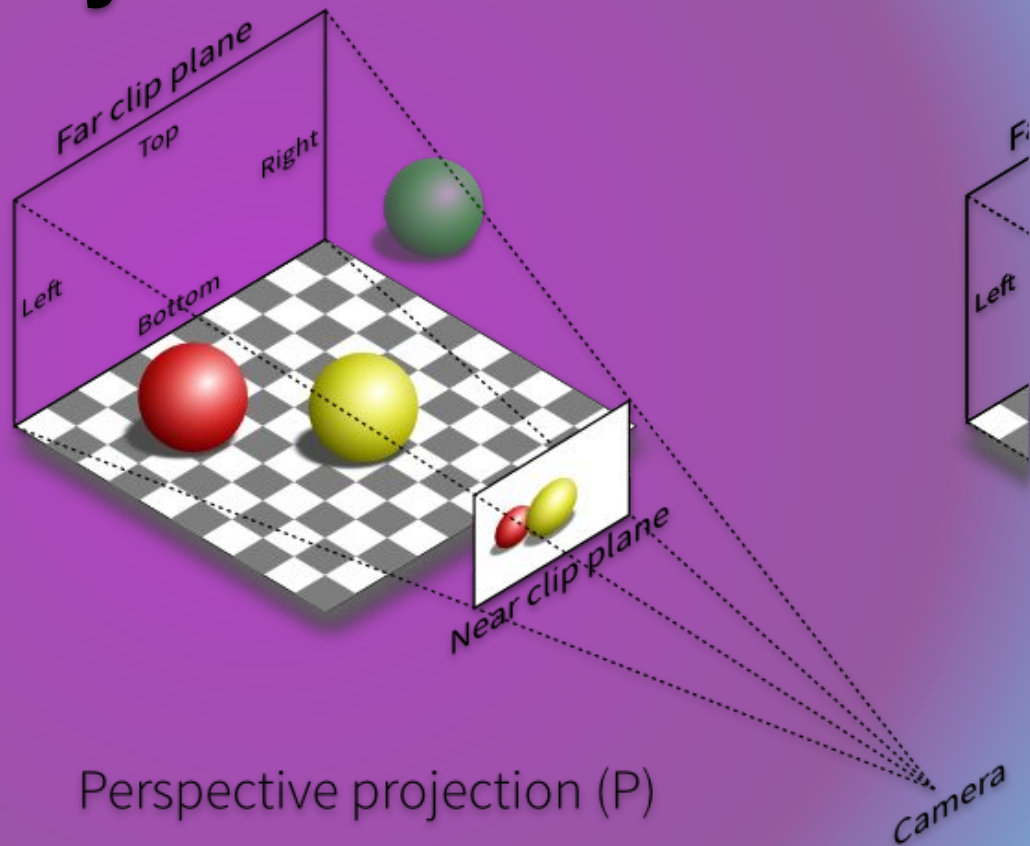
RGB + Depth



Trigonometry



Unprojection



Perspective projection (P)

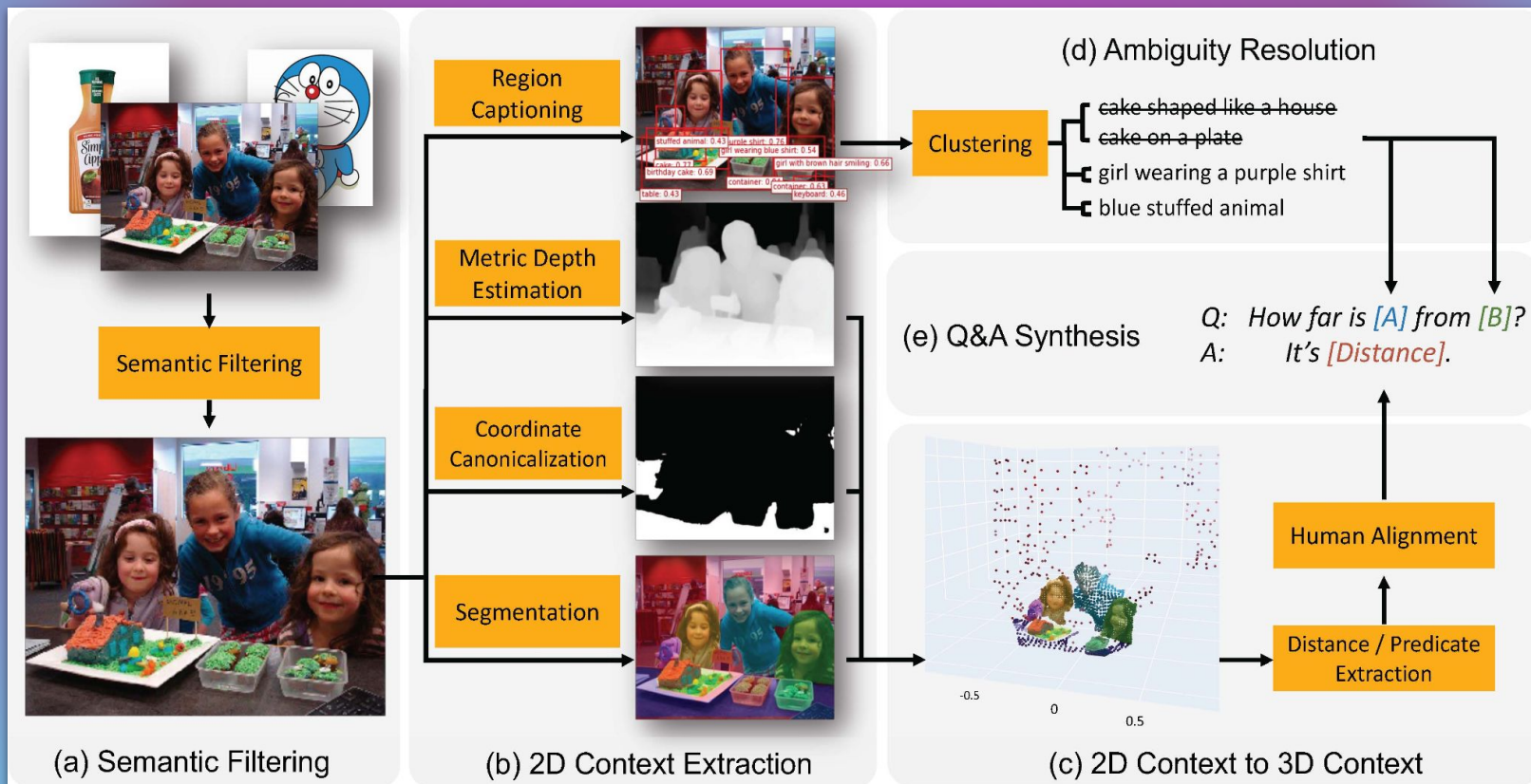
Teaching VLMs to **See** in 3D 🙄

5

- 🌋 Construct the 3D Scene Graph from an Image
- 🌋 Estimate Inter-object Euclidean distances
- 🌋 Synthesize Templated VQA Training Samples

SpatialVLM Data Synthesis Pipeline

6



Replicating **SpatialVLM** in OSS

7



VQASynth - Generate Spatial VQA



LLaVA-1.5-13B for Spatial Reasoning



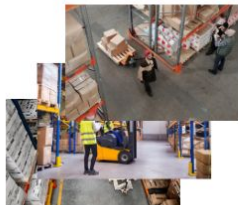
SpaceLLaVA LoRA on 28K samples



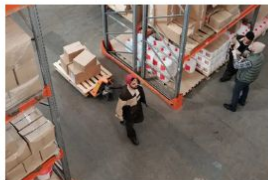
Original VQASynth Pipeline

8

Semantic Filter
CLIP Retrieval

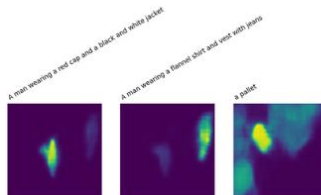


Object Tags & Captions
Recognize Anything
Model (RAM)
LLaVA 1.6 34B

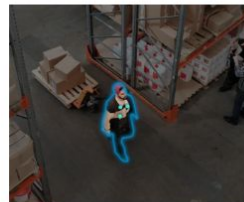


Captions: "A man wearing a red cap and a black and white jacket", "A man wearing a flannel shirt and vest with jeans", "a pallet"

Region Proposal
GroundingDINO
CLIPSeg



Object Segmentation
Segment Anything
Model (SAM)



Monocular Depth Estimation
ZoeDepth



Plane Segmentation
RANSAC



A SpaceLLaVA Response

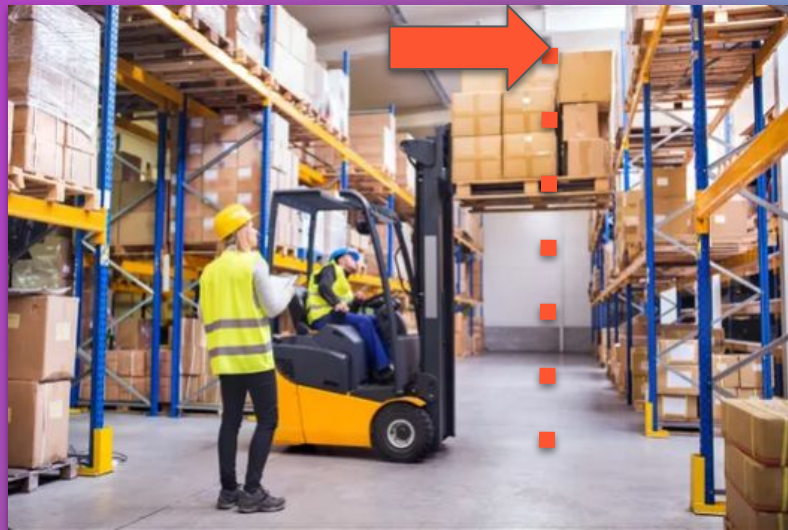
9



How high is the stack of boxes on the pallet held up by the forklift?



SpaceLLaVA - the stack of boxes on the pallet held up by the forklift is 17.32 feet



SpaceQwen Smaller-Faster-Smarter 10



Based on **Qwen2.5-VL-3B-Instruct**

- Updates VQASynth Pipeline



Training Data

- Molmo + SAM2 + VGGT
- 10K VQA samples



SpaceThinker - Reasoning VLM

11

 Based on **VLAA-Thinker-Qwen2.5VL-3B**

- Adds test-time-compute

 Training Data

- Synthetic Reasoning Traces
- 12K VQA Samples



SpaceOm - Best of Qwen & Thinker

12

 Based on **VLAA-Thinker-Qwen2.5VL-3B**

- Targets o_proj

 Training Data

- Adds Robo2VLM-Reasoning
- 18K VQA with Synthetic Reasoning



Evaluating Spatial Reasoning

13



QSpatialBench



SpatialScore



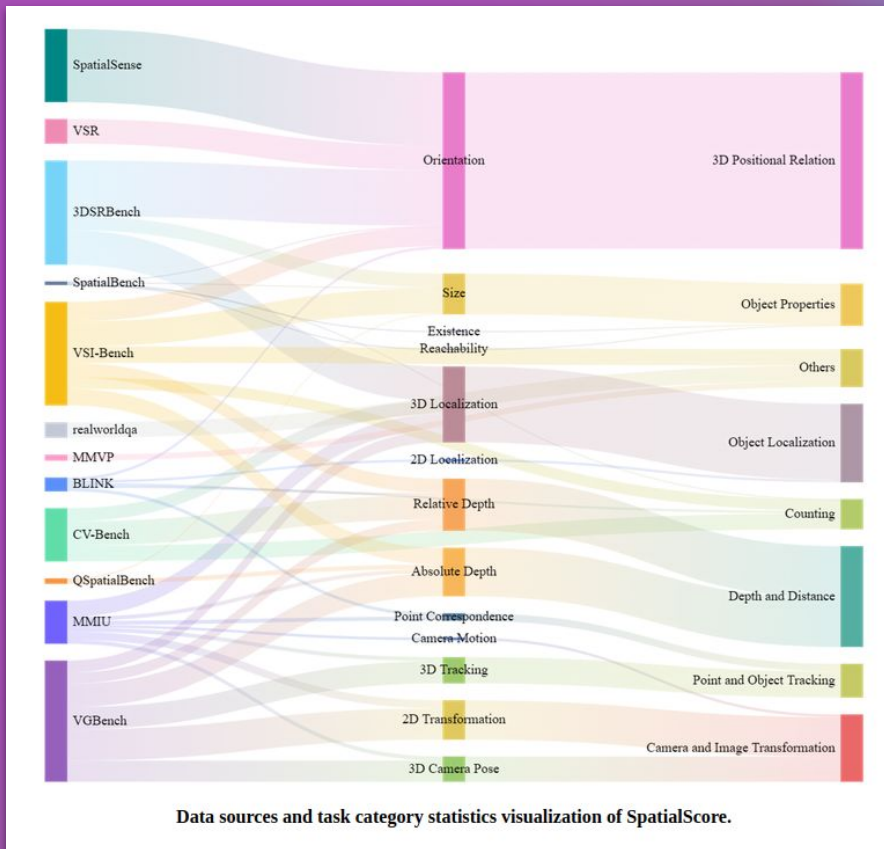
OmniSpatial



SpaCE-10



SIRI-Bench

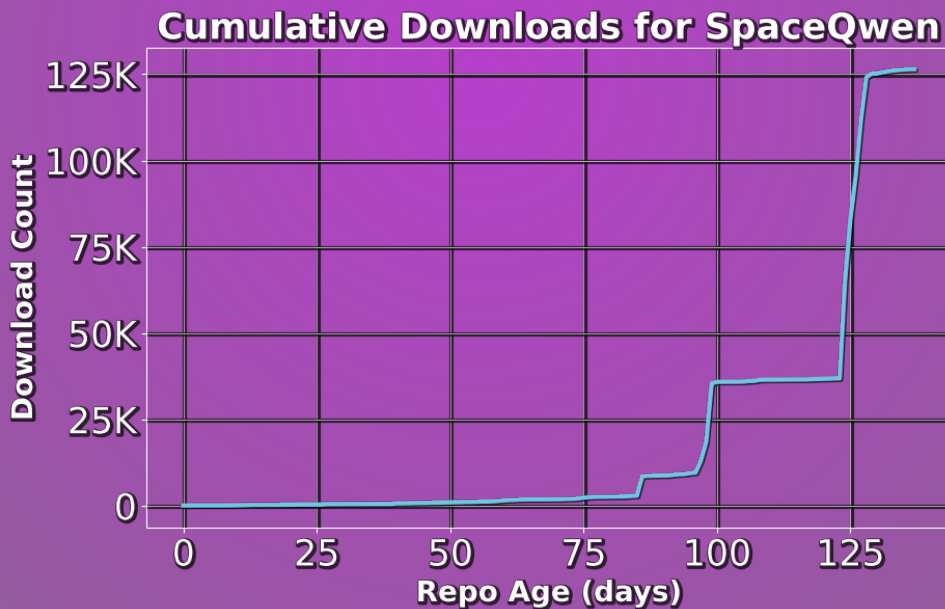


SpaceQwen

14



“... is particularly noteworthy. Despite its small size, it shows competitive performance...” -SIRI-Bench



SpaceThinker

15



QSpatialBench: Beats gpt-4o, near gemini-2.5-pro

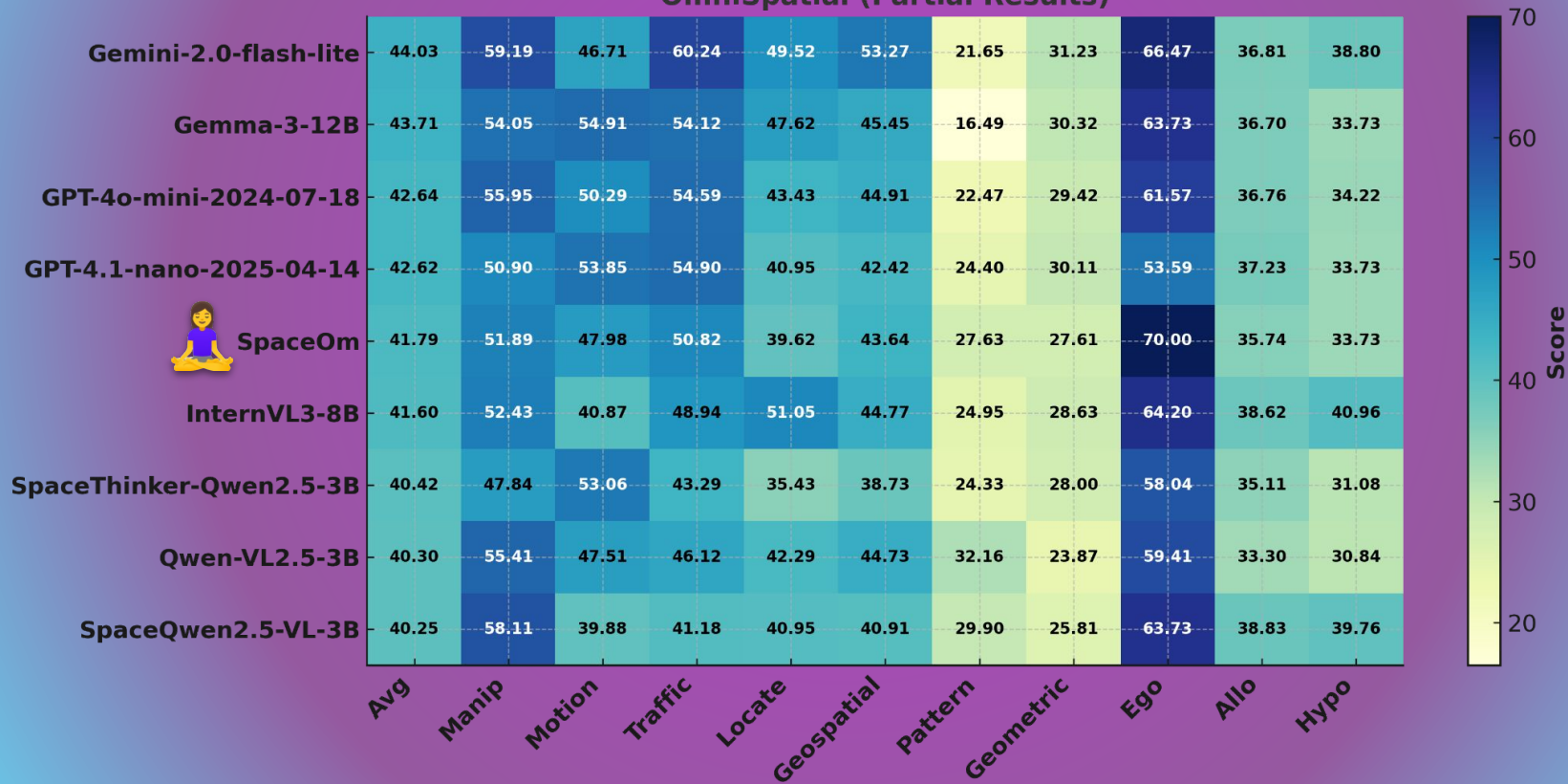


Competitive Distance Estimation

SpaceOm

16

OmniSpatial (Partial Results)



What's Next?

17



Orientation-Aware Scene Graphs like SoFar



Multi-view correspondences like gemini-ER

References

18

-  [SpatialVLM](#)  [VQASynth](#)  [LLaVA](#)
-  [Qwen 2.5-VL](#)  [SpatialScore](#)  [OmniSpatial](#)
-  [VLAA-Thinker-Qwen2.5VL-3B](#)  [SpaCE-10](#)
-  [SpaceQwen2.5-VL-3B-Instruct](#)  [SIRI-Bench](#)
-  [SpaceThinker-Qwen2.5VL-3B](#)  [SpaceOm](#)