



UNIVERSITY OF  
OXFORD



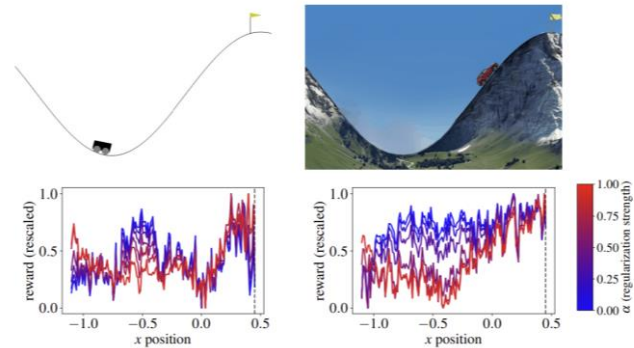
# SpatialPIN: Enhancing Spatial Reasoning Capabilities of Vision-Language Models through Prompting and Interacting 3D Priors

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, Andrew Markham

NeurIPS 2024



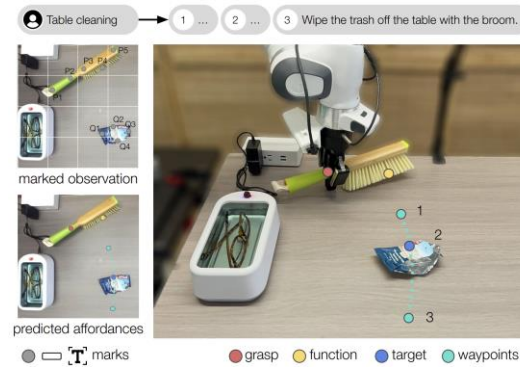
# Motivation



Reward Annotation  
(Rocamonde et al., ICLR 2024)



Robotic Data Generation  
(Wang et al., ICML 2024)



Affordance Grounding  
(Liu et al., RSS 2024)

Equipping vision-language models (VLMs) the capacities of spatial reasoning unlocks exciting applications.



# Motivation



GPT-4V: The stack of towels on the counter appears to be neatly folded and stacked. ...If you can provide any specific details about the size of any item in the image or any other reference, I could attempt a more precise estimate.



Gemini 1.5 Pro: It is impossible to determine the height of the towels with the given image. `\scalar{NaN}` `\distance_unit{NaN}`

<System> ...  
<User> Question: What is the height of the stack of towels on the counter?

Q-Spatial Bench  
(Liao et al., EMNLP 2024)



User  
You are a cleaning robot that is 1 meter wide, now you are standing in a living room and see the image, you want to move from here to the door that leads to the backyard, do you think I can go through the path between the sofa and the table and chairs.



GPT-4V\*  
As an AI, I'm unable to physically interact with environments, but I can provide some insights based on the image provided... From the image, we can make some visual estimations. Judging from the image, it appears that the sofa and the table are quite close to each other, and there may not be enough space to accommodate a 1-meter wide robot...



Spatial-VLM  
[VLM Reasoning] How wide is the path between the sofa and the table and chairs? [VLM Answer] 1.56m [Answer] Yes, the robot can go through the path between the sofa and the table and chairs since it is wider than the robot's width.



User  
Which can is the closest to the yellow finger that is on the right most side of the picture, the blue, orange or the silver one?



GPT-4V  
After estimating the distances by measuring the Euclidean distances between the yellow finger and each of the cans, it is clear that the silver can is the closest to the yellow finger with an approximate distance of 110 units. In comparison, the blue can is about 262 units away, and the orange can is approximately 233 units away from the yellow finger.



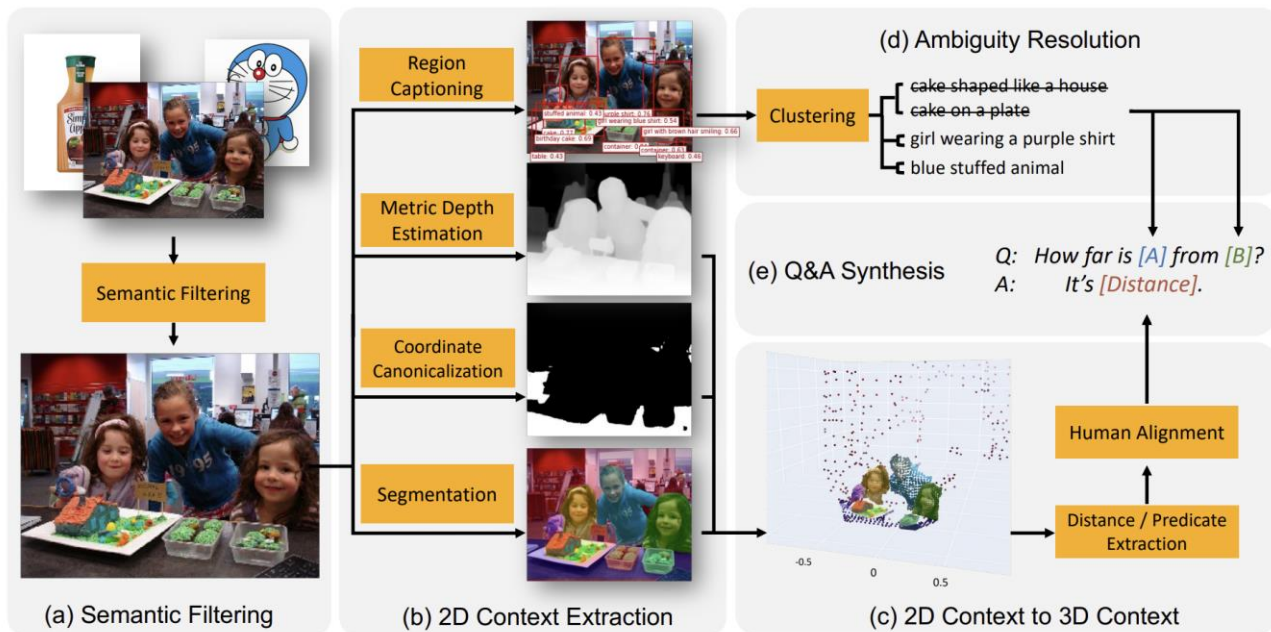
Spatial-VLM  
[VLM Reasoning] How far is the yellow finger from the blue can? [VLM Answer] 0.5m [VLM Reasoning] How far is the yellow finger from the orange can? [VLM Answer] 0.59m [VLM Reasoning] How far is the yellow finger from the silver can? [VLM Answer] 0.21m [Answer] The silver can is the closest to the yellow finger that is on the right most side of the picture.

SpatialVLM  
(Chen et al., CVPR 2024)

However, the spatial reasoning capabilities of VLMs on fine-grained spatial understanding tasks are somewhat limited.



# Related Works

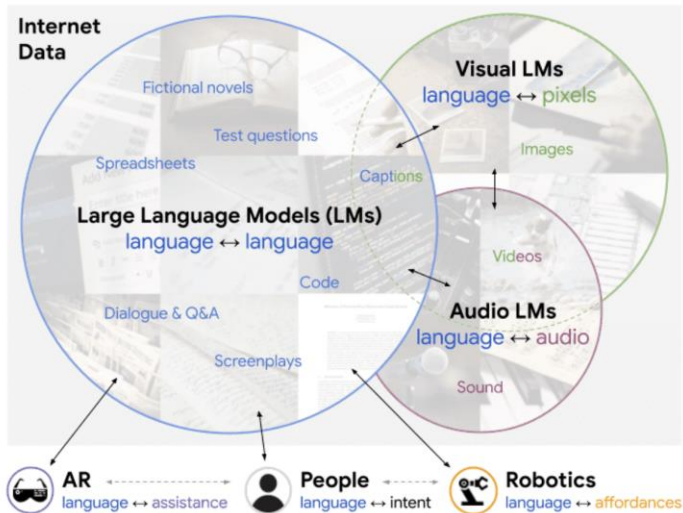


SpatialVLM  
(Chen et al., CVPR 2024)

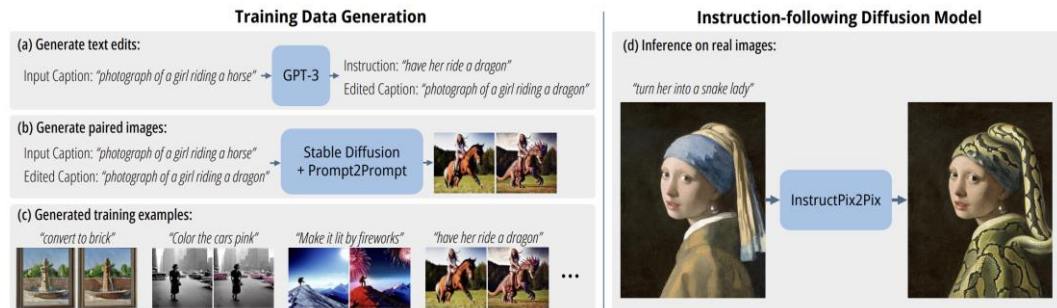
Hypothesis: VLMs finetuned on standard spatial VQA datasets may struggle to generalize beyond their dataset or adapt to more complex spatial tasks.



# Motivation



Socratic Models  
(Zeng et al., ICLR 2023)

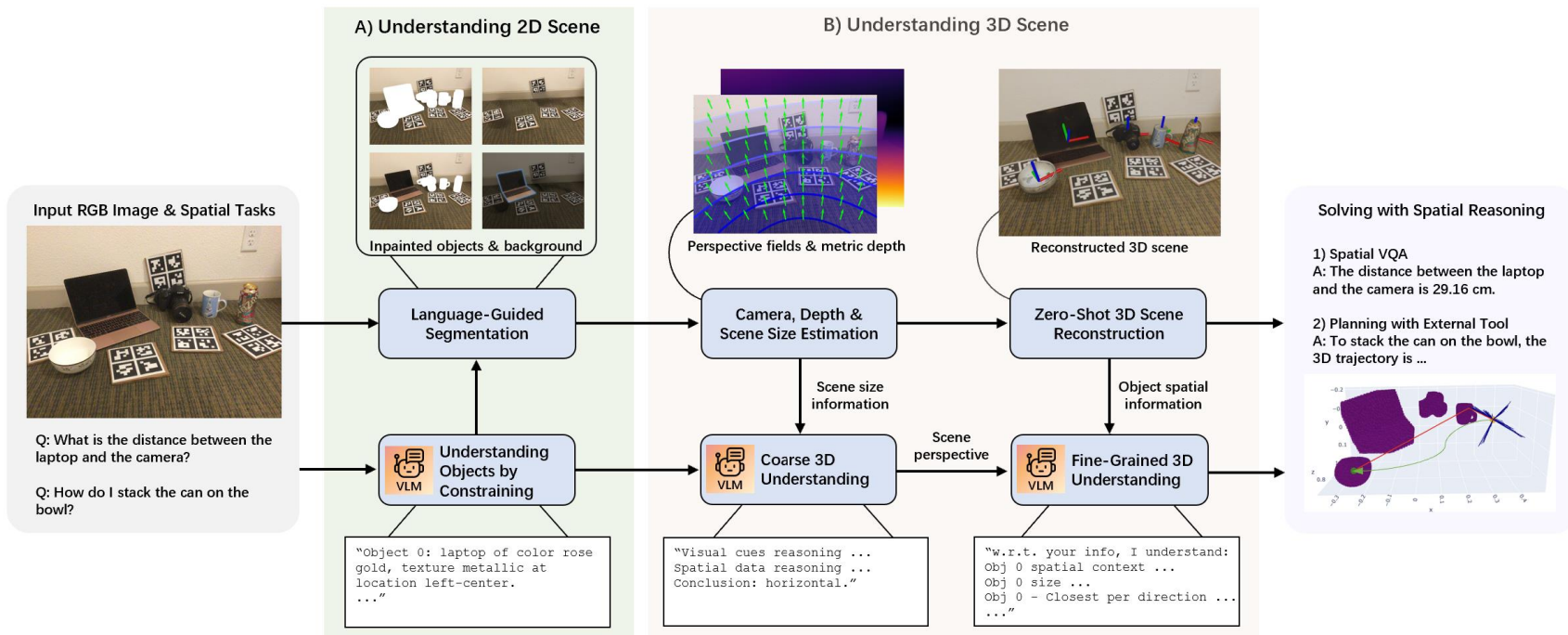


InstructPix2Pix  
(Brooks et al., CVPR 2023)

Can we elevate LLM + multimodal foundation models → VLM + 3D foundation models?



# SpatialPIN

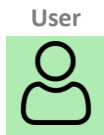


SpatialPIN: A framework that utilizes progressive prompting and interactions between VLMs and 2D/3D foundation models as “free lunch” to enhance spatial reasoning capabilities in a zero-shot, training-free manner.





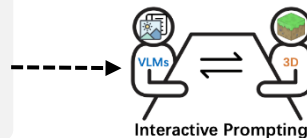
# SpatialPIN



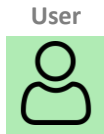
User



You are a robotic grasper. The task is to grasp the carrot. Determine your goal location and closing distance in centimeters to ensure a firm grip on the carrot.



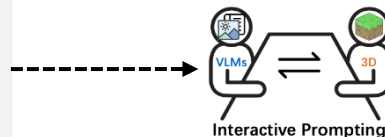
To ensure a firm grip, the grasper should target the center of the carrot at coordinates (0.7399, 0.5185, 0.1503). This midpoint takes into account the carrot's width of 5.44 cm. The grasper should close to approximately 5.3 cm, just less than the carrot's width, for an optimal grasp.



User



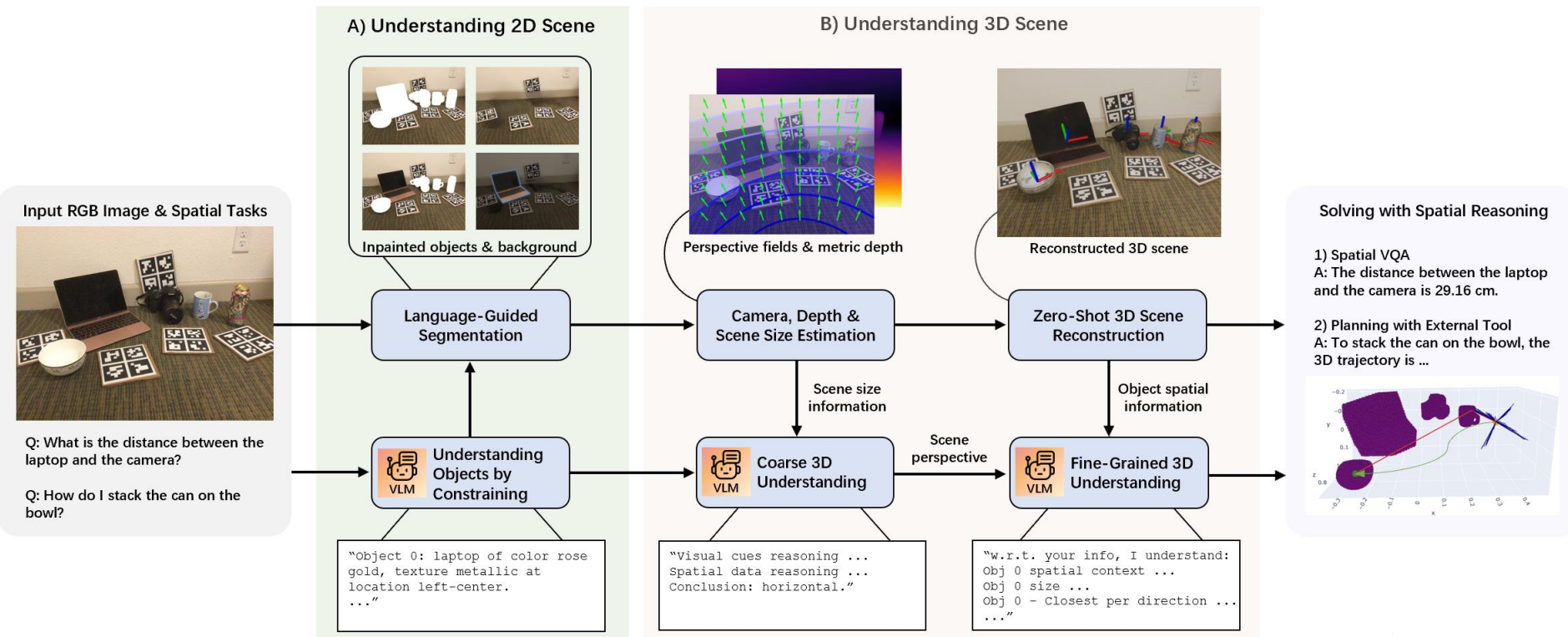
The task is to pick up the plastic bottle and place it inside the trash can. Plan for it.



First, align the plastic bottle with the trash can by rotating the bottle approximately 23.1 degrees around its local z-axis. This adjustment aligns the bottle's x-axis (0.9254, -0.3781, 0.0246) with the trash can's x-axis (0.9982, 0.0398, -0.0447). Next, move the bottle by [0, 0, 5] cm relative to the trash can's 3D center to place it securely inside the trash can. Using RRT\*, the planned 3D trajectory is ..."



# Method Overview





# Method Overview

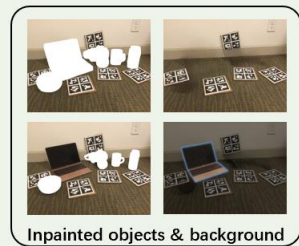
## A) Understanding 2D Scene

Input RGB Image & Spatial Tasks



Q: What is the distance between the laptop and the camera?

Q: How do I stack the can on the bowl?



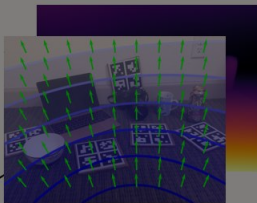
Inpainted objects & background

Language-Guided Segmentation

Understanding Objects by Constraining  
VLM

"Object 0: laptop of color rose gold, texture metallic at location left-center. ..."

## B) Understanding 3D Scene



Perspective fields & metric depth



Reconstructed 3D scene

Camera, Depth & Scene Size Estimation

Scene size information

Coarse 3D Understanding  
VLM

"Visual cues reasoning ...  
Spatial data reasoning ...  
Conclusion: horizontal."

Zero-Shot 3D Scene Reconstruction

Object spatial information

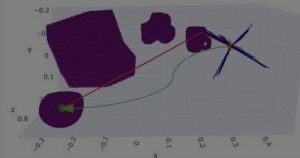
Fine-Grained 3D Understanding  
VLM

"w.r.t. your info, I understand:  
Obj 0 spatial context ...  
Obj 0 size ...  
Obj 0 - Closest per direction ...  
..."

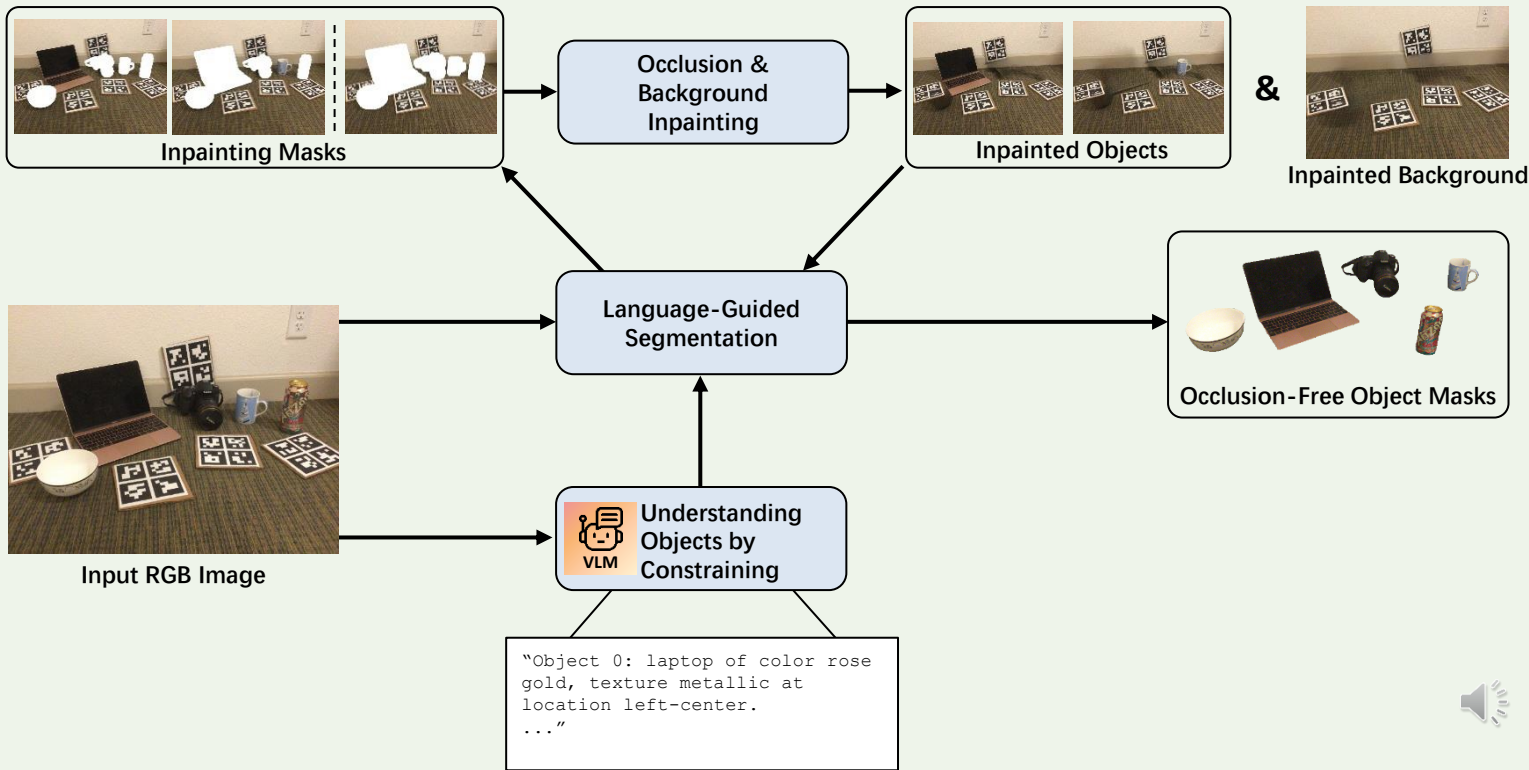
Solving with Spatial Reasoning

1) Spatial VQA  
A: The distance between the laptop and the camera is 29.16 cm.

2) Planning with External Tool  
A: To stack the can on the bowl, the 3D trajectory is ...

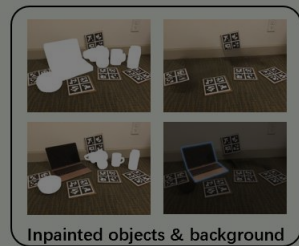


# Understanding 2D Scene



# Method Overview

## A) Understanding 2D Scene

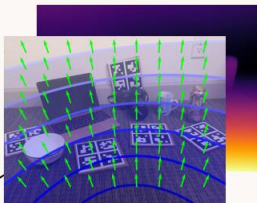


Language-Guided Segmentation

Understanding Objects by Constraining  
VLM

"Object 0: laptop of color rose gold, texture metallic at location left-center. ..."

## B) Understanding 3D Scene



Camera, Depth & Scene Size Estimation

Scene size information

Coarse 3D Understanding  
VLM

"Visual cues reasoning ...  
Spatial data reasoning ...  
Conclusion: horizontal."



Zero-Shot 3D Scene Reconstruction

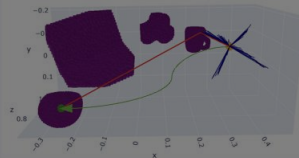
Object spatial information

Fine-Grained 3D Understanding  
VLM

"w.r.t. your info, I understand:  
Obj 0 spatial context ...  
Obj 0 size ...  
Obj 0 - Closest per direction ..."

## Solving with Spatial Reasoning

- 1) Spatial VQA  
A: The distance between the laptop and the camera is 29.16 cm.
- 2) Planning with External Tool  
A: To stack the can on the bowl, the 3D trajectory is ...



## Input RGB Image & Spatial Tasks

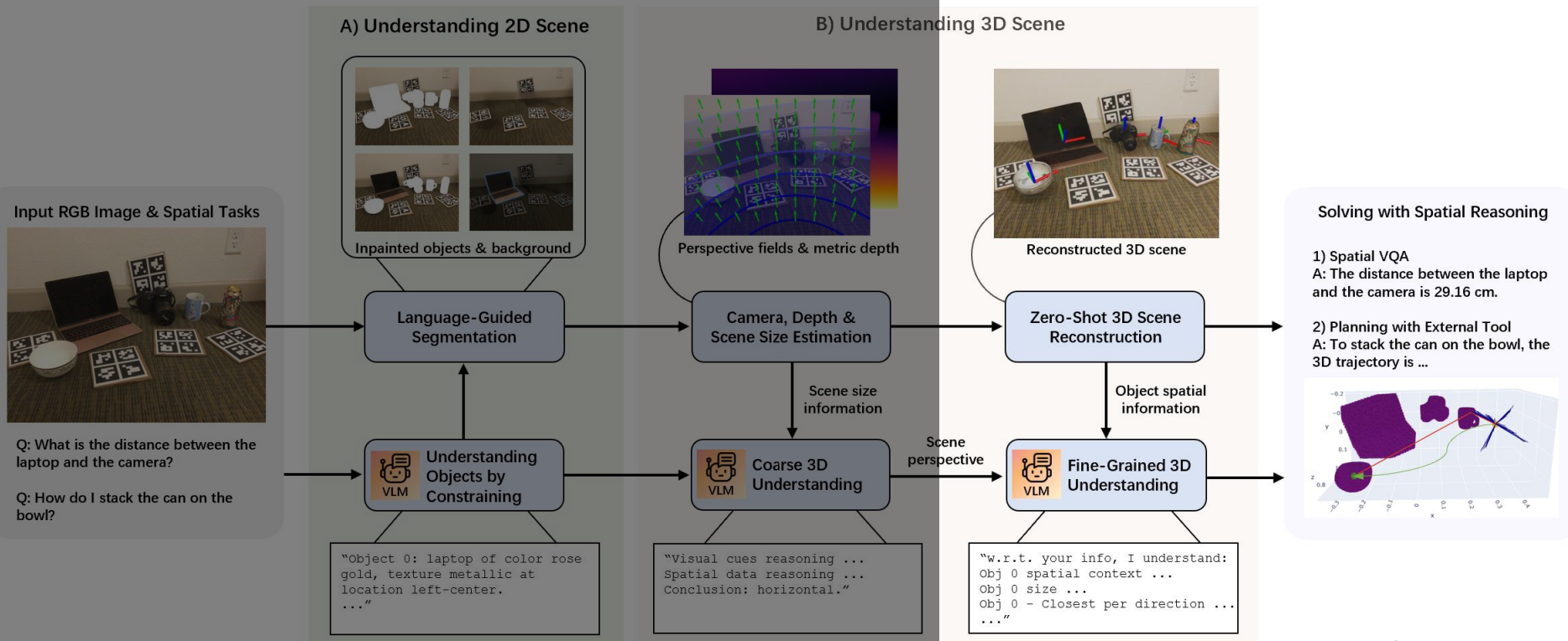


Q: What is the distance between the laptop and the camera?

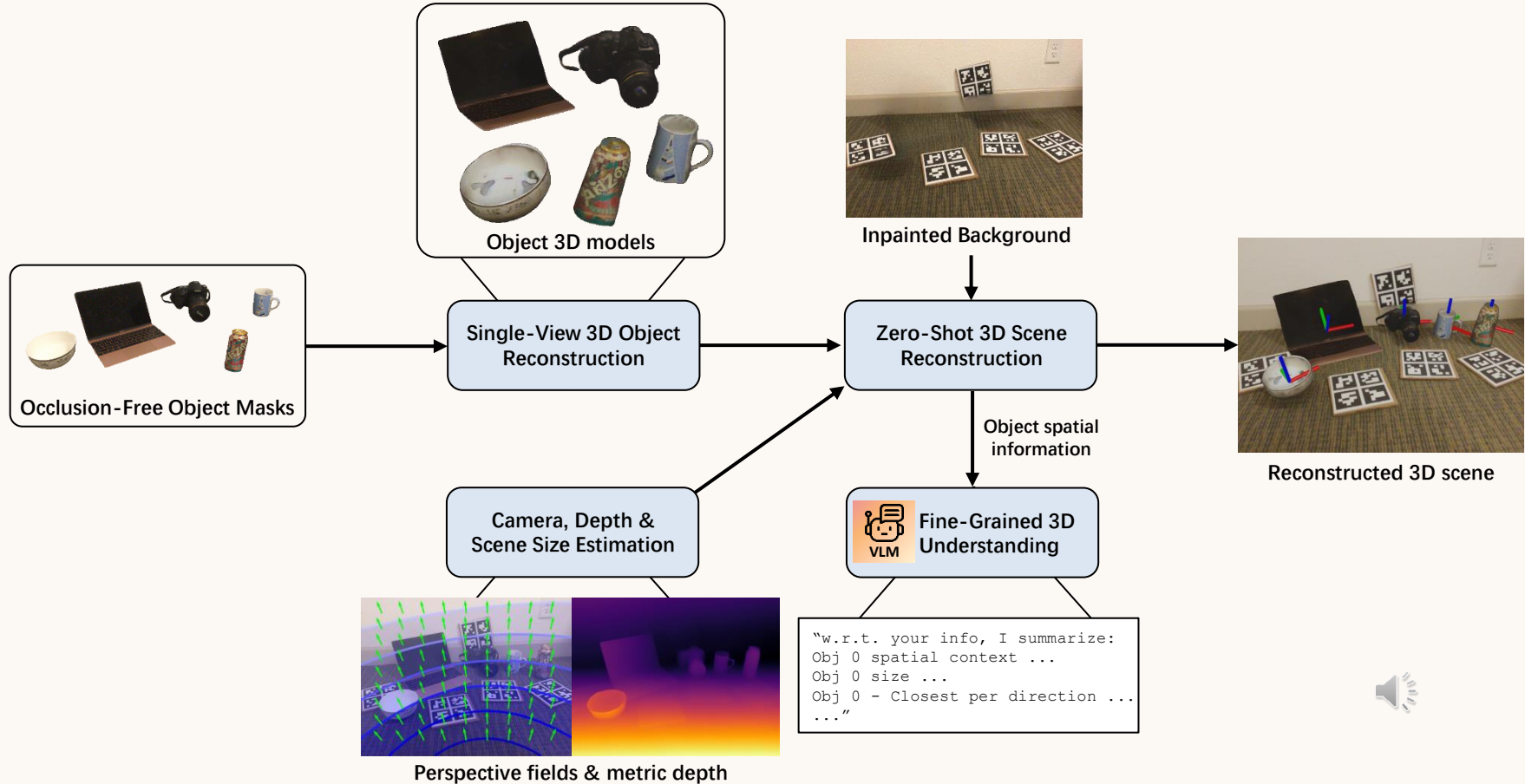
Q: How do I stack the can on the bowl?



# Method Overview



# Fine-Grained 3D Understanding





# Plug-and-Play Framework

## A) Understanding 2D Scene



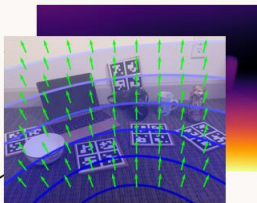
Inpainted objects & background

Language-Guided Segmentation

Understanding Objects by Constraining  
VLM

"Object 0: laptop of color rose gold, texture metallic at location left-center. ..."

## B) Understanding 3D Scene



Perspective fields & metric depth

Camera, Depth & Scene Size Estimation

Coarse 3D Understanding  
VLM

"Visual cues reasoning ...  
Spatial data reasoning ...  
Conclusion: horizontal."



Reconstructed 3D scene

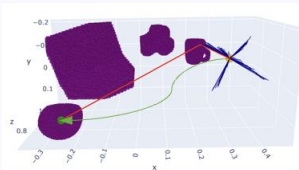
Zero-Shot 3D Scene Reconstruction

Fine-Grained 3D Understanding  
VLM

"w.r.t. your info, I understand:  
Obj 0 spatial context ...  
Obj 0 size ...  
Obj 0 - Closest per direction ..."

## Solving with Spatial Reasoning

- 1) Spatial VQA  
A: The distance between the laptop and the camera is 29.16 cm.
- 2) Planning with External Tool  
A: To stack the can on the bowl, the 3D trajectory is ...



## Input RGB Image & Spatial Tasks



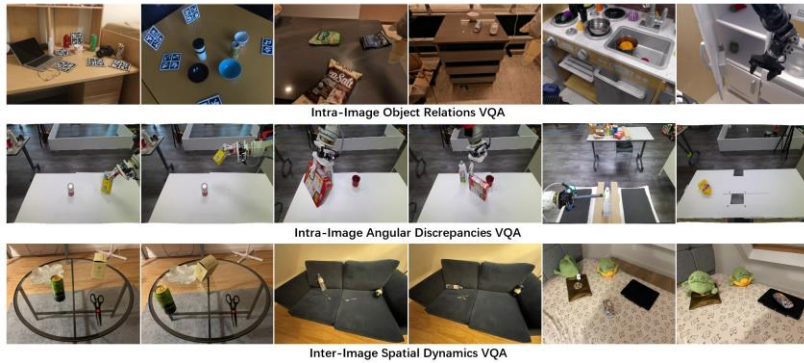
Q: What is the distance between the laptop and the camera?

Q: How do I stack the can on the bowl?

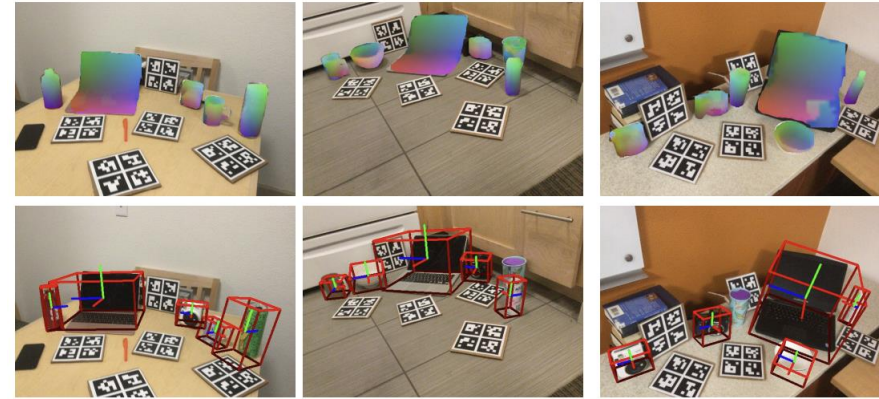




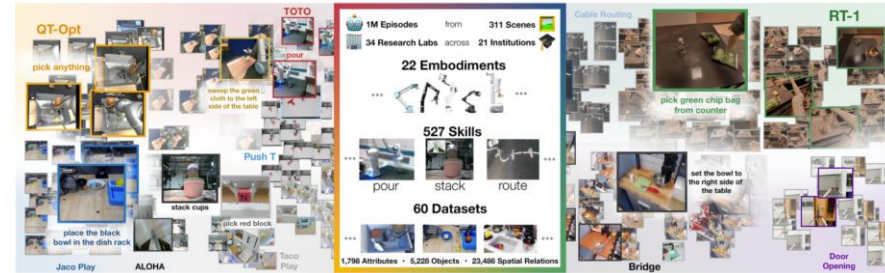
# Experiments



Handcrafted Datasets



NOCS (Wang et al., CVPR 2019)



RT-X (O'Neill et al., ICRA 2023)

Off-the-shelf Datasets

# Experiments: Various Forms of Spatial VQA



User



laOR-VQA

Is the laptop further to camera compared to the mug?

Measure the height difference between two white bowls.



laAD-VQA

Along which axis (W, D, H) is the box more tilted?

What is the relative angle between the box and the bottle?

Which objects rotate across the images?

How far does the milk box move?



lrSD-VQA



SpatialPIN (Ours)

The laptop is 75.0 cm from the camera, while the mug is 54.7 cm from the camera, making the laptop further.

The height difference between bowl 1 (12.0 cm) and bowl 2 (9.4 cm) is approximately 2.6 cm.

The tilt along the z-axis (Height) is the largest (about 0.798), indicating the box is more tilted along the Height (H) axis.

Using the dot product formula, the relative angle between the box and the bottle is approximately 47.41 degrees.

By object principle x-axes, both the knife and the milk rotate around their vertical axes, presenting a different face.

By observing the change in the 3D center, the milk box moves a distance of approximately 22.26 cm.



SpatialVLM

Yes, the laptop is further away from the camera compared to the mug.

The two white bowls on the table have different heights, with one being taller than the other.

The box is more tilted along the horizontal axis, as it is leaning on the table and appears to be at a slight angle.

The box is positioned at a diagonal angle to the bottle on the table.

The knife and the cutting board rotate across the image, while the box of apples remains stationary.

The milk box moves from the counter to the stove, indicating that it is being used for cooking or heating.

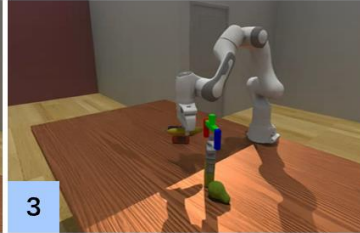
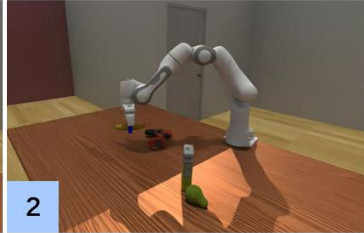
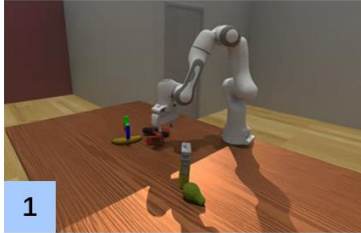


# Experiments: Robotics Pick-and-Stack



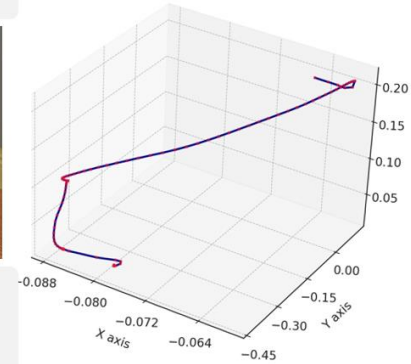
User

"Banana size: 19.8 cm x 6.9 cm x 3.66 cm (WxDxH). Cracker box size: 9.21 cm x 4.51 cm x 17.6 cm (WxDxH). Task: grasp and stack the banana on the cracker box. Determine the grasper's closing distance and the optimal drop height in cm, using the image and spatial information."



SpatialPIN

"The robot should grasp the banana at its thinnest side with a width of approximately 3.5 cm (just less than the banana's 3.66 cm) for a firm grip. It should then lift the banana to 22.26 cm (cracker box height of 17.6 cm plus banana height of 3.66 cm plus a 1 cm margin) and drop it from 22.16 cm to stack it securely on the cracker box. Using RRT\*, the planned 3D trajectory is ..."



# Experiments: Discovering and Planning Robotics Task Trajectories from a Single Image

Input



Output



"I can lift the **chair** upright to position it in front of the table."

Input



Output

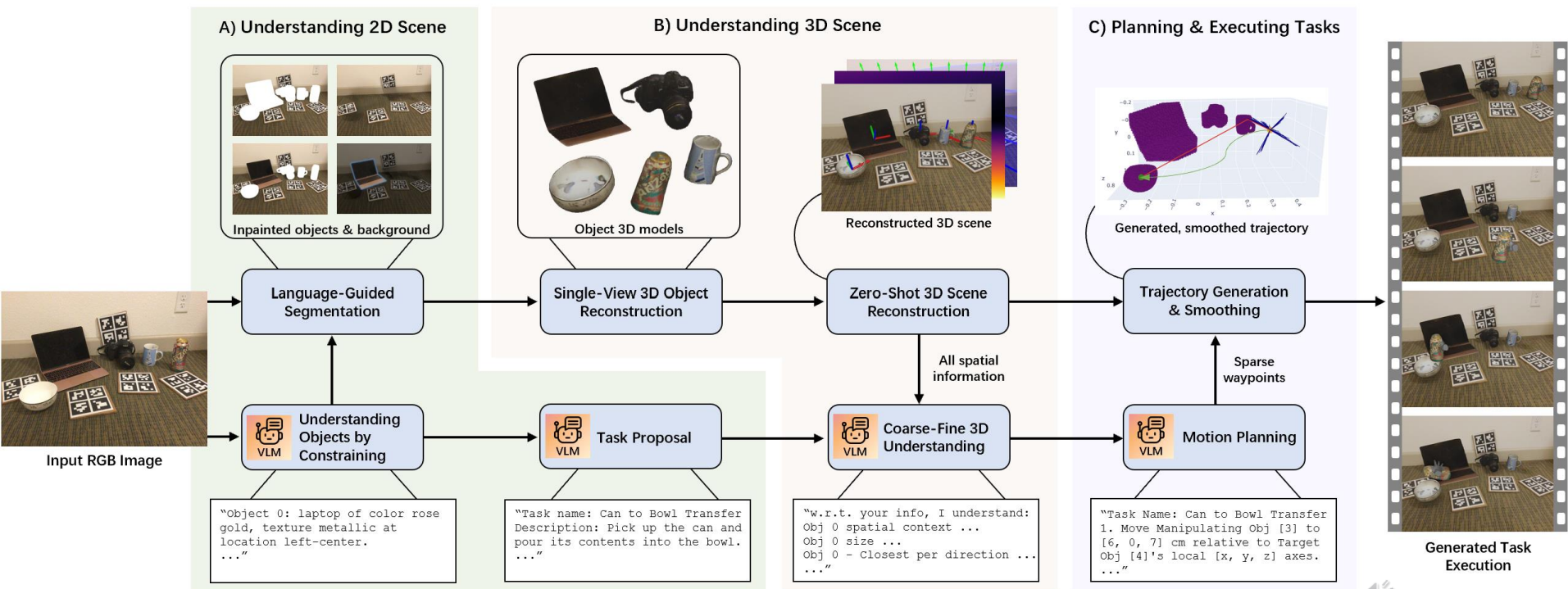


"I can lift the **container** and deposit it into the trash bin."





# Experiments: Discovering and Planning Robotics Task Trajectories from a Single Image



# Conclusion

- We see our work as a step towards equipping VLMs with more generalized spatial reasoning capacities.
- Limitation: Inference speed.





# Thanks for Listening

- Project page: [https://dannymcy.github.io/zeroshot\\_task\\_hallucination/](https://dannymcy.github.io/zeroshot_task_hallucination/)
- Paper: <https://arxiv.org/abs/2403.13438>

