Believing is Seeing

Unobserved Object Detection using Generative Models

CVPR Nashville JUNE 11-15, 2025

Australian National University

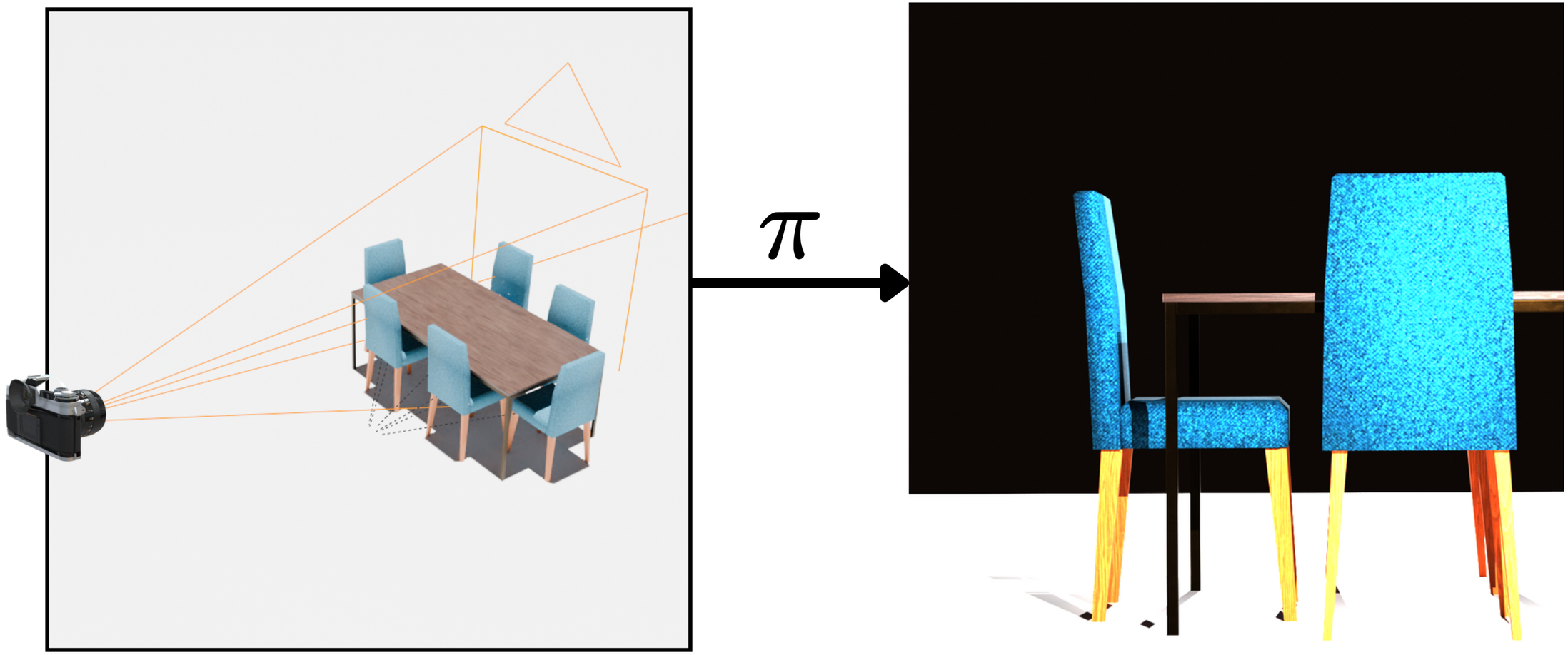
Subhransu S. Bhattacharjee

Dylan Campbell

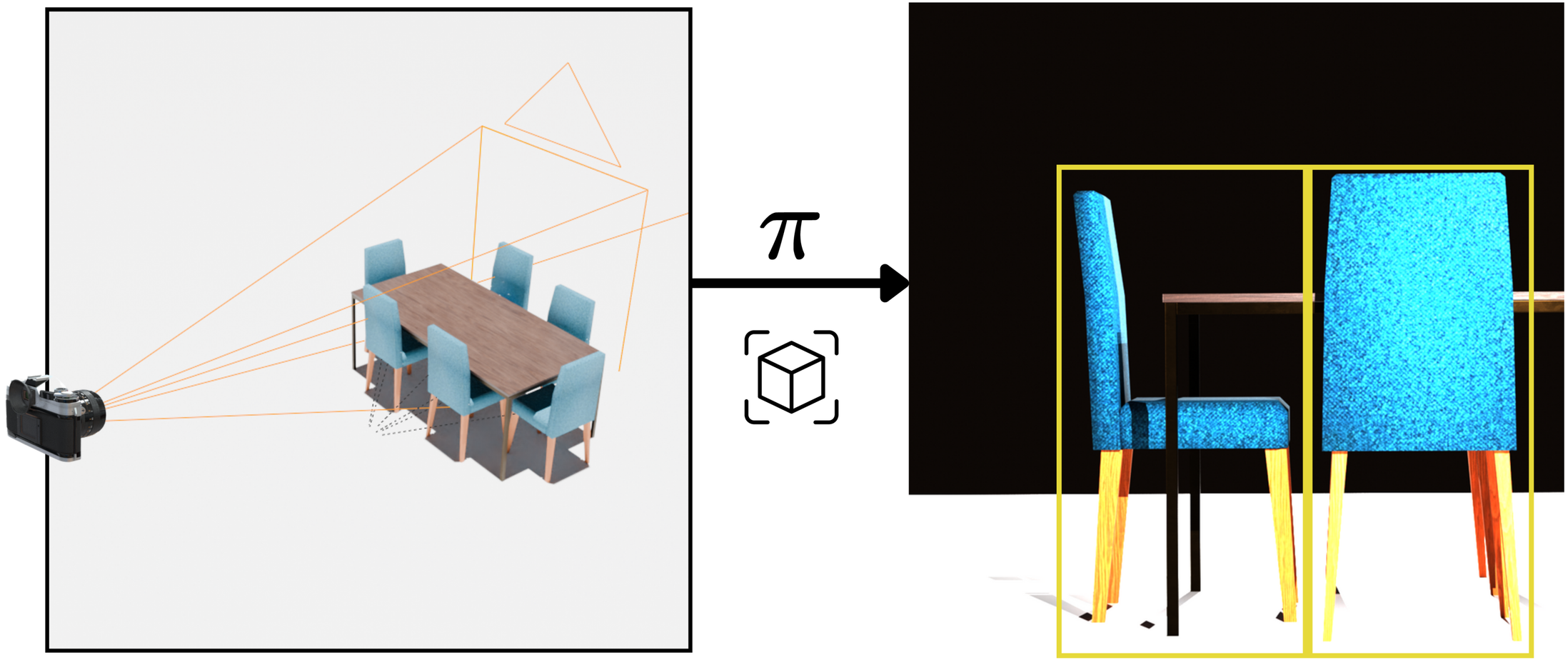
Rahul Shome

School of Computing, The Australian National University, Canberra

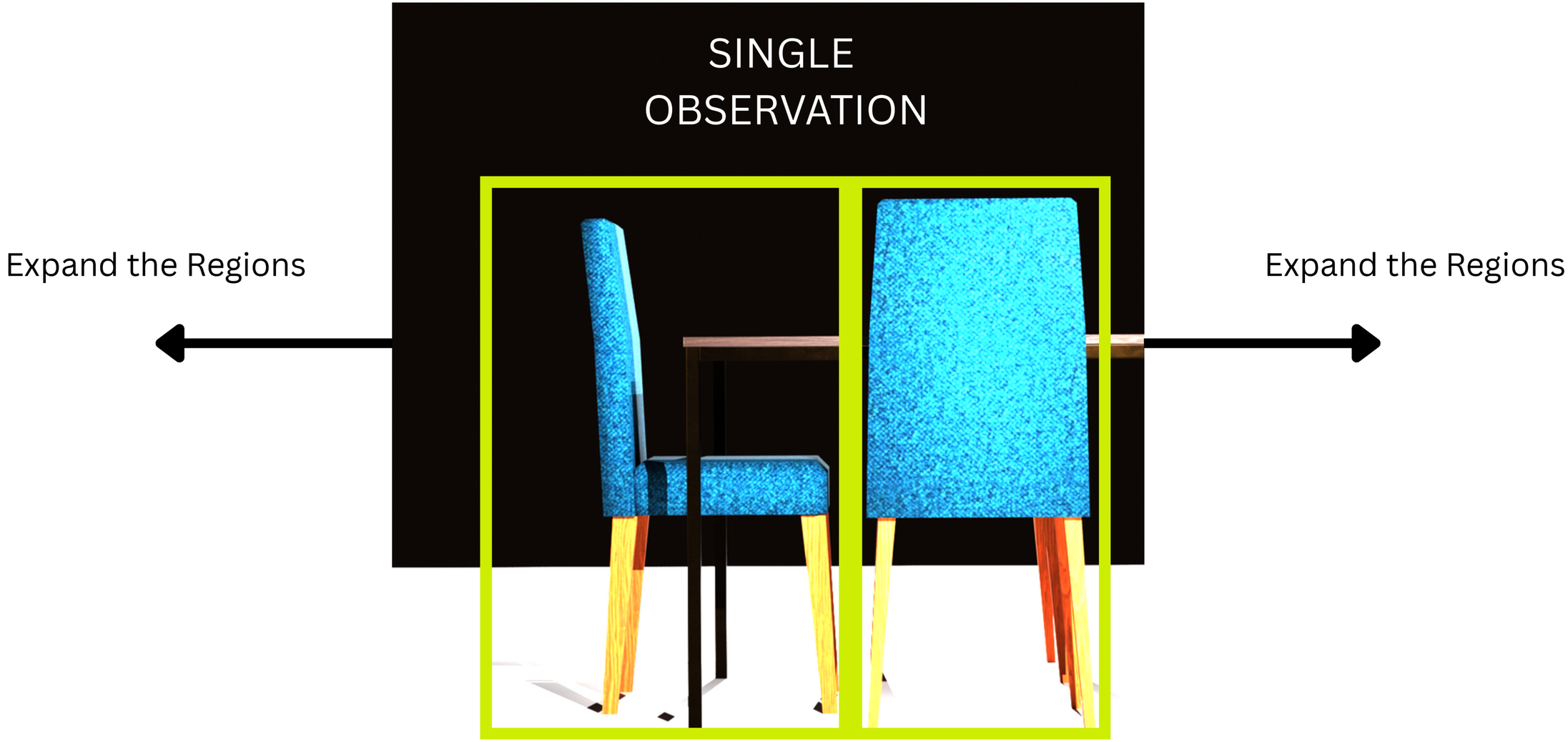
Background: Imaging as a Projection



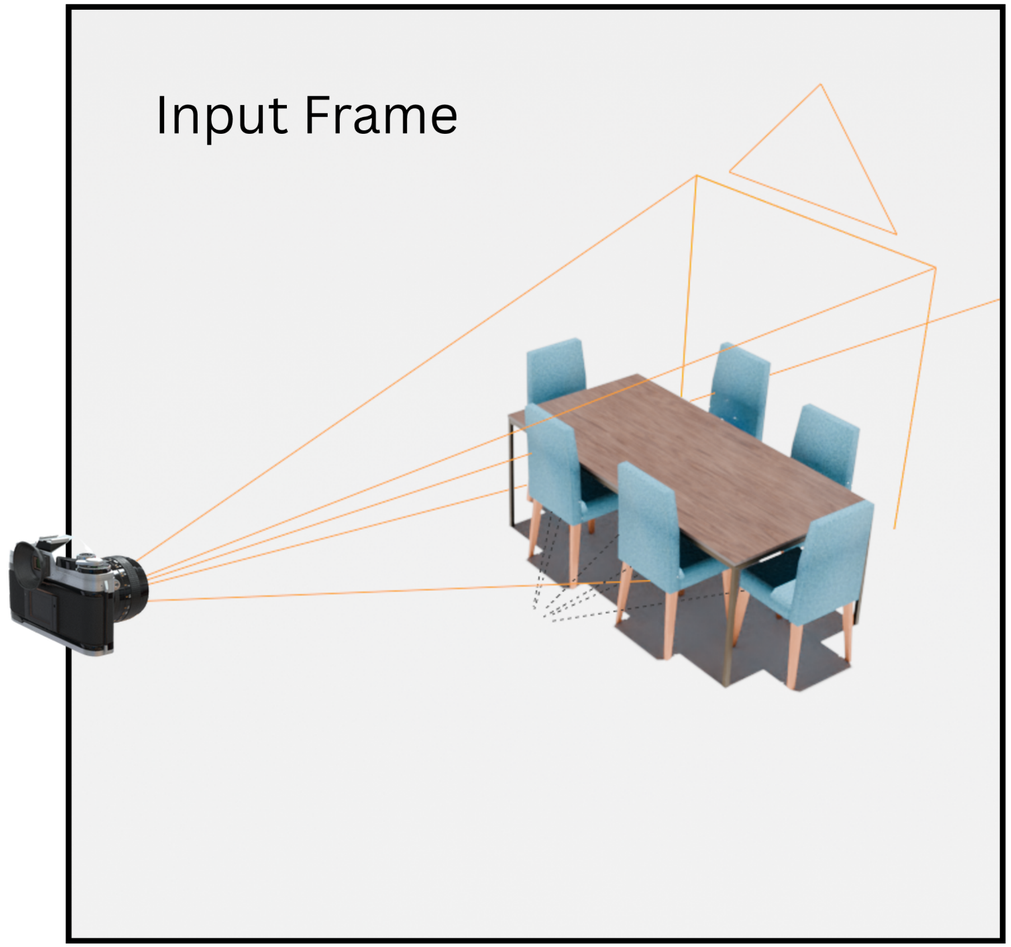
Background: Object Detection

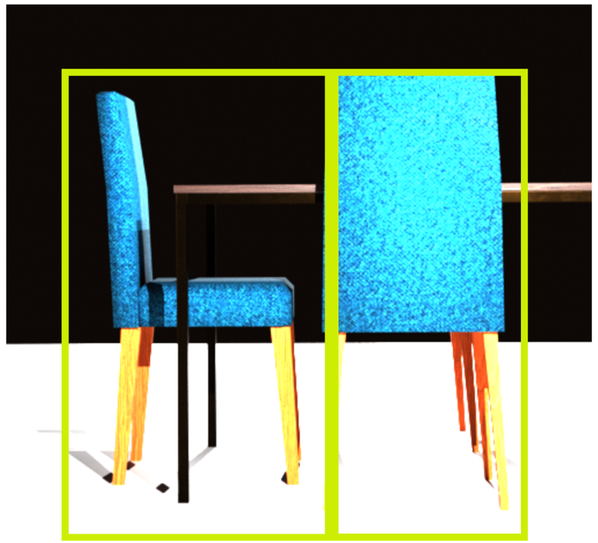


Motivation: What lies beyond this frame?

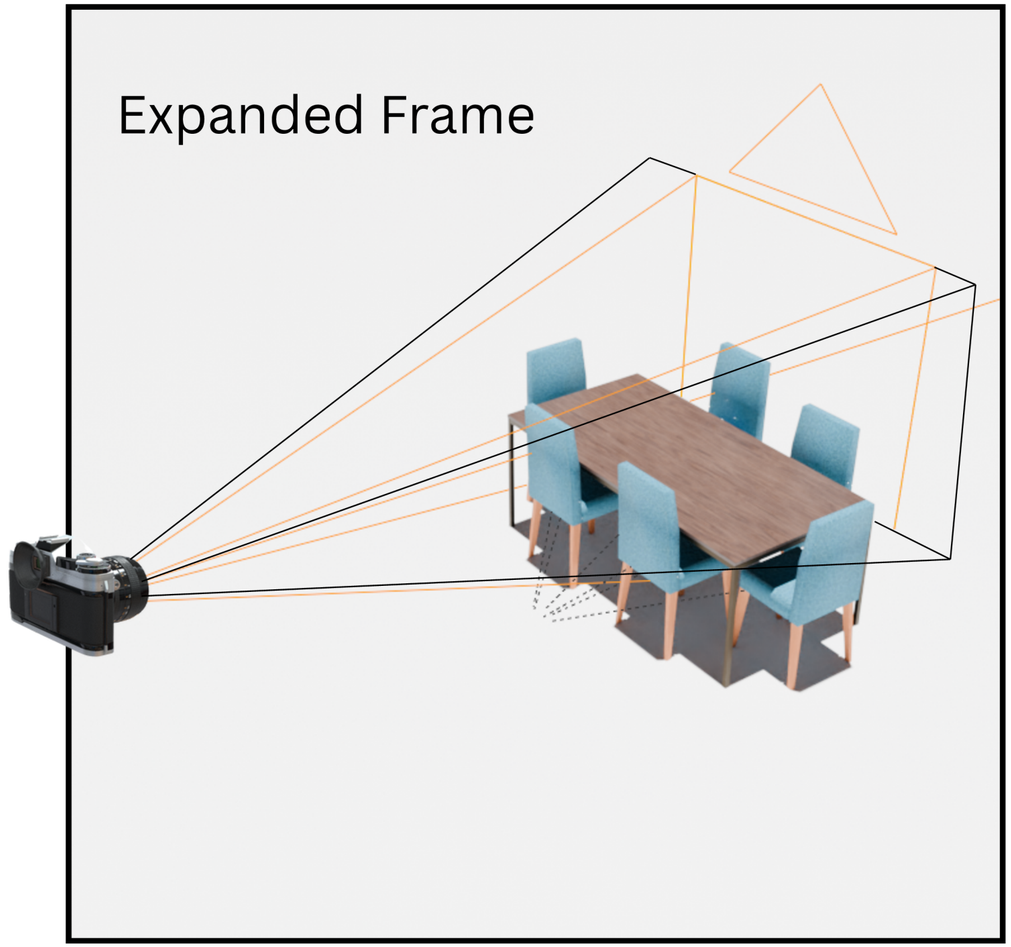


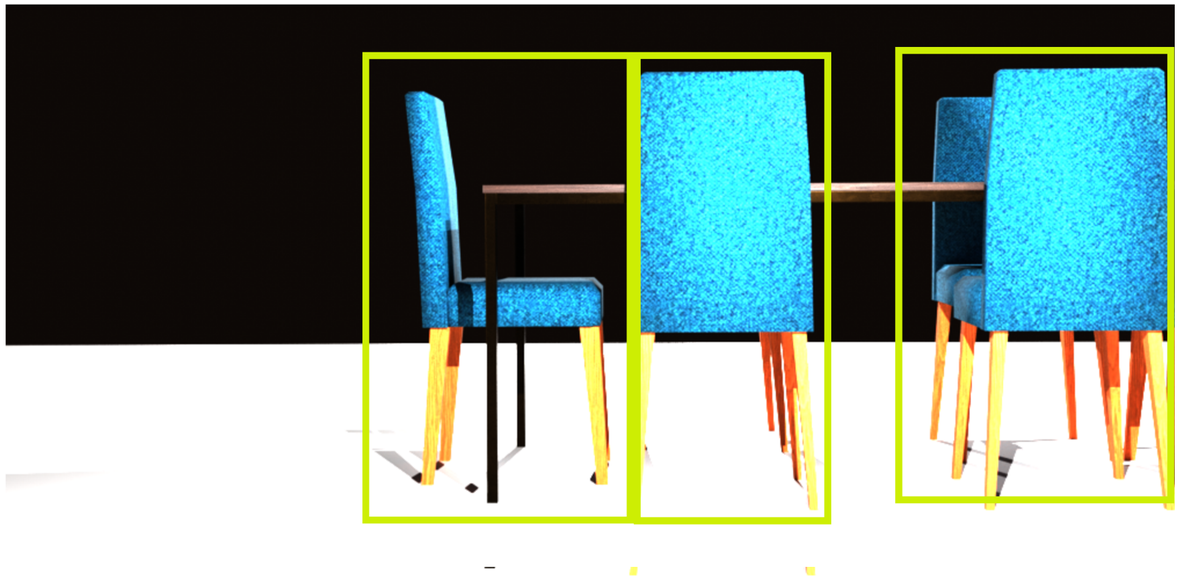
Task: Find chairs



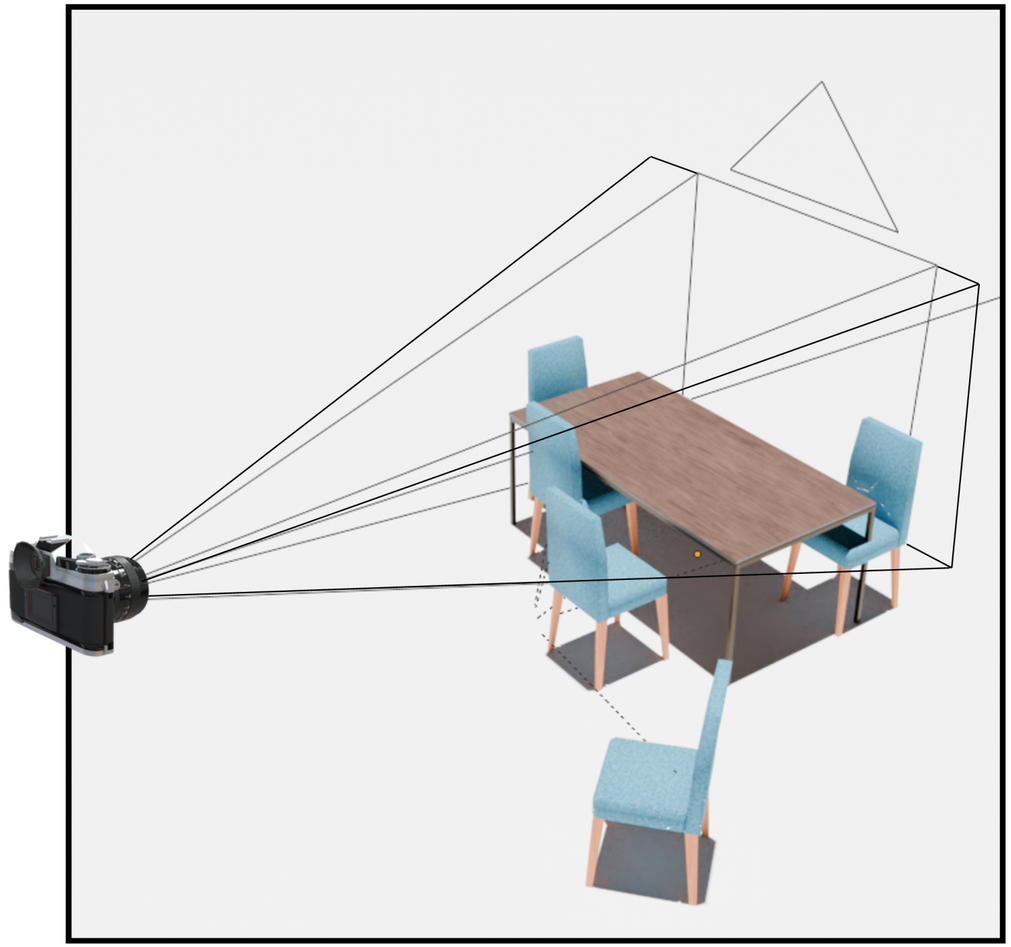


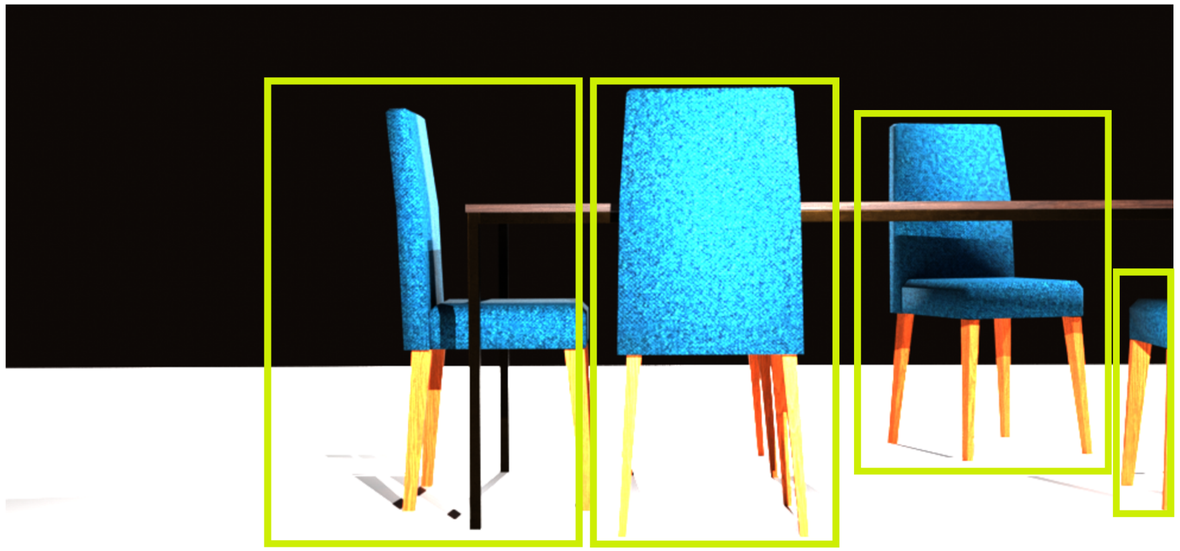
Task: Find chairs





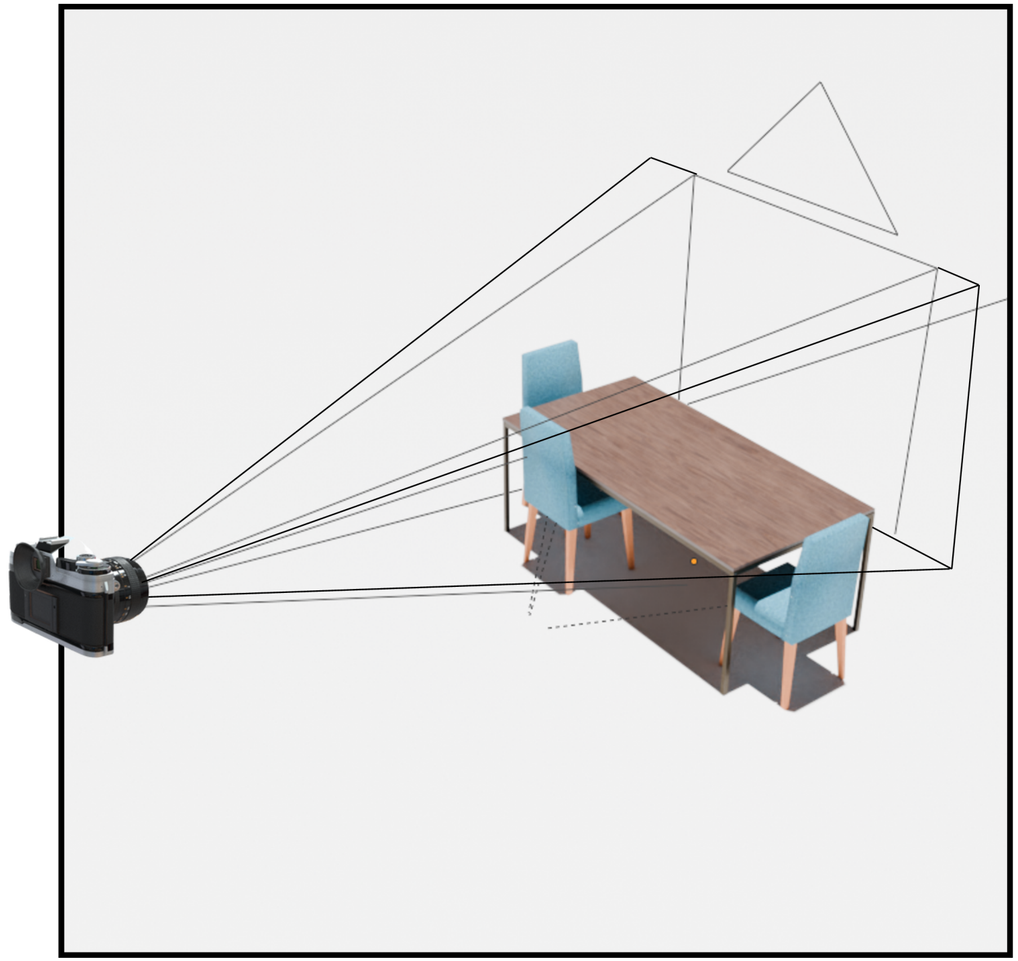
Task: Find chairs

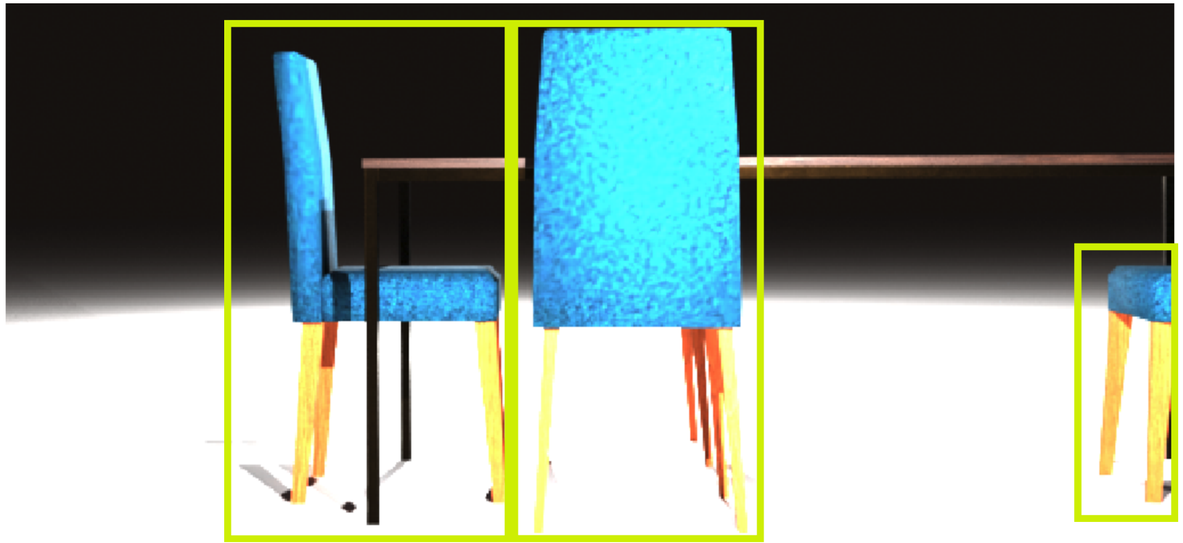




Possible Variations

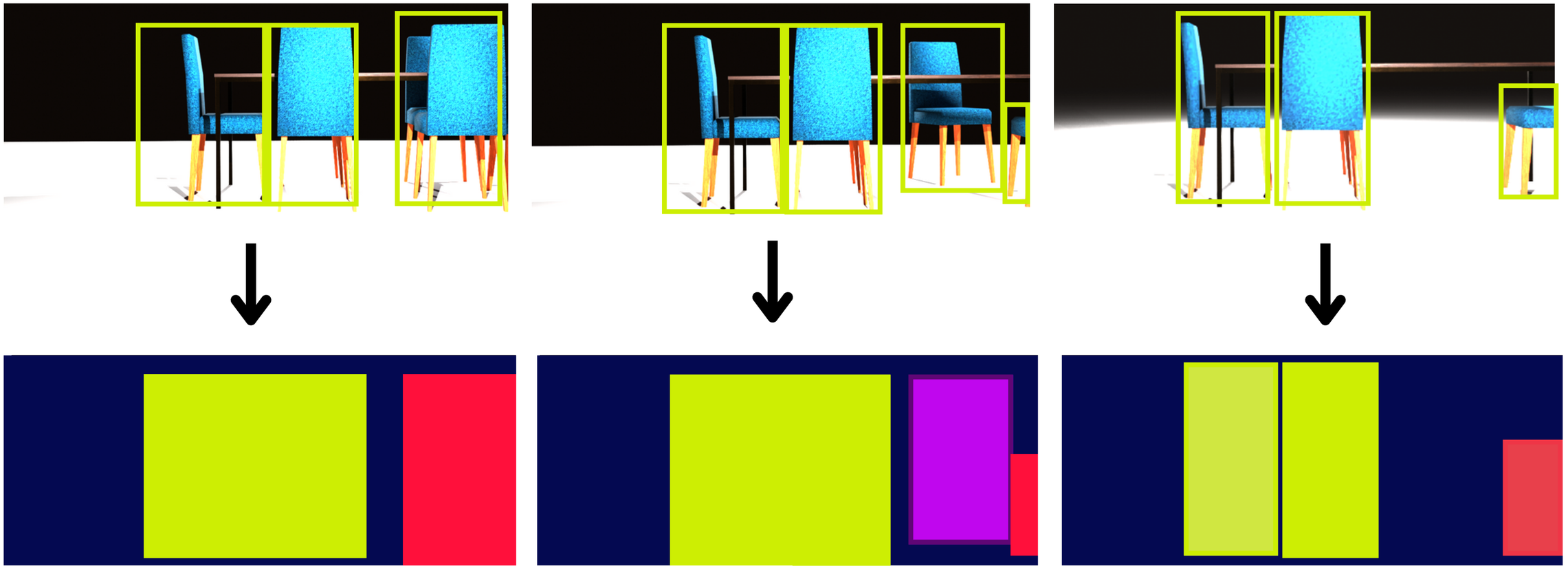
Task: Find chairs





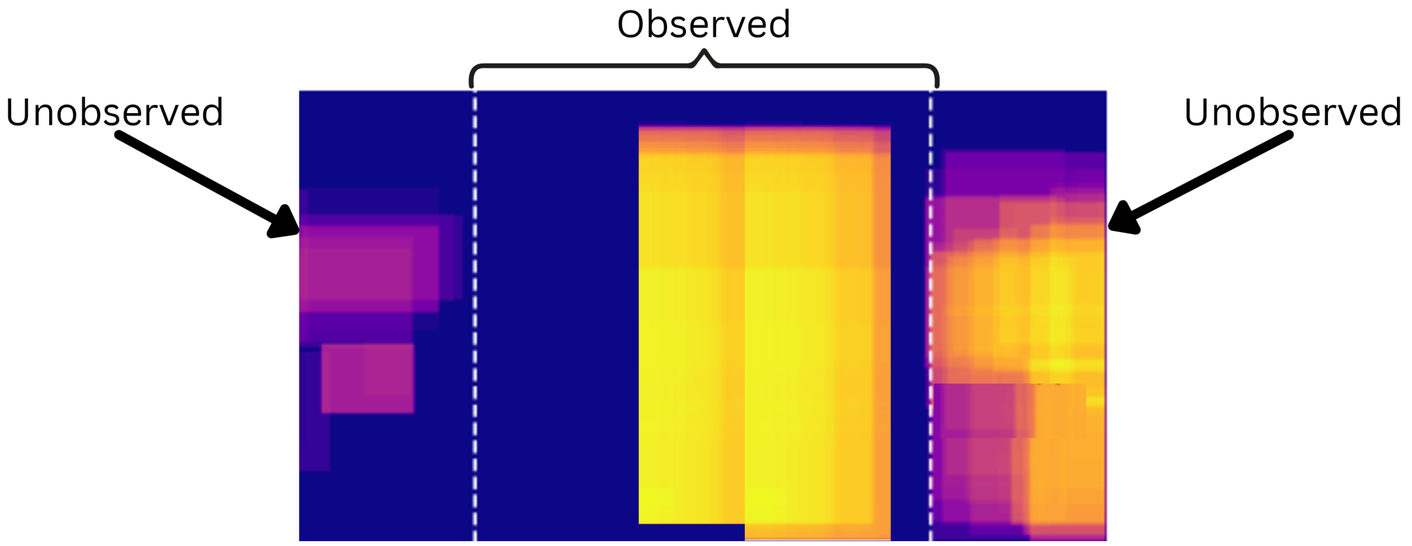
Possible Variations

Task: Unobserved Object Detection



Bounding box representations of the Object Detector

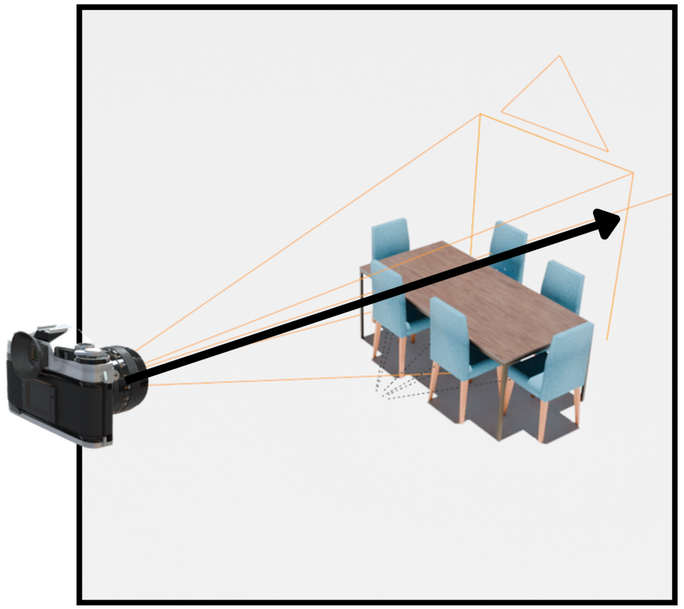
Task: Unobserved Object Detection



Aggregate predictions into a spatio-semantic distribution

Task: Unobserved Object Detection

Definition: Detect objects that are in the vicinity of the camera but not in the visible camera frustum.



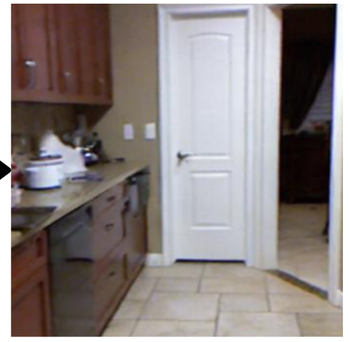
Chair occluded from the field of view

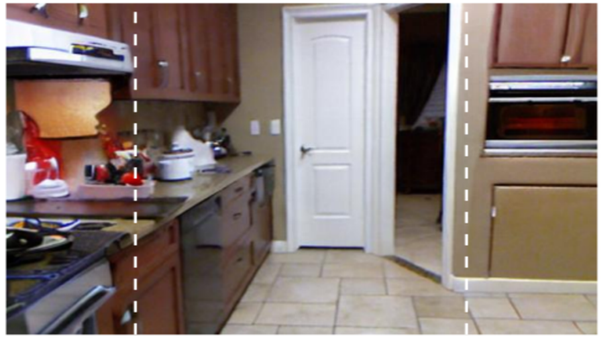
Sampling using Generative Models

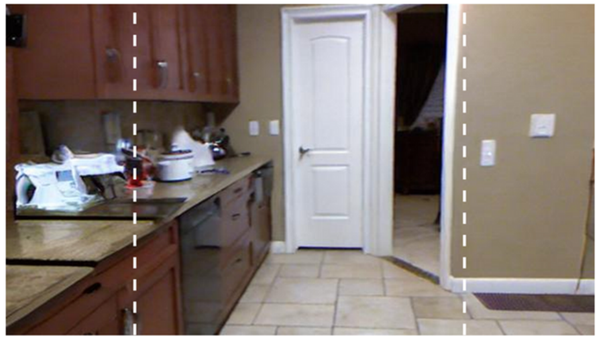
3D Generation

INPUT

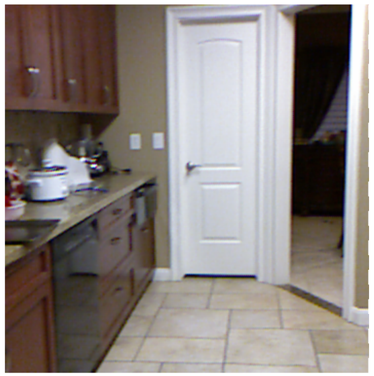
2D Generation







Pipeline: Vision Language Models



Input image

VLM Prompts

Could a **sink** *be 140 pixels to the left/right of the image?*

N samples

Right

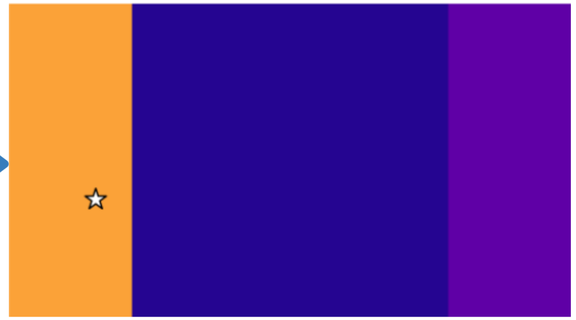
Left

Yes/
No

Yes/
No

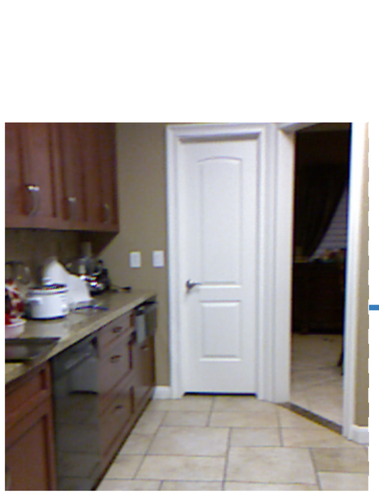
Aggregate

1. ChatGPT-4o
2. Claude-3.5-Sonnet
3. LLaVA-v1.6
4. Gemini-1.5-Ultra



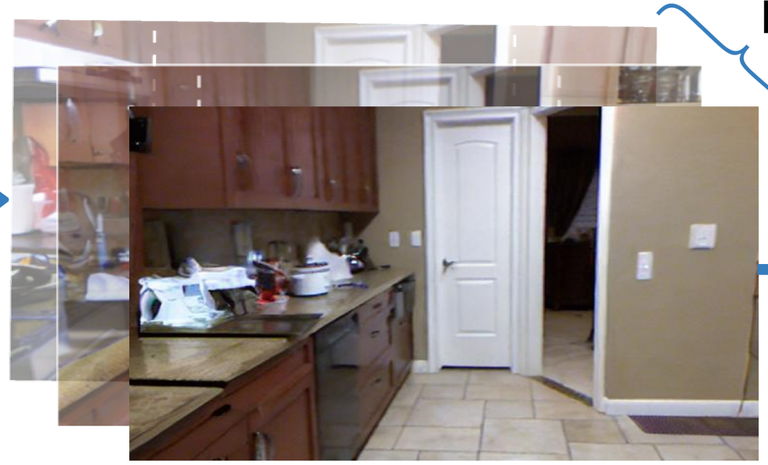
- “Hello GPT-4o” (OpenAI)
- Claude-3.5-Sonnet (Anthropic)
- Gemini-1.5-Ultra (Google)
- H. Liu et al., “LLaVA-1.6:”, NeurIPS, 2023

Pipeline: 2D Diffusion Model



Input image

2D Diffusion Outpainting

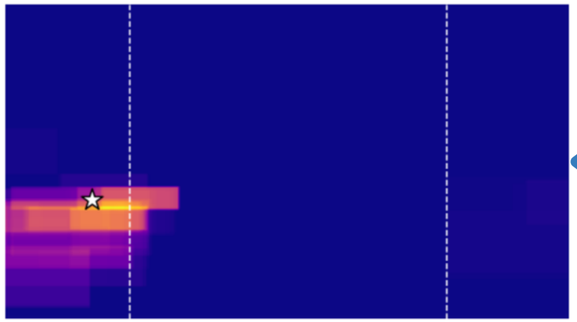


N samples

sink

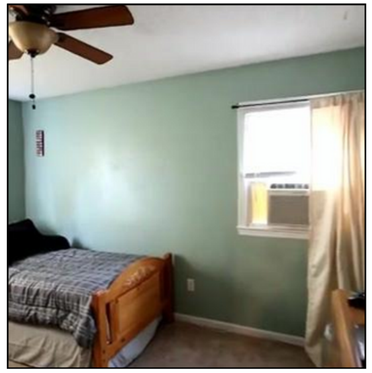
Object Detector: YOLOv8x

Aggregate



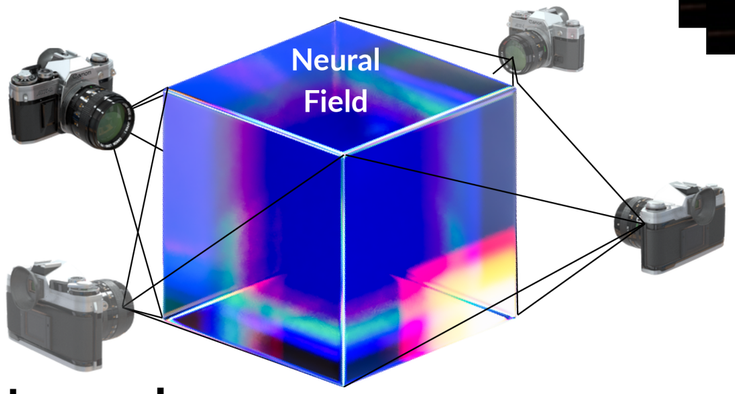
- D. Podell et al., SDXL, ICLR, 2024
- Ultralytics, YOLOv8, 2023

Pipeline: 3D Diffusion Model

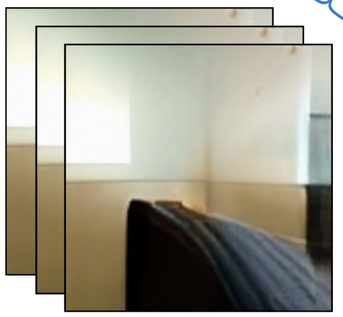


Input image
Pose 1

Diffusion with
Forward Model

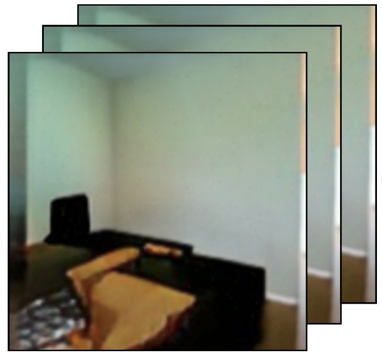


N samples



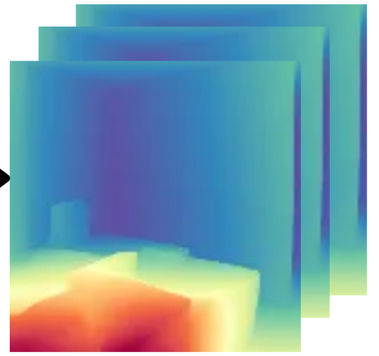
Generated images
from Pose 4

Generated images
from Pose 3



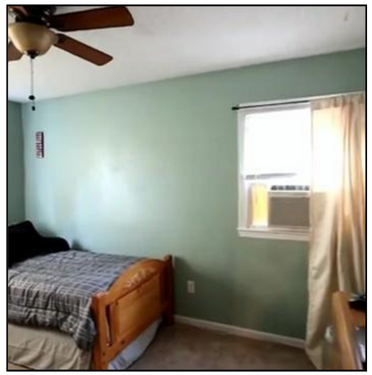
Generated images
from Pose 2

Depth Map



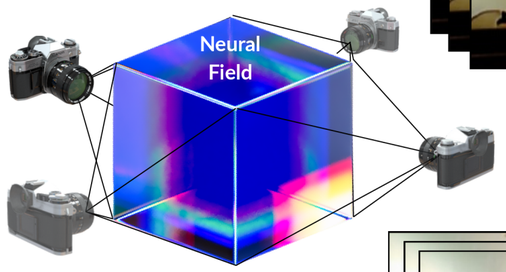
- A. Tewari et al., DFM, NeuIPS, 2023
- Ultralytics, YOLOv8, 2023

Pipeline: 3D Diffusion Model

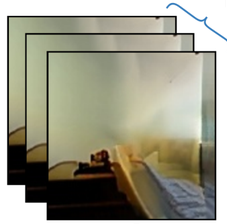


Input image
Pose 1

3D Generation



N samples

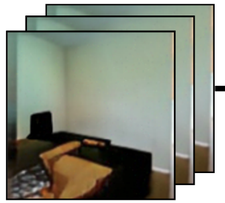


Generated images
from Pose 3

TV

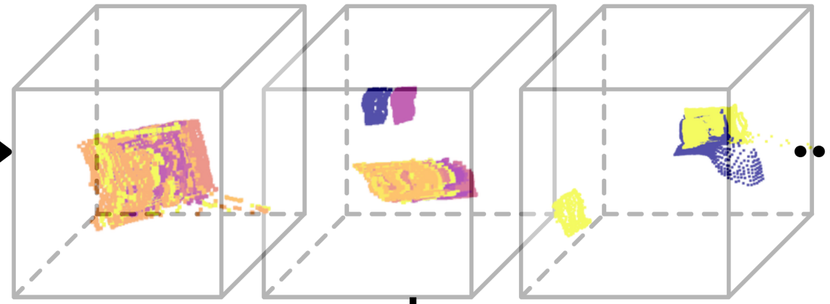
Object Detector:
YOLOv8x

Generated images
from Pose 4



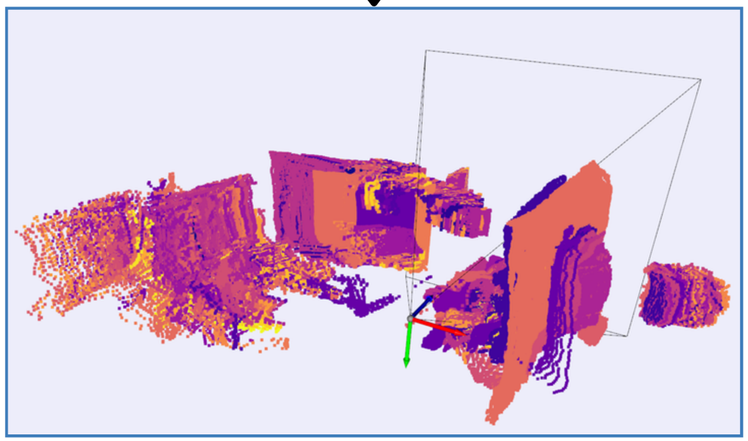
Generated images
from Pose 2

3D back-projection
of object detections

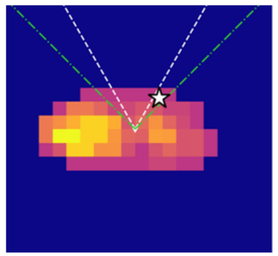


K poses
X
N samples

Aggregate

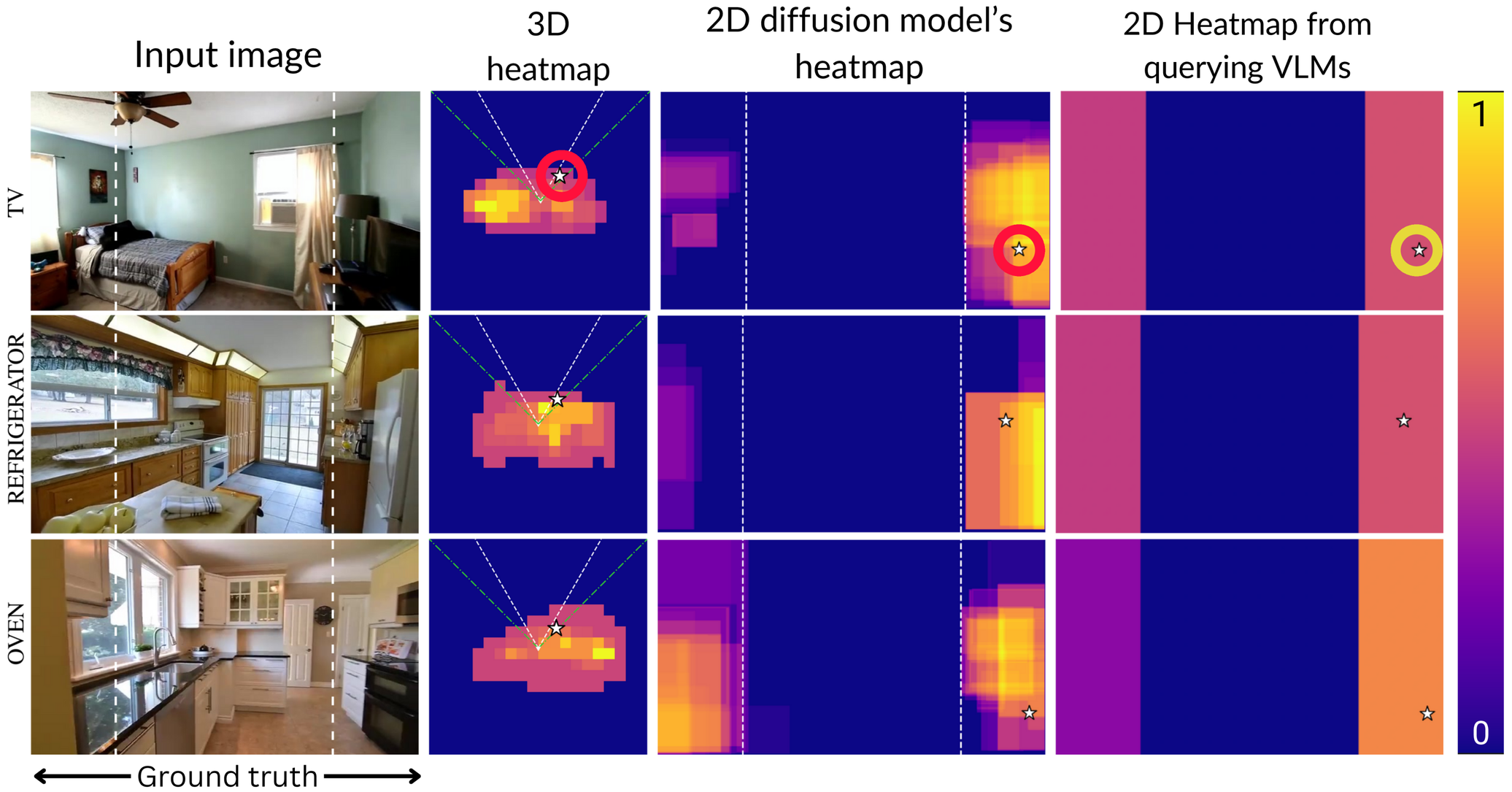


Voxelized top-down view

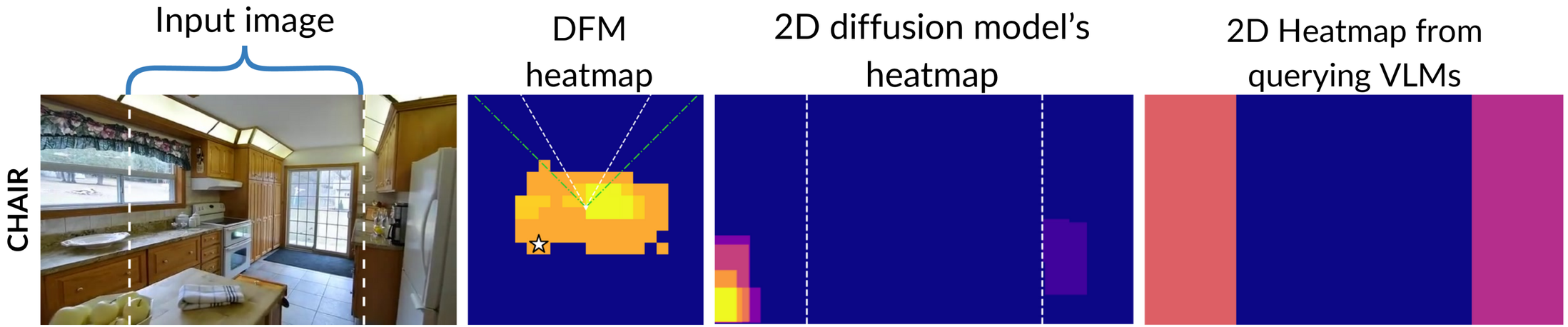


- A. Tewari et al., DFM, NeuIPS, 2023
- Ultralytics, YOLOv8, 2023

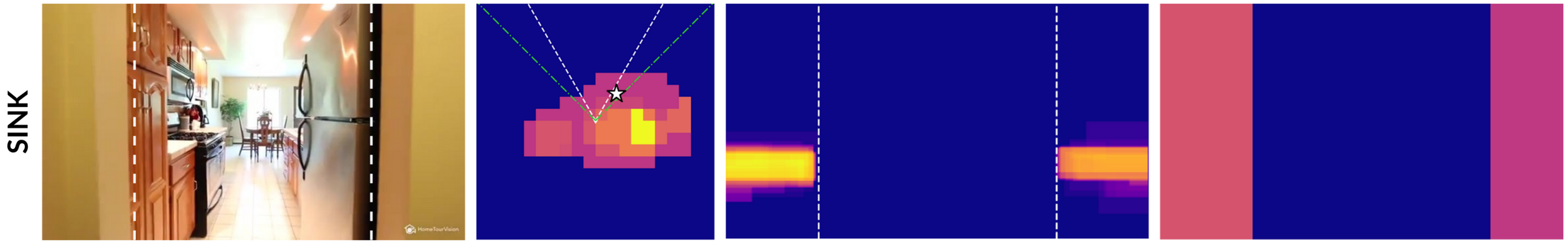
Qualitative Results: Out of Frame



Qualitative Results: Not in 2D



Ground truth chair behind the camera



Ground truth sink occluded behind the refrigerator

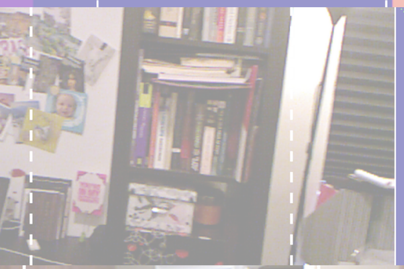
Summary

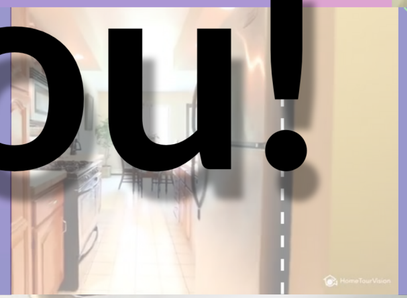
Across the 3 pipelines, tested in terms of accuracy and diversity in *2D, 2.5D and 3D* settings on the **RealEstate10k and NYUDepthv2** datasets, we observe:

- In 3D settings, DFM leads across all metrics, achieving a **0% false negative rate**
- SDXL outpainting with semantic prompts outperforms other pipelines in 2D and 2.5D settings
- Certain VLMs exhibit competitive region-wise predictions in 2D, with 2% **false negative rates**

Zhou et al., Stereomagnification, SIGGRAPH, 2017
Silberman et al., NYU Depth Dataset V2, ECCV, 2012





Thank You!