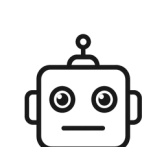


Our Contributions

1. SPLIT is a harness that reconstructs a metric 3-D point cloud per video and gives a planner VLM perception tools to query it. SPLIT achieves the best training-free score on VSIBench and the best published score on VSTIBench and ReVSI.
2. Replacing the tools' estimates with ground-truth measurements lifts SPLIT from 66.9 / 78.3 / 71.5 to 88.8 / 96.7 / 91.0 on VSIBench, VSTIBench, and ReVSI. Perception, not reasoning, limits current VLMs.

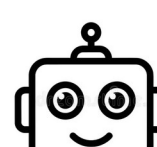
Sample Question

Where is the toilet relative to you when you face the bed from the table?



Baseline

Baseline: front-left ✗



SPLIT

SPLIT: back-left ✓

Pipeline

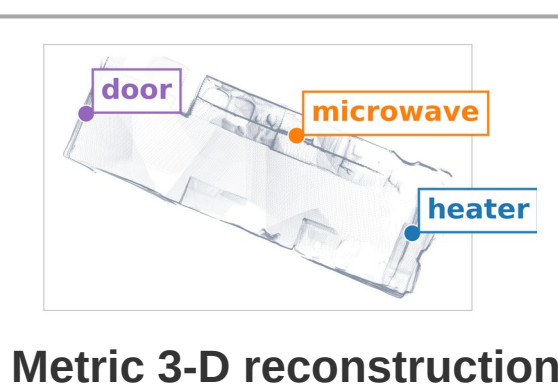
Input video



Tools

Multi-view 3-D reconstruction
(e.g., G3T + UniK3D)2-D detection
(e.g., SAM3)Geometric math
(Python, e.g., $a \times b$)

If I am standing by the heater and facing the microwave, is the door to my front-left, front-right, back-left, or back-right?



Metric 3-D reconstruction

Planner VLM
Gemini 3.1 Pro

Walkthrough of a Question

Question

If I am standing by the heater and facing the microwave, is the door to my front-left, front-right, back-left, or back-right?

Input Video Frames



1 Write a plan

1. Find frames with each object.
2. Localize each object in 2D.
3. Lift 2D localizations to 3D centroids.
4. Compute the relative direction.

2 Find frames with each object

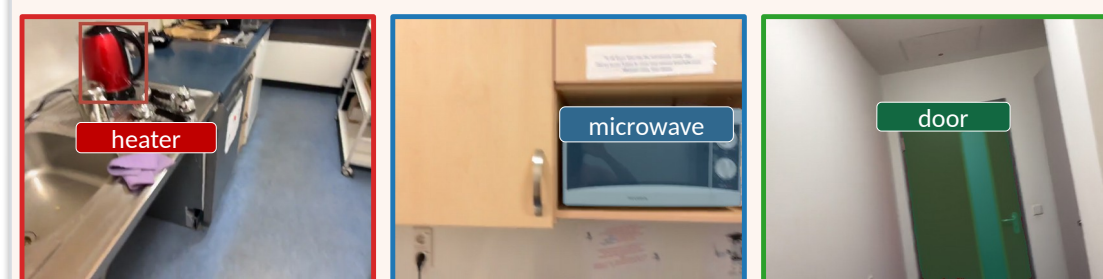
```
h_frames = find_frames_with_object("heater")
m_frames = find_frames_with_object("microwave")
d_frames = find_frames_with_object("door")
```

```
h_frames = [1, 11, 14, 23, 30]
m_frames = [1, 16, 17, 28, 31]
d_frames = [1, 5, 9, 28, 32]
```

3 Ground each object in 2D

```
h_mask = predict_2d_segmentation_masks(
    frame=14, object="heater")
m_mask = predict_2d_segmentation_masks(
    frame=28, object="microwave")
d_mask = predict_2d_segmentation_masks(
    frame=9, object="door")
...
m_mask = None # heater mask empty
# I should reground using the VLM 2D bounding box
h_bbox = predict_2d_bounding_box(
    frame=14, object="heater")
```

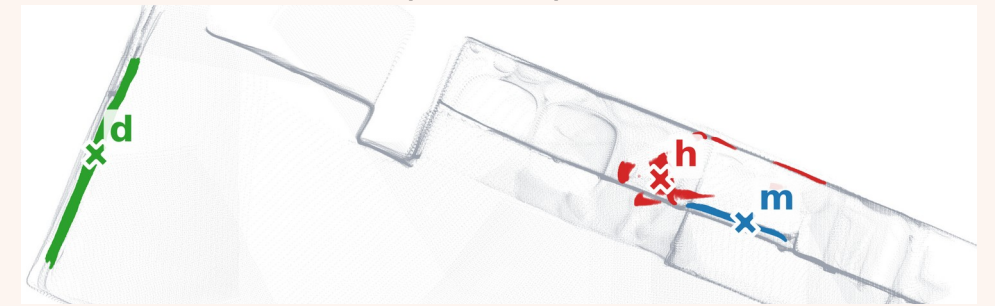
SAM3 heater mask: EMPTY → fall back to a VLM bounding box ✗



4 Lift 2D localizations to 3D centroids

```
h_centroid = get_3d_points_in_bbox(h_bbox).centroid
m_centroid = get_3d_points_in_mask(m_mask).centroid
d_centroid = get_3d_points_in_mask(d_mask).centroid
...
h_centroid = [-0.50, 1.48, -0.61]
m_centroid = [-0.25, 1.97, 0.12]
d_centroid = [-0.64, -1.78, 0.39]
```

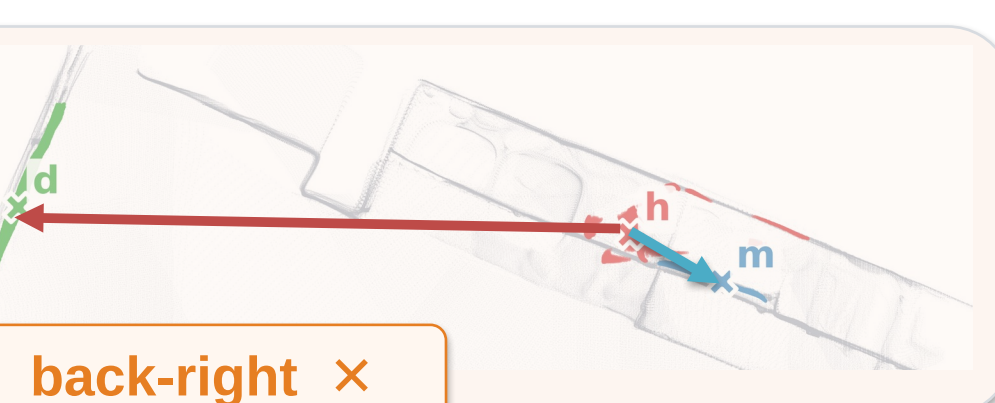
Lifted 3D point map (BEV)



5 Compute provisional direction

```
forward = normalize(m_centroid[:2] - h_centroid[:2])
to_door = normalize(d_centroid[:2] - h_centroid[:2])

dot = np.dot(forward, to_door)
cross = np.cross(forward, to_door)
...
dot = -0.91 → back
cross = -0.42 → right
```



back-right ✗

6 Cross-check heater localization across frames

```
h_bbox_f11 = predict_2d_bbox(frame=11, object="heater")
h_bbox_f1 = predict_2d_bbox(frame=1, object="heater")

h_centroid_f11 = get_3d_points_in_bbox(h_bbox_f11).centroid
h_centroid_f1 = get_3d_points_in_bbox(h_bbox_f1).centroid
...
h_centroid_f11 = [1.48, 4.01, -1.37]
h_centroid_f1 = [1.48, 4.03, -1.41]
h_centroid_f14 = [-0.50, 1.48, -0.61]
# h_centroid_f14 is an outlier, so let's ignore it
h_consensus = [1.48, 4.02, -1.39]
```

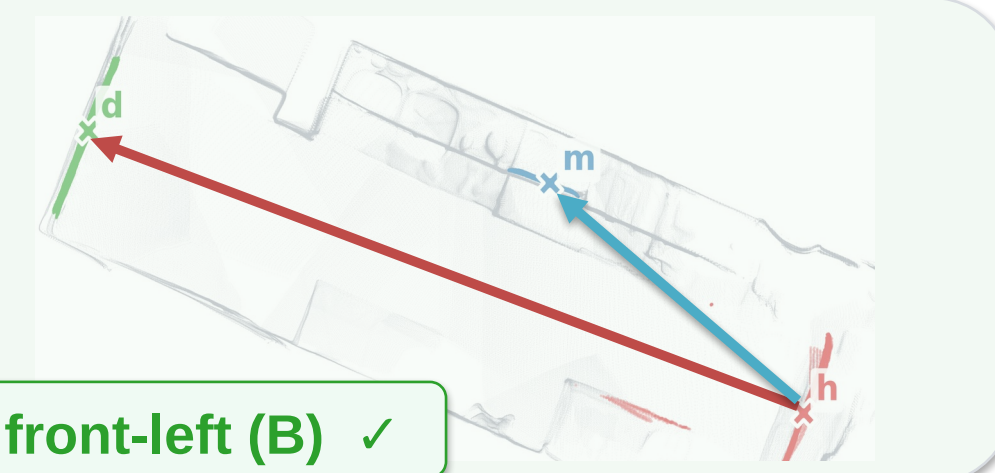


7 Recompute with consensus heater position

```
forward = normalize(m_centroid[:2] - h_consensus[:2])
to_door = normalize(d_centroid[:2] - h_consensus[:2])

dot = np.dot(forward, to_door)
cross = np.cross(forward, to_door)
...
dot = 0.94 → front
cross = 0.34 → left

Answer: front-left
```

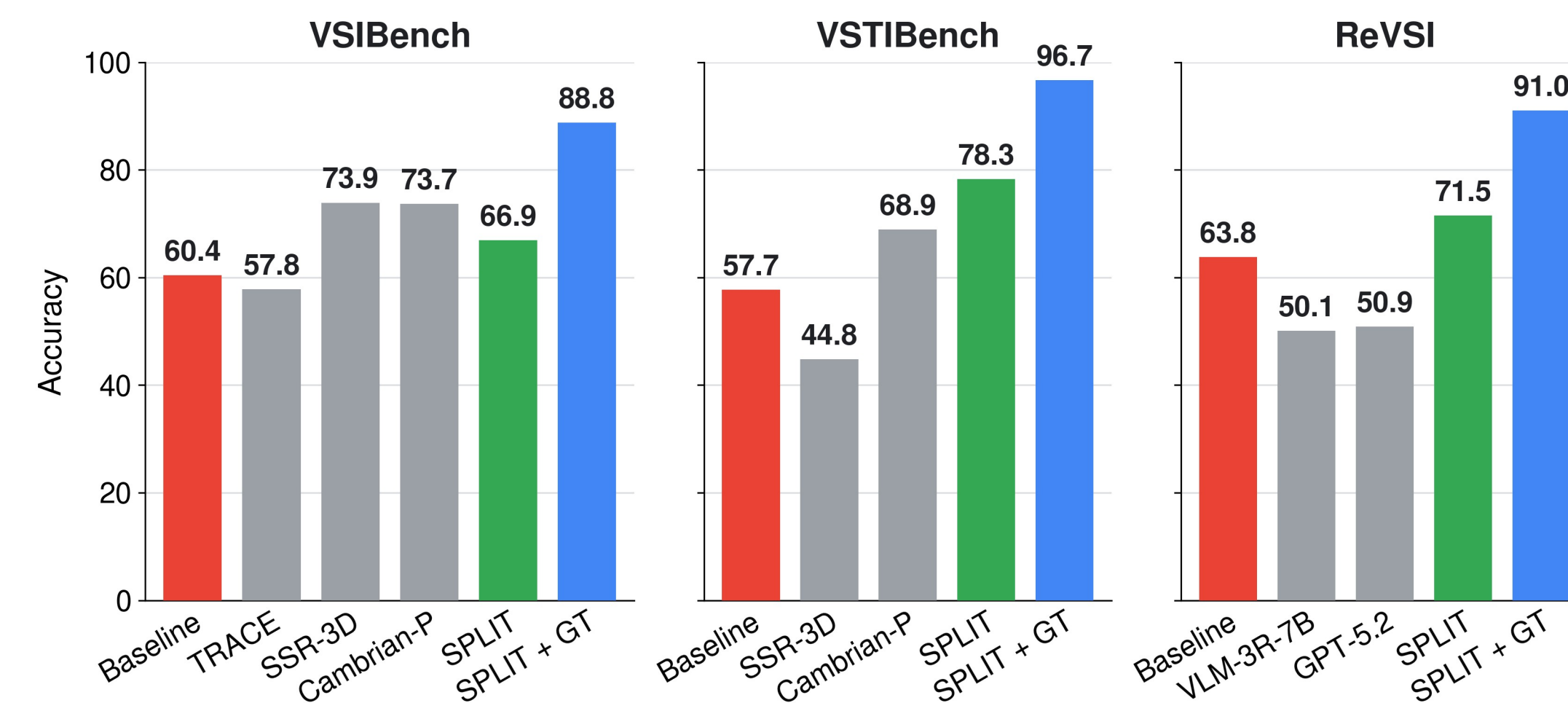


front-left (B) ✓

SPLIT Improves Video Spatial Reasoning

Baseline (Gemini 3.1 Pro) SPLIT (Ours)
Published models SPLIT (Ours) + GT

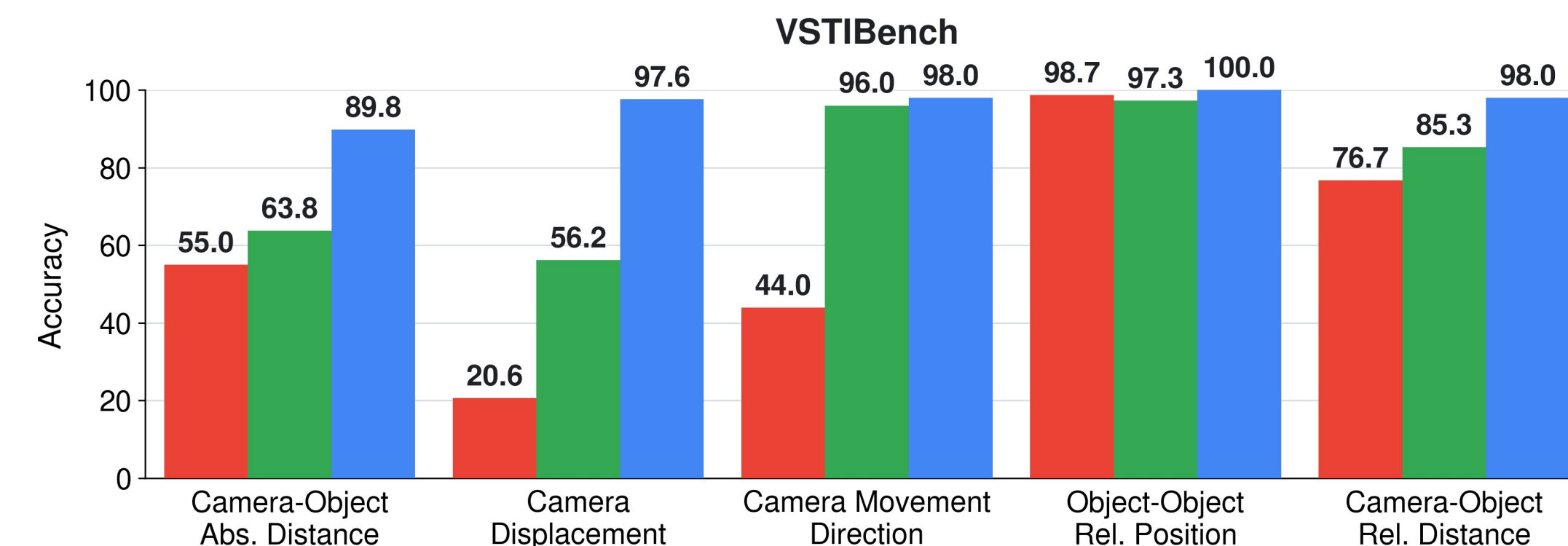
SPLIT (Ours) always uses Gemini 3.1 Pro as the planner VLM.



Ground Truth Closes the Gap in Every Category

Baseline (Gemini 3.1 Pro) SPLIT (Ours) SPLIT (Ours) + GT

SPLIT (Ours) always uses Gemini 3.1 Pro as the planner VLM.



Conclusion

With ground-truth measurements in place of the tools' estimates, SPLIT rises from 78.3 to 96.7 on VSTIBench: perception, not reasoning, is the bottleneck.

SPLIT improves VSIBench, VSTIBench, and ReVSI by 6.5, 20.6, and 7.7 points over the same planner VLM without tools.