SPATIAL REASONING — BENCHMARK RELEASE

Spatial reasoning? Frontier models have some *distance* to go.

We evaluated leading VLMs on 3D spatial reasoning tasks derived from a production Level 4 AV stack. The best model scored 50% — barely above the 25% random baseline. When we checked the reasoning, it dropped to 10%.

50%	10%	25%	10
Best Accuracy	**Best Reasoning**	**Random Baseline**	**Categories**
Sonnet 4.6	Opus 4.7 rationale	4-way MC	Depth, heading, lateral

THE BENCHMARK

Grounded in LiDAR-fused 3D annotations

Every answer comes from calibrated sensor data — sub-meter distances, yaw in radians, verified object labels. Sourced from an established Level 4 AV partner's production perception stack.

Each sample pairs an annotated front-camera image — numbered bounding boxes on detected objects — with a 4-way multiple-choice question. The model must infer depth, lateral position, heading, and object type from a single 2D frame. Questions are programmatically generated from production 3D scene graphs, ensuring metric precision, consistency, and arbitrary scalability.

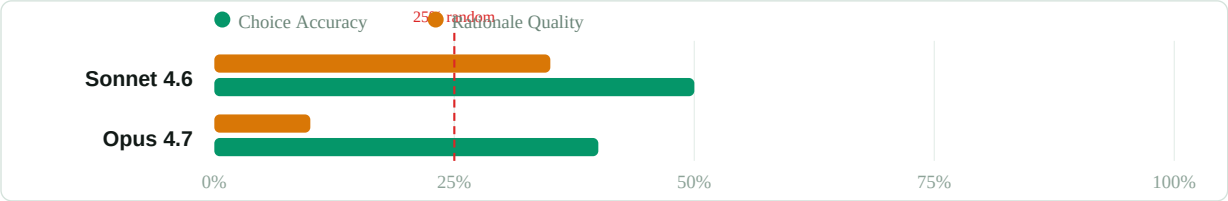
Highway, urban, nighttime, construction. Multiple platforms, 2022–2025.

Category	Task	Ground Truth
`order_closest`	Rank 4 objects by distance from ego	Metric distances (m)
`pick_closer`	Which of two objects is nearer?	Distance pair (m)
`identify_distance_long`	Classify distance: Close / Medium / Far	Metric distance (m)
`identify_nearest_ahead`	Nearest object along the forward axis	Forward projection (m)
`order_leftmost`	Rank 4 objects left-to-right in 3D	Lateral offset (m)
`identify_rightmost`	Furthest-right object in 3D?	Lateral offset (m)
`identify_position`	Classify position (ahead-left, etc.)	Forward + lateral (m)
`identify_heading`	Object heading in clock notation	Yaw angle (rad)
`relative_heading`	Same, opposite, or perpendicular heading?	Yaw diff (deg)
`identify_type`	Classify the object type	3D annotation label

EVAL RESULTS

Barely above random — and the reasoning is worse

Two scoring layers: choice accuracy (right letter?) and rationale match (right reasoning, verified by a judge model against metric ground truth).

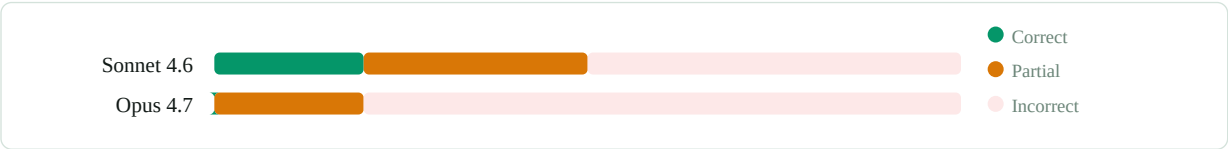


Per-Category Results

Category	Son. Ch	Son. Ra	Opus Ch	Opus Ra
Dist. Order	●	●	●	●
Nearest Ahead	●	●	●	●
Rightmost	●	●	●	●
Dist. Bin	●	●	●	●
Position	●	●	●	●
Pick Closer	●	●	●	●
Lateral Order	●	●	●	●
Obj. Type	●	●	●	●
Heading	●	●	●	●
Rel. Heading	●	●	●	●

● Pass ● Partial ● Fail

Rationale Quality Breakdown (of 10 samples each)



Eval integrity: Oracle runs achieve 100% on both scorers, confirming calibration. Rationale scoring uses a judge model that checks for correct object types, distance orderings, and angular relationships. The eval harness ships with the dataset.

KEY FINDINGS

What the eval reveals

Five patterns — each pointing to a specific, addressable gap in current VLM training.

01 Heading estimation is broken

Both models fail every heading task. Estimating yaw from a monocular frame requires 3D reasoning current VLMs lack entirely.

02 Fragile 2D pixel heuristics

Even correct answers use "lower in image = closer" instead of 3D inference. Breaks on slopes, elevated roads, complex intersections.

03 Object ID degrades with range

Beyond ~50m, models misclassify objects (SUV > car > van). Wrong types cascade into wrong spatial reasoning.

04 Scale doesn't solve it

Opus 4.7 underperforms Sonnet 4.6 (40% vs. 50%). Spatial reasoning requires targeted supervision, not more parameters.

05 Correct answers ≠ correct reasoning

Sonnet: 50% choice, 35% rationale. Opus: 40% choice, 10% rationale. Models are frequently right for wrong reasons — creating false confidence.

THE DATASET

SFT & RL-ready spatial reasoning data

Every sample includes metric-grounded chain-of-thought rationales — the kind of spatial supervision current training corpora lack.

Chain-of-thought supervision

Step-by-step rationales with exact distances, yaw angles, object types. Directly usable for SFT.

Arbitrarily scalable

Programmatic generation from a multi-year driving log archive. This sample is 10; full runs produce thousands.

Production-grade ground truth

LiDAR-camera fusion, sub-meter precision. Deterministic, reproducible, no annotation noise.

Object & scene coverage

Car, SUV, truck, bus, bike, pedestrian, barriers. 10m–200m+. Highway, urban, night, construction.

Eval harness included

4 task variants, oracle validation, judge-model scorer. Ready to run.

Diagnostic granularity

10 categories for precise capability profiling.

Ready to close the spatial reasoning gap?

Expert-curated, RL & SFT-ready spatial training data at the scale and quality frontier labs require.

Contact: reasoncore.ai