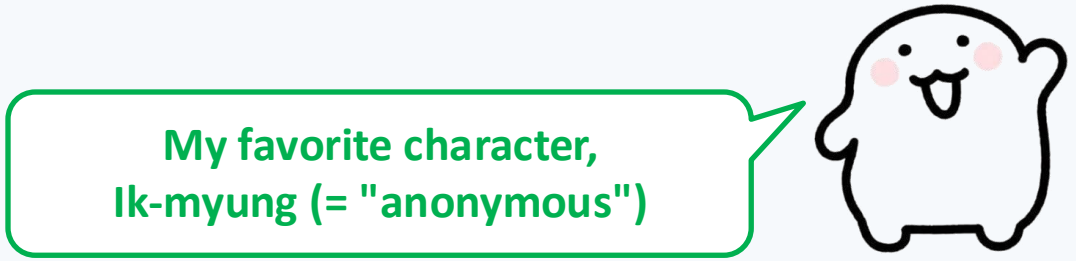
2nd place · ERA Workshop @ CVPR 2026



Eliciting Spatial Skills from Open-Weight VLMs

A zero-shot, per-task recipe for RoboSpatial-Home

Eunsu Baek (Team: Anonymous) Seoul National University

Eliciting a spatial recipe from two open-weight VLMs — a **reasoner** ⊕ a **pointing specialist**

Before we start — I'm not a spatial-reasoning researcher :)

I research **"Adaptive Sensing"**

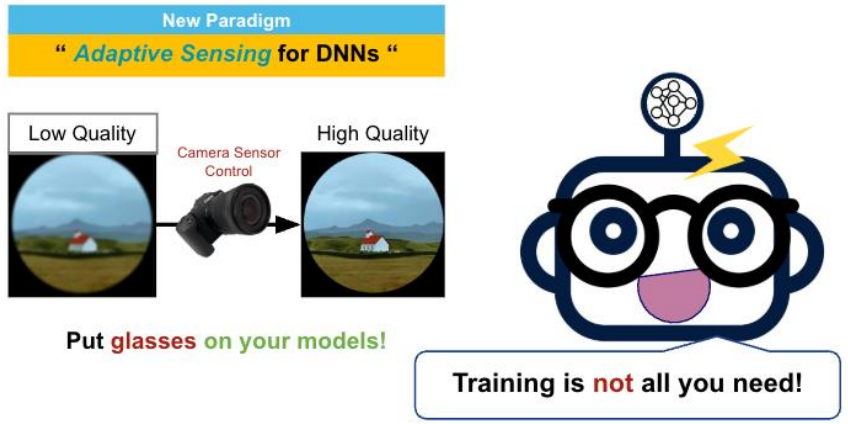
Improve robustness by adapting **how the agent senses** — rather than depending on large-scale training.

Now exploring its extension to robotics / embodied agents.

Joined this challenge to learn, hands-on:

- which spatial problems matter
- how models fail
- what transfers

My Solution - Sensing better than study harder!



"Training is not all you need!" — the same bet, applied to VLMs.

STRATEGY

No training — elicit what the models already know

The constraint

Limited compute — no room to fine-tune a large VLM on spatial data.

The bet

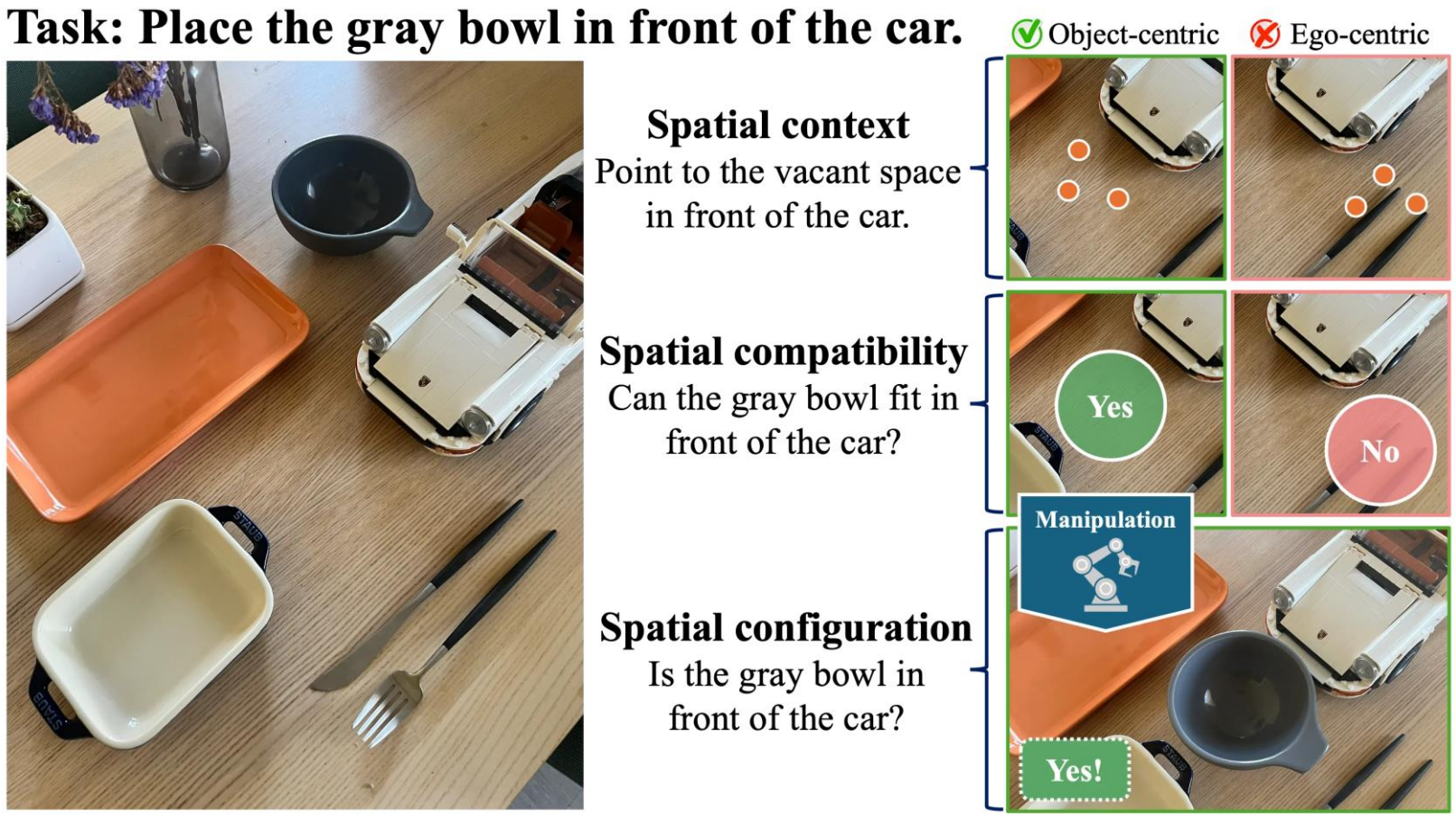
Strong open VLMs already hold more spatial skill than they show — elicit it with inference-time choices instead of training.

Keep the models fixed. Tune the **inference**, per task — and let the model's own reasoning tell us what to choose.

Result: 83.1 overall → 2nd place · zero-shot, open-weight leaderboard

Three spatial tasks, three different skills

One scene, three questions —
and no single recipe fits all three.



Context
pointing · ≤2 points
Localize a point in the correct region.
78.7 %

Compatibility
fit yes/no
Does the object fit in that space?
79.1 %

Configuration
directional yes/no
Does a directional relation hold?
91.5 %

figure: Song et al., RoboSpatial — CVPR 2025

STRATEGY

Two base models · one design space

Base models

Qwen3-VL-32B-Thinking · reasoner

Emits <think> traces · directional & fit reasoning
→ Configuration, Compatibility

RoboBrain2.5-8B-MT · pointing specialist

Trained for point localization
→ Context (⊕ Qwen in a 2-point union)

Consideration axes (chosen per task)

① Model — which VLM answers the task
② Reasoning — <think> on / off
③ Decoding — greedy vs sampling
④ Frame prompt — name a reference frame
⑤ Parser — trust the reasoning over the final token
⑥ Two-model union — combine complementary models

Same knob, opposite effect: greedy wins Config, sampling wins Compat.

STRATEGY

The recipe each task lands on

Task	Approach — no training	Micro	Macro
Configuration	Qwen3-VL-32B-Thinking · object-frame · greedy · reasoning · parser	**91.5**	91.9
Compatibility	Qwen3-VL-32B-Thinking · no-frame · sampling · reasoning · parser	**79.1**	74.1
Context	RoboBrain2.5-8B (greedy) ⊕ Qwen (sampling, object) → 2-point union	**78.7**	72.3
Overall	zero-shot · no training	**83.1**	**79.4**

Micro = plain accuracy. **Macro** = balanced over Yes/No (VQA) · mean over 6 relations (pointing) — guards the 83% Yes-prior.

TASK 1

Configuration

Challenge — the default over-rejects

91.5%

Base model: Qwen3-VL-32B-Thinking · reasoner · with no frame instruction, valid cases get rejected

“Is the monitor behind the sofa?”

DEFAULT reasoning

“The monitor is on the desk to the right of the sofa. So the monitor is not behind the sofa — it’s to the side. Answer: no.”

→ No ✗ correct: Yes

Literal matching — answers about the obvious relation (right), never checks the queried axis (behind).

“Is the microwave above the counter?”

DEFAULT reasoning

“The microwave is sitting on the countertop, so it’s not above the counter — it’s on the counter. Therefore no.”

→ No ✗ correct: Yes

Over-strict — treats “above” as needing a gap, so “resting on top of” fails on a technicality.

Configuration

Candidates — which reference frame?

91.5%

The knob: prepend a one-line reference-frame instruction. We considered four settings —

default

no frame instruction (baseline)

injects

— nothing prepended —

ego-centric

from the CAMERA's viewpoint

injects

"Right of X = to your right in the image."

Object-centric

from the OBJECT's own perspective

injects

"Right of X = to X's own right side."

world-centric

fixed WORLD-axis convention

injects

"Use consistent world axes, regardless of view."

Why this matters: the default reasons literally;

simply naming ***any*** frame pushes it into relative, spatial reasoning that re-verifies before committing.

TASK 1 · STRATEGY

Configuration

Strategy — name the frame

91.5%

The one line we prepend to every question

PROMPT FORMAT

```
[ FRAME INSTRUCTION ]
( blank line )
[ ORIGINAL QUESTION ]
```

OBJECT-CENTRIC ● our choice

"Interpret all spatial directions from the reference object's own perspective — as if you stood at the object, looking outward in its natural front-facing direction. Right of X = to X's own right side."

+

"Is the monitor behind the sofa?"

Any frame wins — the choice barely matters

Frame prompt	Macro
no-frame (default)	81.4
ego-centric	88.1
object-centric ● our choice	88.1
world-centric	89.8

+6.8 pt naming any frame is the win (81.4 → 88.1); which frame barely matters (1.7 pt spread).

Configuration

Why it works — literal → perspective reasoning

91.5%

“Is the monitor behind the sofa?”

Default → No ✗

“…the monitor is not behind the sofa; it’s to the side.”

+ Object frame → Yes ✓

“…from the sofa’s perspective, the coffee table is in front of it, so its back faces the desk — the desk is behind the sofa. So the monitor is behind it. Yes.”

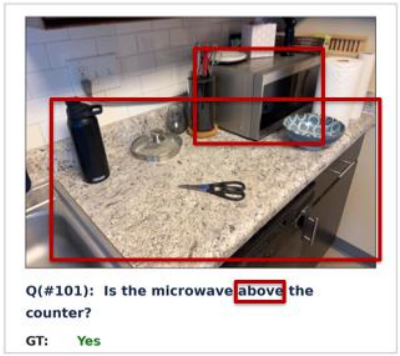
“Is the microwave above the counter?”

Default → No ✗

“…it’s not above the counter; it’s on the counter.”

+ Object frame → Yes ✓

“…from the microwave’s perspective, it sits on top of the counter, meaning the counter is below it. So yes, it’s above the counter.”

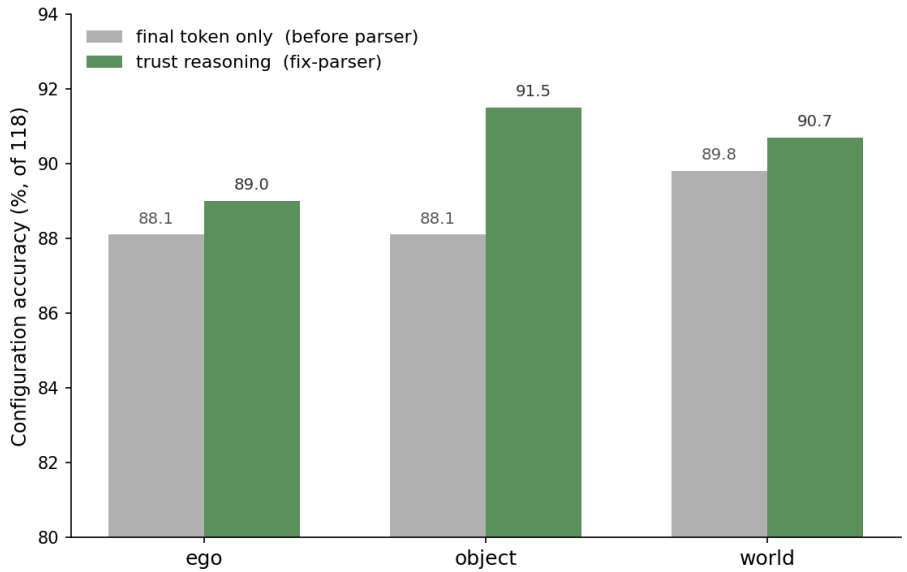


Configuration

Two more choices — parser & decoding

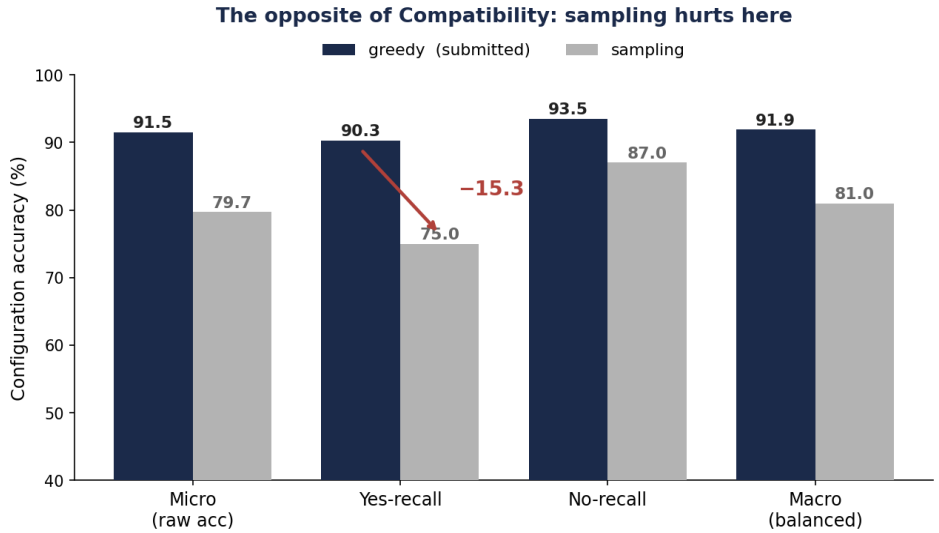
91.5%

② Fix parser — trust reasoning over the stray final token



When they disagree, the reasoning is right:— 4/4 in Config (3/3 in Compat, too). Every mismatch was an over-rejection, so the fix only restores wrongly rejected answers (+3.4 pt).

③ Decoding: greedy — sampling re-opens over-rejection

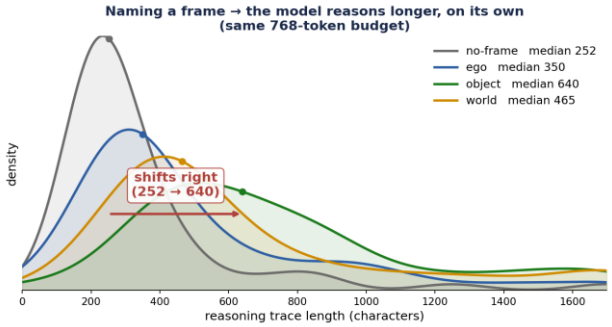


Insight: sampling re-opens over-rejection — Yes-recall 90.3 → 75.0, so micro drops −11.9. Configuration commits with greedy — the opposite of Compatibility.

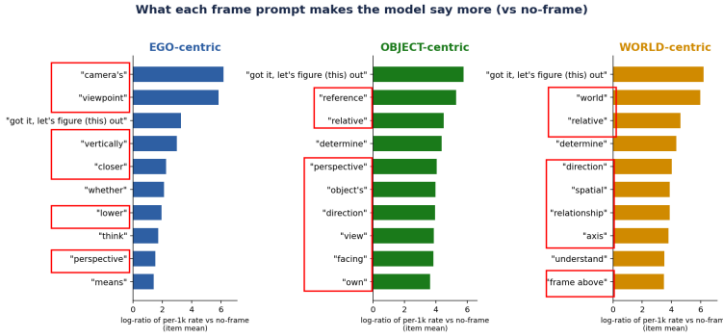
Configuration

Insight — framing makes it reason more, in spatial terms

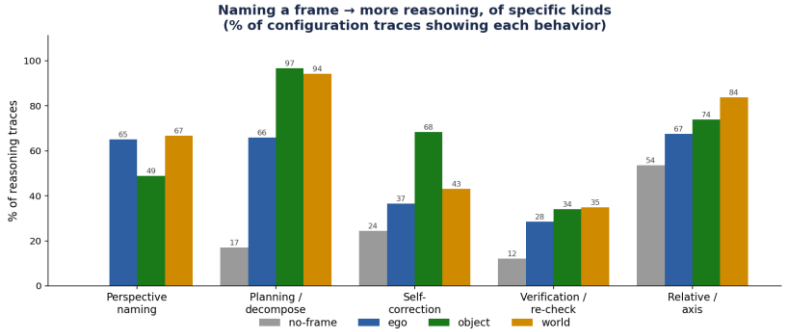
91.5%



① **It reasons longer** — same 768-token budget, but the object frame pushes reasoning length up ~2.5× (median 252 → 640 chars), on its own.



② **Each frame surfaces its own vocabulary** — ego → "camera/viewpoint", object → "reference/relative", world → "axis/world". The model echoes whichever frame it's given.

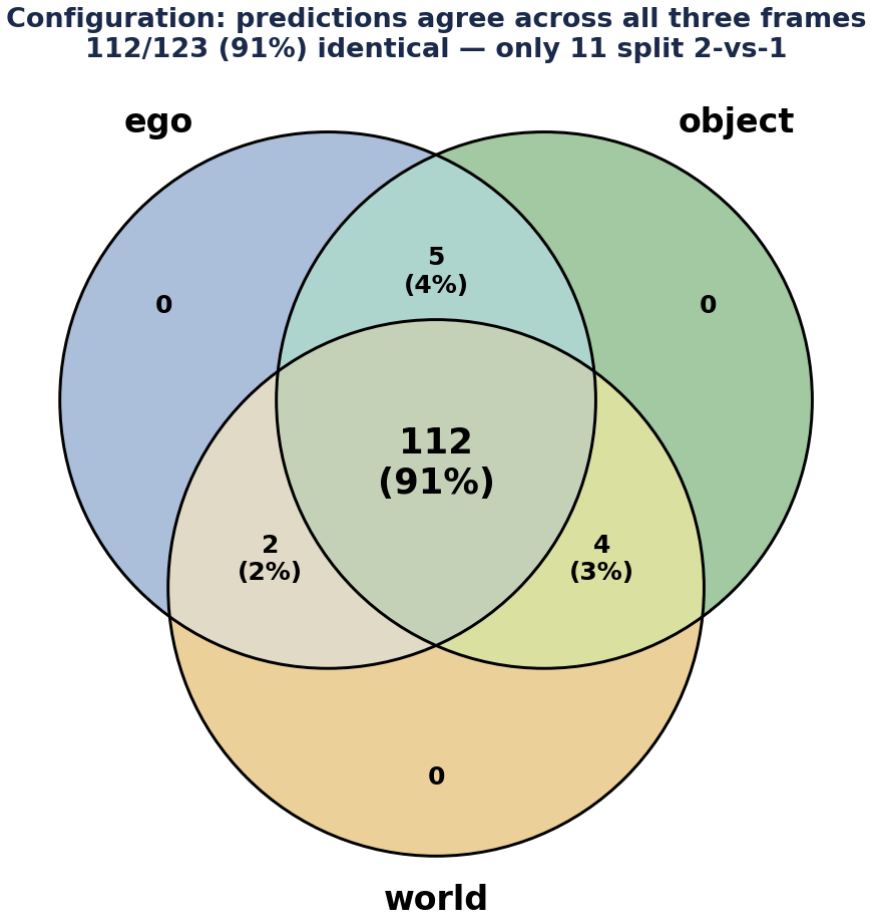


③ **More reasoning, of specific kinds** — a frame sharply raises Planning / decomposition (17 → 95%), with more self-correction, verification, and relative-axis reasoning.

Configuration

Insight — it isn’t true perspective-taking

91.5%



Naming a frame helps — but the model isn’t really disambiguating the frames.

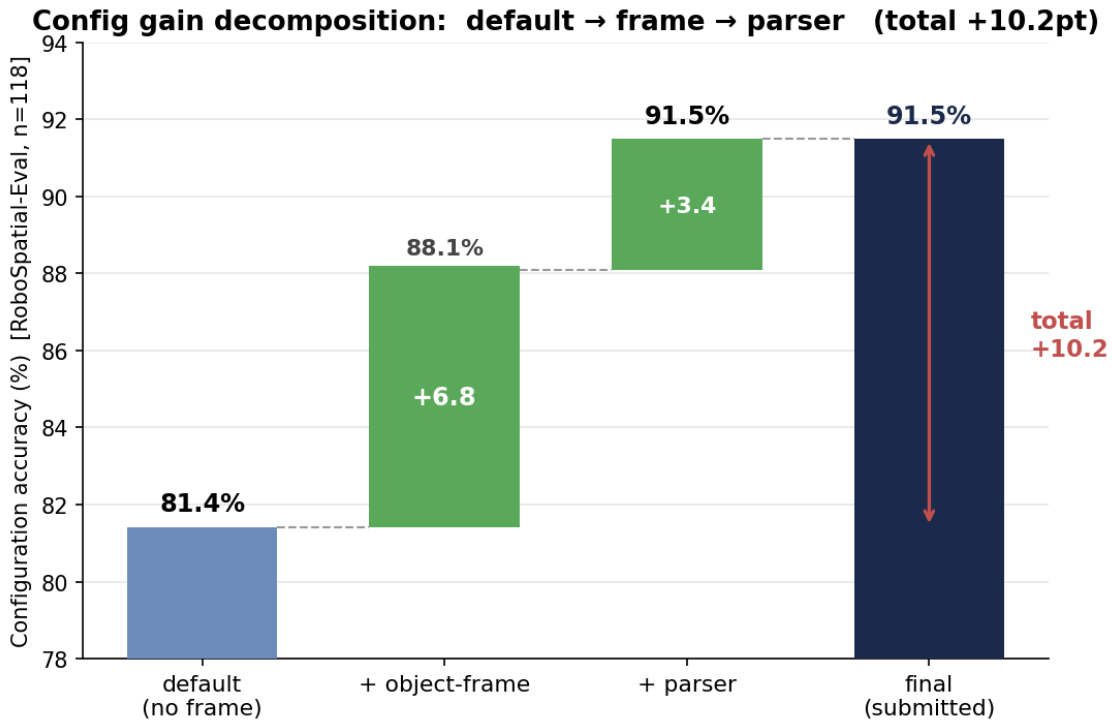
- ego / object / world all land at **88–90** (1.7 pt spread) — world only marginally higher.
- The three frames’ correct-answer sets **overlap heavily** (the Venn) — it re-reasons relationally, not from a specific viewpoint.
- Practical takeaway: the **presence** of a frame instruction is what matters, not **which** one — **the model was never trained to tell them apart.**

Configuration

Gain decomposition — micro and macro agree

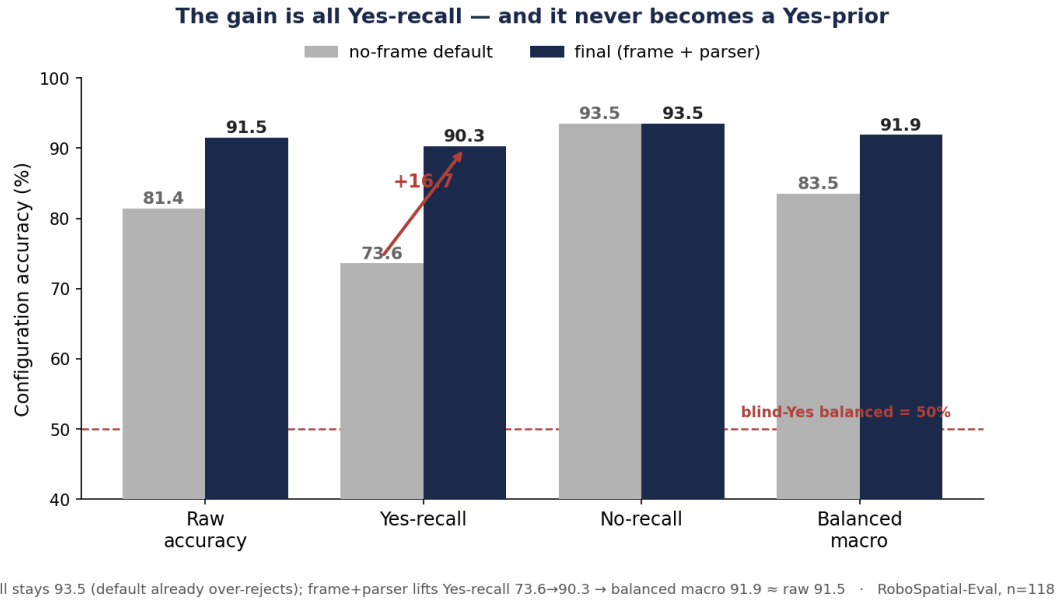
91.5%

Micro — the two fixes stack to 91.5



81.4 default → **+6.8** naming the frame → **+3.4** parser → **91.5** (+10.2 total)

Macro — the gain is real, not a Yes-prior



No-recall stays 93.5 (default already over-rejects); frame+parser lifts Yes-recall 73.6→90.3 → balanced macro 91.9 ≈ raw 91.5 · RoboSpatial-Eval, n=118

The whole gain is **Yes-recall (73.6 → 90.3)**; No-recall never moves off 93.5.

So balanced macro **91.9** ≈ raw **91.5** — the frame fixes over-rejection, the opposite of a Yes-bias.

TASK 2

Compatibility

Raw accuracy (Micro) is gamed — look at Macro

79.1%

105 questions, **83% "Yes"**. A blind "Yes" tops Micro (82.9) yet catches **zero "No"** — Macro 50 = chance.

Model	pred Y/N	Micro	Yes-rec	No-rec	Macro	
blind-Yes · guess	105 / 0	82.9	100	**0**	**50.0**	Yes-biased · gameable
Instruct · no reasoning	30 / 75	41.9	32	89	**60.5**	No-biased
Thinking · greedy	78 / 27	77.1	83	**50**	**66.4**	Misses ½ of No
✓ Thinking · sampling	74 / 31	79.0	82	**67**	**74.1**	**Most balanced**

Even our reasoner with greedy misses **half of all "No"** (No-recall 50). → We optimize Macro, and pick the last row: **Thinking · sampling**.

TASK 2 · STRATEGY

Compatibility — Greedy vs sampling

79.1%

Sampling grounds the answer in the scene — greedy just asserts a fluent conclusion.

It checks **occupancy** and **size** — fixing both directions. No-recall **50 → 67**, balanced (Macro) **66.4 → 74.1** // raw (micro) 77.1 → 79.0.

"Can the headphone fit right of the keyboard?" · GT No

Greedy → Yes ✗ — "If we move the mouse pad, the headphones can fit."

Sampling → No ✓ — "The mouse pad is already there, taking that space, so they can't fit."

"Can the wok fit above the fridge?" · GT Yes

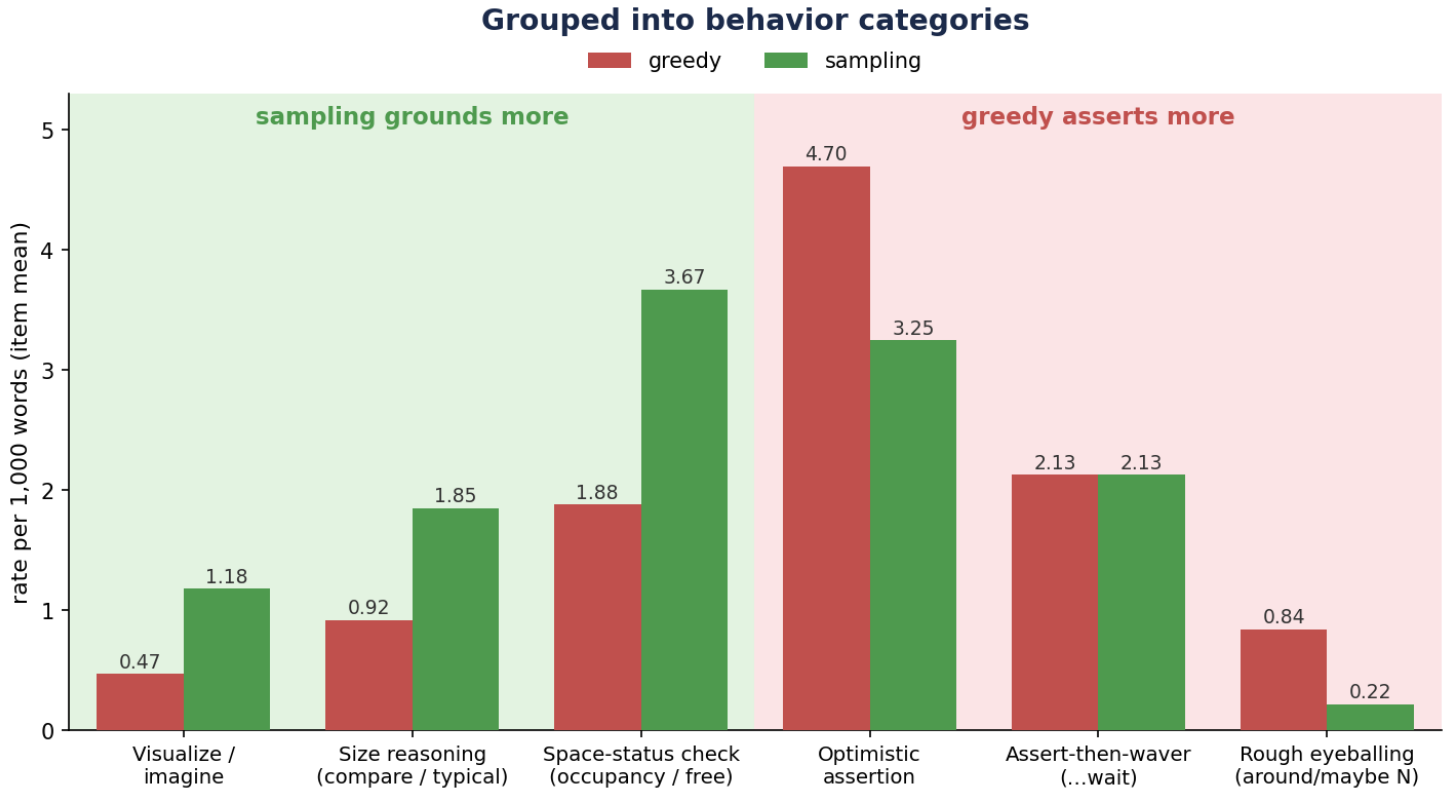
Greedy → No ✗ — "The wok is probably too big … can't fit."

Sampling → Yes ✓ — "The wok is about the size of one burner; the fridge top is wider than it, so it can fit."

Compatibility

Why sampling wins — it grounds, greedy asserts

79.1%



Grouping reasoning traces into behaviors:

sampling does more **visualizing & imagine (×2.5)**, **size reasoning (×2.0)** and **space-status checking (×2.0)**;

greedy does more **optimistic assertion** and **rough eyeballing (×3.8)**.

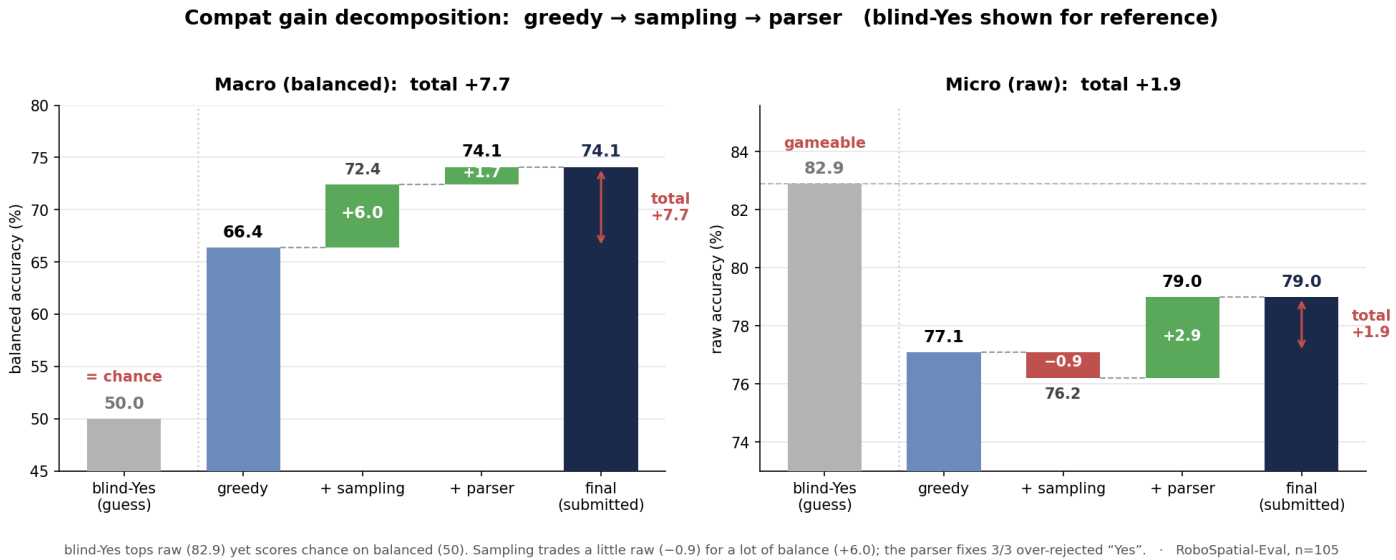
The "…wait" waver is identical (2.13 = 2.13) — that spiral is decoding-independent. Sampling's edge is the **grounding**, not less hesitation.

TASK 2

Compatibility

Gain decomposition

79.1%



blind-Yes tops raw (82.9) yet scores chance on balanced (50). Sampling trades a little raw (−0.9) for a lot of balance (+6.0); the parser fixes 3/3 over-rejected "Yes". · RoboSpatial-Eval, n=105

66.4 greedy → **+6.0** sampling → **+1.7** parser → **74.1**

+7.7 pt total (66.4 → 74.1, balanced).

Sampling recovers the missing "No"; the parser fixes 3/3 over-rejected "Yes".

Micro(raw) lands at **79.0 (submitted)**.

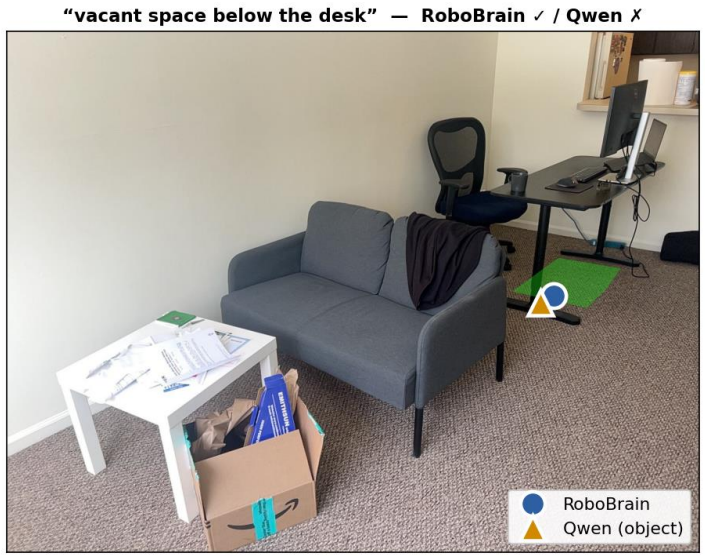
TASK 3

Context

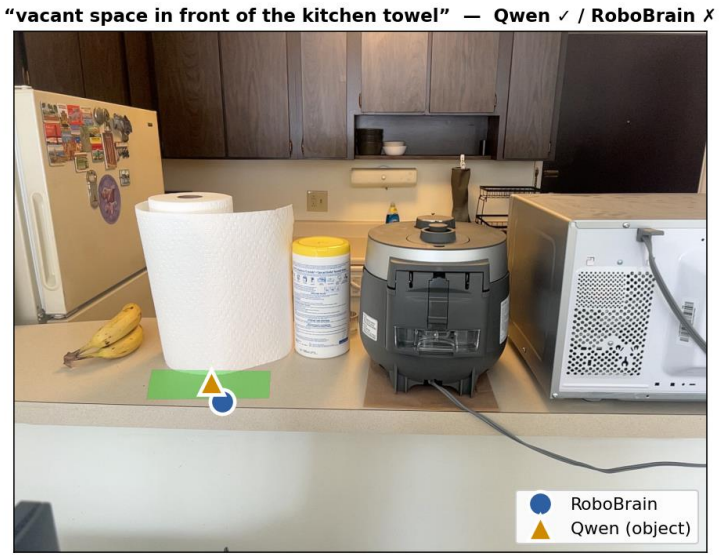
Challenge — one model, one point, often just misses

78.7%

Point at a described region (**≤2 points allowed**) — correct if either point lands inside the GT mask.



"below the desk" — Qwen alone gets this wrong: its point lands just barely outside.



"in front of the kitchen towel" — RoboBrain alone gets this wrong, again just barely outside.

① One point per answer

RoboBrain answers with a single point on 98% of questions; Qwen on 87%. One shot each.

② …and that shot often misses

Best solo: RoboBrain 66.4 · Qwen 61.5 — a miss on 34–39% of questions, many just outside.

③ Half the budget unused

Scoring allows ≤2 points, hit if either lands inside — a solo model wastes the second point.

TASK 3 · STRATEGY

Context

Strategy — use both points: union two models

78.7%

Submit **RoboBrain's point ⊕ Qwen's point** — no averaging, no picking a winner. Members chosen by sweep, not by hand.

Step 1 — sweep each model solo, keep its top-2

Model · decoding	none	ego	object	world
RoboBrain · greedy	**66.4**	62.3	61.5	63.9
RoboBrain · sampling	50.8	51.6	41.0	41.8
Qwen · greedy	52.5	56.6	57.4	50.8
Qwen · sampling	46.7	61.5	**57.4**	51.6

▢ top-2 per model (kept for Step 2) ■ member used by the submitted union

RoboBrain: **greedy** ≫ **sampling** (+15.6), no-frame best → top-2 greedy·{none, world}.

Qwen: needs a frame (+14.8), **sampling** ≥ **greedy** → top-2 sampling·{ego, object}.

Step 2 — union every top-2 × top-2 pair, submit the best

Union (RB ⊕ Qwen, ≤2 pts)	Micro	Macro
✓ RB greedy·none ⊕ Qwen samp·object	**78.7**	**72.3**
RB greedy·none ⊕ Qwen samp·ego	77.0	70.3
RB greedy·world ⊕ Qwen samp·object	76.2	70.3
RB greedy·world ⊕ Qwen samp·ego	76.2	69.2

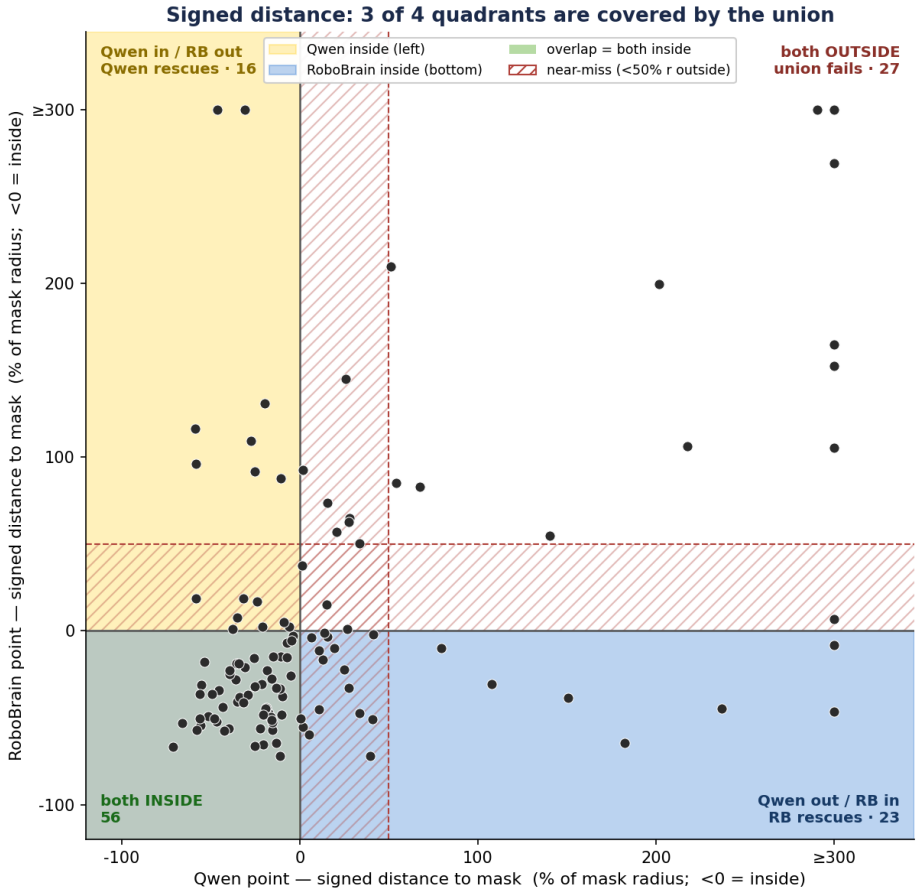
Reasoning is asymmetric — keep it ON for Qwen, OFF for RoboBrain: swapping either lowers the union (−3.3 / −6.6 macro).

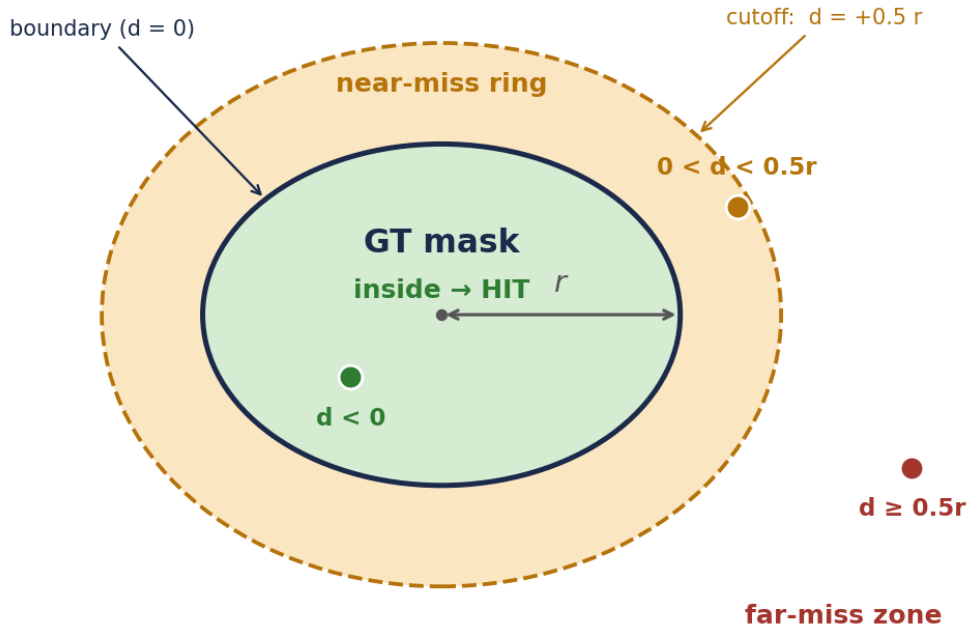
66.4 → 78.7 best solo → union (+12.3) · #1 on both micro and macro · RoboSpatial-Eval, n=122

Context

Findings ① — they rarely miss together

78.7%





Complementary coverage — 3 of 4 quadrants covered by at least one model inside → ~78%.

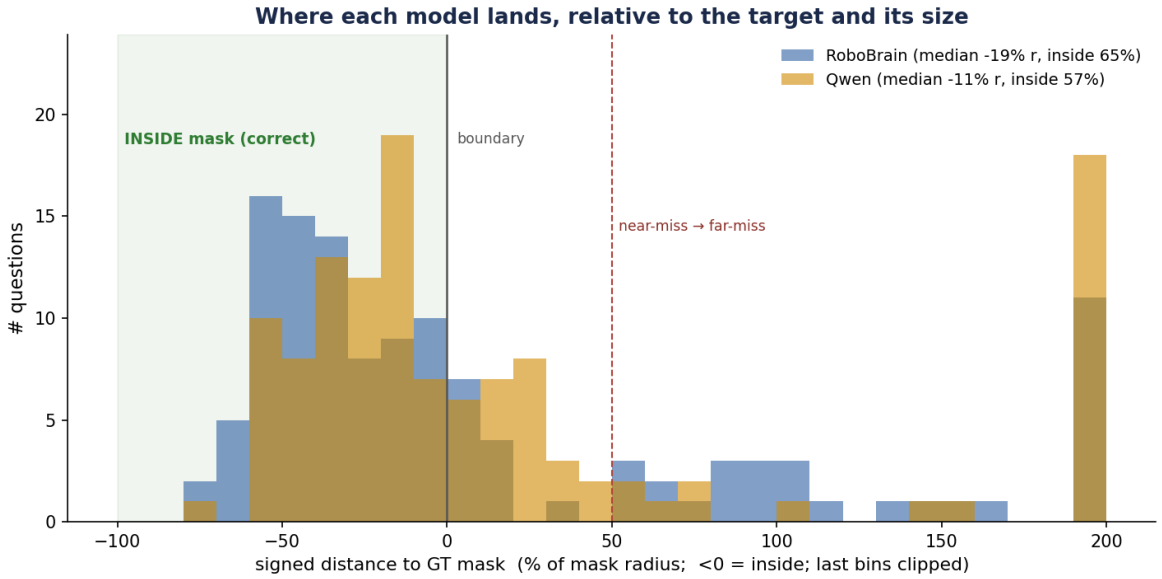
How we read "near" vs "far" — signed distance d to the mask, normalized by mask radius r = √(area/π):

HIT d < 0 **near-miss** 0 ≤ d < 0.5 r **far-miss** d ≥ 0.5 r

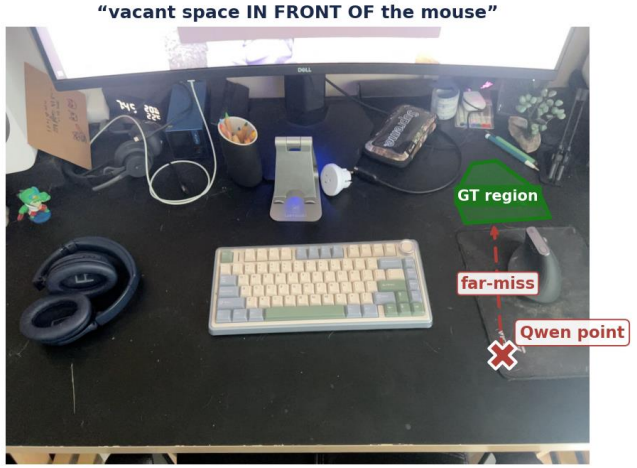
Context

Findings ② — near vs far, and why Qwen far-misses

78.7%



A real far-miss — “vacant space in front of the mouse”



“From the mouse’s perspective, in front is towards the bottom… Wait — … Wait, no. Wait, maybe towards the keyboard… Let’s think again.”

re-decides “front” ~10×, never commits — frame confusion

The distributions differ in the tail

RoboBrain’s misses skew **far** (median ≈ 0.9 r → wrong region); **Qwen**’s hug the boundary (median ≈ 0.4 r → right region, just outside).

Yet Qwen — the reasoner — still has a real far-miss tail. Why?

Far-misses think longer, not less (449 vs 388 words, more hedging) — reasoning helps discrimination, not localization.

RESULTS

Leaderboard — task tuning + union beat every single model

Model	Config	Compat	Context	Overall
Our recipe (2 models)	91.5	79.1	78.7	**83.1** 🥇
RoboBrain2.5-8B-MT	89.8	59.0	66.4	71.7
Qwen3-VL-32B-Thinking	84.8	68.6	58.8	71.1
RoboRefer-8B-SFT	84.8	30.5	61.5	58.0
RoboPoint-13B	71.2	71.4	27.9	56.0
Qwen2-VL	83.0	45.7	1.6	43.4
Molmo-7B	62.7	18.1	7.4	29.4

+11.5 overall

over the best single model (71.7 → 83.1), with no training.

General VLMs collapse on pointing.

Future work

- Multi-point interior estimate (fix near-misses)
- Frame-expert ensemble, weighted by the trace
- Teach fitting behaviors, grounded in depth
- Margin-aware loss — aim inside, not the edge

Takeaway: no single open model wins spatial reasoning — but eliciting and combining what each already knows does.

Thank you!

Eliciting spatial skills from open-weight VLMs — a zero-shot, per-task recipe for RoboSpatial-Home

Eunsu Baek (Team: Anonymous)

Ph.D Student @ Seoul National University

🏠 edw2n.github.io

✉️ beshu9407@snu.ac.kr

Homepage
edw2n.github.io

Project page
edw2n.github.io/robospatial-home