GeoSense

Internalizing Geometric Necessity Perception for Multimodal Reasoning

Core message

MLLMs should not always use 3D geometry. They should learn when geometry is necessary.

ICML 2026 Regular Paper

Ruiheng Liu, Haihong Hao, Mingfei Han, Xin Gu, Kecheng Zhang, Changlin Li, Xiaojuan Chang

Motivation: 3D Is Useful, but Not Universal

Spatial queries

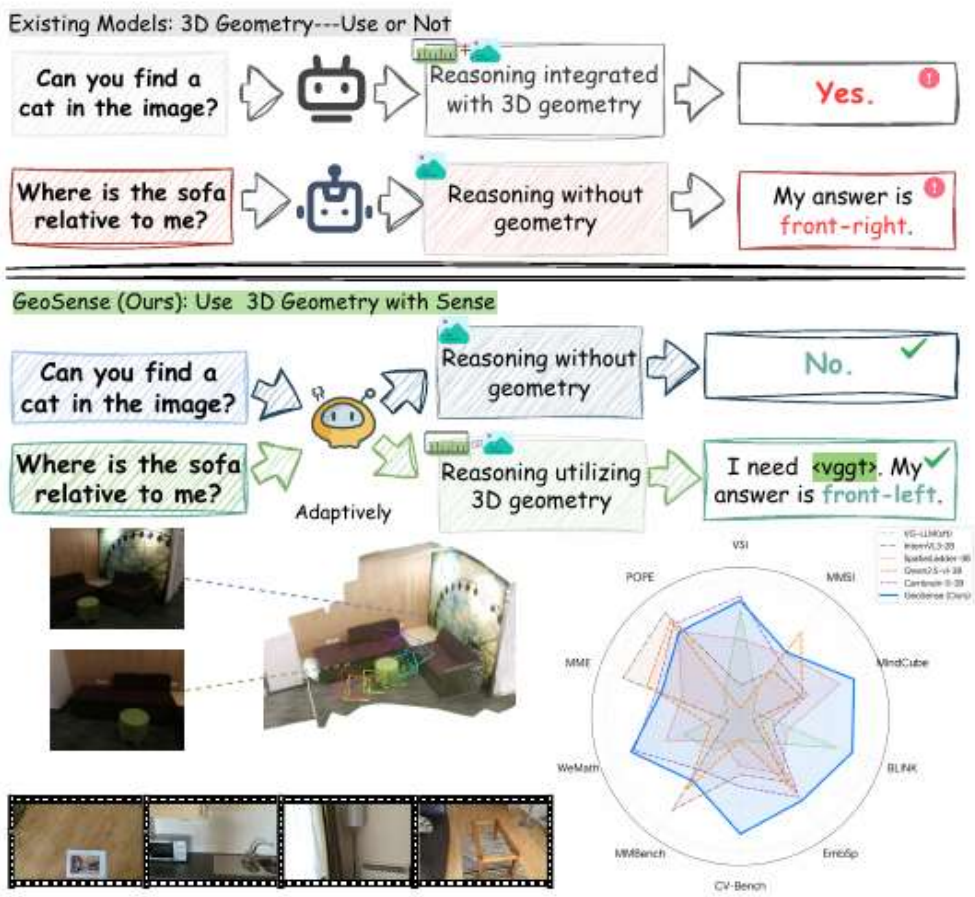
Distance, direction, occlusion, object size and navigation often need geometric cues.

General visual queries

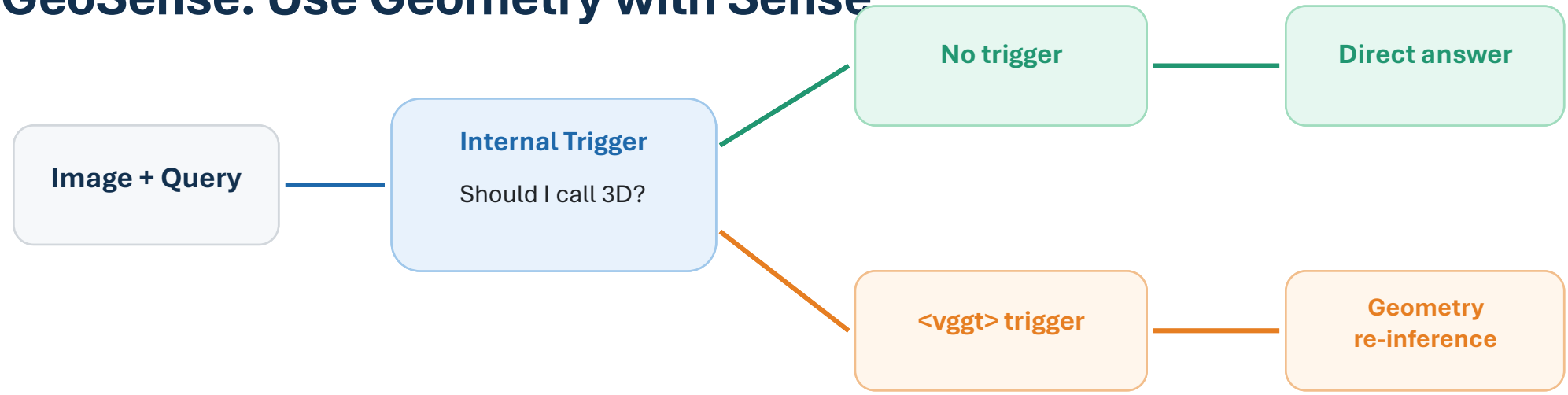
OCR, recognition and many counting questions may not need geometry at all.

Problem

Always-on 3D can add noise and compute; never using 3D misses spatial structure.



GeoSense: Use Geometry with Sense



Design 1

Independent geometry input channel: 3D is an on-demand resource, not a mandatory feature.

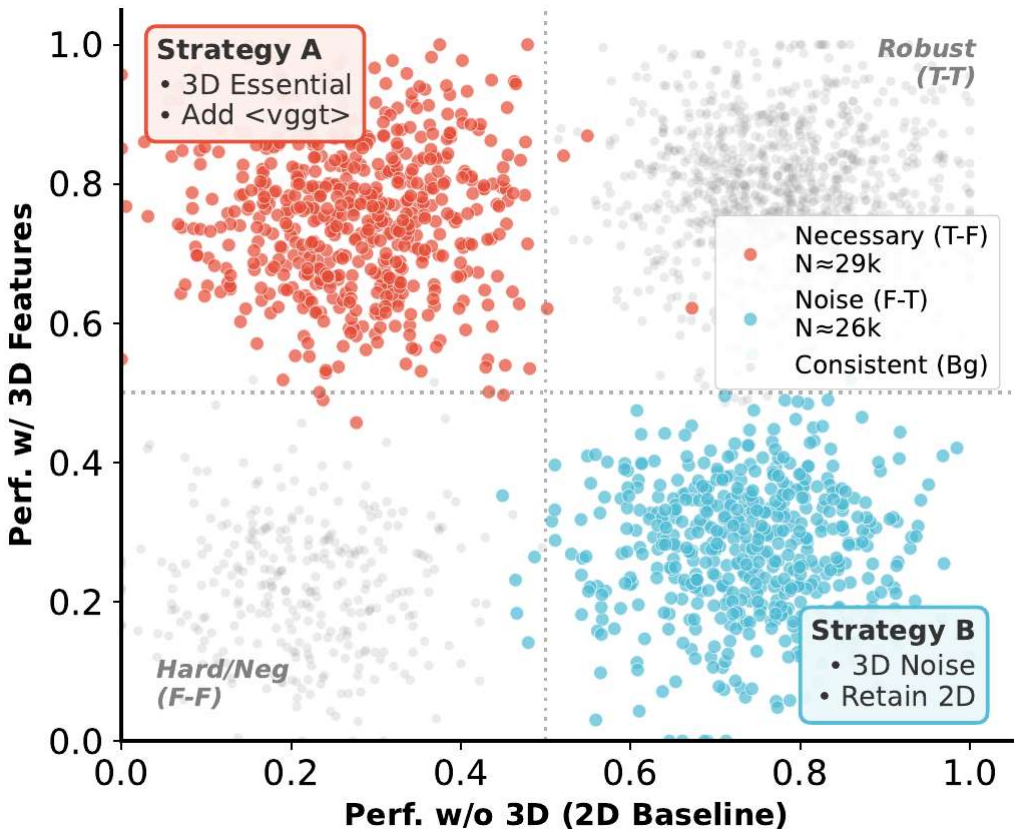
Design 2

Model-adaptive perception tuning: learn empirical priors from with/without 3D discrepancies.

Outcome

Suppress geometry for easy 2D cases; activate it for 3D-dependent reasoning.

Training Signal: Learn Geometry Necessity from Discrepancies



Dual-condition inference

Run each public-data sample twice: with 3D geometry and without 3D geometry.

T-F: 3D helps

Add <vggt> trigger and teach geometry activation.

F-T: 3D hurts

Retain 2D answer and teach trigger suppression.

117K	29K	26K
task-aware samples	geometry-needed cases	geometry-noise cases

Results: Better Spatial Reasoning without Always-On 3D

#1	56.6	78.7	64.3	895ms
overall rank among evaluated models	Spatial Avg. across 6 benchmarks	CV-Bench 3D	EmbSpatial	amortized latency on A100

Inference efficiency

- ~36% of VSI-Bench samples trigger 3D extraction
- 895 ms/sample vs. 950 ms/sample for always-on 3D fusion
- 0.5% extra FLOPs for non-triggered samples

Interpretation

GeoSense does not make 3D reasoning free. It selectively pays the cost when geometry is likely to change the answer.

Efficiency comes from avoiding unnecessary 3D calls.

Takeaway: Perceive the Need for Geometry

Q: Is there only one real cat in this picture? Yes or No?
GT: Yes.

Pred:<vggt> Yes.

(a)

Q: Is there a chair in the image?
GT: No.

Pred:<vggt> No.

(b)

Q: What is the closet distance between the stove and the tv (in meters)?
GT: 4.5

Pred:<vggt> 4.6

(c)

Q:Is this Python code?

Pred: Yes. GT: Yes.

(d)

Q: Is there a snowboard?
GT: Yes.

Pred: Yes.

(e)

Q: Based on these two views showing the same scene: in which direction did I move from the first view to the second view?
GT: Diagonally forward and right.

Pred: Diagonally forward and right.

(f)

1. Adaptive

The model learns to trigger <vggt> only for queries that need geometric cues.

2. Model-adaptive

Training labels come from the model's own with/without 3D performance gaps.

3. Practical

Better spatial reasoning while preserving general visual ability and reducing always-on geometry cost.

One-sentence takeaway: GeoSense teaches an MLLM not just how to use 3D geometry, but when to use it.