SpaRRTa: A Synthetic Benchmark for Evaluating Spatial Intelligence in Visual Foundation Models

Turhan Can Kargin[1,2], Wojciech Jasiński[1,3], Adam Pardyl[1,2,4], Bartosz Zieliński[1], Marcin Przewięźlikowski[1,2]

[1]Jagiellonian University [2]Doctoral School of Exact and Natural Sciences [3]AGH University of Krakow [4]IDEAS NCBR

Winter School on Foundation Models (FoMo) 2026

Project page

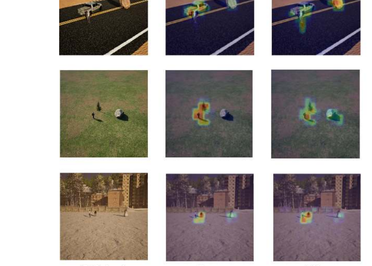
Overview

SpaRRTa probes whether visual foundation models encode object-to-object spatial relations, not just semantics.

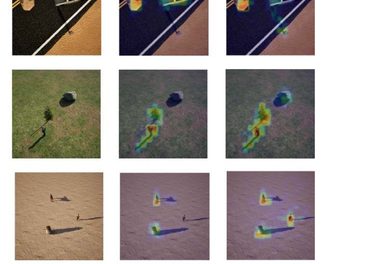
- **Benchmark:** UE5 synthetic scenes with controllable layouts and unambiguous labels.
- **Tasks:** Egocentric (camera view) and allocentric (human view) relation prediction.
- **Scale:** 5 environments, 13+ VFMs, 50K+ images, 3 probing strategies.

Task Formulation

- Predict one of four classes: **left / right / front / back**.
- Ground truth comes from simulator geometry and viewpoint frame.
- Ambiguous boundary cases are filtered out during data generation.



Egocentric. Camera defines the reference frame.



Allocentric. Human viewpoint defines the frame.

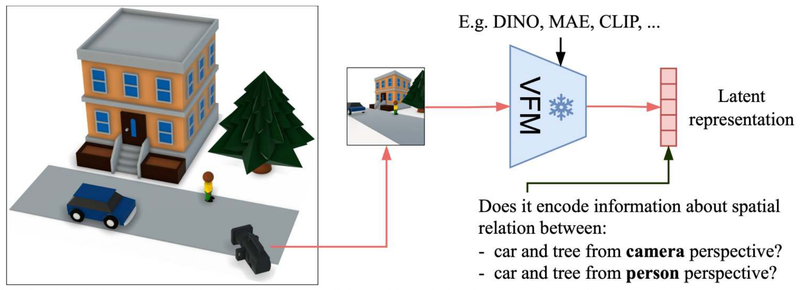
Correlation with Other Metrics

Pattern: SpaRRTa aligns more with geometry/3D-oriented benchmarks than with semantic-only classification metrics.

Efficient Probing Attention

Task-aware focus. Efficient probing attends to source/target regions, and in allocentric cases also emphasizes the viewpoint object.

Problem Illustration

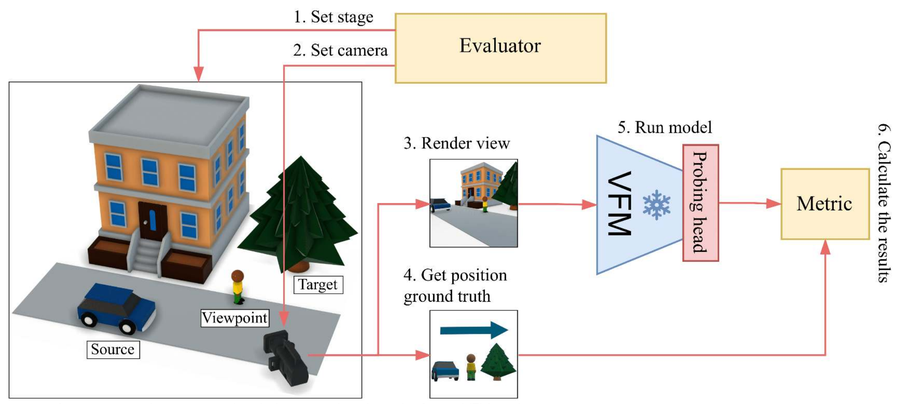


SpaRRTa intuition. Spatial relations change with viewpoint (ego vs. allo), requiring relational scene understanding.

Benchmark Environments

Diverse evaluation worlds. Forest, desert, winter town, bridge, and city setups challenge transfer of spatial reasoning.

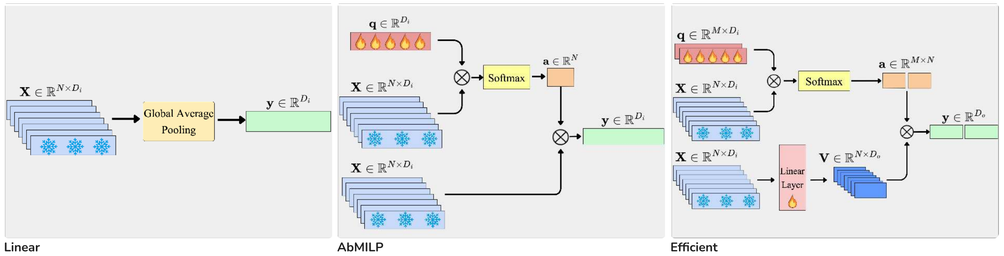
Evaluation Pipeline



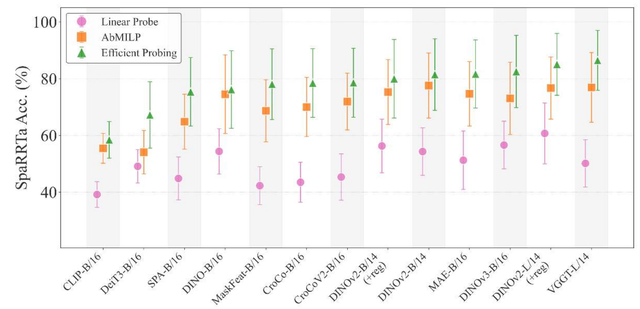
6-step protocol: set stage and camera, render, derive ground truth, probe frozen VFM, and compute accuracy.

Probing Methods

- **Linear Probing (GAP):** pools all patch tokens uniformly.
- **AbMILP:** learns a single attention map to weight informative patches.
- **Efficient Probing:** uses multiple learnable queries for selective aggregation.

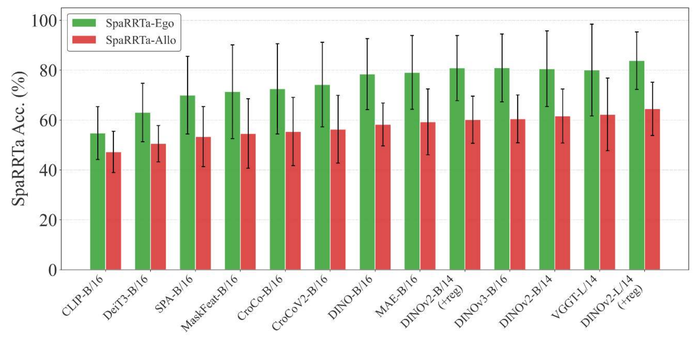


Impact of Probing Strategy



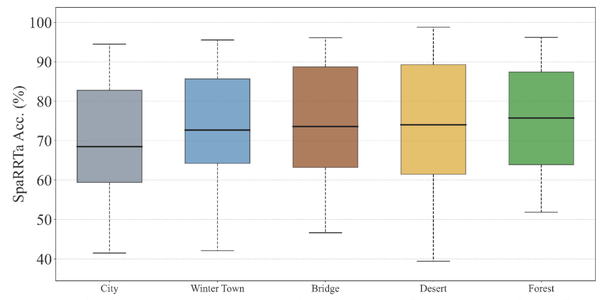
Across backbones: efficient probing consistently unlocks the strongest spatial signal.

Egocentric vs Allocentric



Systematic difficulty gap. All models perform better on egocentric than allocentric relations.

Environment Complexity



Complex scenes are harder. Accuracy drops in cluttered settings such as City compared to Desert/Forest-like scenes.

Key Findings

- **Consistent hierarchy:** Linear < AbMILP < Efficient probing.
- **Allocentric is harder:** viewpoint shift causes a large persistent gap.
- **Patch-level signal:** global pooling hides spatial information.
- **Clutter hurts:** city-like scenes reduce accuracy versus simpler environments.

Takeaways & Links

- SpaRRTa is a dedicated axis for **spatial intelligence** evaluation.
- Use spatially selective probes to reveal hidden geometric structure.

Project: sparrta.gmum.net
Paper: arXiv:2601.11729
Code: github.com/gmum/SpaRRTa
Dataset: huggingface.co/datasets/turhancan97/SpaRRTa