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

Turhan Can Kargin, Wojciech Jasinski, Adam Pardyl, Bartosz Zielinski, Marcin Przewiezlikowski

Jagiellonian University · AGH University of Krakow · IDEAS NCBR

PP-RAI 2026 Krakow

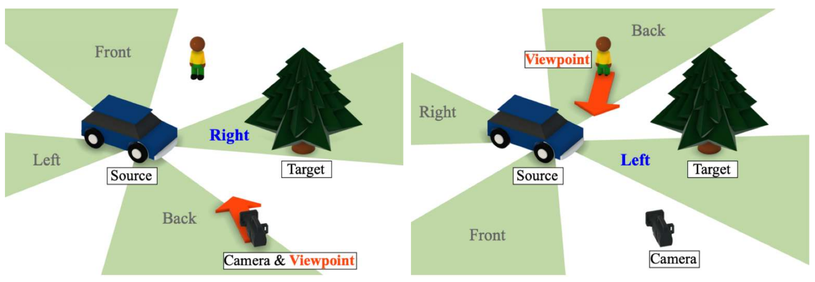
Overview

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.

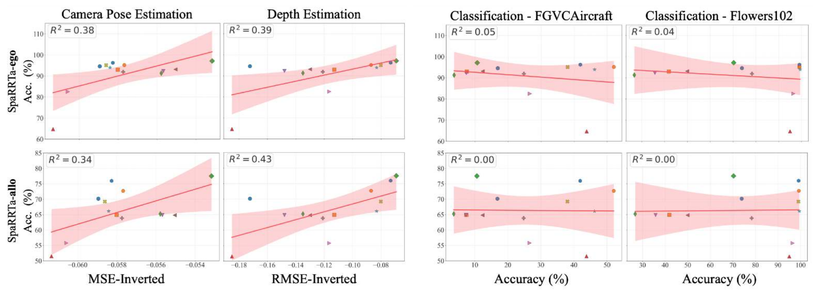
Task Formulation

- 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.



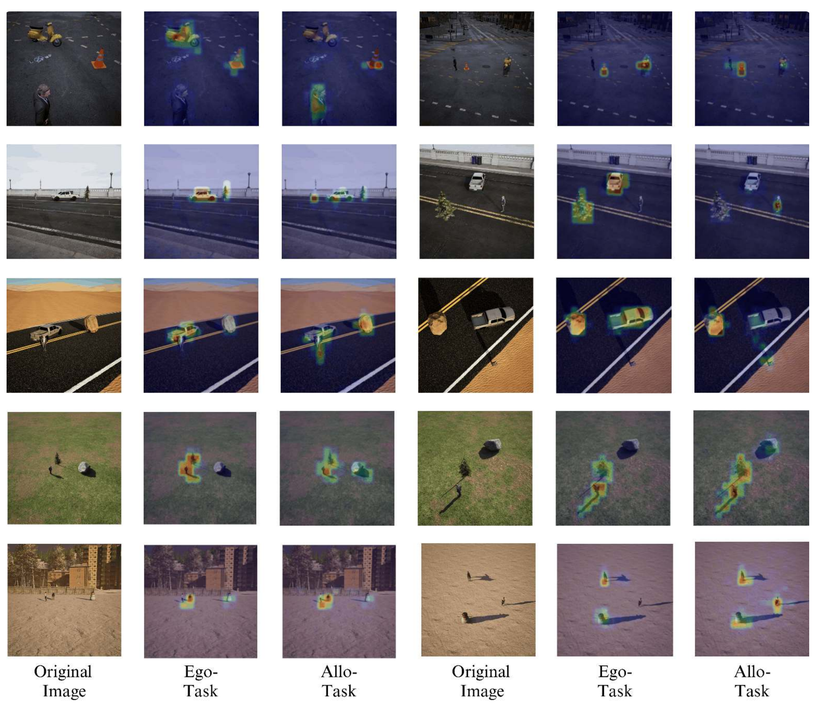
Egocentric: camera defines reference frame. **Allocentric:** human defines reference frame.

Correlation with Other Metrics



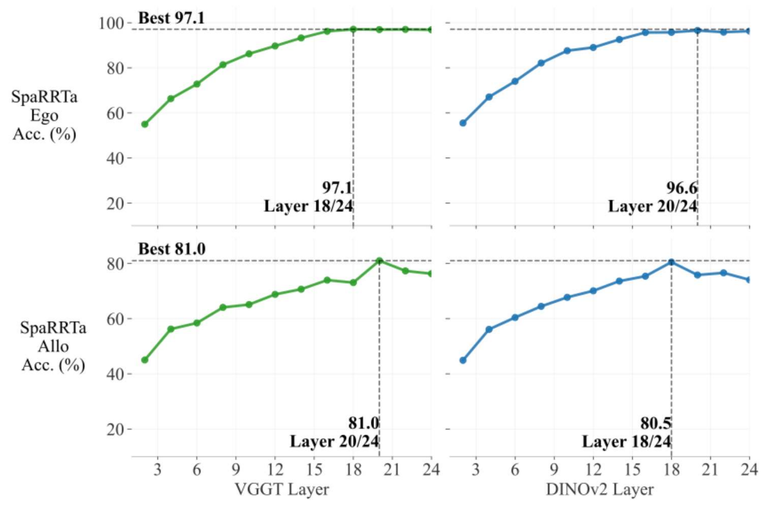
Pattern: SpaRRTa aligns with geometry/3D-oriented benchmarks more than semantic-only classification metrics.

Efficient Probing Attention



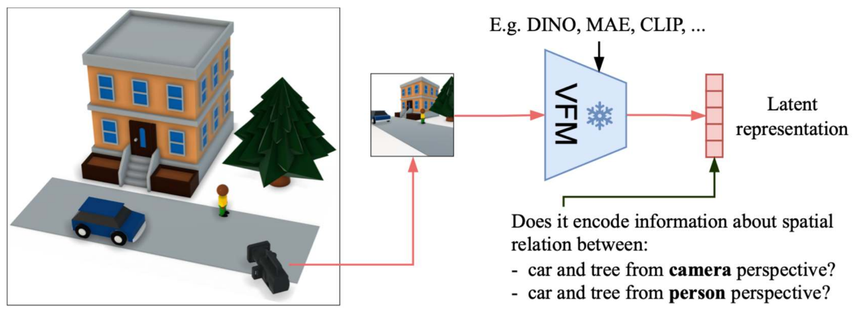
Task-aware focus: efficient probing attends to source/target regions and in allocentric cases emphasizes viewpoint cues.

Layer-wise Probing



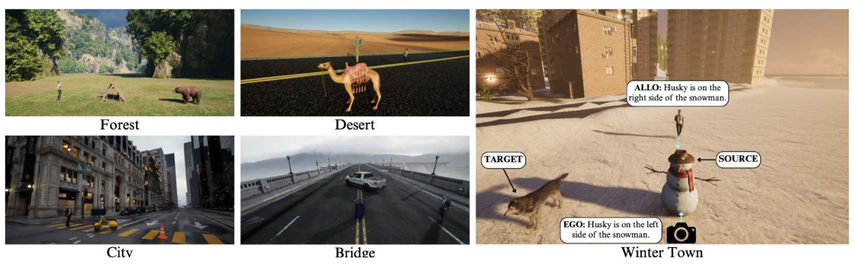
Spatial accuracy rises through early layers, peaks in late-intermediate blocks (around layers 18-20), and then slightly drops at the final layer.

Problem Illustration



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

Benchmark Environments

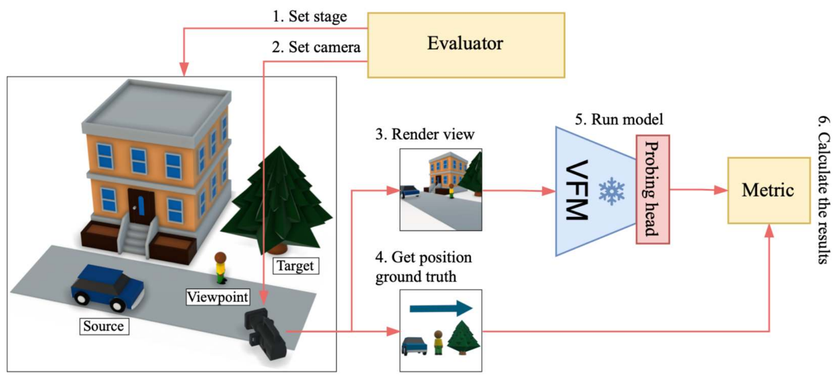


Diverse worlds (forest, desert, city, bridge, winter town) challenge transfer of spatial reasoning.

SpaRRTa Asset Library

SpaRRTa constructs test examples using a curated set of diverse, high-fidelity 3D assets selected based on common ImageNet classes.

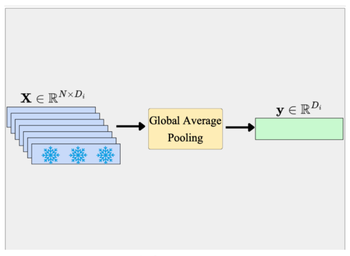
Evaluation Pipeline



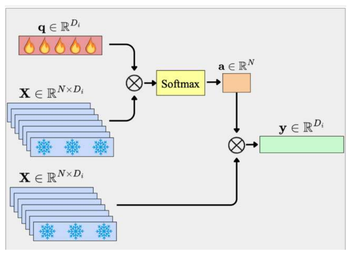
6-step protocol: set stage/camera, render, derive labels, probe frozen VFM, and compute accuracy.

Probing Methods

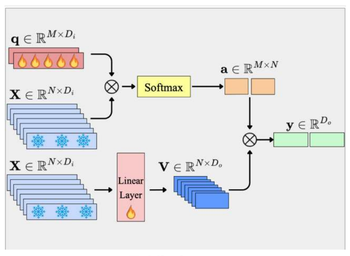
- **Linear Probing (GAP):** pools all patch tokens uniformly.
- **AbMILP:** learns a single attention map to weight informative patches.
- **Efficient Probing:** uses multiple learnable queries for selective aggregation.



Linear

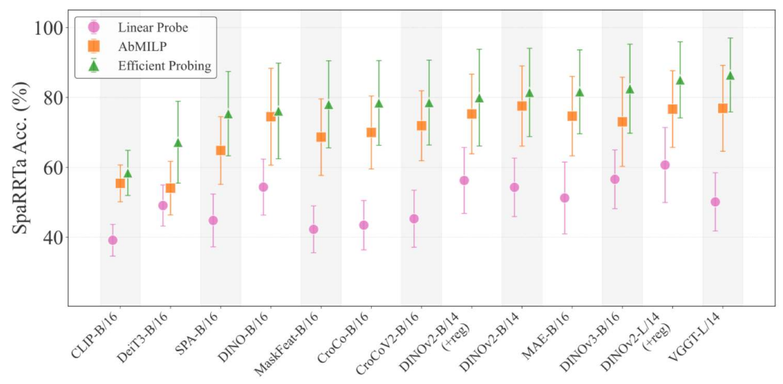


AbMILP



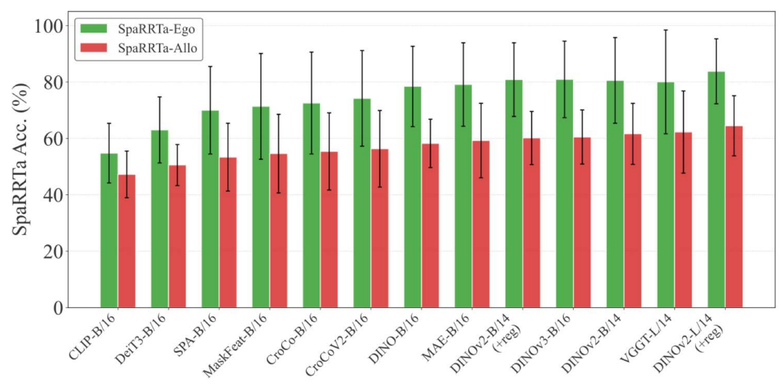
Efficient

Impact of Probing Strategy



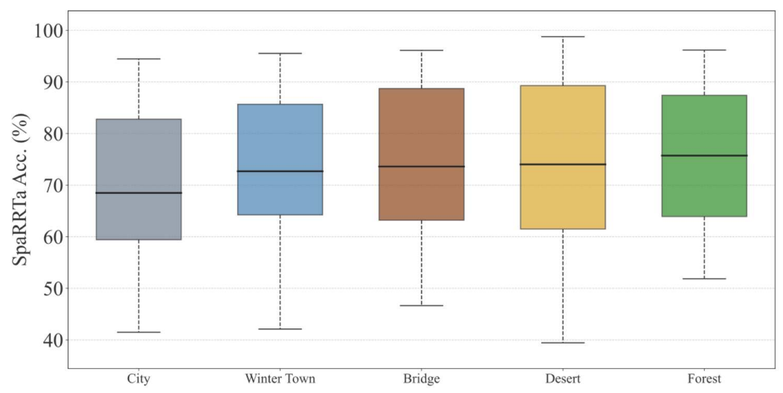
Across backbones, efficient probing consistently unlocks the strongest spatial signal.

Egocentric vs Allocentric



All models perform better on egocentric than allocentric relations.

Environment Complexity



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

Key Findings

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

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

Project QR

Project page: scan to open sparrta.gmum.net.