PROBING THE GENERAL SPATIAL AWARENESS OF VISION ENCODERS

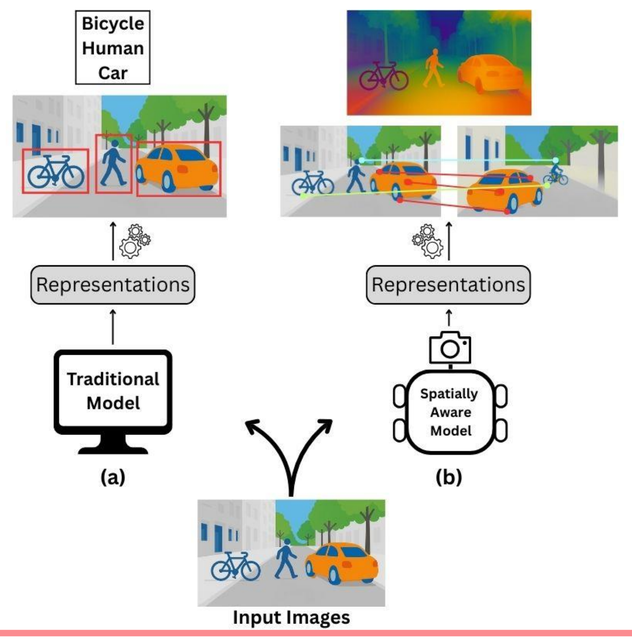
Kargin T.C. Przewięźlikowski M. Jasiński W. Zieliński B.
Jagiellonian University, IDEAS NCBR, AGH University

#TLDR

We evaluate whether vision encoders (e.g., DINO, CLIP, VGGT) inherently understand spatial relationships between objects in 3D scenes. Using probing techniques and synthetic environments, we compare models' ability to classify the relative position of one object with respect to another from camera's perspective. It lays the groundwork for building spatially aware autonomous systems.

Motivation

- AI systems require **3D spatial understanding** for robotics, AR/VR, and autonomous driving. However, existing benchmarks focus on narrow tasks (e.g., depth estimation) and rely on expensive, dataset-specific annotations.
- They often fail to assess abstract or generalizable spatial reasoning.
- This project asks a key question: **Can general-purpose vision models encode spatial awareness directly from 2D images?**
- To answer it, we propose a **systematic, scalable, and task-agnostic evaluation protocol** that probes pretrained vision models for their internal spatial awareness without requiring costly labels.



Solution

We introduce a general-purpose framework to probe spatial understanding in vision encoders using synthetic, photorealistic 3D environments (Unreal Engine 5).

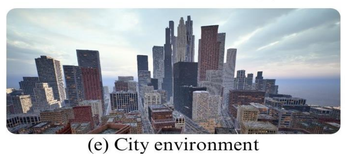
- **Environment**: Fully controllable 3D scenes (clutter, occlusions, camera motion)
- **Data Generation**: Auto-labeled from scene geometry (no manual annotation)
- **Task**: Classify relative position of one object to another (left, right, front, back)
- **Models**: Probed using Linear, ABMILP, and Efficient Probing methods


(a) Desert environment


(b) Forest environment


(c) Bridge environment

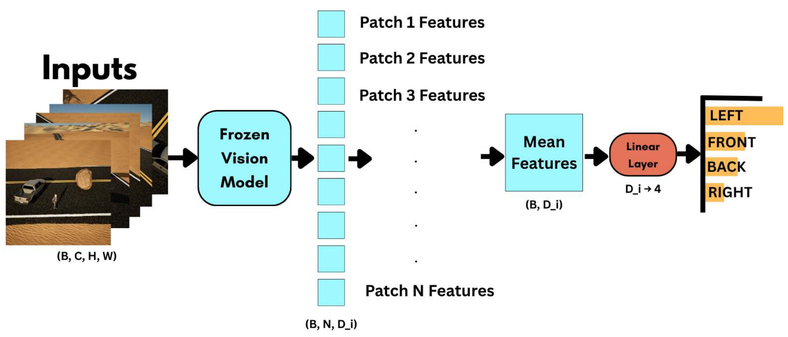

(d) Winter Town environment


(e) City environment

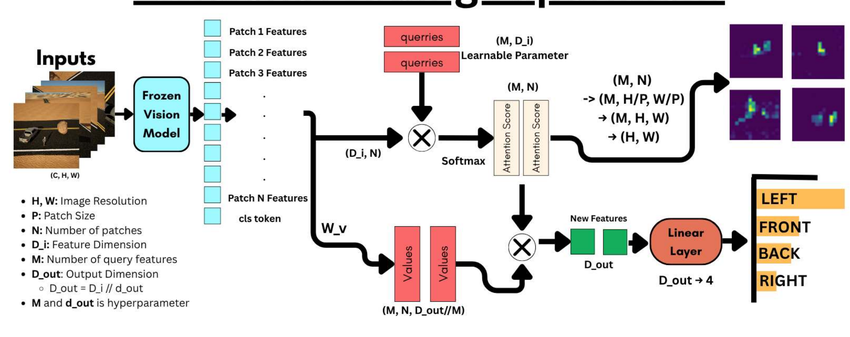

Main Assets

Probing Methods

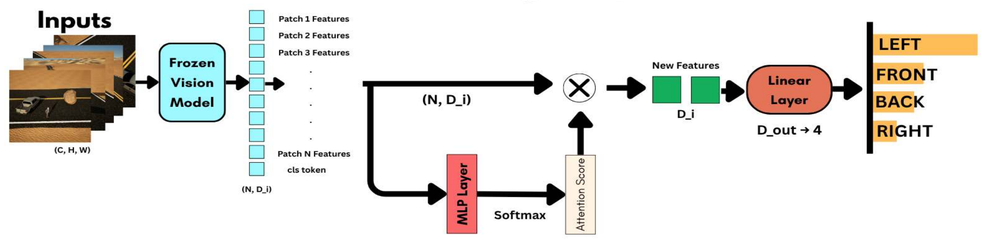
Linear Probing



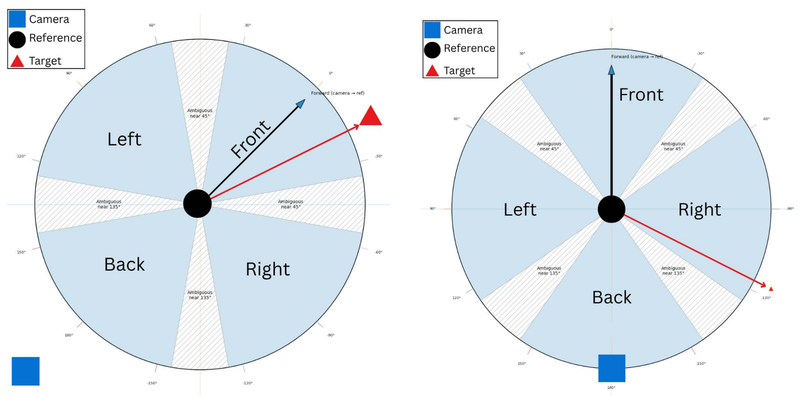
Efficient Probing Experiment



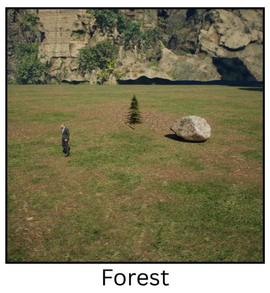
ABMILP Probing Experiment



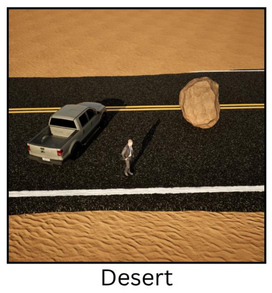
Data Generation Method



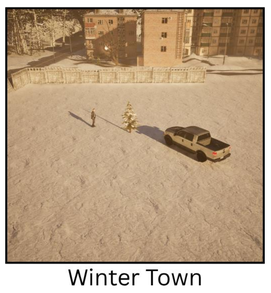
- **Target:** Human
- **Reference:** Tree


Forest

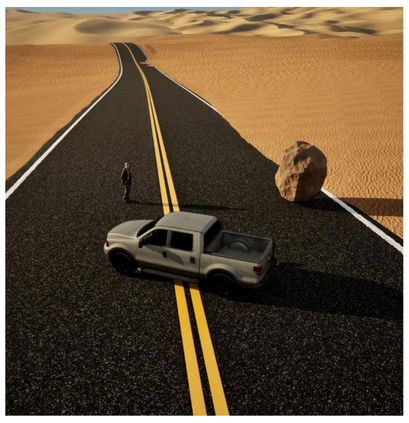
- **Target:** Car
- **Reference:** Rock

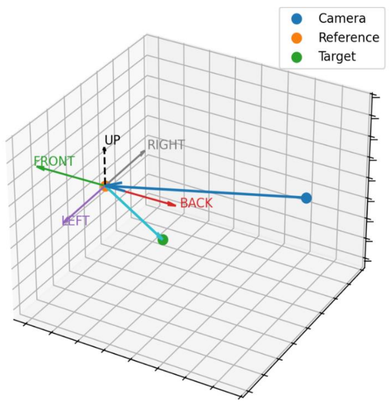

Desert

- **Target:** Car
- **Reference:** Tree


Winter Town

Figrue: **Example generated scenes.** For each environment, we generated 5,000 images and split them as **train, validation and test** set. Then we automatically label the images by determining where **target object** lies relative to **reference object**, from the **camera's perspective**. It classifies the target as Front, Back, Left, Right, or Ambiguous.

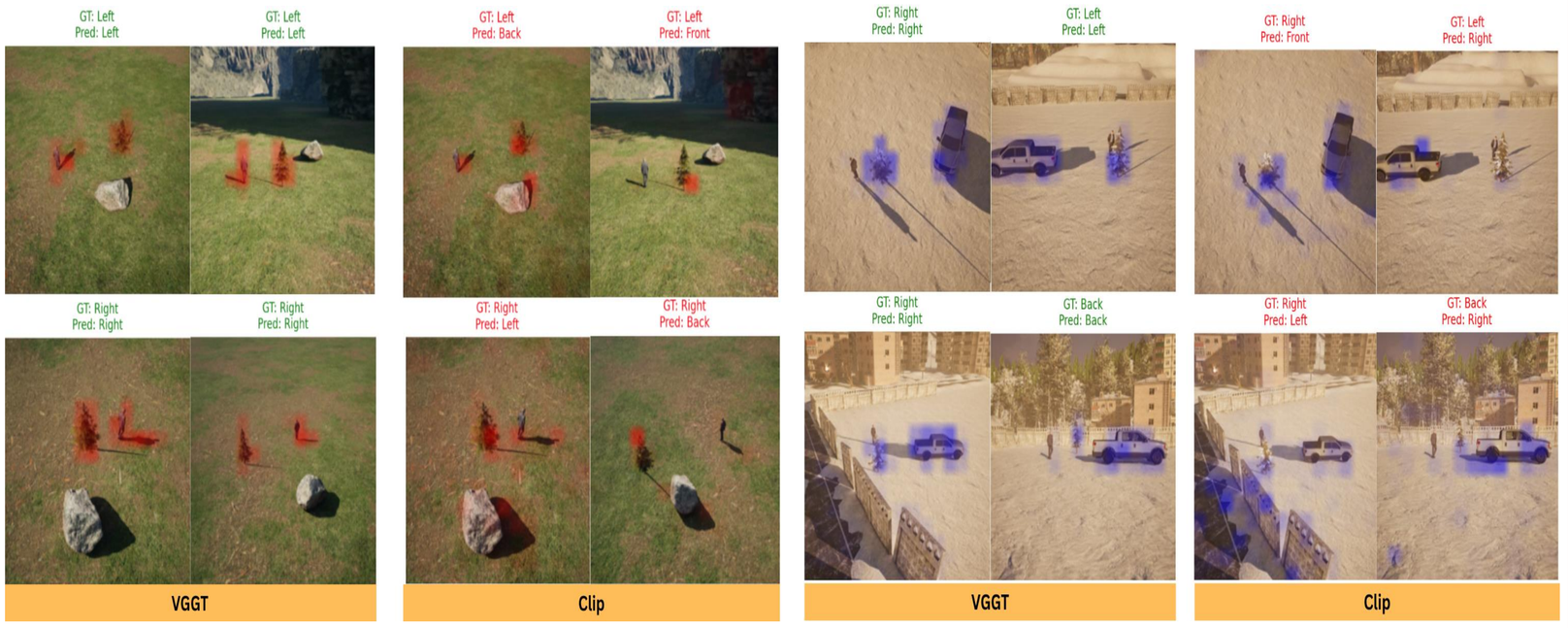




Results

Model	Architecture	FOREST			DESERT			Winter Town		
		Linear	ABMILP	Efficient	Linear	ABMILP	Efficient	Linear	ABMILP	Efficient
DINO	ViT-B/16	56.3	83.83	85.33	60.12	90.80	96.01	70.91	**92.42**	93.94
DINOv2	ViT-B/14	57.19	85.33	91.92	68.71	93.25	97.24	67.60	88.48	96.97
DINOv2_reg	ViT-B/14	64.10	82.34	92.51	78.22	88.65	97.55	76.06	91.82	92.73
DINOv2_reg	ViT-L/14	**64.40**	84.73	94.01	**82.82**	92.33	98.47	**81.00**	86.06	96.67
DINOv3	ViT-B/16	61.70	82.00	93.41	79.45	**94.17**	98.77	78.5	84.24	95.45
VGGT	ViT-L/14	58.38	**88.92**	**95.81**	78.71	92.33	**99.39**	75.8	89.39	**97.27**
SPA	ViT-B/16	49.00	76.95	85.93	57.98	73.93	95.40	67.88	78.18	93.33
Croco	ViT-B/16	40.12	75.15	86.23	51.23	82.52	93.56	62.73	84.85	93.64
Croco v2	ViT-B/16	48.80	77.54	88.02	60.74	78.22	94.79	65.15	86.06	94.17
CLIP	ViT-B/16	31.44	48.50	54.19	38.65	62.27	73.01	54.00	62.12	66.06
DeiT	ViT-B/16	49.40	57.49	71.86	55.21	66.00	87.73	58.50	71.21	85.15
MAE	ViT-B/16	53.00	82.34	90.72	68.71	90.18	96.63	75.80	86.15	96.06
MASKFEAT	ViT-B/16	40.00	65.27	87.43	46.01	89.57	96.93	57.90	80.91	94.24
RANDOM	-	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00

Table: **Results of Spatial Relationship Classification Tasks.** Performance of visual models in three environments using three different probing techniques. Performance was evaluated using Accuracy.



turhancan.kargin@gmail.com

Supported by **NCN Sonata Bis Sustainable Computer Vision for Autonomous Machines** grant

ML in PL CONFERENCE 2025