

Introduction

Robots and self-driving cars must stay safe in worlds too complex to label by hand [1, 2]. They often reuse vision models trained on huge image sets, kept “frozen” (not retrained) [3, 4]. Yet safety depends on how close danger is, not just whether it has happened [5].

Research Question

Can frozen vision models tell how close an agent is to danger, without being taught what danger is?

Hypothesis: DINOv3 [6], which learned from unlabeled images, will beat ResNet-50 [7], which learned from labeled photos.

Publication and Code

Accepted for publication, National High School Journal of Science (2026). Scan for code.



Probing Frozen Visual Representations for Safety-Relevant Structure in MiniGrid LavaGap Navigation: A Geometric Comparison of DINOv3 and ResNet-50

Anastasia Titomirov

Methods

Game. MiniGrid LavaGap [8]: a 7×7 grid where an agent crosses a lava wall through one gap (15 layouts). A randomly moving agent produced 5,000 images (732 unique).

Labels. Safe, near-boundary or unsafe, based on the agent’s position (Fig. 1).

Models. Frozen DINOv3 and frozen ResNet-50 turn each image into a list of numbers (features), never seeing the labels. Comparisons: a ResNet-50 trained on the labels, raw pixels and an untrained network.

Test. After removing duplicates, 3 layouts were hidden for testing. A small classifier [9] predicted labels from the features (Fig. 2).

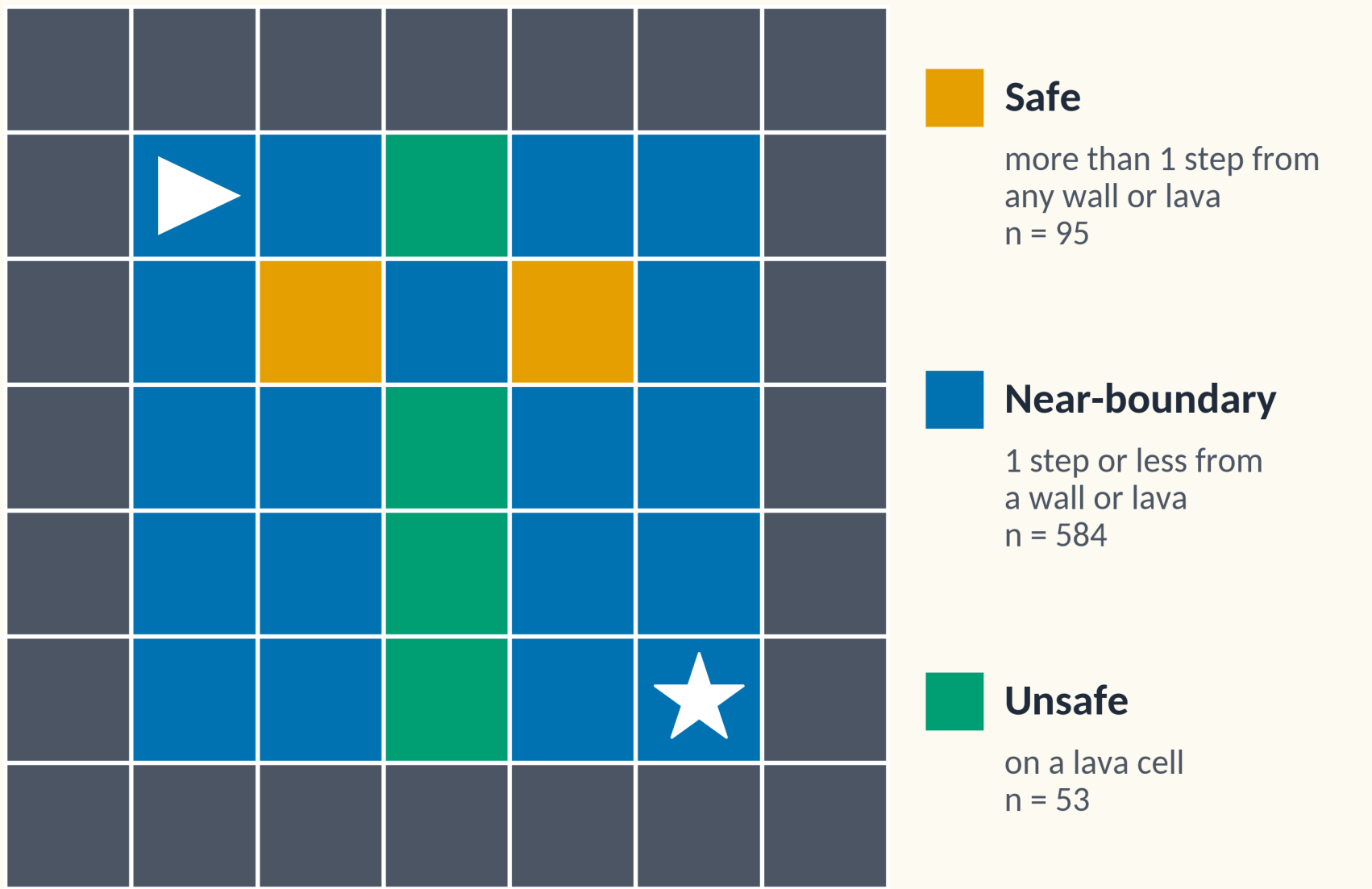


Figure 1. One layout colored by label (triangle: start, star: goal, n: unique images).

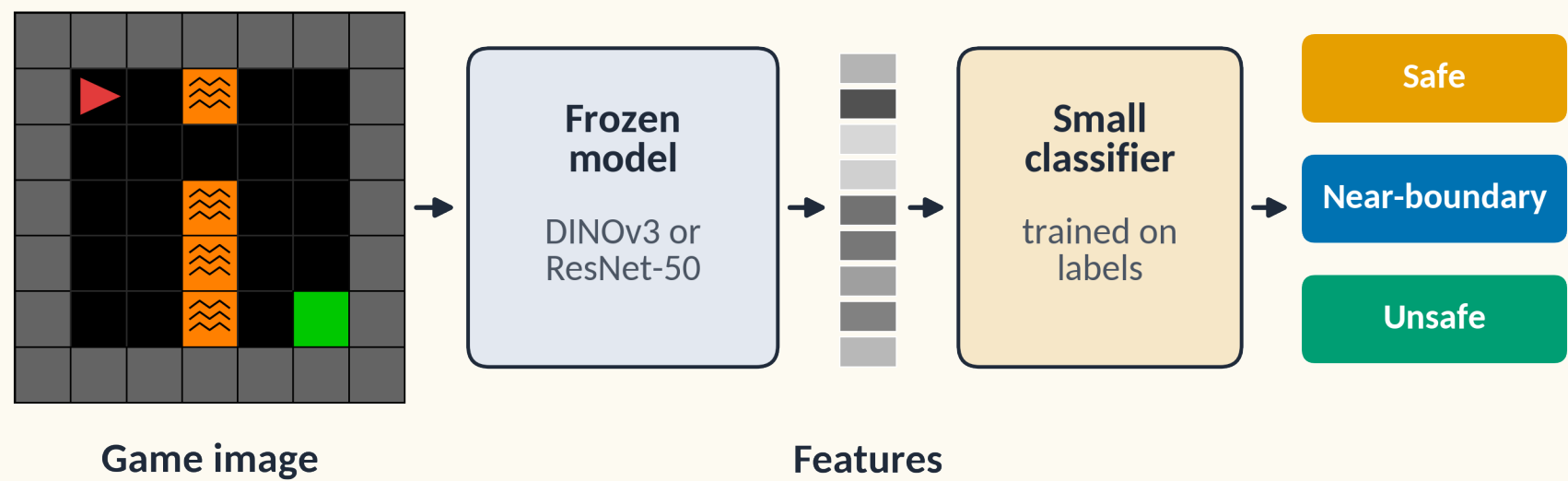


Figure 2. Only the small classifier is trained. The model stays frozen.

Results

Spotting lava is easy: even the untrained network scored 98% (Fig. 3a). The real test is telling safe from near-boundary images (Fig. 3b).

On the hidden layouts, **DINOv3 recognized all 18 safe images (100%)**, while frozen ResNet-50 recognized only 1 (53%, close to guessing).

A detector that never saw labels agreed [10] (Table 1), and on a feature map [11], DINOv3 grouped safe images tightly (Fig. 4).

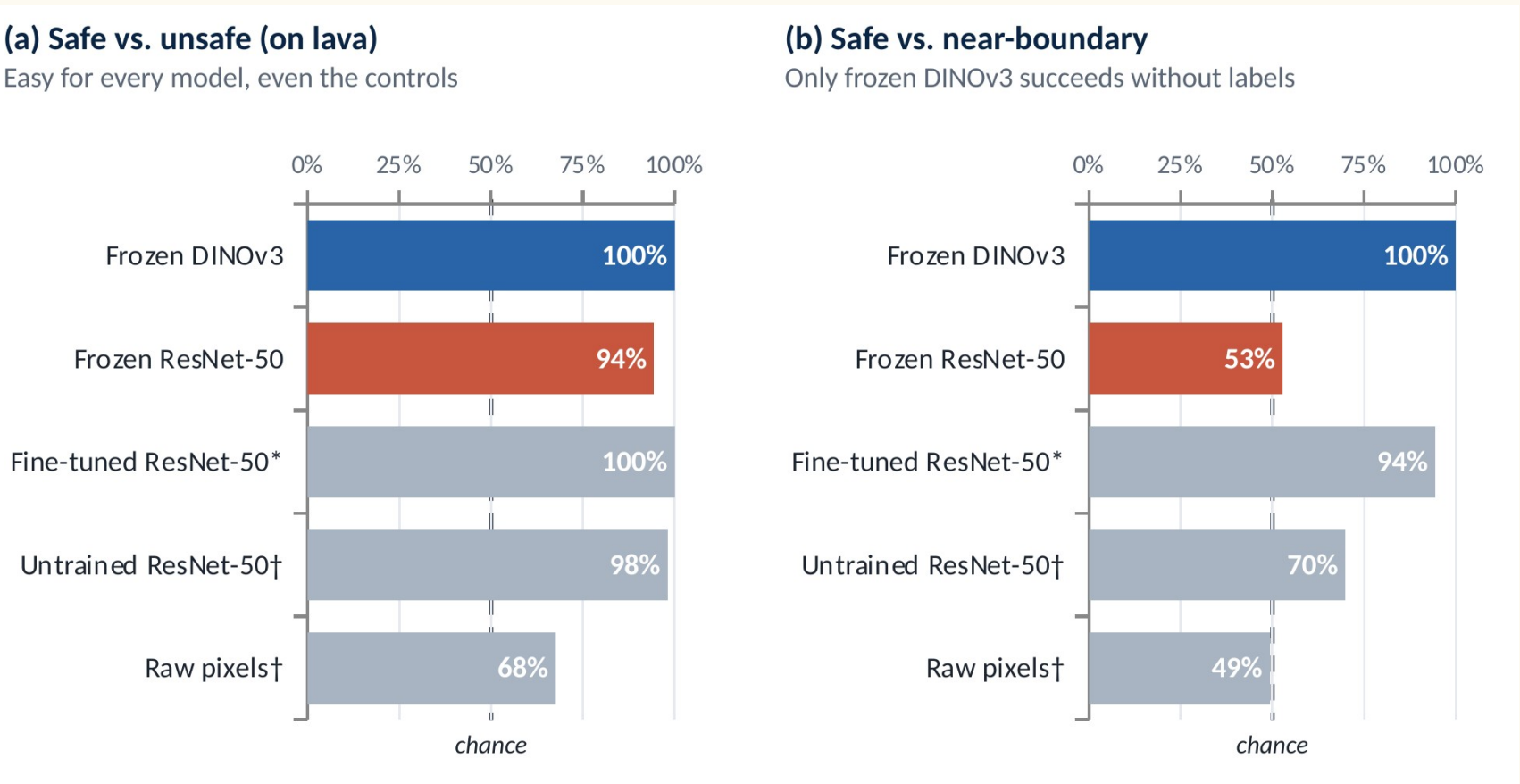


Figure 3. Score on the hidden layouts, each class weighted equally (dashed line: guessing). *Trained on labels. †Baselines.

Model	Detector score
Frozen DINOv3	0.975
Frozen ResNet-50	0.750
ResNet-50 trained on labels (reference)	0.9998

Table 1. Detector trained only on safe images (0.5 = guessing, 1 = perfect).

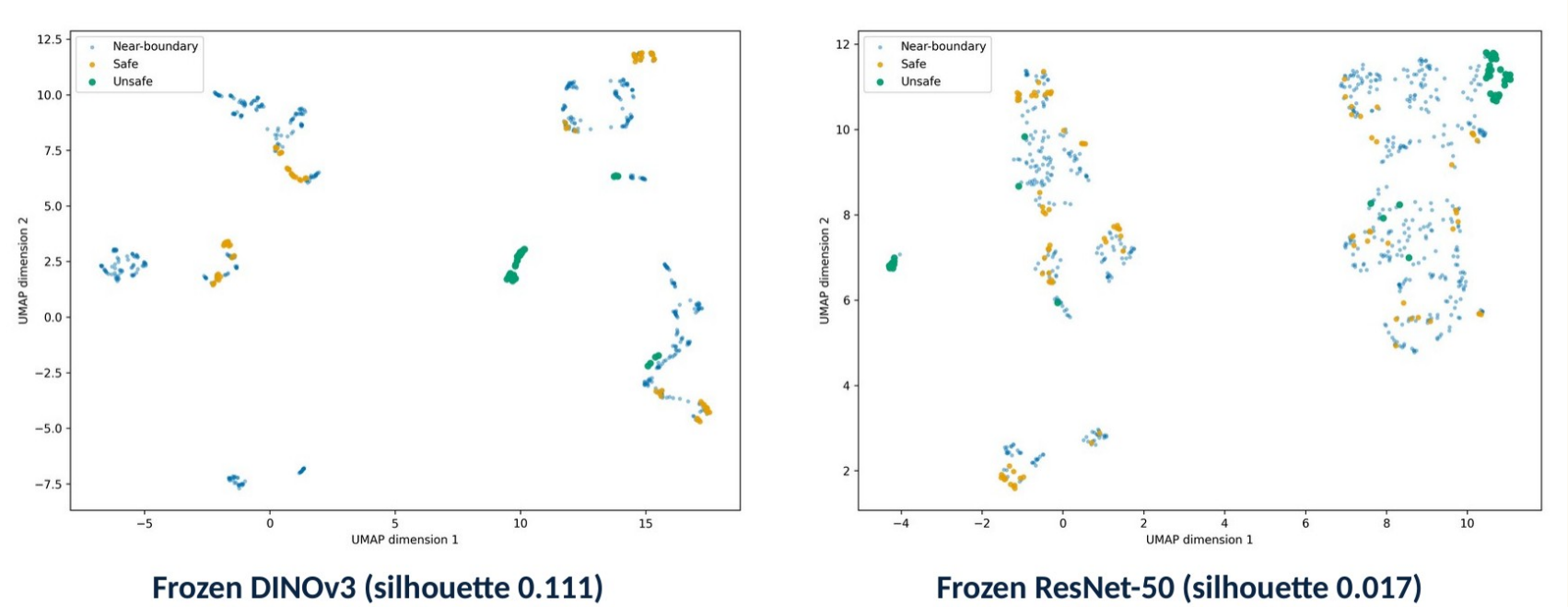


Figure 4. Feature maps of all 732 images (orange: safe, blue: near-boundary, green: unsafe).

Key Findings

- Spotting lava is easy: even an untrained network does it (98%).
- Sensing near danger is hard: on new layouts, DINOv3 found all 18 safe images, ResNet-50 only 1.
- A test without labels agrees.

Discussion

The hypothesis holds, in a narrower form. DINOv3 encodes distance to danger, matching reports that DINO features capture object boundaries [12].

This is not true safety understanding: labels depend only on position, and the models also differ in design and training data.

Test design mattered. With a random split, 95.9% of test images copied training images, so every model looked near perfect [2].

Conclusion

Frozen DINOv3 already “knows” how close an agent is to danger, even on new layouts and without safety labels. Spotting the hazard itself proves little.

Future work: recolor the lava so danger must be inferred from position, and compare models differing only in how they learned.

References

[1] Y. Liu et al. arXiv:2407.06886, 2024.
[2] Z. Wang et al. IJCAI, 2025.
[3] R. Balestriero et al. arXiv:2304.12210, 2023.
[4] S. Nair et al. CoRL, 2022.
[5] K. Nakamura et al. RSS, 2025.
[6] O. Siméoni et al. arXiv:2508.10104, 2025.
[7] K. He et al. CVPR, 2016.
[8] M. Chevalier-Boisvert et al. NeurIPS, 2023.
[9] G. Alain, Y. Bengio. arXiv:1610.01644, 2016.
[10] B. Schölkopf et al. Neural Computation, 2001.
[11] L. McInnes et al. arXiv:1802.03426, 2018.
[12] M. Caron et al. ICCV, 2021.