Is a Linear Probe Evidence of a Linear Representation?

Christian Internò[1] · Nadir Kutluozen[2] · David Klindt[2]

[1] Bielefeld University [2] Cold Spring Harbor Laboratory

Any sufficiently rich feature map admits a linear readout. So a high probe score is consistent with two very different geometries, and on its own certifies neither of them.

THE CLAIM UNDER TEST

The linear world model hypothesis says pretrained vision models encode generative factors as linearly decodable directions, and it licenses a standard method: fit a probe, declare a factor represented when accuracy is high.

Where DINOv3-B/16 actually sits

EACH DOT IS DINOV3 UNDER ONE TEST · POSITIONS SCHEMATIC

KERNEL MACHINE (H2) — z lies linearly inside a much larger curved lift

LINEAR WORLD MODEL (H1) — f is an affine image of z, up to a bounded residual

RANDOM-INIT DINOv3 — TRAINED SIM, λ=1

D1 REVERSE R² — 0.18 CLS / 0.28 patch vs 0.02

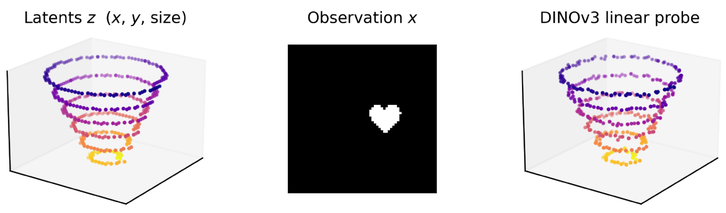
D2 SPECTRAL CONCENTRATION — aligned variance in the top ~32–64 PCs

D3 STYLE TRANSFER — Δρ 0.41 within, 0.42 across

FORWARD DECODING — ρ = 0.94 vs 0.55 random

01 The test everyone runs

Fit a probe $\hat{z} = W f(x)$ on frozen features, report R^2, call the factor represented. The implicit inference: decodable from $f(x)$ means f has learned z as an axis.



The evidence usually offered. A probe on frozen DINOv3-B/16 inverts the rendering pipeline.

02 Forward decoding is asymptotically free

Random feature maps approximate RKHS kernels, so any smooth target is linearly recoverable once there are enough features.

$$\mathbb{E}_x \, \|W f_{\text{rand}}(x) - z(x)\|^2 \le \epsilon \quad \forall \, d \ge d_\epsilon$$

Prop. 1. If the kernel's RKHS contains every latent coordinate, a *random* map is already forward-decodable at $d_\epsilon = O\big(k(R^2{+}K^2)\epsilon^{-1}\log(k/\delta)\big)$: a property of having enough features, not of having learned z.

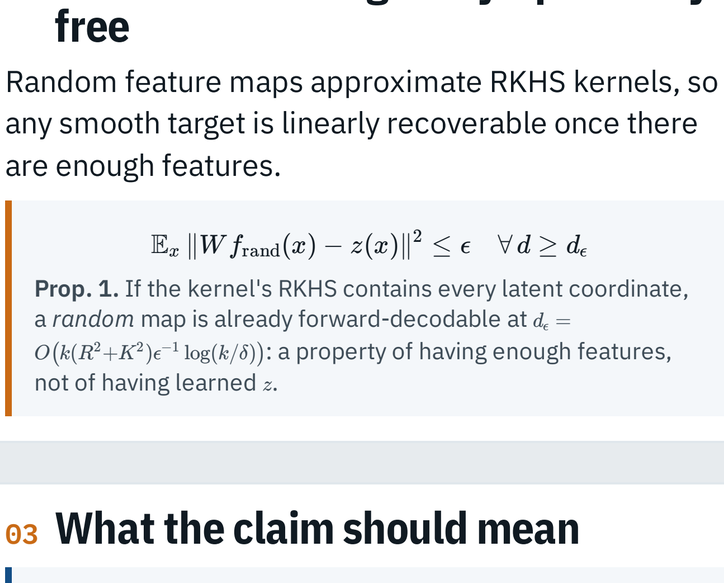
03 What the claim should mean

$$f(x) = W z(x) + \mu + \varepsilon(x)$$
$$\mathbb{E}\big[\varepsilon z^\top\big] = 0, \quad \mathbb{E}\big[\|\varepsilon\|^2\big] \le \sigma^2 \|W\|_F^2$$

Def. 1 (linear world model). f is an affine map of z up to a bounded orthogonal residual.

Stronger than mere decodability: **most of f's variance must be aligned with z**, not just z lying in the span of f.

LINEAR WORLD MODEL (H1) **KERNEL MACHINE (H2)**

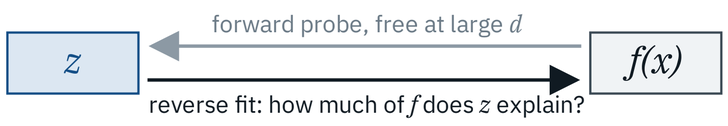


Same grid, two lifts. Cartoon above, measured at $h{=}1024$ below. **Trained is a plane, random is a saddle.**

04 Three asymmetric diagnostics

D1 Reverse predictivity

$$R^2_{\text{rev}} = 1 - \frac{\mathbb{E}\|f - W_{\text{rev}} z - \mu\|^2}{\mathbb{E}\|f - \mathbb{E}f\|^2} = \frac{\operatorname{tr}(A\Sigma_z A^\top)}{\operatorname{tr}(A\Sigma_z A^\top) + \operatorname{tr}\Sigma_\varepsilon}$$

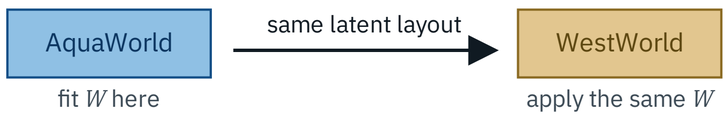


Prop. 2: small whenever a high-dimensional z-independent residual dominates, **even at forward $R^2 = 1$.**

D2 Spectral concentration

A linear world model reaches 1 at $k = k_{\text{lat}}$; a kernel machine **decays as k grows.**

D3 Out-of-distribution transfer

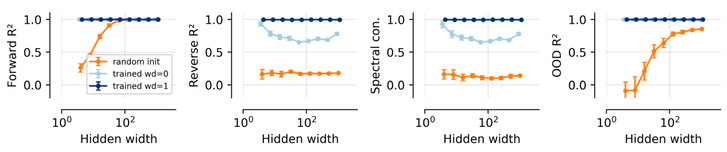


Decodability inside the training support alone is consistent with overfitting a rich kernel there. A real world model **extrapolates.**

Under Def. 1 all three are near-tight; **under a kernel machine** forward R^2 stays ≈ 1 while all three fail.

05 Validation where ground truth is known

2D latents lifted by a two-layer ReLU MLP at matched width h: **random init** against **trained**, weight decay $\lambda \in \{0, 1\}$ standing in for "be a linear world model".



Forward saturates, the rest separate. Forward $R^2 \to 1$ by $h \approx 100$ for random init while the other three stay low. Bars $\pm 1\sigma$, 3 seeds.

Top 2 PCs of f at $h{=}1024$, colored by z_0: random folds the latent square into a blob, trained keeps it axis-aligned.

- Without weight decay, reverse R^2 and spectral concentration sag to ≈ 0.7 while forward and OOD stay pinned at 1. Def. 1 is a **tunable property of training pressure**.

06 SVG-World: one layout, two styles

Both panels show the canonical scene at $z = 0$; arrows mark the two position latents per object. Only the rendering differs.

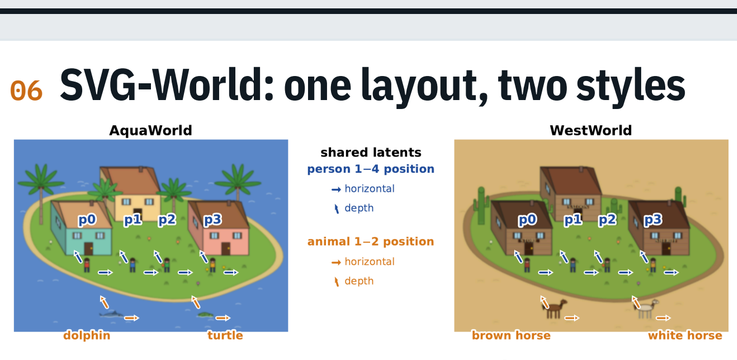
`32-dim latent, 20 active` `4 people: x, z, facing, hue`

`2 animals: x, z` `cairosvg 224 × 224`

`N = 10,000 / style` `12 position latents`

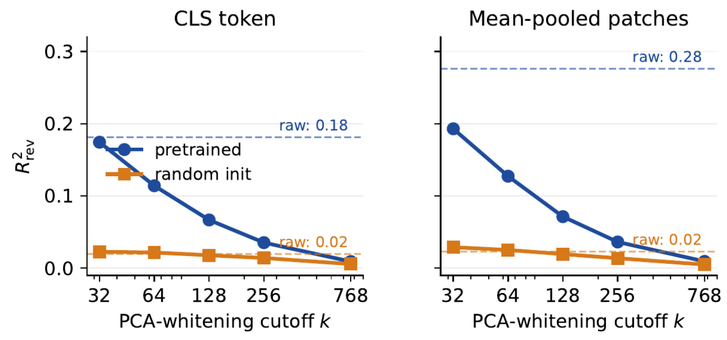
07 DINOv3-B/16 under all four tests

Forward: pretraining, not feature richness



Per-latent ρ on CLS features: mean 0.94 pretrained, 0.55 random. The asymptotic regime does *not* dominate at $d = 768$.

D1 and D2: a 10× gap, and where it lives



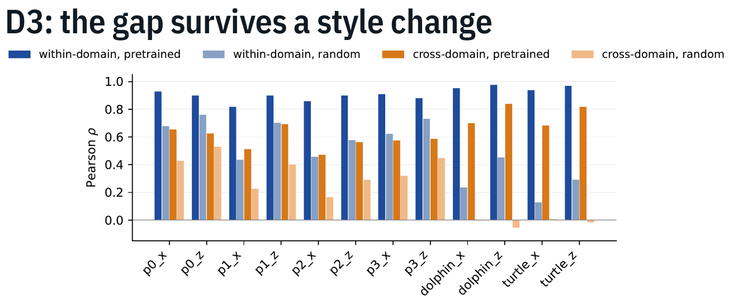
Pretrained's aligned variance sits in the top ~32–64 PCs, near the 20 active latents, and is gone by $k = 768$. Dashed lines are raw features.

10× reverse gap, pretrained vs random

1.7× forward gap, same features

0.01 R^2_{rev} at full whitening

D3: the gap survives a style change



Absolute values fall under a style change (0.91 → 0.65) but the gap does not: $\Delta\rho = 0.41$ within, 0.42 across. **Style invariance is the strongest signal here.**

What this buys, and what it costs

§9 DISCUSSION

Forward decoding does show something

- At $d = 768$ the asymptotic regime does not dominate, so the strongest "kernel machine in disguise" reading is ruled out.
- Sharpest on small foreground objects: animal $\Delta\rho = 0.58$ within-domain, ≈ 0.76 across styles.

It does not certify Def. 1

- DINOv3 sits partway: above a random-feature null on alignment and spectral structure, well below the strict regime.
- Self-distillation applies no pressure to make f an affine image of any external z. Bottlenecked objectives should do better.

Why it matters, and the limits

- The curved structure orthogonal to the probe is where OOD generalization, robustness, and compositional transfer live.
- SVG-World is deliberately simple, the simulation trains f with a supervised loss, and every analysis is ridge based.

Report reverse R^2, the spectral curve, and a cross-condition transfer next to every probe. Forward accuracy on its own is not evidence of a linear representation.

Rahimi & Recht 2007 · Alain & Bengio 2017 · Locatello et al. 2019 · Stringer et al. 2019 · Bernardi et al. 2020 · Schott et al. 2022 · Oquab et al. DINOv3

Full references and appendices in the paper.