SPATIAL MIXING PREDICTS THE READ-OUT DEPTH FOR GEOMETRY IN VISION FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense geometric tasks are read off a vision transformer's output block by convention, under the suspicion that depth trades spatial detail for abstraction. We find instead that the depth at which a linear probe best decodes a geometric attribute rises with the attribute's ground scale, so the best depth belongs to the attribute and not to the network alone. Over 32 backbones this ordering appears in every one whose coarse-scale representation is still improving at the output block, and it transfers to indoor photographs. Our account is that each block mixes a token with its neighbours over a length that grows with depth, so depth loses geometry below that length and keeps what is coarser. We measure that length without labels, as the radius over which a token responds to one occluded input patch. On overhead imagery, where every token has a known ground scale, dividing each target's scale by that radius brings the retention curves of nine matched backbones together for surface normals and aspect and not for band-passed height, and no rival length of the mixing does this. Blurring an intermediate block reproduces the output block's loss at a width the radius predicts. Narrowing late attention as far as semantic accuracy allows returns the lost surface geometry on three of six backbones and worsens it on none. Mixing is therefore part of the cause. An earlier block still holds the lost detail, and on the same imagery warping a tile by a known map names it without labels or training, recovering 80 to 88 % of what a labelled sweep would gain over all 32 backbones.

1 INTRODUCTION

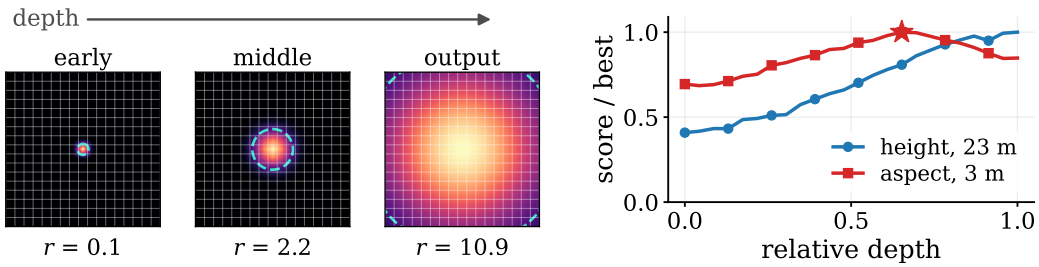
A vision transformer used as a dense backbone is almost always read at its output block (Li et al., 2022). For semantic tasks that convention is well founded (Park, 2026), and for geometric tasks it is widely suspected to be wrong (El Banani et al., 2024), since deep representations trade spatial detail for abstraction (Amir et al., 2021). The suspicion is not actionable. It does not say which attribute is affected, by how much, for which backbone, or what to do instead, and the usual remedy of fusing intermediate blocks lacks a criterion for choosing them.

This paper turns that suspicion into a measurement, then a model, then a rule that needs no labels (Figure 1). We work on overhead imagery, where metric geometry is dense ground truth on the same pixels as semantic labels, over cities that can be held out, and a known ground sampling distance gives every scale in metres. We probe nine ViT-L backbones matched in architecture and ground footprint, block by block, against geometric targets whose scale varies within one family, so that scale moves and the task does not. The finding is that there is no one best block for a network. Which block is best moves with the target's scale, fine geometry peaking at intermediate blocks and coarse geometry at the output block. The same reading repeats over 32 backbones and nine pretraining objectives, and on indoor photographs.

We explain the ordering by spatial mixing. Each block blends a token with its neighbours over a length that grows with depth, so depth does not degrade geometry uniformly. What it loses is geometry below the scale over which the network has begun to mix tokens.

We measure that length without labels, as the radius over which a token responds to one occluded input patch. It is the length at which the ordering turns, and it calibrates the ordering, since dividing a target's scale by it brings different backbones' retention curves together.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) mixing widens with depth

(b) two targets, opposite trends

Figure 1: The cause and the phenomenon it produces, for DINOv3-Web. (a) Occluding one input patch gives a response radius r_ℓ in tokens growing two orders of magnitude with depth, on DFC2019 at an 8 m token, as a schematic with radii measured at the first, middle and output block. (b) Two targets differing only in scale move in opposite directions with depth, on US3D at a 5.84 m token.

Table 1: What the paper claims, what each predicts and where that is tested, in section order. The first row is the observation, the next two its calibration and its cause, and the last the consequence.

claim	prediction	where it is tested
spatial scale selects depth	a coarser target peaks later	Section 5, probes
the radius calibrates the ordering	curves collapse against s/r_L	Section 6, nine backbones
spatial mixing is part of the cause	changing mixing moves fine geometry	Section 7, blur and window
the repair is a choice of block	an earlier block keeps the detail	Section 8, unlabelled warp

Contributions. The paper makes one claim, that a target's spatial scale sets the depth at which it is best read, and Table 1 lists what supports it in the order Sections 5 to 8 test it. We grade that support rather than merge it. The ordering is observed directly. That mixing is part of its cause rests on two manipulations. That the response radius calibrates the ordering is our narrowest claim, established on overhead imagery. Three things are new. The peak moves with scale in each of the 32 backbones that is still abstracting at its output block, and on photographs, and two controls put the loss on spatial scale and not on modality. The length that places the peak is measured without labels, no rival length of the mixing calibrates the ordering, and each quantitative prediction it makes is tested against an exact permutation null. And the block that keeps the detail is named without labels, which is the criterion the fusing remedy lacks, and on overhead imagery that choice holds over all 32 backbones.

Relation to concurrent work. That geometric precision peaks at intermediate blocks before yielding to semantic abstraction has recently been reported on natural images, by Zhou et al. (2026) for DINOv2, MAE and iBOT and by Sanghavi (2026) for a supervised ViT-B/16. We reproduce it in a new domain as validity evidence and claim none of it, since what is new here is that the peak moves with scale. What those reports establish over three backbones and one, we measure over 32, adding the length that places the peak and an unlabelled rule that names the block. Both intervene on feature directions, where we intervene on the mixing itself.

2 Related work

Where information lives, and which blocks make good descriptors. Linear probing of intermediate representations goes back to Alain & Bengio (2016), and units at different depths responding to structure at different scales to Zeiler & Fergus (2014). For vision transformers, Raghu et al. (2021) established that mean attention distance grows with depth so deep blocks aggregate globally, and Park & Kim (2022) characterised self-attention as a low-pass filter, on which Section 4 builds. Both are statements about a network. That a feature has a scale at which it is best detected, and that the scale is what should be selected, is the older idea of scale space (Lindeberg, 1998). The response radius we measure is the transformer counterpart of the effective receptive field (Luo et al., 2016), which is likewise far narrower than the nominal one and grows with depth. That intermediate blocks accordingly make better dense descriptors is established practice, following Amir et al. (2021) and

Table 2: The nine backbones, all ViT-L with 1024 channels and 24 blocks. The response radius r_L at the output block is in tokens on band-limited US3D at a 64×64 grid, which is the protocol this table and every ranking of backbones in this paper are measured under, and it is the length Equation 2 needs. Rows are in ascending r_L. A radius read under another protocol names it where it is used. Matching accuracy is under a 10° rotation on US3D at $\tau = 5.84$ m and real pairs on Maxar at $\tau = 4.88$ m, both as retention of the best block.

backbone	pretraining objective	r_L	matching	real pairs
Scale-MAE-L	masked reconstruction	0.71	0.962	0.947
SatMAE-L	masked reconstruction	0.88	0.911	0.949
MAE-L	masked reconstruction	0.98	0.996	1.000
DOFA-L	masked reconstruction	1.00	0.294	0.179
DINOv3-Sat	self-distillation	1.33	0.847	0.835
CLIP-L	language alignment	6.80	0.254	0.163
DINOv2-L	self-distillation	7.49	0.453	0.375
ViT-L ImageNet-21k	classification	7.63	0.263	0.246
DINOv3-Web	self-distillation	7.67	0.698	0.845

the correspondence literature built on Caron et al. (2021), and fusing several stages is standard in DPT (Ranftl et al., 2021) and hypercolumns (Hariharan et al., 2015). We take the widening aggregation as given and measure what it costs, per scale and per backbone.

Training-free refinement and model selection. Stochastic resonance transformers (Lao et al., 2023) aggregate features over sub-token input shifts and improve dense tasks by up to $14.9\,\%$ without training. That attacks quantisation by tokenisation, whereas the effect we measure and repair is depth-wise mixing, so the two causes are separate. Transferability estimation, such as LEEP (Nguyen et al., 2020) and LogME (You et al., 2021), ranks checkpoints for a task. We measure a length inside one network rather than a score, read against a target's own scale.

3 Setup

Conventions. A backbone has $L + 1$ blocks whose outputs we index by $\ell \in \{0, \dots, L\}$, where $\ell = 0$ is the first block's output and $\ell = L$ the output block's. Relative depth ℓ/L is then comparable across architectures. Throughout a subscript indexes a block and a superscript names a target. Writing q_ℓ^t for the score of target t at block ℓ, its retention is

$$\eta^t = \frac{q_L^t}{\max_\ell q_\ell^t} \quad \text{for scores where larger is better, and} \quad \eta^t = \frac{\min_\ell q_\ell^t}{q_L^t} \quad \text{for error metrics,} \tag{1}$$

where the extremum is over the validation split and every score reported is on the test split. A value $\eta^t = 1$ means nothing was lost by the output block and $\eta^t > 1$ that it beat the block validation chose. The matching accuracy of Section 8 has no such split.

Four symbols are lengths. Two are in metres, τ for the ground size of one token and s for the ground scale of a target. Two are in tokens, r_ℓ for the response radius at block ℓ, defined below, and k for a square window's width. Three more are statistics and carry no block index, ρ for a Spearman correlation, R^2 for a coefficient of determination and p for significance.

Data. The main probe battery runs on the DFC2019 large-scale semantic stereo data (Le Saux et al., 2019; Bosch et al., 2019), resampled to 1.0 m over Jacksonville and Omaha. That gives 800 tiles of 512 pixels with above-ground height and six semantic classes, and probes are trained on Omaha and tested on Jacksonville, so none can win by memorising a location. A 1.0 m repack cannot resolve a roof facet inside one token, so the ordering is tested on US3D-Enhanced at 0.365 m over Jacksonville. That is 400 tiles of 1024 pixels split into disjoint spatial bands. Real bi-temporal pairs come from Maxar Open Data, where every acquisition of an event is orthorectified onto a shared grid at 0.305 m and indexed by quadkey. We take 61 usable pairs, all acquired before the event and a median of 65 days apart, so the ground is unchanged. At a 64×64 token grid the three sources give $\tau = 8.0$, 5.84 and 4.88 m, and DFC2019 at 32×32 gives 16 m.

Backbones. Every backbone is a ViT-L with 1024 channels and 24 blocks, so probe capacity is identical and differences reflect representation content rather than width. The set spans self-distillation on natural images, with DINOv2 (Oquab et al., 2023) and DINOv3-Web pretrained on LVD-1689M (Siméoni et al., 2025), and the same architecture and objective on satellite images, DINOv3-Sat on SAT-493M. It also spans masked reconstruction with MAE (He et al., 2022), Scale-MAE (Reed et al., 2023) and SatMAE (Cong et al., 2022), multimodal masked modelling with DOFA (Xiong et al., 2024), supervised classification on ImageNet-21k (Steiner et al., 2021) and language alignment with CLIP (Radford et al., 2021). The DINOv3 pair is the controlled comparison, since it varies only the pretraining corpus. The DINOv2 checkpoint is the ViT-L/14 with four register tokens (Darcet et al., 2024) and not the original. Each backbone is fed the input size that makes its token grid cover the same ground footprint, so all are scored against identical targets.

Probes and targets. Read-outs are linear, with the ridge coefficient chosen on the validation split and confidence intervals from bootstrapping whole tiles rather than tokens. Semantic targets are the six classes. Coarse geometric targets are the mean above-ground height of a token, that height minus its class-conditional mean, height on building tokens only, and the signed height difference of two tokens, the last three of which a semantic classifier cannot fake. Differential targets are the surface normal of the height field and its aspect, the compass direction of steepest descent, scored only where the slope defines that direction. To vary s within a family we band-pass the height field at five octave-spaced scales and compute normals and aspect at five and three Gaussian widths.

The response radius. We occlude one input patch, recompute the representation, and report the centre of mass of the per-token change over distance on the token grid as r_ℓ in tokens. Distances use channel-whitened features, because raw cosine distance collapses in late blocks where token directions share a dominant component, which alone would make an output block look insensitive. That choice moves a radius by up to 1.06 tokens and moves neither of the two tests that read it as a length (Appendix A). The measurement needs no labels, so it runs on any imagery, and is a property of the backbone rather than the dataset, since its rankings on DFC2019 and Maxar Open Data agree at $\rho = +0.967$.

Which tests the claims rest on. Five tests are confirmatory, and we label them C1 to C5 in the order the sections report them. Section 5 asks whether the best block rises with target scale (C1), whether the premise it defines separates the backbones that order (C2) and whether a photometric target is lost harder than a geometric one of the same scale (C3). Section 6 asks whether the response radius calibrates the ordering (C4), which its collapse and the blur width of Section 7 read four times between them, so C4 enters on the least significant of the four. Section 7 asks whether narrowing late attention returns the surface normal (C5). Corrected for their own multiplicity, all five survive at 0.05 under both the Benjamini and Hochberg procedure and Holm's, which is valid under any dependence. Holm's largest adjusted p is 0.036, on C4. Everything else is secondary and corrected for nothing, since it bounds or describes a claim rather than establishes one. Appendix B maps every claim below to the measurement behind it, and every section has one appendix that reports it in full.

4 Depth as spatial mixing, and the two propositions it yields

We state the account the next three sections test, as a set of sign restrictions we never fit, written compactly as a product. It has two ingredients. The first is abstraction, and we write $g_\ell > 0$ for what the processing up to block ℓ contributes to the score of any target. The second is mixing, since block ℓ also blends each token with its neighbours over the response radius r_ℓ. A target defined at ground scale s then survives that blending in proportion to an attenuation A of the ratio of $r_\ell\tau$ to s. We call $r_\ell\tau$, the ground length block ℓ has mixed over, its aggregation scale. The score of target t read from block ℓ is

$$q_\ell^t = g_\ell \, A\Big(\frac{r_\ell \, \tau}{s^t}\Big), \qquad \text{with } A \text{ non-increasing and } g_\ell \text{ non-decreasing in } \ell. \tag{2}$$

In words, depth helps through g_ℓ and hurts through A, and which of the two wins at a given block is settled by the target's own scale, since s^t enters nowhere else. A coarse target is barely attenuated and rides the abstraction to the output block, and a fine one peaks where the two effects cross, which is the pair of curves Figure 1b draws.

One of the three quantities is measured, namely r_ℓ, the occlusion radius of Figure 1a, and neither g_ℓ nor the shape of A is estimated anywhere below. We test the interaction between depth and target scale that the ratio forces. The equation needs r_ℓ only to increase, and the measured growth is multiplicative rather than additive in depth, which follows if the reach of a block is set by features already pooled to r_ℓ. The blur equivalence of Section 7 shows r_ℓ low-passes the token grid rather than deleting it, an average pool reproducing a later block. That the abstraction factor g_ℓ is not measured costs nothing. Over the grid of blocks against targets the equation is additive in logarithms, with $\log g_\ell$ a row effect depending on the block and not the target. Profiling it out leaves a restriction on the interaction alone, that the two meet only through $r_\ell \tau / s^t$ and in no other way.

That ratio $r_\ell \tau / s^t$ is a difference of logarithms. A power of it puts the maximum of q_ℓ^t at one block for every target, whatever the exponent, so a measured ordering shows A is curved and not only decreasing. A model with no interaction predicts the same, which is the competitor Section 5 tests the ordering against. What is not identified is the shape of A and not the interaction itself (Appendix C). So every test below is ordinal, an ordering, a crossover or the sign of an intervention and never a score. Section 5 tests the ordering, Section 6 the length it turns at, and Section 7 whether mixing at that length is a cause.

Proposition 1. Assume also that A is differentiable and that $x \mapsto x\,|\mathrm{d} \log A / \mathrm{d}x|$ is non-decreasing, which says that one further block of mixing costs proportionally less on a coarser target and holds for every $A(x) = \exp(-c\,x^\beta)$ with $c, \beta > 0$. Then the retention η^t of Equation 1 is non-decreasing in s^t, and so is the block at which q_ℓ^t is maximal. That condition gives $\log q_\ell^t$ increasing differences in the block index and the target scale, so the monotone comparative statics of Topkis (1978) and Milgrom & Shannon (1994) order the maximisers and retention follows as a minimum of differences each non-decreasing in s^t. Appendix C.1 proves it and Section 5 tests it.

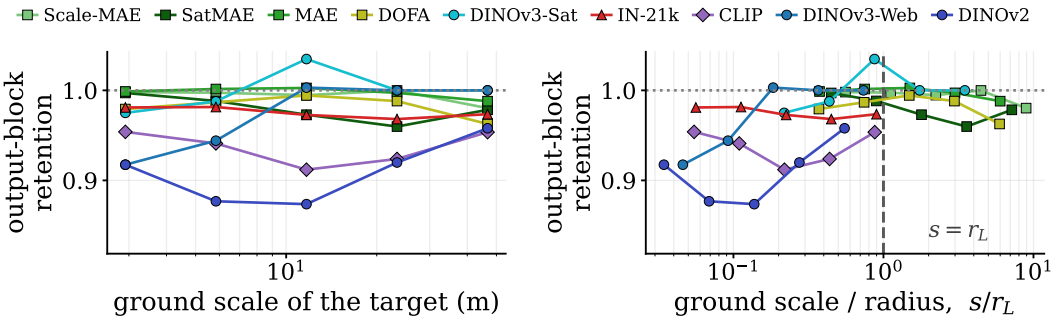
Proposition 2. Let the feature carry the target field through the mixing kernel plus content uncorrelated with it, and let a linear read-out see a square window of k tokens. The best achievable coefficient of determination rises with k but is bounded, for every k, by the Wiener value for an unbounded window (Wiener, 1949). Appendix C.1 states that value, which retains no power at the spatial frequencies where mixing has pushed the target below the uncorrelated content. Whatever the window sees, no linear read-out recovers the target there, and none at all does when field and content are jointly Gaussian, the Wiener estimator being the conditional mean. Widening the window moves the read-out towards that ceiling and cannot move it, so the prediction is a recovery saturating in k near the response radius and below an earlier block, which Section 7 tests.

5 The spatial scale of a target selects the depth that encodes it

Proposition 1 predicts that a coarser target peaks deeper, the paper's central empirical claim. For DINOv3-Web at $\tau = 5.84$ m the relative depth ℓ / L of the best block rises with s and never falls, from 0.65 for the finest normal and aspect to the output block above 23 m. Retention rises monotonically with s in all three families, surface normals from 0.917 to 1.003 over 2.9 to 46.7 m (Figure 4). Only the target changes between points, on the same cached features, so neither model nor probe capacity does. Coarse geometry is therefore not lost at all, against the abstraction story. Figure 2a shows the same curves for all nine backbones, and Appendix C.1 measures how monotone each one is.

A model that keeps the block effect and drops the interaction puts the best block at one block for every target scale, so it forbids the ordering outright. Permuting the target scales over the observed peak blocks tests it exactly in each backbone, and combining the nine by Fisher's method rejects it at $p = 0.0074$ (C1, Appendix C.2).

The premise is measurable, and every backbone satisfying it orders. A coarse target is barely attenuated and follows g_ℓ, so the ordering reaches the output block only in a network whose abstraction is still accumulating there. That is the premise of Equation 2. We read it off the block at which coarse mean height is best decoded, counting it as holding within three blocks of the output. We score the ordering in a backbone by the Spearman ρ between a band-passed target's scale and the depth of its best block, counting it as present at $\rho \geq 0.5$. Over 32 backbones from ViT-S to ViT-g, all 9 that satisfy the premise show the ordering, at a median ρ of $+0.894$ against $+0.105$ for the other 23, an exact $p = 0.0007$ (C2). The separation survives counting each training recipe once,



(a) against ground scale

(b) against scale over radius

Figure 2: One measured length carries the ordering. Surface-normal retention for the nine backbones of Table 2 on US3D, the same points in both panels, against (a) the ground scale s of the target and (b) s divided by that backbone's own response radius r_L at the output block, read on DFC2019 at the analysis grid, where the within-bin standard deviation over six logarithmically spaced bins falls by $24\,\%$. That is an exact $p = 0.0074$ over all 362880 reassignments of the nine radii. Retention reaches the output block's own score at $s \approx r_L$, the crossover Equation 2 places there. Surface aspect tightens by $36\,\%$ at $p = 0.031$ and band-passed height, which the model puts above the aggregation scale, does not tighten at all.

since six of the nine are three capacities each of two recipes ($p = 0.0036$). It also survives reading the premise off the semantic probe instead of height ($p = 0.0062$), so it is not one height signal read twice (Appendix D.3).

The premise is sufficient and not necessary, since 9 of the 23 order anyway, over six objectives and seven recipes, two of them windowed hierarchies and none descended from a checkpoint that satisfies it. That puts the ordering in 18 of the 32, over seven of nine objectives, and in 4 of 6 on indoor photographs (Appendix D.5). There the output block reads the finest surface normal 5.3 to $7.6°$ worse than the best block on three of the six (Appendix F.3).

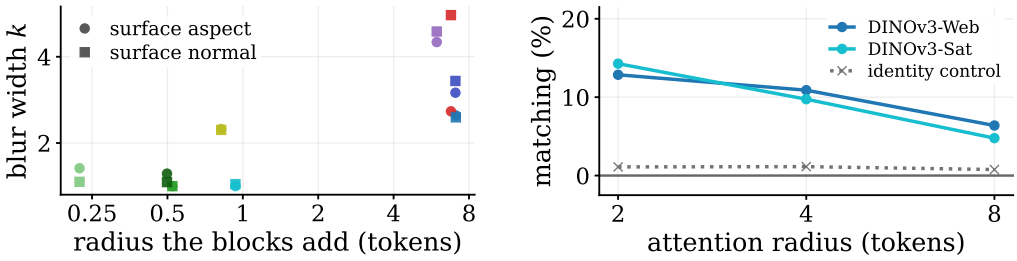
What the loss falls on, and what escapes it. The cheapest deflationary reading of the ordering is that deep blocks are simply worse at fine targets, geometry among them. Two controls pulling opposite ways reject it, on US3D at $\tau = 5.84\,\mathrm{m}$ over the nine. The unchanged semantic head rescored on class-boundary tokens is fine and survives, retaining 0.978 where surface aspect retains 0.921. The response radius predicts aspect retention across the nine at $\rho = -0.667$ and boundary retention at $\rho = +0.067$, so what governs the geometric loss does not reach the semantic one. The photometric analogue of surface aspect is fine, is not geometric, and dies harder, retaining 0.462 against that 0.921, lower in nine of nine at $p = 0.0039$ (C3). It is surface aspect built on the image in place of the height, so the two differ in the field and nothing else. The loss therefore follows spatial scale and not modality, as Equation 2 says.

So the depth at which a target is best read is a property of that target's scale and not of the network alone. Which length that scale is measured against is the next question.

6 The response radius calibrates the ordering

If the ordering is mixing, it should turn at the scale the network has mixed over. Two predictions then use the value of the response radius and not only the order of blocks. Different backbones' retention curves should collapse once each target's scale is divided by that backbone's radius, which this section tests. And a blur that makes an earlier block as bad as a later one should widen with the radius those blocks add, which Section 7 tests, since the blur is also a manipulation of the mixing. Both read one radius, measured once per backbone and never fitted.

On one backbone the turn can be read in metres. DINOv3-Web's best block reaches the output block for targets near 20 m, and its output block's response radius on this imagery is 3.75 tokens, or



(a) a blur adds mixing

(b) a window removes it

Figure 3: Adding mixing and removing it both move fine geometry, in the order this section makes them. (a) The blur width k in tokens that makes block 16 as bad as the output block, against $\sqrt{r_L^2 - r_{16}^2}$, the radius those blocks add, on US3D, in the colours of Figure 2. (b) Narrowing late attention instead returns matching accuracy, in per cent of an unmodified run, against a control that does not move.

21.9 m, where one token is 5.84 m. That radius is read at native detail, on the imagery the probes read, since the comparison is in metres, where Table 2 ranks backbones band limited.

Dividing by the radius collapses the retention curves. Plotting retention against s/r_L rather than s brings different backbones' curves together for the two families the model places near the aggregation scale (Figure 2). Over six logarithmically spaced bins it tightens the within-bin standard deviation of retention by 24 % for surface normals and 36 % for aspect and leaves band-passed height where it was. Dividing by any per-backbone constant can tighten a spread, so the null reassigns the same nine radii to the same nine backbones. Over all 362880 relabelings its median is +7 % where the two measured values sit at exact $p = 0.0074$ and $p = 0.0307$ (C4). The collapse is selective, since one length governs fine surface geometry where coarse height is set by the abstraction a backbone has accumulated, band-passed height sitting at $p = 0.76$. The tightening does not depend on the nine, since thirteen further releases satisfying the same invariant tighten too, by 17.5 and 17.7 % over the twenty-two (Appendix D.1).

Of four label-free lengths of the mixing, the radius is the only one that calibrates. Mean attention distance (Raghu et al., 2021), a gradient receptive field (Luo et al., 2016) and the feature map's autocorrelation length, fixed before measuring and divided in instead, tighten neither family beyond chance and order no blur width (Appendix D.2). The claim is about the ratio of the target scale to the aggregation scale $r_\ell \tau$ and not about depth, so the ordering is visible near $r_\ell \tau$ and not far above. What the radius adds is a unit, since it makes a target's ground scale commensurable with a block, which no index of depth does. The radius is not the rule that picks a block in practice, since Section 8 does that with an unlabelled warp.

7 Mixing at the measured length is a cause of the loss

Sections 5 and 6 are observational, and reading down depth moves the radius and the block order together, so no surface of scores separates them. We therefore hold the block fixed and change the mixing, in an intervention that adds mixing, a read-out control that tests Proposition 2, and an intervention that removes it, one paragraph each in that order. Adding mixing covers every backbone of Table 2, where removing it returns the normal on three of six, each read at the strongest dose semantic accuracy allows, and worsens it on none. The attribution therefore rests on their agreement.

An intervention that adds mixing. Average-pooling a middle block over a $k \times k$ token window and re-probing asks how wide a blur reproduces a later block's score, since Equation 2 treats a further block of depth as a wider blur. On US3D with DINOv3-Web against block 16, the output block's aspect error of 49.18° and normal error of 12.59° are both matched by one width, $k \approx 2.6$. No blur reaches the output on semantics or coarse height, so the equivalence is specific to geometry below the aggregation scale. A width exists in all nine, since blurring acts after the forward pass

and, unlike the intervention below, reaches every backbone. That much any blur gives, since blurring degrades fine geometry, so what is tested is the width's value, the second prediction of Section 6. Kernels compose in quadrature, so the width should track $\sqrt{r_L^2 - r_{16}^2}$, and it does, at $\rho = +0.636$ on aspect and $+0.650$ on the normal, both at an exact p below 0.04 (C4, Figure 3a, Appendix E.1).

A read-out control, and the saturation Proposition 2 predicts. If the output block had merely averaged its neighbours, a read-out seeing that neighbourhood could undo it. Table 18 shows it cannot. A 3×3 window over the output block's own neighbourhood carries $1.32\,\mathrm{M}$ and cuts the surface-normal error by $1.4\,\%$. A bare linear read-out of one token of block 16 carries $0.008\,\mathrm{M}$ and cuts it by $7.7\,\%$. Proposition 2 says more, that widening must not help, and Table 14 confirms it, k from 1 to 9 moving the parameters from 0.27 to $10.75\,\mathrm{M}$ and leaving the error inside the seed deviation at every width. The training loss falls by $3.8\,\%$, so the wider window has nothing to find. Coarse height behaves oppositely, the same widening buying block 16 $30.7\,\%$ (Appendix E.2).

A convolutional decoder does not stand in for the block, since $2.95\,\mathrm{M}$ of it on the output block reaches $9.82°$ where a linear head reaches $9.75°$ once a $0.28\,\mathrm{M}$ branch injects block 16 into the output block. The decoder, two 3×3 convolutions, improves every read-out and reorders none. The decoder is worth about what the block is worth, 12.1 against $12.8\,\%$, and they compose, the decoder over the injected block reaching $9.02°$ (Appendix E.4).

An intervention with a dose-response. Confining self-attention from a chosen block onward to a local window of radius w tokens on the native grid, and changing nothing else, improves surface normals and degrades coarse height at once. Each moves by more than the paired interval on the probe side, a trade-off no confound gives. Over the same 71 held-out tiles in a paired bootstrap, restricting attention from block 16 at $w = 4$ moves surface-normal error by $-0.148°$ at $p = 0.0035$ (C5) and raises coarse-height squared error by $+0.638$, with semantic accuracy unchanged. On the task side of Section 8 output-block matching rises too, and it rises monotonically in the window's tightness, in every warp and from either starting block (Figure 3b), which no story of masking as mere regularisation gives. The control moves by at most 2.3 points against $26\,\%$ of matching. Only the geometric half of the loss is the mixing. The same window does not reliably return photometric aspect, the control of Section 5, moving it by $-1.03°$ at $p = 0.13$, and on DINOv3-Sat no window improves it at all (Appendix E).

The mask is the largest dose of a family, and which dose a backbone is read at is fixed by the semantic control and never by the effect. Penalising attention outside the window rather than deleting it holds that control on three more backbones, CLIP-L returning the normal by $-0.111°$ at $p = 0.011$ at $61\,\%$ of the mask. No admissible cell of the sweep makes the normal worse (Appendix E.3).

The three measurements therefore agree. A blur that adds mixing to an early block reproduces a later one, a window that removes it late gives the fine geometry back, and no read-out of the output block we train recovers what the mixing took.

8 What the loss costs on a task, and how to recover it

The cost is not the probe's, since it shows on a task with no labels and no read-out. We warp a tile by a known affine map and match every token to its nearest neighbour in the warp by cosine similarity. Matching accuracy is the fraction landing within 0.75τ of the mapped centre, and warps are in source raster pixels, so one token step is one ground distance for every backbone (Appendix A). The first half measures the cost and the second recovers most of it, and Appendix F has both in full.

The curve is an inverted U, and two failures have to be separated. The last third gives up $30\,\%$ of the peak, and early blocks fail for the opposite reason, not being invariant to the resampling the warp introduces, so the middle of the network is best (Appendix F.1). The whole-token shift asks only whether a token stays distinguishable from its neighbours, its token identity, and is reported separately because in four backbones it collapses. That is worse than losing sub-token precision, its retention running from 0.204 to 0.711 against 0.965 or better for the other five (Table 20).

Real bi-temporal pairs validate the synthetic warp. Retention on real pairs and under the synthetic rotation, measured on identical tiles, agree at $\rho = +0.967$ with an exact $p = 0.0002$, panel (b)

of Figure 6 in Appendix F.1. The agreement survives changing the dataset, at $\rho = +0.950$ between real Maxar pairs and synthetic warps on US3D.

The repair is a choice of block, and it needs no training and no parameters. The detail the output block has mixed away is still present in an earlier block, so the repair is to read that block. No read-out of the output block we train reaches one token of block 16. We match on the best single block concatenated with the output block, whitening each half separately, per backbone in Table 20. The winning block sits near relative depth 0.43, earlier than the block 16 of Section 7, and the null lands where it should, MAE-L losing nothing and the repair finding none. Where it finds one the recovery is large, reaching $99\,\%$ of the best single block for DINOv3-Web and $94\,\%$ for DINOv3-Sat while keeping the output block, which matters if the same features must serve semantics. On real pairs it beats that block by 6.4 points, and where token identity has collapsed it still recovers 51 to $72\,\%$, the output half then adding nothing distinctive (Appendix F.2).

And without labels. Two questions have to be separated, whether to leave the output block at all and which block to read, and one unlabelled curve answers both. The warp is ours, so the correspondence is known and the curve names a read-out block for a target with no labels yet. That curve rises before the output block below the aggregation scale and not above, which settles the first question, and its maximum settles the second. Scored on surface aspect under each of the four warps, that block recovers 74 to $76\,\%$ of the gap between the output block and the one a labelled sweep would pick. On the four backbones whose gap exceeds 5° it recovers 81 to $86\,\%$ (Appendix F.2).

The choice holds across the model zoo. We fixed in advance how the rule would be scored beyond the nine and applied it unchanged to all 32 backbones of the premise sweep. It recovers 80 to $88\,\%$ of the gap on surface aspect and 68 to $78\,\%$ on surface normals, and 81 to $91\,\%$ on aspect over the 23 it had never been run on, which span all nine objectives, capacities from ViT-S to ViT-g and two windowed hierarchies (Appendix F.3).

With labels the repair becomes a module rather than a choice. Injecting block 16 into a frozen ViT-L through the $0.28\,\text{M}$ branch improves surface normals by $12.8\,\%$ where a random earlier block does not, and where a $7.36\,\text{M}$ hypercolumn over seven blocks reaches $8.1\,\%$ (Appendix E.4).

9 Conclusion and limitations

A target's spatial scale sets the depth at which it is best read. Reading a dense geometric task off the output block is therefore wrong for detail below the scale the backbone has begun to mix over. That scale is a length the network sets, measurable with no labels and manipulable. The same length collapses nine backbones' retention curves and predicts a blur width, and no rival length of the mixing does either. Three things follow that the abstraction story does not. Read-out depth belongs to the target and on overhead imagery is settled before any label exists, since an unlabelled warp names the block and recovers 80 to $88\,\%$ of what a labelled sweep would gain over all 32 backbones, and 71 to $80\,\%$ on the nine whichever of three imagery sources it chooses on. Whether to move it at all is settled the same way, since the criterion pays below the aggregation scale and not above. The loss selects for spatial scale and not modality, since a boundary label survives where the photometric analogue of aspect is lost harder. The mixing account explains one part of that loss, since narrowing late attention returns the surface normal and not the luminance direction.

Limitations. That the best depth for a target moves with its scale is measured directly, in all 9 of the 32 swept whose premise holds and 9 of the 23 whose premise fails. That late mixing is a cause rests on manipulations, the blur equivalence carrying all nine where the intervention is clean on three of six once its dose is titrated. That the length is this response radius, and not a rival length of the mixing, is the narrowest claim, established on overhead imagery where a token has a known ground scale. On photographs the calibration is undecided rather than refuted, five backbones being too few to resolve the overhead effect, and a perspective token covers many ground scales at once, so that boundary is the imaging geometry rather than an outcome. The unlabelled choice of block is established on overhead imagery too, since on indoor photographs the warp names a block earlier than the one a labelled sweep picks. Across architectures the radius ranks no backbone, and a hierarchical backbone needs its output stride read beside it (Appendices A, D and F).

AI USE STATEMENT

Generative AI tools were used to assist with drafting and revising the manuscript. Specifically, they were used to improve the clarity, structure, and presentation of the text during the writing and editing process. The authors reviewed and verified all AI-assisted content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

All measurements in the paper are computed from cached features of publicly released checkpoints on public imagery. The probe targets are computed once per token grid and shared across backbones, so that every model is scored against identical labels. Geographic holdouts, ridge selection, the bootstrap procedure over whole tiles, the exact warps, the occlusion protocol and the offset estimation for real pairs are specified in Appendix A in enough detail to reproduce each number. Code will be released.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic stereo for incidental satellite images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1524–1532. IEEE, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806. IEEE, 2024.

Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 447–456, 2015.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Dong Lao, Yangchao Wu, Tian Yu Liu, Alex Wong, and Stefano Soatto. Sub-token vit embedding via stochastic resonance transformers. *arXiv preprint arXiv:2310.03967*, 2023.

Bertrand Le Saux, Naoto Yokoya, Ronny Hänsch, and Myron Brown. 2019 ieee grss data fusion contest: Large-scale semantic 3d reconstruction. *IEEE Geoscience and Remote Sensing Magazine (GRSM)*, 7(4):33–36, 2019.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pp. 280–296. Springer, 2022.

Tony Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, 1998.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, et al. Tips: Text-image pretraining with spatial awareness. In *International Conference on Learning Representations*, volume 2025, pp. 64256–64279, 2025.

Paul Milgrom and Chris Shannon. Monotone comparative statics. *Econometrica: Journal of the Econometric Society*, pp. 157–180, 1994.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International conference on machine learning*, pp. 7294–7305. PMLR, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Junghwan Park. Frost: Training-free few-shot segmentation with frozen features and nonparametric statistics. *arXiv preprint arXiv:2606.31136*, 2026.

Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4065–4076. IEEE, 2023.

Jainum Sanghavi. From edges to depth: Probing the spatial hierarchy in vision transformers. *arXiv preprint arXiv:2604.23452*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Donald M Topkis. Minimizing a submodular function on a lattice. *Operations research*, 26(2): 305–321, 1978.

Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press, 1949.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International conference on machine learning*, pp. 12133–12143. Pmlr, 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Weichen Zhou, Yawen Zou, Chunzhi Gu, Ran Dong, Haoran Xie, and Chao Zhang. Understanding geometric representations in self-supervised vision transformers via subspace intervention. In *European Conference on Computer Vision*, pp. 357–374. Springer, 2026.

A IMPLEMENTATION DETAIL

Feature extraction. For each backbone we cache the representation of every block as half-precision arrays of shape (blocks, tiles, tokens, channels), so that every probe for one backbone reads the same features. The sweeps that need no depth profile, namely the pooling and local-attention runs, cache the seven blocks 0, 4, 8, 12, 16, 20 and 23 alone. Patch sizes of 14 and 16 are reconciled by choosing the input size that makes the token grid cover the same ground footprint. For a 64×64 grid that means 1024 pixels at patch size 16, and 1022 pixels at patch size 14 followed by interpolation of the token grid from 73 to 64. Because the interpolation is applied to the token grid rather than to the image, the ground footprint of a token is identical across backbones by construction. The response radius is read on the 64×64 grid throughout, including where the retention compared against it is read at $\tau = 16$ m, since the protocol of Table 2 is held fixed wherever a radius is read.

Probes. Regression read-outs are ridge regressions solved in closed form on the GPU, with the ridge coefficient chosen from a logarithmic grid on the validation split of the geographic holdout. The block in the denominator of a retention is chosen on that split too, since selecting it on the split it is scored on would make that denominator the largest of 24 noisy numbers and push every retention below 1. The semantic read-out is a multinomial logistic regression fitted with full-batch AdamW to convergence. Angular targets are handled by regressing the sine and cosine of the angle and reporting circular error. All confidence intervals resample whole tiles with replacement, never tokens, because tokens within a tile are strongly dependent, and the resampling runs on the GPU through a padded index matrix.

Perturbation diagnostics. The occlusion measurement compares an image with a barely perturbed copy, so it runs in single precision, since reduced-precision kernel non-determinism is the same size as the signal, and the reference image passes through the same batched call as the perturbed variants. The floor of that comparison is measured rather than assumed. Pushing two unmodified copies of an image through as variants, one inside the reference's own batch and one in a later batch, gives it directly, and over the nine backbones it runs from 8.6×10^{-9} to 1.0×10^{-8}, flat in depth to within a factor of 1.12. The occlusion far field runs from 576 to 47968 times that floor, and its strength orders the backbones the way the radius does, at $\rho = +0.900$, so the mass-weighted statistic reports a long-range response rather than a level below it. Raghu et al. (2021) establish that mean attention distance grows with depth and report it against no such floor, so what the calibration adds is that the far field is a response and not the arithmetic of the comparison.

Feature distances. Distances use channel whitening with a shrinkage of 0.2, and the raw and layer-normalised variants are recorded alongside it. The choice moves the response radius by up to 1.06 tokens over the nine backbones, which is why the main text states it, and the largest move in proportion is 35.3 % on SatMAE-L. The ordering survives it, at $\rho = +0.983$ against both other variants under the protocol the collapse reads and $+0.933$ and $+0.950$ band limited, since the disagreement is inside the two groups and not between them. Recomputing the two tests that read the radius as a length under each variant, with the retention curves, the blur widths and the permutation nulls held at their published values, rejects in every differential-geometry cell, worst $p = 0.0417$, and leaves every cell of the family Equation 2 excludes null between $p = 0.65$ and $p = 0.76$. The token-identity collapse is not a whitening artefact either, since raw and whitened cosine agree to within 0.021 on every entry of the token-identity column of Table 20.

Real pairs. Two acquisitions of one Maxar quadkey image the same ground, so their correspondence is the identity map up to a residual geolocation error of a median 0.76 token, which is the scale that destroys an accuracy score thresholded at 0.75 token. We remove it as one integer offset per pair, estimated by normalised cross-correlation on the grey images, which never reads a feature and so cannot make the ground truth circular. Two filters the Maxar data forced are worth recording. Same-day acquisitions are rejected, since two strips from one pass differ in look angle but barely in illumination or season. Open water is excluded because waves decorrelate between dates, and it is judged from the first image of each pair only, on coarse-scale structure that water lacks and land has, so that the criterion cannot select for the matchability being measured. Nine of 70 candidate pairs are dropped because no single translation reconciles them, leaving the 61 used in the paper, and the

mean normalised cross-correlation rises from 0.628 to 0.791 once the estimated offset is applied. Absolute accuracy on the pairs is about half the synthetic value, since a 65-day appearance change and parallax floor every block, which is why the paper reports matching as a retention.

Local attention. The intervention replaces the attention logits from a chosen block onward with a masked version that allows only tokens within a Chebyshev radius of w on the native token grid, meaning the grid implied by the input size and the patch size rather than the interpolated grid, and leaves every weight untouched.

B What each claim rests on, and the confirmatory family

The claim map. Table 3 maps every claim Sections 5 to 8 make to the measurement behind it, the value that measurement returns and the appendix that reports it in full. It is a reading aid and introduces nothing, every row being a sentence of the main text with its evidence beside it. The five confirmatory tests are marked C1 to C5 as Section 3 labels them, C4 on both of the rows that read it, and the rows that are not marked are secondary in the sense the last paragraph of this appendix gives.

The confirmatory family. Table 4 is the family of Section 3, with the file each p value is read from. Each of its five sits under one level of Table 1. Under the phenomenon sit the ordering requiring an interaction, the premise selecting the backbones that order and a photometric target of the same scale being lost harder than a geometric one, C1 to C3. Under the calibration sits the collapse of the retention curves on the radius, C4. Under the mechanism sits narrowing late attention returning what mixing removed, C5.

Two rules fix the membership and both are stated here rather than inferred from the table. Each confirmatory claim enters through exactly one test, so a claim read several ways contributes one member and never several. And where a claim is read by more than one test the member is the least significant of them, which is the only direction a selection rule may point once the answers are known. Only C4 is read by more than one test. The claim that the measured length is the one the account needs is read four times, by the collapse of the retention curves on the surface normal and on surface aspect at $p = 0.0074$ and 0.0307, and by the two predictors of the blur width at 0.0120 and 0.0359. It enters at the largest of the four. Taking the most significant instead, which is the reading we do not claim, would put the largest adjusted value of the whole family at 0.015.

Two corrections are reported because they answer different questions and neither is the flattering one by construction. Holm's step-down procedure controls the probability of even one false rejection and is valid under arbitrary dependence, which matters here since several members are computed from overlapping measurements on the same backbones. The Benjamini and Hochberg procedure controls the expected share of false rejections among those made, which is the question five distinct claims raise rather than one. At 0.05 both keep all five.

What is secondary is everything else, and it is worth saying what kind of thing that is, since a long list of uncorrected tests is only a problem if the claims depend on them. They are of four kinds. Scope statements, which say where a claim holds and are read as counts rather than as rejections, the four nested separations inside Appendix D.3 among them. Bounds, which limit a claim the paper makes and can only weaken it, so a correction that made them harder to reject would work in our favour. Sensitivity analyses, which re-run a confirmatory test under a changed setting and are read for whether the answer moves rather than for a level. And negative results, which a multiplicity correction cannot rescue and does not help. The intervention sweep of Appendix E.3 is the largest single group of uncorrected tests, at twenty cells, and it is read by its semantic control rather than by its own p values, which is why no cell of it is confirmatory.

C The model, its proofs and what it identifies

This appendix proves the two propositions of Section 4 and tests the interaction the ordering requires.

Table 3: Every claim of the main text, the measurement behind it and where it is reported. Rows marked C1 to C5 are the confirmatory tests of Table 4 and carry its multiplicity correction, and the rest are secondary. Retentions select their block on validation and are scored on test throughout. Exact p values enumerate the full permutation set.

	claim	measurement	what it reads	in full
Section 5, the ordering, its premise and what the loss falls on				
	ordering	peak block against target scale, per backbone	9 of 9 admitted, 18 of 32	D.3
C1	the interaction is required	scale labels permuted over the observed peak blocks	$p = 0.0074$ over nine, 0.0033 over three	C.2
C2	the premise selects it	coarse-height peak block, exact permutation	$+0.894$ against $+0.105$, $p = 0.0007$	D.3
	and is not height read twice	the same rule on the semantic probe	31 of 32 agree, $p = 0.0062$	D.3
	a boundary label survives	the unchanged semantic head rescored by partition	0.978 against 0.921	D.4
C3	a photometric one does not	the same functions on luminance, matched token fraction	0.462 against 0.921, $p = 0.0039$	E.5
Section 6, the length that calibrates the ordering				
C4	the radius calibrates it	within-bin spread against s and s/r_L, exact 9! null	$+24$ and $+36\,\%$, $p = 0.0074$ and 0.0307	D.1
	and selectively	the same statistic on the excluded family	$-9\,\%$, $p = 0.76$	D.1
	and no rival length does	three label-free lengths fixed in advance	best rival at $p = 0.3796$	D.2
Section 7, that mixing at that length is a cause of the loss				
	a blur imitates depth	block 16 average-pooled until it reaches the output score	a width exists in 9 of 9	E.1
C4	at the width the radius gives	that width against $\sqrt{r_L^2 - r_{16}^2}$, exact 9! null	$+0.636$ and $+0.650$, $p < 0.04$	E.1
	a read-out cannot undo it	a 3×3 window on the output block against one token of block 16	$1.4\,\%$ against $7.7\,\%$	E.2
	nor can a decoder	two 3×3 convolutions on the output block against block 16 injected	9.82° against 9.75°	E.4
	and widening does not help	window k from 1 to 9, three seeds	10.948 to 11.038°	E.2
C5	narrowing attention returns it	paired bootstrap over the same 71 held-out tiles	-0.148°, $p = 0.0035$	E.3
	on three of six once dosed	the same, swept over five penalties per backbone	semantic control holds in 3 of 6	E.3
	and not the photometric one	the same intervention read on that target	-1.03°, $p = 0.13$	E.5
Section 8, the task cost and the repair				
	the last third costs matching	nearest neighbour over the grid under five warps	$30\,\%$ of the peak given up	F.1
	real pairs rank as warps do	two Maxar acquisitions, offset removed from pixels	$\rho = +0.967$, $p = 0.0002$	F.1
	a warp picks the block	the block chosen without labels against a labelled sweep	71 to $80\,\%$ of the gap	F.2
	and across the model zoo	the same rule on all 32 backbones of the premise sweep	80 to $88\,\%$ of the gap	F.3
	and not on photographs	the same rule on six backbones of NYU Depth v2	it costs more than it recovers	F.3

Table 4: The five confirmatory tests, their raw p values and the same values adjusted for the multiplicity of this family alone. Holm controls the family-wise error rate under arbitrary dependence and Benjamini and Hochberg the false discovery rate, both at 0.05. Every raw value is read from the result file named beside it and none is restated by hand.

claim	the test asks	raw p	Holm	Benjamini and Hochberg
C1	the peak block rises with target scale	0.0074	0.0148	0.0092
C2	the premise separates the ordering	0.0007	0.0035	0.0035
C3	the photometric target is lost harder	0.0039	0.0140	0.0065
C4	the radius tightens the pooled spread and orders the blur width	0.0359	0.0359	0.0359
C5	a narrow window returns the normal	0.0035	0.0140	0.0065

C.1 Proofs, and a calibration of the window bound

Proposition 1. Take logarithms in Equation 2 and write $\psi = \log A$, so that $\log q_\ell^t = \log g_\ell + \psi(r_\ell \tau / s^t)$. For a fixed pair of blocks $\ell < m$ the difference

$$\log q_m^t - \log q_\ell^t = \log \frac{g_m}{g_\ell} + \psi\Big(\frac{r_m \tau}{s^t}\Big) - \psi\Big(\frac{r_\ell \tau}{s^t}\Big) \tag{3}$$

depends on s^t only through the last two terms. Differentiating them in s^t gives $-(r_m\tau/(s^t)^2)\psi'(r_m\tau/s^t) + (r_\ell\tau/(s^t)^2)\psi'(r_\ell\tau/s^t)$, which is $(s^t)^{-1}$ times the difference of $x|\psi'(x)|$ evaluated at $r_m\tau/s^t$ and at $r_\ell\tau/s^t$. Since $r_m \geq r_\ell$ and $x|\psi'(x)|$ is non-decreasing by assumption, that difference is at least 0, so Equation 3 is non-decreasing in s^t for every $\ell < m$. This is exactly increasing differences of $\log q_\ell^t$ in the pair (ℓ, s^t), and the maximisers of a function with increasing differences are non-decreasing in the parameter in the strong set order (Topkis, 1978; Milgrom & Shannon, 1994), which gives the second claim. For the first, put $m = L$ in Equation 3 and note that $\log \eta^t = \min_\ell \{\log q_L^t - \log q_\ell^t\}$, a minimum of functions each non-decreasing in s^t, hence non-decreasing in s^t. Targets scored by an error rather than by a score are covered by applying the same argument to the reciprocal, which is the form Equation 1 already takes for them.

Proposition 2. Let the observed feature be $z = K * y + \varepsilon$ on the token grid with y the target field, K the mixing kernel and ε uncorrelated with y, all jointly stationary with power spectra S for y and N for ε. A read-out over a window of k tokens is a linear filter whose transfer function is a trigonometric polynomial of degree $(k-1)/2$. The set of such filters grows with k and exhausts the space of square-summable filters, so the attainable error is non-increasing in k and its infimum is the unconstrained Wiener error (Wiener, 1949), whose spectrum is $SN/(|\hat{K}|^2 S + N)$. Dividing by the target variance $\int S$ and subtracting from one, the best coefficient of determination available to any window at all, written $\bar{R}^2$, is

$$\bar{R}^2 = 1 - \Big(\int S(\omega)\, \mathrm{d}\omega \Big)^{-1} \int \frac{S(\omega)\, N(\omega)}{|\hat{K}(\omega)|^2\, S(\omega) + N(\omega)}\, \mathrm{d}\omega, \tag{4}$$

where S and N are the power spectra of the target and of the uncorrelated content at spatial frequency ω and $\hat{K}$ is the transfer function of the mixing. Where $|\hat{K}(\omega)|^2 S(\omega) \ll N(\omega)$ the integrand is $S(\omega)$, so the whole of the target power at that frequency is lost for every window.

The class of read-outs the bound covers, since the class matters. Equation 4 is proved over linear read-outs, which is the class the window control of Section 7 varies, so the non-linear map of Table 18 recovering $6.1\,\%$ of the surface-normal error sits outside that class rather than against the bound. Two extensions say what the word bound means here. If y and ε are jointly Gaussian the Wiener estimator is the conditional mean, so Equation 4 then bounds every measurable read-out and not only the linear ones. Without that assumption what survives for a general read-out is the frequency-wise statement. A channel of gain $\hat{K}$ carrying noise N conveys at most $\frac{1}{2}\log(1 + |\hat{K}|^2 S/N)$ nats about y at each frequency, and that vanishes wherever the mixing has pushed the target below the uncorrelated content. So the frequencies the mixing removes are gone for any read-out, while the coefficient of determination attainable at the surviving ones is not in general the Wiener value. The paper claims the bound in the linear form it proves and reads the saturation of Table 14 as evidence for that form.

How the response radius accumulates. Section 4 reports the growth of r_ℓ as multiplicative in depth, and this is the fit behind that. Taking the rising stretch of each profile, seven of the nine backbones are better described by $\log r_\ell$ linear in ℓ, at R^2 between 0.90 and 0.99, than by r_ℓ^2 linear in ℓ, which is what independent per-block mixing would give. Multiplicative growth is what follows if the reach of a block is set by features already pooled to radius r_ℓ. The two exceptions stay below 1.4 tokens throughout, so neither law has growth to fit.

Why every test in the paper is ordinal. Section 4 takes from Equation 2 the orderings it implies and never a fitted law, and this is the measurement behind that choice. Past the block at which a target peaks, abstraction has saturated, and the strongest reading of the model asserts that the normalised score there is a function of $r_\ell \tau / s$ and of nothing else, in particular not of the backbone. Collecting every such point over nine backbones and the three target families gives 1089 measurements, and a single attenuation function shared across backbones accounts for $R^2 = 0.016$ of them, rising only to 0.158 once each backbone is allowed its own offset. Backbones whose abstraction saturates at different levels are therefore not poolable in that way. Nor is the failure only the pooling, since Equation 2 does not say the attenuation is shared. Giving every backbone its own exponent as well as its own offset, which is 18 parameters for 1089 points, reaches $R^2 = 0.302$, and three of the nine fitted exponents come out negative, which the model forbids outright because A has to be non-increasing. Those three are the backbones whose radius never leaves the first token, so the ratio has almost no range in them and the fit is reading noise, which makes the fit uninformative there and does not rescue the quantitative form. All three fits assume a power of the ratio, which is the one family Section 4 shows cannot produce a target-dependent peak block at all, so what they reject is narrower than a quantitative reading of Equation 2 (Appendix C.2). So the paper uses only what Proposition 1 needs, which is monotonicity rather than a universal functional form, and tests it with rank statistics throughout.

How exactly monotone is the ordering? Proposition 1 asserts monotonicity for every backbone, which is more exposure than the single-backbone claim of Section 5 takes on, so we check it here. Counting a fall of more than 0.02 in retention between adjacent scales as a violation, eight of the fourteen retention series are strictly monotone, the fourteen being nine backbones by the two target families of fixed population less the four the floor excludes. Every violation that survives lies between 0.026 and 0.074, which is the width of the bootstrap interval on a single probe score, so no series departs from the ordering by more than the precision it is read at. DINOv3-Web, the backbone of Figure 4, is monotone in all three families. Both exclusions are properties of the targets and not choices made after seeing the answer. Aspect is left out because the slope threshold that defines the direction is applied at each smoothing width, so the fraction of tokens it is scored on falls from 64 to 58 to $42\,\%$ over the three widths and a change of width changes the population as well as the target. Putting it back would raise the monotone fraction rather than lower it. The coarsest height band is left out where its best score is below $R^2 = 0.25$, since a retention ratio against a peak that low is not a measurement, and those series carry the four largest apparent violations, from 0.21 to 0.94. The ordering is therefore approximate rather than exact, which is what Proposition 1 claims and no more.

This check does not test how far the ordering separates the backbones, since it runs on DFC2019 at $\tau = 16\,\mathrm{m}$, where Section 5 reports the sweep compressed by the resolution. The sweep that separates them is the one Appendix D.3 reports at $\tau = 5.84\,\mathrm{m}$, where the backbones divide along the premise Section 5 reads, since a backbone whose abstraction stops accumulating early has no reason to order its peaks, and the classification-supervised one does not.

The bound checked on a field whose mixing is known exactly. Because Equation 4 is a statement about spectra rather than about transformers, its shape can be checked directly. We draw a random field with power at a single scale s, mix it with a Gaussian kernel of radius r, add uncorrelated content at a fixed fraction of the field variance, and fit the best ridge read-out over a window of k tokens. Table 5 reports the result. Two features of it are what Section 7 then tests on real features. The ceiling is governed by r/s and falls steeply in it. And widening the window past five tokens moves the ceiling by at most 0.085 at any r/s, so a window read-out that fails to recover a target is not failing because it is too small.

Table 5: Coefficient of determination of the best linear read-out of a field of ground scale s from a representation mixed to radius r, over a square window of k tokens, with content uncorrelated with the field at $1/50$ of the field variance. Averages of three draws on a 96×96 grid with $s = 8$ cells. The ceiling is set by r/s, and past $k = 5$ no row moves by more than 0.085.

	window width k in tokens					
r/s	1	3	5	7	11	17
0.000	1.000	1.000	1.000	1.000	1.000	1.000
0.125	0.746	0.891	0.920	0.926	0.932	0.931
0.250	0.342	0.465	0.516	0.559	0.590	0.601
0.500	0.066	0.068	0.083	0.105	0.104	0.112
0.750	0.005	0.010	0.010	0.008	0.013	0.012
1.000	0.001	0.001	0.001	0.000	0.000	0.000

C.2 The interaction the ordering requires

Section 4 states that Equation 2, read over the grid of blocks against targets, is additive in logarithms and that its content after the block effect is profiled out is a restriction on the interaction, namely that depth and target scale meet only through $r_\ell \tau / s^t$. This appendix reports the test of that restriction against the model that has no interaction at all.

The competitor forbids the ordering, and the ordering is there. The model with no interaction is not a rival account of the ordering. It forbids the ordering. With no interaction the score is $a_\ell + b^t$, a free effect for each block plus a free effect for each target, whose maximum over blocks does not depend on the target at all, so it says the block at which a target peaks is one block for every target. That is a prediction, it needs no functional form to test, and it is false. Permuting the target scales over the observed peak blocks, which is exact at five scales, the peak rises with target scale at $p = 0.0167$ on DINOv3-Sat and on DINOv2-L, $p = 0.050$ on MAE-L and $p = 0.067$ on SatMAE-L, and combining the nine one-sided values gives $p = 0.0074$. Over the three of the nine that satisfy the premise, a condition decided in Appendix D.3 and not here, it gives $p = 0.0033$.

The family this is read on is band-passed height, five target scales by twenty-four blocks for each backbone. Every score is the validation score, which is the split Equation 1 takes its extremum over, and no test score enters this appendix, since a peak block is a location and locating it on the split a series is scored on would select it. Block 0 is dropped for the reason given in Appendix C.1, and cells whose score is below 0.05 are dropped as noise on a logarithmic scale. The script is `scripts/identify_model.py` and the numbers are in `results/identify.json`.

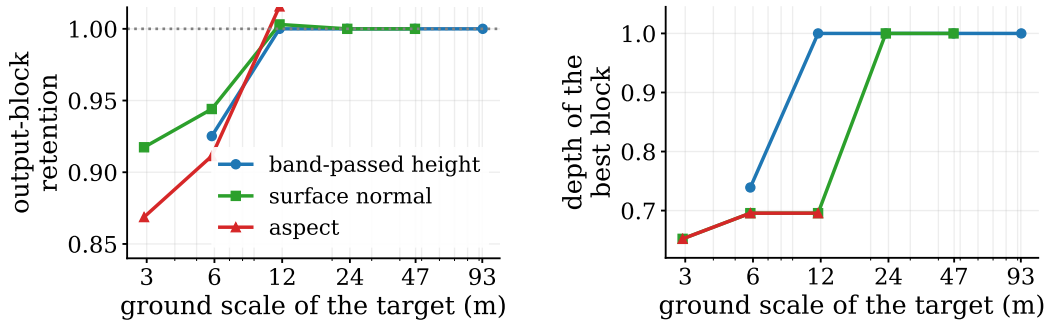
What this establishes and what it leaves open. It establishes that the interaction exists, which is what Proposition 1 needs, and not what shape the attenuation has. Two consequences are taken up in the main text. Fitting a power of the ratio, which is what Appendix C.1 reports, cannot be a test of Equation 2 at all, since that family puts the best block at one place for every target. And a surface of scores read down depth moves the abstraction and the mixing together, so the evidence that the length is what matters has to come from holding the block fixed and moving the radius, which is what the blur equivalence and the attention intervention of Section 7 do.

D The ordering, its premise and its scope

Everything behind Sections 5 and 6 is here, namely the ordering itself and the collapse that calibrates it, three rival lengths that do not, the premise over thirty-two backbones, the per-family peak depths, the same test on indoor photographs and the same test across architectures and capacities.

D.1 The depth of the best block

Figure 4 adds to the retention curves of Section 5 the relative depth at which each target is best decodable, which rises with the ground scale of the target over the same range in which retention rises, so the two views of the ordering agree.



(a) retention

(b) depth of the best block

Figure 4: The scale-depth ordering for DINOv3-Web at $\tau = 5.84$ m. (a) Retention rises monotonically with the ground scale of the target in all three families. (b) The relative depth of the best block rises over the same range, from 0.65 for the finest differential targets to the output block above 23 m.

Two things Sections 5 and 6 state in one line are sized here. The crossover Section 6 reports, where retention reaches the output block's own score near the response radius read at native detail, is measured on one backbone and is not claimed across them. The collapse and the blur width are where a radius is read across nine. And the sweep flattens where the imagery cannot resolve the structure it asks about, the finest aspect target retaining 0.917 at $\tau = 16$ m against 0.869 at $\tau = 5.84$ m, which is the same ratio speaking rather than a different result.

The collapse over every backbone satisfying the main table's invariant. The nine of Figure 2 are Table 2's roster. Thirteen further releases measured for the premise sweep share the invariant that table controls for, a ViT-L tower of 1024 channels and 24 blocks under plain global self-attention at the identical footprint, and the pool rule is that invariant plus the existence of both measurements, never an outcome. Table 6 lists them with the radius the collapse divides by. Over the twenty-two the within-bin tightening is 17.5 % on surface normals and 17.7 % on aspect, at $p = 0.0021$ and $p = 0.0107$ against a null of 10^6 sampled reassignments of the radii, since 22! cannot be enumerated, and the nine alone reproduce Section 6's exact values. So the calibration is not a property of the nine. Band-passed height, untouched over the nine, tightens by 6.0 % at $p = 0.034$ here. We read that after the fact, and the reading consistent with Equation 2 is that the exemption follows the ratio rather than the target family. The added releases carry radii up to 13.2 tokens, or 77 m, which puts two thirds of their band-passed targets at or below the aggregation scale against half for the nine. None of this enters the confirmatory family, which reads the collapse over the nine at its exact null.

Table 6: The thirteen releases added to the collapse, every one a ViT-L with 1024 channels and 24 blocks under plain global self-attention. The response radius r_L at the output block is in tokens on DFC2019 at the analysis grid, the protocol the collapse divides by, so these values are not comparable to Table 2's column, which is band limited.

backbone	pretraining objective	r_L
AIMv2-L	autoregressive	10.60
BEiT-L	masked reconstruction	1.65
BEiTv2-L	masked reconstruction	3.07
DFN-L	language alignment	8.81
Depth-Anything-V2-L	geometry supervision	6.97
DINOv2-L plain	self-distillation	4.45
EVA02-L	masked image modelling	2.93
MetaCLIP-L	language alignment	10.45
OpenCLIP-L	language alignment	7.23
SigLIP-L	language alignment	10.85
SigLIP2-L	language alignment	11.00
TIPSv1-L	self-distillation with language	4.70
TIPSv2-L	self-distillation with language	13.18

D.2 THREE RIVAL LENGTHS OF THE MIXING

The null of the collapse asks whether each radius belongs to the backbone it is assigned to. It does not ask whether another label-free length of the mixing would serve as well, and three are standard. Each was defined in writing before any was measured, in tokens on the analysis grid at block 16 and at the output block, and each is read on exactly the imagery the radius is read on for each test. Mean attention distance is the statistic of Raghu et al. (2021), the attention each patch query pays each patch key times the distance between them, averaged over queries and heads and read from the probabilities every block computes. The gradient field is the effective receptive field of Luo et al. (2016) read in tokens. It is the squared gradient of a whitened random projection of one token's feature with respect to the input image, summed over each token's cell. That map is reduced to a centre of mass over distance exactly as the radius is, at the radius's own occlusion sites on the first 20 of its 40 tiles. The autocorrelation length is the lag at which the token feature map's spatial autocorrelation first falls below $1/e$, which is the statistic of the smoothness-matched width of Appendix E.1. Each enters both tests unchanged, the collapse with its exact null and the blur width against the length it adds between block 16 and the output block.

Table 7: The response radius against three rival label-free lengths of the mixing, all fixed before any was measured. The collapse rows are the within-bin tightening of the nine retention curves of Figure 2 when the ground scale is divided by that length at the output block on DFC2019 at the analysis grid, with the exact p over all 362880 reassignments of the nine lengths. The blur rows are the rank correlation of the widths of Table 13 with the length each adds between block 16 and the output block on band-limited US3D, the square root of the difference of its squares, with the exact one-sided p. Each p is in parentheses. The radius column reproduces the values Sections 6 and 7 report.

test	response radius	mean attention distance	gradient field	autocorrelation length
collapse, surface normals	$+24.1\,\%$ (0.0074)	$+5.9\,\%$ (0.3796)	$+5.9\,\%$ (0.9815)	$+8.4\,\%$ (0.5233)
collapse, surface aspect	$+35.7\,\%$ (0.0307)	$+19.3\,\%$ (0.4259)	$+27.5\,\%$ (0.8972)	$+11.1\,\%$ (0.5970)
collapse, band-passed height	$-8.7\,\%$ (0.7577)	$+16.2\,\%$ (0.0450)	$+17.0\,\%$ (0.3955)	$-21.2\,\%$ (0.9438)
blur width, surface aspect	$+0.636$ (0.0359)	$+0.247$ (0.2582)	-0.118 (0.6253)	-0.183 (0.6911)
blur width, surface normal	$+0.650$ (0.0333)	$+0.254$ (0.2573)	$+0.033$ (0.4687)	-0.237 (0.7464)

Table 7 gives every cell and none of the three rivals calibrates. None tightens surface normals or aspect beyond chance, the best cell reaching $p = 0.3796$, and none orders the blur widths, the best at $\rho = +0.254$. The one rival cell below 0.05 is mean attention distance tightening band-passed height, which is the family the account places above the aggregation scale and the radius leaves where it was. The radius reproduces its published value in every cell, so the comparison is between lengths and not between pipelines.

Why they fail is visible in the lengths. Mean attention distance spans a factor of 2.0 over the nine, from 16.74 to 33.92 tokens, where the radius spans a factor of 16.2, and it runs against the radius at $\rho = -0.167$. DINOv3-Web attends most locally of the nine while its radius, 10.87 on this imagery, is the second widest. Attention weights leave out the residual stream and the values, so the distance a block attends over is not the distance its output responds over. The gradient field ranks the backbones much as the radius does, at $\rho = +0.733$, and still fails, since it puts DINOv3-Sat widest of the nine at 39.90 tokens where its radius is 2.29. We had predicted in writing that the gradient field, as the infinitesimal form of the occlusion, would calibrate as the radius does, and that prediction is wrong. What calibrates is the response to replacing a token's content, which a derivative at the unperturbed image does not read. The autocorrelation length mixes the image's own spatial correlation into the network's and follows the radius at only $+0.217$.

D.3 THE PREMISE OVER THIRTY-TWO BACKBONES

Section 5 reads the premise of Equation 2 off one proxy, the block at which the mean above-ground height of a token is best decoded. This appendix reports what that proxy is worth over 32 backbones, which come in five groups. Nine are Table 2, and seven change the architecture or the capacity and are the subject of Appendix D.6. Nine more were admitted by a rule fixed in writing before any

Table 8: The premise and the ordering over the 32 backbones swept, sorted by whether the premise admits the backbone and then by how strongly the ordering appears. The premise is the block at which coarse mean height is best decoded, counted as holding within three blocks of the output. The ordering is the rank correlation between a band-passed target's ground scale and the depth of its best block, on US3D at a 64×64 grid, and counts as shown at $+0.5$. The two are read off different targets. Pretraining objective is taken from each release's own description.

backbone	pretraining objective	premise	ordering ρ	ordering
DINOv3-Sat	self-distillation	yes	$+0.975$	yes
DINOv2-L reg4	self-distillation	yes	$+0.975$	yes
DepthAnyv2-L	geometry supervision	yes	$+0.975$	yes
DINOv2-B	self-distillation	yes	$+0.894$	yes
DINOv2-g	self-distillation	yes	$+0.894$	yes
DINOv3-S	self-distillation	yes	$+0.894$	yes
DINOv2-L	self-distillation	yes	$+0.821$	yes
DINOv3-Web	self-distillation	yes	$+0.707$	yes
DINOv3-B	self-distillation	yes	$+0.707$	yes
TIPSv1-L	self-distillation with language	no	$+0.975$	yes
Swin-B	supervised	no	$+0.949$	yes
AIMv2-L	autoregressive	no	$+0.949$	yes
TIPSv2-L	self-distillation with language	no	$+0.900$	yes
MAE-L	masked reconstruction	no	$+0.894$	yes
DINOv1-B	self-distillation	no	$+0.866$	yes
SatMAE-L	masked reconstruction	no	$+0.791$	yes
Swin-V2-B	supervised	no	$+0.738$	yes
EVA02-L	masked image modelling	no	$+0.667$	yes
ViT-L IN-21k	supervised	no	$+0.158$	no
OpenCLIP-L	language alignment	no	$+0.154$	no
MetaCLIP-L	language alignment	no	$+0.105$	no
CLIP-L	language alignment	no	$+0.000$	no
DINOv1-S	self-distillation	no	$+0.000$	no
I-JEPA-H	joint-embedding prediction	no	$+0.000$	no
SigLIP2-L	language alignment	no	-0.289	no
Scale-MAE-L	masked reconstruction	no	-0.300	no
SigLIP-L	language alignment	no	-0.316	no
I-JEPA-g	joint-embedding prediction	no	-0.354	no
DFN-L	language alignment	no	-0.564	no
BEiTv2-L	masked reconstruction	no	-0.718	no
DOFAv1-L	masked reconstruction	no	-0.949	no
BEiT-L	masked reconstruction	no	-0.975	no

measurement existed on any of them, namely every publicly released ViT-L of 1024 channels and 24 blocks read by plain global self-attention whose pretraining recipe or corpus differs from all the others, enumerated from one public catalogue. Four were admitted by a second such rule, written to break a confound and stated below. The last three are two TIPS releases (Maninis et al., 2025) and a Depth-Anything encoder (Yang et al., 2024), held out for an analysis that has since left the paper, so their exclusion outlived its reason and they are pooled here.

Neither pre-registered group played any part in choosing the proxy, the threshold or the target, so applying the rule to them is a test rather than a choice. The proxy itself is not pre-registered, having been developed on the nine of Table 2. What was fixed in advance is the pool it was then applied to unchanged, which is the weaker of the two things the word could mean and is the one we claim.

Table 8 gives every backbone. The premise holds in 9 and the ordering holds in all 9 of those. It fails in 23 and the ordering holds in 9 of them, so the reading and the ordering agree in 23 of the 32 and the ordering appears in 18 of the 32 counted without regard to scope.

The separation, which needs no threshold on the ordering. Counting agreements needs a cut on what shows the ordering and discards how strongly each backbone shows it, so we also report the separation itself, where no threshold enters. The nine backbones the premise admits carry a median

rank correlation of $+0.894$ against $+0.105$ for the twenty-three it does not, an exact permutation $p = 0.0007$. The three smaller pools inside it give $+0.894$ against $+0.000$ at $p = 0.0005$ over the twenty-nine, $+0.894$ against $+0.448$ at $p = 0.0162$ over the sixteen and $+0.975$ against $+0.079$ at $p = 0.0595$ over the nine of Table 2 alone. That last pool splits three admitted backbones against six and does not reach the conventional level. So the separation strengthens monotonically with the pool rather than weakening, which is what says it is not a property of the smallest one. Two things keep it from being circular. The premise is read off the block at which coarse mean height peaks and the ordering off the peak depths of band-passed height, so the two are measurements on different targets rather than one quantity compared with itself. And no backbone was reassigned to improve it. The two windowed ones sit in the failing group while carrying $+0.949$ and $+0.738$, which is the assignment their measurement gives at either grid, and that costs the test a factor of six, since reassigning them would give $p = 0.0001$.

The premise is not height read twice. Both readings are still taken on height, the premise off coarse mean height and the ordering off band-passed height, so a separation between them could be one signal read twice. Two further readings remove the shared part, and both were fixed in writing before either was taken. The first reads the premise off the six-class semantic probe instead, by the identical rule at the same threshold of three blocks. It admits the same nine and Scale-MAE-L, which sits at exactly three blocks short and does not order, so the two partitions agree on 31 of the 32. Its separation is $+0.894$ against $+0.130$ at an exact $p = 0.0062$, with the ordering in 9 of the 10 it admits and in 9 of the other 22. The second keeps the published premise and recomputes the ordering without its coarsest band, which is the end of it a late coarse-height peak implies most directly. The nine then carry $+0.775$ against $+0.000$ for the twenty-three at an exact $p = 0.0012$, and all nine still order over the four finest bands. So a backbone whose semantic representation is still improving at its output block is, with one exception, one whose coarse geometry is too, and the ordering the premise predicts is carried by bands the premise never reads.

Counting each training recipe once. A permutation over backbones treats 32 releases as 32 independent draws, and six of the nine the premise admits are rungs of two capacity ladders. So we repeat the test at the coarser unit, every lineage collapsing to one value, its members' median ordering strength, which makes each ladder one observation and pairs the version pair and the architecture pair of the failing group the same way. Five admitted recipes then face seventeen at a median $+0.975$ against $+0.154$, and enumerating all 26334 relabelings of the clusters gives an exact $p = 0.0036$, the ordering holding in all five and in six of the seventeen. That is the harder reading, since it discards the count as a source of power.

The span of the pool. The 32 cover nine pretraining objectives, capacities from ViT-S to ViT-g and both isotropic and hierarchical attention. The ordering appears in seven of the nine objectives, at 9 of 10 for self-distillation, 2 of 6 for masked reconstruction, 2 of 3 for supervised classification, 2 of 2 for self-distillation carrying a language term and 1 of 1 for each of autoregressive pretraining, masked image modelling and dense geometry supervision. Neither hierarchical backbone satisfies the premise and both show the ordering. Two objectives never show it, language-image alignment over six independent releases and joint-embedding prediction over two, which is a result rather than a gap, since those are also two objectives in which the premise never holds. It is consistency and not a prediction the pool tested, since masked reconstruction never satisfies the premise either and orders in two of six. The 9 of 9 is not nine instances, since six of the nine are rungs of two capacity ladders, each holding one objective and one corpus fixed across three sizes, so the numerator is five independent training recipes and not nine.

The premise is a measurement and not the name of an objective. Over the twenty-five that the first three groups supply, the eight the premise admitted were exactly the eight self-distillation backbones, and those eight were exactly the rungs of DINOv2 and DINOv3. Three descriptions named one set, so no permutation on that pool could attribute the ordering to the premise rather than to the objective, and the clustered test does not lift that either, since it only counts recipes. Four releases were therefore admitted on a rule fixed in writing before any of them was fetched. The rule asks for a plain global self-attention transformer at patch 16, so that its native token grid is the analysis grid with no resampling, whose objective is either self-distillation outside the DINOv2 and DINOv3 lineage or a joint-embedding objective that is neither self-distillation nor pixel reconstruction. Ap-

plied to one public catalogue that rule admits four releases and no others, two DINOv1 rungs and two I-JEPA rungs.

The two descriptions come apart in both directions. DINOv1-B and DINOv1-S are self-distillation and fail the premise, peaking 7 and 5 blocks short of the output and reading 0.88 and 0.76 of their own best block there, and both TIPS releases, self-distillation carrying a language term, fail it too. Depth-Anything v2 is trained on dense depth labels, is not self-supervised at all, and satisfies it at 2 blocks short. So the pool holds ten self-distillation backbones of which the premise admits eight, and it admits the one backbone trained on geometry. Two of those comparisons hold one factor fixed. Depth-Anything v2 is the plain DINOv2-L release trained further on 62M dense depth labels, so it changes the objective and keeps the initialisation, and the premise holds on both at 2 blocks short. DINOv1-B against DINOv2-B keeps the objective family and changes the recipe, and the premise holds on one and not the other. The premise is therefore not the objective, and it is not implied by the objective either. We had predicted in writing that both DINOv1 rungs would satisfy the premise, on the reading that it is about what a network is still building late rather than about its recipe, and that prediction is wrong.

The follow-up is whether the objective label separates the ordering as well, and over these 32 it does not. Ten self-distillation releases against twenty-two others give an exact $p = 0.0042$ where the premise's nine against twenty-three give 0.0007, with the ordering in 9 of 10 against 9 of 9. That gap is one backbone wide, since it is Depth-Anything the premise admits and the objective label does not, so we claim the direction and not the size. What the premise still coincides with is narrower and is stated rather than lifted. All nine it admits descend from a DINOv2 or DINOv3 checkpoint, Depth-Anything included, since it is initialised from one. Breaking that needs a release with no such ancestry that satisfies the premise, and these 32 contain none. That bounds the premise and not the ordering.

The four admitted on the second rule change nothing else. No release satisfies the premise and then fails to order, so the conditional regularity is 9 of 9 in scope. DINOv1-S and I-JEPA-H fail the premise and do not order, at $\rho = +0.000$ each, and I-JEPA-g at -0.354, which is agreement in three cases, and DINOv1-B fails it and orders at $+0.866$, which is the sufficiency without necessity that nine of the twenty-three carry. Of the two DINOv1 rungs, the one whose coarse height falls further by the output block is the one that does not order. A negative premise reading would be indistinguishable from a misfed backbone, so all four were checked against the standing diagnostic of this paper, the response radius profile's minimum, which they bring to 0.18, 0.20, 0.27 and 0.30 tokens where a misfed backbone stays above 6. Those radii are a load report and nothing here reads one against a premise or an ordering.

The reading is an argmax and both TIPS releases show what that costs. Each decodes coarse height at the output block to within $0.2\,\%$ of its own best block, which is what the premise is asking about, and each is counted as failing because the argmax itself sits 4 and 6 blocks earlier on a curve that is flat there. Both order, at $+0.975$ and $+0.900$. So the failing group holds cases where the network keeps abstracting and the statistic cannot see it, which makes the separation conservative rather than flattering.

The threshold is not a knob, and the reading says nothing finer than its side. Counting the premise as satisfied within three blocks of the output invites the objection that three was chosen, so we report where the readings sit. They are bimodal over the 32 and the gap between the modes has closed as the pool grew. The nine the premise admits peak between 0 and 3 blocks short of the output, six of them at the output block itself, and the twenty-three it does not peak between 4 and 19 blocks short. So 3 is now the only threshold giving this partition, where over the twenty-five every value from 3 to 8 gave it, seven backbones added since sitting in what used to be the gap. We report that rather than move the cut. What keeps this a test is that 3 was fixed in writing before any of those seven was measured and is applied to them unchanged, and what is weaker than it was is the separate argument that the cut sits in a wide gap in the data. The same measurement bounds the proxy in the other direction. The count orders the ordering at $\rho = +0.509$ over the whole pool, which is the gap speaking, and at $+0.229$ inside the twenty-three and -0.425 inside the nine. Those two run in opposite directions, so the reading is a partition and not a graded predictor, and it says which side a backbone falls on and nothing finer.

All six disagreements run the same way, a backbone failing the premise and ordering anyway, and the sharpest two are the windowed pair. DINOv2-L, admitted from three blocks short of the output, orders at $\rho = +0.975$. Both Swins peak in the first third, which counts as failing the premise, and both show the ordering at $\rho = +0.949$ and $+0.738$. Both lose three quarters of the finest band-passed height in the process, retention running from 0.241 and 0.242 at 5.8 m to 0.952 and 0.989 at 23.4 m, the coarsest band the R^2 floor admits for either of them.

That failure is a property of the network and not of the analysis grid, which is worth establishing because the windowed pair invites the opposite reading. Both emit 32 tokens a side at their last stage where the analysis grid has 64, so their coarse target is interpolated up and its score falls toward the output block instead of rising, from 0.345 to 0.220 of coefficient of determination for Swin-B and from 0.356 to 0.279 for Swin-V2-B. Every one of the eight backbones the proxy admits reads its coarse target at 0.939 of its own best block or better. Read instead at a 32×32 analysis grid, where their last stage emits exactly that grid and nothing is interpolated, Swin-B reads 0.460 of its best block at the output and Swin-V2-B 0.465. Both are lower than at the finer grid and both still peak in the first third, while the DINOv3-Web control peaks at the output block at either grid, and both continue to show the ordering, at $+0.775$ and $+0.949$. So these two are the clearest evidence that the premise is sufficient and not necessary.

A label-free replacement for the proxy was also tried and is withdrawn. It read 6 of 9 in sample against the labelled version and 0 of 5 on backbones outside the main table. So a sound observable for the accumulation of abstraction is an open problem and not something this paper supplies. The premise is therefore a diagnostic that admits few backbones and excludes some the ordering describes well, and it is reported as that rather than as a gate the main claim rests on.

D.4 Peak depth and retention per target family

Table 9 gives the relative depth of the best block and the retention of that block's score at the output, for four target families and all nine backbones, on DFC2019 with the cross-city holdout. The semantic family is the six-class read-out, the metric family is the mean above-ground height of a token, the relational family is the signed height difference of two tokens, and the differential family is the surface normal and aspect.

The three further targets of the specificity control. Section 5 compares boundary-abutting semantic retention against surface aspect, and the comparison is not specific to aspect. Boundary retention also exceeds that of surface tilt and of the height-discontinuity fraction in eight backbones of nine at $p = 0.0078$ and that of within-token height roughness in nine of nine at $p = 0.0039$, all three scored as a coefficient of determination and clearing the dynamic-range floor this appendix states for the family.

Why aspect is the primary differential-geometry instrument. Section 5 compares boundary-abutting semantic retention against surface aspect rather than against the surface normal, because a retention ratio needs a scale on which zero means no information and mean angular error has none. The best constant direction already achieves $13.93°$ on these tiles, where the best probe of the best backbone achieves 10.90 and the worst backbone's best block reaches only 13.84, which is chance. So the usable part of the scale is a fifth of its nominal range, every ratio taken on it is compressed toward unity, and run against the normal the specificity test does not separate, at 0.973 against 0.969 with a two-sided paired Wilcoxon $p = 0.82$. Aspect has the range the comparison needs, a constant azimuth giving $81.25°$ against a best of 37.58. Rescaling the normal to the share of directional variation it explains, $(\mu - \bar{\mu})/(1 - \bar{\mu})$ for a mean cosine μ against the $\bar{\mu} = 0.9448$ a constant direction achieves, the target's resultant length, does make the test separate, in nine backbones of nine at an exact one-sided Wilcoxon $p = 0.0020$. It leaves aspect almost unmoved at 0.821 against 0.912. We do not adopt it. Five of the nine backbones then explain less than a quarter of the directional variation at their best block, from 0.122 to 0.230, below the level at which Appendix C.1 treats a retention ratio as meaningless. Every backbone clears it on aspect, from 0.282 to 0.537. The compression bounds what a normal retention magnitude can be compared against, and not the ordering of a target's peak depths within one backbone, which is what the ordering is read on.

Table 9: Relative depth of the best block and retention at the output block, per target family, on DFC2019 with the cross-city holdout. Retention is defined in Equation 1. The metric column averages the three height targets, except for DOFA-L where it is the mean height alone, since including the other two gives -1.584. At this token footprint the height targets sit near the $R^2 = 0.25$ floor of Appendix C.1 for most backbones, so that column is descriptive.

	semantic		metric		relational		differential	
backbone	depth	retention	depth	retention	depth	retention	depth	retention
DINOv3-Web	1.00	1.000	1.00	1.000	1.00	1.000	0.86	0.934
DINOv3-Sat	0.91	0.986	1.00	1.000	1.00	1.000	0.93	0.988
DINOv2-L	0.74	0.979	1.00	1.000	1.00	1.000	0.76	0.839
MAE-L	0.70	0.975	0.64	0.639	0.65	0.795	0.64	0.961
CLIP-L	0.43	0.901	0.77	0.934	0.65	0.795	0.49	0.842
Scale-MAE-L	0.74	0.996	0.68	0.879	0.87	0.964	0.77	0.930
SatMAE-L	0.57	0.959	0.68	0.715	0.74	0.913	0.65	0.961
DOFA-L	0.26	0.875	0.65	0.461	0.57	0.740	0.36	0.914
ViT-L ImageNet-21k	0.39	0.886	0.93	0.989	0.61	0.822	0.38	0.901

D.5 The ordering on natural photographs

Section 5 is measured on overhead imagery because that is where metric geometry is dense ground truth and a known sampling distance turns every scale into metres. Both are properties of the imaging geometry rather than of the mechanism, so the ordering has to be shown somewhere else before it can be read as a fact about spatial mixing. This appendix shows it on NYU Depth v2, whose 1449 labelled frames are indoor, hand held, perspective and cluttered, with dense metric range on the same pixels.

Only the field changes. The band-passed targets are built by the functions of Appendix A with the range field in place of the height field, on the same token grid and shared across backbones as before. The frames are cropped to the released valid region, squared and resized to 1024 pixels, and the held-out split is a group of whole scene types rather than a city, since images of one room are near duplicates and a random split would put the same room on both sides. Six backbones are run, spanning the response radius from 0.98 to 7.67 tokens over five pretraining objectives, and Appendix F.3 runs the unlabelled choice of block on the same frames.

Two things do not carry over. The scale is read in tokens, since under a perspective camera one image-space scale is many ground scales, and the mixing acts in image space in any case. And the surface aspect needs no metric scale at all, since the arc tangent of a gradient is invariant to any scaling of the field, which is the argument the photometric control of Section 5 rests on.

The ordering of peak depth is what transfers. Table 10 gives it. Read at the analysis grid of 64 the ordering holds in 4 of the 6 at a median ρ of $+0.707$, against 4 of 6 at $+0.801$ for the same six backbones on the overhead imagery, and the two domains agree backbone by backbone on 4 of the 6. Read at grid 32, where a token covers four times the area, it again holds in 4 of 6 and the median rises to $+0.801$, so this is not a property of one grid. MAE-L misses at both grids and is the backbone Section 5 already places outside the ordering's scope, since its retention at the finest band is 1.084 and it therefore has nothing to order.

Retention, and the floor that decides how much of it can be read. A retention is a ratio and a ratio needs a denominator worth dividing by, so the $R^2 = 0.25$ floor of Appendix C.1 is applied before any of these series is read as one rather than after. At grid 64 only two backbones decode four or more bands above the floor, and both order their retention by scale, DINOv3-Web at $+0.949$ and DINOv2-L at $+1.000$, DINOv3-Web rising from 0.853 at one token to 1.000 at eight while its best block moves from 0.65 of the depth to the output block. At grid 32, where the admissible counts run 5, 4, 4, 3, 4 and 2 down the same rows, five backbones have three or more and three of those five order their retention, at a median of $+0.800$. So the retention leg is thinner here than the peak-depth leg, because a range field carries little power at the finest bands a 1024 pixel frame defines.

Table 10: The band-passed range targets on NYU Depth v2, at the analysis grid of 64 unless the last two columns say otherwise. The five score columns are the best coefficient of determination each band reaches, before any ratio is taken, at band scales of 1 to 16 tokens. The admissible column counts the bands whose best score clears the $R^2 = 0.25$ floor of Appendix C.1, which is how many of them can carry a retention at all. The three ordering columns are the rank correlation between a band's scale and the depth at which it is best decoded, within that backbone, on the photographs at each grid and on the overhead imagery.

	best coefficient of determination					grid 64		ordering	
backbone	1	2	4	8	16	admissible	ordering	grid 32	overhead
DINOv3-Web	0.181	0.294	0.394	0.503	0.795	4	$+0.975$	$+0.289$	$+0.707$
DINOv2-L	0.156	0.308	0.417	0.516	0.812	4	$+0.949$	$+0.894$	$+0.975$
CLIP-L	0.058	0.146	0.246	0.380	0.620	2	$+0.707$	$+0.707$	$+0.000$
ViT-L ImageNet-21k	0.049	0.104	0.183	0.293	0.598	2	$+0.224$	$+0.894$	$+0.158$
MAE-L	0.037	0.077	0.132	0.229	0.397	1	$+0.154$	-1.000	$+0.894$
DINOv3-Sat	0.062	0.097	0.138	0.203	0.384	1	$+0.707$	$+0.949$	$+0.975$

The radius is readable here, and the unit the calibration needs is not. The occlusion diagnostic needs no labels, so it runs on any images at all, and the natural question is whether the length that calibrates the ordering overhead does so here. Equation 2 settles half of that before any measurement, since the collapse divides a target's ground scale by the radius and the two must be in one unit. A nadir token has a fixed ground scale where a perspective token does not, which is a fact about the projection and not about any outcome. We measure both halves anyway. At native detail five of these six backbones bring their radius below 1.011 tokens at some block, which is what says the occluded patch has an effect local enough for the profile to place, and only DINOv2-L does not, never falling below 4.77 at either source resolution tried. The length is therefore readable on five and the calibration can be tested rather than declined.

The test does not decide it, and the reason is the instrument's resolution and not its outcome. Repeating the collapse of Section 5 on those five with the same statistic and the same six bins, the rescaling tightens the within-bin spread by $19.2\,\%$ on band-passed range, $8.0\,\%$ on the surface normal and $7.4\,\%$ on aspect, against null medians of $+7.5$, $+13.0$ and $+12.4\,\%$ from reassigning the same five radii to the same five backbones. Enumerating all 120 relabelings puts the three at exact $p = 0.24$, 0.74 and 0.75. The same enumeration says what this instrument could have found. With 120 relabelings a tightening must exceed 42.5, 43.9 and $44.4\,\%$ to reach $p = 0.05$, where the nine backbones of Section 5 need 19.7 and 32.9. The overhead effects of 24.1 and $35.7\,\%$ fall below every one of those three thresholds, so this test would have missed the overhead result itself. A null here is therefore not evidence that the calibration is absent, by the rule this paper applies to any comparison whose outcome has no range in it. What the point estimates do say is that band-passed range sits above its own null median while the surface normal and aspect sit below theirs, so there is no hint of the effect on two of the three families and one too small to test on the third. Four to eight bins move the band figure between 15.4 and $22.3\,\%$ and never below $p = 0.15$, and reading the radii band limited leaves four backbones admissible at $+9.4\,\%$ and $p = 0.50$.

Two measurements say why the reading is this weak. The photographic radius is not the overhead radius of the same backbone, ordering it at $\rho = +0.143$ against the imagery the collapse takes its overhead radii from and at $\rho = -0.143$ against the band-limited protocol every other ranking uses, both nothing over six backbones. And one image-plane scale is many physical scales here, since a token covers its range times the pixel pitch over the released focal length, which runs 0.0153 to 0.0720 metres over the corpus and spreads within a single frame by a factor of 2.7 at the median and 5.0 at the ninetieth percentile. Overhead imagery lacks that term by construction, which is why the paper writes a target scale in metres there and in tokens here. So the length that calibrates the ordering on nadir imagery is one this instrument cannot decide on photographs, and this appendix claims the ordering and nothing else.

The family that does not transfer, and why that is measurable too. The differential-geometry family does not order on these photographs, and the reason is that it barely spans a range of scales here, so it has nothing to order. Smoothing the field at 1, 2, 4 and 8 tokens turns the aspect direction

by a median of 8.2, 15.2, 25.4 and 37.9° away from its unsmoothed value, against 20.0, 41.7, 65.3 and 80.2 on the overhead imagery, where 90° is chance for a circular difference. An indoor range field is close to piecewise planar, so smoothing it mostly preserves the surface orientation. The overhead protocol pays for its range in the other currency, its sloped token fraction falling from 0.634 to 0.007 over the same widths where this one holds at 0.924 to 0.888, which is why Appendix C.1 leaves aspect out of its monotonicity check. The two protocols therefore fail in opposite directions and neither is the instrument for both, so this appendix claims the ordering on the band-passed family and nothing about the differential-geometry one.

D.6 Changing the architecture and the capacity

Every backbone in Table 2 is a 24 block ViT-L read by global self-attention, so that probe width and spatial mixing are identical across models and a difference can only come from what pretraining put in them. The same choice is why the main text cannot say whether the account is about spatial mixing, about one width, or about that one architecture. Seven further backbones settle the two questions separately, all outside the main table and reported only here, and they run from $L = 11$ to $L = 39$ where every backbone of Table 2 has $L = 23$. Swin-B and Swin-V2-B, windowed and hierarchical, change the mixing and nothing else that can be held, both being 24 blocks ending at 1024 channels so that depth and output width match a ViT-L exactly. That match is what makes them a test of the mixing rather than one more pair of models. One property of them runs against the prediction being tested. Token width quadruples down the stages, so a late block is read by a wider probe than an early one, which can only help a late block and makes an early peak a conservative observation.

The three released plain DINOv2 checkpoints change the capacity and nothing else, holding the objective, the corpus and the training code fixed at 768, 1024 and 1536 channels over 12, 24 and 40 blocks. The same conservatism applies, since the wider rungs are read by wider probes, which can only raise a late-block score and makes a flat retention the harder observation. One ladder is one recipe, so a second answers that, namely the released DINOv3 checkpoints trained on LVD at 384, 768 and 1024 channels, whose top rung is DINOv3-Web of Table 2 and which therefore adds two rows rather than three. Table 11 gives all seven.

Table 11: The seven backbones outside the main table, the first two changing the mixing and the two groups below them the capacity, at patch 14 and at patch 16. The response radius r_L is on the same protocol Table 2 states. Retention is at the output block on US3D at $\tau = 5.84$ m, matching is under the 10° rotation and identity under the token-aligned shift, and the ordering column is the rank correlation between the ground scale of a band-passed height target and the depth at which it is best decoded, within that backbone. The aspect column beside it is the best block's absolute error in degrees, since a retention near one is cheap for a backbone that is weak everywhere. The three DINOv2 rows are the released plain checkpoints, so the middle one is not the register-token variant Table 2 reports and the two are separate training runs.

backbone	mixing	blocks	channels	r_L	aspect	best aspect	identity	ordering
Swin-B	window	24	1024	3.95	0.770	52.97	0.265	+0.949
Swin-V2-B	window	24	1024	4.29	0.814	56.20	0.262	+0.738
DINOv2-B/14	global	12	768	4.08	0.929	48.66	0.439	+0.894
DINOv2-L/14	global	24	1024	3.69	0.918	45.04	0.430	+0.821
DINOv2-g/14	global	40	1536	6.85	0.908	43.55	0.482	+0.894
DINOv3-S/16	global	12	384	9.97	0.969	54.42	0.829	+0.894
DINOv3-B/16	global	12	768	7.09	0.916	49.81	0.939	+0.707

The ordering survives both changes. Both windowed backbones order their peak depths by target scale, at +0.949 and +0.738, all three rungs of the patch-14 ladder do, at +0.894, +0.821 and +0.894, and both new rungs of the patch-16 one do, at +0.894 and +0.707. So the ordering of Section 5 does not depend on having 24 blocks of global attention, nor on one token width. These seven are also where the premise of Equation 2 mostly holds, since only 3 of the nine backbones of Table 2 are in its scope at all and all three of those show the ordering. That is why the counts of

Appendix D.3 are stated over its whole pool. Both new rungs satisfy the premise at the output block itself, which takes it to 8 admitted backbones, and every one of those 8 shows the ordering.

Capacity buys skill and does not buy retention. If fine geometry at the output were something a larger model supplies, the ladder would narrow its radius and raise its retention as it grows. Neither happens. The radius is not even monotone in the size, the 1024 rung reading 3.69 tokens against 4.08 for the 768 one before the 1536 one widens to 6.85, a span of 3.16 tokens with no ordering by capacity in it. Aspect retention stays inside 0.021 over the three rungs and normal retention inside 0.029, which is narrower than the bootstrap interval on a single probe score. Over exactly the same rungs the absolute aspect error falls monotonically from 48.66 to 43.55°. So the ladder gains 5.12° of skill and no retention, which is exactly what Equation 1 is built to divide out.

The second ladder makes that claim stronger and bounds the instrument. The DINOv3 ladder agrees with the first on the claim and disagrees with it on the form. Its radius is again out of order with the size, the 384 rung reading 9.97 tokens where the 768 one reads 7.09 and the 1024 one 7.67, and its absolute aspect error again falls monotonically over the same rungs, from 54.42 to 40.94°. But its aspect retention is not flat, falling from 0.969 to 0.916 to 0.850 as the model grows, a span of 0.118 against the first ladder's 0.021. So the claim two ladders support is that capacity never buys retention, and not that retention is constant across capacity, which only one of them shows. The instructive cost is that DINOv3-S has the widest radius of anything measured here, 9.97 tokens under the protocol Table 2 states, while retaining 0.969 of its aspect, so for this recipe a wide response and a well kept fine target coexist. Part of the reason is that retention near one is cheap for a backbone weak everywhere, its absolute error being 54.42°, but only part, since CLIP-L is no better in absolute terms at 54.68 and retains only 0.797.

The three links a wide radius runs through, since they are measurable. Reading a wide radius as less fine geometry at the output runs through three links, a wide radius meaning more mixing, more mixing meaning a token less distinguishable from its neighbours, and that in turn meaning the target is gone. For this family the first link holds and the second does not. The radius is not a whitening artefact, since the whitened, raw and layer-normalised distances of Appendix A put DINOv3-S at 9.97, 9.70 and 9.72 tokens. It is produced in a single block, since the three rungs read 3.34, 3.20 and 6.00 tokens at the penultimate block, which is ordered by size, and the output block then multiplies them by 2.99, 2.22 and 1.28, a factor that falls as capacity rises. And that far field costs this family almost none of its token identity, since of the eight backbones past six tokens the three DINOv3 rungs are the top three under the token-aligned shift, at 0.965, 0.939 and 0.829, where the other five run from 0.204 to 0.711. Register tokens are not the explanation, since the register and plain DINOv2-L releases read 0.433 and 0.430. So for this recipe the mass-weighted radius reports a response the probe is largely indifferent to, which bounds what a radius alone can mean and is why every use of it in this paper is either inside one network or beside the token identity Section 8 measures directly.

A windowed hierarchy loses more fine geometry than its radius predicts, and that is instructive. The prediction was that attention confined to a shifted window cannot grow a wide response radius, so a windowed backbone should sit at small radius and high retention, and half of it holds. The radius, under the protocol Table 2 states, is smaller than any global-attention backbone of comparable absolute skill, 3.95 and 4.29 against 6.80 to 7.67 for the supervised, language-aligned and self-distilled ViT-Ls, so the window does constrain the mixing. But the two Swins take two of the three lowest aspect retentions of all 16 backbones, 0.770 and 0.814 with CLIP-L's 0.797 between them, and their token identity collapses to 0.265 and 0.262, so they lose fine geometry more badly than their radius alone would say.

Two properties of a hierarchy account for it and neither contradicts the mechanism. A window is a small footprint only in the units of the stage it sits in. Swin's window of eight tokens at the 32 pixel stride of its last stage reaches 256 pixels, which is 16 tokens of the 16 pixel grid the response is read on, against 64 for a globally attending ViT-L at the same input size. And a hierarchy discards spatial resolution outright, its last stage producing 32 tokens a side where the analysis grid has 64, so structure below two analysis tokens is not represented at all. The response radius does not see the second of these, because it reports how far a response spreads and not how coarsely the result is sampled. So aggregation is what removes fine geometry, here in a second and more explicit form

than attention, and on a hierarchical backbone the radius understates the loss, so the output stride has to be read beside it. The main text's nine backbones all share a stride, which is why the question does not arise there.

What a radius alone predicts across architectures, which sizes that scope. Two cases are not a scope, so the statement is sized over every backbone measured for this paper, the nine of Table 2, the seven of this appendix and the three of the controlled supervision pair, which is 19. Table 12 compares three predictors of output-block aspect retention. Two of them need no labels, the response radius and the token identity under the token-aligned shift, and the third is architecture metadata rather than a measurement, namely the analysis grid divided by the grid the backbone emits, which in this pool separates the two hierarchical backbones from the seventeen isotropic ones. Rank correlation says whether a predictor orders the pool and leave-one-out prediction says whether it is worth conditioning on, the second being the one to read, since at 19 points and up to three predictors an in-sample coefficient of determination measures nothing.

The radius orders the pool at $\rho = -0.442$ with a permutation p of 0.059 and cuts the error of predicting the pool mean by $6.2\,\%$, so taken alone and across architectures it is worth close to nothing. Token identity orders it at $+0.733$ at $p = 0.0006$ and cuts that error by $31.8\,\%$, and adding the radius to it buys a further 1.2 points. The three links this appendix names are where that comes from, since it is the first of them that fails, a wider response not meaning a token less distinguishable from its neighbours once the pool spans architectures, at $\rho = -0.267$ and $p = 0.27$. A radius is therefore a length inside one network, where it calibrates the ordering, sets the crossover of Section 6 and is what the intervention of Section 7 moves, and not a statistic that ranks independent backbones, which is the scope the conclusion states.

The same choice tested forward, where the outcome has almost no range. Choosing among three predictors on the pool one then reports is how a statistic comes to look better than it is, so the choice is tested forward, on the 9 further backbones that carry the band-limited radius, the probes and the token-aligned shift and were not in that pool. Nothing is refitted, and the ordering does not hold outside the set it was read on. The radius orders them at -0.250 and token identity at $+0.250$, both with a permutation p above 0.5, and the fit made on the 19 is worse out of sample than that set's own mean by a factor of about three for either predictor. The reason is measurable and is the first thing to report, since these nine span the radius from 1.27 to 13.66 tokens and token identity from 0.069 to 0.967, which is the whole range the pool spans, while their aspect retention spans 0.068 against the pool's 0.230. There is almost nothing there to predict, and a comparison with no range in its outcome is not evidence either way, which is the rule this paper applies to a predictor with no range in it. What the nine do establish is a bound on that table. Token identity is the better of the three inside the pool, and neither it nor the radius is offered as a general predictor of fine-geometry retention.

Table 12: Three predictors of aspect retention at the output block, over the 19 backbones measured for this paper. The response radius and the token identity are label-free measurements and the stride is architecture metadata, the analysis grid of 64 divided by the grid the backbone emits. Rank correlation carries a two-sided permutation p over 2×10^5 draws, since 19 backbones cannot be enumerated the way nine can. Prediction error is the mean absolute leave-one-out error of a linear model, and the cut is against predicting the pool mean, so a predictor worth nothing reads zero there.

predictor	rank correlation	permutation p	prediction error	cut
none, the pool mean			0.0587	
response radius	-0.442	0.059	0.0551	$6.2\,\%$
token identity	$+0.733$	0.0006	0.0401	$31.8\,\%$
output stride	-0.501	0.023	0.0500	$14.9\,\%$
radius and identity			0.0394	$32.9\,\%$
all three			0.0378	$35.6\,\%$

E THE MECHANISM, ITS CONTROLS AND ITS SPECIFICITY

Everything behind Section 7 is here, namely the blur that adds mixing, the read-out controls that cannot undo it, the attention intervention and its dose, the trained read-outs and the photometric control that says what the mixing does not explain.

E.1 THE BLUR EQUIVALENCE ON A SINGLE SCALE

The equivalence on every backbone of the main table. Section 7 reports the score-matched width on DINOv3-Web. Unlike the attention intervention, which acts on the forward pass, this perturbs the representation after it, so it cannot move a network off the distribution its weights were trained on and it extends to every backbone with the instrument intact. Table 13 gives the width at which blurring block 16 degrades it to the output block's own score on each differential-geometry target. The direction is worth stating, since on these targets block 16 is the better one and the question is how wide a blur makes it as bad. A width is reached in 9 backbones of 9 on both targets.

Equation 2 predicts more than that the width exists. The output block differs from block 16 by the mixing performed in between, and independent kernels compose in quadrature, so the blur that imitates the difference should widen with $\sqrt{r_L^2 - r_{16}^2}$. That quantity was fixed from the model before the sweep finished and it orders the measured widths at $\rho = +0.636$ on aspect and $+0.650$ on the surface normal over the nine, at exact one-sided $p = 0.0359$ and $p = 0.0333$ over all 362880 relabelings, while the simpler ratio r_L/r_{16} orders them at $+0.753$ and $+0.750$ at $p = 0.0120$ and $p = 0.0127$. Neither was chosen on the correlation it reports and we do not claim between them, and both reject the hypothesis that the width the blur needs is unrelated to the radius the backbone gained. So the width needed to imitate depth by blur tracks the response radius the backbone actually gained, across four pretraining objectives and a radius range from 0.71 to 7.67 tokens, and that quantitative content of the equivalence is measured on every backbone rather than on one.

The coarse and semantic targets are not a control here in the way they are on DINOv3-Web. On a target depth helps, the output block is the better one and blurring an earlier block can only move it further away, so no width can imitate it and the statement is about the sign of the gap. Depth helps both coarse targets in 3 of the 9, and on the other six the output block is not the best block for coarse height either, so those cells say nothing about specificity in either direction and we do not read them.

Table 13: The width k in tokens at which average-pooling block 16 over a $k \times k$ token window degrades it to the output block's own score, on US3D at $\tau = 5.84$ m, interpolated between the swept widths 1, 3, 5, 7 and 11. The response radius is in tokens under the band-limited protocol of Table 2, at block 16 and at the output block, and the added radius is $\sqrt{r_L^2 - r_{16}^2}$. Rows are in the order of Table 2.

backbone	r_{16}	r_L	added	k aspect	k normal
Scale-MAE-L	0.68	0.71	0.22	1.41	1.10
SatMAE-L	0.72	0.88	0.50	1.29	1.09
MAE-L	0.83	0.98	0.52	1.00	1.00
DOFA-L	0.57	1.00	0.82	2.32	2.30
DINOv3-Sat	0.95	1.33	0.93	1.00	1.05
CLIP-L	3.30	6.80	5.95	4.34	4.59
DINOv2-L	2.48	7.49	7.06	3.17	3.44
ViT-L ImageNet-21k	3.51	7.63	6.78	2.74	4.97
DINOv3-Web	2.92	7.67	7.09	2.64	2.59

A second width, matched on smoothness rather than on score. The score-matched blur width of Section 7 answers how wide a blur of block 16 reproduces a later block's probe score. A second and independent question is how wide a blur reproduces that block's own smoothness, which we measure as the spatial autocorrelation length of the token feature map over 60 tiles. For DINOv3-Web at $\tau = 5.84$ m that length is 0.99, 1.04, 1.21, 1.70 and 5.31 tokens at blocks 8, 12, 16, 20 and 23, while blurring block 16 gives 1.21 at $k = 1$, 3.90 at $k = 3$ and 5.86 at $k = 4$, so the output block's feature map is as smooth as block 16 blurred by $k \approx 3.7$. The score-matched width of $k \approx 2.6$ and the smoothness-matched width of $k \approx 3.7$ therefore agree in order and in direction.

The remaining gap has the sign the account requires, since the deep block also gained context on the way, so slightly less blur matches its score than matches its smoothness.

E.2 The window read-out against the window width

Table 14 is the dose-response behind Section 7. The module is the same one the read-out table uses, a low-rank projection of the block followed by a mixing of the $k \times k$ window, so that only k changes down a column. Run at $k = 3$ on the output block it reproduces the window row of Table 18 to five decimal places, at $11.02539°$ against $11.02548°$, which confirms that this is the same instrument at a different width. The final training loss is quoted because it is what separates the two ways a wider window can fail to help, since a window that overfits still drives the training loss down whereas a window with nothing further to extract cannot. Read down the surface-normal columns, the output block does not move, at $+0.6\,\%$ from $k = 1$ to $k = 9$ against a training loss that falls by $3.8\,\%$. Block 16 reads $6.4\,\%$ worse over the same range, but its scatter over seeds reaches $0.70°$ at a single width, more than the whole range over widths, so at that block the absence of a gain is resolvable and a trend in k is not, as Appendix E.4 shows. Neither block buys recovery from width, which is Proposition 2. Read down the height columns, the output block gains nothing on test while its training loss falls by $10.4\,\%$, so a block that already aggregates globally is only overfitting, whereas block 16 gains $30.7\,\%$ on test, a gain that separates on every seed, because the window supplies the global context it lacks.

Table 14: Surface-normal error and coarse-height R^2 of a read-out over a window of k tokens, for a frozen DINOv3-Web on the DFC2019 token grid, with the training loss the same run ended at in parentheses. Every entry is the mean over three seeds. Parameters include the head and grow with k, so the comparison is not capacity-matched and is not meant to be.

		surface normal (°)		height R^2	
k	parameters	output block	block 16	output block	block 16
1	0.27 M	10.966 (0.0949)	10.388 (0.0935)	0.396 (1.351)	0.246 (1.591)
3	1.32 M	11.025 (0.0914)	10.782 (0.0905)	0.406 (1.268)	0.281 (1.446)
5	3.41 M	10.948 (0.0913)	10.523 (0.0898)	0.400 (1.240)	0.300 (1.382)
7	6.56 M	11.038 (0.0913)	11.026 (0.0896)	0.402 (1.224)	0.310 (1.335)
9	10.75 M	11.032 (0.0917)	11.049 (0.0894)	0.397 (1.211)	0.322 (1.300)

E.3 The intervention grid

Table 17 gives the full grid that Figure 3b of Section 7 summarises, where each radius averages the injection blocks available at it and the token-identity control asks only whether a token remains distinguishable from its neighbours. The ordering in w is monotone for every warp, from both injection blocks and in both backbones, and the token-aligned control never moves by more than 2.3 percentage points.

On the probe side, the paired bootstrap of Section 7 over the 71 held-out tiles gives a surface-normal difference of $-0.148°$ with a $95\,\%$ interval of $[-0.258, -0.040]$ and $p = 0.0035$, a coarse-height squared-error difference of $+0.638$ with an interval of $[+0.111, +1.271]$, and semantic accuracy of 0.798 against 0.803.

How far the intervention reaches across backbones, and what decides a cell. Section 7 intervenes on two backbones and both are DINOv3, so the identical intervention was run on four more, spanning language alignment, supervised classification, a second self-distillation release and masked reconstruction, and covering the response radius from 0.98 to 7.63 tokens. Both arms of every pair dump per-token errors, the baseline included, and the pipeline reproduces the published DINOv3-Web cell exactly, so a null elsewhere is not an artefact of re-running it. Table 15 gives every cell.

At this mask the extension buys one prediction confirmed and three cells lost to the instrument, and the control is what separates the two. Section 7 states the effect as three things happening together, the surface normal returning, coarse height degrading and semantic accuracy not moving, and the

third is what says the mask narrowed the aggregation rather than damaging the representation. It holds in three of the nine cells, the published one and both MAE-L cells, and fails in all six on CLIP-L, DINOv2-L and the supervised ViT, where the mask moves semantic accuracy by more than its own interval at either width. That is what a mask imposed at inference on weights trained with global attention can do. Those six cells are uninformative rather than negative, and the dose sweep recovers two of the three backbones by making the mask smaller, while taking the confirmation back.

Where it worked the probe arm agrees with the account in both directions. MAE-L has a response radius of 0.98 tokens, so Equation 2 says there is no mixing at the output block for a window to undo and the intervention should return nothing. It returns -0.045° with an interval spanning zero at both widths, where the same measurement on DINOv3-Web at a radius of 7.67 returns the normal at $p = 0.0035$. That agreement does not survive a softer version of the same mask, and the dose sweep reports where it goes.

The task arm of the same experiment then disagrees on that backbone, and the disagreement is worth more than the agreement. Running the identical windows and scoring output-block matching under the 10° rotation, with the token-aligned control Section 7 reports it under, MAE-L gains 4.0 to 9.4 points at all four settings while its control moves by 0.1 of a point or less. It retains 0.996 of its own best block on that task, so there was nothing for a window to give back, and every one of those four cells puts it above the best block it has. So narrowing late attention does two separable things, one of them returning what mixing removed and one of them not, and MAE-L separates them because it is the only backbone here with no loss for the first to act on. That is the sharp form of the caveat Section 7 carries, that the intervention licenses the mixing as the cause and not the response radius as the whole of it, and it is why the probe arm and the task arm are reported separately. On the other three the task arm adds nothing either way. The control moves past the 2.3 points Section 7 bounds it by on DINOv2-L and the supervised ViT, and on CLIP-L the window either lowers matching by more than it moves the control or raises the two together, never separating them by more than 0.03 of a point.

One rescue for the three lost cells was available in principle and measurement rules it out. A window can only narrow an aggregation wider than itself, so the natural reading would be that it failed to bind there. It binds at least as hard, since at the injection block the band-limited radius is 2.92 tokens for DINOv3-Web against 3.30 for CLIP-L and 3.51 for the supervised ViT.

Table 15: The probe intervention of Section 7 on five backbones, each a paired bootstrap over the same 71 held-out tiles against that backbone's own unmodified run, with attention restricted from block 16 to a window of radius w tokens. Surface-normal error is in degrees and lower is better, so a negative difference is the predicted direction. The semantic column is the control the effect is defined with, and a cell whose control moves by more than its own interval is uninformative about the mechanism rather than evidence against it.

backbone	r_L	window	surface normal	95 % interval	semantics	control
DINOv3-Web	7.67	$w = 4$	-0.148	$[-0.258, -0.040]$	-0.0046	holds
MAE-L	0.98	$w = 4$	-0.045	$[-0.139, +0.048]$	-0.0034	holds
MAE-L	0.98	$w = 2$	-0.047	$[-0.160, +0.071]$	-0.0061	holds
CLIP-L	6.80	$w = 4$	$+0.023$	$[-0.077, +0.129]$	-0.0188	moves
CLIP-L	6.80	$w = 2$	$+0.077$	$[-0.050, +0.212]$	-0.0188	moves
DINOv2-L	7.49	$w = 4$	$+0.121$	$[-0.022, +0.251]$	-0.0178	moves
DINOv2-L	7.49	$w = 2$	$+0.190$	$[+0.044, +0.340]$	-0.0258	moves
ViT-L ImageNet-21k	7.63	$w = 4$	$+0.021$	$[-0.076, +0.118]$	$+0.0162$	moves
ViT-L ImageNet-21k	7.63	$w = 2$	$+0.022$	$[-0.083, +0.124]$	$+0.0298$	moves

The mask has a size, and most of what was lost to the instrument was the size. The hard mask deletes every attention pair outside the window, which is the largest member of a family and not the only one. Subtracting a finite number of nats from the logit of each pair outside the window instead gives the same intervention a dose, with zero being no intervention and a large value being the hard mask exactly. The script `scripts/check_local_attn_dose.py` verifies both limits at a maximum absolute difference of zero. How far a dose moves the output block towards the hard

mask depends on how peaked the backbone's attention already is, so it is measured per backbone from the cached features the probes read and is the third column of Table 16. The largest admissible dose is defined by the control and never by the effect, as the largest dose whose semantic interval spans zero and all of whose smaller doses also span it.

The empty dose is a control rather than padding. Passing any mask changes which attention kernel runs, worth a mean relative difference of 3.4×10^{-3} in the output block, and that difference sits inside every published cell because the baseline is unmasked. Every target is null at the empty dose on three of the five backbones. CLIP-L's surface normal moves there by $-0.0012°$, a hundredth of what it shows at its admissible dose. MAE-L's semantic control fires at $+0.0008$ of accuracy with an interval whose lower end is $+0.0000$, so there the control sits at its own resolution and a failure carries no information. Re-reading all twenty cells against the empty dose instead of the unmasked baseline changes one verdict, DINOv2-L's.

What the dose buys is that two of the three backbones lost to the instrument were lost to its size. On CLIP-L the control holds to $61\,\%$ of the hard mask, where the surface normal returns by $-0.111°$ at $p = 0.011$, and the return is ordered by dose at $\rho = -1.000$ over the three doses at or below that ceiling. At $97\,\%$ and at the mask itself the control moves and the effect is gone. On DINOv2-L the control holds to $42\,\%$, where the normal returns by $-0.126°$ at $p < 0.0001$. The causal leg is therefore clean on three backbones rather than one, and the two it gains are the language aligned and second self-distillation recipes. Read at each backbone's own ceiling the surface normal is right-signed on all five and no interval there lies wholly above zero, so no admissible cell of the sweep makes the normal worse. On the supervised ViT the control holds to $45\,\%$ and the return is right-signed and ordered at $\rho = -1.000$ but no dose resolves it, at $-0.021°$ with an interval from -0.080 to $+0.038$. DINOv3-Web holds at every dose and its published cell is past its own optimum, since $76\,\%$ of the mask returns $-0.215°$ where the mask returns $-0.148°$.

Two things are reported against the account. DINOv2-L is the one verdict that depends on the baseline, its semantic interval at that dose reading $[-0.0119, +0.0005]$ against the unmasked baseline and $[-0.0125, -0.0001]$ against the empty dose, so we keep the published convention for the number and note that its admissibility turns on the convention. And MAE-L, the one backbone whose radius leaves the account predicting no return and which the hard mask confirmed, does not survive a softer dose. At the mask it returns $-0.045°$ with an interval spanning zero, which is the published null, and at $41\,\%$ of that dose it returns $-0.0624°$ with an interval from -0.1056 to -0.0188 at $p = 0.0030$. So the null on the one backbone the account predicts nothing for is a property of the dose and not of the backbone. Selecting the strongest cell over doses is the direction it is safe to select in here, since it is the direction that costs the account a confirmation, and it is done for MAE-L and for nothing else.

Two further readings are not evidence, by rules this paper applies elsewhere. The size of the return at each backbone's best admissible dose orders by the response radius at $\rho = -0.400$, but four of these five radii sit inside 6.80 to 7.67 tokens against MAE-L's 0.98, so the predictor has almost no range. And the coarse-height arm is descriptive, since per-token errors are dumped for the surface normal, surface aspect and the semantic label but not for coarse height, and an unpaired interval over seventy-one tiles is several times the effect. It falls at the top of the range on three of the five, by 0.028, 0.015 and 0.057 of coefficient of determination, and at CLIP-L's admissible dose it does not fall. So the dose sweep re-establishes two parts of Section 7's three-part signature and leaves the third where that section measured it.

The intervention moves the radius itself, when it can. Equation 2 is written in the response radius and this appendix narrows attention inside one backbone, so the account would be unconnected to the intervention if the intervention did not move that radius. We measure it under each setting, on the same 40 tiles and 8 occlusion sites as the unmodified baseline. A local window can only shrink the aggregation if it is narrower than the aggregation the backbone already performs, and that is what happens. DINOv3-Web has a radius of 10.87 tokens, so all four windows are narrower and all four cut it, to 5.35, 4.24, 5.05 and 4.50 tokens for the settings in the order of Table 17. DINOv3-Sat has a radius of 2.29, so windows of 8 and 4 tokens are wider than its response and leave it at 2.61, 2.30 and 2.44, and the one window narrower than it, $w = 2$, is the one that reduces it, to 1.92. The rule that the radius falls exactly when the window binds holds in all eight cells, and on DINOv3-Web the radius that results tracks how far the mask lets a token reach, which is w times the number of masked

Table 16: The attention intervention given a dose. Attention outside the window of radius 4 from block 16 loses the stated number of nats from its logit before the softmax instead of being deleted, so 0 is no intervention and the hard row is the published mask. The second numeric column is how far that dose moves the output block towards the hard mask, measured on the cached features per backbone. The surface-normal column is the baseline error minus the intervened error over the same 71 held-out tiles in a paired bootstrap, so a negative value is a return of geometry. The semantic column is the same comparison on the six-class semantic head, which is the control that says the mask narrowed the aggregation rather than damaging the representation, and the last column reads it. Coarse height is not dumped per token and is discussed in the text instead.

	dose		surface normal, degrees			semantic control	
backbone	nats	% of hard	baseline minus intervened	95% interval	p	accuracy change	verdict
DINOv3-Web	0	0	-0.0004	$[-0.002, +0.001]$	0.2760	-0.0001	holds
	2	43	-0.1536	$[-0.213, -0.091]$	0.0000	$+0.0052$	holds
	4	76	-0.2148	$[-0.317, -0.116]$	0.0000	$+0.0019$	holds
	8	99	-0.1539	$[-0.263, -0.046]$	0.0035	-0.0046	holds
	hard	100	-0.1480	$[-0.258, -0.040]$	0.0035	-0.0046	holds
CLIP-L	0	0	-0.0012	$[-0.002, -0.000]$	0.0150	-0.0003	holds
	2	22	-0.0190	$[-0.055, +0.015]$	0.1410	-0.0039	holds
	4	61	-0.1107	$[-0.203, -0.017]$	0.0110	-0.0060	holds
	8	97	$+0.0061$	$[-0.090, +0.106]$	0.5675	-0.0197	moves
	hard	100	$+0.0232$	$[-0.077, +0.129]$	0.6850	-0.0188	moves
DINOv2-L	0	0	$+0.0009$	$[-0.000, +0.002]$	0.9315	$+0.0007$	holds
	2	42	-0.1259	$[-0.192, -0.066]$	0.0000	-0.0056	holds
	4	79	-0.0905	$[-0.224, +0.027]$	0.0655	-0.0117	moves
	8	99	$+0.1135$	$[-0.031, +0.242]$	0.9455	-0.0176	moves
	hard	100	$+0.1206$	$[-0.022, +0.251]$	0.9575	-0.0178	moves
ViT-L ImageNet-21k	0	0	-0.0002	$[-0.001, +0.000]$	0.2190	$+0.0005$	holds
	2	14	-0.0067	$[-0.027, +0.014]$	0.2600	$+0.0017$	holds
	4	45	-0.0207	$[-0.080, +0.038]$	0.2460	$+0.0105$	holds
	8	90	$+0.0017$	$[-0.092, +0.093]$	0.5200	$+0.0178$	moves
	hard	100	$+0.0212$	$[-0.076, +0.118]$	0.6785	$+0.0162$	moves
MAE-L	0	0	$+0.0001$	$[-0.001, +0.001]$	0.5800	$+0.0008$	moves
	2	41	-0.0624	$[-0.106, -0.019]$	0.0030	$+0.0002$	holds
	4	82	-0.0680	$[-0.150, +0.010]$	0.0440	-0.0024	holds
	8	100	-0.0466	$[-0.141, +0.046]$	0.1550	-0.0027	holds
	hard	100	-0.0453	$[-0.139, +0.048]$	0.1645	-0.0034	holds

blocks, at $\rho = +0.80$. This calibrates the radius against a manipulation rather than against another correlation. It does not make the radius the whole story for matching, since matching accuracy rises in all eight cells including the three where the radius does not fall.

E.4 The trained read-outs

Table 18 is the read-out comparison that Sections 7 and 8 draw on. It, the window table of Appendix E.2 and the decoder column of Table 19 are the only measurements in the paper that train anything, so they are the only ones whose numbers depend on an initialisation. Each quotes a mean over the same three seeds, 0, 1 and 2. Everything else in the paper is a closed-form ridge or logistic fit, or a forward pass, and carries no seed, with sampling uncertainty reported instead as a bootstrap over whole tiles.

The same comparison with a convolutional decoder instead of a linear head. Proposition 2 is proved over linear read-outs and Table 18 tests linear read-outs, which invites the question of whether a decoder with spatial extent and two orders of magnitude more parameters closes the gap that a block of depth opens. It does not. Table 19 repeats five rows of that comparison with the head replaced by two 3×3 convolutions of 256 channels, over the same three seeds and nothing

Table 17: Output-block matching accuracy on US3D at $\tau = 5.84$ m under local attention imposed from a given block with window radius w in tokens. Every row, including the unmodified one, is measured on the same 40 tiles. The final column is the token-aligned control.

backbone	configuration	half-token shift	rotation 10°	scaling 1.1	control
DINOv3-Web	unmodified	0.686	0.581	0.604	0.966
	from block 12, $w = 2$	0.748	0.656	0.657	0.987
	from block 12, $w = 4$	0.734	0.634	0.630	0.989
	from block 16, $w = 4$	0.750	0.655	0.651	0.988
	from block 16, $w = 8$	0.723	0.618	0.619	0.982
DINOv3-Sat	unmodified	0.561	0.603	0.635	0.997
	from block 12, $w = 2$	0.705	0.690	0.711	0.998
	from block 12, $w = 4$	0.661	0.661	0.677	0.997
	from block 16, $w = 4$	0.634	0.663	0.678	0.997
	from block 16, $w = 8$	0.597	0.632	0.651	0.997

Table 18: What the read-out is allowed to see, for a frozen DINOv3-Web with a 1×1 head on the DFC2019 token grid. Parameters include the head, quoted for the largest of the three read-outs, percentages are relative to the first row, and mIoU denotes mean intersection over union of the six classes. Every entry is the mean of three runs, whose largest standard deviation is 0.40° on surface normals and 0.016 on height.

what the read-out sees	parameters	surface normal (°)	height R^2	mIoU
output block, one token	0.008 M	11.177	0.279	0.651
output block, 3×3 window	1.320 M	11.025 (−1.4 %)	0.406 (+45.5 %)	0.656
output block, non-linear map	0.271 M	10.500 (−6.1 %)	0.572 (+105 %)	0.669
block 16, one token	0.008 M	10.317 (−7.7 %)	0.180 (−35.5 %)	0.638
seven blocks, hypercolumn	7.362 M	10.274 (−8.1 %)	0.409 (+46.6 %)	0.668
a random earlier block, injected	0.275 M	11.210 (+0.3 %)	0.253 (−9.3 %)	0.659
block 16, injected	0.275 M	9.748 (−12.8 %)	0.461 (+65.1 %)	0.671
three blocks, routed by scale	0.820 M	9.747 (−12.8 %)	0.543 (+94.6 %)	0.681

else changed. Every read-out improves and the order of the five is exactly the order the linear head gives, the output block staying worst of them. The decoder buys the output block 12.1 % and the block choice buys a linear head 12.8 %, so the two are worth about the same and they compose, the decoder on the diagnosed block reaching 9.018°. The sharp form is that 2.952 M of decoder on the output block reaches 9.824° where 0.275 M injecting block 16 under a linear head already reaches 9.748°. A decoder is therefore not a substitute for reading the right block, which is the practical content of Proposition 2 and is what the abstract claims in place of the unqualified statement that no read-out recovers the loss.

Table 19: Five read-outs of a frozen DINOv3-Web under two heads, a 1×1 linear read-out and a decoder of two 3×3 convolutions at 256 channels, scored by per-token surface-normal error in degrees on DFC2019, lower being better. Parameters are for the decoder column and include the head. Each entry is the mean of seeds 0, 1 and 2, with the standard deviation beside it. The concatenation was run at one seed under the linear head and its entry there is omitted rather than quoted from one run.

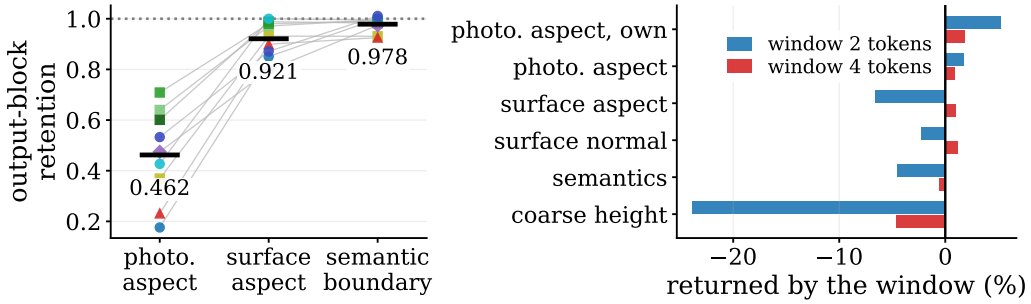
what the read-out sees	parameters	linear head	convolutional decoder
output block	2.952 M	11.177 ± 0.148	9.824 ± 0.027
block 16	2.952 M	10.317 ± 0.002	9.640 ± 0.247
seven blocks, hypercolumn	10.306 M	10.274 ± 0.495	9.357 ± 0.037
output block and block 16	5.314 M		9.311 ± 0.048
block 16, injected	3.219 M	9.748 ± 0.048	9.018 ± 0.040

Three things in those runs bear on how the results should be read. The hypercolumn is the least stable read-out in the set, at a standard deviation of 0.40° on surface normals against 0.04° for the

injected block, and its mean is carried by two of its three runs. The diagnosed block beats it on every seed and not only on the mean, which is what licenses the comparison Section 8 draws. In the window table the surface-normal read-out of block 16 scatters by up to $0.70°$ at one width, so the apparent worsening in k there is not resolvable, as Appendix E.2 states. Coarse height at block 16 is the opposite case, since its worst run at $k = 1$ reads 0.268 and its best run at $k = 9$ reads 0.317, so the gain holds on every pairing of runs and is the one width effect in the table that does.

E.5 The photometric control, and what the window does not return

Sections 5 and 7 rest on a target built to be everything surface aspect is except geometric, and this appendix states how it is built and what it reads.



(a) what survives

(b) what the window returns

Figure 5: (a) Output-block retention of three fine-grained targets, one marker per backbone with the mean marked and the pairing drawn faintly. All three are defined inside one token, so what separates them is modality and redundancy and not fineness. (b) What forcing the late blocks into a window of 2 and of 4 tokens returns for DINOv3-Web, as a percentage of the unmodified score, ordered by scale.

Construction, which introduces no free parameter. Surface aspect is the compass direction of steepest descent of the height field, taken as $\operatorname{atan2}(n_y, n_x)$ of the token-pooled surface normal. The photometric target calls the same two functions on the luminance field of the same image, so the two differ in the field and in nothing else. Three properties make it a control rather than an analogy. The direction is invariant to any scaling of the field, so no gain or unit has to be chosen, where a magnitude would have needed one. It is read on the identical token population, since the probe gives it the same mask the geometric target takes, namely a surface slope above 0.15. A second arm repeats it on the tokens where the luminance field itself varies most, cut at the quantile that keeps the same fraction the geometric mask keeps rather than at a tuned threshold. And its chance level, the error of the best constant direction, is $82.06°$ against surface aspect's 81.25 on those same tokens and 89.46 on its own, so a retention on it is read against the same scale.

It is decoded, and then given up. A retention on a target at chance means nothing, so the absolute scores are what license the comparison. Over the nine backbones the best block reads 9.49 to $34.12°$ against that chance of 82.06, which is 58 to $88\,\%$ of the usable scale, and the output block reads 13.56 to 71.93. The peak sits at relative depth 0.00 to 0.17 in every backbone, so the finest target in the paper is best decoded at the first blocks and is then lost monotonically, which is the extreme case of Section 5 rather than an exception to it. Retention averages 0.462 on the shared mask and 0.354 on its own, against 0.921 for surface aspect, lower in nine backbones of nine at a two-sided paired Wilcoxon $p = 0.0039$ for both arms. The two coarser photometric targets, the mean luminance gradient and the within-token luminance spread, retain 0.867 and 0.847, so the panel is ordered by scale inside one modality as well as across the two.

What the window returns, and what it does not. Figure 5b reports the output-block change under each window and the paired bootstrap over the same held-out tiles gives the intervals. On DINOv3-Web the window of radius 4 from block 16 returns the surface normal by $-0.148°$ with an interval of $[-0.258, -0.040]$ at $p = 0.0035$, and the same window returns photometric aspect

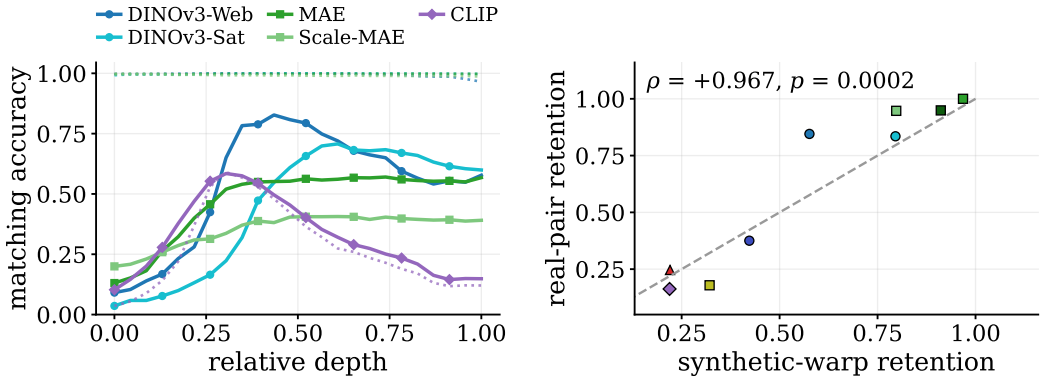
by $-0.546°$ with an interval of $[-1.987, +0.777]$ at $p = 0.23$. The tightest window, radius 2 from block 12, returns it by $-1.035°$ with $[-2.875, +0.711]$ at $p = 0.13$. On DINOv3-Sat no window improves it at all, the five settings costing 1.7 to 16.9 per cent. Coarse height degrades under every window, most at the tightest, which is the sign the account requires and the evidence the intervention is doing what it claims. So the photometric arm decays with depth, is not recovered by narrowing the mixing, and is therefore lost for a reason this paper's mechanism does not supply, most plausibly an invariance the pretraining objective rewards. We report it as a bound on the mechanism and as a control on the intervention, since an intervention that improved every target that decays would be evidence of much less.

One tension worth stating. Across the nine backbones the response radius tracks photometric retention at $\rho = -0.783$ with an exact $p = 0.017$, more strongly than it tracks surface aspect at -0.667, and yet the intervention does not return the photometric target inside a network. The across-backbone association and the within-network causation therefore come apart here, most likely because a backbone with a wide response has also abstracted away appearance, and the intervention is the test that separates them. It is the reason the radius is read in this paper as the length Equation 2 needs and not as evidence on its own that mixing caused any particular loss.

F THE TASK COST AND THE REPAIR

Everything behind Section 8 is here, namely matching accuracy against depth on synthetic warps and on real pairs, the training-free repair per backbone, and how far the unlabelled choice of block reaches.

F.1 MATCHING AGAINST DEPTH



(a) matching against depth

(b) real pairs against synthetic

Figure 6: (a) Matching accuracy against relative depth on US3D at $\tau = 5.84$ m under a $10°$ rotation, for one backbone per pretraining objective, with the dotted curves giving the token-aligned control. (b) Retention on real bi-temporal acquisitions of the same ground against retention under the synthetic rotation on identical Maxar tiles, with the identity line, which is the check Section 8 reports. Points are the backbones of Table 2 in that order, with marker shape the pretraining objective.

Figure 6a gives the depth curves summarised in Section 8. Matching accuracy is an inverted U, since early blocks are not invariant to the resampling the warp introduces while late blocks no longer report what lies under a token, and the dotted control shows that token identity is at ceiling from the first block in the five backbones that retain it. The cost grows as the imagery gets finer, as Section 5 expects, DINOv3-Web retaining 0.776 under the $10°$ rotation at $\tau = 16$ m and 0.698 at 5.84 m. Token identity also predicts matching on the real pairs, at $\rho = +0.900$ with an exact $p = 0.0020$.

F.2 THE TRAINING-FREE CONCATENATION, PER BACKBONE

Table 20 is the per-backbone form of the repair reported in Section 8. Its token-identity column separates the two regimes discussed there, since the four backbones whose token identity has collapsed are also the four where keeping the output half costs the concatenation most of the available gain.

On real pairs the concatenation exceeds the best single block. Table 20 is measured under synthetic warps, where for DINOv3-Web the concatenation reaches $99\,\%$ of what the best single block gets and therefore does not beat it. On the real bi-temporal Maxar pairs it does. Reading matching accuracy in percentage points, the output block alone gives 35.85, the single block at relative depth 0.43 gives 42.41, and carrying that block alongside the output gives 48.82. Comparing two read-outs of one backbone over 61 tiles needs a paired test, since the tile-to-tile variance of matching accuracy is several times the difference between read-outs and the separate intervals only just fail to overlap. Resampling the same tiles for both, and with the carried block fixed at the depth named above rather than chosen on these tiles, the concatenation beats the output block by 12.97 percentage points with a $95\,\%$ interval of $[10.80, 15.11]$ and beats the best single block by 6.41 with an interval of $[5.41, 7.45]$. Choosing the block on these tiles instead moves the second figure to 6.56 with an interval of $[4.94, 8.15]$, so the conclusion does not rest on the choice.

The token-aligned shift also tells the two regimes apart before the repair is run, as a tendency and not a classifier, with group medians of $92.5\,\%$ and $56\,\%$ of the best block's gain recovered at a one-sided Mann-Whitney $p = 0.029$.

Table 20: A repair that costs nothing. Matching accuracy in percentage points, averaged over the four non-trivial warps on US3D at $\tau = 5.84$ m, for the output block alone, the best single block, and that block concatenated with the output block after whitening each half separately. Both of those columns choose the block per warp on the metric, so the last column, the concatenation's gain as a fraction of the best block's gain, is unaffected. Token identity is the control of Section 8.

backbone	token identity	output block	best block	concatenated	recovered
MAE-L	0.999	53.6	53.7	54.2	no gap
SatMAE-L	0.999	43.3	47.1	46.0	71 %
DINOv3-Sat	0.998	59.9	67.6	67.1	94 %
Scale-MAE-L	0.990	43.6	45.5	45.4	91 %
DINOv3-Web	0.965	62.1	81.6	81.5	99 %
ViT-L ImageNet-21k	0.711	14.9	46.2	31.2	52 %
DOFA-L	0.607	13.2	41.7	30.2	60 %
DINOv2-L	0.433	44.4	82.2	71.8	72 %
CLIP-L	0.204	18.9	63.3	41.4	51 %

The block chosen without labels, per backbone. Table 21 gives the block the label-free criterion of Section 8 picks under each of the four warps, beside the block a labelled surface-aspect sweep would have picked and the error that separates the labelled choice from the output block. The chosen block sits 3.3 to 4.3 blocks from the labelled one on average and never deeper than the output block, and the four warps agree with each other more closely than any of them agrees with the oracle, which is what a common cause of the two rankings predicts. On surface normals the same rule recovers 49 to $79\,\%$ rather than 74 to $76\,\%$. The reason is the compression of Appendix D.4. No backbone's gap between the oracle and the output block reaches 1° on that target, so the pooled share is a ratio of differences the instrument can barely resolve.

Selecting on one region and scoring on another. A criterion that consumes no ground truth can still be optimistic if it is exercised on the same imagery the result is read from, so we move the selection off that imagery entirely. Three selection sources are used against one fixed evaluation, the surface aspect probe on US3D at $\tau = 5.84$ m, and the pooled recovery is 74 to $76\,\%$ selecting in domain, 72 to $80\,\%$ selecting on 61 Maxar tiles of Florida, and 71 to $80\,\%$ selecting on DFC2019 read at a 16 m token against the evaluation's 5.84 m. The three agree inside the spread that the four warps already produce, so the choice of block is insensitive to where it is made, and the Maxar case is the strong form of that statement, since Maxar carries no height product at all and no pixel of it enters the evaluation. On the Maxar tiles the real bi-temporal pair can also stand in for the synthetic

Table 21: The read-out block chosen with no labels, against the block chosen with them. Backbones are in the order of Table 2. The oracle column is the block that minimises surface-aspect error on US3D at $\tau = 5.84$ m and the four warp columns are the blocks that maximise matching accuracy on the same tiles, each of which needs no ground truth. The last column is the aspect error in degrees that the oracle block saves over the output block, which is what the label-free choice is trying to capture and is under $1.2°$ for the two backbones where the choice costs rather than saves.

backbone	oracle	half-token shift	rotation 3°	rotation 10°	scaling 1.1	oracle gain
Scale-MAE-L	13	10	12	14	11	1.56
SatMAE-L	14	18	17	17	18	0.57
MAE-L	18	23	23	17	23	1.03
DOFA-L	12	0	3	9	6	4.37
DINOv3-Sat	23	23	14	14	13	0.00
CLIP-L	6	8	7	7	7	14.14
DINOv2-L	14	11	10	10	10	6.91
ViT-L ImageNet-21k	4	7	3	6	3	6.41
DINOv3-Web	16	12	10	10	10	7.22

warp, and it gives $79\,\%$, so the criterion does not even need a warp of our own when two dates of the same place are available.

Two limits belong beside that. The rule is not safe on a backbone whose output block is already the best one, since it then moves the read-out for nothing. What says which case one is in is not the response radius, which Appendix D.6 shows does not order backbones across architectures, and the paragraph on the rule's condition names it. And running the evaluation on DFC2019 at $\tau = 8$ m instead gives a pooled share with no denominator worth dividing by, since only five of the nine backbones have labels there and three of those five have an oracle gain under $1.5°$. We report per-backbone differences there and no ratio.

The condition for applying the rule, since it is not the radius. The criterion maximises matching accuracy over blocks, so the margin it wins by, namely that accuracy at the block it picks minus the accuracy at the output block, is a number the rule already has and that needs no labels. Over the 288 cells of nine backbones, eight targets and four warps it picks the output block itself in 32, where it is a no-op and its cost is exactly zero. On the 256 where it moves, that margin orders how much the move was worth on surface aspect at $\rho = +0.740$ and on no other target, the median over the eight being -0.252. That is the answer rather than a failure of it. The criterion is a localisation score, so Equation 2 says it should track the benefit on a target defined below the aggregation scale and say nothing about one defined above it. The split it produces is the paper's own division between differential geometry and the coarse and semantic targets rather than a threshold fitted here. The condition for applying the rule is therefore a property of the target and not of the backbone, which is what makes it usable, since the target's scale is known before any measurement is made.

The rule on the second differential-geometry target, where it recovers less. The block the warp names is scored above on surface aspect, which is the paper's primary retention instrument for the reason Appendix D.4 gives. On the surface normal the same choice recovers 49 to $79\,\%$ of the oracle gap against aspect's 74 to $76\,\%$, the four warps giving 79, 50, 56 and $49\,\%$. The gap between the two targets is the compression Appendix D.4 measures rather than a different block, since the packed normal metric leaves about a fifth of its nominal scale usable and a share of a gap on it is correspondingly noisier. The smaller figure is the lower bound on what the rule delivers inside differential geometry.

The rule outside differential geometry. Scored on the other seven targets of the probe battery the same choice recovers 67 to $108\,\%$ of the oracle gain on the backbones whose gain exceeds the resolution of the target. A share above $100\,\%$ is correct and is not clipped, since the oracle here is the block the validation split names and the warp can land on a block that beats it on the test split, which is the same fact as the retention above 1 that Equation 1 allows. On the backbones with no gap to recover it costs instead, on one to five of the nine depending on the target, and the largest single cost anywhere in the seven is 0.222 of coefficient of determination on coarse height. Those

costs are reported in the units of the target rather than as a share of a gap, for the same reason, the gap being exactly zero for two of the nine on coarse height. That target is the extreme case, since its oracle block improves on the output block by less than 0.05 of coefficient of determination for seven of the nine and is the output block itself for two of those seven, so moving the read-out earlier is a pure cost. This is a consistent picture rather than a mixed one. The criterion finds the block that preserves token-level detail, which is the right block exactly when detail is what the target needs and the backbone has lost it. The target's scale says the first in advance, and the margin the criterion wins by reports the second.

On a trained read-out the radius resolves and the choice of block does not. A linear probe cannot compensate for a poor choice of block and a trained head can, so the same comparison was run where something is trained. For each of the four backbones with a dense feature cache we trained the low-rank injection of Table 18 at every mid block the cache holds, which is 0, 4, 8, 12, 16 and 20, at two seeds. The output-block baseline is read from the same two seeds, so the reference is a trained oracle rather than a probe standing in for one. The metric is the per-token angular error the read-out table reports. The block the warp names and the block the labelled sweep names differ on all four backbones, so nothing here is settled by the two happening to agree.

The pooled within-cell seed standard deviation is $0.519°$, so a difference between a pair of two-seed means carries a standard error of 0.519, and we call a difference resolved when it reaches two of those. Against that, one backbone resolves and three do not. Radii here are the ones Table 2 states. DINOv3-Web, at 7.67 tokens, gains $1.504°$ from its best mid block, $13.4\,\%$ and 2.9 standard errors. Its best mid block is block 16, so that cell is the one Table 18 also reports, and the $12.8\,\%$ there against the $13.4\,\%$ here is the same comparison read from the three seeds that table runs rather than from the two this sweep runs at every block. DINOv3-Sat, MAE-L and Scale-MAE-L read 1.33, 0.98 and 0.71 tokens and gain -0.237, $+0.551$ and $+0.058°$, none of them past one and a half standard errors, and the first of those is negative because its output block is already better than every mid block we tried. So the trained read-out tracks the radius rather than the block choice, since the repair pays on the one backbone the radius says will lose fine geometry and on none of the three it says will not.

Which block is injected is a further question and two seeds cannot answer it, since on the three small-radius backbones the six candidates span 0.216 to $0.512°$, which is at or below the seed standard deviation. So we report no pooled share of the oracle gain here, and the block-selection comparison stays on the linear probe, where the gap between a good and a bad choice is large relative to the noise, which is the same reason the linear read-out is the instrument of record throughout this paper.

F.3 How far the unlabelled choice of block reaches

The rule of Section 8 was written on the nine backbones of Table 2 and on overhead imagery. Before running it anywhere else we fixed how it would be scored, and applied it unchanged in two places, the four warps, the distance, the oracle chosen on the validation split and the pooled share of the oracle's gain all as in Appendix F.2. Table 22 gives every pooled share and Table 23 the blocks chosen on photographs.

Across the thirty-two backbones of the premise sweep. Twenty-eight of the thirty-two already had a warp curve on US3D, and the four of the confound pool were measured with the command every other curve was made with. The nine inside the pool reproduce Appendix F.2 exactly. Over all thirty-two the choice recovers 80 to $88\,\%$ of the oracle's gain on surface aspect and 68 to $78\,\%$ on the surface normal, and over the twenty-three it had never been run on, 81 to $91\,\%$ on aspect and 74 to $77\,\%$ on the normal. Those twenty-three span all nine pretraining objectives, capacities from ViT-S to ViT-g and both windowed hierarchies, so the criterion is not a property of the nine it was written on. Under any one warp it costs on at most three of the thirty-two, by at most $3.15°$ of aspect error, and on at most one of the twenty-three, by at most $0.25°$. The prediction we wrote down, at least one half on aspect over both pools, holds.

On indoor photographs the warp names a block too early. The six backbones of Appendix D.5 were scored on the finest surface normal and the finest aspect of that battery, with warp curves measured on sixty frames of the scene groups the probes train on, so that no pixel of the evaluation

Table 22: The unlabelled choice of block beyond the nine it was written on. Each cell is the share, in per cent, of the gap between the output block and the block a labelled sweep would choose that the block the warp names recovers, pooled over the backbones of the row, under each of the four warps of Table 21. The overhead rows choose and score on US3D at $\tau = 5.84$ m. The photographic rows choose on frames of the scene groups the probes are trained on and score on the held-out group. A negative share means the choice costs more than it recovers.

backbones	target	half-token shift	rotation 3°	rotation 10°	scaling 1.1
the nine of Table 2	surface aspect	76	76	76	74
the twenty-three added	surface aspect	81	91	90	91
all thirty-two	surface aspect	80	88	88	88
the nine of Table 2	surface normal	79	50	56	49
the twenty-three added	surface normal	77	76	75	74
all thirty-two	surface normal	78	70	71	68
six on NYU Depth v2	finest aspect	7	−18	−16	−36
six on NYU Depth v2	finest normal	−9	−36	−31	−56

takes part in the choice. There the choice recovers -56 to $-9\,\%$ of the normal's gap and -36 to $7\,\%$ of the aspect's, so it costs more than it recovers, and the prediction we wrote down beforehand, that it would recover at least half, is wrong. The third, that it would cost under a degree where the gap is under one, holds on MAE-L and fails on DINOv3-Sat, at $1.54°$ under three of the four warps.

Table 23: Where the warp goes wrong on photographs. For the six backbones of Table 10, in the order of Table 2, the block a labelled sweep of the finest surface normal picks on NYU Depth v2, the block each of the four warps names, and the normal error in degrees that the labelled block saves over the output block.

backbone	labelled	half-token shift	rotation 3°	rotation 10°	scaling 1.1	saving
MAE-L	19	23	23	12	12	0.51
DINOv3-Sat	18	17	13	13	13	0.13
CLIP-L	10	8	7	7	7	5.29
DINOv2-L	15	10	9	9	8	7.65
ViT-L ImageNet-21k	13	0	3	5	3	2.07
DINOv3-Web	16	10	9	9	8	5.40

Table 23 says where it goes wrong. On the three backbones whose gap exceeds $5°$ the warp names blocks 7 to 10 where the labelled sweep names 10 to 16, and the finest normal is still improving steeply between them, DINOv3-Web's error falling from $33.9°$ at block 10, the latest block any warp names for it, to $24.9°$ at block 16. We read that after the fact. The warp is a localisation score, so it pays on a target that carries detail below a token, and the finest normal of an indoor range field carries little, since smoothing that field over one token turns its aspect by $8.2°$ against $20.0°$ overhead (Appendix D.5). On photographs the block is therefore a labelled choice, and a sweep of linear probes makes it, saving 5.3 to $7.6°$ of surface-normal error over the output block on three of the six.