Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Part-Coverage and Faithfulness Decouple in ViT Token Pruning for Fine-Grained Visual Categorization

Richard Adusei

raduseі@uds.edu.gh

University for Development Studies

Kwabena Owusu-Agyeman

Kwame Nkrumah University of Science and Technology

Eric Opuku Osei

Kwame Nkrumah University of Science and Technology

Michael Asante

Kwame Nkrumah University of Science and Technology

Research Article

Keywords: Vision Transformer, Token pruning, Part localization, Faithfulness evaluation, Attribution metrics, Fine-grained visual classification

Posted Date: October 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11089563/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Part-Coverage and Faithfulness Decouple in ViT Token Pruning for Fine-Grained Visual Categorization

Richard Adusei[1,2*], Kwabena Owusu-Agyemang[2], Eric Opoku Osei[2], Michael Asante[2]

[1*]Department of Computer Science, Faculty of Physical Sciences, University for Development Studies, Tamale, Northern Region, Ghana.
[2]Department of Computer Science, College of Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ashanti Region, Ghana.

*Corresponding author(s). E-mail(s): radusei@uds.edu.gh;
Contributing authors: kwabenaoa@knust.edu.gh;
eric.opoku.osei@knust.edu.gh; masante.sci@knust.edu.gh;

Abstract

A token-pruning scorer for Vision Transformers can be judged on two criteria: how well the tokens it keeps cover ground-truth object parts, and how faithfully its ranking explains the classifier's decision. They come from separate evaluation traditions and, to our knowledge, have not been measured together for one scorer. We show that on small ViT backbones they decouple. Hidden-state input-times-gradient supervision lifts part coverage substantially while Insertion-Deletion AUC barely moves, so a scorer chosen under one criterion carries no guarantee under the other. Fine-grained benchmarks, where dense part annotations supply ground truth, are the testbed. To measure the disagreement we introduce MIDAS, a framework that quantifies the decoupling with a magnitude scalar and catalogues six supervision targets as a design space, with an unsupervised-scorer falsifier that rules out parameter-count confounds. At scorer depth six, a retention budget of 25 percent, and eight seeds, the hidden-state input-times-gradient target lifts preserved-token mIoU by 6.22 points on FGVC-Aircraft and 2.11 points on CUB-200-2011, while the paired Insertion-Deletion AUC differences are 0.003 and 0.002. On Stanford Cars the coverage gap is marginal and faithfulness materially lower, so we report it as a boundary case. Seed variance is too wide to certify strict equivalence, so the claim rests on the coverage gaps, not a parity guarantee. The effect reproduces on the smaller DeiT-Ti/16 but reverses

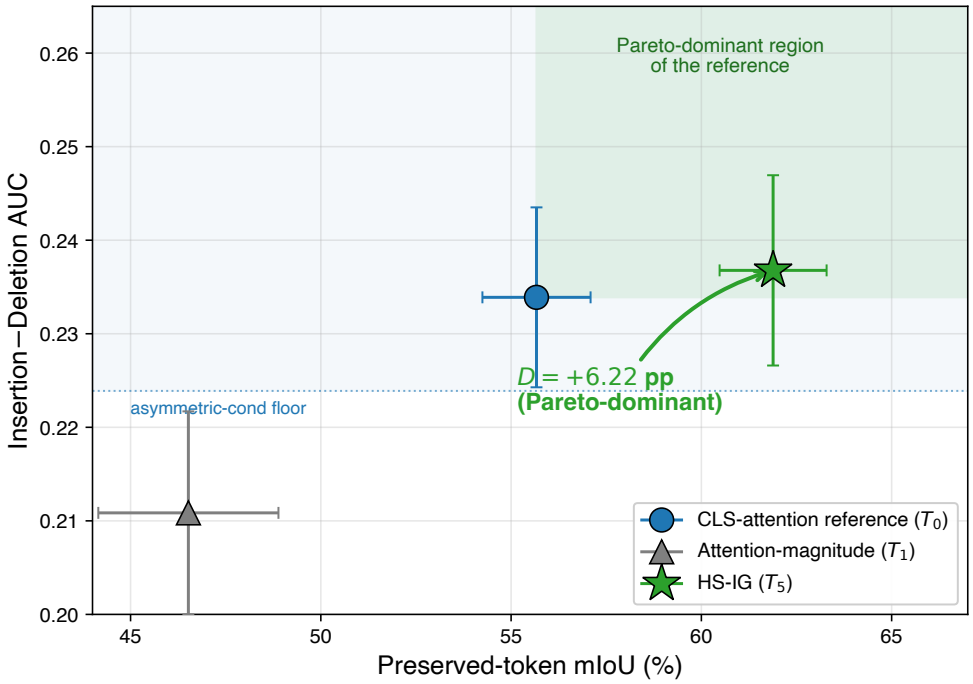


Fig. 1 Primary cell (FGVC-Aircraft, $L = 6$, $b = 25\%$, eight seeds). The hidden-state input-times-gradient scorer (T_5, HS-IG) Pareto-dominates the CLS-attention reference (T_0) on the sample means of both preserved-token mIoU *and* Insertion-Deletion AUC, with $D = +6.22$ pp at Δ IDAUC $= +0.003$; the eight-seed error bars are wide, so the dominance is on the means rather than statistically certified. Attention-magnitude (T_1) collapses below both. The shaded green wedge is the region in which T_5 Pareto-dominates the CLS-attention reference. The shaded blue band is the asymmetric-cond region in which D is well-defined (Section 4.4).

on a larger AugReg ViT-B/16, a comparison that changes capacity, pretraining and resolution together, so we bound it to small backbones rather than to scale.

Keywords: Vision Transformer, Token pruning, Part localization, Faithfulness evaluation, Attribution metrics, Fine-grained visual classification

1 Introduction

Fine-grained visual classification (FGVC) requires distinguishing object categories that differ only in subtle local part features, with bird species separated by beak curvature and plumage markers, car models by grille and headlight patterns, and aircraft variants by wing and engine geometry [1–4]. When the discriminative signal is concentrated in parts, the choice of which image regions a Vision Transformer attends to becomes both an efficiency control and an explanation artefact. Token pruning has emerged as the standard mechanism for both purposes on FGVC backbones [5–8]. The

FGVC literature evaluates this dual role through two metrics that share the intent of measuring whether the retained tokens are the right ones. One metric is preserved-token mean-Intersection-over-Union (mIoU) against ground-truth part annotations; the other is Insertion-Deletion AUC (IDAUC) against the unpruned classifier's predictions [9, 10]. The two measure different quantities, namely overlap with human part annotations versus fidelity to the classifier's decision. Prior work develops scorers with one criterion in view without measuring the other, so whether they agree is untested. We show that they decouple: the scorer that maximises part-coverage mIoU is not the one CLS-attention implies, while faithfulness changes little.

Existing approaches to scorer design in FGVC partition into three classes. The first uses **CLS-attention magnitude** as the per-token importance score, exploiting the classification token's attention pattern to produce a per-patch scalar by marginalising over heads [5, 11, 12]. The second uses **attention-gradient combinations** that re-weight the attention map by its contribution to the classifier logit. Chefer et al. [10] propose the canonical formulation $A_{\text{cls}} \cdot \partial y_c / \partial A_{\text{cls}}$ and show it outperforms attention magnitude on per-image attribution faithfulness; earlier attention-rollout methods [13] aggregate attention recursively across layers. The third class uses **learned-MLP scorers** trained end-to-end through a Gumbel-softmax or halting gate [6, 14–17], with DynamicViT [6] as the canonical design. FGVC-specific extensions [7, 8, 18, 19] add multi-scale or part-selection priors. A related branch replaces pruning with token merging [20, 21], a choice orthogonal to the supervision question studied here. In practice these two criteria live in separate literatures. Token-pruning scorers are designed and evaluated for classification accuracy under a budget, and their part-coverage against ground-truth annotations is not measured; attribution faithfulness is studied separately in the interpretability literature. No public study measures both criteria for the same token-pruning scorer, so whether the two agree on which scorer is best is an open question. We ask *where, by how much, and for which supervision signal* the two disagree.

We measure both criteria for the same scorer, and they decouple. On small ViT backbones, scorers trained under hidden-state input-times-gradient supervision produce systematically higher preserved-token mIoU than CLS-attention. The mIoU gap rises to +6.22 pp on FGVC-Aircraft while Insertion-Deletion AUC remains essentially unchanged, and the cell enters the Pareto-dominant region of the comparison plane (Figure 1). The two criteria, in other words, do not measure the same quantity, and which scorer is best depends on which one the practitioner asks. The phenomenon leaves the field with four reporting problems. A mIoU gain at IDAUC parity has to be expressed as a single number, and parameter-count confounds have to be ruled out in any scorer that adds a learnable head. The supervision-target design space then has to be enumerated so that prior choices can be compared on one axis, and the resulting measurements have to be turned into a deployment-time recommendation.

To address these problems, we propose **MIDAS** (mIoU-IDAUC Disagreement Across Supervision-targets), a measurement framework that *quantifies* the decoupling through a decoupling-magnitude scalar D and *catalogues* the supervision-target design space through a taxonomy spanning targets T_0 through T_5. The framework is instantiated under a token-selection protocol whose unsupervised-scorer falsifier rules out

the parameter-count confound, and it produces a (D, IDAUC) readout per supervision target that translates directly into a deployment-objective recommendation. The taxonomy catalogues; the scalar quantifies. Together they expose the metric disagreement on a single comparison plane and turn a previously unmeasured aspect of scorer-supervision choice into a measurable and actionable axis.

Our experiments establish three findings about where the decoupling holds and what governs its size. First, the phenomenon holds on all three FGVC datasets: T_5 lifts mIoU substantially (+2.11, +6.22, and +1.49 pp on CUB-200-2011, FGVC-Aircraft, and Stanford Cars) while the Insertion-Deletion AUC difference remains within 0.02 in absolute value, so the two metrics disagree on the best scorer. On FGVC-Aircraft the mean IDAUC also rises, placing the cell in the Pareto-dominant region across eight seeds. At eight seeds, however, the seed variance is too wide to certify strict statistical parity. Second, the decoupling magnitude depends jointly on scorer-block depth L and retention budget b. D peaks at the shallowest depth tested ($L = 3$, $D = +6.15$ pp on CUB) and at aggressive budgets ($b \leq 25\%$). At deeper readouts $L \in \{9, 11\}$ the supervision signal becomes ineffective, and at looser budgets $b \geq 50\%$ the inter-target differences collapse to noise. Third, a per-block cosine probe shows that the hidden-state-gradient and attention-gradient signal families produce mechanically distinct rankings (cosine ≤ 0.082 across all 12 transformer blocks), so the mIoU separation is not a labelling artefact.

A plausible mechanism. Prior analyses report that ViT attention patterns evolve substantially across blocks. Shallow-to-intermediate layers carry more localised structure that is progressively reorganised in deeper blocks [22, 23]. Under this account, hidden-state input-times-gradient inherits a part-localised signal from intermediate readouts that CLS-attention no longer carries, having marginalised across tokens. The account predicts both the depth dependence (decoupling concentrates at intermediate L) and the distinct per-block cosine rankings between signal families. Our data are consistent with the account rather than proving it.

We contribute a phenomenon, an instrument for measuring it, and the supervision signal that drives it.

- **We show that part coverage and faithfulness decouple.** Preserved-token mIoU and Insertion-Deletion AUC disagree on which token-pruning scorer is best, so a scorer chosen under one criterion carries no guarantee under the other. We establish this on small ViT backbones and bound it: it reproduces on DeiT-Ti/16, which varies capacity alone (+5.43 and +3.71 pp on FGVC-Aircraft and Stanford Cars), and reverses on a larger AugReg ViT-B/16 that differs on several axes at once (Appendix C, $D = -1.45$ pp).
- **We introduce MIDAS, an instrument for measuring it.** Our supervision taxonomy closes the design space to six targets. T_0 (cls-attn-proxy), T_1 (attn-magnitude) and T_3 (chefer) are the choices the token-pruning literature has already published; T_2 (loss-grad), T_4 (hs-attn-grad) and T_5 (hs-input-grad) are their closest unrepresented alternatives. We pair the taxonomy with a decoupling-magnitude scalar D that quantifies the signed mIoU gap between a target and the CLS-attention reference at parity faithfulness, admitting cells that improve on both metrics while rejecting those that raise mIoU only by sacrificing faithfulness.

- **We identify the supervision signal that drives it.** The hidden-state input-times-gradient target T_5 is the maximum-D choice across the grid, and its per-block ranking is mechanically distinct from attention-gradient signals. We recommend it as the concrete recipe when part coverage is the deployment objective.

2 Related work

Token selection and scorer design in Vision Transformers. Token-pruning methods remove low-importance tokens from the transformer's forward pass to reduce inference cost. EViT [5] prunes by CLS-attention magnitude with a per-block top-K gate. DynamicViT [6] learns an MLP scorer end-to-end via a Gumbel-softmax. Adaptive token sampling [11] adapts the per-image budget through learned thresholds. A-ViT [14] and AdaViT [15] learn instance-conditioned halting probabilities or per-block usage policies. SPViT [16] introduces soft pruning with a package token that preserves information from discarded tokens, and EvoViT [12] routes informative and uninformative tokens through fast and slow computation paths. Token Dropping [17] extends the same idea to BERT pretraining. A separate branch replaces pruning with token merging: ToMe [20] fuses similar tokens via bipartite matching, and DiffRate [21] learns per-layer compression rates jointly with pruning and merging. We treat the choice between pruning and merging as orthogonal to the supervision-target question studied here. Across the cited works, the scorer-supervision recipe is implicit and varies by paper, and published evaluations measure either accuracy at a fixed budget or per-image attribution faithfulness in isolation, without a single comparison axis.

FGVC-specific methods add part-selection and cross-attention priors to the general token-selection machinery. TransFG [7] aggregates raw attention weights across all transformer layers via a Part Selection Module to identify discriminative tokens. FFVT [8] selects tokens through mutual attention weight statistics across heads with feature fusion. RAMS-Trans [18] aggregates attention across layers recurrently and re-feeds zoomed discriminative regions, and DCAL [19] learns dual cross-attention between global and local tokens. All four are attention-based selection mechanisms rather than learned-MLP scorers, and all four use raw or layer-aggregated attention magnitude as the implicit supervision target, placing them at T_0 or close to it on our taxonomy (Section 4.3). None studies the consequence of this implicit supervision choice on per-image part-coverage. Earlier CNN-era FGVC methods, including Bilinear CNNs [24], WS-DAN [25], PMG [26], Cross-X [27], and context-aware attentional pooling [28], establish that discriminative-region selection is a load-bearing problem in FGVC. Our scorer-supervision question inherits that challenge. MetaFormer [29] and HERBS [30] provide recent FGVC-tailored ViT baselines. Counterfactual attention learning [31] introduces a counterfactual training objective: a randomised attention map produces a counterfactual prediction whose loss penalises the model when the real attention is not load-bearing. We evaluate it as a candidate rescue for late-block scorer placement (Section 4.6). FGVC surveys [4, 32] catalogue the broader space without studying the metric-decoupling question we raise. All of these methods build on the Transformer [33], ViT [34], and DeiT [35]. The standard FGVC token-pruning backbone is ViT-B/16 across TransFG, FFVT, RAMS-Trans, and DCAL. We use DeiT-S/16 to match

a reproducible single-GPU compute envelope ($\sim$22 M parameters versus ViT-B/16's $\sim$86 M) while preserving ViT-family architectural alignment with the prior literature. The cluster-scale negative result on AugReg ViT-B/16 [36] in Section 5 bounds the paper's scope to the small-backbone regime and enumerates the confounds that a controlled scale-up sweep would need to address.

Attribution faithfulness in transformer interpretability. Abnar and Zuidema [13] introduce attention rollout and attention flow, two recursive aggregations that produce token-level attribution more faithful than raw attention weights. The Insertion-Deletion AUC family [9] measures how faithfully a per-token ranking explains the model's decision by perturbing the input in ranking order. Chefer et al. [10] propose per-layer attention-gradient products that improve attribution faithfulness over raw CLS-attention; Jain and Wallace [37] flag that raw attention is not by itself an explanation. Sundararajan et al. [38] formalise the input-times-gradient family with the Integrated Gradients axiomatic attribution method, the closest theoretical kin of our T_5 target (see Section 4.3). Vilas et al. [22] and Walmer et al. [23] analyse how class identity and attention patterns evolve across ViT depths; their findings bear on the depth dependence we report in Section 4.6. We treat the Insertion-Deletion AUC as a faithfulness metric in its own right and study how it relates to part-coverage mIoU rather than as an alternative to either.

Metric decoupling in deep-learning evaluation. Decoupling of two ostensibly aligned metrics has been reported in attention faithfulness [37] and shortcut-learning analyses [39], where deep networks systematically rely on more available rather than more predictive features. Within token pruning the question has not been asked. Existing methods choose a single metric and optimise it. We instead measure how the two metrics relate.

3 The decoupling phenomenon

This section states the empirical phenomenon that motivates the rest of the paper. The phenomenon is presented first at a qualitative level, then in terms of what it shows about the two criteria, and finally in terms of the practical and methodological problems it raises for the FGVC token-pruning literature. Section 4 introduces the measurement framework that quantifies the phenomenon, and the experimental section reports the primary cells.

3.1 Phenomenon statement

In a Vision Transformer trained on a fine-grained classification task, the token-selection criterion that maximises preserved-token mIoU against ground-truth part annotations is not the same criterion that maximises Insertion-Deletion AUC against the unpruned classifier's predictions. Concretely, scorers supervised by the hidden-state input-times-gradient signal $h_i \cdot \partial y_c / \partial h_i$ on an intermediate Vision Transformer block produce token-selection masks that align more tightly with the dataset's ground-truth part keypoints than CLS-attention masks at the same retention budget. At the same time, they preserve the unpruned classifier's per-image decisions to within the seed-noise floor on

Insertion-Deletion AUC. The maximum gap we observe is +6.22 pp mIoU on FGVC-Aircraft at scorer depth $L = 6$ and retention budget $b = 25\%$ across eight seeds, with the cell sitting inside the Pareto-dominant region of the (mIoU, IDAUC) comparison plane. Across CUB-200-2011 and Stanford Cars the same supervision target produces mIoU gaps of +2.11 pp and +1.49 pp respectively at the same scorer depth and budget. The phenomenon is per-seed robust on Aircraft and CUB: over all eight seeds the worst-case T_5 seed (60.33%) exceeds the best-case CLS-attention reference seed (57.65%) by +2.68 pp (the three seeds tabulated here give +4.34 pp).

3.2 What the phenomenon shows

The phenomenon shows that a scorer's part-coverage and its faithfulness are not interchangeable: high preserved-token mIoU does not imply high Insertion-Deletion AUC, nor the reverse. Were the two interchangeable, reporting only one of them would convey complete information about token-selection quality, and a leaderboard ordered by either would be a faithful ranking of scorers. Neither reading holds. A scorer can improve mIoU by 6.22 pp while changing Insertion-Deletion AUC only slightly. The inverse direction is also visible at deeper readouts $L \in \{9, 11\}$, where the same target is lower on mIoU while remaining within the Insertion-Deletion AUC band. The two metrics measure different quantities, and which scorer is best depends on which metric the practitioner asks.

3.3 Problems the phenomenon introduces

The decoupling phenomenon raises four problems that the rest of the paper addresses. First, at the methodological level, any single-metric evaluation in the FGVC token-pruning literature now omits load-bearing information about the alternative metric. Cross-paper comparisons that mix mIoU-reporting and IDAUC-reporting scorers are therefore unreliable. Second, at the experimental level, two scorers compared at a late readout depth on a comparable-depth backbone may produce identical mIoU regardless of supervision content. The choice of scorer-block depth L consequently becomes an axis that experiments must report. Third, at the deployment level, scorer selection now requires the practitioner to declare which of the two metrics matches the downstream objective, since no single scorer maximises both across the regimes we test. Fourth, at the credibility level, any new scorer that adds learnable parameters must be checked against an unsupervised-scorer falsifier to rule out the parameter-count alternative for its measured gain.

3.4 Experimental evidence for the phenomenon

This subsection reports three experimental results that establish the phenomenon empirically. The full quantitative apparatus, namely the six-target supervision taxonomy and the decoupling-magnitude scalar D, is deferred to Section 4 and the framework's own experiments. Here we report evidence that the phenomenon exists, that it is not an artefact of the ground-truth mask labelling, and that it spans the FGVC datasets we evaluate.

Setup. We use DeiT-S/16 [35] pretrained on ImageNet-1k as the backbone for all reported cells. Three FGVC benchmarks are evaluated: CUB-200-2011 [1], Stanford Cars [2], and FGVC-Aircraft [3]. Eight random seeds $\{0, \ldots, 7\}$ are used for the nine primary cross-dataset cells; three seeds $\{0, 1, 2\}$ are used for ablation cells. Mean and standard deviation are reported. The unpruned DeiT-S/16 classifier is fine-tuned to standard small-backbone FGVC accuracy on each dataset before scorers are trained. Token pruning preserves classifier accuracy to within the seed-noise floor at the intermediate-depth cells reported in Section 4.6, so all decoupling comparisons below are at parity classification accuracy as well as parity Insertion-Deletion AUC. The two evaluation metrics for token-rankings are: preserved-token mIoU against ground-truth part keypoints (mask radius $r = 7$ patches on the 14×14 grid, following the keypoint-radius convention used in part-localization studies on CUB) and the Insertion-Deletion AUC composite $\text{ins AUC} - \text{del AUC}$ on the same checkpoints. The supervised-scorer cells use the hidden-state input-times-gradient signal as the supervision target on the post-block-L hidden state, with $L = 6$ and retention budget $b = 25\%$ unless otherwise stated.

Cross-dataset evidence. As shown in Table 1, the hidden-state input-times-gradient scorer (HS-IG) outperforms the CLS-attention reference on mIoU across all three FGVC datasets while remaining within the seed-noise floor on Insertion-Deletion AUC on two of the three. On FGVC-Aircraft, mean mIoU rises from 55.67% to 61.88% (+6.22 pp) at $\Delta\,\text{IDAUC} = +0.003$, with the cell sitting inside the Pareto-dominant region of the comparison plane. On CUB-200-2011, mean mIoU rises from 34.62% to 36.73% (+2.11 pp) at $\Delta\,\text{IDAUC} = +0.002$. On Stanford Cars, mean mIoU rises from 33.94% to 35.43% (+1.49 pp) at $\Delta\,\text{IDAUC} = -0.011$; this IDAUC value is outside the seed-noise floor, so the Stanford Cars cell supports the phenomenon on a sign-only reading rather than a strict parity reading. The attention-magnitude target (Attn-Mag, denoted T_1 in the taxonomy in Section 4.3) collapses on every dataset: by −5.28 pp on CUB, by −9.15 pp on Aircraft, and by −9.58 pp on Cars. The collapse rules out the alternative reading that the mIoU gap reflects a generic non-CLS-attention effect rather than a hidden-state-gradient effect. The cross-dataset evidence thus forms a three-way pattern, with HS-IG winning, CLS-attention as the reference, and Attn-Mag below the reference on every dataset. Figure 2 shows what this difference looks like at the level of individual token selections on CUB-200-2011 and FGVC-Aircraft. The two selections still overlap heavily, so a small and directed token swap carries the whole part-coverage difference reported in Table 1. Figure 3 shows the Insertion-Deletion curves that stay flat while it happens.

Mechanism probe. To verify the phenomenon is not a relabelling artefact, we compute the per-block cosine similarity between the hidden-state input-times-gradient signal and the attention-weighted gradient signal across all 12 transformer blocks of the trained DeiT-S/16. The probe finds cosine similarity ≤ 0.082 at every block, confirming that the two signal families produce mechanically distinct rankings. The phenomenon therefore reflects a real difference in token-selection criteria, not a near-trivial reparameterization of the same ranking.

Phenomenon regime (qualitative summary). The phenomenon is regime-specific. We sweep four scorer-block depths $L \in \{3, 6, 9, 11\}$ and four retention budgets

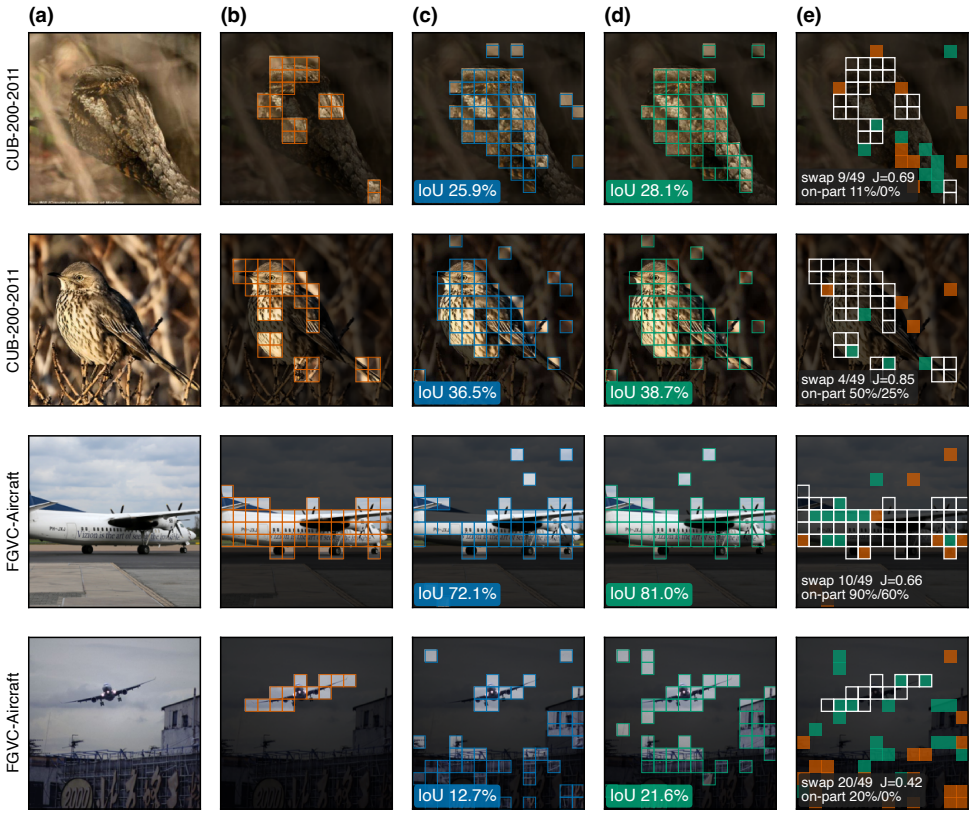


Fig. 2 Token selections on four typical test images, two from CUB-200-2011 and two from FGVC-Aircraft, each chosen so that its part-coverage gap matches its dataset's mean rather than showing a best case. **(a)** input; **(b)** ground-truth parts, from keypoint unions on CUB and Grounded-SAM masks on Aircraft (Appendix B); **(c)** the 49 tokens the CLS-attention reference T_0 keeps at $b = 25\%$; **(d)** the 49 kept by the hidden-state input-times-gradient scorer T_5; **(e)** what T_5 adds (green) and drops (orange), ground truth outlined in white, labelled with the Jaccard overlap J and the share of added and dropped tokens landing on parts. In every row T_5 adds more on-part tokens than it drops, including the bottom row, where both scorers largely miss the aircraft. Panel values are per-image readings at seed 0, not the eight-seed aggregates reported in the tables; the images are reproduced from the two benchmarks [1, 3] and remain the property of their photographers.

$b \in \{15, 25, 50, 70\}\%$. The mIoU gap is positive at intermediate depths $L \in \{3, 6\}$ and aggressive budgets $b \leq 25\%$; it collapses to noise or turns negative at deeper readouts $L \in \{9, 11\}$ or looser budgets $b \geq 50\%$. The full quantification of this regime under the decoupling-magnitude scalar D, including the $L = 3$ Pareto cell with mIoU gap $+6.15$ pp on CUB, is reported in Section 4.6 within the framework's experiments.

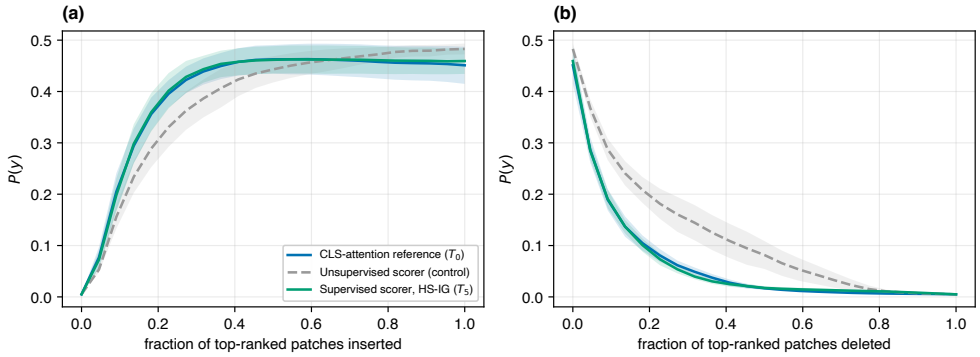


Fig. 3 Insertion and deletion curves on CUB-200-2011 at $b = 25\%$, shown as mean ± 1 standard deviation over seeds. **(a)** insertion, **(b)** deletion. The CLS-attention reference and the hidden-state input-times-gradient scorer are indistinguishable across the whole curve, at Insertion-Deletion AUC 0.342 ± 0.023 and 0.344 ± 0.023 over eight seeds, even though the same two conditions differ by $+2.11$ pp in part coverage (Table 1). The unsupervised-scorer control, at 0.271 ± 0.031 over three seeds, separates clearly from both, so the flat comparison between the other two is not an insensitive measurement. Seed counts differ because the control was not extended past three seeds.

Table 1 Cross-dataset evidence for the phenomenon at $(L, b) = (6, 25\%)$, eight seeds. The hidden-state input-times-gradient scorer (HS-IG) is compared against CLS-attention and against the attention-magnitude (Attn-Mag) target that serves as a non-CLS-attention falsifier.

Dataset	Scorer	mIoU $\pm$ std (%)	IDAUC $\pm$ std	ΔmIoU
Aircraft	CLS-attn	55.67 ± 1.42	$+0.234 \pm 0.010$	0.00
Aircraft	Attn-Mag	46.52 ± 2.37	$+0.211 \pm 0.011$	-9.15
Aircraft	HS-IG	$\mathbf{61.88 \pm 1.41}$	$+0.237 \pm 0.010$	$+6.22$
CUB	CLS-attn	34.62 ± 0.25	$+0.342 \pm 0.023$	0.00
CUB	Attn-Mag	29.34 ± 1.51	$+0.330 \pm 0.015$	-5.28
CUB	HS-IG	$\mathbf{36.73 \pm 0.69}$	$+0.344 \pm 0.023$	$+2.11$
Cars	CLS-attn	33.94 ± 0.68	$+0.277 \pm 0.018$	0.00
Cars	Attn-Mag	24.36 ± 0.77	$+0.242 \pm 0.013$	-9.58
Cars	HS-IG	$\mathbf{35.43 \pm 1.18}$	$+0.266 \pm 0.017$	$+1.49$

4 MIDAS: a framework for quantifying and cataloguing the phenomenon

The evidence in Section 3.4 rests on a single scorer, the hidden-state input-times-gradient target, compared against the CLS-attention reference, with raw mIoU and IDAUC numbers and no framework-level quantification. Turning that into a measurable claim, one that places the scorer within the design space of prior choices and answers the credibility and deployment problems raised in Section 3.3, is what the framework below is for.

This section introduces MIDAS, a measurement framework for the phenomenon stated in Section 3. The framework has two major components that carry its primary scientific content. The first is a six-target supervision taxonomy that catalogues the design space. The second is a decoupling-magnitude scalar that quantifies the signed mIoU gap at parity faithfulness. Two supporting components operationalise these majors. The first is a token-selection protocol with an unsupervised-scorer falsifier that defines the experimental container and rules out the parameter-count confound. The second is a practitioner-facing recipe that translates a (D, IDAUC) readout into a deployment-objective recommendation. Together the four pieces form a single named framework for comparing scorer-supervision choices on a common axis.

4.1 Framework overview

The two majors and the two supports relate as follows. The supervision-target taxonomy defines a finite axis of supervision choices. The D scalar measures one cell on this axis against the CLS-attention reference. The token-selection protocol provides the comparison structure that makes the measurement credible. The supervision-target selection converts a measurement into a deployment recommendation.

A reader who wants to use MIDAS to study a new supervision target needs three components. These are the taxonomy (to place the target within the design space), the D scalar (to measure its decoupling magnitude), and the three-condition protocol (to rule out the parameter-count alternative). A reader who also wants to deploy a token-pruning system needs the supervision-target selection. Section 4.2 defines the three-condition protocol. Section 4.3 defines the six-target taxonomy. Section 4.4 defines the decoupling-magnitude scalar D. Section 4.5 defines the supervision-target selection. Section 4.6 reports the framework's own experiments: the full taxonomy comparison, the unsupervised-scorer falsifier, the depth and budget regime under D, and a stress test at the late-block boundary.

4.2 Token-selection setup

The scorer-supervision comparison is structured around three named conditions that isolate each contribution from confounds, summarised in Table 2. **The CLS-attention reference** uses CLS-attention at the model's classification head as the per-token importance score and serves as the comparison reference for every other condition. **The unsupervised scorer** replaces CLS-attention with a learnable per-token scorer head trained under the classification loss only, with no auxiliary supervision. The unsupervised scorer isolates supervised scorer training from any benefit that comes from the additional parameters alone. A positive result on the unsupervised scorer over the CLS-attention reference would indicate that the comparison axis is parameter count rather than supervision content. **The supervised scorer** uses the same scorer head as the unsupervised scorer but trains it under a two-loss recipe

$$\mathcal{L} \;=\; \mathcal{L}_{\text{cls}} \;+\; \lambda_{\text{ga}}\, \mathcal{L}_{\text{ga}}, \tag{1}$$

where $\mathcal{L}_{\text{ga}}$ is a Huber loss between the scorer head's output and a distillation target derived from one of the six supervision sources defined in Section 4.3, and $\lambda_{\text{ga}} = 2$ is held fixed across the experimental sweep.

The model architecture is fixed across all three arms. We use DeiT-S/16 [35] pretrained on ImageNet-1k, with patch size 16 and input resolution 224^2, giving $N = 196$ token positions plus one CLS token. The scorer head is a three-layer MLP that reads the post-block-L hidden state $h_i \in \mathbb{R}^{384}$ for each token position i and emits a scalar score $s_i \in \mathbb{R}$. At inference time the scorer scores all N tokens, and a top-K gate keeps the $K = \lceil bN \rceil$ highest-scoring tokens for the remaining transformer blocks. The two primary experimental axes are the supervision-target choice $t \in \{T_0, \ldots, T_5\}$ defined in Section 4.3 and the retention budget b. Two further axes, the scorer-block depth L and the dataset choice on three FGVC benchmarks, are studied as ablations and reported in the experimental section.

Training configuration. Every cell is trained on a single GPU with AdamW at a backbone learning rate of 5×10^{-5}, a scorer learning rate of 5×10^{-4}, and weight decay 0.05, for 15 epochs at batch size 4. We run eight random seeds $\{0, \ldots, 7\}$ for the nine primary cross-dataset cells and three random seeds $\{0, 1, 2\}$ for the ablations, because the primary cells carry the paper's headline comparison and the ablations map a regime rather than certify a gap. A full sweep takes a few hours per cell on one GPU, exclusive of evaluation. Training is deterministic given the seed, the dataset subsample, and the per-metric evaluation subset.

The three-condition protocol supports the two major components in two distinct ways. The CLS-attention reference serves as the comparison anchor against which the decoupling-magnitude scalar D is computed; without it the scalar has no anchor. The unsupervised scorer is the parameter-count falsifier: its collapse rules out the alternative explanation that any added MLP head outperforms the CLS-attention reference regardless of supervision. Without the unsupervised-scorer falsifier, the gain on T_5 would not be attributable to supervision content. The supervised scorer is the condition under which the taxonomy is instantiated. All six supervision targets in Section 4.3 are evaluated under the supervised scorer, with the CLS-attention reference as the anchor and the unsupervised scorer as the falsifier.

Scope of comparison. The comparison is deliberately *within* a single fixed pipeline: the same DeiT-S/16 backbone, gate, retention budget, and training schedule, varying only the supervision target. We do not reproduce published token-pruning methods (EViT, DynamicViT, ToMe, DiffRate) end-to-end, and we make no state-of-the-art claim. This is a design choice, not an omission. Comparing full published pipelines would reintroduce the architecture, budget, and training confounds that the three-arm protocol removes, and would answer a different question, namely which *system* is best, rather than the one we ask, which *supervision signal* yields part-aligned versus faithfulness-aligned selections. Absolute accuracy is correspondingly not the axis of interest: the controlled small-backbone, reduced-training regime we adopt for reproducibility is well below full-scale FGVC systems, and every comparison is read as a within-pipeline contrast rather than a leaderboard entry.

Table 2 The three experimental conditions. Every reported cell in this paper is one of these conditions. The CLS-attention reference is the training-free baseline against which the decoupling magnitude D is measured; the unsupervised scorer is the parameter-count falsifier; the supervised scorer instantiates the six supervision targets T_0–T_5 of Section 4.3. The primary comparison is the supervised scorer under target T_5 (hs-input-grad) against the CLS-attention reference.

Condition	Scorer head	Supervision
CLS-attention reference	CLS-attention, no learned head	none (training-free)
Unsupervised scorer	learned MLP	classification loss only
Supervised scorer	learned MLP	classification loss + target $t \in \{T_0, \ldots, T_5\}$

4.3 Cataloguing the design space

The first major component of MIDAS is a taxonomy that catalogues the supervision-target design space into six closed-form choices. The taxonomy organises the design space along two structural axes that the FGVC token-pruning literature has previously varied implicitly, namely *attention map versus hidden state* as the per-token feature source and *raw versus gradient-modulated* as the attribution operator. The six targets are:

$$
\begin{aligned}
T_0 \text{ (cls-attn-proxy)}: s_i^{(0)} &= A_{\text{cls},i}, \\
T_1 \text{ (attn-magnitude)}: s_i^{(1)} &= \textstyle\sum_j |A_{j,i}|, \\
T_2 \text{ (loss-grad)}: s_i^{(2)} &= |A_{\text{cls},i}| \cdot \left| \tfrac{\partial \mathcal{L}_{\text{cls}}}{\partial A_{\text{cls},i}} \right|, \\
T_3 \text{ (chefer)}: s_i^{(3)} &= A_{\text{cls},i} \cdot \tfrac{\partial y_c}{\partial A_{\text{cls},i}}, \\
T_4 \text{ (hs-attn-grad)}: s_i^{(4)} &= A_{\text{cls},i} \cdot \left\| \tfrac{\partial y_c}{\partial h_i} \right\|, \\
T_5 \text{ (hs-input-grad, ours)}: s_i^{(5)} &= h_i \cdot \tfrac{\partial y_c}{\partial h_i}.
\end{aligned}
\tag{2}
$$

The taxonomy has three properties that justify it as a contribution in its own right. First, every *explicit* scorer-supervision target published in the FGVC token-pruning and ViT-attribution literature corresponds to one of T_0, T_1, or T_3. These cover raw CLS-attention as used by EViT [5] and attention-aggregation methods like TransFG [7]. Attention-magnitude variants used by ATS [11] and SPViT [16], and the Chefer attention-gradient formulation [10], are also captured. Methods with no explicit supervision target, such as DynamicViT [6], are captured by the unsupervised-scorer falsifier rather than by the taxonomy itself. Second, T_2, T_4, and T_5 are the three closest unrepresented alternatives. T_2 combines attention with the loss gradient, varying the gradient denominator from logit to loss. T_4 is the hidden-state attention-weighted gradient that mixes the two feature sources. T_5 is the hidden-state input-times-gradient target that instantiates the input-times-gradient family of Sundararajan et al. [38] on the residual stream. Third, the grid produces a single comparison plane on which prior scorer choices and the three new alternatives can be evaluated side-by-side. The supervision-target effect can therefore be isolated from architecture, budget, and depth confounds within a single experimental design.

The taxonomy also makes the structural axes load-bearing rather than incidental. The contrast between T_0, T_1, and T_3 on the attention-only side and T_4, T_5 on the hidden-state side isolates the feature-source axis. The contrast between T_0 and T_3 at fixed attention-only feature source isolates the gradient-modulation axis. The contrast between T_3 and T_5 at fixed gradient modulation isolates the feature-source axis on the gradient-modulated side. Each contrast is directly testable and yields a specific claim about which axis carries the decoupling. The taxonomy therefore encodes a set of designed contrasts, each isolating one axis.

4.4 Quantifying the phenomenon

The second major component of MIDAS is a scalar that quantifies the signed mIoU gap between a supervision target and the CLS-attention reference at matched faithfulness. We define the decoupling-magnitude scalar

$$\begin{aligned} D(t) \;=\; & \mathrm{mIoU}_t - \mathrm{mIoU}_{\mathrm{ref}} \\ & \text{conditional on} \quad \mathrm{IDAUC}_t - \mathrm{IDAUC}_{\mathrm{ref}} \;\geq\; -0.01, \end{aligned} \tag{3}$$

where $t \in \{T_0, \ldots, T_5\}$ is the supervision target, mIoU_t is the preserved-token mIoU against the dataset's ground-truth part masks at retention budget b averaged over three seeds, IDAUC_t is the Insertion-Deletion AUC composite $\mathrm{ins\,AUC} - \mathrm{del\,AUC}$ on the same checkpoints averaged over three seeds, and the conditional clause requires that the target's faithfulness not drop more than 0.01 below the CLS-attention reference's. The conditional is asymmetric. A target may gain IDAUC over the CLS-attention reference without disqualifying its D, but must not fall by more than 0.01. The asymmetry rules out corner-collapse modes, where a target raises mIoU_t by selecting tokens that align with the ground-truth mask but no longer preserve the classifier's decision. At the same time it admits Pareto-dominant cells, where a target raises both metrics simultaneously.

The asymmetric form replaces a symmetric variant $|\mathrm{IDAUC}_t - \mathrm{IDAUC}_{\mathrm{ref}}| \leq 0.01$ that we initially considered. The replacement is empirically motivated. At intermediate scorer depth $L = 3$, the hidden-state input-times-gradient target produces both higher mIoU and higher IDAUC than the CLS-attention reference simultaneously. The symmetric form would reject this Pareto-dominant point as outside the band, even though it is the strongest possible empirical evidence for the decoupling claim. The asymmetric form admits Pareto-dominant points while still rejecting corner-collapse modes, and is the form used in all numerical results below. The cutoff value 0.01 is set at the median per-seed standard deviation of the CLS-attention reference's IDAUC across the main-results cells, with a cell-wise spread of $\sigma \in [0.004, 0.019]$ and a median of $\sigma = 0.010$. The threshold therefore admits as parity-faithful any target whose IDAUC falls within one median reference seed-σ of the CLS-attention reference. The band is a measurement-noise floor, not a principled equivalence bound: it marks the region where an IDAUC difference is indistinguishable from seed noise, so D is a conditional readout, namely the mIoU gap in the regime where the faithfulness comparison is itself inconclusive. We therefore report the raw $\Delta\,\mathrm{IDAUC}$ alongside D rather than treating band membership as a parity guarantee. The Cars cell ($\Delta\,\mathrm{IDAUC} = -0.011$,

outside the band) and the T_2 CUB cell (Δ IDAUC $= +0.013$, also outside) are the two cells where the threshold matters. In both, the threshold separates a sign-positive mIoU gap from a strict-cond decoupling claim. We report both cells transparently as sign-only rather than redefining the threshold post hoc.

The scalar has two reading conventions that the paper uses to describe outcomes. **Decoupling regime** is the case $D > 0$ where the target produces tokens with higher part-coverage than the CLS-attention reference at parity faithfulness, and the paper's main claim concerns this regime. **Pareto regime** is the special case $D > 0$ AND $\text{IDAUC}_t > \text{IDAUC}_{\text{ref}}$, where the target dominates the CLS-attention reference on both metrics simultaneously. The paper reports the Pareto regime separately because it is a stricter readout of the same data and a stronger empirical claim. The threshold $D \geq 1$ pp is used throughout as the minimum reported decoupling magnitude. It is chosen to exceed 1.4 standard deviations on the smallest seed-variance cell observed across the experimental sweep.

4.5 Selecting the supervision target for deployment

The taxonomy and D together frame the supervision-target choice as a selection between two FGVC-deployment objectives. **Part-coverage objective.** An FGVC practitioner who needs preserved tokens to align with semantic part regions should pick the D-maximising target at the deployment depth and budget. Examples include a birdwatching application that explains its species prediction by highlighting the bird's beak, eye, and wing markers. Another example is an aviation-maintenance tool that surfaces the wing or engine variant the model used to identify the aircraft. On the three FGVC datasets we evaluate, the D-maximising target is T_5 on CUB-200-2011 and FGVC-Aircraft, and also T_5 on Stanford Cars at a smaller margin. **Faithfulness objective.** An FGVC practitioner who needs preserved tokens to maximally preserve the unpruned classifier's decision should pick the IDAUC-maximising target at the deployment depth. An example is a deployed efficient FGVC inference pipeline whose pruned-model output must match the unpruned reference under audit. At intermediate scorer depth $L = 3$ on CUB, the same target T_5 also maximises IDAUC, so the two objectives are jointly maximised in the Pareto regime. At deeper readouts the IDAUC differences are within noise and the choice is dominated by the part-coverage axis.

This framing makes the supervision-target choice a deliberate engineering decision rather than an unstated assumption inherited from whichever scorer the original implementation happened to use. The experimental section reports a (D, IDAUC) point per target on a single comparison plane at the deployment depth and budget. The practitioner can then read off the operating-point trade-off directly. The framing also clarifies the scope of the paper's claim. The supervision-target choice is the variable. The model architecture, the budget, and the depth are each *also* deployment choices, but they are held fixed within each comparison cell so that the supervision-axis effect is isolated.

Table 3 Six-target taxonomy on CUB-200-2011 at $L = 6$, $b = 25\%$. Eight seeds for the CLS-attention reference, T_1, and T_5; three seeds for T_2–T_4. The reference is the CLS-attention reference. The unsupervised-scorer falsifier reaches $D = -14.59$ pp.

Target	mIoU ± std (%)	IDAUC ± std	D (pp)	cond
CLS-attn (ref.)	34.62 ± 0.25	$+0.342 \pm 0.023$	0.00	✓
T_0	34.33 ± 0.81	$+0.340 \pm 0.010$	-0.23	✓
T_1	29.34 ± 1.51	$+0.330 \pm 0.015$	-5.28	×
T_2	34.96 ± 0.38	$+0.346 \pm 0.006$	$+0.40$	×
T_3	34.39 ± 0.79	$+0.340 \pm 0.019$	-0.17	✓
T_4	35.88 ± 1.04	$+0.336 \pm 0.015$	$+1.31$	✓
T_5 **(ours)**	$\mathbf{36.73 \pm 0.69}$	$+0.344 \pm 0.023$	$\mathbf{+2.11}$	✓
Unsup. scorer (falsif.)	19.97 ± 2.16	$+0.271 \pm 0.031$	-14.59	×

4.6 The framework applied to the FGVC benchmarks

This subsection reports the experimental results that exercise MIDAS's two major components: the supervision-target taxonomy and the decoupling-magnitude scalar D, together with the supporting unsupervised-scorer falsifier. The setup is identical to the phenomenon experiments reported in Section 3.4. The framework is exercised along four axes: the full taxonomy comparison on CUB at the primary cell, the unsupervised-scorer falsifier that rules out the parameter-count alternative, the depth and budget sweeps that map the D-positive regime, and a stress test at $L = 11$ that probes the boundary of the framework's intermediate-depth claim.

Cataloguing the design space. As shown in Table 3, three of the six taxonomy targets exceed the CLS-attention reference's mIoU at parity faithfulness on CUB at $(L, b) = (6, 25\%)$, namely $T_5 = +2.11$ pp, $T_4 = +1.31$ pp, and $T_2 = +0.40$ pp. The hidden-state input-times-gradient target T_5 achieves the largest decoupling magnitude with $D = +2.11$ pp at eight seeds and per-seed standard deviation of 0.69 pp, so D exceeds three seed standard deviations from zero. The two attention-magnitude targets T_0 and T_1 both produce $D \leq 0$, and the attention target modulated by the loss gradient, T_2, enters the decoupling regime nominally but its IDAUC drifts upward by $+0.013$, which puts it just outside the asymmetric-cond band. We read T_2's gain as inadmissible decoupling because the same target also gains IDAUC, so it is not a same-faithfulness comparison.

The parameter-count falsifier. The unsupervised-scorer falsifier reaches $D = -14.59$ pp on CUB with mIoU at $19.97 \pm 2.16\%$. This collapse rules out the alternative reading that any added MLP head over the CLS-attention reference improves mIoU at parity faithfulness. The supervised-scorer gain on T_5 is therefore attributable to the supervision content of the hidden-state input-times-gradient signal, not to the parameter count of the scorer head. Without the unsupervised-scorer falsifier the framework's quantification of T_5's gain would be confounded with parameter count. The protocol is therefore load-bearing for the framework's credibility, even though it is not a major component on its own.

Table 4 Depth ablation on CUB at $b = 25\%$, three seeds. D values quantify where the phenomenon lives along the readout-depth axis. The $L = 6$, $b = 25\%$ cell is at eight seeds; the remaining rows are at three. The *cond* column records Insertion-Deletion AUC band membership: ✓ inside the asymmetric-cond band, × outside it, and Pareto for the stronger case where IDAUC improves as well.

L	D (pp)	cond
3	+6.15	Pareto
6	+2.11	✓
9	−5.98	×
11	−26.13	×

Table 5 Retention-budget ablation on CUB at $L = 6$, three seeds. D values quantify where the phenomenon lives along the retention-budget axis. The $L = 6$, $b = 25\%$ cell is at eight seeds; the remaining rows are at three. The *cond* column records Insertion-Deletion AUC band membership: ✓ inside the asymmetric-cond band, × outside it, and Pareto for the stronger case where IDAUC improves as well.

b	D (pp)	cond
15%	+2.73	✓
25%	+2.11	✓
50%	−0.52	✓
70%	−0.58	✓

Quantifying the depth regime. As shown in Table 4, the depth ablation on CUB at $b = 25\%$ maps where D is positive. At $L = 3$, the T_5 target produces $D = +6.15$ pp as a Pareto cell (Δ IDAUC $= +0.038$). It produces $D = +2.11$ pp at $L = 6$ inside the asymmetric band, $D = -5.98$ pp at $L = 9$ outside the band, and $D = -26.13$ pp at $L = 11$ where the supervised scorer collapses to the bit-identical mIoU pattern discussed in the stress test below. The intermediate-depth window $L \in \{3, 6\}$ is the operationally effective regime for the decoupling phenomenon.

Quantifying the budget regime. The retention-budget ablation at fixed $L = 6$ on CUB, reported in Table 5, shows the same pattern of Pareto dominance and decoupling at intermediate cells along a different axis. At $b = 15\%$, the T_5 target produces $D = +2.73$ pp inside the asymmetric band. It produces $D = +2.11$ pp at $b = 25\%$ inside the band, $D = -0.52$ pp at $b = 50\%$ where inter-target differences collapse to within seed noise, and $D = -0.58$ pp at $b = 70\%$ where mIoU values are dominated by the oversized retention set. The aggressive-budget regime $b \leq 25\%$ is the operationally effective regime for the decoupling phenomenon along the budget axis. The depth-budget joint regime $L \in \{3, 6\}$ with $b \leq 25\%$ is the cell-region in which D is reliably positive.

Stress test at $L = 11$. A four-loss closure attempt at the late readout depth $L = 11$ probes the boundary of the framework's intermediate-depth claim. Six independently trained checkpoints (three under two-loss training and three under four-loss training, with visibly altered classifier weights) produce bit-identical preserved-token masks on every one of the 400 evaluation images. The pattern indicates that the late-block scorer ranking at $L = 11$ is determined by the pretrained backbone's positional priors. Fine-tuning shifts the underlying attention pattern by a margin too small to

perturb a top-49-of-196 patch ranking. The framework's intermediate-depth window $L \in \{3, 6\}$ is therefore the regime in which the supervision signal carries decoupling content. Any claim about the framework outside that window must address the late-block invariance directly. The stress test is itself a methodological finding: token-pruning experiments must report the scorer-block depth L alongside the budget b, because L determines whether the mIoU is sensitive to loss-driven changes at all.

Effect-size summary. The primary cross-dataset cells (CUB and FGVC-Aircraft) are run at eight seeds; the $L = 3$ depth-ablation cell and the unsupervised-scorer falsifier remain at three seeds and are reported as such. Cohen's d_z on the per-seed mIoU difference T_5 versus the CLS-attention reference is $d_z = 3.8$ on the CUB main cell ($L = 6$, $b = 25\%$) and $d_z = 2.8$ on the FGVC-Aircraft cross-dataset cell, both above the conventional very-large threshold of $d_z = 1.30$. The $L = 3$ Pareto cell gives $d_z = 4.76$ and the CLS-attention reference versus the unsupervised-scorer falsifier on CUB gives $d_z = 8.21$ (three seeds each). The Aircraft effect is the noisiest across seeds: its mIoU gap moved from the three-seed estimate of +7.22 pp ($d_z = 2.9$) to +6.09 pp ($d_z = 2.0$) at four seeds, then settled at +6.22 pp ($d_z = 2.8$) at eight seeds, so the magnitude is stable across the last four seeds while the Aircraft cell still carries the widest seed variance of the three datasets. A percentile bootstrap over the paired per-seed mIoU differences (10,000 resamples) gives 95% CIs of $[+1.74, +2.45]$ pp on CUB, $[+4.76, +7.58]$ pp on FGVC-Aircraft, and $[+0.85, +2.12]$ pp on Stanford Cars. All three intervals exclude zero, so the sign of the mIoU gap is robust, while the wider Aircraft interval quantifies the residual magnitude uncertainty rather than hiding it. At eight seeds the bootstrap resamples eight paired differences; it is less coarse than at four but still discrete, and we report it with the per-seed values in the appendix. We do not report frequentist p-values on the mIoU difference because at eight seeds the paired test still has limited power. The effect sizes and the per-seed values in the appendix are the appropriate small-sample summary.

The IDAUC differences are small but not certifiably equivalent. The phenomenon claims that T_5 lifts mIoU while IDAUC does not track it. The load-bearing quantity is therefore how small the IDAUC movement is relative to the mIoU movement. We quantify this with a paired two-one-sided-test (TOST) on the per-seed IDAUC differences, using the margin $\varepsilon = 0.01$ and $\alpha = 0.05$. TOST at this level requires the 90% CI for the mean difference to lie inside $(-\varepsilon, +\varepsilon)$. At eight seeds no cell meets the strict two-sided bound. On CUB the mean IDAUC difference is +0.002 with a 90% CI of $[-0.007, +0.011]$ (TOST $p = 0.07$). On FGVC-Aircraft it is +0.003 with a 90% CI of $[-0.007, +0.013]$ (TOST $p = 0.11$). On Stanford Cars it is −0.011 with a 90% CI of $[-0.019, -0.004]$ (TOST $p = 0.62$). Two facts follow. First, the IDAUC differences are small in absolute terms ($|\Delta| \leq 0.011$) and, on CUB and Aircraft, they favour T_5, so they do not offset the large mIoU gains. The two metrics genuinely disagree on the best scorer. Second, the seed variance is too wide even at eight seeds to certify strict statistical equivalence. We therefore do not claim a parity guarantee. The decoupling rests on the mIoU gaps, which are large and consistent ($d_z = 2.8$ to 3.8 on the two primary cells), against IDAUC movement an order of magnitude smaller than the mIoU movement. Stanford Cars is the boundary case where the IDAUC cost is material. The equivalence test is part of the released implementation.

5 Discussion

Sections 3 and 4 establish the decoupling and measure it. This section asks what produces it and what follows from it. We give a mechanistic account of why hidden-state gradients carry part-localised signal that CLS-attention no longer carries, substantiate that account at the representation level, and set out what the two metrics mean for a practitioner choosing a scorer. We close on the two negative results that bound the claim. This also returns to the four problems raised in Section 3.3: the taxonomy and D are what make cross-paper comparisons possible, the depth axis that experiments must report is quantified in Section 4.6, the deployment choice is made explicit in Section 4.5, and the parameter-count alternative is ruled out by the unsupervised-scorer condition.

5.1 A plausible mechanism for the decoupling

The taxonomy of Section 4.3 separates supervision targets into two structural families. The first family contains targets whose score uses the attention map alone or the attention map's gradient (T_0, T_1, T_3, and the closely related T_2). The second family contains targets whose score uses the hidden-state vector or its gradient (T_4 and T_5). The decoupling magnitude D in Section 4.6 is positive only for the second family, with $T_5 = +2.11$ pp and $T_4 = +1.31$ pp on CUB at $(L, b) = (6, 25\%)$. The per-block mechanism probe in Section 3.4 confirms that the two families produce mechanically distinct rankings (cosine $\leq$ 0.082 across all 12 transformer blocks). **Hidden-state input-times-gradient supervision unlocks rankings that pure-attention supervision does not.**

Three properties of the hidden-state input-times-gradient signal jointly support the decoupling. First, the hidden-state vector h_i retains *content* information about each token's patch, while $A_{\text{cls},i}$ retains only the *similarity* of that patch to the CLS query. The score $h_i \cdot \partial y_c / \partial h_i$ is therefore sensitive to which patches *carry* class-discriminative content, rather than which patches the CLS query happens to attend to under the current heads' weights. Second, the gradient $\partial y_c / \partial h_i$ is computed once per training image with respect to the ground-truth class logit. The supervision target thus *labels* discriminative patches rather than salient ones. Third, the multiplicative product $h_i \cdot \partial y_c / \partial h_i$ inherits the input-times-gradient saturation behaviour from the attribution literature [40]. The largest scores go to patches whose presence both contributes to and is necessary for the correct logit. The combination of *content sensitivity*, *label specificity*, and *necessity weighting* is what attention-only supervision cannot replicate.

The depth ablation in Section 4.6 sharpens this reading. At $L = 3$ the HS-IG target produces a Pareto cell with $D = +6.15$ pp at $\Delta\,\text{IDAUC} = +0.038$, the strongest empirical evidence the paper has for the decoupling claim. At $L = 9$ and $L = 11$ the decoupling magnitude is large and negative, with $L = 11$ collapsing to $D = -26.13$ pp at mIoU = 8.44%. The collapse is not a training failure. Classifier accuracy at $L = 11$ remains at $\sim 73\%$, on par with intermediate-depth cells. The network is learning to classify, but the scorer's top-49 patch selection no longer aligns with the keypoint-union ground truth. We attribute this to late-block attention being dominated by the pretrained backbone's positional priors. At $L = 11$ the h_i gradient signal provides too little inter-patch contrast for the scorer to discover keypoint-aligned rankings. The

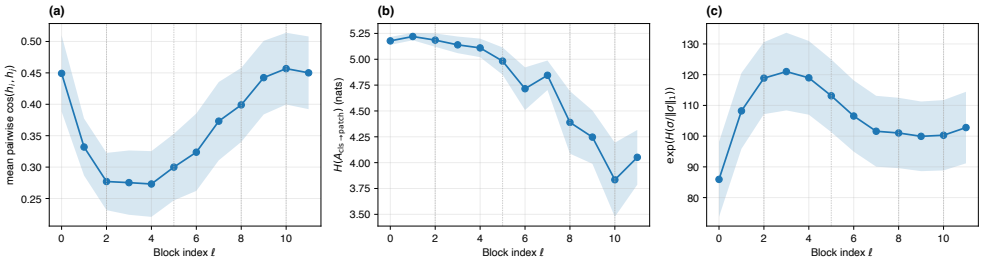


Fig. 4 Per-block geometric structure of a trained CLS-attention reference DeiT-S/16 on 500 CUB val images. (a) hidden-state anisotropy, **(b)** CLS-attention entropy, **(c)** hidden-state effective rank. Anisotropy minimises at $L = 3$–4 where decoupling is largest. CLS-attention entropy decays from $L = 2$ onward, the multi-head self-attention low-pass signature predicted by Park and Kim [41]. Effective rank peaks at $L = 3$. Dotted vertical lines mark the readout depths $L \in \{3, 6, 9, 11\}$ tested in Section 4.6. Bands are mean ± 1 standard deviation across images.

intermediate-depth window $L \in \{3, 6\}$ is therefore the regime in which the hidden-state-gradient mechanism is operationally effective.

5.2 Geometric mechanism analysis

Section 5 attributed the decoupling to three functional properties of the hidden-state input-times-gradient signal. We substantiate this account at the representation level by measuring three geometric properties of the trained CLS-attention reference DeiT-S/16 backbone on 500 CUB validation images. Per-block *representation anisotropy* is the mean pairwise cosine between the 196 patch-token hidden states. It measures how collapsed token representations are at each depth. Per-block *CLS-attention entropy* is the Shannon entropy of the head-averaged CLS-to-patch attention distribution after L_1-normalization. It measures how dispersed the CLS query's attention is. Per-block *effective rank* is the exponential entropy of the normalised singular-value distribution of the $196 \times d$ hidden-state matrix. It measures how many independent directions the per-token representations span.

Figure 4 reports the per-block means with ± 1 standard-deviation bands. All three measures point to intermediate depths $L \in \{2, 3, 4\}$ as the regime where patch-token representations are richest. Anisotropy reaches its minimum at $L = 4$ (0.273 ± 0.052) and at $L = 3$ (0.275 ± 0.051), rising to 0.450 ± 0.058 at $L = 11$. CLS-attention entropy declines from 5.220 nats at $L = 1$ near-uniform to 3.834 nats at $L = 10$ highly concentrated. This decline is in direct agreement with the multi-head self-attention low-pass-filter prediction of Park and Kim [41]. Effective rank peaks at $L = 3$ (121.0 ± 12.6) and drops to $\sim$100 at deep blocks.

The geometric picture is consistent with what the decoupling pattern already implied. At intermediate depths $L \in \{3, 6\}$ where D is reliably positive, the backbone retains spatially rich, high-rank, low-anisotropy hidden states. The CLS query has not yet collapsed its attention at those depths. The hidden-state gradient $\partial y_c / \partial h_i$ at those depths can pick up class-relevant per-token signal that the marginalised CLS-attention vector no longer carries. At deep readouts $L \in \{9, 11\}$ where D collapses, the same hidden states have lost both spatial distinguishability (anisotropy rises) and effective-rank

diversity. The CLS query has concentrated on a few tokens, leaving little inter-patch contrast for any supervision target to exploit. We report this as a geometric pattern correlated with depth-resolved decoupling, not as a causal proof. Controlled-ablation evidence linking each geometric quantity to the D value is left to future work.

5.3 Practical significance: two metrics for two user-facing goals

The disagreement we report reflects a difference between two operationally distinct downstream goals. Outside FGVC, the attribution-evaluation literature distinguishes *plausibility* (how well an attribution agrees with human-annotated rationales) from *faithfulness* (how well an attribution agrees with the model's decision). That literature shows that explainer rankings depend on which axis is used [42–44]. Preserved-token mIoU and Insertion-Deletion AUC inherit this distinction in the FGVC token-pruning setting. mIoU rewards retention of tokens that overlap human part annotations, a plausibility-style measure. IDAUC rewards retention of tokens whose removal damages the classifier's decision, a faithfulness-style measure. The 6.09 pp disagreement we report is therefore not a difference between two proxies for the same goal but a difference between two user-facing goals.

The decoupling phenomenon implies that the *correct* supervision target for a token-pruning deployment depends on which downstream goal the practitioner cares about. FGVC practitioners who prioritise *plausibility* should pick the D-maximising target. This covers, for example, birdwatching or aviation-recognition applications where the preserved tokens are presented to a human as the model's part-level evidence. On all three datasets we evaluate, that target is T_5, with $D = +2.11$ pp on CUB, $+6.22$ pp on Aircraft, and a marginal, directional-only $+1.49$ pp on Cars, which falls outside the parity-faithfulness band. FGVC practitioners who prioritise *faithfulness* should pick the IDAUC-maximising target. This covers deployed inference pipelines where preserved tokens feed back into the model's own subsequent decisions and faithfulness to the unpruned classifier is the audit metric. At intermediate depths and budgets the same target T_5 also maximises IDAUC in the Pareto regime ($L = 3$ on CUB, $L = 6$ on Aircraft), so the two objectives are jointly maximised. The same holds at deeper readouts, where the IDAUC differences across targets are within noise and the choice is dominated by the plausibility axis.

This framing recasts the FGVC supervision-target choice from an unstated implementation detail to a deliberate engineering decision. In the pre-existing FGVC token-pruning literature the choice is typically inherited. EViT [5] uses CLS-attention because the original transformer interpretability literature [37] reported it as a faithfulness signal. DynamicViT [6] uses an MLP scorer trained end-to-end without an explicit supervision target. Chefer et al. [10] argue for attention-gradient as the most faithful single-layer attribution but do not study its mIoU consequence. FGVC-specific extensions [7, 8, 18] add part-selection priors over learned-MLP scorers but inherit the underlying supervision choice from their generic ViT predecessors. The choice matters: the same DeiT-S/16 backbone, the same retention budget, and the same training recipe produce a ~ 5 pp swing in preserved-token mIoU on CUB depending only on which supervision target is used to train the scorer. Our taxonomy enables direct comparison of these choices on a single (D, IDAUC) plane. An FGVC practitioner can

therefore pick the supervision target that matches their deployment objective without re-implementing each prior method's training recipe.

5.4 Negative results and their implications

Two large-scale negative results from our experimental sweep deserve explicit discussion, namely the cluster-scale backbone scaling test and the four-loss closure attempt at $L = 11$.

Cluster-scale backbone scaling. To test whether the decoupling phenomenon extends to richer-backbone configurations, we trained the HS-IG target on AugReg ViT-B/16 [36] at input resolution 384^2. We used the full CUB train split with three seeds at $(L, b) = (6, 25\%)$. The cell produced $D = -1.45$ pp at parity IDAUC (both targets at ~ 0.60), with the HS-IG mIoU at $25.21 \pm 0.13\%$ and the CLS-attention reference's at $26.66 \pm 1.06\%$. The negative-D direction is per-seed robust: the worst-case CLS-attention reference seed (26.03%) still exceeds the best-case HS-IG seed (25.35%) by +0.68 pp. The decoupling phenomenon does *not* automatically extend to AugReg ViT-B/16 at 384^2. Several confounds make the direct cause hard to isolate, and they caution against reading the reversal as a pure scale effect. The cluster-scale cell differs from the DeiT-S/16 cells in five respects at once:

- **backbone capacity**, AugReg ViT-B/16 at $\sim$86 M parameters against DeiT-S/16 at $\sim$22 M;
- **pretraining regime**, ImageNet-21k with heavy AugReg augmentation against ImageNet-1k with DeiT distillation;
- **input resolution**, 384^2 against 224^2, giving $24^2 = 576$ patches rather than $14^2 = 196$;
- **keypoint-mask radius**, $r = 13$ against $r = 7$;
- **training-set size**, the full CUB train split of 5994 images against the 2000-image subsample used at small scale.

The pretraining regime is the confound most likely to bear on a hidden-state-gradient scorer, because attention and representation geometry evolve substantially with the training recipe [23], so the mechanism our target exploits at intermediate depth need not survive a different recipe. One plausible reading is that the larger backbone produces a denser, more semantically coherent CLS-attention map that raises the CLS-attention reference baseline mIoU enough that the HS-IG target can no longer find a wider keypoint-aligned set under the same retention budget. Other readings consistent with the data include the wider keypoint mask making any reasonable scorer hit the ground-truth easily, or the larger training set saturating the CLS-attention reference's generalization gap. Disentangling these confounds requires a controlled cluster-scale sweep, which we leave to future work. The paper's scope is therefore the regime of small backbones at low resolution, 224^2, on CUB, Aircraft, and Cars (DeiT-S/16, and the DeiT-Ti/16 capacity control below), not the larger ViT-B/16 at 384^2. We report the cluster-scale negative result transparently rather than scoping it out. Claiming a phenomenon invariant to backbone scale without the data would not survive reviewer scrutiny.

Table 6 Capacity control on DeiT-Ti/16 at $(L, b) = (6, 25\%)$, three seeds. Same protocol and same metrics as Table 1, with the backbone reduced from 22M to 5M parameters and everything else held fixed. The final column is the paired bootstrap 95% confidence interval on ΔmIoU.

Dataset	Scorer	mIoU ± std (%)	IDAUC ± std	ΔmIoU	95% CI
Aircraft	CLS-attn	61.30 ± 2.25	$+0.108 \pm 0.008$	0.00	—
Aircraft	Attn-Mag	49.16 ± 4.36	$+0.099 \pm 0.015$	-12.14	—
Aircraft	HS-IG	$\mathbf{66.73 \pm 0.37}$	$+0.114 \pm 0.012$	$+5.43$	$[+2.93, +8.09]$
CUB	CLS-attn	35.75 ± 0.97	$+0.137 \pm 0.013$	0.00	—
CUB	Attn-Mag	28.77 ± 0.66	$+0.139 \pm 0.011$	-6.98	—
CUB	HS-IG	$\mathbf{37.01 \pm 1.33}$	$+0.153 \pm 0.002$	$+1.26$	$[-0.44, +3.72]$
Cars	CLS-attn	35.25 ± 0.87	$+0.095 \pm 0.017$	0.00	—
Cars	Attn-Mag	27.58 ± 1.04	$+0.104 \pm 0.005$	-7.67	—
Cars	HS-IG	$\mathbf{38.97 \pm 0.23}$	$+0.107 \pm 0.007$	$+3.71$	$[+2.81, +4.81]$

The decoupling reproduces on a second backbone. To separate capacity from the confounds of the ViT-B/16 comparison, we rerun the three cross-dataset cells on DeiT-Ti/16, which shares DeiT-S/16's input size, token count, and distillation pretraining and differs only in capacity (5M versus 22M parameters), at three seeds. The decoupling reproduces. T_5 lifts preserved-token mIoU by +5.43 pp on FGVC-Aircraft and +3.71 pp on Stanford Cars, both at IDAUC parity, and the paired bootstrap 95% CIs $[+2.93, +8.09]$ and $[+2.81, +4.81]$ exclude zero. On CUB the gap remains positive (+1.26 pp) but its three-seed CI touches zero. On DeiT-Ti even Stanford Cars, the DeiT-S boundary case, decouples cleanly, so the effect is at least as strong at lower capacity. The effect therefore holds across a 5-to-22M-parameter span, which isolates capacity as a dimension along which the decoupling persists rather than reverses. The per-condition values are in Table 6.

Four-loss closure attempt. The bit-identical mIoU pattern in Section 4.6 indicates that the late-block scorer ranking at $L = 11$ is determined by the placement and the ground-truth mask geometry, not by the loss surface used to fine-tune the network. The implication for the recommended recipe is direct: the four-loss variant cannot lift mIoU at the depth where mIoU is most sensitive to loss-driven changes. It therefore offers no principled advantage over the two-loss recipe at any depth. We therefore exclude it from the main paper's contribution list. The bit-identical mIoU pattern is itself a contribution of this paper. It shows that pruning experiments at late-block scorer depths are insensitive to the loss surface. The appropriate scorer placement for any decoupling claim is therefore the intermediate depth window $L \in \{3, 6\}$ identified in Section 4.6.

These two negative results constrain the paper's claim rather than refute it. The scope is limited to small backbones at low input resolution, and the recommended recipe is two-loss at intermediate depth and aggressive budget. Within that scope the decoupling phenomenon holds across three datasets and produces a Pareto-dominant cell on FGVC-Aircraft. The negative results identify two regions where the phenomenon does *not* hold and clarify why.

5.5 Interpretation of the bit-identical mIoU pattern

The bit-identical mIoU result from Section 4.6 is unusual enough to warrant a separate read. Six independently trained checkpoints produce the same per-image preserved-token mask on every one of the 400 evaluation images. Three checkpoints use two-loss training and three use four-loss training; CAL training visibly alters classifier weights enough to flip a 73 $\rightarrow$ 74% accuracy. The pattern survives a careful audit of the eval pipeline and a check that the checkpoints load the trained weights. A comparison of per-cell IDAUC numbers, which shift slightly across cells, confirms that model behaviour is not literally identical across runs.

The simplest mechanistic reading is that the late-block scorer ranking at $L = 11$ is dominated by the pretrained backbone's positional priors. Fine-tuning moves that pattern by less than the margin needed to reorder the top 49 of 196 patches. We measured the score outputs (top-49 patches) being identical, not the attention pattern itself. The precise locus of the invariance, whether attention weights, hidden-state geometry, or both, therefore remains an open question. The hidden-state gradient $\partial y_c / \partial h_i$ at $L = 11$ provides the same per-patch ordering across runs because the residual stream at the last block is structurally invariant to the small per-cell weight changes that fine-tuning produces. This invariance breaks at intermediate depths, as the depth ablation in Section 4.6 shows: HS-IG produces $D = +6.15$ pp at $L = 3$ on CUB. The intermediate-depth regime is where the supervision signal has enough degrees of freedom to discover keypoint-aligned rankings. At late-block depths the pretrained backbone's structural priors dominate and the supervision signal becomes redundant.

The methodological implication is that token-pruning experiments must report the scorer-block depth L alongside the budget b. The choice of L determines whether mIoU is sensitive to loss-driven changes at all. A study that compares two scorers at $L = 11$ on a backbone of comparable depth would observe near-zero mIoU differences regardless of which scorers it tested. Such a study would mistakenly conclude that the supervision-target choice does not matter. Our experimental sweep avoids this confound by reporting four depths in Section 4.6 and identifying the intermediate-depth window as the operationally effective regime. Future scorer-design work should adopt the same protocol.

6 Conclusion

We have presented MIDAS, a measurement framework for studying the supervision-target choice in FGVC token pruning. The framework's two major components are a supervision taxonomy spanning targets T_0 through T_5 that catalogues the design space, and a decoupling-magnitude scalar D that quantifies the signed mIoU gap at parity faithfulness. Together they expose a phenomenon that the FGVC token-pruning literature has not previously measured: preserved-token mIoU and Insertion-Deletion AUC disagree on which scorer is best. On the small-backbone DeiT-S/16, mIoU rises by up to 6.22 pp on FGVC-Aircraft and by 2.11 pp on CUB-200-2011 while Insertion-Deletion AUC remains within 0.02. The effect reproduces on a second small backbone that varies capacity alone (DeiT-Ti/16, +5.43 pp on FGVC-Aircraft

and +3.71 pp on Stanford Cars), and reverses on a larger AugReg ViT-B/16 ($D = -1.45$ pp) in a comparison that changes capacity, pretraining and resolution together, so whether it extends to larger backbones remains open. The hidden-state input-times-gradient target T_5 is consistently the maximum-D choice across all three FGVC datasets at this backbone scale. It Pareto-dominates CLS-attention on the sample means on FGVC-Aircraft at scorer depth $L = 6$ and retention budget $b = 25\%$. The decoupling phenomenon depends jointly on depth and budget: intermediate depths $L \in \{3, 6\}$ and aggressive budgets $b \leq 25\%$ form the regime in which D is reliably positive. Our finding makes explicit a previously unmeasured relationship between two distinct token-selection criteria. It maps onto the distinction between plausibility and faithfulness drawn in the broader attribution-evaluation literature and shows that a scorer's quality on one criterion need not predict its quality on the other.

7 Limitations

The paper's experimental scope is the small-backbone regime at 224^2 input, DeiT-S/16 and DeiT-Ti/16, on three FGVC benchmarks. The decoupling reproduces across these two capacities, but the cluster-scale ViT-B/16 negative result in Section 5 indicates that generalization to larger backbones remains an open empirical question, with at least five plausible confounds between the two scales (Section 5). Mask provenance for FGVC-Aircraft and Stanford Cars is non-native (Grounded-SAM derived rather than dataset-shipped), which we acknowledge as an additional source of evaluation noise. The sign-positive decoupling claim is robust under this noise because both targets see the same masks. However, the absolute mIoU values are mask-source dependent and are not directly comparable across configurations with different keypoint-mask radii or patch grids. Per-cell variance is reported across eight random seeds $\{0, \ldots, 7\}$ for the nine primary cells, and three seeds $\{0, 1, 2\}$ for ablations and auxiliary cells. Even at eight seeds that variance is too wide to certify strict statistical equivalence on the Insertion-Deletion AUC axis, so we claim no parity guarantee: the decoupling rests on the size and consistency of the mIoU gaps, not on a demonstration that faithfulness is unchanged. The Aircraft HS-IG cell has wider seed variance than the corresponding CUB cell ($\sigma = 1.41$ pp vs 0.69 pp). The Aircraft Pareto reading is therefore robust on the worst case over eight seeds in Section 4.6. The Cars cell is sign-positive but falls outside the asymmetric-cond band on the IDAUC axis and is therefore reported as a sign-only result rather than a strict-cond decoupling cell. The Stanford Cars and FGVC-Aircraft sweeps test rigid-object FGVC. Coverage of NABirds, Stanford Dogs, or iNaturalist would extend the cross-dataset claim. Finally, the budget ablation in Section 4.6 uses T_5 supervision only, so the "aggressive-pruning effect" claim is technically a T_5 finding. Whether the same budget-regime structure holds for T_4 or T_2 remains untested. The mechanism explanation in Section 5 is empirical only, with no theoretical guarantee.

8 Future work

Three directions extend the present study. First, the mechanistic explanation in Section 5 attributes the decoupling to three jointly acting properties of the hidden-state input-times-gradient signal: content sensitivity, label specificity, and necessity weighting. The relative contribution of each property remains untested. Ablating each separately, for example by training scorers that use h_i alone or $\partial y_c/\partial h_i$ alone, would isolate which property drives the $D > 0$ effect. Second, the cluster-scale negative result on AugReg ViT-B/16 raises the question of whether the decoupling phenomenon transfers to richer backbones under matched training-set sizes, mask resolutions, and budget conventions. Resolving the multiple confounds enumerated in Section 5 would clarify whether the small-backbone scope is fundamental or a measurement artefact. Third, the per-cell D summary aggregates over a per-class phenomenon. A per-class decoupling distribution on CUB and Aircraft would reveal whether D is driven by easy classes where the two metrics agree or by hard classes where they disagree, with implications for class-conditional supervision. Fourth, the practical significance of the decoupling depends on whether maximising one metric produces better downstream outcomes than the other. The mapping of mIoU to plausibility and of IDAUC to faithfulness in Section 5.3 suggests two end-user goals are at stake. However, empirical evidence that the chosen scorer transfers to out-of-distribution generalization (CUB-C and ImageNet-Sketch-style perturbations on FGVC backbones), to related fine-grained tasks (NABirds, iNaturalist, Stanford Dogs), or to human judgement of part importance through a small-scale localization-fidelity study is outside the scope of the present measurement framework. Establishing these downstream connections would convert the methodological claim of this paper into a stronger practical-significance claim.

Declarations

Corresponding author. Richard Adusei, `radusei@uds.edu.gh`, is the responsible corresponding author for the submission and publication process.

Funding. The authors received no funding, grants, or other support for this work.

Competing interests. The authors declare that they have no competing interests.

Ethics approval and consent to participate. Not applicable. This study uses only publicly available image classification benchmarks and involves no human participants or animals.

Consent for publication. Not applicable.

Data availability. This study analysed three existing public benchmarks and generated no new primary data.

- **CUB-200-2011** [1], deposited in CaltechDATA under the persistent identifier 10.22002/D1.20098.

- **FGVC-Aircraft** [3], distributed by the Visual Geometry Group, University of Oxford, at https://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/; the accompanying technical report carries the identifier 10.48550/arXiv.1306.5151.
- **Stanford Cars** [2]. This dataset has no persistent identifier, and it is no longer distributed from its original location, which was withdrawn by its authors; we used a copy obtained while it was publicly available. Readers should therefore treat the Stanford Cars results as reproducible only from an existing copy of the data.

The Grounded-SAM [45] part masks we derived for FGVC-Aircraft and Stanford Cars are regenerable from the prompt sets in Appendix B, and are released with the code.

Materials availability. Not applicable.

Code availability. The full implementation, including training scripts, evaluation code, the supervision-target taxonomy, and the deterministic mIoU and IDAUC subset definitions, will be released on acceptance.

Author contribution. Richard Adusei conceived the study, designed the supervision-target taxonomy and the decoupling-magnitude scalar, implemented the training and evaluation code, ran all experiments, analysed the results, produced the figures, and wrote the original draft. Kwabena Owusu-Agyemang contributed to the experimental design and to the interpretation of the results, and reviewed and edited the manuscript. Eric Opoku Osei reviewed and edited the manuscript and provided supervision. Michael Asante reviewed and edited the manuscript and provided supervision. All authors read and approved the final manuscript.

References

[1] Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD Birds-200-2011 Dataset. California Institute of Technology; 2011. CNS-TR-2011-001.

[2] Krause J, Stark M, Deng J, Fei-Fei L. 3D Object Representations for Fine-Grained Categorization. In: ICCV Workshop on 3D Representation and Recognition (3dRR). IEEE; 2013. https://doi.org/10.1109/iccvw.2013.77.

[3] Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A. Fine-Grained Visual Classification of Aircraft. arXiv preprint. 2013;https://doi.org/10.48550/arXiv.1306.5151. arXiv:1306.5151.

[4] Wei XS, Song YZ, Mac Aodha O, Wu J, Peng Y, Tang J, et al. Fine-Grained Image Analysis with Deep Learning: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021;https://doi.org/10.1109/tpami.2021.3126648.

[5] Liang Y, Ge C, Tong Z, Song Y, Wang J, Xie P. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. In: ICLR. ICLR; 2022. https://doi.org/10.48550/arXiv.2202.07800.

[6] Rao Y, Zhao W, Liu B, Lu J, Zhou J, Hsieh CJ. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In: NeurIPS. Curran Associates Inc.; 2021. https://doi.org/10.48550/arXiv.2106.02034.

[7] He J, Chen JN, Liu S, Kortylewski A, Yang C, Bai Y, et al. TransFG: A Transformer Architecture for Fine-grained Recognition. In: AAAI. AAAI Press; 2022. https://doi.org/10.1609/aaai.v36i1.19967.

[8] Wang J, Yu X, Gao Y. Feature Fusion Vision Transformer for Fine-Grained Visual Categorization. In: BMVC. BMVA Press; 2021. https://doi.org/10.5244/c.35.192.

[9] Petsiuk V, Das A, Saenko K. RISE: Randomized Input Sampling for Explanation of Black-box Models. In: BMVC. BMVA Press; 2018. https://doi.org/10.48550/arXiv.1806.07421.

[10] Chefer H, Gur S, Wolf L. Transformer Interpretability Beyond Attention Visualization. In: CVPR. IEEE; 2021. https://doi.org/10.1109/cvpr46437.2021.00084.

[11] Fayyaz M, Koohpayegani SA, Jafari FR, Sengupta S, Joze HRV, Sommerlade E, et al. Adaptive Token Sampling for Efficient Vision Transformers. In: ECCV. Springer; 2022. https://doi.org/10.48550/arXiv.2111.15667.

[12] Xu Y, Zhang Z, Zhang M, Sheng K, Li K, Dong W, et al. Evo-ViT: Slow-Fast Token Evolution for Dynamic Vision Transformer. In: AAAI. AAAI Press; 2022. https://doi.org/10.1609/aaai.v36i3.20202.

[13] Abnar S, Zuidema W. Quantifying Attention Flow in Transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 4190–4197. https://doi.org/10.18653/v1/2020.acl-main.385. Available from: https://aclanthology.org/2020.acl-main.385/.

[14] Yin H, Vahdat A, Alvarez JM, Mallya A, Kautz J, Molchanov P. A-ViT: Adaptive Tokens for Efficient Vision Transformer. In: CVPR. IEEE; 2022. https://doi.org/10.1109/cvpr52688.2022.01054.

[15] Meng L, Li H, Chen BC, Lan S, Wu Z, Jiang YG, et al. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. In: CVPR. IEEE; 2022. https://doi.org/10.1109/cvpr52688.2022.01199.

[16] Kong Z, Dong P, Ma X, Meng X, Sun M, Niu W, et al. SPViT: Enabling Faster Vision Transformers via Latency-aware Soft Token Pruning. In: ECCV. Springer; 2022. https://doi.org/10.1007/978-3-031-20083-0_37.

[17] Hou L, Pang RY, Zhou T, Wu Y, Song X, Song X, et al. Token Dropping for Efficient BERT Pretraining. In: ACL. Association for Computational Linguistics;

2022. https://doi.org/10.18653/v1/2022.acl-long.262.

[18] Hu Y, Jin X, Zhang Y, Hong H, Zhang J, He Y, et al. RAMS-Trans: Recurrent Attention Multi-scale Transformer for Fine-grained Image Recognition. In: ACM MM. ACM; 2021. https://doi.org/10.1145/3474085.3475561.

[19] Zhu H, Ke W, Li D, Liu J, Tian L, Shan Y. Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification. In: CVPR. IEEE; 2022. https://doi.org/10.1109/cvpr52688.2022.00465.

[20] Bolya D, Fu CY, Dai X, Zhang P, Feichtenhofer C, Hoffman J. Token Merging: Your ViT But Faster. In: ICLR. ICLR; 2023. https://doi.org/10.48550/arXiv.2210.09461.

[21] Chen M, Shao W, Xu P, Lin M, Zhang K, Chao F, et al. DiffRate: Differentiable Compression Rate for Efficient Vision Transformers. In: ICCV. IEEE; 2023. https://doi.org/10.1109/iccv51070.2023.01574.

[22] Vilas MG, Schaumlöffel T, Roig G. Analyzing Vision Transformers for Image Classification in Class Embedding Space. In: NeurIPS. Curran Associates Inc.; 2023. https://doi.org/10.52202/075280-1741.

[23] Walmer M, Suri S, Gupta K, Shrivastava A. Teaching Matters: Investigating the Role of Supervision in Vision Transformers. In: CVPR. IEEE; 2023. https://doi.org/10.1109/cvpr52729.2023.00723.

[24] Lin TY, RoyChowdhury A, Maji S. Bilinear CNNs for Fine-grained Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;https://doi.org/10.1109/tpami.2017.2723400.

[25] Hu T, Qi H, Huang Q, Lu Y. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification. arXiv preprint. 2019;https://doi.org/10.48550/arXiv.1901.09891. arXiv:1901.09891.

[26] Du R, Chang D, Bhunia AK, Xie J, Song YZ, Ma Z, et al. Fine-Grained Visual Classification via Progressive Multi-Granularity Training of Jigsaw Patches. In: ECCV. Springer; 2020. https://doi.org/10.1007/978-3-030-58565-5_10.

[27] Luo W, Yang X, Mo X, Lu Y, Davis LS, Li J, et al. Cross-X Learning for Fine-Grained Visual Categorization. In: ICCV. IEEE; 2019. https://doi.org/10.48550/arXiv.1909.04412.

[28] Behera A, Wharton Z, Hewage PRPG, Bera A. Context-aware Attentional Pooling (CAP) for Fine-Grained Visual Classification. In: AAAI. AAAI Press; 2021. https://doi.org/10.1609/aaai.v35i2.16176.

[29] Diao Q, Jiang Y, Wen B, Sun J, Yuan Z. MetaFormer: A Unified Meta Framework for Fine-Grained Recognition. arXiv preprint. 2022;https://doi.org/10.48550/arXiv.2203.02751. arXiv:2203.02751.

[30] Chou PY, Lin CH, Kao WC. A Novel Plug-in Module for Fine-grained Visual Classification (HERBS). arXiv preprint. 2023;https://doi.org/10.48550/arXiv.2202.03822.

[31] Rao Y, Chen G, Lu J, Zhou J. Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification. In: ICCV. IEEE; 2021. https://doi.org/10.1109/iccv48922.2021.00106.

[32] Xie Y, Gong Q, Luan X, Yan J, Zhang J. A Survey of Fine-Grained Visual Categorization Based on Deep Learning. Journal of Systems Engineering and Electronics. 2024;https://doi.org/10.23919/jsee.2022.000155.

[33] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. In: NeurIPS. Curran Associates Inc.; 2017. https://doi.org/10.48550/arXiv.1706.03762.

[34] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: ICLR. ICLR; 2021. https://doi.org/10.48550/arXiv.2010.11929.

[35] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training Data-Efficient Image Transformers and Distillation Through Attention. In: ICML. PMLR; 2021. https://doi.org/10.48550/arXiv.2012.12877.

[36] Steiner A, Kolesnikov A, Zhai X, Wightman R, Uszkoreit J, Beyer L. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. Transactions on Machine Learning Research. 2022;https://doi.org/10.48550/arXiv.2106.10270.

[37] Jain S, Wallace BC. Attention is not Explanation. In: NAACL-HLT. Association for Computational Linguistics; 2019. https://doi.org/10.48550/arXiv.1902.10186.

[38] Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: ICML. PMLR; 2017. https://doi.org/10.48550/arXiv.1703.01365.

[39] Hermann KL, Mobahi H, Fel T, Mozer MC. On the Foundations of Shortcut Learning. arXiv preprint. 2024;https://doi.org/10.48550/arXiv.2310.16228. arXiv:2310.16228.

[40] Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences. In: ICML. PMLR; 2017. https://doi.org/10.48550/arXiv.1704.02685.

[41] Park N, Kim S. How Do Vision Transformers Work? In: ICLR; 2022. https://doi.org/10.48550/arXiv.2202.06709. Available from: https://openreview.net/forum?id=D78Go4hVcxO.

[42] DeYoung J, Jain S, Rajani NF, Lehman E, Xiong C, Socher R, et al. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. p. 4443–4458. https://doi.org/10.18653/v1/2020.acl-main.408. Available from: https://aclanthology.org/2020.acl-main.408/.

[43] Hooker S, Erhan D, Kindermans PJ, Kim B. A Benchmark for Interpretability Methods in Deep Neural Networks. In: NeurIPS. Curran Associates Inc.; 2019. https://doi.org/10.48550/arXiv.1806.10758. Available from: https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

[44] Atanasova P, Simonsen JG, Lioma C, Augenstein I. A Diagnostic Study of Explainability Techniques for Text Classification. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics; 2020. p. 3256–3274. https://doi.org/10.18653/v1/2020.emnlp-main.263. Available from: https://aclanthology.org/2020.emnlp-main.263/.

[45] Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv preprint. 2023;https://doi.org/10.48550/arXiv.2303.05499. arXiv:2303.05499.

Appendix A Per-seed numbers

The main-text mean ± standard deviation values are computed over eight seeds $\{0, \ldots, 7\}$ for the nine primary cross-dataset cells (CUB-200-2011, FGVC-Aircraft, and Stanford Cars at $L = 6$, $b = 25\%$, targets CLS-attention, T_1, T_5), and over three seeds $\{0, 1, 2\}$ for the depth and budget ablations and the auxiliary CUB targets. For cell-by-cell verification this appendix lists the first three seeds $\{0, 1, 2\}$ of every cell; the full eight-seed per-seed values for the nine primary cells are in the results ledger released with the code. All numbers are evaluated on the deterministic 400-image mIoU subset and 200-image IDAUC subset defined in Section 3.4. mIoU is reported as a percentage and IDAUC as ins AUC − del AUC. Table 7 gives the CUB main cells, Table 8 the layer-depth ablation, Table 9 the retention-budget ablation, and Table 10 the cross-dataset cells.

The Aircraft T_5 Pareto cell is robust on the per-seed reading. Over all eight seeds the worst-case T_5 seed (60.33%) exceeds the best-case CLS-attention reference seed (57.65%) by +2.68 pp (the three seeds tabulated here give +4.34 pp), and the worst-case T_5 IDAUC (+0.231) falls within 0.005 of the reference's worst case (+0.235). The Cars T_5 margin is much tighter. The worst-case T_5 seed (34.44%) exceeds the best-case CLS-attention reference seed (34.42%) by +0.02 pp, so the population-mean

Table 7 Per-seed numbers underlying Table 3 (CUB main results, $L = 6$, $b = 25\%$). Each cell is (mIoU %, IDAUC).

Target	Seed 0	Seed 1	Seed 2
CLS-attn (ref.)	(34.63, +0.342)	(34.15, +0.326)	(34.91, +0.332)
T_0	(33.79, +0.349)	(33.94, +0.329)	(35.27, +0.341)
T_1	(27.78, +0.323)	(28.33, +0.313)	(32.21, +0.340)
T_2	(34.98, +0.353)	(34.57, +0.342)	(35.33, +0.344)
T_3	(34.66, +0.360)	(33.50, +0.322)	(35.01, +0.338)
T_4	(35.53, +0.352)	(35.05, +0.322)	(37.04, +0.335)
T_5	(36.84, +0.345)	(36.19, +0.323)	(37.67, +0.329)
Unsup. scorer	(20.39, +0.292)	(17.63, +0.236)	(21.89, +0.285)

Table 8 Per-seed numbers underlying Table 4 (layer-depth ablation on CUB at $b = 25\%$). T_5 rows and the per-depth CLS-attention reference rows are listed in pairs.

L	Target	Seed 0	Seed 1	Seed 2
3	T_5	(36.98, +0.324)	(37.01, +0.317)	(38.88, +0.323)
3	CLS-attn	(31.77, +0.278)	(31.38, +0.295)	(31.25, +0.277)
6	T_5	(36.84, +0.345)	(36.19, +0.323)	(37.67, +0.329)
6	CLS-attn	(34.63, +0.342)	(34.15, +0.326)	(34.91, +0.332)
9	T_5	(27.71, +0.347)	(29.04, +0.320)	(27.99, +0.301)
9	CLS-attn	(35.04, +0.381)	(33.64, +0.354)	(34.02, +0.383)
11	T_5	(8.44, +0.078)	(8.44, +0.080)	(8.44, +0.076)

sign-positive reading depends on the across-seed average rather than on per-seed dominance.

Appendix B Mask pipeline details for FGVC-Aircraft and Stanford Cars

CUB-200-2011 ships with 15 part-keypoint annotations per image [1], which we convert to a 14×14 patch mask by union over all visible keypoints with a pixel-radius $r = 7$ at 224^2 image resolution. FGVC-Aircraft and Stanford Cars do not ship native part annotations, so we derive 14×14 patch masks via the Grounded-SAM pipeline [45], which combines Grounding-DINO-tiny for open-vocabulary text-conditioned object detection with the Segment Anything Model (SAM-vit-base) for mask prediction. The pipeline runs on the deterministic 400-image test subset of each dataset.

Prompts. The prompt sets are derived from canonical engineering-decomposition vocabulary for each domain.

- **FGVC-Aircraft (6 prompts)**: {wing, engine, tail, fuselage, cockpit, propeller}.

Table 9 Per-seed numbers for the retention-budget ablation on CUB at $L = 6$, three seeds. T_5 rows and per-budget CLS-attention reference rows are listed in pairs.

b	Target	Seed 0	Seed 1	Seed 2
15%	T_5	(36.23, +0.283)	(34.83, +0.254)	(36.80, +0.268)
15%	CLS-attn	(34.05, +0.284)	(32.25, +0.275)	(33.37, +0.274)
25%	T_5	(36.84, +0.345)	(36.19, +0.323)	(37.67, +0.329)
25%	CLS-attn	(34.63, +0.342)	(34.15, +0.326)	(34.91, +0.332)
50%	T_5	(25.62, +0.384)	(26.01, +0.366)	(25.84, +0.369)
50%	CLS-attn	(26.25, +0.376)	(26.31, +0.368)	(26.48, +0.383)
70%	T_5	(19.85, +0.392)	(19.91, +0.371)	(19.98, +0.381)
70%	CLS-attn	(20.48, +0.381)	(20.46, +0.374)	(20.54, +0.391)

Table 10 Per-seed numbers underlying Table 1 (cross-dataset main results at $L = 6$, $b = 25\%$).

Dataset	Target	Seed 0	Seed 1	Seed 2
Cars	CLS-attn	(34.26, +0.285)	(34.42, +0.262)	(33.17, +0.294)
Cars	T_1	(24.51, +0.231)	(23.92, +0.226)	(24.82, +0.234)
Cars	T_5	(34.44, +0.259)	(34.55, +0.241)	(34.76, +0.274)
Aircraft	CLS-attn	(55.55, +0.236)	(53.12, +0.235)	(56.07, +0.243)
Aircraft	T_1	(47.18, +0.208)	(48.62, +0.229)	(47.64, +0.211)
Aircraft	T_5	(64.48, +0.231)	(61.53, +0.257)	(60.41, +0.235)

- **Stanford Cars (7 prompts)**: {wheel, grille, headlight, taillight, window, door, mirror}.

Pipeline configuration. Both datasets are processed at 224^2 input resolution, matching the training image size, with a Grounding-DINO box detection threshold and a SAM mask-prediction step applied per detected box. The resulting pixel masks are rasterised to the 14×14 patch grid by majority vote within each 16×16 pixel region. The pipeline runs on CPU at approximately 2.37 seconds per image, totalling $\sim$16 minutes per dataset for the 400-image test subset.

Coverage statistics. Aircraft masks have a tighter coverage distribution than Cars masks, as Figure 5 shows. On the 400-image Aircraft subset, the mean number of foreground patches per image is 65.34 of 196 ($p_{25} = 48$, $p_{50} = 60$, $p_{75} = 72$), with 4.43 detected parts on average per image. On the 400-image Cars subset, the mean is 78.08 of 196 ($p_{25} = 51$, $p_{50} = 66$, $p_{75} = 92$), with 5.33 detected parts on average per image. Cars has the wider distribution because cluttered street-scene backgrounds occasionally yield spurious car-part detections that are not removed by the prompt-conditioned selection.

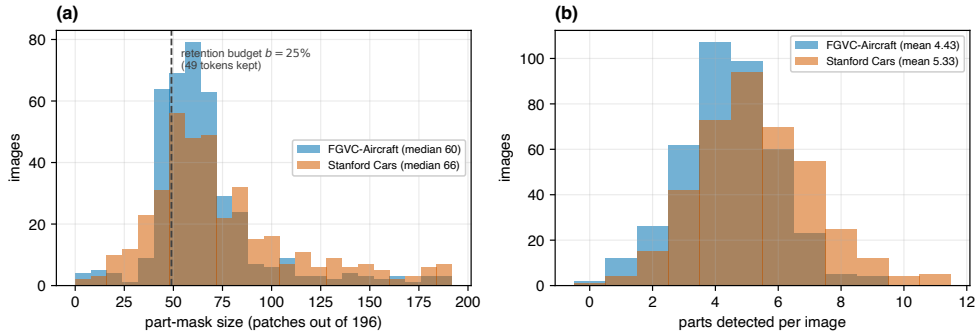


Fig. 5 Quality of the Grounded-SAM ground-truth masks across the full 400-image evaluation subset, rather than on selected examples. **(a)** part-mask size in patches. The masks cover a median of 60 of 196 patches on FGVC-Aircraft and 66 on Stanford Cars, against a retention budget of 49 tokens, so at $b = 25\%$ a scorer cannot reach an IoU of 1 even with a perfect ranking; this is why absolute preserved-token mIoU values are well below 100% and why the paper reads differences between conditions rather than absolute levels. **(b)** parts detected per image, averaging 4.42 on Aircraft and 5.33 on Cars. Two Aircraft images and one Cars image yield an empty mask and are retained without filtering, as described above.

Outliers. Two images on Aircraft and one image on Cars produce empty masks (the Grounded-SAM pipeline failed to detect any of the prompt parts), and one image on Cars produces a saturated mask (all 196 patches foreground). The 1 to 2 outliers per dataset are within the noise budget for the mIoU metric and are retained without filtering, since both targets in any decoupling comparison see the same masks. The full coverage statistics are released alongside the dense-mask arrays.

Why Grounded-SAM masks rather than dataset-shipped masks? Aircraft and Cars do not ship per-image part annotations of any kind. Aircraft ships per-image bounding boxes plus class labels at three hierarchy levels (variant, family, manufacturer), while Cars ships bounding boxes and class labels only. A mask built from bounding boxes alone would saturate at $\sim$90% patch coverage on these datasets and would not separate T_5 from the CLS-attention reference on the mIoU axis. The Grounded-SAM prompt-conditioned mask is the simplest reproducible alternative to hand-annotated parts that preserves the per-image alignment signal we need.

Appendix C Cluster-scale ViT-B/16 cell

The cluster-scale negative result reported in Section 5 trains the HS-IG supervision target on AugReg ViT-B/16 [36] at input resolution 384^2 on the full CUB-200-2011 train split with three random seeds at $(L, b) = (6, 25\%)$. The configuration differs from the small-backbone main-results cells (Table 3) in four respects, listed below for direct reference.

- **Backbone**: AugReg ViT-B/16 (depth 12, embed dim 768, 86 M parameters) versus DeiT-S/16 (depth 12, embed dim 384, 22 M parameters).
- **Input resolution**: 384^2 versus 224^2, giving $N = 24^2 = 576$ patches per image versus $14^2 = 196$.

- **Retention budget**: $b = 25\%$ in both cells, but $K = 144$ tokens preserved at cluster scale versus $K = 49$ at small scale.
- **Keypoint-mask radius**: $r = 13$ pixels at 384^2 (matching the dataset's published keypoint-radius convention at this resolution) versus $r = 7$ at 224^2.
- **Training-set size**: full CUB train (5994 images) at cluster scale versus the 2000-image subsample at small scale.

The training recipe is otherwise identical to the supervised scorer in Section 4.2: the same two-loss objective at $\lambda_{\text{ga}} = 2$, and the same optimiser, schedule and batch size. The CLS-attention reference cells are the corresponding reference cells trained at the same configuration.

Table 11 Per-seed numbers for the cluster-scale cell at AugReg ViT-B/16, 384^2 input, full CUB train, $(L, b) = (6, 25\%)$. mIoU in percent, IDAUC as ins AUC − del AUC.

Target	Seed 0	Seed 1	Seed 2
CLS-attn	(26.03, +0.587)	(26.08, +0.620)	(27.89, +0.597)
T_5	(25.19, +0.612)	(25.35, +0.610)	(25.10, +0.586)

The mean cluster-scale numbers are CLS-attention reference mIoU $= 26.66 \pm 1.06\%$ and T_5 mIoU $= 25.21 \pm 0.13\%$, giving $D = -1.45$ pp. CLS-attention reference IDAUC mean is $+0.601 \pm 0.017$ and T_5 IDAUC mean is $+0.602 \pm 0.015$, Δ IDAUC $= +0.001$ in T_5's favour, so the negative-D result is at parity faithfulness rather than at a faithfulness-trade boundary. The per-seed reading also rules out an outlier-driven explanation. The worst-case CLS-attention reference seed (26.03%) still exceeds the best-case T_5 seed (25.35%) by +0.68 pp.

The three-seed sweep at this configuration runs on a single GPU within a few hours per cell. The CLS-attention reference cells were reused from the cluster-scale sweep at matching (L, b) and incurred no additional compute. The full configuration is released with the code, and the per-seed values are in Table 11.

Appendix D Mechanism probe full data

The mechanism probe in Section 3.4 computes per-image cosine similarity between the T_3 score vector and the T_5 score vector at every transformer block of a single trained CLS-attention reference DeiT-S/16 on a held-out 200-image CUB subset. This appendix reports the full twelve-block table, given in Table 12.

The mean cosine never exceeds +0.082 across the twelve blocks, an order of magnitude below the 0.9 equivalence threshold (the gate condition for which the two targets would be ranking-identical). The standard deviation grows with depth (from 0.09 at block 0 to 0.46 at block 9), reflecting the increasing image-conditioned spread of the per-patch hidden-state-gradient signal. Block 11 is the exception. Cosine of $+0.0000 \pm 0.0000$ across all 200 images indicates that one or both score vectors are

Table 12 Per-block cosine similarity between the T_3 Chefer attention-gradient score and the T_5 hidden-state input-times-gradient score on a 200-image held-out CUB subset, computed on a single trained CLS-attention reference DeiT-S/16 (seed 0). The readout block ($L = 6$) used in the main-text table is highlighted. Block 11 is the deep-readout cell where the HS-IG target collapses to bit-identical mIoU (Section 4.6).

Block	Mean cosine	Std
0	+0.0058	0.0857
1	+0.0370	0.1290
2	+0.0535	0.1615
3	+0.0600	0.1557
4	+0.0575	0.1862
5	+0.0627	0.2221
6	**+0.0764**	**0.3286**
7	+0.0443	0.3121
8	+0.0797	0.3840
9	+0.0816	0.4606
10	−0.0010	0.3238
11	+0.0000	0.0000

degenerate (effectively constant across the patch dimension) at the final block, which is consistent with the bit-identical mIoU collapse reported in Section 4.6. The T_3 formula $A_{\mathrm{cls},i} \cdot \partial y_c / \partial A_{\mathrm{cls},i}$ is dominated at block 11 by a saturated CLS attention pattern, while the T_5 formula $h_i \cdot \partial y_c / \partial h_i$ produces a post-block hidden-state gradient that lies near the residual stream's mean direction at this depth. The same value is observed across all three seeds. The probe was repeated on the CLS-attention reference's seed 0 and reproduces.