WHERE YOU SCORE MATTERS MORE THAN WHAT YOU SCORE

Anonymous authors
Paper under double-blind review

ABSTRACT

Token reduction in vision transformers decides which tokens to keep, merge, or drop. Most existing methods rank tokens by a score computed once, and assume the ranking is independent of the reference set. But the same scoring rule produces different rankings when the reference set changes. We call this the reference-set shift. Changing the reference changes the outcome: relative to full-reference scoring, scoring on the candidate set improves Top-1 on ImageNet-1K by $+10.6$ points, with a mean normalized loss improvement of 0.305 $[0.303, 0.306]$. Because the task loss is a set function, a token's marginal effect depends on interactions with tokens that will be removed, and this dependence is not recoverable from the token's own features. Recovering it requires re-evaluating the candidate set, and matching-based methods do exactly that: they re-evaluate as the set shrinks, so they succeed where scorers computed once cannot. The reference set is therefore an integral part of the method.

1 INTRODUCTION

Vision transformers process an image as a set of tokens, and reducing that set makes inference cheaper (Dosovitskiy et al., 2021; Rao et al., 2021; Liang et al., 2022; Bolya et al., 2023). Most methods rank tokens by a score computed once, and assume the ranking is independent of the reference set. Attention weights in EViT (Liang et al., 2022) and learned importance heads in DynamicViT (Rao et al., 2021) are computed on the full set, but evaluated on the candidate set. We argue that this evaluates the wrong object: the candidate set changes what a reducer keeps, and no scorer computed once on a fixed reference can be robust to that change.

The reason is that the task loss is a set function. The marginal effect of token i is $d_i(R) = L(R \setminus \{i\}) - L(R)$, so a token's score is defined only relative to a reference set. Moving the reference from the full token set V to the candidate set C, the set from which tokens are actually selected, shifts this value. We call the vector of shifts the *reference-set shift*, $\Delta = d(C) - d(V)$. We measure the effect of this shift on ImageNet-1K: scoring the same candidate set under C rather than V changes the selected subset for a third of tokens, and candidate-set scoring improves Top-1 by $+10.6$ points over full-reference scoring while yielding a mean normalized loss improvement of 0.305 $[0.303, 0.306]$. The effect is consistent across models, losses, candidate sets, and budgets (Figure 1).

The shift is not a property of the score, but of when and on which set the score is computed. What matters, then, is not the score but the decision process on a shrinking candidate set: where and when each reduction happens, and on which set each score is taken. Re-evaluating the score on the current candidate set consistently beats computing it once at the start. The same principle applies to matching-based methods, which do not score tokens at all. Our matched decomposition explains why a matching-based reducer such as ToMe (Bolya et al., 2023) wins: its advantage comes from the reduction process itself, not from preserving merged content, and it beats exact per-token scoring computed once in every tested setting.

Our contributions are:

- **We quantify the shift.** Moving the scoring reference from the full sequence to the candidate set changes the selected subset for a third of tokens. Candidate-set scoring improves Top-1 by $+10.6$ points on the full ImageNet-1K validation set and yields a mean normal-

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

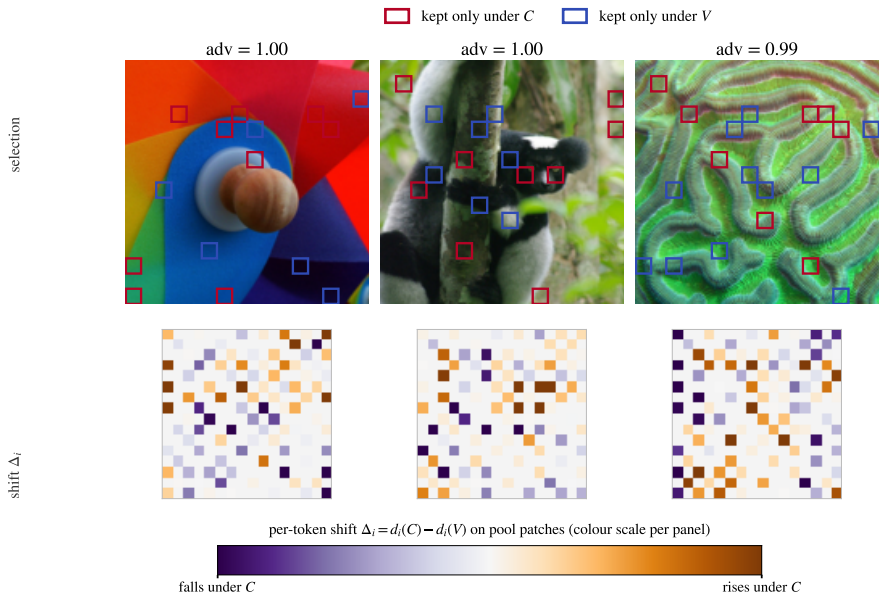


Figure 1: Reference-set shift on ImageNet-1K validation images. *Top:* the same candidate set scored under two references; vermilion outlines mark patches kept only under candidate-set scoring $d(C)$, indigo outlines patches kept only under full-reference scoring $d(V)$. *Bottom:* per-token shift $\Delta_i = d_i(C) - d_i(V)$ on candidate-set patches, with orange marking utility that rises under the candidate-set reference and purple marking utility that falls. Same image, same candidate set: different reference, different value, different subset.

ized loss improvement of 0.305 $[0.303, 0.306]$. The effect holds across models, losses, candidate sets, and budgets.

- **We prove the limit.** A token's marginal effect depends on interactions with tokens that will be removed, and this dependence is not contained in the token's own features. A scorer that reads only per-token features, without re-evaluating the candidate set, cannot represent it.
- **We trace the advantage.** In deployed reducers, the advantage comes from the reduction process, not the score. Class attention is essentially uncorrelated with $d(C)$, and a matched decomposition shows that preserving merged content adds nothing to the advantage of a matching-based reducer.

2 The Reference Set Changes the Selection

2.1 Problem Setup

An input image x is tokenized into a set V of N tokens: a class token plus 16×16 or 14×14 patch tokens at 224^2 resolution. A pretrained backbone computes logits from any retained subset $R \subseteq V$ that includes the class token, inducing a set loss

$$L(R) \;=\; \mathrm{KL}(p_{\mathrm{dense}}(\cdot \mid V) \,\|\, p(\cdot \mid R)) \quad \text{or} \quad L(R) \;=\; -\log p(y_x \mid R), \tag{1}$$

where the first loss, Kullback–Leibler (KL) divergence to the dense prediction (Hinton et al., 2015), never reads the label, and the second, negative log-likelihood (NLL), uses ground truth. Given a candidate set $C \subseteq V$ with $|C| = n$ patch tokens—the class token is always retained and is never a candidate—and a budget $k < n$, a selector must output $S \subseteq C$, $|S| = k$. Under full-reference scoring, marginal effects are computed once and decoded by top-k:

$$d_i(R) \;=\; L(R \setminus \{i\}) - L(R), \qquad S_R \;=\; \arg\operatorname*{top\text{-}k}_{i \in C} \, d_i(R). \tag{2}$$

2.2 The Shift

Reference-set shift is the vector $\Delta = d(C) - d(V)$: the change in every candidate marginal effect when the reference used to compute it is moved from the full token set V to the candidate set C

itself. $\Delta \approx 0$ means the two references are decision-equivalent; $\Delta \neq 0$ means they rank tokens differently. Because L is a set function, token loss is a property of the set over which it is measured, not of the token; the reference is a decision variable of the selection. In particular, the candidate set C plays two roles: it restricts which tokens may survive the budget, and it can also supply the reference on which survival is scored. Δ changes the reference while this restriction stays fixed.

Token importance is thus defined relative to the reference set and to what the decision does with the tokens it does not keep. Evaluations report only the objective and leave the reference implicit. Figure 1 illustrates the reference intervention.

2.3 How we quantify the shift

We measure the shift on the full ImageNet-1K validation set, and verify it on ImageNet-V2 matched-frequency (Recht et al., 2019), on ADE20K segmentation and NYU Depth V2 regression (Silberman et al., 2012), across three ViT backbones, two losses, four candidate-set constructions of 96 patches each—TOKEN-NORM, SPATIAL-COVER, RANDOM, and GLOBAL-KL, which ranks tokens by their full-reference marginal effect—and three budgets $k \in \{32, 64, 80\}$. The full grid is in Appendix C.

Normalized loss improvement. For the treatment reference R_a and the baseline reference R_b with selections S_{R_a}, S_{R_b} in the same cell, we report

$$A(R_a, R_b) \;=\; \frac{L(S_{R_b}) - L(S_{R_a})}{|L(S_{R_a}) - L(V)| + |L(S_{R_b}) - L(V)| + \epsilon}, \tag{3}$$

where $L(V)$ is the loss of the full, unpruned token set and $\epsilon = 10^{-12}$. The numerator is the per-image loss improvement of R_a over R_b; the denominator normalizes by the combined degradation both selectors incur relative to not pruning, so A is bounded and symmetric, positive when S_{R_a} has the lower loss. A is computed per image and summarized with image-clustered bootstrap 95% confidence intervals (Efron & Tibshirani, 1994).

2.4 Scoring on the candidate set changes decisions

If token loss were a property of the token, scoring the same candidate set under V and under C would select the same subset. We test this on the full ImageNet-1K validation set. The main experiment crosses 2 models, 2 losses, 4 candidate sets, and 3 budgets (Table 4, Appendix B). Scoring on the candidate set wins with a mean normalized loss improvement of 0.305 $[0.303, 0.306]$, positive in 45 of 48 cells and in every stratum: 0.392 $[0.390, 0.394]$ on DINOv2 versus 0.218 $[0.216, 0.220]$ on DeiT-III, and 0.426 $[0.425, 0.428]$ under label-NLL versus 0.184 $[0.182, 0.186]$ under teacher-KL. The three exceptions share one setting, the GLOBAL-KL candidate set under the teacher-KL loss, where the shift is significantly negative (down to -0.10 for DINOv2 at k=64). The two references agree on only 66.5% of selected tokens. The effect grows with budget, $0.253 \rightarrow 0.356$, consistent with the stability margin of Proposition 1: at larger k more candidates sit near the decision boundary where the shift exceeds the margin.

The candidate set gradient confirms the prediction: the GLOBAL-KL candidate set, already screened by full-reference first-order evidence, shows the smallest effect, 0.118, and the only three negative cells are GLOBAL-KL cells under teacher-KL. A candidate set formed by the full-reference marginal effect has already absorbed most of the benefit of re-evaluation. The weakest regime is the one that has already paid the cost candidate-set scoring removes: GLOBAL-KL is the only candidate set whose construction requires a full-reference sweep, while the three candidate sets a deployed pipeline can afford—TOKEN-NORM, SPATIAL-COVER, RANDOM—all sit in the strong regime, 0.285–0.441.

2.5 Set content is the cause

The shift comes from the content of the reference set. The overlap experiment on a 5,000-image subset of ImageNet-1K holds reference size fixed at 96 patches and varies the fraction q of decision-support tokens drawn from the true candidate set: for each candidate i,

$$R_i(q) \;=\; \{i\} \cup A_i(q) \cup D_i(q), \qquad A_i(q) \subseteq C \backslash \{i\}, \quad D_i(q) \subseteq V \backslash C, \quad |A_i(q)| + |D_i(q)| = 95, \tag{4}$$

with three independent context replicates, so only the fraction of decision-support tokens drawn from the true candidate set varies.

Varying q (Table 6) establishes the effect and its structure: the endpoint improvement is 0.320 $[0.316, 0.324]$, positive in 44 of 48 cells, and the overlap slope is 0.372 $[0.323, 0.409]$, positive in 43 of 48 cells. Across context replicates, score Spearman correlation averages 0.305: non-candidate identity still matters. Overlap is a robust explanatory coordinate, not a sufficient statistic for reference quality.

We also test whether a static correction computed once under V can explain the effect. It cannot. With the correction strength fixed on a 5,000-image calibration subset, scoring on the candidate set retains a 0.303 $[0.298, 0.308]$ improvement on a second, disjoint 5,000-image subset, and remains ahead, 0.107 $[0.102, 0.112]$, against an upper bound that grants each instance its best possible correction strength: the interactions that matter are conditioned on the candidate-set context, not merely present in it. Evaluation is a decision process, not a one-time computation: backward elimination that re-evaluates marginal effects every 4 deletions beats $d(C)$ computed once at the start by 0.538 $[0.493, 0.586]$, positive in 96.9% of cells.

The effect also holds across distribution shift, task change, and model scale. On the full ImageNet-V2 matched-frequency test set, the effect is 0.297 $[0.293, 0.300]$, with per-model, per-loss, candidate set, and budget gradients reproduced. An ADE20K segmentation probe yields 0.417 $[0.409, 0.426]$, and an NYU Depth V2 regression probe yields 0.253 $[0.227, 0.279]$, both positive in all nine candidate-set-by-budget cells. Scaling to DINOv2 ViT-B/14 replicates the endpoint at 0.273 $[0.240, 0.307]$ (Table 6).

3 THE SHIFT IS NOT CONTAINED IN PER-TOKEN FEATURES

The shift described in Section 2 has two candidate sources: the loss function, and the scoring rule. We first bound the regime in which the shift leaves the selection unchanged. Proposition 1 gives this bound. For the loss source, Proposition 2 shows that even when the loss is nearly additive, reference choice can still change the decision. For the scoring source, Proposition 3 shows that no per-token scorer can represent the shift, however it is trained. Propositions 2 and 3 together locate the shift in the set-valued structure of the loss, not in the quality of the scorer.

When the shift leaves the selection unchanged.

Proposition 1 (Ranking stability) *Let* $\tau = \frac{1}{2}\min\{\, d_i(V) - d_j(V) \,:\, i \in S_V,\ j \in C \setminus S_V \,\}$ *be half the top-k margin under the full reference, and let* $\delta = \Delta = d(C) - d(V)$. *If* $\|\delta\|_\infty < \tau$ *then* $S_C = S_V$.

The proof is immediate: for $i \in S_V$ and $j \notin S_V$, $d_i(C) - d_j(C) = \big(d_i(V) - d_j(V)\big) + (\delta_i - \delta_j) \geq 2\tau - 2\|\delta\|_\infty > 0$.

Reference choice under an approximately additive loss. A near-additive loss is one natural candidate for why the reference would not matter: if token contributions are essentially independent, the set over which they are measured should not change the ranking. Proposition 2 shows the opposite.

Proposition 2 (Separation) *For every* $M \geq 1$ *there is a ground set* V, *candidate set* $C \subseteq V$, *budget* k, *and set loss* L_M *such that, under the uniform measure on subsets of* V, *the mean-squared error of the best additive approximation of* L_M *is* $O(1/M) \to 0$ *and the largest pairwise interaction weight is* $O(1/M) \to 0$, *yet top-k selection under* $d(V)$ *incurs excess loss bounded below by a positive constant independent of* M, *while top-k selection under* $d(C)$ *is optimal.*

The construction binds the failure to candidate-set formation and fixed-budget selection: a candidate interacts with many partners outside the candidate set whose *individual* influence vanishes, so averaged additive behavior is essentially exact; but $d(V)$ charges the candidate for all of that mass at once, because the reference V still contains every partner that selection is about to discard. The first candidate source is ruled out.

The representational limit of per-token scoring. The second candidate source is the scorer: a better scorer—trained on more data, with more parameters, or on better features—might represent the shift. Proposition 3 shows that this is not possible in principle: there are candidate sets in which two candidates have *identical* features yet every reference disagrees about their order, so any per-token scorer is forced to err.

Proposition 3 (Impossibility of reference-robust per-token scoring) *There exist a ground set V with feature map h, a candidate set $C \subseteq V$, and two set losses L^+, L^- such that (i) $h(i_1) = h(i_2)$ for two candidates $i_1, i_2 \in C$; (ii) under L^+, $d(V)$ ranks i_1 strictly above i_2 while $d(C)$ ranks i_2 strictly above i_1, and the $d(V)$-top-2 selection has loss exceeding the $d(C)$-top-2 selection by $u > 0$; (iii) under L^- both comparisons reverse, and again the $d(V)$-top-2 selection has loss exceeding the $d(C)$-top-2 selection by u; (iv) both losses are additive except for one candidate-conditional edge and interactions with tokens outside the candidate set of weight $a/M \to 0$. Consequently, every scoring rule of the form $s_i = g(h_i)$ incurs excess loss of at least u on at least one of the two instances, while the top-2 selection under $d(C)$ is optimal under both.*

The construction is a mirror pair: in L^+ a candidate's support outside the candidate set makes it valuable under V but not under C; in L^- the support belongs to the other candidate. Because the two candidates share features, any per-token scorer must rank them the same way in both instances, and therefore contradicts one reference's strict order, while scoring on the candidate set is optimal in both. The learning experiments of Section 5 confirm the impossibility on two backbones, DeiT-III Small/16 and DINOv2 ViT-S/14: a per-token ridge head reaches a candidate-set rank correlation of only 0.013/0.011 with exact $d_i(C)$, and a 182,657-parameter set-attention scorer with full visibility of all 96 candidates performs comparably to a parameter-matched per-token control.

4 Related work

Token reduction as an implicit reference choice. Since ViT (Dosovitskiy et al., 2021) and its data-efficient recipes (Touvron et al., 2021; 2022), token reduction has been pursued by learned pruning (Rao et al., 2021), token reorganization (Liang et al., 2022), adaptive halting (Yin et al., 2022; Meng et al., 2022), sampling (Fayyaz et al., 2022), merging (Bolya et al., 2023), pruning–merging hybrids (Kim et al., 2024), globally informed scorers such as V-Pruner (Yao et al., 2026), content-adaptive tokenizers such as DART (Yin et al., 2025). These works differ in how losses are predicted but share one implicit choice: the reference context in which losses are evaluated. None of these works quantifies how that choice changes the selection. That choice is the object of this paper.

Importance depends on a reference. A related line of work provides the precedent: attribution methods treat importance as not a property of a feature alone. Shapley values average over reference sets (Lundberg & Lee, 2017), integrated gradients chooses a baseline (Sundararajan et al., 2017), influence functions use leave-one-out perturbations (Koh & Liang, 2017), and attention flow propagates context (Abnar & Zuidema, 2020). The shift we identify is related to, but not the same as, reference dependence in attribution. Attribution allocates credit to features or inputs; a selector acts on a counterfactual marginal whose reference is fixed by the candidate set being chosen. The reference is therefore a design variable of the decision itself. Propositions 1–2 show that changing it can flip the chosen subset.

5 Re-evaluation decides

The advantage of a reducer comes from when and on which set its scores are computed.

5.1 The advantage comes from the decision process

To make this precise, we describe a reducer as a sequence of rounds: each round computes its proposals on some reference state and then drops, merges, or halts, until k tokens remain. A reducer is *re-evaluated* when its proposals are recomputed on the current candidate set. This re-evaluation is what drives the advantage: re-evaluating on the candidate set as it shrinks consistently beats computing the proposals once at the start. Three controlled contrasts test this.

Table 1: Equal-token ToMe comparison. Attention, $d(V)$, and $d(C)$ prune within the attention-formed candidate set of 96 patches; ToMe merges the full patch sequence to the same final count. All arms use the same 50,000 ImageNet-1K validation images, pretrained backbone, dense teacher, teacher-KL loss, and post-hoc Top-1. These results correspond to the three contrasts below.

		Attention		$d(V)$		$d(C)$		ToMe	
Model	k	KL	Top-1	KL	Top-1	KL	Top-1	KL	Top-1
DeiT-III Small	32	1.695	0.483	0.959	0.638	0.778	0.677	**0.386**	**0.777**
	64	0.623	0.708	0.438	0.749	0.362	0.761	**0.183**	**0.796**
	80	0.434	0.747	0.353	0.765	0.234	0.782	**0.134**	**0.800**
DINOv2 ViT-S/14	32	1.558	0.510	1.334	0.552	1.158	0.590	**0.290**	**0.743**
	64	0.560	0.696	0.478	0.714	0.484	0.710	**0.164**	**0.774**
	80	0.371	0.733	0.337	0.742	0.246	0.756	**0.130**	**0.781**

Note: ToMe improves over attention and over $d(C)$ in all six cells; its normalized advantage over $d(C)$ is 0.021–0.468, with every confidence interval above zero. ToMe merges the full sequence to the same final count, so it is an equal-count comparison.

Refresh. Re-probing the surviving strata after each drop, with candidate set, strata, actions, objective, and budget held fixed, improves over computing them once at the start in all six model–budget cells: teacher-KL loss 0.789 $[0.636, 0.957]$, 0.162 $[0.112, 0.220]$, 0.024 $[0.016, 0.034]$ at k=32/64/80.

Update semantics. We next hold the schedule fixed and vary only whether a destination receives the merged source content. Keeping contextual keys, matching edges, rounds, and removed sources identical leaves the outcome unchanged, 0.007 $[-0.024, 0.038]$: the matching schedule carries the ToMe control's advantage, and preserving the merged content adds nothing to it.

System level. At the system level, ToMe improves over exact $d(C)$ computed once at the start in all six cells, by 0.021–0.468 in normalized advantage, with every confidence interval above zero (Table 1). This is the empirical form of Proposition 3.

5.2 Re-evaluation in deployed reducers

The same pattern appears in deployed reducers. We audit those that do not re-evaluate. On the full 50,000-image ImageNet-1K validation set, EViT-style class attention is essentially uncorrelated with $d_i(C)$, with rank correlation 0.002 on DeiT-III and 0.012 on DINOv2. Substituting exact $d(C)$ improves over attention in all six model–budget cells: by 0.264–0.318 on DeiT-III and 0.062–0.142 on DINOv2, with every confidence interval above zero (Table 1). Per-cell numbers are in Appendix B.

Since exact scoring is expensive, we ask whether a learned scorer can recover the candidate-set marginal effect from either per-token features or full candidate-set context. Two matched learning experiments (Appendix C) answer no. A closed-form ridge head on per-token features reaches candidate-set-level rank correlation of only 0.030/0.017 with exact $d_i(C)$; a 182,657-parameter set-attention scorer given full visibility of all 96 candidates performs comparably to a parameter-matched per-token control, with aggregated advantage -0.003 $[-0.035, 0.029]$. Proposition 3 explains why: $d_i(C)$ depends on which co-members survive, information that per-token features do not contain. The cost of exact scoring is bounded: batched evaluation of all 96 marginal effects takes 60–65 ms against 11 ms for a single forward, affordable for offline and latency-tolerant selection.

Finally, we check iterative re-evaluation directly. On a separate 5,000-image subset with teacher-KL selection, iterative re-evaluation improves over full-reference scoring by 0.735 $[0.731, 0.740]$ and over candidate-set scoring computed once at the start by 0.633 $[0.629, 0.638]$ at k=64, with the same ordering at k=32 and k=80; ground-truth Top-1 improves by $+15.5$ pp $[+14.8, +16.2]$ and prediction retention by $+23.6$ pp $[+22.8, +24.4]$.

6 Discussion

Reference choice matters for selection. Reference-set shift makes reference choice an explicit part of every top-k pruning decision. Once a token's marginal effect is recognized as a property of

the set over which it is measured, the relevant design question becomes which reference the rule is evaluated on. Reducers that re-evaluate on the candidate set as it shrinks answer this question by construction, and their advantage comes from the decision process rather than from a better score.

Why per-token scoring is limited. The limitation of per-token scoring is structural rather than a matter of capacity. A token's marginal effect depends on which co-members survive, and that information is not contained in the token's own features. Per-token scorers—including trained ones with more parameters—cannot represent the shift, which is why matching-based reducers that never compute a score can outperform exact per-token scorers. The relevant distinction is between scoring rules that re-read the candidate set and rules that do not.

Reference quality varies with content. Reference quality varies with content, not with distance along any deletion order. The fixed-length overlap experiment has a positive average overlap slope, but replicate score correlation remains only 0.305: which tokens sit beside a candidate matters in addition to their overlap with the candidate set. Reference quality spans both endpoints and controlled intermediate references, and comparisons across reference designs need to cover the full range.

Open questions. The mechanism we identify has an empirical scope bounded by this paper: three ViT backbones, four candidate set constructions, three budgets, and teacher-preservation or label losses. Extending the mechanism to language, retrieval, further distribution shift, other architectures, and jointly trained token-pruning pipelines is open. Whether the same decision-process advantage carries to those settings, and whether reference-aware training can recover what per-token scorers cannot, are also open.

7 Conclusion

Token reduction is not only a scoring problem. The reference set on which a score is taken causally changes which tokens survive. A scorer computed once on a fixed reference is not robust to that change, because the shift is not contained in per-token features. What determines fidelity is the decision process: re-evaluating the score on the candidate set beats computing it once at the start, and the reducers we tested operate this way. The set on which a score is taken is part of the method.

AI use statement

In this work, we used generative AI tools only for language editing and polishing of the manuscript, a recommended disclosure task. We have not used generative AI tools for any required disclosure tasks, including but not limited to: generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, assisting in the writing of proofs, designing or providing feedback on research methodology or experiments, implementing methods, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting results. All required disclosure tasks are not applicable to this work.

We reviewed all AI-assisted text to ensure accuracy, consistency, and fidelity to our scientific claims. All authors verified the final manuscript. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

Experimental details, including dataset splits, model backbones, candidate-set constructions, budgets, and evaluation protocols, are in Appendix C. Code and configuration files are included as anonymized supplementary material.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sarbajit Sengupta, Hamid Sajedi, and Juergen Gall. ATS: Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, pp. 396–414, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2603–2612, 2024.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894, 2017.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. EViT: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774, 2017.

Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. AdaViT: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12309–12318, 2022.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13937–13949, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*, pp. 746–760, 2012.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 4798–4807, 2017.

Hugo Touvron, Matthieu Cord, Matthieu Douillard, Piotr Bojanowski, Armand Joulin, and Marc Larussi. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357, 2021.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. *arXiv preprint arXiv:2204.07118*, 2022.

Guangzhen Yao, Jiayun Zheng, Zezhou Wang, Wenxin Zhang, Renda Han, Chuangxin Zhao, Zeyu Zhang, and Runhao Liu. V-pruner: A fast and globally-informed token pruning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34396–34404, 2026.

Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10809–10818, 2022.

Shicheng Yin, Kaixuan Yin, Yang Liu, Weixing Chen, and Liang Lin. DART: Differentiable dynamic adaptive region tokenizer for vision foundation models. *arXiv preprint arXiv:2506.10390*, 2025.

A FULL FACTORIAL RESULTS

Table 2 and Table 3 report every cell of the main experiment—normalized loss improvement of $d(C)$ over $d(V)$, image-clustered 95% CIs. Of the 48 cells, 45 are positive. The three negative cells share one setting: the GLOBAL-KL candidate set under the teacher-KL loss.

Table 2: Per-cell normalized loss improvement on the full ImageNet-1K validation set: DeiT-III Small.

Task	Candidate set	k=32	k=64	k=80
NLL	GLOBAL-KL	0.062 [0.057, 0.068]	0.167 [0.163, 0.171]	0.248 [0.244, 0.251]
NLL	RANDOM	0.248 [0.242, 0.254]	0.372 [0.367, 0.377]	0.444 [0.440, 0.448]
NLL	SPATIAL-COVER	0.318 [0.313, 0.324]	0.431 [0.426, 0.435]	0.499 [0.495, 0.504]
NLL	TOKEN-NORM	0.180 [0.175, 0.186]	0.301 [0.297, 0.306]	0.394 [0.390, 0.398]
KL	GLOBAL-KL	-0.007 [-0.011, -0.002]	0.048 [0.043, 0.053]	0.116 [0.111, 0.121]
KL	RANDOM	0.111 [0.107, 0.116]	0.162 [0.158, 0.167]	0.225 [0.221, 0.230]
KL	SPATIAL-COVER	0.163 [0.158, 0.167]	0.177 [0.172, 0.181]	0.232 [0.227, 0.236]
KL	TOKEN-NORM	0.064 [0.059, 0.068]	0.104 [0.099, 0.109]	0.175 [0.170, 0.179]

Table 3: Per-cell normalized loss improvement on the full ImageNet-1K validation set: DINOv2 ViT-S/14.

Task	Candidate set	k=32	k=64	k=80
NLL	GLOBAL-KL	0.286 [0.281, 0.291]	0.301 [0.297, 0.306]	0.325 [0.322, 0.329]
NLL	RANDOM	0.573 [0.569, 0.578]	0.707 [0.704, 0.711]	0.704 [0.701, 0.707]
NLL	SPATIAL-COVER	0.659 [0.655, 0.663]	0.758 [0.755, 0.761]	0.740 [0.737, 0.743]
NLL	TOKEN-NORM	0.411 [0.406, 0.416]	0.547 [0.543, 0.551]	0.548 [0.544, 0.552]
KL	GLOBAL-KL	0.013 [0.008, 0.019]	-0.103 [-0.109, -0.098]	-0.038 [-0.043, -0.033]
KL	RANDOM	0.316 [0.312, 0.321]	0.298 [0.292, 0.303]	0.346 [0.342, 0.351]
KL	SPATIAL-COVER	0.435 [0.431, 0.440]	0.426 [0.421, 0.430]	0.456 [0.452, 0.460]
KL	TOKEN-NORM	0.207 [0.202, 0.212]	0.211 [0.205, 0.216]	0.275 [0.270, 0.280]

B ADDITIONAL RESULTS

Reference depth analysis. On 40 images per model, changing the layer at which the reference is read changes the top-k selection: top-k disagreement is 0.270 $[0.262, 0.279]$ for DeiT-III Base and 0.282 $[0.275, 0.288]$ for DINOv2 ViT-S/14. The backward re-evaluation holdout in Table 6 uses a refresh step of four deletions and spans both losses, all four candidate sets, and all three budgets.

C EXPERIMENTAL DETAILS

Evaluation subsets. The full factorial experiment and deployed-reducer audit use the full ImageNet-1K validation set (50,000 images); the ImageNet-V2 replication uses the full matched-frequency test set (10,000 images); the dense-prediction validations use the full ADE20K validation set (2,000 images) and the full NYU Depth V2 validation set (654 images). The mechanism experiments (overlap experiment, static second-order calibration and validation) use class-balanced 5,000-image subsets of ImageNet-1K, disjoint across stages. All subset assignments are determined by a fixed random seed.

Table 4: Full factorial experiment: normalized loss improvement of candidate-set scoring $d(C)$ over full-reference scoring $d(V)$ (Eq. 3); image-clustered 95% CIs.

Contrast	Improvement	95% CI
Overall	0.305	[0.303, 0.306]
Model		
DINOv2 ViT-S/14	0.392	[0.390, 0.394]
DeiT-III Small	0.218	[0.216, 0.220]
Utility		
label-NLL	0.426	[0.425, 0.428]
teacher-KL	0.184	[0.182, 0.186]
Candidate set		
SPATIAL-COVER	0.441	[0.439, 0.443]
RANDOM	0.376	[0.373, 0.378]
TOKEN-NORM	0.285	[0.282, 0.287]
GLOBAL-KL	0.118	[0.116, 0.121]
Budget k		
32	0.253	[0.251, 0.254]
64	0.307	[0.305, 0.308]
80	0.356	[0.354, 0.357]

Table 5: Iterative group-probe elimination at equal final token count: teacher-KL loss, post-hoc Top-1. Iterative = group-probe elimination with re-probing of the surviving strata inside the new candidate set after each drop: 26/45/71 forwards per image at k=80/64/32; the static variant costs 15 group probes.

		Attention		$d(V)$/$d(C)$		Iterative group	
Model	k	KL	Top-1	KL	Top-1	KL	Top-1
DeiT-III Small	32	1.108	0.500	1.038	0.610	**0.325**	**0.750**
	64	0.662	0.660	0.405	0.730	**0.157**	**0.790**
	80	0.494	0.690	0.352	0.710	**0.157**	0.780
DINOv2 ViT-S/14	32	1.748	0.460	1.163	0.510	**0.371**	**0.730**
	64	0.518	0.650	0.435	0.710	**0.129**	**0.780**
	80	0.363	0.700	0.271	0.750	**0.122**	**0.770**

Note: the $d(V)$ and $d(C)$ columns show the two fixed-reference scoring rules; iterative group probing at 26–71 forwards (budget-dependent) improves over per-token scoring computed once in all twelve model-by-budget conditions.

Deployed scoring rules experiment. The class-attention rule of Section 5 uses final-block CLS-to-patch attention averaged over heads, computed from the last transformer block's own queries and keys at the standard resolution; the candidate set is the top-96 patches by that attention, and all three selection rules act within this candidate set at budgets $\{32, 64, 80\}$. Teacher-KL is recomputed by a forward pass on every selected subset; Top-1 is the pruned model's argmax against the folder label, evaluated post hoc. The $d(V)$ and $d(C)$ marginal effects use the same evaluator as every other experiment in the paper.

Cross-task dense-prediction experiments. Both dense-prediction validations use the same pretrained DINOv2 ViT-S/14 backbone and attach a trained linear head read at decision layer 3, then advance through the pretrained suffix per query. *ADE20K segmentation*: a 150-class linear probe trained with cross-entropy on 2,000 training images, with a fixed feature cache; the utility loss is per-patch cross-entropy over pixels assigned to their nearest retained patch, and realized metrics are patch-level pixel accuracy and mIoU. *NYU Depth V2 regression*: a linear log-depth head trained with the scale-invariant log loss on 2,000 training frames from the official train split; the utility loss is log-RMSE over pixels assigned to their nearest retained patch, and realized metrics are meter-scale RMSE and δ_1, with threshold 1.25. Both use three candidate sets—token-norm, spatial-cover, random—budgets $\{32, 64, 80\}$, and the same normalized loss improvement and image-clustered

Table 6: Full validation set results. Normalized loss improvements compare candidate-set scoring with full-reference scoring; CIs are image-clustered 95% intervals.

Validation	Estimate	95% CI	Positive cells
Overlap experiment improvement	0.320	[0.316, 0.324]	44/48
Overlap slope	0.372	[0.323, 0.409]	43/48
ADE20K segmentation	0.417	[0.409, 0.426]	9/9
NYU Depth V2 regression	0.253	[0.227, 0.279]	9/9
DINOv2 ViT-B/14 improvement	0.273	[0.240, 0.307]	—
Backward re-evaluation holdout	0.538	[0.493, 0.586]	48/48

bootstrap as every other experiment. Ground-truth annotations are read only through the utility definition; no selection reads a label except the label-NLL arms.

Equal-token ToMe comparison. On the same 50,000 ImageNet-1K validation images, model weights, dense teacher, and budgets, we implement training-free ToMe bipartite soft matching from each block's head-averaged keys. Prefix tokens are protected; multiple sources assigned to one destination are accumulated; token sizes persist across blocks; and the DINOv2 patch readout is weighted by those sizes. A deterministic 12-block schedule reaches exactly 32, 64, or 80 final patch tokens. Paired percentile-bootstrap intervals use 10,000 image resamples. ToMe vs. $d(C)$ normalized loss improvement at $k{=}32/64/80$ is 0.110 $[0.105, 0.115]$, 0.052 $[0.047, 0.057]$, and 0.021 $[0.016, 0.027]$ on DeiT-III, and 0.468 $[0.464, 0.473]$, 0.357 $[0.352, 0.362]$, and 0.242 $[0.237, 0.247]$ on DINOv2. ToMe merges full-sequence information, so it is interpreted as an equal-final-count comparison.

Matched learned-scorer intervention. The set-attention scorer and its per-token control share architecture capacity of 182,657 parameters, training data of 300 images per model, three fixed seeds, the token-norm candidate set, budget $k{=}64$, and teacher-KL utility; the only difference is whether self-attention sees all 96 candidates or each token independently. The ridge baseline of the same section uses closed-form ridge regression on pretrained per-token features. Both interventions test whether the candidate-set marginal effect can be learned from per-token features under controlled budgets. Per-model teacher-KL for global/local/contextual/per-token: $1.033/0.647/1.440/1.532$ on DINOv2, $0.715/0.541/0.971/0.928$ on DeiT-III.

Bootstrap. All reported quantities are per-image paired differences; CIs are percentile bootstrap intervals resampling images as clusters. Models, tasks, candidate sets, and budgets are reported as stratifications, never as independent replicates.

Multiplicity and robustness. The 48 cells are stratifications of a single factorial design. Applying Holm correction to two-sided image-level sign-flip tests leaves all 48 cells significant at family-wise $\alpha = 0.05$ (45 positive, 3 negative); the main result is the aggregated image-clustered contrast.