GRADSELECT: OUTPUT-TARGETED PATCH RANKING UNDER EXACT-K VISION TRANSFORMER PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

How should a frozen vision transformer choose K patches when the pruning layer and token budget are fixed? This paper isolates patch ranking from system choices such as pruning schedules, routing modules, and classifier changes. GradSelect scores each image by first running the full model to obtain its own predicted class, backpropagating that output to the cut layer, keeping positive first-order support for each patch, and fusing this signal with CLS attention after per-image z-scoring. The fusion weight is selected on a lock split, and all final comparisons use the same exact-K intervention. At 25% retention after block 8, GradSelect has the highest point estimate among primary selectors in all nine ImageNet-1K, Food-101, and SUN397 cells across CLIP, DINOv2, and DeiT, including five cells not used to choose the recipe. Against the stronger per-cell opponent drawn from either the primary suite or the attribution/removal stress suite, it wins 6/9 cells with +2.24-point mean and +2.53-point median accuracy margins; exact single-token removal remains stronger on supervised DeiT. On Caltech-101, GradSelect is best on CLIP and DINOv2 and 0.22 points below DivPrune on DeiT. Because scoring requires a full forward pass and one backward pass, GradSelect is best understood as an offline output-aware patch selector rather than an online accelerator.

1 INTRODUCTION

Patch pruning in vision transformers is often reported as an end-to-end efficiency result. Yet a large part of the decision is simpler and more diagnostic: after choosing a layer and a token budget, which K patches should remain? Existing systems commonly change several ingredients at once, including the pruning schedule, router, classifier, token operation, endpoint, and execution path. Those choices matter for deployment, but they make it hard to tell whether one ranking signal actually selected a better subset of patches.

We isolate this ranking question with a shared exact-K intervention. A frozen backbone is cut after a fixed block; each selector receives the same CLS and patch tokens; exactly K patches are kept; the selected sequence is restored to spatial order; and the same frozen suffix and head or probe produce the final prediction on the same held-out split. Under this interface, the comparison is not a complete-system leaderboard. It asks a narrower question that a reader can inspect directly: when everything after selection is matched, which scoring rule best preserves top-1 accuracy?

GradSelect answers the question with the model's own output. For each image, the unpruned frozen model first predicts a class. One backward pass to the cut then measures how each patch locally supports that predicted output. GradSelect keeps only positive signed support, standardizes it within the image, fuses it with standardized CLS attention, and selects the exact Top-K patches. The fusion coefficient is chosen once on a selector-lock split and is fixed before final evaluation. Labels train probes, choose the global coefficient, and measure accuracy, but they do not score tokens for an individual image. Figure 1 shows the matched intervention and the qualitative difference in selected patches.

The main result is a controlled ranking result, with explicit reuse boundaries. At block 8 and 25% exact-K retention, GradSelect has the highest primary-suite point estimate in all nine combinations of ImageNet-1K, Food-101, and SUN397 with CLIP-B/16, DINOv2-B/14, and DeiT-B/16. Four of these cells helped form the recipe; the other five method cells were not inspected when choosing predicted targeting, positive support, z-score fusion, or the coefficient grid. The narrowest primary

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.

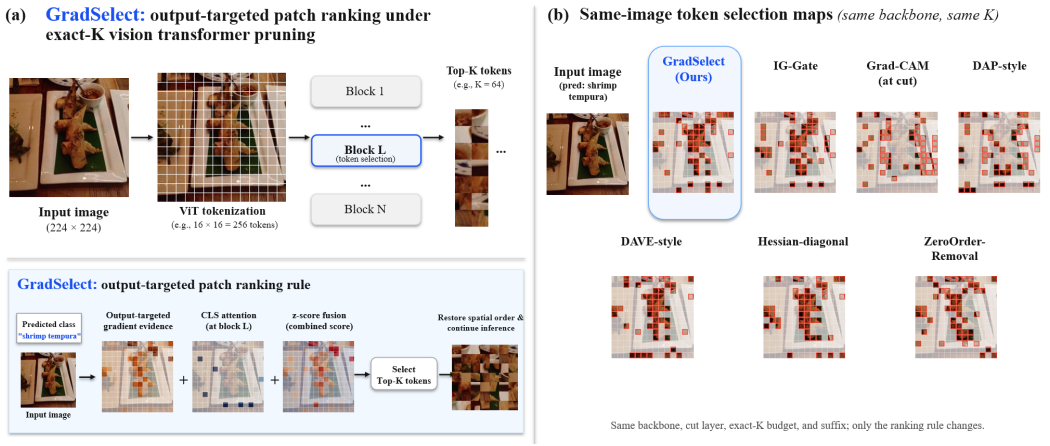

Figure 1: Overview of the controlled exact-K protocol and same-image selections under a matched backbone, cut, token budget, and downstream computation.

comparison is ImageNet/DeiT, where GradSelect exceeds TRIO-NMS by exactly 0.244 points over 50,000 paired predictions, with Holm-adjusted exact McNemar $p = 0.0325$.

We also test where the ranking claim weakens. A separate offline stress suite adds integrated gradients, Grad-CAM, decision-aware rules, Hessian-diagonal scoring, and exact single-token removal as exact-K rankers. Against the stronger opponent drawn from either the primary suite or this stress suite in each cell, GradSelect wins 6/9 cells, with +2.24-point mean and +2.53-point median margins. Exact removal is stronger in all three supervised-DeiT stress cells. On Caltech-101 transfer, GradSelect is best on CLIP and DINOv2 but is 0.22 points below DivPrune on DeiT.

The design evidence narrows the mechanism rather than expanding the claim. In the four recipe-forming cells, GradSelect is the best locked factorial recipe. At common coefficients, all 16 target-by-sign difference-in-differences estimates are positive, supporting an interaction: predicted-class targeting helps most when the score retains positive signed support. This does not establish an unconditional target effect or a causal attribution explanation. Because GradSelect requires a full forward pass and one backward pass before pruning, it should be read as an offline output-aware selector, not an online accelerator.

Our contributions are:

- We formulate frozen-model patch selection as an exact-K ranking problem with a shared cut, representation, suffix, head or probe, and final split.
- We introduce GradSelect, a lock-tuned fusion of predicted-class positive support and CLS attention, and separate four recipe-forming cells from five uninspected method cells in the main 3×3 evaluation.
- We evaluate the ranking claim against access-stratified primary and stress suites, Caltech-101 transfer, and retention, cut-depth, representation, removal, and cost diagnostics that expose both gains and boundaries.

The remainder defines the exact-K interface and GradSelect, positions it relative to pruning and attribution work, and evaluates the ranking claim and its limits. The supplement contains complete matrices, paired tests, fixed-λ statistics, transfer endpoints, sensitivity analyses, and provenance checks.

2 GradSelect

2.1 Exact-K intervention

The method starts from a matched intervention rather than from a pruning system. Split a frozen vision transformer after block L into a prefix $F_{\leq L}$ and suffix $F_{>L}$. For image x, the prefix returns a

CLS token z_0 and N patch tokens

$$[z_0; Z] = F_{\leq L}(x), \qquad Z = [z_1, \ldots, z_N] \in \mathbb{R}^{N \times d}. \tag{1}$$

A selector receives these tokens and returns a subset $S \subseteq \{1, \ldots, N\}$ with $|S| = K$. We keep the CLS token, restore the selected patch tokens Z_S to their original spatial order, and evaluate

$$p_S(c \mid x) = \mathrm{softmax}(h(F_{>L}([z_0; Z_S])))_c \tag{2}$$

with the unchanged frozen suffix and fixed classification head or probe h. Thus every final comparison shares the cut, representation, token count, sequence order, suffix, head, and final split. Only the score used to choose S changes. The main setting uses $L = 8$ and $K = \lfloor 0.25N \rfloor$. Figure 2 summarizes both the scoring path and the downstream exact-K path.

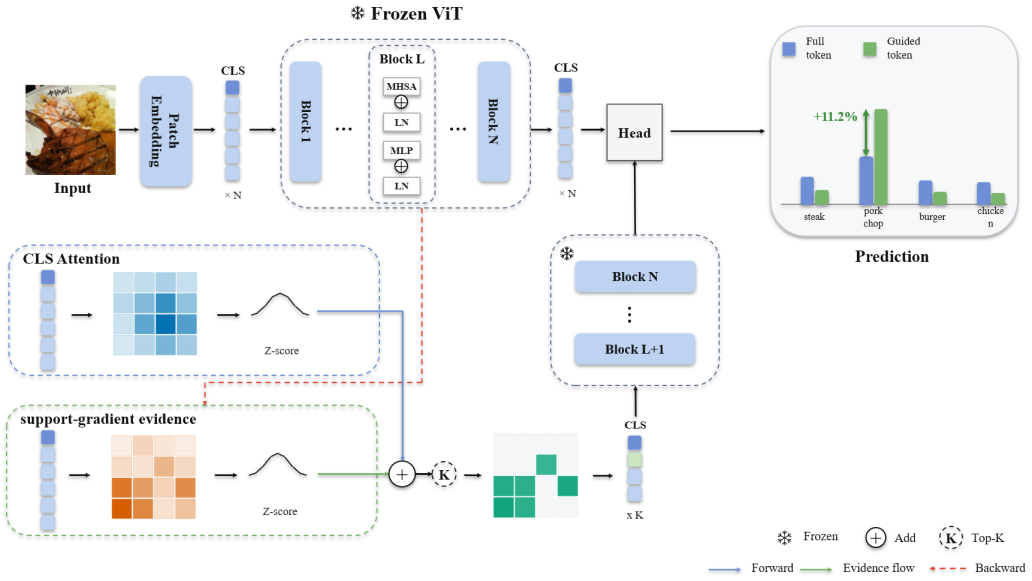

Figure 2: GradSelect in a frozen vision transformer. At block L, CLS attention and output-targeted support-gradient evidence are standardized per image and fused with a lock-selected coefficient. Exact Top-K retains the selected patches together with CLS, restores their spatial order, and passes them through the unchanged frozen suffix and head. The full-token forward/backward scoring path makes the selector offline.

2.2 Predicted Target and Positive Support

GradSelect first asks which output the full frozen model is already making. Let $p_{\mathrm{full}}(c \mid x) = p_{\{1,\ldots,N\}}(c \mid x)$ be the full-token prediction. The image-specific target is

$$\hat{y} = \arg\max_c p_{\mathrm{full}}(c \mid x), \qquad t(x) = \log p_{\mathrm{full}}(\hat{y} \mid x). \tag{3}$$

The argmax is detached before differentiation. This target is label-free at selection time: labels are used to train the probe, lock the global fusion coefficient, and evaluate final accuracy, but not to score patches for a given image. The target also comes from the same frozen model, not from an external teacher.

For each patch, GradSelect uses a local first-order gate sensitivity and keeps only positive support:

$$u_i = \langle z_i, \nabla_{z_i} t(x) \rangle = \sum_{j=1}^{d} z_{ij} \frac{\partial t(x)}{\partial z_{ij}}, \qquad g_i^+ = \max(0, u_i). \tag{4}$$

The feature channels are summed before rectification, so positive and negative channel contributions can cancel inside a token. Equivalently, if a scalar gate replaces z_i by $m_i z_i$, then $u_i = \partial t / \partial m_i|_{m_i=1}$. The score is therefore a local first-order proxy for how the token supports the predicted-class log probability. It is related to gradient attribution and Taylor criteria (Ancona et al., 2018; Molchanov et al., 2019), but it is not the causal effect of deleting the token and reprocessing the shortened sequence.

2.3 CLS FUSION AND EXACT TOPK

Positive support is output-specific, while CLS attention provides a forward, output-agnostic view of which patches the model aggregates near the cut. Let a_i be the CLS-to-patch attention adjacent to the cut, averaged across heads. Since a_i and g_i^+ live on different scales, GradSelect standardizes each vector within the image:

$$\mathcal{Z}(v_i) = \frac{v_i - \bar{v}}{\max\left(\sqrt{N^{-1}\sum_{r=1}^{N}(v_r - \bar{v})^2}, \epsilon\right)}, \qquad \bar{v} = N^{-1}\sum_{r=1}^{N} v_r. \tag{5}$$

The final score is

$$s_i^{\mathrm{GS}} = \lambda \mathcal{Z}(a_i) + (1-\lambda)\mathcal{Z}(g_i^+), \qquad S = \mathrm{TopK}\big(s^{\mathrm{GS}}, K\big). \tag{6}$$

We set $\epsilon = 10^{-8}$. The coefficient λ is the CLS weight: $\lambda = 0$ gives support-only selection and $\lambda = 1$ gives CLS-TopK. Exact TopK enforces the same retained-token count for all selectors; ties follow deterministic implementation order, and selected patch indices are spatially sorted before entering the suffix.

2.4 LOCKING AND VARIANTS

For each dataset/backbone setting, a selector-lock split chooses $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. We maximize lock-split top-1 accuracy and break ties by macro-F1 and then by the smaller coefficient. The chosen coefficient is frozen before final evaluation. Thus λ is a protocol hyperparameter, not a learned backbone parameter or per-image label input. Probe training and final evaluation never overlap, and final data are separate from selector locking. Section 4.2 discloses the earlier Food-101 and SUN397 reuse of the lock partition for probe-epoch selection.

The design factorial keeps the intervention fixed while crossing predicted versus deterministic non-predicted target, absolute $|u_i|$ versus positive $\max(0, u_i)$ evidence, and empirical-rank versus z-score fusion. Each recipe locks its own λ. GradSelect refers specifically to the predicted-target, positive-support, z-score recipe. Because independent locking can confound factor contrasts, the experiments also compare target-by-sign effects at common coefficients.

2.5 COMPUTATIONAL SCOPE

GradSelect performs one full-token forward pass, backpropagates the predicted-class target to block L, computes the N token scores, and then runs the suffix on the selected K patches. The full forward and backward happen before pruning. Therefore the shortened suffix does not make GradSelect an online accelerator. We treat it as an offline output-aware selector whose rankings can be cached or used as teacher targets for a cheaper router; router distillation is outside the scope of this paper.

3 RELATED WORK

Frozen vision transformers as a controlled substrate. Vision Transformers make patch tokens explicit intermediate units (Dosovitskiy et al., 2021). Supervised DeiT, language-supervised CLIP, and self-supervised DINOv2 induce different token spaces, but each can be used as a frozen backbone for downstream classification (Touvron et al., 2021; Radford et al., 2021; Oquab et al., 2024). This diversity is useful for our question because the representation changes across backbones while the intervention can remain fixed. Within each cell, we hold the backbone, cut, suffix, head or probe, and token budget constant, and study only the ranking rule.

Token reduction as system design. Efficient-ViT methods often combine token importance with execution choices. Learned sparsification methods train routers, reorganize tokens, sample or pool patches, or progressively discard tokens together with the network (Rao et al., 2021; Liang et al., 2022; Fayyaz et al., 2022; Marin et al., 2022; Pan et al., 2021). Training-free and merging methods exploit pretrained attention or aggregate tokens instead of deleting them (Wang et al., 2024; Bolya et al., 2023). These systems answer deployment questions: how should a model run faster under a

particular architecture and schedule? Our work asks a narrower diagnostic question. Because TopK rules can remain competitive under a common intervention (Haurum et al., 2023), system-level gains alone do not identify which ranking signal selected the better exact-K subset.

Recent ranking rules and adaptation limits. Recent reduction methods derive selection from representation shift, task-conditioned objectives, attention correction, or auxiliary cropping modules (Choi et al., 2025; Zhang et al., 2026; Ichikawa et al., 2026; Bergner et al., 2025). Vision-language pruning systems use attention, adaptive merging, diversity, or saliency–coverage criteria before multimodal decoding (Chen et al., 2024; Shang et al., 2025; Alvar et al., 2025; Deng et al., 2025). We adapt RepShift, DivPrune, and TRIO-NMS only where their score or selection rule can be placed at the same cut and forced to return exactly K patches. FastV, PruMerge, Token Cropr, SCOPE, ToMe, and HAC-ToMe remain complete-system context because reproducing them would change the task interface, schedule, endpoint, token operation, or runtime stack. The distinction is important: our tables compare rankers under a declared interface, not every original method as deployed.

Attribution, sensitivity, and removal. First-order saliency methods measure local output sensitivity, while integrated gradients and Grad-CAM aggregate gradient information along a path or into spatial maps (Simonyan et al., 2013; Ancona et al., 2018; Sundararajan et al., 2017; Selvaraju et al., 2017). Transformer attribution propagates attention or relevance through token interactions; newer ViT methods add decision-aware or distribution-aware gradient structure (Abnar & Zuidema, 2020; Chefer et al., 2021; Jo et al., 2026; Wrobel et al., 2026). Hessian and Taylor pruning criteria connect sensitivity to structural or token decisions (Yang et al., 2021; Uddin et al., 2025; Molchanov et al., 2017; 2019). Reliability studies caution that explanations should be tested against behavior (Adebayo et al., 2018; Hooker et al., 2019). GradSelect uses predicted-output first-order support as a ranking signal, not as a causal explanation; exact single-token removal supplies a higher-cost boundary, and the full forward/backward scoring path makes the method offline rather than an online accelerator.

4 Experiments

4.1 Protocol

The experiments test a ranking claim under a fixed intervention: once the cut, budget, suffix, head or probe, and final split are matched, does GradSelect choose a better exact-K subset? The main matrix crosses ImageNet-1K, Food-101, and SUN397 with frozen CLIP-B/16, DINOv2-B/14, and DeiT-B/16 (Russakovsky et al., 2015; Bossard et al., 2014; Xiao et al., 2010; Radford et al., 2021; Oquab et al., 2024; Touvron et al., 2021). Unless stated otherwise, we cut after block 8 and retain 25% of patches: 49/196 for CLIP and DeiT, and 64/256 for DINOv2. Every displayed comparison uses the same cut representation, retained-token count, suffix/probe, and final split. Complete systems that change the schedule, endpoint, classifier, token operation, or hardware are kept protocol-separated in the supplement.

Each frozen backbone feeds one affine linear probe trained on its final CLS feature. We use AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 512, at most 50 epochs, and patience 8, selecting the epoch by validation accuracy. ImageNet uses 40,000 probe-training, 10,000 probe-validation, 10,000 selector-lock, and 50,000 final images. Food-101 uses 65,650/10,100/25,250 train/lock/final images, and SUN397 uses 15,880/3,970/19,850. In the earlier Food and SUN runs, the lock split also selected the probe epoch. Because the probes are data-limited and frozen-feature only, Full accuracy is not intended to match full-data probes, fine-tuned models, or public heads. We interpret within-cell paired selector comparisons.

The primary endpoint is final-split top-1 accuracy. Caltech-101 also reports macro-F1 because its class sizes differ. We do not use KL, target-logit, or calibration endpoints to rank selectors.

4.2 Main matrix: recipe-forming and uninspected cells

This subsection tests the central primary-suite claim. Four cells were recipe-forming: Food/CLIP, Food/DeiT, SUN/DeiT, and SUN/DINOv2. Their final predictions were inspected while choosing predicted-class targeting, positive support-gradient evidence, z-score fusion, the λ grid, and tie handling. We then fixed that recipe before evaluating five uninspected method cells: all three

Table 1: Matched-selector matrix on the complete registered final splits at block 8 and 25% exact-K retention. Four cells (Food/CLIP, Food/DeiT, SUN/DeiT, SUN/DINO) contributed to recipe formation; five did not. Values are top-1 accuracy (%); DINO abbreviates DINOv2 and bold marks the best selector in this suite.

	ImageNet-1K			Food-101			SUN397		
Selector	CLIP	DINO	DeiT	CLIP	DINO	DeiT	CLIP	DINO	DeiT
Full	72.04	77.71	80.77	92.38	91.82	77.51	77.32	75.22	62.93
CLS-TopK	53.69	70.34	77.47	56.42	80.17	67.57	54.66	65.68	55.90
Gradient-L_2	37.39	42.38	79.03	42.88	45.87	73.03	43.71	60.85	56.78
Center+CLS	61.62	70.35	77.28	82.93	82.41	68.66	63.35	65.67	55.90
Uniform+CLS	61.18	71.89	78.43	82.55	85.43	72.71	67.01	68.47	59.94
RepShift score	60.85	66.34	77.24	76.05	81.44	66.46	62.69	64.61	54.75
DivPrune rule	58.40	71.17	78.59	80.79	84.99	71.47	64.94	69.08	58.69
TRIO-NMS rule	44.13	44.76	79.42	52.29	46.97	73.92	47.75	61.67	57.49
GradSelect (Ours)	**69.72**	**75.17**	**79.67**	**90.38**	**90.32**	**76.99**	**75.63**	**72.90**	**61.31**

Table 2: Conservative strongest-opponent summary at block 8 and 25% exact-K retention. The opponent is the higher-accuracy method from either the primary matched suite or the added attribution/sensitivity suite in each cell. Values are top-1 accuracy (%); Δ is GradSelect minus that opponent.

Dataset	Backbone	GradSelect	Ranker	Best	Δ
ImageNet-1K	CLIP	69.72	IG-Gate	66.77	$+2.95$
ImageNet-1K	DINOv2	75.17	Uniform+CLS	71.89	$+3.29$
ImageNet-1K	DeiT	79.67	ZeroOrder	80.68	-1.02
Food-101	CLIP	90.38	Center+CLS	82.93	$+7.45$
Food-101	DINOv2	90.32	Uniform+CLS	85.43	$+4.89$
Food-101	DeiT	76.99	ZeroOrder	77.43	-0.44
SUN397	CLIP	75.63	IG-Gate	73.90	$+1.72$
SUN397	DINOv2	72.90	IG-Gate	70.37	$+2.53$
SUN397	DeiT	61.31	ZeroOrder	62.53	-1.22

ImageNet cells, Food/DINOv2, and SUN/CLIP. ImageNet images had appeared in earlier internal analyses of predecessor recipes, so these are uninspected method cells rather than a new-dataset test.

Table 1 reports the primary matched suite, and the supplement gives the longer matrix and paired statistics. GradSelect has the highest point estimate in all nine cells[1], including the four recipe-forming cells. The narrowest comparison is ImageNet/DeiT: over 50,000 paired predictions, GradSelect exceeds TRIO-NMS by $(1663-1541)/50000 = +0.244$ points, with Holm-adjusted exact McNemar $p = 0.0325$ across the nine primary comparisons. This is a small but paired positive result, not a large-effect claim.

4.3 Stress test and Caltech transfer

The next two evaluations ask whether the main result survives stronger or newer settings. The stress suite adds integrated gradients (Sundararajan et al., 2017), Grad-CAM (Selvaraju et al., 2017), DAP (Jo et al., 2026), DAVE (Wrobel et al., 2026), Hessian-guided sensitivity (Uddin et al., 2025), and exact single-token removal. These rows are exact-K adaptations, not complete-system reproductions.

Because the added stress rows are unfused and not lock-tuned with CLS attention, the +4.20-point mean against only the strongest added stress row is descriptive and asymmetric. Table 2 therefore uses a stricter comparison: in each cell, GradSelect is compared with the stronger opponent from either

[1] RepShift, DivPrune, and TRIO-NMS enter as scoring or selection rules inside the fixed single-cut protocol, not as complete-system reproductions.

the primary suite or the stress suite. Under that rule, GradSelect wins 6/9 cells, with a +2.24-point mean and +2.53-point median margin. The boundary is clear: exact single-token removal is stronger on all three supervised-DeiT cells by 0.44–1.22 points.

Caltech-101 tests transfer to a new dataset with non-overlapping probe-training, probe-validation, selector-lock, and final roles. CLIP and DeiT were evaluated first; DINOv2 was added under a locked addendum before its final results were opened. We therefore report the three-backbone outcome without treating Caltech as uniformly prospective. CLIP and DeiT retain 49/196 patches, DINOv2 retains 64/256, and selector-lock chooses every tunable coefficient.

Table 3 shows the transfer result. GradSelect is best on CLIP and DINOv2, reaching 94.72 and 94.56 top-1; the supplement reports macro-F1 and paired details. On DeiT, GradSelect is 0.216 points below DivPrune (0.22 rounded): 37 images are correct only for GradSelect and 41 only for DivPrune, with exact McNemar $p = 0.734$. This cell is statistically unresolved and remains in the transfer conclusion.

Table 3: Caltech-101 transfer on the complete held-out final split. Values are top-1 accuracy (%). GradSelect uses the same scoring rule as Table 1, with λ selected only on selector-lock; bold marks the best exact-K selector for each backbone.

Backbone	Full	Strongest comparator	Comp.	GradSelect
CLIP	96.23	RepShift	93.91	**94.72**
DINO	96.98	Uniform+CLS	92.67	**94.56**
DeiT	94.07	DivPrune	**92.99**	92.78

4.4 Design factorial

This subsection tests why the chosen recipe is plausible, not whether it is prospectively confirmed. The factorial crosses target identity, attribution sign, and fusion normalization in the four recipe-forming cells. Figure 3 shows that GradSelect is the best lock-selected recipe in all four. To remove independent coefficient locking as a confound, we also compare targets and signs at common $\lambda \in \{0.25, 0.5\}$. The difference-in-differences $(\mathrm{Pred} - \mathrm{Nonpred})_{+} - (\mathrm{Pred} - \mathrm{Nonpred})_{\mathrm{abs}}$ is positive in all 16 cell/fusion/coefficient comparisons, ranging from +6.86 to +34.44 points. This supports a target-by-sign interaction in the inspected cells; it does not establish an unconditional target main effect or a causal attribution claim.

4.5 Fidelity boundary and cost

Exact token removal asks whether a score tracks the Full model's predicted-class log-probability drop when one token is removed. Table 4 compares GradSelect with CLS attention on all three datasets for the two 196-patch backbones; DINOv2 is absent because its cut exposes 256 patches. GradSelect has higher signed Spearman correlation in five of six final cells. This is a consistency check, not causal validation: the score is a first-order proxy for local removal. Deletion AUC, which removes groups of high-ranked patches, splits by backbone—lower for GradSelect on CLIP and higher on DeiT. Together with the DeiT exact-removal stress result, removal is a boundary diagnostic rather than the definition of selector quality.

Table 4: Removal-fidelity boundary on six complete final splits. Entries are GradSelect minus CLS attention. Higher signed Spearman ($\Delta\rho$) and larger deletion AUC drop (ΔDel.) are better.

	CLIP		DeiT	
Dataset	$\Delta\rho$	ΔDel.	$\Delta\rho$	ΔDel.
ImageNet-1K	$+0.191$	-0.148	-0.078	$+0.185$
Food-101	$+0.220$	-0.404	$+0.160$	$+0.527$
SUN397	$+0.275$	-0.214	$+0.068$	$+0.295$

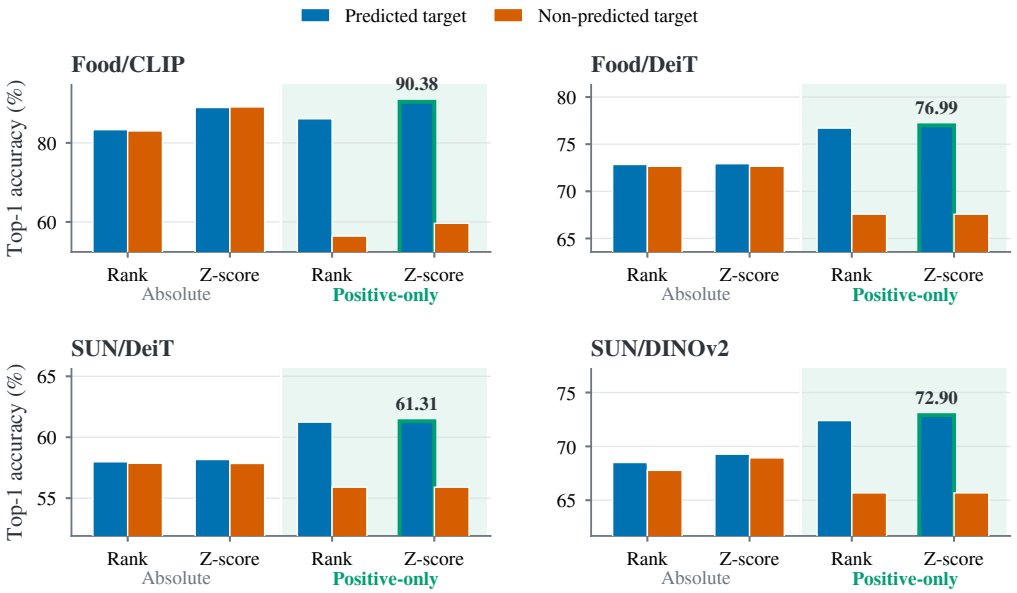

Figure 3: Post-hoc design factorial on four recipe-forming final cells. Blue and coral bars use predicted and non-predicted targets; the green field marks positive-only attribution, and the outlined bar is GradSelect. Each recipe locks λ independently. These four values also appear in the matrix and are mechanism evidence, not an independent evaluation.

Table 5: Measured end-to-end cost on an RTX 3090 at block 8 and 25% retention. Latency is median ms per batch of 16 after 30 warm-up iterations and 200 paired repeats. GradSelect includes target-defining Full inference, one backward pass, selection, and shortened-suffix inference. Throughput is images/s (Full / GradSelect).

Dataset	Backbone	Full ms	GradSelect ms	Ratio	Throughput
ImageNet-1K	DeiT	13.12	20.71	1.58×	1219 / 773
Food-101	DINOv2	26.02	39.16	1.50×	615 / 409
SUN397	CLIP	18.72	29.18	1.56×	855 / 548

Table 5 includes the Full forward pass that defines the target, one backward pass to block 8, TopK selection, and the shortened suffix/probe pass. Across three anchors, GradSelect takes 1.50–1.58× the median latency of ordinary Full inference and delivers 63–66% of Full throughput. These measurements rule out an online acceleration claim; GradSelect produces an offline ranking that can be cached or used as a teacher target.

4.6 Sensitivity and the DeiT boundary

The final main-paper diagnostics vary the operating point without changing the selection interface. At block 8, the lock-selected mixture exceeds CLS-TopK in all 27 retention cells. At 25% retention across blocks 4, 8, and 10, it is better in 23 cells and tied in four late-cut cells where selector-lock chooses the CLS-only endpoint. Because CLS-TopK is the $\lambda = 1$ endpoint of the tuned family, this sweep maps the operating envelope rather than an independent external-baseline test. Representation preservation is mixed but mostly favorable: GradSelect improves median sample-cosine preservation in 15/27 cells and class-geometry preservation in 19/27.

Figure 4 shows the largest differences at tighter budgets and earlier cuts. The four zeros occur when selector-lock selects the CLS-only endpoint; no coefficient is retuned on the final split.

Why DeiT is tighter. ImageNet/DeiT is the smallest-margin primary cell. Under three independently initialized probes, GradSelect reaches 79.67, 79.74, and 79.81 top-1, with lock curves favoring

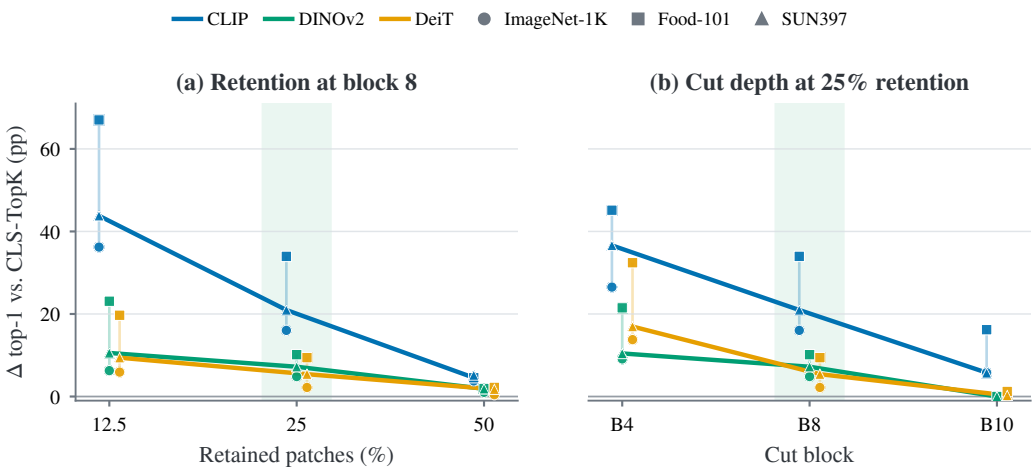

Figure 4: GradSelect minus CLS-TopK top-1 accuracy across retention and cut depth. Colored curves are dataset medians for each backbone; bars span the dataset range, and markers show individual datasets. CLS-TopK is the $\lambda = 1$ endpoint of the lock-tuned GradSelect family.

$\lambda = 0.25$–0.5. This checks GradSelect's probe sensitivity, but TRIO-NMS was not rerun with all three probes, so it does not establish pairwise-ordering stability. Exact single-token removal is stronger in every supervised-DeiT stress cell, although it requires evaluating every candidate removal. DeiT therefore marks the clearest empirical boundary of the ranking claim.

5 Conclusion

GradSelect isolates a basic question inside ViT patch pruning: under a fixed cut, exact token budget, frozen suffix, head or probe, and final split, which rule should rank the patches? The answer proposed here is to combine the model's own predicted-output positive support with CLS attention, lock the fusion coefficient before final evaluation, and compare all selectors through the same exact-K intervention. In the primary matched suite, GradSelect has the highest point estimate in all nine dataset–backbone cells, while the manuscript separates the four recipe-forming cells from the five uninspected method cells.

The evidence also defines the method's limits. Against the stronger per-cell opponent drawn from the primary or attribution/sensitivity stress suite, GradSelect wins 6/9 cells with +2.24-point mean and +2.53-point median margins; exact single-token removal is stronger in all three supervised-DeiT cells. Caltech-101 transfer is positive on CLIP and DINOv2, but GradSelect is 0.22 points below DivPrune on DeiT with an unresolved paired test. The method is therefore a controlled offline selector, not a causal explanation or an online accelerator. Its natural next use is as a cached ranking signal or as supervision for a cheaper router that would have to be evaluated as a separate deployment system.

Reproducibility statement. Experiments track probe training, probe validation, selector locking, and final evaluation under a fixed exact-K intervention; final evaluation never overlaps probe training or selector locking. The supplement records the complete protocol, including earlier probe-validation reuse, matrices, and diagnostics. Code will be released upon acceptance.

AI use statement. OpenAI Codex and other large-language-model tools assisted with literature search, experiment execution, code and figure production, reference organization, and editing. No synthetic datasets or AI-generated proofs were used; all numbers come from stored artifacts. The authors take responsibility for the manuscript and its claims.

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. doi: 10.18653/v1/2020.acl-main.385.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.

Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Token cropr: Faster ViTs for quite a few tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9740–9750, 2025. doi: 10.1109/CVPR52734.2025.00910.

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.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461, 2014. doi: 10.1007/978-3-319-10599-4_29.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35, 2024. doi: 10.1007/978-3-031-73004-7_2.

Joonmyung Choi, Sanghyeok Lee, Byungoh Ko, Eunseo Kim, Jihyung Kil, and Hyunwoo J. Kim. Representation shift: Unifying token compression with FlashAttention. In *IEEE/CVF International Conference on Computer Vision*, pp. 20456–20466, 2025. doi: 10.1109/ICCV51701.2025.01902.

Jinhong Deng, Wen Li, Joey Tianyi Zhou, and Yang He. SCOPE: Saliency-coverage oriented token pruning for efficient multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2025.

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.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Juergen Gall. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, pp. 396–414, 2022. doi: 10.1007/978-3-031-20083-0_24.

Joakim Bruslund Haurum, Sergio Escalera, Graham W. Taylor, and Thomas B. Moeslund. Which tokens to use? investigating token reduction in vision transformers. In *IEEE/CVF International Conference on Computer Vision Workshops*, pp. 773–783, 2023. doi: 10.1109/ICCVW60793.2023.00085.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Yuki Ichikawa, Masato Motomura, Thiem Van Chu, and Daichi Fujiki. Efficient vision transformers via token merging with head-wise attention correction. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3908–3917, 2026. doi: 10.1109/WACV61042.2026.00381.

Sehyeong Jo, Gangjae Jang, and Haesol Park. Decision-aware attention propagation for vision transformer explainability. *arXiv preprint arXiv:2604.18094*, 2026.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16062–16072, 2022.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2017.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11264–11272, 2019.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. IA-RED^2: Interpretability-aware redundancy reduction for vision transformers. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

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*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE/CVF International Conference on Computer Vision*, pp. 618–626, 2017.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *International Conference on Computer Vision*, 2025.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328, 2017.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357, 2021.

Mohammad Helal Uddin, Liam Seymour, and Sabur Baidya. HEART-ViT: Hessian-guided efficient dynamic attention and token pruning in vision transformer. *arXiv preprint arXiv:2512.20120*, 2025.

Hongjie Wang, Bhishma Dedhia, and Niraj K. Jha. Zero-TPrune: Zero-shot token pruning through leveraging of the attention graph in pre-trained transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16070–16079, 2024. doi: 10.1109/CVPR52733.2024.01521.

Adam Wrobel, Siddhartha Gairola, Jacek Tabor, Bernt Schiele, Bartosz Zielinski, and Dawid Rymarczyk. DAVE: Distribution-aware attribution via ViT gradient decomposition. In *International Conference on Machine Learning*, 2026.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

Huanrui Yang, Hongxu Yin, Maying Shen, Pavlo Molchanov, Hai Li, and Jan Kautz. Global vision transformer pruning with hessian-aware saliency. *arXiv preprint arXiv:2110.04869*, 2021.

Haokui Zhang, Congyang Ou, Dawei Yan, Peng Wang, Qingsen Yan, Yu Zhang, Ying Li, and Rong Xiao. TRIO: Token reduction via inference-objective guidance for efficient vision-language models. *arXiv preprint arXiv:2602.04657*, 2026.