WHAT SURVIVES TOKEN MERGING? MERGE-INVARIANT LOW-RANK ATTENTION FOR VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Token reduction and low-rank factorization offer complementary ways to accelerate vision transformers, yet their composition introduces dependencies that neither addresses independently. Factorizing the joint query–key product changes the representations used for token merging, while merging destroys the stable positional correspondence required by conventional logit-residual correction. We resolve both incompatibilities without retraining. First, we characterize fixed logit corrections invariant to patch relabelling: their space is five-dimensional, but two directions are invisible to the softmax, leaving exactly three effective degrees of freedom. These three coefficients per head recover 93.9% and 103.7% of the accuracy recovered by a $38{,}809$-entry position-indexed correction on DeiT-Small and DeiT-Tiny, respectively, while remaining applicable after merging. Second, we decouple token matching from the current query–key factorization by computing similarity from normalized block inputs, improving matching stability under rank reduction. We further identify an accuracy–compute crossover: mild token reduction favors merging alone, whereas more aggressive reduction benefits from combining both compression axes. Beyond this crossover, our method improves accuracy by up to 4.34 percentage points at matched compute and achieves up to $1.06\times$ throughput at matched accuracy, with no retraining or gradient computation. Experiments across DeiT scales, ViT-L, and an alternative token-reduction rule establish that the composition holds beyond a single setting, and results on ADE20K segmentation and five CLIP image encoders examine the broader applicability of low-rank correction. Code and compressed checkpoints will be released to support reproduction and further research.

1 INTRODUCTION

Vision transformers achieve strong performance across recognition and dense prediction, but their inference cost grows rapidly with the number of image tokens (Dosovitskiy, 2020; Touvron et al., 2021; Ranftl et al., 2021). Token reduction addresses this cost by pruning, sampling, or merging intermediate representations (Rao et al., 2021; Liang et al., 2022; Fayyaz et al., 2022; Bolya et al., 2022). Subsequent work has improved compression-rate selection, preservation of token structure, and selection of salient content (Chen et al., 2023; Tran et al., 2024; Bergner et al., 2025; Xie et al., 2026), while recent training-free methods further adapt how tokens are combined (Lee & Choi, 2025; Zeng et al., 2025). These approaches primarily reduce the number of tokens participating in subsequent computation. A complementary opportunity lies in the dimension of the query–key interactions: computing attention scores costs $N^2 d$ multiply–accumulates per head, where N is the token count and d is the head dimension.

Low-rank factorization reduces this second source of cost. The query and key projections determine the logits through their bias-augmented joint product $\tilde{M} = \tilde{W}_q \tilde{W}_k^\top$, whose rank is at most d (Bhojanapalli et al., 2020; Elhage et al., 2021). Approximating this product at rank $r < d$ reduces score computation to $N^2 r$. Joint-product factorization has been studied for vision transformers (Xiao et al., 2023), and activation-aware compression has established the value of accounting for input statistics when selecting low-rank approximations (Yuan et al., 2023; Wang et al., 2025). We study how such approximations compose with token reduction. Although the two reduce different dimen-

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.

sions of attention, their interaction affects both the correction of truncation error and the geometry used to select tokens for merging.

The first interaction concerns the structure of the correction. A mean logit-residual table estimated from calibration data can compensate for systematic truncation error, but its entries refer to particular query–key positions. After input-dependent merging, an index instead denotes a group of patches whose membership varies across images. A position-indexed table therefore has no direct correspondence to the reduced sequence. This raises a structural question: which fixed additive corrections remain well-defined under patch relabelling?

We characterize this class exactly. For a distinguished class token and at least two patch tokens, the fixed matrices satisfying $PCP^\top = C$ for every patch permutation P form a five-dimensional space. Two directions are row-wise shifts and hence leave the softmax unchanged. The resulting quotient has exactly three degrees of freedom: patch self-interaction, patch-to-class interaction, and class self-interaction. These interactions depend on token roles rather than spatial indices, providing a correction that can be applied to sequences of different lengths. On DeiT-Small and DeiT-Tiny at rank 32, three coefficients per head recover 93.9% and 103.7%, respectively, of the accuracy gain obtained with a $38{,}809$-entry position-indexed table. The characterization also makes a prediction that admits no tolerance: one direction lies in the softmax gauge, so adding it with any coefficient must leave every prediction unchanged. It alters none of $45{,}000$ test predictions, in every backbone and rank configuration we evaluate.

The second interaction concerns token similarity. Key-space matching, as used in ToMe (Bolya et al., 2022), depends on representations that factorization itself changes. For any invertible R, $AB^\top = (AR)(BR^{-\top})^\top$: identical attention logits need not imply identical key-space cosine similarities, so even an exact refactorization can change merge decisions. We instead compute similarity from the normalized block input, removing its direct dependence on the current block's query–key factors.

Together, the structured correction and block-input matching enable a calibrated composition without retraining or gradient computation. The two reductions have opposite cost profiles, and the composition is worthwhile only past their crossover. Below it, a compute budget is better spent on token reduction alone, and we report the settings where our composition is behind. Above it, gains reach 4.34 percentage points over token reduction alone at matched model FLOPs, and 2.67 points, or 1.06 times the throughput at equal accuracy, on locked-clock hardware. We evaluate composition across DeiT scales and ViT-L, with both merging and token reorganization, and measure end-to-end throughput alongside model FLOPs. Separate evaluations on ADE20K segmentation and CLIP image encoders examine the applicability of rank reduction and correction beyond image classification (Ranftl et al., 2021; Radford et al., 2021). **Our core contributions are:**

- **A characterization of position-free logit correction.** We prove that fixed corrections invariant to patch relabelling have exactly three degrees of freedom modulo the softmax gauge (Proposition 1), and derive their transformation under token averaging (Proposition 2).
- **An analysis of factorization-dependent matching.** We identify why preserving $AB^\top$ does not preserve key-space cosine similarity, and use normalized block-input features to remove the direct dependence of matching on the chosen query–key factorization.
- **An empirical account of when the two reductions compose.** We locate the crossover past which combining them improves the accuracy–cost trade-off, and evaluate the composition across model scales, token-reduction rules, and compression budgets using accuracy, model FLOPs, and measured throughput, with additional tests on dense prediction and vision–language encoders.

2 Related work

Token reduction. Vision transformer acceleration commonly reduces sequence length through learned pruning (Rao et al., 2021; Bergner et al., 2025), attention-guided reorganization (Liang et al., 2022), or adaptive sampling (Fayyaz et al., 2022). ToMe (Bolya et al., 2022) instead merges similar tokens through bipartite matching without retraining. Subsequent methods optimize layer-wise reduction rates (Chen et al., 2023), preserve structural information (Tran et al., 2024), incorporate saliency (Xie et al., 2026), or improve training-free merging (Lee & Choi, 2025; Zeng et al.,

2025). These methods primarily reduce sequence length N, while retaining the original query–key dimensionality. In similarity-based merging, the representations used for matching are themselves produced by the projections being compressed.

Low-rank attention and factorization. Low-rank attention methods approximate the sequence-level attention operator through learned projections or structured approximations (Wang et al., 2020; Choromanski et al., 2020; Xiong et al., 2021). A complementary direction exploits redundancy in the query–key projections (Bhojanapalli et al., 2020; Elhage et al., 2021). COMCAT (Xiao et al., 2023) investigates low-rank compression of attention-based vision models. For language models, Fisher-weighted and activation-aware factorization account for input-dependent approximation error (Hsu et al., 2022; Yuan et al., 2023), while SVD-LLM (Wang et al., 2025) additionally uses parameter updates to compensate for truncation. Our approach instead factorizes the joint query–key product using calibration statistics and compensates for systematic error directly in the attention logits. This introduces a distinct challenge under token merging: a correction indexed by the original token positions cannot be transferred unchanged to image-dependent groups of patches.

Composing compression axes. Token reduction and query–key factorization act on complementary dimensions of attention, N and d, but their composition introduces two dependencies. First, a fixed position-indexed logit correction loses its correspondence to patches after input-dependent merging. Second, preserving the query–key product does not preserve cosine similarity between its factors, so even exact refactorization can change key-based merge decisions. We characterize the fixed corrections invariant to patch relabelling, modulo softmax-invisible row shifts, and decouple matching from the current block's query–key factors using normalized block inputs. We then establish empirically when combining the two reductions improves the accuracy–compute trade-off over token reduction alone.

3 METHOD

We combine query–key rank reduction with token merging to reduce computation in pretrained vision transformers without retraining. Rank reduction lowers the cost of each attention score; merging reduces the number of tokens. We couple the two with a position-free logit correction and compute merge similarities from the normalized block input. Figure 1 summarizes the method.

3.1 SETUP

Let $X \in \mathbb{R}^{N\times D}$ enter a transformer block with h heads of dimension $d = D/h$. Write $\bar{X} = \mathrm{LN}(X)$ and augment it as $\tilde{X} = [\bar{X}\ \mathbf{1}_N] \in \mathbb{R}^{N\times(D+1)}$ to absorb projection biases. For head i, define $\tilde{W}_q^{(i)} = [W_q^{(i)\top}\ b_q^{(i)}]^\top \in \mathbb{R}^{(D+1)\times d}$ and $\tilde{W}_k^{(i)}$ analogously, with biases written as column vectors. Position 0 is the class token; $1,\dots,p$ with $p = N-1$ are patch tokens. Let e_j denote the standard basis vectors of $\mathbb{R}^N$ and $v_p = \sum_{j=1}^{p} e_j$ the patch indicator.

3.2 TRUNCATING THE QUERY–KEY PRODUCT

Queries and keys determine the attention scores through their product (Bhojanapalli et al., 2020; Elhage et al., 2021):

$$L^{(i)} = s\tilde{X}\tilde{M}^{(i)}\tilde{X}^\top, \qquad \tilde{M}^{(i)} = \tilde{W}_q^{(i)}\tilde{W}_k^{(i)\top}, \qquad s = d^{-1/2}. \tag{1}$$

Since $\mathrm{rank}(\tilde{M}^{(i)}) \le d$, we approximate it by $A^{(i)}B^{(i)\top}$ with $A^{(i)}, B^{(i)} \in \mathbb{R}^{(D+1)\times r}$ and $r \le d$:

$$\hat{L}^{(i)} = s(\tilde{X}A^{(i)})(\tilde{X}B^{(i)})^\top. \tag{2}$$

The scale remains $d^{-1/2}$ and the value dimension is unchanged. Per head, score computation decreases from N^2d to N^2r multiply–accumulates and query/key projections from $2NDd$ to $2NDr$.

Fitting the factorization. We weight the approximation by calibration activations, following activation-aware compression (Yuan et al., 2023; Wang et al., 2025). Suppress the head index and let Σ_0 be the empirical second moment of augmented normalized tokens. For independent draws $\tilde{x}, \tilde{x}'$ from this token pool and $E = \tilde{M} - \hat{M}$,

$$\mathbb{E}\big[(\tilde{x}^\top E\tilde{x}')^2\big] = \mathrm{tr}(E\Sigma_0 E^\top\Sigma_0) = \|S_0 E S_0^\top\|_F^2, \qquad \Sigma_0 = S_0^\top S_0. \tag{3}$$

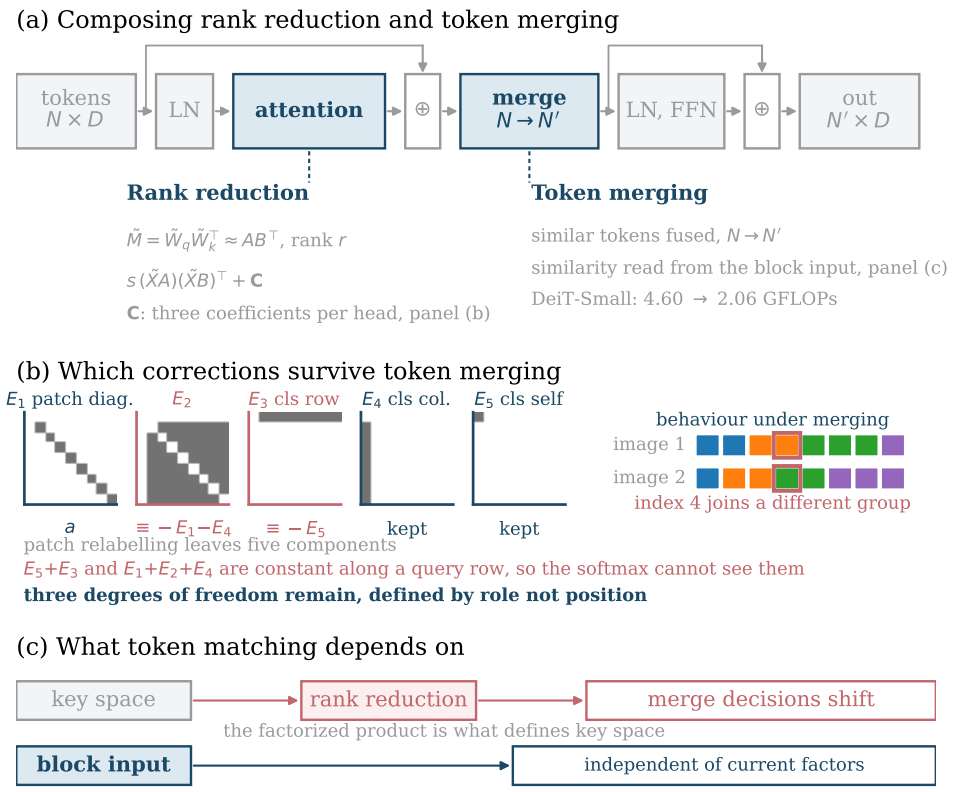

Figure 1: **Composing rank reduction and token merging.** **(a)** A calibrated rank-r factorization replaces the query–key product; token merging follows attention. **(b)** Patch relabelling leaves five components. Two combinations, $E_5{+}E_3$ and $E_1{+}E_2{+}E_4$, are constant along a query row and invisible to the softmax, leaving three degrees of freedom defined by role rather than position. **(c)** Matching on the block input removes the dependence on the current block's factors.

This independent-token objective is a surrogate for logit reconstruction error, since tokens within an image are dependent. We adopt it because it is minimized in closed form and because the subspace it selects is worth 6.60 points on DeiT-Small at $r = 16$ over activation-agnostic SVD (Table 17).

We floor the eigenvalues of Σ_0 at 10^{-6}, giving $\Sigma_\epsilon = S^\top S$ with S invertible, and minimize $\|S(\tilde{M} - \hat{M})S^\top\|_F^2$ over rank at most r. The closed-form solution is

$$U_r \Lambda_r V_r^\top = (S\tilde{M}S^\top)_r, \qquad A = S^{-1}U_r\Lambda_r^{1/2}, \qquad B = S^{-1}V_r\Lambda_r^{1/2}, \tag{4}$$

where $(\cdot)_r$ denotes a truncated SVD (Appendix A.1). The factors are fitted offline without gradient updates.

3.3 Position-free logit correction

Let Δ denote the mean calibration residual $L - \hat{L}$ for one head. A fixed $N \times N$ table corrects this residual at the original token positions. After merging, the same index can represent different patch groups across images, so the table cannot be reused unchanged. We instead consider fixed, image-independent corrections satisfying

$$PCP^\top = C \qquad \text{for every permutation } P \text{ of the patch tokens,} \tag{5}$$

which we call *position-free*.

Proposition 1 (The position-free correction space)**.** *For $p \geq 2$,* (i) *the matrices satisfying Equation 5 form the five-dimensional space spanned by*

$$E_1 = \sum_{j>0} e_j e_j^\top, \quad E_2 = v_p v_p^\top - E_1, \quad E_3 = e_0 v_p^\top, \quad E_4 = v_p e_0^\top, \quad E_5 = e_0 e_0^\top. \tag{6}$$

(ii) *Its softmax-invisible row shifts span* $\{E_5 + E_3,\ E_1 + E_2 + E_4\}$. *The quotient has dimension three, with basis* $\{[E_1], [E_4], [E_5]\}$.

We therefore use, for each head,

$$C = aE_1 + bE_4 + cE_5, \tag{7}$$

three coefficients adjusting patch self-scores, patch-to-class scores and the class self-score, at $O(N)$ application cost.

Fitting the coefficients. We project Δ onto the invariant space and remove its row shifts. Equivalently, with $\mu_u = (p-1)^{-1} \sum_{j>0,\, j\neq u} \Delta_{uj}$ for patch rows and $\mu_0 = p^{-1} \sum_{j>0} \Delta_{0j}$,

$$a = \frac{1}{p} \sum_{u>0} (\Delta_{uu} - \mu_u), \qquad b = \frac{1}{p} \sum_{u>0} (\Delta_{u0} - \mu_u), \qquad c = \Delta_{00} - \mu_0. \tag{8}$$

The estimate is invariant to per-row constant offsets (Appendix A.2); direct Frobenius projection onto $\mathrm{span}\{E_1, E_4, E_5\}$ would not implement this gauge choice. Coefficients are calibrated on the unmerged model with truncation installed in all blocks (Appendix B.2).

Transfer across sequence lengths. At inference we reuse a, b, c and construct the correction at the current token count. The following describes how a fixed correction behaves under the averaging that merging performs.

Proposition 2 (Transport under averaging)**.** *Let T average each group of a partition π uniformly, with the class token a singleton and $s_j = |\pi_j|$. For $C = aE_1 + bE_4 + cE_5$, the transported correction $TCT^\top$ preserves b and c and scales the diagonal coefficient of group j to a/s_j.*

Transport respects softmax equivalence, so the result does not depend on the representative chosen, and generalizes to the size-weighted averaging we apply (Appendix A.3). It does not establish optimality for the merged network, since averaging does not commute with layer normalization. The size-scaling ablation shows differences of mixed sign and within split resolution (Appendix D.4). We apply the scaling of Proposition 2 wherever tokens are merged; without reduction every group has $s_j = 1$ and the scaling is the identity.

3.4 Where token similarity is measured

ToMe matches tokens by cosine similarity of head-averaged keys (Bolya et al., 2022). However, $AB^\top = (AR)(BR^{-\top})^\top$ for any invertible R: attention scores can remain unchanged while key-space cosine similarities change. Even an exact factorization therefore need not preserve key-based merge decisions.

We instead match using all D channels of $\bar{X}$, avoiding direct dependence on the current block's query–key factorization. Section 4.5 evaluates the resulting merge stability. We retain ToMe's bipartite matching and size-weighted averaging, merging after the attention residual addition and before the MLP, with the class token protected. Token sizes accumulate across blocks. For incoming sizes m_j the corrected logits are $\hat{L}_{uj} + C_{uj} + \log m_j$: the correction precedes the key-axis size bias and the row-wise softmax.

3.5 Cost

For a block receiving N tokens and emitting N', rank reduction changes score computation from $N^2 D$ to $N^2 rh$ and query/key projection from $2ND^2$ to $2NDrh$. The value and output projections precede merging and retain N; the MLP and subsequent blocks benefit from N'. The attention–value product costs $N^2 D$ and is unaffected by r, which is why rank reduction alone saturates near 17% of model compute on DeiT-Small and why composing rather than reducing rank alone is the operating point. We count one multiply–accumulate as one FLOP.

Reported FLOPs exclude matching, as in Bolya et al. (2022). Its similarity matrix costs $n_a n_b c_{\mathrm{match}}$ for bipartite sets of sizes n_a, n_b, with $c_{\mathrm{match}} = D$ for block inputs, d for original head-averaged keys, and r for compressed ones. On DeiT-Small at $N = 197$ this is 24.7% of the attention scores for block inputs against 4.1% for keys, so block-input matching carries an additional cost not reflected in the FLOP comparisons. End-to-end throughput measurements include matching and merging.

Table 1: Correction decomposition without token reduction, test split. *recov.* is the share of the truncation gap closed, relative to the full N^2-entry table. All eight configurations appear in Table 9.

	Tiny, r=32		Small, r=32		Base, r=12	
correction	top-1	recov.	top-1	recov.	top-1	recov.
dense	72.19	–	79.80	–	81.91	–
truncated	70.49	0.0%	78.94	0.0%	76.92	0.0%
+ row shift	70.49	0.0%	78.94	0.0%	76.92	0.0%
+ E_1 (patch diagonal)	70.90	84.4%	79.09	42.4%	77.92	52.9%
+ E_4 (class column)	70.50	3.7%	78.99	15.2%	77.04	6.3%
+ E_5 (class self)	70.55	13.8%	79.15	58.8%	77.00	3.8%
+ E_1, E_4	70.96	98.2%	79.10	46.1%	78.02	58.2%
+ E_1, E_4, E_5 (ours)	**70.99**	**103.7%**	**79.28**	**93.9%**	**78.07**	**60.4%**
+ full table	70.97	100%	79.30	100%	78.82	100%

4 Experiments

4.1 Setup

We evaluate on ImageNet-1k Russakovsky et al. (2015) with DeiT-Tiny, Small and Base (Touvron et al., 2021) and ViT-L/16 at 224 and 384 pixels (Dosovitskiy, 2020). The validation set is split once into 5,000 development and 45,000 test images, with calibration drawn from the training split; results are on the test split unless marked. Paired comparisons use McNemar's exact test, reported with discordant counts and two-sided p-values. Throughput is fp16 on an RTX 3090 with locked clocks. FLOP counts follow Bolya et al. (2022) and exclude token matching, which throughput includes; Appendix B gives the full protocol.

4.2 Where each reduction is expensive

Figure 2 plots the accuracy given up per GFLOP saved as each reduction is pushed further. Token reduction is inexpensive at first and then steep, rising from 2.7 points per GFLOP at moderate rates to 46 at rate 24; rank reduction is free to $r = 48$ and rises to 36 by $r = 16$, its cost falling with width (Table 8). The two are steep in different regions, which is the condition under which a budget is better divided than spent on either alone.

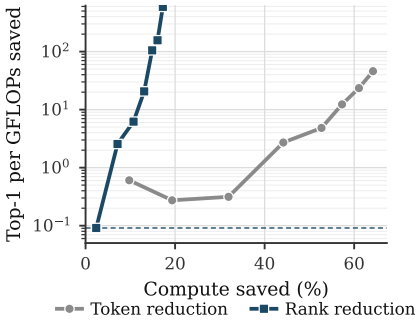

Figure 2: Top-1 given up per GFLOP saved, DeiT-Small, against the fraction of model compute removed.

4.3 Three coefficients replace a position-indexed table

Proposition 1 gives a five-dimensional invariant space with two softmax-invisible directions, leaving three degrees of freedom. Table 1 tests both: a row shift from that subspace alters $n_{01} = n_{10} = 0$ predictions in all eight backbone–rank configurations, on 45,000 images each, while three coefficients recover 103.7% of what the 38,809-entry table recovers on DeiT-Tiny and 93.9% on DeiT-Small ($p = 1{\times}10^{-7}$), the weight carried by the patch diagonal on Tiny and the class self-score on Small. The correction repairs the error rank reduction introduces, so its effect scales with that error: on DeiT-Base the recovered share rises monotonically from -60.0% at $r = 32$, where the gap is 0.40 points and nothing reaches significance, to 60.4% at $r = 12$, where it is 4.99 (Table 9). Coefficients calibrated without reduction transfer to reduced models, gaining $+0.23$ to $+0.85$ points at reduction rate 16 wherever significant (Table 10); a position-indexed table calibrated at rate 0 and applied at rate 20 changes as many predictions in each direction as it leaves alone.

4.4 Composing the two reductions

Table 2 compares the composition against token reduction alone at matched compute. The crossover depends on backbone and rank: at mild reduction rates, the composition provides little benefit or trails token reduction alone, while its advantage generally grows as reduction becomes more aggressive. At rate 18, gains reach $+4.34$ points on DeiT-Tiny and $+3.17$ on DeiT-Base. thereafter to $+4.34$ on DeiT-Tiny and $+3.17$ on DeiT-Base (Figure 3). It holds at 305M parameters: ViT-L/16-384 at rate 22, r=48 reaches 87.24 at 95.3 GFLOPs against 86.92 at 100.2 (Table 13).

Table 2: Top-1 change against token reduction alone at equal compute, test split; the baseline matches on the block input. Absolute accuracies and compute in Tables 11 and 12.

	DeiT-Small				DeiT-Tiny		DeiT-Base
rate	r=48	r=40	r=32	r=24	r=48	r=32	r=48
0	+0.00	−0.11	−0.40	−1.11	−0.19	−0.76	+0.04
8	+0.03	+0.03	−0.16	−1.02	+0.01	−0.57	+0.20
12	+0.19	+0.17	+0.15	−0.56	+0.18	+0.01	+0.38
16	+0.60	+0.84	+1.06	+0.69	+1.12	+2.12	+1.42
18	+1.17	+1.41	+1.93	+2.90	+2.11	+4.34	+3.17

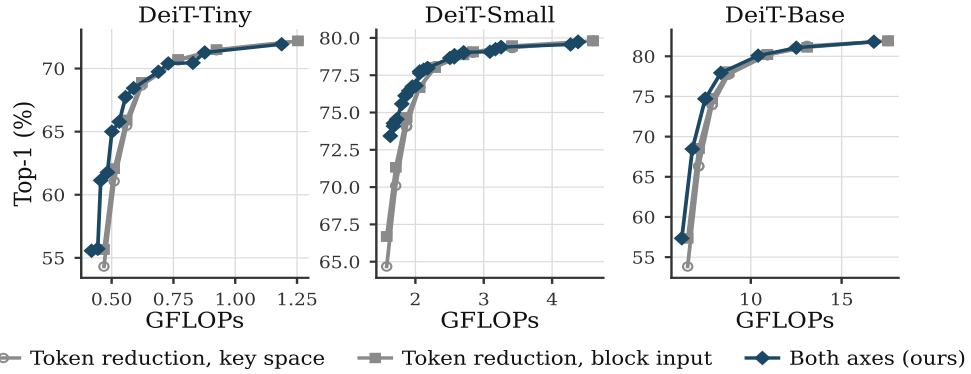

Figure 3: Accuracy against compute, test split. Token reduction alone under both similarity sources, and the front of the composition over ranks 24 to 48.

Matched throughput. FLOPs are not time: at batch 1 all three DeiT models take 3.5 ms despite a 14× range in arithmetic. Table 3 repeats the comparison against measured images per second, including the matching cost that FLOP counts omit, and the crossover survives. Rank reduction exceeds its own FLOP saving in wall clock, 12.0% more throughput at rate 0 and $r = 32$ for a 9.5% reduction in multiply–accumulates; applying the correction costs 1.9–2.1%.

Table 3: DeiT-Small at matched throughput, development split. Against token reduction alone: Δtop-1 at equal img/s, *speed-up* at equal top-1.

setting		measured		vs. baseline	
rate	r	top-1	img/s	Δtop-1	speed-up
12	40	79.58	3759	+0.12	1.02×
16	40	78.58	4538	+0.45	1.04×
18	32	77.02	5158	+1.74	1.06×
20	32	74.52	5572	+2.67	1.06×

4.5 Where token similarity is measured

Section 3.4 argues that truncation reshapes key space and that merge decisions computed there move with it. Table 4 varies only the tensor the similarity is read from. The block input is best at every rate, by +0.84, +1.84 and +1.14 points over the default at equal compute, and the $r = 64$ block isolates the effect from compression: with truncation exact, the change alone improves token reduction by +0.32, +0.30 and +0.20. The benefit tracks the reduction rate and is absent at mild rates, running from −0.18 at rate 8 to +2.19 at rate 20 on DeiT-Base.

Figure 4 measures the stability behind those numbers. Comparing each similarity source against its own grouping at full rank, key-space matching loses 58–64% of its agreement between $r = 48$ and $r = 8$ while block-input matching loses 31–41%, and is more stable at all eighteen combinations of rank and reduction rate.

4.6 Comparison with prior work

Table 5 groups post-hoc compression of DeiT-S by what each method requires beyond a pretrained checkpoint. Within attention factorization, the retained subspace matters more than the low-rank structure: a random subspace falls to 2–3% (Table 17), while activation-aware selection reaches 78.94% and the three coefficients further improve accuracy by 0.34 points to 79.28% at $r = 32$. Without reduction, the position-indexed table and the three coefficients are equivalent (79.30 against 79.28, Table 1); under reduction only the three remain defined. Using both axes, reduction rate 8 with $r = 48$ reaches 79.38% at 3.25 GFLOPs, while rate 13 with $r = 48$ reaches 78.85% at 2.57

Table 4: Similarity source, DeiT-Small, development split. Compute is identical within each block except where noted. At r=64 truncation is exact, isolating the similarity source.

		top-1 at reduction rate		
r	similarity read from	16	18	20
32	compressed keys (default)	77.36	75.18	73.38
	uncompressed keys†	78.14	76.44	74.26
	values	77.52	76.14	73.72
	block input (ours)	**78.20**	**77.02**	**74.52**
64	key space	78.30	77.18	74.82
	block input (ours)	**78.62**	**77.48**	**75.02**

†retains the full key projection; 2.256 rather than 2.064 GFLOPs at rate 16.

Table 5: Token reduction and attention factorization on ImageNet-1K, grouped by optimization requirements. N and d mark the reduced dimension of attention. Training-free ToMe and ToFu results are reported on AugReg ViT-S, so Δdense gives the drop from each method's own baseline. †reimplemented here.

Method	Venue	Backbone	N	d	Top-1 (%)	Δdense	GFLOPs
DeiT-S (Touvron et al., 2021)	ICML'21	DeiT-S			79.80	–	4.60
AugReg ViT-S (Steiner et al., 2021)	TMLR'22	ViT-S			81.40	–	4.60
Token reduction with backbone training or fine-tuning							
DynamicViT (Rao et al., 2021)	NeurIPS'21	DeiT-S	✓		79.30	−0.50	2.90
A-ViT (Yin et al., 2022)	CVPR'22	DeiT-S	✓		78.60	−1.20	2.90
SP-ViT (Kong et al., 2021)	ECCV'22	DeiT-S	✓		79.30	−0.50	2.60
EViT (Liang et al., 2022)	ICLR'22	DeiT-S	✓		79.50	−0.30	3.00
ToMe (Bolya et al., 2022)	ICLR'23	DeiT-S	✓		79.40	−0.40	2.70
BAT (Long et al., 2023)	CVPR'23	DeiT-S	✓		79.60	−0.20	3.00
MCTF (Lee et al., 2024)	CVPR'24	DeiT-S	✓		80.10	+0.30	2.60
Token reduction with frozen backbone and learned compression rates							
DiffRate (Chen et al., 2023)	ICCV'23	DeiT-S	✓		79.54	−0.26	2.91
Token reduction without gradient computation							
Zero-TPrune (Wang et al., 2024)	CVPR'24	DeiT-S	✓		79.40	−0.40	2.97
ToMe (Bolya et al., 2022)	ICLR'23	ViT-S	✓		79.30	−2.10	2.70
ToFu (Kim et al., 2024)	WACV'24	ViT-S	✓		79.60	−1.80	2.70
ToMe, our reproduction	–	DeiT-S	✓		78.98	−0.82	2.70
Attention factorization without gradient computation							
SVD†	–	DeiT-S		✓	78.38	−1.42	4.16
ASVD† (Yuan et al., 2023)	arXiv'23	DeiT-S		✓	78.94	−0.86	4.16
SVD-LLM† (Wang et al., 2025)	ICLR'25	DeiT-S		✓	78.94	−0.86	4.16
Ours, r=32	–	DeiT-S		✓	79.28	−0.52	4.16
Both axes, without gradient computation							
Ours, rate 8, r=48	–	DeiT-S	✓	✓	79.38	−0.42	3.25
Ours, rate 13, r=48	–	DeiT-S	✓	✓	78.85	−0.95	2.57

GFLOPs. Methods that fine-tune the backbone remain ahead at the lowest budgets, the expected cost of requiring no gradients.

4.7 Beyond classification

Three settings test whether the components generalize. EViT (Liang et al., 2022) ranks patches by class-token attention and fuses the discarded ones, a criterion unrelated to pairwise similarity; composed with $r = 32$ and the correction it gains $+1.43$ and $+3.29$ points on DeiT-Tiny at keep rates 0.70 and 0.60, with the crossover depending on backbone and rank (Table 16). Rank reduction removes no token, so it applies where token reduction does not: DPT fuses features from four encoder taps, each requiring full spatial coverage, yet ADE20K Zhou et al. (2017) mIoU is unchanged at $r = 48$ and falls by 0.22 at $r = 32$ (Table 6), with whitening worth $+1.50$ mIoU at $r = 24$ and $+5.78$ at $r = 16$ over an activation-agnostic SVD (Table 14). Across five CLIP towers spanning three scales and two pretraining corpora, the correction is positive at nineteen of twenty settings,

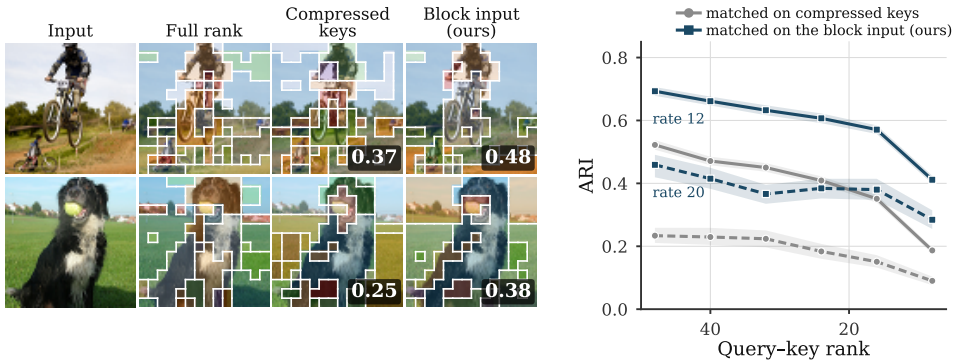

Figure 4: **Rank reduction moves the merge decision.** Left: patches sharing a colour were fused into one token; DeiT-Small at reduction rate 16, r=32. The number is the adjusted Rand index against the full-rank grouping in the second column. Right: the same comparison over 256 development images, at two reduction rates, with bootstrap 95% intervals. Further images in Figure 5.

Table 6: Rank reduction beyond classification. Upper: ADE20K mIoU with DPT-Large, single-scale 480×480, 2,000 validation images; encoder compute saved is 5.2%, 10.4%, 13.1% and 15.7% at the four ranks. Lower: CLIP zero-shot ImageNet top-1; vision-tower compute saved is 4.5%, 9.0%, 11.2% and 13.5%. Each cell is without and with the three-coefficient correction. Full sweeps in Tables 14 and 15.

		r=48		r=32		r=24		r=16	
	dense	plain	$+C$	plain	$+C$	plain	$+C$	plain	$+C$
ADE20K semantic segmentation, mIoU									
DPT-Large	45.92	45.93	45.93	45.66	45.70	45.02	45.20	42.04	42.70
CLIP zero-shot ImageNet, top-1									
ViT-B/32	63.76	63.42	63.40	61.42	62.02	59.42	61.26	53.50	58.42
ViT-B/16	69.18	68.58	68.66	66.78	67.62	63.78	65.28	55.68	59.94
ViT-L/14	76.04	75.96	76.16	74.20	74.58	72.30	72.90	66.14	68.76
ViT-L/14-336	76.92	76.64	76.74	74.80	75.10	72.68	73.64	66.74	69.30
ViT-L/14 LAION	75.44	75.32	75.36	73.76	74.28	72.10	72.50	66.42	68.72

growing with compression and larger on narrower towers as on DeiT, reaching +4.92 on ViT-B/32 and +2.56 on the ViT-L/14-336 tower used by LLaVA-1.5 (Liu et al., 2024).

Scope. The correction is measurable in proportion to the truncation error it repairs, so at ranks where factorization is already free there is little to recover, and the composition is worthwhile only past the crossover of Section 4.4. The method operates within a chosen model: on ImageNet, DeiT-Base under aggressive compression does not overtake dense DeiT-Small.

5 Conclusion

We investigated when token reduction and query–key rank reduction can be combined in pretrained vision transformers. Although they compress different dimensions of attention, their interaction is not independent: merging invalidates position-indexed logit corrections, while factorization can change the key-space similarities that determine which tokens are merged. We addressed the first by characterizing all fixed corrections invariant to patch relabelling, which modulo softmax-invisible row shifts leave exactly three degrees of freedom and yield a compact correction that remains defined as token identities and sequence lengths change; and the second by computing merge similarities from normalized block inputs, eliminating their dependence on the current factorization. Together these enable composition without retraining, though not universally: our experiments locate a crossover between the marginal costs of the two reductions, below which merging alone is preferable and above which their combination improves the accuracy–compute frontier on every backbone evaluated, an advantage that persists in measured throughput despite the additional cost of block-input matching. The broader implication is that compression methods acting on distinct dimensions need not preserve each other's assumptions; our characterization concerns permutation equivariance rather than images, leaving open which other inference-time sequence transformations admit similarly compact corrections.

AI USE STATEMENT

We used a general-purpose large language model as a writing assistant, for improving clarity, concision and organization of the text. All statements, claims and numerical results were written and verified by the authors against the underlying experimental logs, and the model was not used for research ideation, experimental design, or the generation or modification of any result. We have reviewed all AI-assisted text and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work compresses publicly available pretrained vision models and evaluates them on public benchmarks. It involves no human subjects, no new data collection, and no release of personal or sensitive data. We foresee no direct ethical concerns beyond those already attached to the underlying models and datasets, whose known biases our method neither introduces nor mitigates.

REPRODUCIBILITY STATEMENT

The method requires no training, so reproduction reduces to applying a closed-form factorization and three scalar coefficients per head to public checkpoints. Appendix B specifies the split construction, calibration protocol, numerical precision, statistical tests and timing methodology, and Table 7 lists every fixed setting and symbol. Every run asserts two exactness identities before producing numbers, described in the same appendix. Complete proofs of both propositions, including the assumptions each requires, appear in Appendix A. Full sweeps behind every main-paper table are given in Appendices C and D. Code and the compressed checkpoints will be released.

REFERENCES

Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Token cropr: Faster vits for quite a few tasks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9740–9750. IEEE, 2025.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*, pp. 864–873. PMLR, 2020.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Mengzhao Chen, Wenqi Shao, Peng Xu, Mingbao Lin, Kaipeng Zhang, Fei Chao, Rongrong Ji, Yu Qiao, and Ping Luo. Diffrate: Differentiable compression rate for efficient vision transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17118–17128. IEEE, 2023.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *European conference on computer vision*, pp. 396–414. Springer, 2022.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. *arXiv preprint arXiv:2207.00112*, 2022.

Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1372–1381. IEEE, 2024.

Zhenglun Kong et al. Spvit: enabling faster vision transformers via soft token pruning (2022). *URL https://arxiv. org/abs/2112.13890*, 2, 2021.

Jaeyeon Lee and Dong-Wan Choi. Lossless token merging even without fine-tuning in vision transformers. *arXiv preprint arXiv:2505.15160*, 2025.

Sanghyeok Lee et al. Multi-criteria token fusion with one-step-ahead attention for efficient vision transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15741–15750. IEEE, 2024.

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. *arXiv preprint arXiv:2202.07800*, 2022.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Sifan Long et al. Beyond attentive tokens: Incorporating token importance and diversity for efficient vision transformers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10334–10343. IEEE, 2023.

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

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

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Olga Russakovsky et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Hoai-Chau Tran, Duy Nguyen, Duy M Nguyen, TrungTin Nguyen, Ngan Le, Pengtao Xie, Daniel Sonntag, James Zou, Binh T Nguyen, and Mathias Niepert. Accelerating transformers with spectrum-preserving token merging. *Advances in Neural Information Processing Systems*, 37: 30772–30810, 2024.

Hongjie Wang et al. Zero-tprune: Zero-shot token pruning through leveraging of the attention graph in pre-trained transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16070–16079. IEEE, 2024.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression. In *International Conference on Learning Representations*, volume 2025, pp. 19299–19319, 2025.

Jinqi Xiao, Miao Yin, Yu Gong, Xiao Zang, Jian Ren, and Bo Yuan. Comcat: Towards efficient compression and customization of attention-based vision models. *arXiv preprint arXiv:2305.17235*, 2023.

Weiying Xie, Xiaoyu Chen, Xin Zhang, Chenhe Hao, Jitao Ma, Yunsong Li, and Leyuan Fang. Saliency-driven token merging for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32184–32193, 2026.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14138–14148, 2021.

Hongxu Yin et al. A-vit: Adaptive tokens for efficient vision transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10799–10808. IEEE, 2022.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*, 2023.

Fanhu Zeng, Deli Yu, Zhenglun Kong, and Hao Tang. Token transforming: A unified and training-free token compression framework for vision transformer acceleration. *arXiv preprint arXiv:2506.05709*, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5122–5130. IEEE, 2017.

A MATHEMATICAL DETAILS

Equation 1 follows by writing $\tilde{x}_u = [\bar{x}_u^\top \; 1]^\top$ and $\tilde{W}_q = [W_q^\top \; b_q]^\top$, so that $s(\bar{x}_u^\top W_q + b_q^\top)(\bar{x}_v^\top W_k + b_k^\top)^\top = s\,\tilde{x}_u^\top \tilde{W}_q \tilde{W}_k^\top \tilde{x}_v$; the rank bound holds because both factors have d columns.

A.1 ACTIVATION-WEIGHTED APPROXIMATION

For independent draws $\tilde{x}, \tilde{x}'$ with second moment $\Sigma_0 = S_0^\top S_0$ and fixed $E = \tilde{M} - \hat{M}$, taking the inner expectation first gives

$$\mathbb{E}\big[(\tilde{x}^\top E \tilde{x}')^2\big] = \mathbb{E}_{\tilde{x}}\big[\tilde{x}^\top E \Sigma_0 E^\top \tilde{x}\big] = \operatorname{tr}(E\Sigma_0 E^\top \Sigma_0) = \|S_0 E S_0^\top\|_F^2. \tag{9}$$

Flooring the eigenvalues of Σ_0 at ϵ yields a positive definite $\Sigma_\epsilon = S^\top S$ with S invertible and the objective $\|S(\tilde{M} - \hat{M})S^\top\|_F^2$. Writing $G = S\tilde{M}S^\top$, the map $\hat{M} \mapsto S\hat{M}S^\top$ is a bijection on matrices of rank at most r, so Eckart–Young gives $S\hat{M}S^\top = U_r \Lambda_r V_r^\top$ and hence the factors of Equation 4. When $r \geq \operatorname{rank}(\tilde{M})$ the reconstruction is exact.

A.2 POSITION-FREE CORRECTIONS

Proof of Proposition 1. Write a patch permutation as $P = \operatorname{diag}(1, Q)$ and partition C into class and patch blocks,

$$C = \begin{pmatrix} \gamma & w^\top \\ z & Y \end{pmatrix}. \tag{10}$$

Then $PCP^\top = C$ for every Q is equivalent to $Qw = w$, $Qz = z$ and $QYQ^\top = Y$, with γ free, so w and z are constant vectors. Simultaneous permutations act on the patch block with two orbits, diagonal and ordered off-diagonal entries, so Y takes one value on each. These five free values give the basis of Equation 6, whose members have disjoint nonempty supports for $p \geq 2$.

Two logit vectors have the same softmax if and only if they differ by a constant, so the additive matrices invisible to row-wise softmax are exactly $u\mathbf{1}_N^\top$. Such a matrix satisfies Equation 5 exactly when $u = (\alpha, \beta, \ldots, \beta)^\top$, giving

$$u\mathbf{1}_N^\top = \alpha(E_5 + E_3) + \beta(E_1 + E_2 + E_4). \tag{11}$$

The invariant gauge subspace therefore has dimension two and the quotient dimension three, with $[E_3] = -[E_5]$ and $[E_2] = -[E_1] - [E_4]$, so $[E_1], [E_4], [E_5]$ span it. They are independent: if $aE_1 + bE_4 + cE_5 = u\mathbf{1}_N^\top$ is invariant, an off-diagonal patch entry forces $\beta = 0$, whence $a = b = 0$; a class-to-patch entry forces $\alpha = 0$, whence $c = 0$. □

Gauge invariance of the estimator. Equation 8 subtracts each row's off-diagonal patch mean before reading a coefficient, so adding any row shift $u\mathbf{1}_N^\top$ to Δ leaves it unchanged: each patch row's diagonal entry, class-column entry and off-diagonal mean all increase by u_i, and the class self-entry and class-to-patch mean both increase by u_0. Direct Frobenius projection onto $\operatorname{span}\{E_1, E_4, E_5\}$ would not have this property, since those matrices are not orthogonal to the gauge subspace.

A.3 TRANSPORT UNDER AVERAGING

Proof of Proposition 2. By definition $(TCT^\top)_{jk} = (s_j s_k)^{-1} \sum_{u \in \pi_j} \sum_{v \in \pi_k} C_{uv}$. The class group is a singleton, so the class self-entry remains c. For a patch group j and the class column, the sum contains s_j copies of b, giving b after division by s_j. For a patch group's self-entry, s_j of the s_j^2 ordered pairs are diagonal and contribute a, giving a/s_j. Distinct groups have disjoint supports, so all remaining entries vanish. □

Because $T\mathbf{1}_N = \mathbf{1}_{N'}$, transport maps row shifts to row shifts: $T(C + u\mathbf{1}_N^\top)T^\top = TCT^\top + (Tu)\mathbf{1}_{N'}^\top$, so the transported correction does not depend on which representative of the quotient is used. Under the size-weighted averaging we apply, the diagonal multiplier is $\sum_{u \in \pi_j} m_u^2 / (\sum_{u \in \pi_j} m_u)^2$ for current token sizes m_u, reducing to $1/s_j$ when those sizes are equal. Proposition 2 concerns transport of the correction alone and does not imply optimality for the merged network, since averaging does not commute with layer normalization.

B EXPERIMENTAL DETAILS

B.1 PROTOCOL

The ImageNet-1k validation set is partitioned once with a fixed seed into $5{,}000$ development and $45{,}000$ test images. Calibration sets are drawn from the training split, so nothing the compression is fitted on is scored. All index lists are stored and hash-verified at the start of every run. Our dense baselines match published top-1 to within 0.2 points on all five backbones.

The second moment Σ_0 is accumulated in float64 at the input to each block's attention. Equation 4 is computed in float64, batched over the heads of a block, using triangular solves rather than forming S^{-1}. Correction coefficients are calibrated on a held-out subset of the calibration split, on the unmerged model with truncation installed in all blocks, so the residual at block k reflects truncated blocks $1, \ldots, k-1$ as it does at inference. Calibration uses no size bias.

Token reduction uses bipartite soft matching (Bolya et al., 2022): tokens are split by parity, each source proposes its most similar destination under cosine similarity, and the strongest proposals are selected according to the per-block merging budget. The class token is protected. Merged tokens are size-weighted averages of the original tokens they represent, with sizes accumulating across blocks, and $\log m_j$ is added to the corresponding logit column before the softmax. Our variant changes only the tensor on which cosine similarity is computed, from head-averaged keys (d channels) to $\bar{X}$ (all D channels). Because the size bias is a per-key additive term, logits are materialized for every reported configuration, including the baselines.

Accuracy differences between models evaluated on identical images use McNemar's exact test on discordant pairs, reported with n_{01}, n_{10} and two-sided p-value. Throughput is measured with clocks locked and no other process on the device, as the median of 200 timed iterations after 50 warm-up iterations; unlocked clocks varied by up to 9% between repeated measurements.

Every run asserts two identities before producing numbers: token reduction at rate 0 and truncation at $r = d$ must each reproduce the unmodified model. Both hold to 10^{-6} in maximum absolute logit deviation on all five classification backbones, on DPT-Large and on all five CLIP checkpoints. Software: PyTorch 2.x, timm, HuggingFace Transformers for DPT-Large, OpenCLIP for the CLIP towers.

B.2 NOTATION AND SETTINGS

Table 7: Symbols and fixed settings.

architecture		
N, N'	tokens entering, leaving a block	197; 577 for ViT-L/384
D, h, $d{=}D/h$	width, heads, head dimension	$d = 64$ throughout
$p = N - 1$	patch tokens	class token at index 0
compression		
$r \leq d$	retained rank	swept 4 to 64
$\tilde{M} = \tilde{W}_q \tilde{W}_k^\top$	query–key product	$(D{+}1)^2$, rank $\leq d$
A, B	rank-r factors	$(D{+}1) \times r$
Σ_0, $\Sigma_\epsilon = S^\top S$	second moment, floored	$\epsilon = 10^{-6}$
a, b, c	coefficients of E_1, E_4, E_5	one triple per head, scaled by $1/s_j$
π, s_j, m_u	partition, group size, token size	
calibration		
covariance / correction split	training images	4,096 / 128
factorization	precision, device	float64, CPU, batched over heads
reduction		
similarity	tensor compared	$\bar{X} = \mathrm{LN}(X)$, all D channels
schedule	tokens removed per block	constant across blocks
EViT	keep rate applied at blocks	3, 6, 9
evaluation		
batch size	classification	128; 32 for ViT-L/384
throughput	batch, precision	256, fp16, locked clocks
iterations	warm-up / timed	50 / 200
segmentation	resolution, calibration	480^2, 128 ADE20K train images
CLIP	prompts	80 templates, averaged

C Additional results

C.1 Rank sweep

Table 8 extends Section 4.2 over the full rank range. Compute saved saturates near 17% because rank reduction leaves the attention–value product, the value and output projections, and the MLP untouched. The cost of a given rank falls with width at every rank, and the correction contributes more where there is more error to repair, reaching 14.78 points on DeiT-Base at $r = 4$.

Table 8: Truncation across the rank range, development split, no token reduction. Dense is 80.32 on DeiT-Small and 82.68 on DeiT-Base.

	DeiT-Small			DeiT-Base		
r	GFLOPs	plain	$+C$	GFLOPs	plain	$+C$
64	4.599	80.32	80.32	17.564	82.68	82.68
48	4.380	80.32	80.30	16.777	82.60	82.70
32	4.161	79.36	79.68	15.991	81.96	82.10
24	4.051	78.52	78.84	15.597	81.54	81.72
16	3.942	74.58	76.80	15.204	80.46	80.92
12	3.887	66.48	71.04	15.007	77.72	79.26
8	3.832	55.40	62.46	14.811	70.48	75.14
4	3.778	23.38	30.14	14.614	39.46	54.24

C.2 Correction decomposition

Table 9 gives every configuration behind Table 1. At DeiT-Base $r = 32$ and $r = 24$ the truncation gap is 0.40 and 1.06 points, and the full table itself reaches significance only at $r = 24$ ($p = 3{\times}10^{-5}$), not at $r = 32$ ($p = 0.08$). We report these columns rather than restricting the table to ranks where the correction is measurable.

Table 9: Recovery of the truncation gap relative to the full position-indexed table, test split, all eight configurations. Negative values mean the correction moved accuracy the wrong way.

	Tiny		Small		Base			
correction	r=32	r=24	r=32	r=24	r=32	r=24	r=16	r=12
truncation gap	1.70	3.42	0.86	2.02	0.40	1.06	2.70	4.99
row shift	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E_1	84.4	72.8	42.4	55.0	−68.6	15.2	42.6	52.9
E_4	3.7	8.9	15.2	1.1	2.9	9.1	8.4	6.3
E_5	13.8	8.9	58.8	26.7	−105.7	−1.0	7.0	3.8
E_1, E_4	98.2	85.6	46.1	58.1	−77.1	26.3	52.9	58.2
all three	**103.7**	**96.8**	**93.9**	**87.2**	−60.0	6.1	**49.6**	**60.4**

Table 10 reports the transferred correction under reduction. E_5 is not uniformly beneficial: on Tiny at $r = 24$ the pair is marginally better than the triple, while on Small at $r = 32$ the triple is required to reach significance. The three DeiT-Base rows with near-equal discordant counts are those where truncation costs too little for any correction to register.

Table 10: Transferred correction under token reduction at rate 16, test split. Significance is McNemar against the uncorrected model of the same row.

		top-1				
backbone	r	uncorrected	E_1, E_4	all three	n_{01}/n_{10}	p
Tiny	32	67.26	67.71	**67.74**	1303/1517	6×10^{-5}
Tiny	24	65.88	**66.73**	66.73	1417/1797	2×10^{-11}
Small	32	77.36	77.50	**77.76**	712/ 890	1×10^{-5}
Small	24	76.37	76.71	**76.80**	900/1092	2×10^{-5}
Base	32	77.68	77.67	77.60	937/ 899	0.39
Base	24	77.34	77.28	77.26	998/ 963	0.44
Base	16	76.32	76.26	76.25	1036/1003	0.48
Base	12	74.62	74.78	**74.85**	1162/1266	0.037

C.3 Composition grids

Tables 11 and 12 give the absolute accuracies and compute behind Table 2. Compute depends only on the reduction rate and rank, so it is tabulated once. The $r = 64$ column is omitted: truncation is exact there and reproduces the baseline. Development sweeps covering 110 configurations on DeiT-Small, 88 on DeiT-Tiny and 77 on DeiT-Base, at ranks 24 to 56 and rates 0 to 22, are released with the code.

Table 11: DeiT-Small, test split. *Alone* is token reduction with block-input matching.

rate	alone	r=48	r=40	r=32	r=24
GFLOPs					
0	4.599	4.380	4.270	4.161	4.051
8	3.416	3.253	3.172	3.090	3.008
12	2.844	2.706	2.637	2.569	2.500
16	2.291	2.177	2.120	2.064	2.007
18	2.059	1.956	1.905	1.853	1.801
20	1.869	1.775	1.728	1.680	1.633
top-1, no correction					
0	79.80	79.73	79.47	78.94	77.78
8	79.46	79.40	79.17	78.80	77.70
12	79.06	79.02	78.70	78.55	77.36
16	78.03	78.01	77.80	77.36	76.37
18	76.66	76.66	76.51	76.10	75.17
20	74.66	74.54	74.24	74.17	73.30
top-1, with correction					
0	–	79.74	79.56	79.28	78.48
8	–	79.38	79.24	79.07	78.12
12	–	78.99	78.85	78.70	77.86
16	–	77.96	77.87	77.76	76.80
18	–	76.74	76.45	76.12	75.58
20	–	74.46	74.55	74.28	73.43

Table 12: DeiT-Tiny and DeiT-Base, test split, top-1 without / with correction. Dense is 72.19 and 81.91.

	DeiT-Tiny			DeiT-Base	
rate	alone	r=48	r=32	alone	r=48
0	72.19	71.77 / 71.92	70.49 / 70.99	81.91	81.85 / 81.81
8	71.50	71.09 / 71.27	69.94 / 70.45	81.12	81.07 / 81.07
12	70.72	70.29 / 70.40	69.24 / 69.74	80.20	80.12 / 80.08
16	68.91	68.41 / 68.44	67.26 / 67.74	78.05	77.85 / 77.94
18	65.92	65.69 / 65.79	64.82 / 65.00	74.75	74.62 / 74.70
20	62.06	62.24 / 61.77	61.40 / 61.14	68.50	68.35 / 68.46
22	55.66	56.07 / 55.72	56.16 / 55.56	not swept	

C.4 ViT-Large

Table 13 gives the ViT-L/16-384 sweep; the row at rate 22 is the configuration quoted in Section 4.4. At 224 pixels the sequence is short enough that rates above 8 exhaust it within 24 blocks, and there rank 48 with correction gives 84.60 at 29.56 GFLOPs against reduction alone at 84.38 and 30.96.

Table 13: ViT-L/16-384, development split. Dense is 87.40 at 191.07 GFLOPs. Rates above 26 exhaust the sequence.

	reduction alone		+ r=48		+ r=32	
rate	GFLOPs	top-1	GFLOPs	top-1	GFLOPs	top-1
16	124.1	87.20	118.1	87.06	112.2	86.88
20	108.1	87.20	102.9	87.08	97.6	86.74
22	100.2	86.92	95.3	**87.24**	90.4	86.70
24	92.4	86.74	87.9	86.54	83.4	86.40
26	85.5	85.94	81.3	85.94	77.1	85.60

C.5 SEGMENTATION

Table 14 extends Table 6 to a second input resolution and to the activation-agnostic factorization. Truncation is lossless to $r = 40$ at 480 pixels and to $r = 32$ at 384. Whitening is worth $+1.50$ mIoU at $r = 24$ and $+5.78$ at $r = 16$, a wider margin than on classification at the same rank. Dense mIoU of 45.92 is below the 47 reported for this checkpoint because we use single-scale square-resized inference rather than multi-scale sliding-window; all rows share that protocol.

Table 14: ADE20K with DPT-Large, 2,000 validation images, single-scale inference. Encoder GFLOPs; the decoder is untouched. Dense is 45.92 mIoU at 480 pixels and 43.89 at 384. The 384 and SVD columns were swept at ranks 48, 32, 24 and 16.

	480^2, whitened			384^2, whitened			480^2, SVD	
r	GFLOPs	mIoU	$+C$	GFLOPs	mIoU	$+C$	mIoU	$+C$
64	312.7	45.92	45.92					
56	304.5	45.93	45.93					
48	296.4	45.93	45.93	181.8	43.94	43.93	45.93	45.93
40	288.2	45.92	45.92					
32	280.1	45.66	45.70	172.5	43.92	43.91	45.27	45.42
24	271.9	45.02	45.20	167.8	43.62	43.74	43.52	44.18
16	263.7	42.04	42.70	163.1	41.39	41.93	36.26	39.49

C.6 CLIP ZERO-SHOT

Table 15 covers all five checkpoints. The correction grows with compression on every tower, from at most $+0.20$ at $r = 48$ to between $+2.30$ and $+4.92$ at $r = 16$, and is larger on narrower towers. The LAION checkpoint follows the same trend as its OpenAI counterpart, though their absolute accuracies differ. ViT-L/14 at $r = 56$ gives 75.80 and 75.94; ViT-L/14-336 at $r = 56$ gives 76.80 and 76.80, at $r = 40$ gives 76.20 and 76.12. An activation-agnostic SVD on ViT-L/14 reaches 66.36 and 30.70 at $r = 32, 16$.

Table 15: CLIP zero-shot ImageNet, development split, top-1 without / with correction.

		r=48		r=32		r=24		r=16	
tower	dense	plain	$+C$	plain	$+C$	plain	$+C$	plain	$+C$
ViT-B/32	63.76	63.42	63.40	61.42	62.02	59.42	61.26	53.50	**58.42**
ViT-B/16	69.18	68.58	68.66	66.78	67.62	63.78	65.28	55.68	**59.94**
ViT-L/14	76.04	75.96	76.16	74.20	74.58	72.30	72.90	66.14	**68.76**
ViT-L/14-336	76.92	76.64	76.74	74.80	75.10	72.68	73.64	66.74	**69.30**
ViT-L/14 LAION	75.44	75.32	75.36	73.76	74.28	72.10	72.50	66.42	**68.72**

C.7 EVIT

Table 16 gives both ranks. At $r = 32$ gains first turn positive at keep rate 0.80 on Tiny and Base and at 0.70 on Small; at $r = 48$ Small and Base already benefit at 0.90. The crossover depends on backbone and rank.

Table 16: EViT composed with truncation and the correction, development split. Change against EViT alone at equal compute.

	$r = 48$				$r = 32$			
keep rate	0.90	0.80	0.70	0.60	0.90	0.80	0.70	0.60
DeiT-Tiny	-0.06	$+0.23$	$+0.93$	$+1.92$	-0.47	$+0.32$	$+1.43$	$+\mathbf{3.29}$
DeiT-Small	$+0.20$	$+0.15$	$+0.38$	$+0.91$	-0.37	-0.27	$+0.22$	$+1.09$
DeiT-Base	$+0.09$	$+0.34$	$+0.56$	$+1.68$	-0.25	$+0.34$	$+1.10$	$+2.99$

D ABLATIONS

D.1 FACTORIZATION BASIS AND CONDITIONING

Table 17 compares three ways of selecting the retained subspace. At $r = 48$, both whitened and activation-agnostic SVD retain most of the dense accuracy, with whitening ahead by 0.28 points on DeiT-Small but behind by 0.06 points on DeiT-Base. The advantage of whitening grows as rank decreases, reaching 6.60 points on DeiT-Small and 6.88 points on DeiT-Base at $r = 16$. The gap emerges earlier on the narrower model.

Table 17: Factorization basis, development split, top-1 without / with correction.

		r=48		r=32		r=16	
backbone	basis	plain	$+C$	plain	$+C$	plain	$+C$
DeiT-Small	whitened	**80.32**	**80.30**	**79.36**	**79.68**	**74.58**	**76.80**
	SVD	80.04	80.02	78.38	78.76	67.98	71.50
	random	5.54	5.52	3.28	3.36	2.26	2.40
DeiT-Base	whitened	82.60	**82.70**	**81.96**	**82.10**	**80.46**	**80.92**
	SVD	**82.66**	82.68	81.36	81.32	73.58	74.58
	random	3.42	3.44	2.96	2.96	2.40	2.46

Table 18 sweeps the eigenvalue floor. Accuracy is unchanged from 10^{-8} to 10^{-3}, improves slightly at 10^{-1}, then declines at larger values. At $\epsilon = 10^5$ the whitened factorization reproduces the activation-agnostic SVD exactly, matching Table 17 to three decimals: $\Sigma_\epsilon \to \epsilon I$ makes S a multiple of the identity and Equation 4 an ordinary SVD. The floor interpolates between the two bases, and the default sits well inside the flat region.

Table 18: Eigenvalue floor ϵ, DeiT-Small, development split, no correction.

ϵ	10^{-8}	10^{-6}	10^{-4}	10^{-3}	10^{-1}	10^{1}	10^{3}	10^{5}
$r = 32$	79.36	79.36	79.36	79.36	79.44	79.16	78.40	**78.38**
$r = 16$	74.58	74.58	74.58	74.58	74.70	72.74	68.10	**67.98**

D.2 CALIBRATION SIZE

Table 19 varies the number of images used to estimate Σ_0. Reducing $4{,}096$ to 512 images changes uncorrected accuracy by -0.06 and -0.02 points at $r = 32$ and 24; with correction, -0.06 and $+0.22$. Σ_0 is a $(D{+}1)^2$ matrix and each image contributes $N = 197$ tokens, so 512 images already supply roughly 10^5 samples. Fitting the compression from 512 images takes minutes on one GPU.

Table 19: Images used to estimate Σ_0, DeiT-Small, development split, top-1 without / with correction.

	$r = 32$		$r = 24$	
images	plain	$+C$	plain	$+C$
512	79.30	79.62	78.50	**79.06**
4,096	79.36	79.68	78.52	78.84

D.3 SIMILARITY SOURCE

Figure 4 shows that the block input is more stable at all eighteen combinations of rank and reduction rate, by 0.13 to 0.23 ARI. Both sources degrade, since the block input at layer k depends on earlier attention outputs, which are also truncated; the block input degrades roughly half as fast.

At $r = 64$, where truncation is exact, the change in similarity source alone gives $+0.10$, $+0.32$, $+0.30$, $+0.20$, $+0.42$ and $+1.28$ points at rates 12, 16, 18, 20, 22 and 24 on DeiT-Small. On DeiT-

Base the same comparison gives -0.18 at rate 8, $+0.06$ at 12, $+0.32$ at 16, $+0.80$ at 18 and $+2.19$ at 20: the benefit is absent where reduction is mild.

D.4 Size scaling

Table 20 compares the scaled and unscaled diagonal coefficient. Differences of -0.08, -0.20 and $+0.22$ points, of mixed sign. We apply the scaling, since it is what Proposition 2 derives for transport under averaging; it is the identity in every result reported without token reduction.

Table 20: Diagonal coefficient with and without $1/s$ scaling, DeiT-Small, $r = 32$, development split, matching on compressed keys. The uncorrected row is identical for both by construction.

	rate 16	rate 18	rate 20
uncorrected	76.98	75.30	72.06
unscaled	**77.44**	**75.38**	73.16
scaled by $1/s$	77.36	75.18	**73.38**

D.5 Activation-agnostic factorization under reduction

Table 21 substitutes the activation-agnostic factorization into the composition. Whitening is worth 0.36 to 1.06 points at every rate, with no trend toward zero as reduction hardens.

Table 21: Composition with each factorization, DeiT-Small, $r = 32$ with correction, development split.

rate	0	8	12	16	18	20
SVD	78.84	78.98	78.48	77.64	75.96	74.16
whitened	**79.68**	**79.66**	**79.10**	**78.20**	**77.02**	**74.52**
margin	$+0.84$	$+0.68$	$+0.62$	$+0.56$	$+1.06$	$+0.36$

E Qualitative results

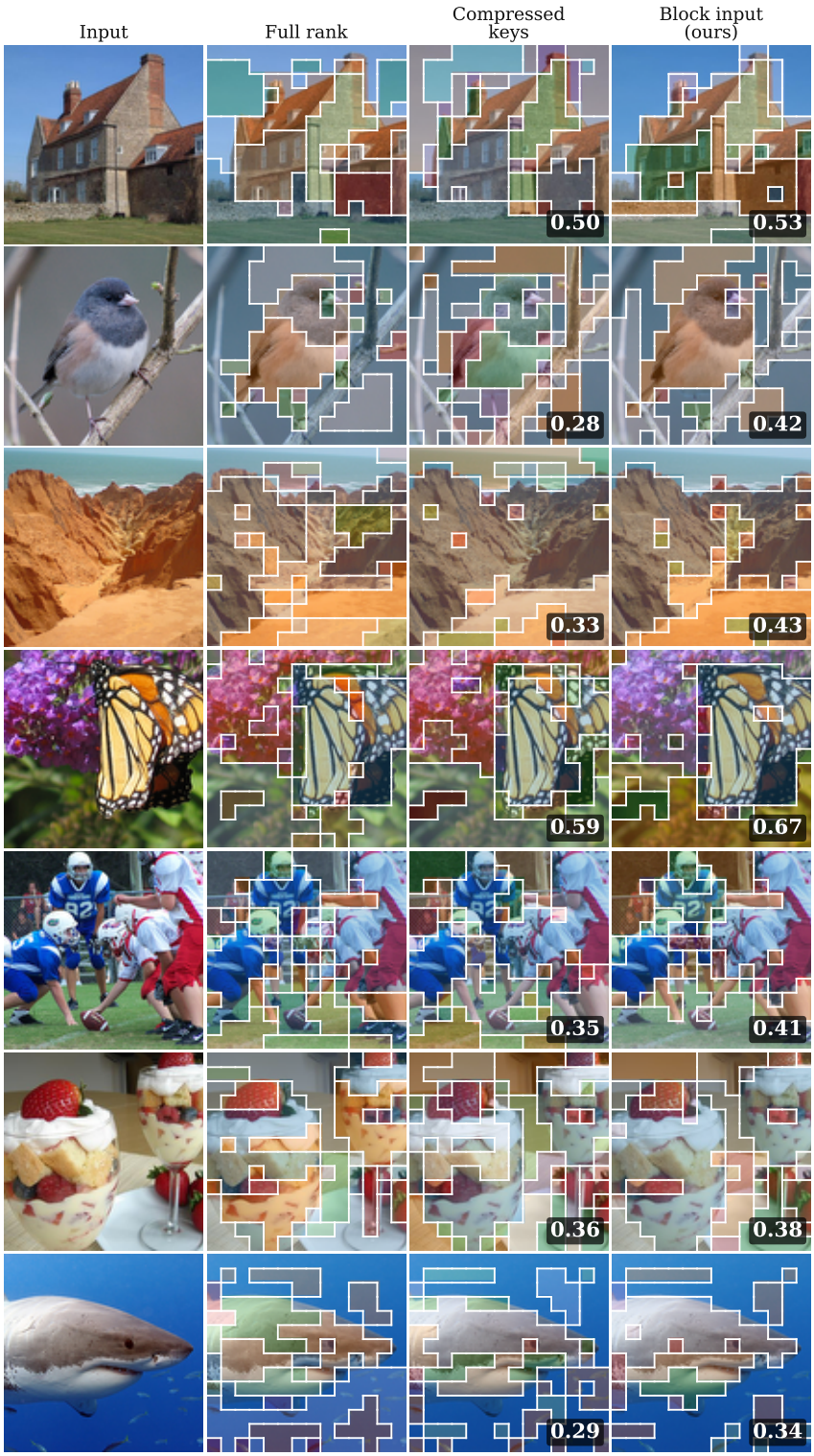

Figure 5: Merge decisions on further images, settings as Figure 4. Block-input matching scores a higher ARI on five of the seven; averaged over 256 images it is higher at all eighteen settings.



Figure 6: ADE20K predictions at four ranks on eight validation images, DPT-Large at 480×480.

F STRUCTURE OF THE TRUNCATION RESIDUAL

The three invariant components account for 20–38% of the squared magnitude of the residual across blocks and backbones while recovering 94–104% of the accuracy the full table recovers on DeiT-Tiny and DeiT-Small. The remainder exceeds a uniform distribution by a factor of only 2.1 to 3.0, so it is not concentrated on a few tokens; its dominant structure is the spatial banding visible in Figure 7, which is bound to position and outside the invariant space. This is also where the DeiT-Base columns of Table 9 sit: when the banded component dominates, a position-indexed table can represent it and a position-free correction cannot.

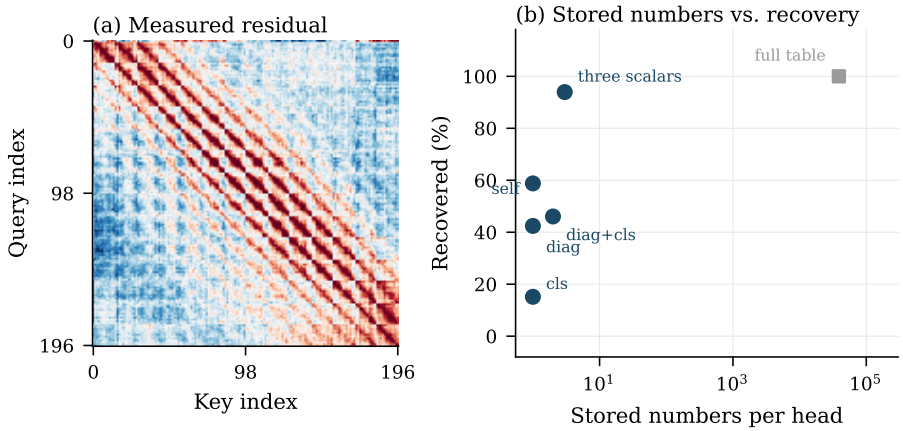

Figure 7: (a) One head of block 6, DeiT-Small at $r = 32$, with the row shift removed. The patch diagonal and class column appear in the data, as do bands at displacements ± 1 and ± 14, the four-neighbourhood of the 14×14 patch grid. (b) Accuracy recovered against stored values per head, test split.

G HYPOTHESES TESTED AND REJECTED

The residual does not concentrate on a small number of register-like tokens; it is diffuse and spatially banded (Appendix F).

Larger models do not uniformly suffer more from token reduction. At reduction rate 24 the losses are 26.6, 15.2 and 45.1 points for Tiny, Small and Base. Truncation, by contrast, costs less on wider models at every rank (Table 8).

A position-indexed table does not transfer across widely separated reduction rates: calibrated at rate 0 and applied at rate 20 it performs at chance, while the three coefficients transfer to moderate rates (Table 10).