DISTILLATION REPAIRS WHAT POST-HOC K=V SURGERY BREAKS IN ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sharing key and value projections ($K{=}V$) works when chosen *before* training: Kayyam et al. (2026) show it matches full QKV attention when trained jointly from scratch. We ask what happens when the same equality is instead imposed *after* training, by merging an already-trained model's independent K/V projections with no retraining. The two settings are qualitatively different: a single attention sub-block's output is provably confined to the convex hull of the shared K/V vectors regardless of attention weights, and this post-hoc surgery collapses every routing-dependent task toward chance — across five synthetic tasks, four vision datasets, a 300M-parameter language model, and an ImageNet-1k ViT-S/16 teacher — while tasks solvable without cross-token content routing are unaffected. The damage is *cheap to repair*: a short distillation pass (5–10 epochs) against the original model recovers nearly every configuration to within a point of the from-scratch ceiling, with one consistent exception: reusing the key projection alone (KEEP_K) recovers markedly slower and less completely than reusing the value projection or averaging the two. Mechanism tests (a content probe, a random-projection baseline, an addressing/payload decomposition, and a learned $K \to V$ basis correction) trace this to a representation-*compatibility* problem, not lost content: K and V are equally decodable, but a cheap, closed-form change of basis closes most of KEEP_K's gap with no retraining. Finally, we confirm on the same ImageNet-1k family that choosing $K{=}V$ *before* training lands within a point of this recovered ceiling directly, no repair needed — one picture: the constraint is easy to accommodate when the network is optimized around it from the start, and only becomes a repair problem when imposed after independent specialization.

1 INTRODUCTION

Kayyam et al. (2026) evaluate projection-sharing attention — Q≠K=V, Q=K≠V, and Q=K=V, de-symmetrized where needed by a 2D positional correction — and show that, trained jointly from a random initialization, these variants match standard QKV attention at small-to-medium scale. Gao and Xu (2026) re-derive the same mechanism from a value-space-routing motivation and generalize it via depth-m factorization. Both results are about a constraint chosen *before* training starts. This paper asks a different question about the same equality $K{=}V$: what happens if it is imposed *after* training, on a model whose K and V projections were independent throughout — and, more importantly, **how cheaply can the resulting damage be repaired?**

Joint training lets addressing and content pressures negotiate a shared representation from the start, under gradient descent's full control. Post-hoc surgery — taking an already-trained QKV checkpoint and setting its shared projection to KEEP_K ($\leftarrow c_k$), KEEP_V ($\leftarrow c_v$), or AVG ($\leftarrow \frac{1}{2}(c_k + c_v)$), with no retraining — is a different intervention entirely: it takes two representations optimized under no such joint pressure and imposes the constraint externally, with no opportunity for the rest of the network to adapt. Sections 3–4.1 develop a convex-hull argument for why this is a fundamentally harder operation than choosing the constraint up front, test the resulting collapse-and-recovery pattern across five scales — five synthetic routing tasks, four vision datasets, a 300M-parameter language model, and (new to this study) an ImageNet-1k ViT-S/16 teacher — and investigate the mechanism behind both the collapse and its repair. Section 5 then closes the loop with a short confirmatory study: trained on the same ImageNet-1k QKV teacher family, choosing $K{=}V$ *before* training reaches, directly and with no repair step, within a point of the ceiling distillation only

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.

QKV	**Post-hoc $K{=}V$ (surgery)**	**From-scratch $K{=}V$**
$x_j \to K_j$ (address), $x_j \to V_j$ (content)	$x_j \to KV_j$: the *same* vector forced to serve both roles, no retraining	$x_j \to KV_j$: one projection, but the whole network trained around it
$y_i = \sum_j a_{ij}(Q_i, K_j)\, V_j$; two independently-trained projections	$y_i \in \mathrm{conv}\{KV_j\}$; rest of the network frozen (Prop. 1)	$y_i \in \mathrm{conv}\{KV_j\}$; same bound, but c_q, c_{proj}, MLPs, head all adapted to it

Table 1: The same architecture, $K{=}V$, reached two different ways — the contrast this paper is built around. Left: the QKV teacher, with independent K/V free to specialize. Middle: post-hoc surgery (Sections 3–4.1) forces one already-specialized representation to do both jobs, nothing else adjusted. Right: choosing $K{=}V$ before training (Section 5) lets every other parameter co-adapt to it from the start — the same bound (Proposition 1) applies to both middle and right, but only the right panel has had a chance to unlearn nothing.

reaches *after* surgery — evidence that the constraint itself is not the problem; only its post-hoc imposition is.

Contributions. (i) A convex-hull proposition (Section 3) confining post-hoc $K{=}V$ tying to the span of the shared vectors, and an account of which tasks are immune. (ii) Empirical validation across five synthetic tasks, four vision datasets, a 300M-parameter LLM, and a new ImageNet-1k QKV teacher (Section 4), with one consistent exception (KEEP_K) that motivates (iii) a battery of mechanism tests (Section 4.1) — a linear content probe, a random-projection baseline, an addressing/payload decomposition, and a learned $K \to V$ basis correction — tracing that exception to representation *incompatibility* rather than lost content, confirmed by direct intervention. (iv) A six-condition confirmatory sweep training ViT-S/16 from scratch on full ImageNet-1k (Sec. 5), extending prior joint-training evidence beyond MNIST/CIFAR scale and revising an earlier symmetric-vs-asymmetric accuracy verdict from that sweep. Full setup and per-task tables are in Appendices A–C.

2 Setup and Notation

We use the attention parameterization of the base paper throughout: for input $x \in \mathbb{R}^{T\times d}$, single-head[1] self-attention computes $q_i = c_q(x_i)$, $k_j = c_k(x_j)$, $v_j = c_v(x_j)$, $a_{ij} = \mathrm{softmax}_j(q_i \cdot k_j/\sqrt{d})$, $y_i = \sum_j a_{ij} v_j$, where c_q, c_k, c_v are learned linear maps (equivalently $q_i = x_i W^Q$ in matrix notation). Variants are recovered by sharing these maps: QKV (none shared), Q=K≠V, Q ≠K=V, and Q=K=V, each optionally with a 2D positional injection (the "+" variants) that de-symmetrizes the resulting attention map by adding a fixed 2D relative-position sinusoidal term to the raw score matrix, indexed by true ViT patch row/column rather than a synthetic reindexing of a 1D sequence (full form in Appendix A). QVV(3), used in Section 5, is the depth-2-factored variant of Q≠K=V from Gao and Xu (2026), matching standard attention's parameter count.

For the post-hoc surgery experiments, given an *already-trained* QKV checkpoint, we construct a Q≠K=V student by setting the shared projection c_{kv} to one of KEEP_K (c_k), KEEP_V (c_v), or AVG ($\frac{1}{2}(c_k + c_v)$), copying every other parameter verbatim from the teacher; joint training lets every other parameter co-adapt to the constraint from the start, while post-hoc surgery imposes it externally with no such opportunity (Appendix A gives the full construction and a schematic).

3 A Convex-Hull Obstruction

Proposition 1 (Convex-hull confinement under K=V)**.** *If $k_j = v_j$ for every position j, then for every query position i, $y_i \in \mathrm{conv}\{k_1, \ldots, k_T\} \subseteq \mathrm{span}\{k_1, \ldots, k_T\}$, regardless of the attention weights a_{ij}.*

Proof. $a_{ij} \geq 0$, $\sum_j a_{ij} = 1$ by the softmax, so $y_i = \sum_j a_{ij} k_j$ is a convex combination of $\{k_j\}$. □

Corollary 1. *The attention sub-block can only make available to position i information already representable in* $\mathrm{span}\{k_1, \ldots, k_T\}$*. If the trained K-projection specialized for* addressing *without*

[1]The multi-head case is identical after splitting d into H blocks; every argument below applies unchanged within each head.

Dataset	Mode	QKV	Q≠K=V (scratch)	Zero-shot surgery	Distilled (5 ep)	Δ% vs. scratch	Distilled (10 ep)
MNIST	keep_k	0.981	0.978	0.114	0.979	+0.1%	–
	keep_v	”	”	0.151	0.982	+0.4%	–
	avg	”	”	0.910	0.983	+0.5%	–
FMNIST	keep_k	0.887	0.882	0.084	0.881	−0.1%	–
	keep_v	”	”	0.105	0.883	+0.1%	–
	avg	”	”	0.811	0.888	+0.7%	–
CIFAR-10	keep_k	0.696	0.698	0.076	0.631	−9.6%	0.666
	keep_v	”	”	0.153	0.686	−1.7%	0.693
	avg	”	”	0.368	0.696	−0.3%	**0.702**
CIFAR-100	keep_k	0.437	0.445	0.007	0.384	−13.7%	0.411
	keep_v	”	”	0.021	0.424	−4.7%	0.430
	avg	”	”	0.109	0.432	−2.9%	0.435
ImageNet-1k	keep_k	0.7438	0.7167	0.0008	0.7101	−0.9%	–
	keep_v	”	”	0.0191	0.7193	+0.4%	–
	avg	”	”	0.0066	0.7187	+0.3%	–

Table 2: Vision classification (ViT, top-1 accuracy). Every dataset shows some collapse under KEEP_K/KEEP_V (unlike the two pointwise synthetic tasks): these ViTs classify from a learned CLS token, itself an attention-weighted, content-routing operation. ” marks a value identical to the row above (QKV and the Q≠K=V scratch ceiling do not depend on surgery mode). Δ% vs. scratch is the 5-epoch distilled accuracy's relative distance from the from-scratch Q≠K=V ceiling, (distilled − scratch)/scratch; near 0% means full recovery. CIFAR was rerun at 10 epochs since KEEP_K was still improving at 5 (AVG/CIFAR-10 slightly exceeds the from-scratch ceiling). The last three rows are the new ImageNet-1k scale point, evaluated with the same protocol (no 10-epoch rerun attempted, given compute cost).

pressure to remain injective in token content — since that pressure was carried by the independent V-projection — k_j need not determine x_j, and routing k_j itself can lose content the downstream task needs. This is a hypothesis stated before we test it; Section 4.1 finds the picture more specific than “K lacks content” as stated.

Remark 1 (Scope)**.** *Proposition 1 constrains y_i, one attention sub-block's output — not $h_i^{out} = h_i^{in} + y_i$, the MLP applied to it, or any later layer. A model with $K{=}V$ in every layer is* not *restricted to one layer's convex hull; the proposition says a single $K{=}V$ operation cannot independently choose the content-transport representation from the addressing representation — what reaches position i must be a convex combination of the very vectors that determined* where *to look. (Normalization invariance and the full proof discussion are in Appendix A.)*

Two of five synthetic tasks are pointwise (SUB: $y_i = 9 - x_i$; COPY: $y_i = x_i$): each block's residual connection preserves a direct additive path from position i's own input, so attention's specific content-routing is not necessary and Proposition 1's obstruction is vacuous. Both surgeries indeed leave SUB/COPY at 1.000 accuracy (Table 7, Appendix B), while three routing-dependent tasks (REVERSE, SORT, SWAP) collapse sharply under the same surgeries.

4 Empirical Validation of Post-Hoc Surgery

We evaluate zero-shot surgery (no fine-tuning) and 5-epoch distillation recovery across five synthetic tasks, four vision datasets, a 300M-parameter FineWeb-Edu language model, and — new to this study — the ImageNet-1k QKV teacher from Section 5. Distillation is data-dependent throughout, not data-free: the student trains on the same real training split used to train the teacher (real images/tokens and their real labels), minimizing hard-label cross-entropy against those labels plus a KL term against the teacher's soft logits ($T{=}2$); we did not run a data-free variant (e.g. synthetic or teacher-generated inputs with no real labels), which would test a different, weaker-supervision question and is left to future work. The synthetic-task table is in Appendix B (Table 7); the vision result, extended with the new ImageNet-1k row, is Table 2 above.

AVG is the safer zero-shot mode on every dataset except ImageNet-1k, where KEEP_V is marginally safer (1.91% vs. 0.66%, both far above KEEP_K's 0.08%) — the first such departure, though all three collapse to the same order of magnitude above chance either way. After distillation, ImageNet-1k's KEEP_V and AVG both slightly *exceed* the 71.67% ceiling, echoing AVG/CIFAR-10's own near-complete recovery, and KEEP_K — the mode with a persistent, only-half-closed gap even after 10 epochs on CIFAR — reaches 71.01% here, just 0.66pt behind, with per-epoch gains decelerating sharply enough by epoch 5 that a 10-epoch extension was not attempted. Both the zero-shot severity and the post-distillation KEEP_K penalty are markedly smaller at ImageNet-1k scale than at every other scale in this table — an open question (more distillation steps per epoch on 1.28M images

Mode	QKV teacher	Q≠K=V (scratch)	Zero-shot surgery	Distilled (500M tok)	Δ% vs. scratch
keep_k	20.82	21.28	13345.19	29.00	+36.3%
keep_v	”	”	7347.14	27.00	+26.9%
avg	”	”	6430.36	22.87	+7.5%

Table 3: FineWeb-Edu language modeling, 300M params, validation perplexity (lower is better). Same protocol as Table 2, but distillation uses a 500M-token budget (about 5% of the from-scratch pretraining budget) rather than a fixed epoch count. ” marks a value identical to the row above. Unlike Table 2, higher Δ% here is *worse* (higher perplexity than the from-scratch ceiling), since lower perplexity is better: KEEP_K remains 36.3% above the ceiling even after distillation, versus 7.5% for AVG.

Task/dataset	KEEP_K	random c_{kv} (mean ± std, 5 seeds)	z
REVERSE	0.100	0.100 ± 0.0004	$+1.25$
SORT	0.583	0.584 ± 0.0099	-0.12
SWAP	0.088	0.100 ± 0.0010	-12.30
MNIST	0.114	0.091 ± 0.0124	$+1.92$
FMNIST	0.084	0.118 ± 0.0278	-1.23
CIFAR-10	0.076	0.112 ± 0.0210	-1.72
CIFAR-100	0.007	0.011 ± 0.0024	-1.88
ImageNet-1k	0.0008	0.0010 ± 0.0003	-0.82

Table 4: KEEP_K vs. a random, untrained c_{kv} baseline (SUB/COPY omitted: always 1.000). Only SWAP is a large, unambiguous effect; every other task, including the new ImageNet-1k row, is within about 2 standard deviations of the random baseline's own spread — not reliably distinguishable from noise.

than CIFAR's 50k, more redundant capacity to route around a mismatched basis, or something classification-specific) this single scale point cannot settle. Full per-epoch recovery curves for every (task/dataset, surgery-mode) configuration are in Appendix F; a continuous sweep of c_{kv} between KEEP_V and KEEP_K, against a random- initialization baseline, is in Appendix G. At LLM scale (Table 3), zero-shot perplexity worsens by two to three orders of magnitude rather than the roughly one-order-of-magnitude accuracy collapse above, consistent with perplexity's exponential sensitivity to per-token log-loss rather than a qualitatively different failure mode; distillation again recovers AVG > KEEP_V > KEEP_K, reproducing the vision-scale ordering, with KEEP_K again the mode furthest from its ceiling (29.00 vs. 21.28 perplexity after 500M distillation tokens).

4.1 MECHANISM

Corollary 1's premise — that a trained K-projection specialized for addressing at some cost to content — has not, until now, been tested directly. We test both halves on the five synthetic-task teachers, where the correct routing target at each position is known exactly. A closed-form ridge-regression probe from k_j (resp. v_j) to token identity shows *both* are decodable at or near 1.000 on every task, with no consistent gap — contradicting the corollary's content-loss story as stated, at least at this scale. The same probe on the 300M-parameter LLM teacher tells a different story: V probes noticeably better than K at every one of 20 layers (85.3% vs. 80.3% mean accuracy, a modest but universal gap), closer to the original content-loss account — a genuine point of scale-dependent divergence we do not resolve.

If KEEP_K collapses only because *any* unfamiliar c_{kv} throws the frozen downstream network off distribution — with the specific content of K irrelevant — then KEEP_K should perform no worse than a c_{kv} with no relationship to the teacher at all. We test this with a **random baseline**: c_{kv} left at its fresh, untrained `nn.Linear` initialization (independent of the teacher's trained K or V), every other parameter — c_q, c_{proj}, MLPs, layer norms, embeddings, head — copied verbatim from the teacher exactly as in KEEP_K/KEEP_V/AVG, evaluated zero-shot with no retraining, averaged over 5 random-initialization seeds:

Read together with this random baseline (Table 4), this points to a more specific mechanism than "K lost the content": outside SWAP's dramatic exception, KEEP_K's specialized K-derived representation confers no detectable advantage over uninformative noise to the frozen downstream network, even though that representation is just as linearly content-rich as V's. This is consistent with a **basis-mismatch** account rather than information loss: k_j carries identity in a representation shaped by the addressing objective, not the one c_{proj}/MLP/head were calibrated against.

Task/dataset	KEEP_K	KEEP_K + linear	KEEP_K + Procrustes	KEEP_V
REVERSE	0.100	0.118	0.107	0.123
SORT	0.583	**0.740**	0.710	0.717
SWAP	0.088	0.118	0.107	0.122
MNIST	0.114	**0.151**	0.113	0.151
FMNIST	0.084	**0.109**	0.100	0.105
CIFAR-10	0.076	0.152	**0.170**	0.153
CIFAR-100	0.007	0.023	**0.025**	0.021
ImageNet-1k	0.0008	0.0145	0.0083	0.0191

Table 5: Zero-shot accuracy correcting KEEP_K's c_{kv} toward V's basis, no retraining. Bold: matches or exceeds KEEP_V. On SORT and every CIFAR-scale/LLM dataset, either correction closes the entire gap and reaches or exceeds KEEP_V — direct manipulation of the hypothesized mechanism, not inference from a probe and a baseline. ImageNet-1k (last row) is the one exception at any scale tested: both corrections help substantially (an order of magnitude over KEEP_K) but neither closes the gap to KEEP_V, unexplained.

We add two zero-shot experiments that manipulate this hypothesis directly rather than inferring it. **Addressing/payload decomposition.** Writing $A(S, P)$ for attention with scores from S and payload from P, the only genuinely new cell beyond the teacher/KEEP_K/KEEP_V triple is $A(Q, V)K$ (scores from V, payload from K): it tracks $A(Q, K)K$=KEEP_K closely on SORT and both CIFAR datasets — evidence that payload identity, not addressing identity, drives the damage — and gives the single cleanest such signal at any scale on the new ImageNet-1k teacher ($A(Q, V)K$=0.10%, almost exactly KEEP_K's 0.08%, far from KEEP_V's 1.91%; the LLM point is less clean, falling between the two). **K→V alignment.** A learned linear (ridge) or orthogonal (Procrustes) map from K's basis to V's is a more direct test. One map is fit per attention block, on the full n_{embd}-dimensional $c_k(x)/c_v(x)$ output *before* the per-head split (so heads within a block share one correction, not one each), from a single held-out training-split batch (2000 sequences/512 images/512 images at synthetic/vision/ImageNet-1k scale respectively, ∼16k tokens at LLM scale) disjoint from the accuracy-reporting test set, then applied zero-shot with *no retraining of anything downstream*:

on SORT and every CIFAR-scale/LLM dataset, this cheap, closed-form fix closes the *entire* KEEP_K gap and matches or exceeds KEEP_V (Table 5); at LLM scale it does so on a 389M-parameter model trained on 10B tokens, two orders of magnitude more parameters than the vision ViTs, weighing against reading the small-scale result as an artifact of model or data size. ImageNet-1k is the only scale where this does not fully replicate. The fitted linear map's singular values suggest this correction could be made cheaper still: their mean sits well below 1 (0.33 at synthetic/vision scale, 0.016 at LLM scale) rather than clustered near 1 as a near-rotation would be, indicating the useful $K \to V$ relationship is concentrated in a low-dimensional subspace rather than requiring a full change of basis (Appendix C).

A further, more direct test of the same question — does the correction need the real image distribution, or just some input diverse enough to reveal the $K \to V$ relationship? — refits both corrections on the ImageNet-1k teacher using i.i.d. Gaussian noise batches in place of real images (5 noise seeds, everything else identical, including evaluation on the real validation set). The data-free linear correction reaches 0.79% top-1 and data-free Procrustes reaches 0.31%, both far above raw KEEP_K (0.08%) but recovering less of the KEEP_K-to-KEEP_V gap than their real-data counterparts (linear: 39% of the gap data-free vs. 75% on real images; Procrustes: 13% vs. 41%) — real image statistics still matter, more so for the rotation-only Procrustes map than for the affine linear one. Results are essentially seed-invariant (std ≤ 0.0001 across the 5 noise draws), consistent with the fitted map mostly recovering the network's own weight geometry rather than memorizing anything about the specific fitting batch.

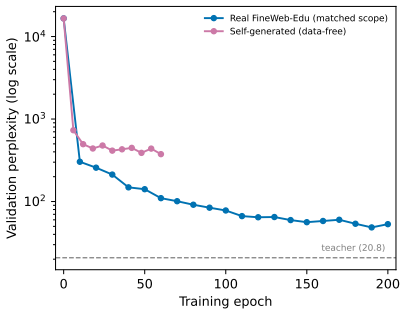


Figure 1: KEEP_K validation perplexity (log scale) vs. training epoch, KD-only recovery restricted to the tied c_k projection. Self-generated: 65,536 tokens, 60 epochs. Real FineWeb-Edu (matched scope): same-size corpus, 200 epochs. Dashed line: teacher perplexity (20.82).

Unlike a vision classifier, an LLM teacher can generate its own plausible training data by construction, so we test a data-free *recovery* (gradient-based KD training, not a closed-form fit) at LLM scale directly: freezing every parameter except each layer's tied c_k projection (21M of 300M parameters, the same restricted scope as the vision correction above) and training via KD alone (no hard la-

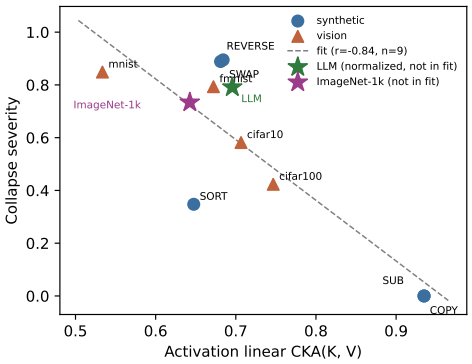


Figure 2: K–V activation CKA vs. accuracy drop under surgery, 5 synthetic tasks + 4 vision datasets (dots; Pearson $r = -0.84$ on these $n = 9$ points, dashed fit line). Stars: the LLM teacher and the new ImageNet-1k teacher, computed after the $n = 9$ fit was already reported and left out of the regression itself (so the originally-reported $r = -0.84$, $n = 9$ statistic does not silently change) rather than for any data-quality reason — both land inside the same scatter as the fit predicts, and the LLM star's y-coordinate is a log-loss-normalized severity (Section 4.1 has no perplexity-scale collapse-severity number to plot directly, so this is the closest analogue), not the same quantity as the accuracy-drop y-axis elsewhere.

bels exist in a data-free setting) against either a small corpus the teacher generates about itself with no real text anywhere, or an equally-sized, equally-scheduled corpus of real FineWeb-Edu text — isolating the data-source variable alone.

Both recover a large fraction of the KEEP_K collapse relative to zero-shot (16589.7), but the two corpora behave qualitatively differently, not just quantitatively (Figure 1): the self-generated corpus plateaus by roughly epoch 18–30 and then oscillates in a noisy band with only a slow net downward drift, bottoming out at 375.3 across 60 epochs ($44.2\times$ below zero-shot). Real text keeps improving far longer, reaching 109.9 by epoch 60 ($150.9\times$, a $3.4\times$ gap over data-free at the same epoch) and continuing to a best of 48.6 at epoch 190 ($342\times$, within $2.3\times$ of the teacher) before its own later plateau around epoch 110–150 — restricting training to c_k alone on a tiny, fixed real corpus eventually approaches the teacher far more closely than a 60-epoch snapshot would suggest, whereas the self-generated corpus's early plateau is a genuine ceiling, not a mid-training artifact. Self-generated text carries real but bounded signal, quickly exhausted, whereas an equally sized real corpus keeps teaching the restricted parameter for far longer. This is a genuine asymmetry with the vision result above, where even the closed-form data-free correction transfers a majority (linear) or substantial minority (Procrustes) of the real-data effect in a single fit, with no analogous training-length axis to plateau along.

The basis-mismatch account makes a further prediction: the teacher's own K and V representations, *before any surgery*, should already look more different (in a basis-invariant sense) on tasks/datasets where surgery does more damage. We test this with **linear CKA (Centered Kernel Alignment)** (Kornblith et al., 2019), a similarity score between two sets of activations that — unlike raw weight comparison — is invariant to rotation and isotropic rescaling, so it can detect that two representations compute the same information even when their parameters look nothing alike. For each teacher (pre-surgery), we capture $c_k(x)$ and $c_v(x)$ — the same K/V projection outputs referenced throughout this section — via forward hooks on a held-out batch, compute CKA between them at every attention layer, and average across layers into one CKA(K,V) score per teacher.

A K–V activation-CKA diagnostic (Figure 2) correlates with collapse severity across the other 9 tasks/datasets ($r = -0.84$, $p = 0.005$, $n = 9$; weaker and non-significant under Spearman's rank correlation, $\rho = -0.56$, $p = 0.12$, and within either task-type group alone at these sample sizes) and places the ImageNet-1k teacher (CKA$= 0.64$) and the LLM teacher (CKA$= 0.70$, log-loss-normalized since perplexity has no shared scale with accuracy) inside that same scatter, though we do not treat it as a validated general predictor — notably, CKA *rises* rather than falls with classification difficulty across the four vision datasets even as collapse severity also rises, the opposite of the intuitive direction. A per-layer breakdown (Appendix C) localizes this reversal rather than resolving it away: it is carried almost entirely by the layer feeding the CLS token, where CKA falls monotonically as dataset difficulty rises (0.33 MNIST $\rightarrow$ 0.70 CIFAR-100), while the earlier layer shows the opposite, intuitive ordering — a single-layer effect, not a dataset-wide one. Full breakdown and robustness checks (RBF kernel, weight-cosine control) are in Appendix C.

5 Confirmation: Joint Training Needs No Repair

Everything above concerns a constraint imposed *after* training. As a direct check that the constraint itself is not the problem, we train ViT-S/16 from scratch on full ImageNet-1k, 100 epochs, DeiT-style recipe, across six projection-sharing conditions — the same QKV/Q≠K=V checkpoint family

Variant	Top-1	Top-5	Δ Overfit	Cache	Params
QKV (baseline)	74.38%	91.93%	+0.0010	100%	22.05M
Q≠K=V	71.67%	90.32%	−0.0004	50%	20.28M
Q=K=V$^+$	66.76%	87.14%	−0.0014	33%	18.50M
QVV(3)	72.55%	90.81%	+0.0009	50%	22.05M
Q=K≠V$^+$	69.45%	89.03%	−0.0007	50%	20.28M
Q=K=V (no $^+$)	69.26%	88.94%	−0.0001	33%	18.50M

Table 6: ImageNet-1k, ViT-S/16, 100 epochs, one run per condition, all six at batch 128. Top-1/Top-5 at best validation epoch; Δ Overfit is final-epoch minus best-epoch validation loss (negative: still falling at epoch 100).

used as the ImageNet-1k teacher and ceiling throughout Sections 4–4.1 (full setup, compute budget, and four accuracy/overfitting/symmetry hypotheses tested in Appendix A).

The result that matters for this paper's central claim: Q≠K=V — chosen *before* training on exactly this architecture and dataset — reaches 71.67% top-1, landing within 0.7pt of what KEEP_V/AVG *recover to via distillation* after post-hoc surgery on the QKV teacher built from the same recipe (Table 2, 71.93%/71.87%) — close, though the distilled numbers are consistently a little *higher*, not identical, one seed each and no controlled comparison of the two paths.

We read this as the constraint itself not being the obstacle (the joint-trained number needed no repair step to land in this range at all), not as evidence the two paths converge to one exact value. Q≠K=V and QVV(3) close most but not all of the gap to QKV (-2.71 and -1.83 pts), while the fully symmetric Q=K=V$^+$ lags well behind (-7.62 pts); every ΔOverfit is within 0.0014 nats of zero, and none of the six models overfits at this schedule. A follow-up condition, Q=K≠V$^+$, closes a gap in the original design and revises an earlier reading: it matches Q≠K=V's parameter count and cache target exactly, and its gap to Q≠K=V is 2.22pt — still less than half of Q=K=V$^+$'s gap to the same reference (4.91pt), so collapsing V into the shared projection, not symmetry per se, looks like the larger cost. A second, unplanned ablation removing the positional correction from Q=K=V raised accuracy from 66.76% to 69.26%, a 2.50pt swing — still the largest positional-correction effect in the table, and still substantial. Full per-epoch loss curves are in Appendix D; full hypothesis statements and results discussion in Appendix A.

6 Discussion

Distillation as the practical fix, and what its one failure mode reveals. Recovery is fast — most configurations reach the from-scratch ceiling within 1–5 epochs versus the 10–50 needed to train from scratch, consistent with distillation searching for a single linear map rather than re-deriving the whole network. KEEP_K is the counter-case: it starts from the same reduced parameter count as KEEP_V/AVG but is consistently slowest and least complete at CIFAR/LLM scale, less so at ImageNet-1k (Section 4) — an asymmetry more consistent with the basis-mismatch account (Section 4.1) than with Corollary 1 as originally stated: distillation is not restoring content KEEP_K discarded (the probe shows it was never gone), it is re-calibrating downstream weights to an unfamiliar basis. The K→V alignment result sharpens what "re-calibrating to an unfamiliar basis" means concretely: a full distillation pass updates every downstream parameter against a much richer signal (hard labels plus soft logits, many epochs), while the alignment correction updates *nothing* downstream and is a single closed-form fit on far less data — that it still gets most of the way to KEEP_V at nearly every scale suggests the bulk of the deficit is a change-of-basis problem, not one requiring generic re-training. That distillation still converges slower for KEEP_K despite this cheap fix existing is itself informative: gradient descent does not find it directly, or finds it more slowly than KEEP_V/AVG need.

The joint-training confirmation (H1–H4; full statements and evidence in Appendix A). The ImageNet-1k sweep tested four hypotheses: **H1**, shared-projection variants match QKV's accuracy within noise; **H2**, they overfit less, replicating a language-modeling regularization effect; **H3**, symmetric-attention variants underperform less in vision than in language; **H4**, ViT classification parity does not guarantee detection-transfer parity. H1 is not supported (no variant reaches parity); H2 is untestable (no condition overfits under the DeiT augmentation stack, so there is no gap to reduce); H3 — the one hypothesis Section 5's follow-up condition speaks to directly — comes out mixed: collapsing V, not symmetry per se, looks like the larger cost; H4 was not tested (Ap-

pendix A). None of this changes the central point: Q≠K=V trained jointly lands, directly and with no repair step, within a point of what post-hoc KEEP_V/AVG need a distillation pass to reach.

Why this is one picture, not two coincidences. The convex-hull bound (Proposition 1) applies identically to post-hoc surgery and joint training (Remark 1), so the from-scratch number and the recovered ceiling were never guaranteed to land near each other. On the same ImageNet-1k QKV family: choosing $K = V$ *before* training (Q≠K=V, 71.67%) lands within 0.3pt of what merging the same teacher's independent K/V *after* training only reaches after 5 epochs of distillation (KEEP_V: 71.93%, slightly higher, not identical, one seed each and no controlled comparison of the two paths), having first collapsed to near-chance (KEEP_K: 0.08%) the moment the merge happens. Both paths end near the same value, subject to the same bound; only one needed a repair step to get there, and Section 4.1 identifies why: not information loss, but a downstream network calibrated to a basis the merge never produced. We did not run a dedicated ablation isolating this comparison beyond the single side-by-side reading above, so we treat it as a coherence check on the paper's central claim, not an independent controlled result.

What does the jointly-trained shared projection actually resemble? A natural follow-up question is whether a from-scratch c_{kv} converges toward the post-hoc surgery's better mode (KEEP_V) or its catastrophic one (KEEP_K). At synthetic/vision scale the answer is unhelpfully mixed (7 of 18 layer comparisons closer to K, 11 to V, no consistent pattern) — joint training finds its own synthesis rather than converging on either teacher projection. At LLM scale the picture is more consistent, and points the *opposite* way from what KEEP_V's post-hoc advantage might suggest: c_{kv} is closer to the teacher's K in 16 of 20 layers (mean CKA 0.79 vs. 0.67). We read this as reinforcing rather than undermining the central claim: a K-like representation works fine once every downstream weight co-adapts to it during training, while the identical representation is catastrophic under KEEP_K precisely because nothing downstream is allowed to adapt (Appendix C).

7 Related Work

Distillation for repair, not compression. Our recovery protocol (hard-label CE plus KL divergence against soft teacher logits) follows Hinton et al. (2015); that work, and most of the compression literature built on it, uses distillation to *shrink* a model while preserving its architecture class. Here distillation instead repairs a capacity loss introduced by an architectural edit made *after* training, on a student that is the *same size* as (a projection-tied variant of) the teacher — a different use of the same mechanism, closer in spirit to self-distillation or fine-tuning against a stronger reference than to model compression. The central empirical finding of this paper is how cheaply that repair works, and where it does not (KEEP_K).

Projection sharing in attention. Kayyam et al. (2026) systematically evaluate three projection-sharing constraints — Q=K≠V, Q≠K=V, and Q=K=V — and show the last two, which produce symmetric attention maps, can be de-symmetrized via a 2D sinusoidal positional correction; Q≠K=V is highlighted as the most favorable trade-off, 50% KV-cache reduction at a 3.1% perplexity cost. Gao and Xu (2026) re-derive the same base mechanism from a value-space-routing motivation and generalize it via depth-m attention factorization, introducing QVV(3), which matches standard attention's parameter count and, in their experiments, matches or exceeds it on perplexity across five models while halving the KV cache — a result that does not reproduce here for ImageNet-1k top-1 (Section 6). Borji (2023) is an earlier, independent precursor examining a KV-only transformer formulation.

Attention as constrained computation. Tsai et al. (2019) reframe self-attention generally as a kernel smoother, which already implies Proposition 1 as a special case once the vectors being smoothed and the vectors producing the kernel weights are forced to be the same object; our contribution is narrower and architecture-specific — showing what that hull becomes and why it is a genuine bottleneck for some tasks and not others. Dong et al. (2021) show that stacked self-attention without skip connections or MLPs collapses token representations to rank 1 doubly exponentially with depth, a degeneracy unrelated to Proposition 1's single-layer, depth-independent obstruction (our checkpoints keep both skip connections and MLPs throughout). Raganato et al. (2020) show replacing most encoder attention heads with fixed, non-learned patterns barely hurts translation quality — an empirical, training-time finding about how much freedom the *weights* need, whereas ours is a weight-independent statement about what the *value space* can carry once $K = V$, holding for every possible a_{ij}, learned or fixed.

Shared projections in vision transformers. Lee et al. (2021) share nonlinear embedding layers across Q, K, V in tiny ImageNet-scale ViTs (3.1M params) with a code-parameterized sharing scheme; MASA (MASA, 2025) decomposes attention projections into a shared dictionary of matrix atoms, cutting attention parameters by 66.7% with on-par performance. Neither targets KV-cache reduction or ViT-B/L-scale training on full ImageNet-1k, and neither reports an overfitting analysis.

KV-cache reduction and KV-sharing architectures. Multi-Query (Shazeer, 2019) and Grouped-Query (Ainslie et al., 2023) Attention reduce the KV cache by sharing keys/values across heads, and Multi-Head Latent Attention (DeepSeek-AI, 2024) compresses keys and values into a shared low-rank latent — orthogonal to projection sharing, which instead removes redundancy between the key and value *projections themselves*; the two families compose. $K{=}V$ tying is a special case of this broader, well-established family; what this paper adds is a mechanism for *why* imposing that same sharing structure *after* independent training is a qualitatively different (and, we show, cheaply repairable) operation from choosing it *before* training starts.

Representational similarity. Our CKA measurements follow Kornblith et al. (2019); that raw weight-cosine similarity is uninformative while activation-level CKA is not is the expected consequence of CKA's basis invariance. We adopt training conventions from Dosovitskiy et al. (2021) and Touvron et al. (2021) for the ImageNet-1k confirmatory sweep (Sec 5); DeiT's data-efficient recipe is essential for training a ViT from scratch on ImageNet-1k without ImageNet-21k pretraining.

8 Limitations

Post-hoc surgery and distillation recovery. The K–V CKA diagnostic is not a validated general predictor of collapse severity beyond the $n = 9$ synthetic/vision tasks it was fit on: the correlation weakens under rank statistics, is not separately significant within either task-type group, and reverses direction *within* the vision datasets, unexplained. Two of six mechanism tests (the linear content probe, the continuous α-sweep beyond its endpoint) were not extended to ImageNet-1k; the four that were, and the ImageNet-1k surgery result itself, are each a single scale point, one seed. The markedly smaller KEEP_K gap at ImageNet-1k scale is reported without an explanation.

Joint-training confirmation. The ImageNet-1k sweep (Section 5) covers a single architecture (ViT-S/16), one seed per condition, and one compressed schedule (100 epochs); the ImageNet-100 pilot, the ViT-B/16 confirmatory run, and the detection-transfer experiment (H4) were all skipped, and we cannot report seed variance or confirm findings hold at larger scale or the full 300-epoch DeiT recipe. All six conditions are matched at batch 128 (Table 6). As with Kayyam et al. (2026), the compressed schedule means absolute accuracy should not be compared to published SOTA or ImageNet-21k-pretrained baselines; the same compression likely explains the null H2 result. The cross-study observation in Sec. 6 is a coherence check on existing tables, not a controlled experiment isolating how the joint-training gap and the post-hoc collapse-and-recovery pattern relate quantitatively.

9 Conclusion

Post-hoc $K{=}V$ surgery on an independently-trained QKV model is catastrophic — Proposition 1's elementary but consequential observation is that a single attention sub-block cannot then choose its content-transport representation independently of its addressing representation, and every routing-dependent task we test collapses toward chance the moment the merge happens. But that damage is cheap to repair: a short distillation pass against the original model recovers nearly every (task, mode) configuration to within a point of the from-scratch ceiling in 5–10 epochs, at every scale from five synthetic tasks up to a 300M-parameter language model and a new ImageNet-1k ViT teacher. The one consistent exception, KEEP_K, is the paper's most informative result precisely because of *why* it is slow: not information loss (a linear probe shows the content was never gone) but representation incompatibility, confirmed directly by a cheap, closed-form $K \to V$ basis correction that recovers most of the gap with no retraining at all. A confirmatory ImageNet-1k sweep closes the loop: the same constraint, chosen *before* training on the same architecture and data, lands within a point of this recovered ceiling directly, with no repair pass needed. Read together, the evidence is consistent with one picture — a shared $K{=}V$ projection is easy for a network to accommodate when it is optimized around it from the start, and becomes a repair problem, not an unsolvable one, only when imposed on representations that specialized apart under no such pressure.

AI USE STATEMENT

We used generative AI tools (Claude, Anthropic) to assist with drafting and editing manuscript text, merging two independently-authored source documents into this paper, writing and adapting the Python analysis scripts that extend mechanism tests to the ImageNet-1k teacher (Section 4.1), and reformatting to ICLR requirements. We checked the merged text against both originals for numerical consistency and sanity-checked analysis scripts against known baselines (e.g., re-deriving the unmodified teacher's recorded accuracy) before reporting their outputs. We did not use generative AI to design the core experiments, state the hypotheses, or derive Proposition 1 and its proof. All AI-assisted content was reviewed by the author(s), who take responsibility for the final work.

REPRODUCIBILITY STATEMENT

The ViT / attention implementation, training recipe, and all analysis scripts referenced in this paper (CKA similarity, $K \to V$ alignment, addressing/payload decomposition, random-projection baseline) will be made available at an anonymized code link: `[ANONYMIZED CODE URL -- TO BE ADDED BEFORE SUBMISSION]`. Section 5 and Appendix A give the full training configuration (architecture, optimizer, schedule, augmentation, hardware, batch sizes) needed to reproduce Table 6, as well as the post-hoc surgery and $K \to V$ alignment construction details; Appendix B gives the full synthetic-task results table. Full per-epoch training curves for all six ImageNet-1k conditions are in Appendix D; per-epoch distillation recovery curves and the continuous zero-shot α-sweep are in Appendices F and G.

REFERENCES

J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *EMNLP*, 2023.

A. Borji. Key-value transformer. *arXiv preprint arXiv:2305.19129*, 2023.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Y. Dong, J.-B. Cordonnier, and A. Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *ICML*, 2021.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

X. Gao and X. Xu. Keyless attention: Value-space routing and value-only caching for efficient transformers. *arXiv preprint arXiv:2606.21848*, 2026.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

A. Kayyam, A. Madan Gopal, and M. A. Lewis. Do transformers need three projections? Systematic study of QKV variants. In *Forty-third International Conference on Machine Learning (ICML)*, 2026.

S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. In *ICML*, 2019.

J. Lee et al. Rethinking query, key, and value embedding in vision transformer under tiny model constraints. *arXiv preprint arXiv:2111.10017*, 2021.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Anonymous. Share your attention: Transformer weight sharing via matrix-based dictionary learning. *arXiv preprint arXiv:2508.04581*, 2025.

A. Raganato, Y. Scherrer, and J. Tiedemann. Fixed encoder self-attention patterns in transformer-based machine translation. In *Findings of EMNLP*, 2020.

N. Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *EMNLP*, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

A Full Experimental Setup

Architectures. ViT-S/16 (22M params, 12 layers, hidden 384, 6 heads) is the primary model for the joint-training sweep. ViT-B/16 (86M) and ViT-L/16 (307M) confirmatory runs at larger scale were part of the original design but descoped for this iteration (Section 8).

Datasets and schedule. The original two-phase plan — an ImageNet-100 pilot sweep over five conditions, then an ImageNet-1k confirmatory run on the 1–2 best pilot variants at ViT-B/16(-L/16) scale, 100–150 epochs — was descoped before execution. What was actually run: only the ImageNet-1k phase, only ViT-S/16, one seed per condition. Of the five originally-planned conditions, four were run in the first batch — QKV, Q≠K=V, Q=K=V^+, QVV(3) — omitting Q=K≠V^+ to keep the sweep to what fits on four GPUs at once (RTX 2080 Ti, 11GB, one condition per GPU, batch 128, the largest that fit). Measured throughput was ∼575 img/s/GPU, ∼3320s/epoch, ∼92 GPU-hours per condition; running all four concurrently kept total wall-clock to that same ∼92 hours rather than $4\times$ it. A second batch trained the two remaining conditions — Q=K≠V^+ and an unplanned ablation removing the positional correction from Q=K=V — sequentially on a single GPU, also at batch 128, matching the first four directly.

Evaluation metrics. Top-1/top-5 accuracy at best-validation epoch (mirrors Table 1 of Kayyam et al. 2026 and Gao and Xu 2026); overfitting gap Δ = final-epoch minus best-epoch validation loss, directly testing H2; activation/cache memory reduction, both absolute and as a fraction of standard-attention memory (target: 50% for Q≠K=V/QVV(3), as in prior work); inference throughput (images/sec, fixed batch), matching the decode-throughput methodology of Gao and Xu (2026) adapted to non-autoregressive ViT inference.

Hypotheses. H1 (accuracy parity). Q≠K=V and QVV(3) match standard ViT-B/16 top-1 accuracy on ImageNet-1k within noise ($\lesssim$0.5pt), consistent with the language-modeling degradation being small (3.1%) and vision transformers being comparatively more over-parameterized relative to ImageNet-1k than LLMs are relative to their pretraining corpora. **H2 (overfitting reduction transfers).** Shared-projection variants show a smaller train/val generalization gap than standard QKV at matched epoch count, replicating the gradient-entanglement regularization effect from Gao and Xu (2026) outside the causal LM setting. **H3 (symmetric variants underperform more in vision than language).** Because 2D spatial relationships in images are less inherently directional than causal token order in language, we expect the gap between corrected symmetric variants (Q=K≠V^+, Q=K=V^+) and asymmetric variants (Q≠K=V, QVV(3)) to be *smaller* in vision than reported in Kayyam et al. (2026) for language. **H4 (transfer to detection is architecture-sensitive).** We expect classification parity to *not* guarantee detection-transfer parity, since localization loss is more sensitive to fine spatial attention structure.

Per-hypothesis evidence. H1 — not supported at this schedule. None of the three shared-projection variants comes within the $\lesssim$0.5pt target. In relative terms, Q≠K=V's 2.71pt drop (3.6% relative) is close to the 3.1% perplexity cost reported for the analogous language-modeling variant in Kayyam et al. (2026), and QVV(3)'s 1.83pt drop (2.5% relative) is similarly in that range — but Gao and Xu (2026) report QVV(3) *matching or exceeding* standard attention on language perplexity, which does not reproduce here for ImageNet top-1. Q=K=V^+'s 7.62pt drop (10.2% relative) clearly exceeds the LM-scale gap. We cannot rule out that a full 300-epoch DeiT schedule (rather than the 100-epoch compressed one used here) would narrow these gaps. **H2 — untestable at this schedule.**

Every condition, including the QKV baseline, shows $|\Delta\text{Overfit}| \leq 0.0014$ nats (Table 6): validation loss is still flat-to-falling at epoch 100 across all six batch-128 models. The DeiT augmentation stack evidently suppresses overfitting at this schedule regardless of variant, unlike the 36-layer LM setting in Gao and Xu (2026) where standard attention showed a substantially larger gap (0.8404 vs. 0.6866 for QVV(3)); since the baseline itself doesn't overfit, there is no gap for shared-projection attention to reduce, and a lower-augmentation or longer-schedule run would be needed to test H2 properly. **H3 — mixed; collapse, not symmetry, looks like the driver.** Q=K≠V$^+$'s gap to its matched-capacity asymmetric analog Q≠K=V is 2.22pt, still less than half of Q=K=V$^+$'s gap to the same reference (4.91pt) despite both producing a symmetric attention map — closer to H3's original motivation than the all-symmetric-variants verdict a single data point previously supported. **H4 — untested.** No detection experiment was run this iteration; H4 remains open.

2D positional correction. Kayyam et al. (2026) introduce a 2D sinusoidal positional correction to de-symmetrize the Q=K≠V and Q=K=V attention maps in language modeling, where the 2D structure is an artificial construction on a 1D sequence. In a ViT, patches are natively on a 2D grid ($H/p \times W/p$ for patch size p), so the correction uses the model's actual patch coordinates: $\tilde{S}_{ij} = S_{ij} + \mathrm{Conv}_{1\times1}(P(\mathrm{row}_i, \mathrm{col}_i, \mathrm{row}_j, \mathrm{col}_j))$, where P is a fixed 2D relative-position sinusoidal embedding and S is the symmetric raw score matrix — a stronger test of the correction's design intent than the language setting, and comparable to 2D relative-position-bias schemes in windowed ViTs (e.g. Swin, Liu et al. 2021).

Post-hoc surgery construction (detail). Given a trained `QKV` checkpoint, we construct a `Q≠K=V` student by setting the shared projection c_{kv} to KEEP_K ($c_{kv} \leftarrow c_k$), KEEP_V ($c_{kv} \leftarrow c_v$), or AVG ($c_{kv} \leftarrow \frac{1}{2}(c_k + c_v)$, averaging the two Linear layers' weights and biases — since both are affine maps this is identical to averaging their outputs); every other parameter (c_q, c_{proj}, MLPs, layer norms, embeddings, head) is copied verbatim.

Remark 2 (Invariance to normalization)**.** *Proposition 1's proof uses only that $a_{ij} \geq 0$ and $\sum_j a_{ij} = 1$; it places no assumption on how k_j is computed, so it holds unchanged whether k_j is a raw linear projection, an ℓ_2-normalized one (as in QK-normalized attention), or any other fixed transformation applied before the weighted sum. In the checkpoints studied here, LayerNorm is applied pre-block to the residual stream, not to k_j or v_j themselves, so this is a non-issue for our experiments specifically; we state the general invariance because it is easy to construct architectures where it would otherwise look like a threat to the argument.*

B Full Per-Task Table: Synthetic Tasks

Task	Mode	QKV	Q≠K=V (scratch)	Zero-shot surgery	Distilled (5 ep)
REVERSE	keep_k	1.000	1.000	0.100	1.000
	keep_v	1.000	1.000	0.123	1.000
	avg	1.000	1.000	0.861	1.000
SORT	keep_k	0.998	0.996	0.583	0.965
	keep_v	0.998	0.996	0.717	0.996
	avg	0.998	0.996	0.751	0.996
SUB	any	1.000	1.000	1.000	1.000
SWAP	keep_k	1.000	1.000	0.088	1.000
	keep_v	1.000	1.000	0.122	1.000
	avg	1.000	1.000	0.906	1.000
COPY	any	1.000	1.000	1.000	1.000

Table 7: Synthetic sequence tasks (token accuracy). Routing-dependent tasks (REVERSE, SORT, SWAP) collapse under surgery and recover under distillation; pointwise tasks (SUB, COPY) are unaffected by either, as Section 3 predicts. 14 of 15 (task, mode) configurations reach the from-scratch ceiling within 5 epochs; the one exception, SORT/KEEP_K, reaches 0.965 against a 0.996 target and is still improving at epoch 5.

C Extended Mechanism Discussion

What does the K→V correction actually do? We check by computing activation CKA between the corrected K and the untouched V on held-out data disjoint from whatever the map was fit on.

The linear map raises CKA(K,V) from a baseline of 0.73 to 0.99 at synthetic/vision scale, and from 0.69 to 0.92 at LLM scale — essentially saturating the measure. Procrustes, by contrast, moves it far less (0.73→0.76 small-scale, 0.69→0.72 at LLM scale) despite comparable or better zero-shot recovery (it beats the linear map outright on CIFAR-10/100) — representational similarity and functional recovery are not the same axis here. The unconstrained linear map's singular values are informative about why it reaches such high CKA: their mean is well below 1 (synthetic/vision: 0.33; LLM: 0.016), rather than clustered near 1 as a near-rotation would be — the fit aggressively projects onto and rescales a small number of directions rather than finding a full change of basis, consistent with the useful $K \to V$ relationship living in a low-dimensional subspace.

Does joint training's shared projection resemble K, V, or neither? At synthetic/vision scale the answer is unhelpfully mixed: across 9 tasks/datasets (18 layer comparisons), the from-scratch model's c_{kv} is closer to K in 7 and to V in 11, with no consistent pattern, and sometimes less similar to either than K and V are to each other — joint training finding its own synthesis. At LLM scale the picture is more consistent, and points the other way from what the KEEP_V-favoring results might suggest: c_{kv} is closer to the teacher's K in 16 of 20 layers (mean CKA 0.79 vs. 0.67). We read this as informative but not contradictory: a K-like representation works fine when every downstream weight co-adapts to it during training, while the identical representation fails catastrophically under KEEP_K precisely because nothing downstream is allowed to adapt.

CKA layer-wise breakdown. Every model in the CKA diagnostic has exactly two attention layers. In the three routing-dependent synthetic tasks, CKA is *lower* in layer 0 than layer 1 (REVERSE 0.54→0.82; SORT 0.56→0.74; SWAP 0.54→0.83): K/V specialize early and converge back toward redundancy. In vision the trend reverses — high in layer 0, dropping in layer 1, the layer feeding the CLS readout (MNIST 0.74→0.33; FMNIST 0.77→0.57; CIFAR-10 0.83→0.58; CIFAR-100 0.79→0.70), and layer 1's ordering is exactly monotonic in dataset difficulty — the across-dataset trend in Figure 2 is carried almost entirely by the layer adjacent to classification. One plausible (unverified) reading: harder classification tasks lean more on the class token aggregating broad, low-precision context, a use of attention Proposition 1 restricts regardless of K/V similarity. ViT-S/16's 12 layers make this breakdown impractical to report compactly for the ImageNet-1k teacher.

Robustness. Recomputing every CKA measurement with an RBF kernel (median pairwise-distance bandwidth) instead of linear leaves the global correlation essentially unchanged ($r = -0.87$ vs. -0.84, $n = 9$, both $p < 0.01$) and preserves the vision layer-1 ordering. Weight-space cosine similarity between the raw (W_k, b_k) and (W_v, b_v) parameters is uninformative in every task (all values within ± 0.005 of zero) — expected, since two linear maps can compute redundant *functions* while living in unrelated parameter bases; only the representation-level (CKA) comparison is basis-invariant.

Addressing diagnostic. For each synthetic-task layer, we take the model's actual attention weights and check whether $\arg\max_j a_{ij}$ matches the task's ground-truth source position. On the two one-to-one permutation tasks this is strong (REVERSE 1.00→0.83 layer 0→1; SWAP 1.00→0.78); on SUB/COPY it is near chance at both layers, consistent with the pointwise-immunity argument of Section 3. SORT is also near chance under this diagnostic despite collapsing under surgery like REVERSE/SWAP, which we read as a limitation of "does argmax match one ground-truth index" as an operationalization for a task whose natural solution is a distributed rank computation, not evidence about SORT's mechanism.

D IMAGENET-1K TRAINING CURVES

Table 6 reports only best- and final-epoch accuracy; Figure 3 shows the full 100-epoch trajectory for all six conditions. Four conditions' per-epoch CSV logs were truncated after epoch 11 by a logging issue unrelated to training itself (the runs completed all 100 epochs, and their final printed top-1 matches Table 6 exactly), so those four curves are parsed from the raw stdout training logs instead; the two follow-up conditions' CSVs were unaffected and are read directly.

E LLM DISTILLATION CURVES

Table 3 reports only zero-shot and final-budget perplexity; Figure 4 shows the full recovery trajectory for all three surgery modes.

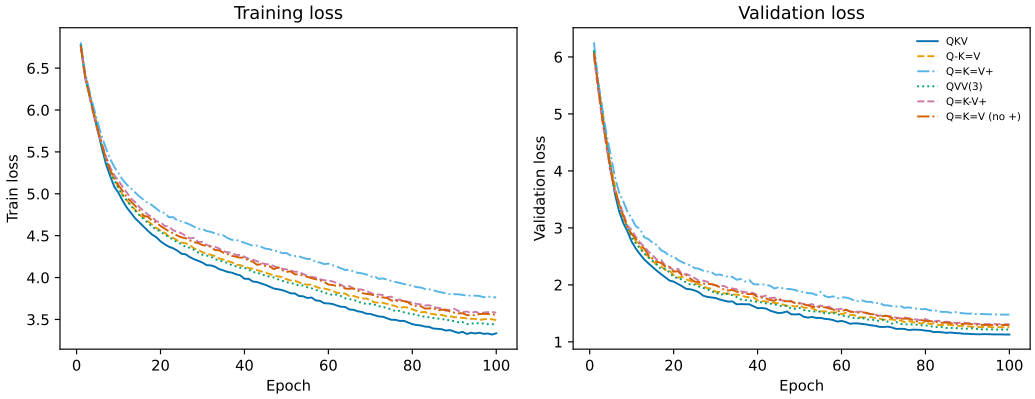


Figure 3: Training loss (left) and validation loss (right) per epoch, all six ImageNet-1k conditions. The ordering tracks final top-1 accuracy throughout training: QKV is lowest-loss almost everywhere, QVV(3) closely trails it, Q≠K=V/Q=K≠V$^+$/Q=K=V (no $^+$) cluster in the middle, and the fully symmetric Q=K=V$^+$ is highest-loss from early on — the accuracy gap is visible almost from the start of training, not a late divergence.

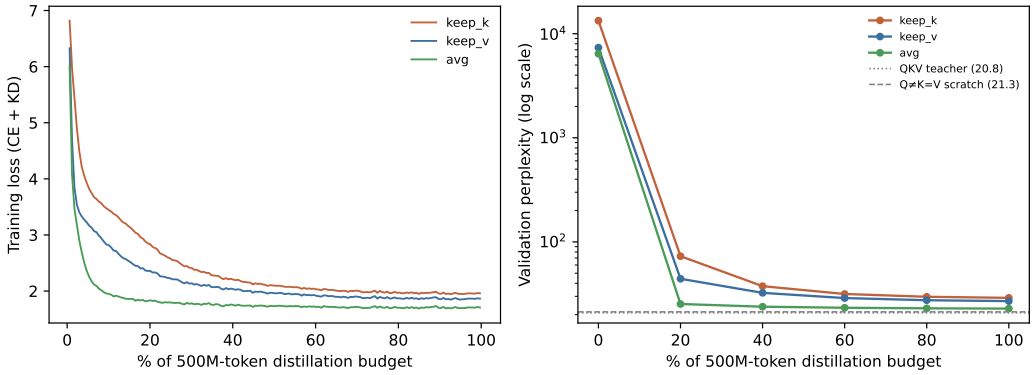


Figure 4: Left: training loss (hard-label CE + KD against the QKV teacher's soft logits) over the 500M-token distillation budget. Right: validation perplexity at the zero-shot point and each of 5 evaluation checkpoints (log scale), against the QKV teacher and from-scratch ceiling (reference lines). All three modes make most of their recovery in the first 20% of the budget; AVG and KEEP_V track closely throughout, while KEEP_K remains visibly separated at every checkpoint.

F Synthetic and Vision Distillation Curves

Tables 7 and 2 report only endpoint accuracies; Figure 5 shows the full per-epoch recovery curve for every (task or dataset, surgery-mode) configuration referenced above (synthetic/vision scale only; the ImageNet-1k addition was not separately curve-logged). REVERSE/SUB/SWAP/COPY and MNIST are already flat by epoch 1 (the surgery was mild or the gap closes almost immediately), SORT/KEEP_K and every CIFAR/KEEP_K curve are still rising through their full epoch budget without having visibly plateaued as early as epoch 2–3, and KEEP_V/AVG on CIFAR flatten out well before their budget ends — consistent with KEEP_K being a harder optimization problem rather than merely a deeper hole to climb out of.

G Zero-Shot Alpha Sweep

Tables 7 and 2 report only three points along the line $c_{kv} = \alpha W_K + (1-\alpha)W_V$ (α=1: KEEP_K, α=0: KEEP_V, α=0.5: AVG). Figure 6 sweeps $\alpha \in \{0, 0.1, \ldots, 1.0\}$ and adds a fourth reference point: c_{kv} left at its fresh, untrained initialization, with no relationship to the teacher's K or V at all — reported as a mean $\pm$ 1 std. over 5 random seeds, not a single number, since the base-

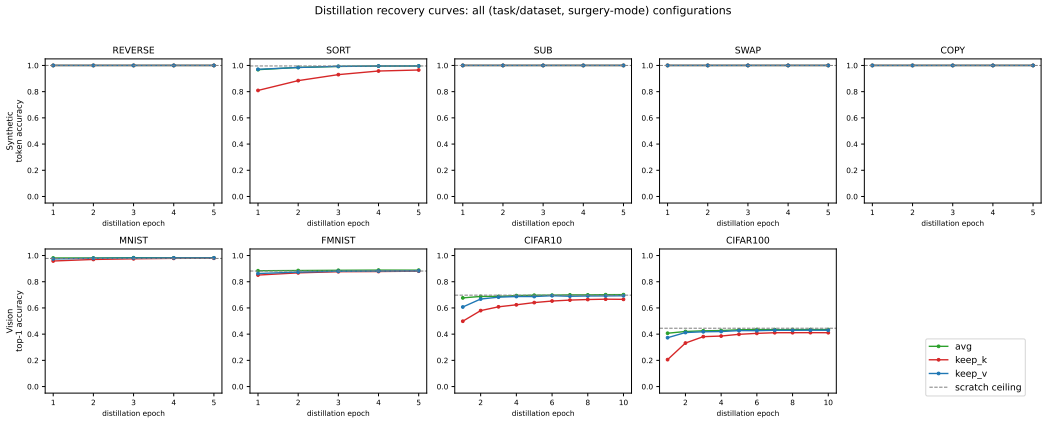


Figure 5: Per-epoch distillation accuracy for every (task/dataset, surgery-mode) configuration in Tables 7–2 (dashed line: the from-scratch Q≠K=V tied baseline). CIFAR-10/CIFAR-100 panels show the full 10-epoch runs; all other panels are the 5-epoch runs (not rerun longer since already converged).

line's own sampling variance turns out to matter for what can be claimed about it. (The continuous sweep was run at synthetic/vision scale only; the random-baseline *endpoint* — 5 seeds, no interpolation — was separately extended to the ImageNet-1k QKV teacher and is reported in Table 4's last row.) The curve is neither monotonic nor symmetric around AVG: REVERSE, SWAP, and all four (synthetic/vision-scale) vision datasets peak near $\alpha \approx 0.4$–0.5, while SORT peaks earlier, around $\alpha \approx 0.2$–0.3, and degrades monotonically from there to KEEP_K. SUB and COPY are flat at 1.000 across every α, including every random seed.

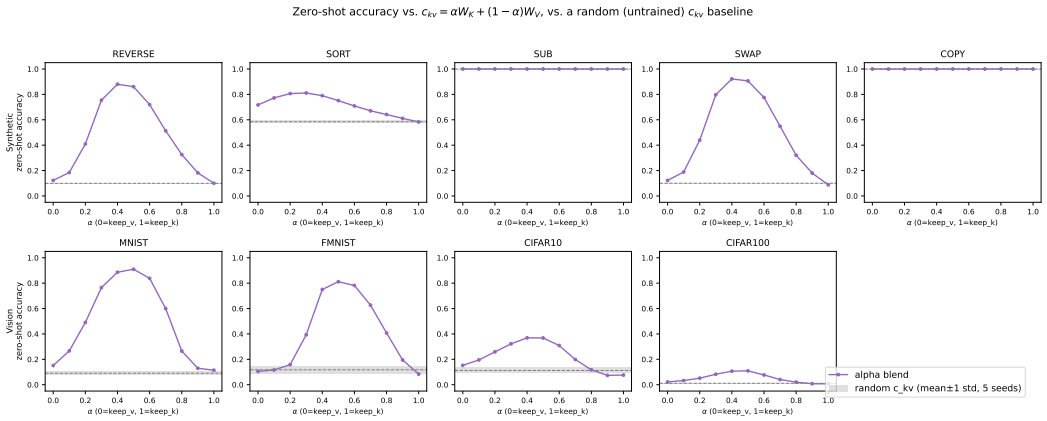


Figure 6: Zero-shot accuracy as c_{kv} is swept from KEEP_V (α=0) to KEEP_K (α=1), vs. a random-initialization c_{kv} baseline (shaded: mean $\pm$ 1 std. over 5 seeds). No retraining anywhere in this figure.