What Transfers from Non-Visual Procedural Warm-Up? Attention Operators as Structured Initializers for Vision Transformers

Anonymous authors
Paper under double-blind review

Abstract

Non-visual procedural warm-up can improve subsequent Vision Transformer (ViT) training despite having no visual or semantic overlap with images. We ask which initialization structure retains this benefit. A ViT is first warmed up for 15,000 steps on DYCK-64, after which we progressively remove warm-up parameters. The resulting initializer derives fresh projection matrices from selected late-block, bias-free query–key (QK) and output–value (OV) weight operators and randomly initializes the remaining model. Across three matched source/downstream seeds, it reaches $75.17 \pm 0.30\%$ on CIFAR-100 and $60.41 \pm 0.64\%$ on Tiny-ImageNet, versus $74.74 \pm 0.51\%$ and $60.30 \pm 0.70\%$ for full DYCK transfer. Relative to scratch, these point estimates correspond to 121.00% and 103.00% recovery of the full-warm-up gain. A ViT-Base/ImageNet-1K coverage screen finds that the last 6, 8, 10, and 12 blocks reach 79.86%, 80.40%, 80.27%, and 80.54% Top-1, versus 77.72% from scratch and 79.50% from full warm-up; three-seed confirmation of last-eight yields $80.32 \pm 0.12\%$. Supervised-source controls show that equal operators can yield different outcomes under alternative factor coordinates, while the same QK+OV intervention recovers only 15.02% of mature visual transfer. Together, the results identify a source-dependent operator initializer that retains the benefit of short non-visual procedural warm-up from CIFAR-100 to ImageNet-1K.

1 Introduction

Vision Transformers (ViTs) are normally initialized randomly or from a model trained on natural images (Dosovitskiy et al., 2021; Touvron et al., 2021). Procedural warm-up offers a striking third route: before seeing any image, a transformer is briefly trained on sequences generated by a formal grammar. This small, non-visual prefix improves subsequent image training, including at ImageNet-1K scale (Shinnick et al., 2025). Related effects now also appear in language-model training, where small amounts of abstract procedural data improve later learning on language, code, and mathematics (Jiang et al., 2026). Because a DYCK grammar shares neither pixels nor semantic labels with image datasets, it provides a controlled setting for asking what useful initialization structure survives the modality change.

Yet the downstream effect alone does not reveal what the warm-up changes. A full checkpoint entangles attention maps, MLPs, normalization, embeddings, and their interactions, so successful full-model reuse cannot identify which structure carries the gain. Correlations and descriptive weight statistics are likewise insufficient: a component may change during warm-up without being responsible for improved downstream optimization. We therefore ask three linked questions: how much of the gain remains after most warmed parameters are removed; whether the retained structure can be expressed as attention operators and refactorized into newly initialized projections; and whether the resulting initializer remains effective as the target grows from CIFAR-100 and Tiny-ImageNet to ImageNet-1K.

We study that question by intervention. Starting from a ViT warmed up for 15,000 steps on DYCK-64, we discard progressively more parameters and retain only selected self-attention weight products. For an attention head, the query–key (QK) product controls the bias-free bilinear logits and the output–value (OV) product controls the bias-free value-to-residual linear map (Elhage et al., 2021). We

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

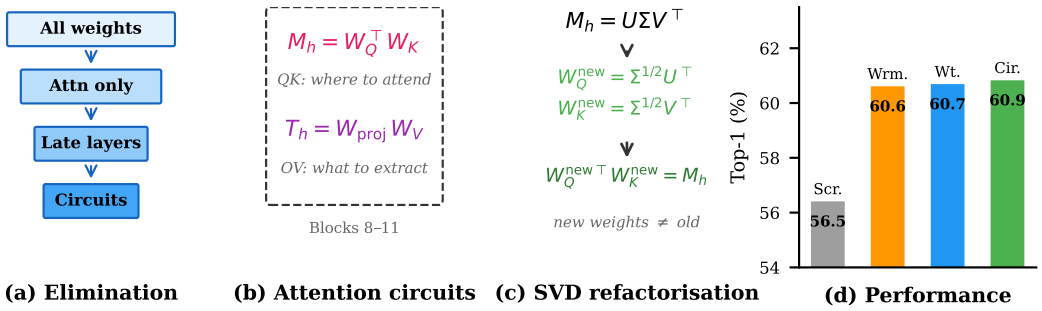


Figure 1: **Overview.** Elimination narrows the retained initialization from the full warm-up to late attention products. QK and OV products are refactorized into fresh projection matrices before downstream image training. Panel (b) follows the checkpoint's stored-matrix convention; Equations 1–2 use the equivalent forward-pass orientation. "Circuits" in the graphic denotes these products, not a complete causal circuit. Panel (d) summarizes intervention selection; confirmatory results appear in Table 1.

extract these operators and factor them into new projection matrices. All other target parameters start randomly. This separates retaining an operator from copying its original coordinates or the rest of the warmed-up model.

QK and OV are useful intervention units because they are head-local, functionally interpretable, and admit factorizations that preserve their bias-free maps while replacing the underlying projection matrices. The resulting test is stronger than comparing checkpoints descriptively: if a downstream benefit persists after the rest of the model is reset, the retained structure is sufficient to convey that benefit under the evaluated training protocol. Comparing sources and alternative exact factorizations then tests whether this explanation is source-independent or invariant to parameter coordinates.

Our evidence gives two complementary answers. First, across three matched seeds, late QK+OV initialization reproduces the full DYCK warm-up point estimate on CIFAR-100 and Tiny-ImageNet. On ViT-Base/ImageNet-1K, extending coverage beyond the last four blocks closes the initial last-four gap, and a selected last-eight setting is stable across three downstream seeds. Second, this is not a generic account of conventional visual pretraining: applied to a supervised ImageNet DeiT checkpoint, late QK+OV retains only a small part of full-backbone transfer. Alternative exact factorizations further show that a fixed operator does not uniquely determine training behavior, and a random-direction control prevents attributing the supervised-source effect to learned directions.

Taken together, these results support a deliberately scoped claim: a short DYCK warm-up yields a reusable attention-based initializer. Transferring and refactorizing only late-layer QK and OV operators retains the warm-up benefit while the rest of the target model is randomly initialized. This conclusion concerns short procedural warm-up rather than all forms of pretraining: supervised visual pretraining exhibits a different transfer pattern, and exact-gauge controls show that downstream optimization depends on both the transferred operators and their factor coordinates.

Our contributions are:

- We formulate operator refactorization as a controlled intervention on the initialization left by short non-visual procedural warm-up, with an explicit bias limitation and a recovery metric relative to scratch and full warm-up.
- We provide three-seed evidence on CIFAR-100 and Tiny-ImageNet, a 300-epoch ViT-Base/ImageNet-1K coverage screen, and three-seed confirmation of its last-eight condition. QK+OV-derived initialization reproduces the observed DYCK warm-up benefit from small datasets to ImageNet-1K.
- We identify empirical boundaries: suffix coverage is non-monotonic; exact operator-preserving gauges yield different downstream outcomes; and QK+OV retains only 15.02% of mature supervised-visual transfer gain.

2 Related work

Procedural warm-up. Procedural data can teach useful computation without natural-data content. A brief abstract warm-up improves later vision training (Shinnick et al., 2025), and a small procedural prefix similarly benefits language, code, and mathematics (Jiang et al., 2026). Curricula can reshape optimization (Hacohen & Weinshall, 2019), while formal-language studies characterize what transformers learn and generalize (Bhattamishra et al., 2020; Delétang et al., 2023). We instead use DYCK-64 as a controlled source and intervene on its weights before image training.

What transfers from a trained network? Transfer studies examine feature reuse, layer replacement, and representation similarity (Yosinski et al., 2014; Kornblith et al., 2019; Neyshabur et al., 2020; Raghu et al., 2021). Prior work reuses fixed deep activations (Donahue et al., 2014), maps relations among visual tasks (Zamir et al., 2018), and shows that fine-tuning can distort useful features under shift (Kumar et al., 2022). We instead reinitialize nearly the whole target and retain only selected projection-weight products.

Vision transformers and pretraining. ViT adapts the Transformer to image patches (Vaswani et al., 2017; Dosovitskiy et al., 2021). Its outcomes depend on data, regularization, and scale (Steiner et al., 2022; Zhai et al., 2022), while Swin extends the family hierarchically (Liu et al., 2021). Visual pretraining spans supervised, contrastive, masked-image, and image–language objectives (Touvron et al., 2021; Chen et al., 2021; Bao et al., 2022; He et al., 2022; Radford et al., 2021); self-supervised attention can also exhibit object structure (Caron et al., 2021). These mature sources provide a boundary for, rather than an assumption about, procedural transfer.

Attention operators and circuits. Circuit notation separates QK and OV operators (Elhage et al., 2021; Olsson et al., 2022). Mechanistic work treats circuits as computational subgraphs (Olah et al., 2020), identifies them manually, tracks their formation, or automates their discovery (Wang et al., 2023; Nanda et al., 2023; Conmy et al., 2023); related analyses assign roles to feed-forward layers (Geva et al., 2021). Here, QK/OV products are controlled intervention targets, not a claimed complete causal circuit for image predictions.

Initialization and parameterization. Initialization controls signal propagation (Glorot & Bengio, 2010), and deep linear models expose nontrivial factor-level dynamics (Saxe et al., 2014). Deep matrix factorization and low-rank adaptation are broader factorized-optimization examples (Arora et al., 2019; Hu et al., 2022). Function-preserving symmetries distinguish functions from coordinates (Godfrey et al., 2022); permutation alignment likewise changes weight-space connectivity (Ainsworth et al., 2023). Our gauges test this distinction specifically for QK/OV factors.

3 Setup and intervention

3.1 Warm-up and downstream training

Our controlled source is a standard 12-block ViT warmed up on sequences from the DYCK-64 balanced-parenthesis language. The transformer receives token embeddings directly, bypassing the visual patch embedding, and is optimized by masked-token prediction for 15,000 steps. Source AdamW uses $(\beta_1, \beta_2) = (0.9, 0.999)$, cosine decay, peak learning rate 5×10^{-4}, and global batch 256; its token embedding and prediction head are never transferred.

Downstream models replace the token interface with a visual patch embedding and train for 300 epochs. Main experiments use AdamW, base learning rate 2×10^{-3}, weight decay 0.05, cosine decay, 50 warm-up epochs, gradient clipping at 1.0, and global batch 512. DeiT-style augmentation (Touvron et al., 2021) combines Mixup 0.8, CutMix 1.0, label smoothing 0.1, and RandAugment (two operations, magnitude 9) on 224×224 bicubic crops. Comparisons match architecture, schedule, augmentation, and data order. Following the same protocol for every method, we report the best recorded Top-1 on the canonical evaluation split (every epoch for the smaller datasets and every 10 epochs for ImageNet) over 300 epochs. We call this stage *downstream image training*, because most target parameters start from scratch.

The primary small-dataset architecture is ViT-Tiny/16 (12 blocks, three heads, embedding width 192). The scale study uses ViT-Base/16 (12 blocks, 12 heads, width 768) on ImageNet-1K (Deng et al., 2009). CIFAR-100 follows its public 50k/10k train/test split (Krizhevsky, 2009); Tiny-ImageNet uses its public 100k-image training and 10k-image, 200-class validation splits (Le & Yang, 2015), mapping validation annotations to class IDs before the same transforms. ImageNet-1K uses the standard approximately 1.28M/50k train/validation split. Evaluation images are never optimized on.

3.2 Bias-free attention weight operators

For input tokens $X \in \mathbb{R}^{n\times d}$ and head width d_h, use $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_h}$ and the head's output block $W_O \in \mathbb{R}^{d_h\times d}$. Omitting biases, the head contributes

$$A(X) = \mathrm{softmax}\left(\frac{X\,M^{\mathrm{QK}}X^\top}{\sqrt{d_h}}\right), \qquad Y(X) = A(X)X\,M^{\mathrm{OV}}, \tag{1}$$

where

$$M^{\mathrm{QK}} = W_Q W_K^\top \in \mathbb{R}^{d\times d}, \qquad M^{\mathrm{OV}} = W_V W_O \in \mathbb{R}^{d\times d}. \tag{2}$$

For a fused multi-head output projection $P = [P_1^\top; \ldots; P_H^\top]$ under an implementation's storage convention, W_O is the corresponding per-head block P_h after transposition as needed. All extraction and audits use the actual forward-pass orientation, avoiding ambiguity from storage layout.

These two products determine the head's *bias-free weight operations*, not the entire attention module. The source has QKV and output biases. We do not copy them; target biases retain their default initialization. Positional and class tokens, patch embedding, LayerNorm, MLPs, and classifier are also randomly initialized unless a baseline explicitly transfers the full model.

3.3 Refactorization and controls

Each per-head product has rank at most d_h. Let $M \approx U_r\Sigma_r V_r^\top$ be its thin SVD with $r = d_h$, the architectural rank bound. The balanced construction sets

$$L = U_r\Sigma_r^{1/2}, \qquad R = \Sigma_r^{1/2}V_r^\top, \quad \text{so that} \quad LR \approx M \tag{3}$$

We assign (L, R) to $(W_Q, W_K^\top)$ for QK and to (W_V, W_O) for OV. All head-width components are retained. Applying this construction independently to each selected head while leaving other parameters random defines *QK+OV initialization*. QK-only and OV-only conditions assign only the corresponding pair.

The factorization is non-unique. For any invertible G, $(LG)(G^{-1}R) = M$. We draw $G = Q_1\,\mathrm{diag}(s)Q_2^\top$ from seeded Gaussian QR factors. Setting $s = 1$ gives an orthogonal gauge; log-spaced s with ratio 4 gives the condition-number-4 gauge. A separate control retains each product's singular values and balanced factor geometry but independently randomizes its left and right singular subspaces. It therefore changes the operator. Factor draws use a dedicated seed and do not alter target-model or data-order randomness.

3.4 Why equal products can optimize differently

Refactorization preserves the initial operator but not necessarily its optimization geometry. Let $M = LR$ and let $H = \nabla_M \ell$ denote the loss gradient with respect to the product. One gradient-descent step on the two factors gives, to first order in the learning rate η,

$$\Delta M \approx -\eta\big(HR^\top R + LL^\top H\big). \tag{4}$$

Thus, two exact factorizations can induce different product updates through their factor Gram matrices. A non-orthogonal gauge changes both $R^\top R$ and $LL^\top$, acting as a left–right preconditioner on H even though the forward operator is unchanged at initialization. Balanced SVD controls this geometry by assigning equal square-root singular scales to the two factors.

For plain gradient descent an orthogonal gauge leaves the first-order terms in Equation 4 unchanged. Our downstream optimizer is AdamW, however, whose coordinatewise moment normalization is not equivariant to a general orthogonal mixing of factor coordinates. Corresponding factor updates

can therefore diverge immediately, and nonlinear coupling to the rest of the randomly initialized network can amplify the difference. For non-orthogonal gauges, changed factor norms and decoupled weight decay provide an additional distinction. This motivates the exact-gauge experiment in Section 6: equality of the initial QK/OV maps is a controlled functional constraint, not a claim that every factorization follows the same training trajectory.

3.5 Baselines, metrics, and inference scope

Scratch uses the target's default random initialization. *Full DYCK* transfers the compatible warmed-up transformer backbone. *Raw late-attention-related subset* copies the loader-defined late subset without refactorization; Section 6 lists its extra parameters. The conventional-source boundary uses a supervised ImageNet-1K DeiT-Tiny checkpoint.

For accuracy a_s from scratch, a_f from full-source transfer, and a_c from an intervention, the recovered fraction of the full gain is

$$\mathcal{R}(a_c) = 100\,\frac{a_c - a_s}{a_f - a_s}\,\%. \tag{5}$$

Recovery above 100% means the intervention finishes above full transfer relative to the matched scratch reference. Reported small-dataset summaries are mean ± sample standard deviation over three matched source/downstream seeds (42, 43, 44). The ImageNet study uses a fixed-randomness coverage screen followed by independent downstream-seed confirmation of the selected last-eight condition. The elimination study selects the late-product intervention; the suffix screen reports every completed coverage point, including the non-monotonic ones.

4 What survives procedural warm-up?

4.1 From elimination to an operator intervention

We first use a systematic elimination study to generate the intervention. The study successively retains single-matrix spectra, module families, layer ranges, and finally attention products. Individual-matrix statistics retain little of the original benefit, whereas late attention parameters retain most of it. Refactorizing the late QK and OV products then matches or exceeds raw late-attention copying in the intervention-selection study. Table 3 reports every tested condition.

This progression localizes a useful intervention within the tested DYCK setting. The ImageNet study below further shows that the preferred suffix is target-dependent. We therefore use "late QK+OV" as an empirically selected initializer and evaluate its coverage explicitly at scale.

4.2 Matched-seed replication on two datasets

Table 1 repeats the central comparison with matched DYCK-source and downstream seeds 42, 43, and 44. QK+OV recovers 121.00% and 103.00% of the full-warm-up gains on CIFAR-100 and Tiny-ImageNet.

QK+OV exceeds full warm-up in all six matched pairs. This consistency, together with the mean and standard deviation in Table 1, supports robust retention of the warm-up benefit.

Table 1: Matched three-seed procedural-source results (best Top-1 mean ± sample SD). Recovery uses unrounded means.

Dataset	Scratch	Full DYCK	QK+OV	Recovery
CIFAR-100	72.68 ± 0.03	74.74 ± 0.51	75.17 ± 0.30	121.00%
Tiny-ImageNet	56.75 ± 0.14	60.30 ± 0.70	60.41 ± 0.64	103.00%

4.3 ViT-Base on ImageNet-1K

The last-four-block ViT-Base condition reaches 78.53% Top-1, recovering 45.51% of the full-warm-up gain. We directly test whether broader operator coverage closes this gap. The screening stage holds the DYCK source, factorization, training randomness, and 300-epoch recipe fixed while initializing progressively broader contiguous suffixes.

Table 2: ViT-Base/ImageNet-1K coverage study. All conditions complete 300 epochs. Scratch and full DYCK are matched reference runs; the coverage screen holds all non-intervention settings fixed. Recovery uses unrounded checkpoint values.

Initialization	Best Top-1	Recovery
Scratch	77.72	0.00%
Full DYCK warm-up	79.50	100.00%
QK+OV, last 4 blocks	78.53	45.51%
QK+OV, last 6 blocks	79.86	120.22%
QK+OV, last 8 blocks	80.40	150.45%
QK+OV, last 10 blocks	80.27	143.48%
QK+OV, last 12 blocks	**80.54**	**158.20%**

Table 3: Elimination study on Tiny-ImageNet/ViT-Tiny. Recovery uses the matched scratch (56.45) and full-DYCK (60.65) references; confirmatory results are in Table 1.

Retained source property	Top-1	Recovery
None (scratch)	56.45	0.00%
Full DYCK backbone	60.65	100.00%
Singular values, random directions	57.40	22.62%
Feature directions, random-scale factors	56.30	−3.57%
Matched Frobenius norms	57.10	15.48%
All attention weights	58.10	39.29%
All MLP weights	56.10	−8.33%
All weights, blocks 0–3	55.30	−27.38%
All weights, blocks 4–7	56.80	8.33%
All weights, blocks 8–11	59.40	70.24%
Raw attention weights, blocks 8–11	60.73	101.90%
QK only, blocks 8–11	57.15	16.67%
OV only, blocks 8–11	59.44	71.19%
QK+OV, blocks 8–11	60.87	105.24%
QK+OV, all blocks	58.75	54.76%

Every broader suffix exceeds the 79.50% full-warm-up reference. The complete curve is non-monotonic—last-10 is below last-8 before last-12 rises again—showing that sufficient coverage matters but more transferred blocks do not guarantee a monotonic improvement.

We then confirm last-8 over three downstream seeds while holding the DYCK source and factorization fixed. It reaches $80.32 \pm 0.12\%$, establishing that the ImageNet-scale result is stable to downstream training randomness.

5 Diagnostic ablations

The matched-seed and ImageNet experiments establish the main result, while the following intervention-selection studies explain why QK+OV was selected and test how component choice, retained rank, and warm-up dynamics affect it.

5.1 How the intervention was localized

Table 3 summarizes the elimination path. Preserving the singular values of individual matrices while randomizing their directions recovers 22.62% of the matched full-warm-up gain. Retaining all attention parameters improves on this control, and retaining raw attention parameters only in blocks 8–11 reaches the full-warm-up point estimate. The balanced QK+OV construction reaches a similar point using fresh factors. In the same batch, applying QK+OV to every block is worse than restricting it to the late four blocks.

These rows nominate an intervention rather than prove a unique locus. Each selective copy changes interactions with randomly initialized parameters; a weak MLP-only condition, for example, does not imply that MLPs are irrelevant inside the full warmed-up network. Likewise, the ImageNet sweep

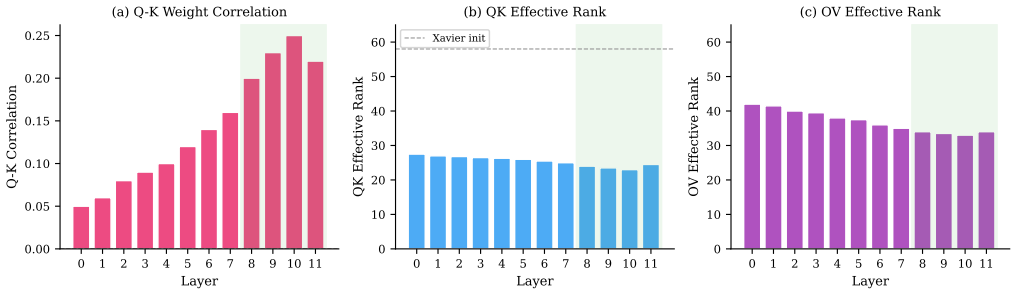


Figure 2: **Per-layer attention diagnostics for the DYCK-warmed ViT-Tiny.** From left to right: Q–K weight correlation, QK effective rank, and OV effective rank. Shading marks blocks 8–11, the suffix used in the Tiny-ImageNet intervention-selection study.

in Table 2 directly shows that four late blocks are not a general optimum. We therefore interpret the elimination as conditional localization within one source, architecture, and target recipe.

Figure 2 characterizes the warmed source before downstream image training. Q–K weight correlation increases in later blocks, QK effective rank remains compressed, and OV effective rank is higher than QK rank. The shaded blocks 8–11 are the suffix selected by the Tiny-ImageNet elimination study. These source-side profiles are consistent with that selection, but the downstream interventions in Table 3 provide the localization evidence.

5.2 Component and rank controls

The intervention-selection study suggested a larger OV than QK contribution (Table 3). We retest that ordering in independent matched three-seed CIFAR-100 batches for both DYCK and supervised-DeiT sources. For DYCK, QK-only and OV-only reach 72.43 ± 0.67 and 73.33 ± 0.14; for supervised DeiT they reach 72.82 ± 0.38 and 74.04 ± 0.34. OV-only exceeds QK-only and scratch in all six paired runs, whereas the QK-only difference from scratch has mixed signs. This supports an OV/QK asymmetry in these two settings, without establishing a fixed ratio or an additive decomposition.

The rank control asks whether only the leading singular directions of the operator are enough. On Tiny-ImageNet, ranks 4, 8, 16, and 32 reach 55.00%, 56.40%, 57.10%, and 57.80%, respectively, well below the 60.87% full-rank condition. This motivates the full head-rank construction used in our intervention.

5.3 The final operators continue to evolve

Figure 3 places downstream and warm-up time courses side by side. The downstream curves show the optimization behavior of raw late-attention and refactorized QK+OV initialization; final comparisons are reported in Tables 1 and 3. During warm-up, early scalar signatures such as effective rank and Q–K correlation can appear stable while the operator itself is still moving. To test product movement directly, we measure normalized Frobenius distance from intermediate late-block products to their 15,000-step values. At step 5,000, QK and OV remain 0.84 ± 0.26 and 0.84 ± 0.33 away from their final products. By step 10,000 the distances fall to 0.21 ± 0.07 and 0.20 ± 0.10.

The continued movement rules out treating an early plateau in a scalar summary as completion of the transferred operator and motivates evaluating the final 15,000-step state. We also measured Q–K correlations and effective ranks in public MAE, DINO, and DINOv2 checkpoints (He et al., 2022; Caron et al., 2021; Oquab et al., 2024). Because we did not run QK/OV transfer interventions from those sources, we use these statistics only to motivate future source comparisons.

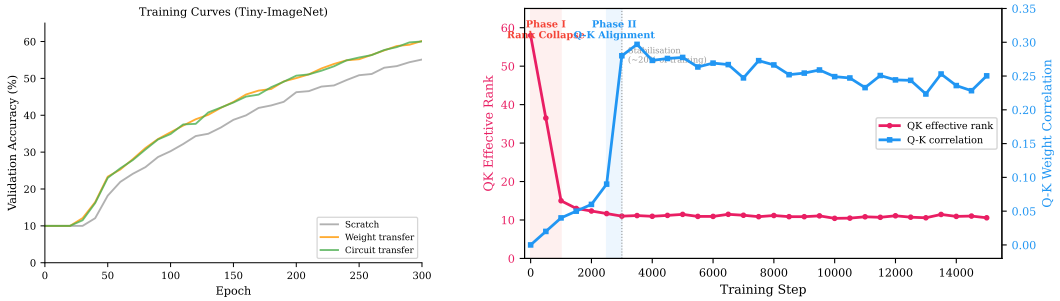


Figure 3: **Training dynamics associated with the intervention.** Left: intervention-selection trajectories on Tiny-ImageNet over 300 epochs for scratch, raw late-attention weight transfer, and refactorized late QK+OV transfer; final aggregate comparisons appear in Tables 1 and 3. Right: QK effective rank and Q–K weight correlation during the 15,000-step DYCK warm-up, averaged across blocks 8–11.

6 Source and parameterization boundaries

6.1 A mature visual source behaves differently

Does high recovery identify a generic transfer mechanism, or a property of short procedural warm-up? We replace the DYCK source with a public supervised ImageNet-1K DeiT-Tiny checkpoint and train CIFAR-100 targets using three downstream seeds.

Table 4 reveals a clear source-dependent contrast. Full-backbone transfer reaches 83.97 ± 0.29%, whereas late QK+OV reaches 74.37 ± 0.33%: above scratch in every seed and 15.02% of the full-source gain. Copying a broader raw late-attention-related subset retains 40.82%, still well below full transfer. For fidelity to the evaluated loader, this raw control also carries blocks 8–11 normalization parameters and shared class/position tokens; it is not a pure attention-only intervention. Increasing QK+OV coverage does not close this conventional-source gap; the best suffix recovers 19.95% of the full-backbone gain. This contrast with the DYCK/ImageNet coverage study makes source type, rather than suffix size alone, central to our conclusion.

Table 4: Supervised ImageNet-1K DeiT-Tiny source to CIFAR-100 (best Top-1 mean ± sample SD over three seeds). Recovery uses unrounded means.

Initialization	Top-1	Recovery
Scratch	72.68 ± 0.03	0.00%
Full DeiT backbone	83.97 ± 0.29	100.00%
Raw late subset	77.29 ± 0.37	40.82%
Late QK+OV	74.37 ± 0.33	15.02%
QK+OV, last 9 blocks	74.93 ± 0.31	19.95%
QK+OV, last 10 blocks	74.41 ± 0.14	15.38%
QK+OV, last 11 blocks	74.57 ± 0.44	16.74%
QK+OV, last 12 blocks	73.96 ± 0.44	11.36%

Thus, the procedural result is not a replacement for feature-rich visual pretraining. A short feature-disjoint warm-up can leave a large fraction of its useful effect in these operators, whereas a mature visual checkpoint stores most of its transferable value elsewhere—potentially in representations, MLPs, normalization, embeddings, biases, or interactions among them. Transfer from MAE, DINO, and DINOv2 is untested; descriptive weight statistics from those models cannot substitute for transfer interventions.

6.2 Equal operators, different coordinates

Balanced SVD is a convenient construction, but Equation 3 does not select a unique factor pair. We therefore use the supervised-DeiT source to compare two exact gauge transformations while holding the QK/OV products fixed. Both preserve those products within 1.34×10^{-6}.

Table 5: Supervised-DeiT QK+OV factorization controls. Accuracy is three-seed mean ± sample SD; reconstruction entries are maximum relative errors over blocks 8–11.

(a) Downstream accuracy

Parameterization	Top-1
Balanced learned QK+OV	74.37 ± 0.33
Same products, orthogonal gauge	74.85 ± 0.53
Same products, condition-4 gauge	74.49 ± 0.39
Matched geometry, random directions	74.42 ± 0.62

(b) Numerical reconstruction audit

Reconstruction check	Max. error
QK products	1.06×10^{-6}
OV products	1.08×10^{-6}
Bias-free logits	1.03×10^{-6}
Bias-free outputs	1.45×10^{-6}
Biased output error (range)	0.71–0.76
Exact-gauge products	$\leq 1.34\times10^{-6}$

Despite this equality, the orthogonal gauge exceeds the balanced construction in all three paired runs, by +0.44, +0.70, +0.29 points. A condition-number-4 gauge produces another intermediate result. Factor coordinates therefore influence downstream optimization even when the initial bias-free products are fixed.

The random-direction control matches the learned-operator mean closely, with mixed paired differences. For the mature visual source, this shows that learned directions are not required to obtain the modest QK+OV benefit. More generally, worthwhile alternatives to balanced SVD are exact gauges with controlled orthogonality or condition number, data-aware factorizations that match activation scale, and constrained factorizations that preserve the product while matching target initialization statistics. The exact-gauge results show why these methods must be compared under product and scale audits rather than treated as algebraically interchangeable. Applying the same matched controls to DYCK is a natural next step for separating operator directions from spectrum and factor geometry.

The numerical audit in Table 5 distinguishes product and bias-free-module equality from reconstruction of the complete biased module. Only the first two are preserved: the 0.71–0.76 biased-output error arises because source QKV/output biases and the rest of the block are not copied.

7 Discussion and conclusion

Summary. A 15,000-step DYCK curriculum leaves a useful initialization effect after the token interface and most warmed-up parameters are discarded. Across three matched seeds, QK+OV-derived initialization attains 75.17 ± 0.30% on CIFAR-100 and 60.41 ± 0.64% on Tiny-ImageNet, retaining 121.00% and 103.00% of the corresponding full-warm-up gains and exceeding full warm-up in every matched pair. At ImageNet-1K scale, every suffix broader than four blocks reaches or exceeds the full-DYCK reference; the confirmed last-eight condition reaches 80.32 ± 0.12%. Together with the elimination and rank studies, these results support late, full-head-rank QK+OV as a reusable structured initializer. Exact-gauge controls qualify this interpretation: both the operators and their factor coordinates affect subsequent optimization.

Potential use. The method applies when a transformer architecture is fixed before final images or labels are available. An abstract curriculum can warm up the transformer first; its attention operators can then initialize a target ViT whose input interface, classifier, and remaining weights are task-specific. Operator extraction also provides a practical test of whether a curriculum has installed reusable initialization structure rather than only source-specific parameters.

Scope and next steps. The supervised-DeiT result establishes a source boundary: late QK+OV recovers 15.02% of mature visual transfer, whose useful information is distributed more broadly through the backbone. The present evidence covers one procedural family, two plain ViT sizes, image classification, and three datasets. Testing other curricula, architectures, and dense-prediction tasks will determine the breadth of the effect. Matched DYCK gauge and random-direction controls, together with activation and gradient measurements, could further separate operator content from factor geometry and explain the effectiveness of late-layer coverage.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript. The authors independently verified the reported values against executed experiments and records, reviewed all AI-assisted text, and take responsibility for the final content.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7096–7116, 2020. doi: 10.18653/v1/2020.emnlp-main.576. URL https://aclanthology.org/2020.emnlp-main.576/.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the chomsky hierarchy. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WbxHAzkeQcn.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 647–655, 2014.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pp. 249–256, 2010.

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2535–2544, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Liangze Jiang, Zachary Shinnick, Anton van den Hengel, Hemanth Saratchandran, and Damien Teney. Procedural pretraining: Warming up language models with abstract data. *arXiv preprint arXiv:2601.21725*, 2026. URL https://arxiv.org/abs/2601.21725.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford University, CS231N, 2015.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128, 2021.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Zachary Shinnick, Liangze Jiang, Hemanth Saratchandran, Damien Teney, and Anton van den Hengel. Can you learn to see without images? procedural warm-up for vision transformers. *arXiv preprint arXiv:2511.13945*, 2025. URL https://arxiv.org/abs/2511.13945.

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. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=4nPswr1KcP.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 Small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.