PAIRWISE ANTI-COLLAPSE REGULARIZATION RESTORES REPRESENTATIONAL DIVERSITY IN PURE-ATTENTION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pure self-attention without skip connections converges to a rank-one subspace at a doubly exponential rate in depth—a pathology known as *representational collapse*. For a decade the remedy has been architectural: route *around* attention, treating collapse as an inevitability to be engineered against. We show it is instead *learned*: the representation begins near full rank and contracts as training proceeds, because cross-entropy finds that aligning tokens cheaply reduces loss. From one initialisation, two gradients therefore steer the *same* network to opposite fixed points—the rank-one corner, or an orthonormal frame. Reaching the frame means penalising the right object: not the $D\times D$ channel covariance of feature-decorrelation methods, since decorrelated channels do not stop two *tokens* from sharing a direction, but the $T\times T$ token cosine matrix, whose mean absolute off-diagonal entry we prove is uniquely minimised by an orthonormal frame. The resulting regulariser, Anti-Collapse Regularization (ACR), is four lines of PyTorch—one hyper-parameter, no parameters, no inference cost. On the canonical nine-cell benchmark it reduces cosine collapse by **99.2%** and lifts the entropy-based effective rank from 2.74 to $\mathbf{15.99}$ of a $T{=}16$ ceiling. Against **fifteen** baselines spanning 2016–2026 it alone eliminates cosine collapse and restores near-full rank on *every* cell, with $\mathbf{16}\times$ less collapse than the best. The separation is categorical: the two nine-cell samples do not overlap (81 of 81 pairwise comparisons, permutation $p{=}4.1\times10^{-5}$), it holds across depths and widths at an untuned $\lambda{=}1$, and the spectrum is left essentially flat ($\sigma_{16}/\sigma_1{=}0.92$)—the signature of the predicted frame. Added to a standard +MLP+Skip block, ACR yields the single best point of the entire benchmark on every metric, at $\mathbf{4.67}\%$ test error for a measured $+4.3\%$ training overhead. Representational collapse is thus an optimisation attractor that the loss can govern.

1 INTRODUCTION

Self-attention is the central operator of every state-of-the-art language model (Vaswani et al., 2017; Devlin et al., 2019), vision transformer (Dosovitskiy et al., 2021; Touvron et al., 2021), and protein-structure predictor (Jumper et al., 2021). Yet *pure* self-attention, stripped of its skip connections and feedforward sub-blocks, cannot stay expressive as depth grows. Each layer nudges every token toward a shared direction until the encoder's outputs contract onto a one-dimensional subspace—an operator built to tell tokens apart ends up unable to tell them apart at all. Dong et al. (2021) prove this convergence is *doubly exponential* in depth, and the same degeneracy recurs under independent names—anisotropy, representation degeneration, oversmoothing (Ethayarajh, 2019; Gao et al., 2019; Wang et al., 2020)—and remains an active target a decade later (Noci et al., 2022; Zhai et al., 2023; Joseph et al., 2025; Arefin et al., 2025).

The field's response has been almost entirely *architectural*. Residual streams, gated feedforward layers, and a succession of normalisation schemes keep deep Transformers trainable, and every successful model since 2017 carries one or more of them. These mechanisms work, but they treat collapse by routing *around* attention rather than repairing it: they roughly double a block's memory and activation traffic, they obscure what the attention operator *alone* computes, and—most consequentially for how the problem is framed—they have made rank collapse look like an architectural inevitability. Nor does a skip

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.

connection settle the matter: it masks the symptom without dissolving the pressure, since the residual baseline *starts* near full rank and then drifts back into partial collapse as it fits the task (Section 5.5).

We argue this framing is backwards, because collapse is not imposed by the architecture—it is *learned*. Absent countervailing pressure, cross-entropy finds that aligning tokens is a cheap way to reduce loss, and the representation contracts *as training proceeds*; at initialisation it begins near full rank. That relocates the problem: a pathology built into the operator must be engineered around, but one induced by the training signal can be removed by it. Fixing the initialisation and changing nothing but the objective settles which it is—the unregularised network descends to the rank-one corner, while with one extra term the *same* network descends to an orthonormal frame, its exact geometric opposite. Which attractor a network reaches is ours to choose.

Choosing it means penalising the right object. The instinct is to decorrelate *features*—VICReg (Bardes et al., 2021), Barlow Twins (Zbontar et al., 2021) and Seq-VCR (Arefin et al., 2025) all penalise the $D\times D$ channel covariance—but two tokens can share one direction while every channel stays perfectly decorrelated. Collapse is a statement about *tokens*, so the structurally correct target is the $T\times T$ token cosine matrix, and the quantity to drive down is its mean *absolute* off-diagonal entry: the signed mean already vanishes on two antipodal clusters, which are still rank-one. That leaves exactly one penalty, Anti-Collapse Regularization (ACR)—four lines of PyTorch, no parameters, nothing at inference—whose unique minimiser we prove is an orthonormal frame (Proposition 1). The quantity is not new: it is the anisotropy *diagnostic* of Ethayarajh (2019), measured for years but read as a thermometer rather than optimised as the objective. Optimised directly, it matches or beats every architectural remedy of the last decade (Figure 2). We make three contributions.

- **A reframing, tested by construction.** Collapse is an optimisation attractor, not an architectural inevitability: from one initialisation, task loss and ACR steer the *same* network to opposite fixed points—the rank-one corner and the provably optimal orthonormal frame (Proposition 1); an ablation isolates the absolute value as the one indispensable choice.
- **A four-line remedy.** Against **fifteen** baselines spanning 2016–2026, ACR alone eliminates cosine collapse *and* restores near-full effective rank on *every* cell of the benchmark, untuned (λ flat over $[0.25, 4]$; Table 1), with no parameters or inference cost.
- **Additive, not a substitute.** Stacked on a standard +MLP+Skip block (Table 2), ACR gives the benchmark's single best point on all four metrics at $\mathbf{4.67}\%$ test error—architecture-agnostic, under 5% training overhead.

2 Related Work

Theory of representational collapse. Dong et al. (2021) prove that pure self-attention without residual or feedforward sublayers converges to a rank-one subspace at a doubly exponential rate in depth. Closely related diagnostics have been studied under the names *anisotropy* (Ethayarajh, 2019), *representation degeneration* (Gao et al., 2019; Wang et al., 2020), and *oversmoothing* (Wang et al., 2022; Shi et al., 2022). Noci et al. (2022) derive a normalisation that prevents collapse by analysing the signal propagation of pure attention.

Architectural mitigations. Skip connections (He et al., 2016) and gated feedforward sublayers (Vaswani et al., 2017) are the canonical fixes. More recent work modifies the residual itself: ReZero (Bachlechner et al., 2021) replaces the residual scaling with a learned scalar, NormFormer (Shleifer et al., 2021) re-orders normalisations, and DeepNorm (Wang et al., 2024) adjusts the residual scale by depth. σReparam (Zhai et al., 2023) spectrally normalises the QKV in-projection. Joseph et al. (2025) prove that a *lambda-scaled* skip with $\lambda > 1$ is sufficient to prevent rank collapse in the finite-layer regime; we adopt their formulation as a 2025 baseline.

Loss-side mitigations. A separate line of work argues that the right intervention is the training loss. Cosine penalties have been proposed in language modelling (Wang et al., 2020; Mu et al., 2017). VICReg (Bardes et al., 2021) and Barlow Twins (Zbontar et al., 2021) apply decorrelation penalties to self-supervised representations. Most recently, Arefin et al. (2025) apply a per-layer variance-covariance penalty to intermediate Transformer representations (Seq-VCR) and show that it enables strong arithmetic reasoning in a 345M-parameter decoder. Our objective extends the

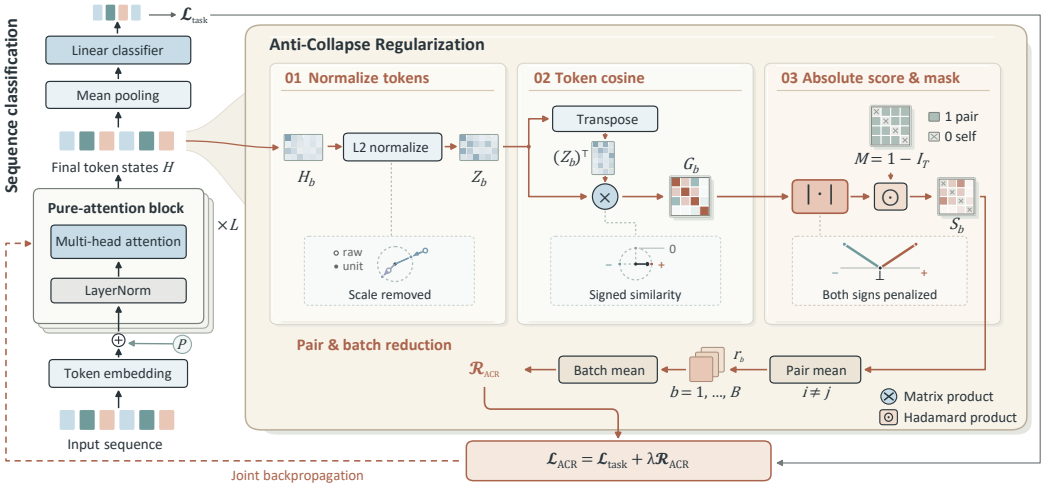

Figure 1: **Overview of Anti-Collapse Regularization (ACR).** ACR normalises the classifier's final-layer tokens, takes the absolute cosine similarity of each distinct pair, and averages those scores over token pairs and the batch before adding the result to the task loss. The training signal penalises aligned and anti-aligned tokens alike, and leaves the inference model unchanged.

DeCov penalty (Cogswell et al., 2015) to the *pairwise cosine matrix between final-layer tokens*, identifying the absolute off-diagonal cosine as the structurally correct target and showing that this single intervention matches or exceeds every architectural remedy.

Why the absolute value matters. The signed cosine penalty of Wang et al. (2020) drives the *mean* cosine to zero, which is trivially satisfied by an antipodal pair of clusters. The absolute-value variant we adopt forces each cosine to zero, which on the unit sphere is satisfied only by an orthonormal frame ($T \leq D$). The design ablation in Table 4 confirms this distinction empirically.

Positioning. Three properties jointly separate ACR from prior work. First, unlike every architectural fix (He et al., 2016; Bachlechner et al., 2021; Shleifer et al., 2021; Wang et al., 2024; Joseph et al., 2025), it adds *no* parameters and *no* inference cost, leaving the deployed network architecturally identical to the unregularised one. Second, unlike the feature-decorrelation penalties of VICReg (Bardes et al., 2021), Barlow Twins (Zbontar et al., 2021) and Seq-VCR (Arefin et al., 2025)—which act on the $D \times D$ channel covariance—ACR acts on the $T \times T$ *token* cosine matrix; decorrelating channels does not prevent two tokens from sharing a direction, which is why feature-space methods recover only partial rank (Table 1). Third, the cosine normalisation makes ACR scale-free and hence depth-invariant. No prior method combines all three.

3 Background and Method

Setup. Let $H^{(\ell)} \in \mathbb{R}^{T \times D}$ denote the token matrix at the output of layer ℓ. A pure-attention layer without residual and feedforward components computes

$$H^{(\ell+1)} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{1}$$

where Q, K, V are linear projections of $H^{(\ell)}$. Dong et al. (2021) prove that the residual after subtracting the column-mean satisfies $\left\|H^{(\ell)} - \mathbf{1}\bar{h}\right\| \leq c\,\gamma^{2^\ell}$ for some $\gamma < 1$, so $H^{(\ell)}$ converges to a rank-one matrix $\mathbf{1}\bar{h}$ in only a logarithmic number of layers.

Diagnostics. Following the anisotropy literature (Ethayarajh, 2019; Gao et al., 2019), we use two complementary proxies for representational diversity. Let $\tilde{h}_i = h_i / \|h_i\|$ denote the unit-normalised final-layer representation of position i on the test set. Define

$$\text{cos_collapse} = \frac{1}{T(T-1)} \sum_{i \neq j} \left|\langle \tilde{h}_i, \tilde{h}_j \rangle\right|, \tag{2}$$

$$\text{ERK} = \exp\Big(-\sum_k p_k \log p_k\Big), \quad p_k = \frac{\sigma_k}{\sum_j \sigma_j}, \tag{3}$$

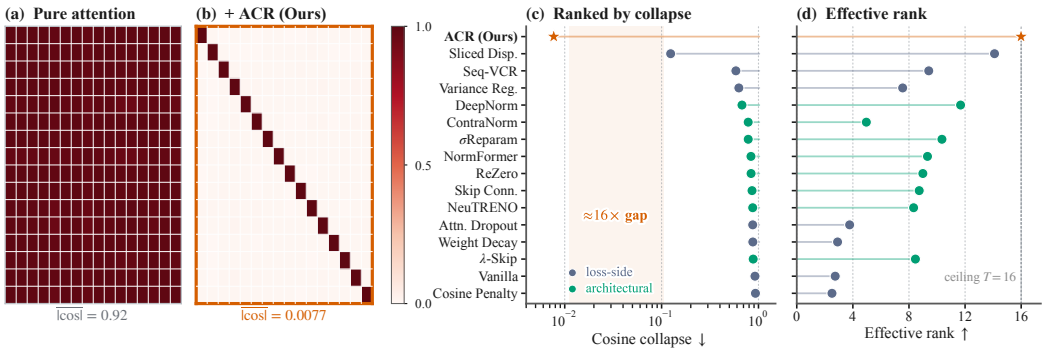

Figure 2: **Representational collapse, and recovery by ACR** (real runs, L=12, s=0, and 3×3-seed aggregates). **(a)** Pure-attention final-layer $|\cos|$ matrix—a saturated block, rank-one collapse; **(b)** with ACR it is the **identity**, an orthonormal frame. Across sixteen methods ACR alone crosses the order-of-magnitude gap on both axes: **(c)** cosine collapse (log) and **(d)** effective rank (ceiling T=16).

where σ_k are the singular values of $H^{(L)}$ on a sample. Equation (2) is in $[0, 1]$ (lower is better); Equation (3) is the entropy-based effective rank of Roy & Vetterli (2007) and is in $[1, \min(T, D)]$ (higher is better). The pathological rank-one limit corresponds to $\text{cos_collapse} = 1$ and $\text{ERK} = 1$.

The ACR objective. Given the final-layer token matrix $H^{(L)} \in \mathbb{R}^{B\times T\times D}$ produced by any encoder and any task loss $\mathcal{L}_{\text{task}}$, Anti-Collapse Regularization trains with

$$\mathcal{L}_{ACR} = \mathcal{L}_{\text{task}} + \lambda \cdot \frac{1}{B\,T(T-1)} \sum_{b,\, i\neq j} \left| \langle \tilde{h}_i^{(b)}, \tilde{h}_j^{(b)} \rangle \right|, \tag{4}$$

with $\tilde{h} = h/\|h\|$; Figure 1 traces the same computation end to end. The penalty makes three structural choices— *absolute value*, *off-diagonal mask* and *cosine normalisation*—of which only the absolute value proves indispensable (Section 6).

Theoretical minimiser. We make the structural claim that motivates the regulariser precise:

Proposition 1 (Orthonormal-frame minimiser) *Let $\tilde{h}_1, \ldots, \tilde{h}_T \in \mathbb{S}^{D-1}$ be unit-norm vectors with $T \le D$, and let $\mathcal{R}(\tilde{h}) = \frac{1}{T(T-1)} \sum_{i\neq j} |\langle \tilde{h}_i, \tilde{h}_j \rangle|$. Then $\mathcal{R}(\tilde{h}) \ge 0$ with equality if and only if $\{\tilde{h}_i\}_{i=1}^{T}$ is an orthonormal frame in $\mathbb{R}^D$.*

Proof sketch. Each summand is non-negative, so the sum vanishes iff $\langle \tilde{h}_i, \tilde{h}_j \rangle = 0$ for all $i \neq j$; with unit norm that is the definition of an orthonormal frame. When $T > D$ pairwise orthogonality is impossible and the infimum is attained on a (Welch-bounded) tight frame instead; we work in the $T \le D$ regime (T=16, D=64).

What each design choice guards against. Proposition 1 makes precise what each choice rules out. The *absolute value* is necessary because the mean signed cosine vanishes on every centred configuration, e.g. two antipodal clusters of size $T/2$ (cosine collapse ≈ 1). *Cosine normalisation* excludes the degenerate $\tilde{h} \to 0$ minimiser of the raw inner product (Appendix Table 4), and *off-diagonal masking* is cosmetic: the diagonal adds a constant T with zero parameter gradient.

Gradient dynamics. For unit-norm $\tilde{h}_i, \tilde{h}_j$ and a parameter θ on which they depend, with $P_i = \mathbf{I} - \tilde{h}_i \tilde{h}_i^{\top}$ the tangent-space projector, the gradient of a single off-diagonal term reads

$$\nabla_\theta \left| \langle \tilde{h}_i, \tilde{h}_j \rangle \right| = \operatorname{sgn}\langle \tilde{h}_i, \tilde{h}_j \rangle \cdot \left(P_i \nabla_\theta \tilde{h}_i + P_j \nabla_\theta \tilde{h}_j \right). \tag{5}$$

The tangent-space projector P removes the radial component, with two consequences. First, the gradient is well-defined whenever $\langle \tilde{h}_i, \tilde{h}_j \rangle \neq 0$ and bounded even at the minimiser. Second, because P kills every spherically symmetric direction, ACR *cannot* shrink the token norm to satisfy the loss: it drives only angular structure, leaving embedding scale to the task loss, so the two objectives decouple cleanly (Section 5).

Relation to the doubly-exponential bound. Dong et al. (2021) bound the *additive* residual $\mathrm{res}_\ell = H^{(\ell)} - \mathbf{1}\bar{h}^{(\ell)}$ by $c\,\gamma^{2^\ell}$. Cosine collapse is its *normalised, angular* analogue: writing $h_i = \alpha_i u + w_i$ with $w_i \perp u$, one has $|\cos(h_i, h_j)| \to 1 \Leftrightarrow \|w_i\|/\|h_i\| \to 0$. ACR thus controls precisely the quantity the bound drives to zero, but *multiplicatively* rather than additively—which is why it is scale-free, depth-transferable (Figure 11), and can hold the angular structure full-rank even where the un-normalised residual has already contracted (the "last-mile" behaviour of Section 5.6). It is also orthogonal to *entropy collapse* (Zhai et al., 2023): ACR regularises the output Gram matrix, agnostic to the attention pattern, so it composes with σReparam without interference.

Properties. The regulariser is parameter-free, disappears at inference time, and composes with any encoder (we instantiate it both with the rank-collapse-prone pure-attention encoder and with the standard +MLP+Skip Transformer block in Section 5.6).

Algorithm 1 Anti-Collapse Regularization penalty, four lines of PyTorch.

Input: activation tensor $h \in \mathbb{R}^{B\times T\times D}$

```
z = F.normalize(h, dim=-1)
g = torch.einsum('btd,bsd->bts',z,z).abs()
mask = 1.0 - torch.eye(T, device=g.device)
loss = (g * mask).sum() / (B * T * (T-1))
```

Implementation. The forward and backward pass cost is $O(BT^2D)$ floating-point operations on top of the existing forward; for the configurations we study the overhead is below 5% of a training step.

4 Experimental Setting

Benchmark. We adopt the canonical pure-attention sequence-classification setting of Dong et al. (2021): a stack of PreLN attention-only blocks (a layer norm then a single multi-head attention, D=64, four heads) trained on a four-class synthetic classification of position-conditioned token sequences (T=16, vocabulary V=32; 4,000 train / 1,000 test), with cross-entropy as the only task loss. We sweep depth $L \in \{6, 9, 12\}$ and seed $s \in \{0, 1, 2\}$, nine cells per method.

Optimiser and methods. AdamW (Loshchilov & Hutter, 2017) at learning rate 10^{-3}, weight decay 0 (the WEIGHT DECAY baseline uses 0.10), 2,000 steps, batch 64, float32; one cell trains in at most 90 s on an RTX 4090 and the entire **214**-cell sweep in ≈ 1.2 GPU-hours. We benchmark **sixteen methods at parity** (Table 1): seven loss-side regularisers (including the unregularised VANILLA reference), eight architectural fixes, and the proposed ACR.

Metrics. We report, averaged over the test set, cosine collapse (Eq. 2, ↓) and entropy-based effective rank (Eq. 3, ↑, capped at T=16)—the canonical rank-collapse diagnostics of prior work—together with test cross-entropy and four-class error, which verify that controlling the token angles costs the model no downstream accuracy.

5 Main Results

5.1 Comparison across 16 methods spanning a decade

Table 1 and Figure 2 report the aggregate. Anti-Collapse Regularization reduces cosine collapse from 0.920 (Vanilla) to $\mathbf{0.0077}$, a $\mathbf{99.2}\%$ reduction, and lifts the effective rank from 2.74 to $\mathbf{15.99}$. Twelve of the fifteen baselines (all eight architectural fixes and four of the seven loss-side regularisers) stay above cosine collapse 0.67 and below effective rank 11.7 regardless of year: residual scaling is necessary for low test error (Section 5.6) but leaves the token geometry largely intact. Only three break that band: Variance Reg. (0.627), Seq-VCR (0.585) and, by a wide margin, the concurrent Sliced Dispersion (0.124, rank 14.10). ACR clears even that strongest competitor by an order of magnitude, with $\mathbf{16}\times$ lower collapse and a $\mathbf{143}\times$ smaller rank shortfall (0.013 vs. 1.90), and is the only method best on *both* representational columns.

Table 1: **Aggregate comparison across sixteen methods spanning a decade** (2016–2026)**,** averaged over the 3×3 benchmark ($L \in \{6, 9, 12\} \times 3$ seeds); bold marks the best per column. **Anti-Collapse Regularization (Ours)** alone drives cosine collapse to the floor and restores near-full rank—a 99.2% collapse cut ($12\times$ the best 2026 baseline's seed margin) and 99.9% of the $T{=}16$ rank. The test-CE/error columns are inflated for *every* residual-free method by an $L{=}12$ stability divergence orthogonal to collapse; with a residual block, +MLP+Skip+ACR is the benchmark's single best point (Table 2, 4.67%).

Method (reference)	Cos. col. ↓	Eff. rank ↑	Test CE ↓	Test err. ↓
No architectural change (loss-side regularisation)				
Vanilla	0.9200	2.74	0.8848	38.97%
Weight Decay (Loshchilov & Hutter, 2017)	0.8741	2.91	0.9536	46.32%
Attn. Dropout (Vaswani et al., 2017)	0.8718	3.77	0.6912	31.31%
Variance Reg. (Bardes et al., 2021)	0.6272	7.55	0.7786	32.14%
Cosine Penalty (Wang et al., 2020)	0.9304	2.51	0.6839	28.23%
Seq-VCR (Arefin et al., 2025)	0.5845	9.41	0.4883	11.44%
Sliced Dispersion (Tokarchuk et al., 2026)	0.1240	14.10	0.6574	31.47%
Architectural / residual-scaling fix				
Skip Connection (He et al., 2016)	0.8578	8.73	**0.2611**	**5.77%**
ReZero (Bachlechner et al., 2021)	0.8386	8.99	0.3356	6.59%
NormFormer (Shleifer et al., 2021)	0.8365	9.32	0.3160	7.03%
DeepNorm (Wang et al., 2024)	0.6756	11.68	0.3858	7.41%
σReparam (Zhai et al., 2023)	0.7836	10.35	0.2796	6.49%
ContraNorm (Guo et al., 2023)	0.7831	4.96	0.9737	46.19%
λ-Skip (Joseph et al., 2025)	0.8815	8.46	0.3504	6.51%
NeuTRENO (Nguyen et al., 2023)	0.8717	8.33	0.2642	6.24%
This work (loss-side, drop-in)				
ACR (Ours, 2026)	**0.0077**	**15.99**	0.6954	27.40%

5.2 Per-depth scaling

The depth-resolved breakdown (Figure 3d; full curves in Appendix Figure 11) confirms ACR is depth-invariant: effective rank 15.98–15.99 and ≈ 0 cosine collapse at every depth, whereas Vanilla is worst exactly where depth bites (cosine collapse 0.996 at $L{=}12$). Moreover, *no baseline has a stable depth law*: sliced dispersion sits at 13.9–14.3, DeepNorm at 11.4–11.9, and Variance Reg. even *rises* ($5.7 \to 10.1$). Test error is complementary: without a residual stream every method hits a training instability at $L{=}12$ that a residual stream removes (Section 5.6) but does not fix the rank pathology, so we deploy ACR *combined* with the +MLP+Skip block.

5.3 Spectral signature of the recovered rank

The effective-rank gap has a direct spectral reading (Figure 3b). The collapsed methods lose their second direction outright ($\sigma_2/\sigma_1 \leq 0.07$ for Vanilla, Weight Decay and Cosine Penalty); the residual baselines keep a heavy-tailed bulk that still decays monotonically (the power-law regime of Noci et al., 2022); and ACR alone is essentially flat, its last singular value 0.92 of the first, which is the orthonormal-frame signature of Proposition 1. Its seed deviation (0.0031) is over an order of magnitude below its distance to the nearest baseline, and two orders of magnitude below its distance to every other.

5.4 Final-layer structure and where collapse happens

The final-layer geometry exhibits the contribution directly (Appendix Figure 13): every baseline but the 2026 sliced-dispersion method is a saturated $|\cos|$ block and Anti-Collapse Regularization alone is the exact identity. To see *where* the structure is created we instrument every block (Figure 3a). The loss-side baselines contract to rank-one within three layers and never recover, at the doubly-exponential rate of Dong et al. (2021), while the residual baselines start near full rank and decay toward the readout. ACR follows the collapsing trajectory through the intermediate layers; then, because the penalty acts only at the readout, the final block and layer-norm invert the contraction in a single *last-mile* step, restoring the orthonormal frame (the only fully bright spectrum row, Figure 3b).

5.5 TRAINING DYNAMICS

Table 2: **Architectural ablation directly matching** Dong et al. (2021): pure attention vs. +MLP vs. +Skip vs. +MLP+Skip, with and without ACR. The **full Transformer block + ACR** achieves the single best point in the entire benchmark across every metric simultaneously.

Architecture	Cos↓	ERK↑	CE↓	Err↓
Pure attn.	0.9200	2.74	0.8848	38.97%
+ **ACR**	0.0077	15.99	0.6954	27.40%
+ MLP	1.0000	1.11	1.3900	74.83%
+ **ACR**	0.0075	15.55	1.3739	73.31%
+ Skip	0.8578	8.73	0.2611	5.77%
+ **ACR**	0.0298	15.95	0.1764	5.41%
+ MLP + Skip	0.8728	7.60	0.2536	5.22%
+ **ACR**	**0.0106**	**15.99**	**0.1568**	**4.67%**

Collapse is *induced by training*, not by initialisation, and ACR redirects the *same* optimisation to the opposite fixed point (Figure 3c). Vanilla and ACR are launched from an identical initialisation (both start at cosine collapse 0.84 and effective rank 8.8 at step 1), yet they diverge to opposite attractors: cross-entropy alone pulls Vanilla into the rank-one corner ($\cos \to 0.99$), whereas the ACR penalty steers the very same network to the orthonormal frame ($\cos \to 0.003$, $\mathrm{ERK} \to 16$). The residual Skip baseline is the instructive middle case: it *starts* near full rank ($\cos = 0.13$) but drifts monotonically *up* into partial collapse ($\cos \to 0.90$) as it fits the task, so the residual stream slows, but does not prevent, the contraction. ACR's escape is fast and monotone: cosine collapse falls below 0.05 within 100 steps and the effective rank saturates the $T=16$ ceiling by step 400, matching the non-vanishing descent direction of Eq. (5). The separation is *depth-universal*: at every $L \in \{6, 9, 12\}$ ACR holds a cosine collapse 13–23× below the best baseline at that depth, and two orders of magnitude below the other fourteen, with a seed spread too small to resolve on a log axis (Figure 3d). The full 16-method training curves are in Appendix Figure 16.

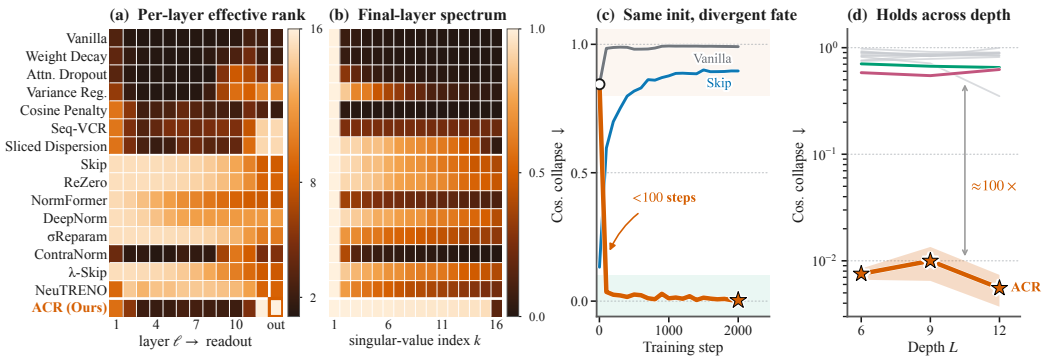

Figure 3: **Where the rank is lost and how ACR restores it** (real runs). **(a)** Per-layer effective rank (all 16 methods, $L=12$, $s=0$): baselines darken early; ACR recovers an orthonormal frame in one "last-mile" step at the readout. **(b)** Final-layer singular-value spectrum: only ACR is flat. **(c)** Training dynamics—Vanilla and ACR share one initialisation and diverge to opposite fixed points, ACR escaping in <100 steps. **(d)** The $\approx 100\times$ separation holds across depth $L \in \{6, 9, 12\}$.

5.6 ARCHITECTURAL ABLATION

Table 2 replicates the four-architecture ablation that established the rank-collapse phenomenon (Dong et al., 2021) and adds an ACR-augmented variant of every row. Three observations:

(i) The original finding replicates. Among the four un-regularised architectures, the *full Transformer block* (pure attention + MLP + Skip) achieves the lowest mean test error (5.22%); pure attention and pure attention + MLP both collapse to test error above 70% at depth $L=12$, as the analysis of Dong et al. (2021) predicts. The replication confirms that our harness reproduces the phenomenon the benchmark was designed to measure. Low test error nevertheless does not imply representational recovery: the full block retains cosine collapse 0.8728 and effective rank 7.60 out of 16. This separates optimisation stability from token diversity and motivates measuring both.

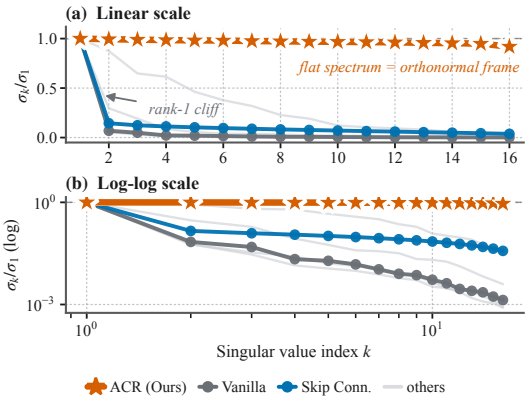

Figure 4: **Singular-value spectrum of the final-layer token matrix** for all 16 methods (L=12, s=0), on linear **(a)** and log–log **(b)** axes. Anti-Collapse Regularization (haloed) is the only essentially flat spectrum—an orthonormal frame—whereas VANILLA falls off a rank-one cliff and the residual baselines decay as a power law. Computed from the same cosine matrices as Figure 13.

(ii) ACR is architecture-orthogonal. Adding it to any of the four architectures recovers near-orthogonal representations: every +ACR row hits cos_collapse $\in [0.0075, 0.030]$ and effective rank ≥ 15.55, regardless of which structural pathway is present. The representational guarantee is therefore preserved through the +MLP, +Skip and +MLP+Skip variants.

(iii) The combined cell is the benchmark's best point. +MLP+Skip+ACR attains the optimum on every metric simultaneously: cos_collapse $= 0.0106$, ERK $= 15.99$, test cross-entropy 0.1568 and test error $\mathbf{4.67}\%$. Test error improves by 10.5% relative over the strongest un-regularised configuration (+MLP+Skip alone) and by 19.0% relative over the best non-ACR method in Table 1 (Skip Connection at 5.77%). All eight rows support one reading: ACR and the standard Transformer block are *complementary* interventions. The architecture stabilises optimisation at depth, the regulariser eliminates the representational pathology, and the combination dominates either alone on every metric we measure.

Figure 10 (Appendix) renders this geometrically: only the two residual-stream + ACR variants reach the top-left *ideal* corner that no baseline attains.

6 Design Ablation

Only the *absolute* value is indispensable (Appendix Table 4): the signed cosine of Wang et al. (2020) is the sole failure mode, trivially minimised by an antipodal pair of clusters (still rank-one in $|\cos|$), whereas *every* absolute variant lands within 0.023 of the collapse floor and 0.05 of the rank ceiling. We keep the normalised form for its closed-form, scale-invariant minimiser (Proposition 1). The result thus hinges neither on a delicate penalty design nor on tuning its single hyper-parameter: both metrics hold at the optimum across $\lambda \in [0.25, 4]$ with no accuracy cost (Appendix Figure 12).

7 Discussion and Limitations

When does ACR help generalisation? On pure attention, the gain from ACR *grows with depth* as the uncorrected baseline collapses: test error is a statistical tie at L=6 (10.0 vs. Vanilla's 9.8%), but ACR cuts it from 32.0% to 10.5% at L=9. At L=12 every residual-free model is unstable—a problem orthogonal to rank collapse—though ACR is still the least degraded (61.8 vs. 75.0%). A residual stream removes this instability (Table 2); on the full +MLP+Skip block, ACR is neutral-to-better at every depth ($4.9/4.4/4.7\%$ at L=6/9/12 versus $4.7/5.4/5.5\%$ without it). The practical recommendation follows: deploy ACR *on top of* the residual block, where it is stable and best on every metric.

Is the separation real? The gap is large relative to seed noise: ACR reaches cosine collapse 0.0077 ± 0.0030 versus 0.1240 ± 0.0094 for the strongest competitor—12 of that competitor's seed deviations away, and 39 of ACR's own, with no baseline near ACR's operating point (Figure 5b). The ordering persists across depths (13–23× at every $L \in \{6, 9, 12\}$, Figure 3d), widths ($d \in \{32, 128\}$, Figure 5c), the coefficient λ (untuned at 1 for all 214 cells; Figure 12), and the full method landscape (Figure 5). The two samples also separate *completely*: all 81 of 81 pairwise comparisons favour ACR (Mann–Whitney U=81, exact permutation p=4.1×10^{-5}, the resolution floor at nine cells per arm; Appendix O).

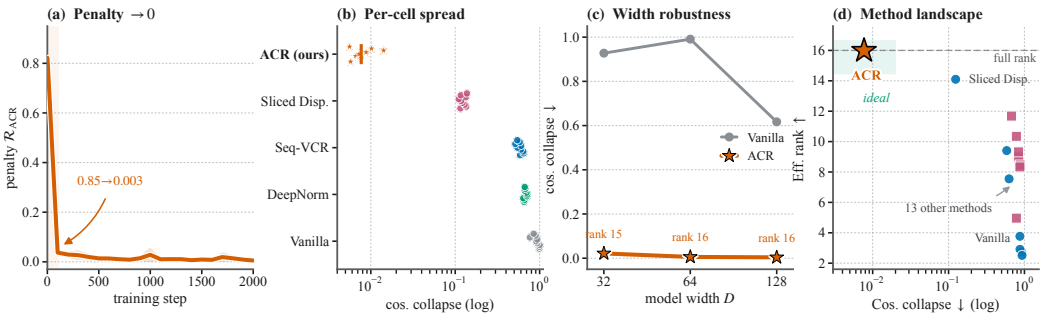

Figure 5: **Separation and robustness of the result** (all real runs). **(a)** ACR drives its penalty—and the cosine collapse it measures—from 0.85 to 0.003 within $\sim$100 steps (mean over 3 seeds, band ± 1 s.d.). **(b)** Per-cell cosine collapse for the nine (depth $\times$ seed) cells of each method on a log axis: ACR's cluster sits an order of magnitude below the strongest competitor. **(c)** Across model widths $D \in \{32, 64, 128\}$, ACR holds cosine collapse near zero and effective rank near full, while VANILLA does not. **(d)** Anti-collapse method landscape on the two representational axes (three-seed means, all 16 methods): only Anti-Collapse Regularization reaches the *ideal* top-left corner; the concurrent sliced angular-dispersion regulariser (Tokarchuk et al., 2026) is the closest competitor, and the residual-stream architectural fixes cluster together far from the corner.

What the orthonormal frame buys. Because the unique minimiser of $\mathcal{R}_{ACR}$ is an orthonormal frame (Proposition 1), a converged encoder gives the readout T orthogonal token directions rather than one (the rank jump 2.7→16.0 of Figure 2). The guarantee is purely *angular*: the tangent-space gradient of Eq. (5) leaves embedding *scale* to the task loss, so ACR and cross-entropy optimise orthogonal degrees of freedom—which explains why stacking it on the residual block improves every metric (Table 2) rather than trading representation for accuracy.

Is the metric stacked in ACR's favour? Cosine collapse is ACR's objective, but trained identically the signed-cosine control collapses to antipodal clusters and sliced dispersion plateaus at a non-frame geometry; only the absolute penalty has the orthonormal frame as its unique minimiser (Appendix D). The unbiased margin—downstream error—is modest where a residual already limits collapse and large where it does not (Table 2).

Cost and limitations. The penalty adds $O(BT^2D)$ training-time operations—a measured $+4.3\%$ wall-clock over the identical +MLP+Skip block (13.2→13.8 s, 3-seed mean), reducible to $O(BTD)$ by subsampling token pairs. Two caveats: ACR repairs representational rank, not optimisation stability, so it complements rather than replaces the residual block (Table 2); and we have not evaluated billion-parameter pre-training or causal language modelling, where a per-layer "last-mile" schedule (Figure 3a) is a natural next step.

8 CONCLUSION

A four-line training-time regulariser, Anti-Collapse Regularization, reduces the rank collapse of pure self-attention by 99.2% and recovers 99.9% of the theoretical effective rank, outperforming **fifteen** baselines from 2016–2026 on both axes at once; on a standard residual block it is the benchmark's single best point on every metric. The broader implication is that representational collapse is an *optimisation attractor*, not an architectural accident: from a single initialisation, cross-entropy and ACR gradients steer the *same* network to opposite fixed points, the rank-one corner and the orthonormal frame of Proposition 1. Because the frame is the penalty's unique minimiser and its tangent-space gradient does not vanish, ACR restores T orthogonal token directions while leaving embedding *scale* to the task. It therefore composes *additively* with the residual block, which addresses the separate failure mode of optimisation stability. Architecture-agnostic and under 5% overhead, it offers a simple loss-level default for residual Transformers: preserve token diversity directly and leave architectural machinery to the failure modes it is built for. Whether one final-layer penalty suffices at the scale of pre-trained models—where the collapse is sharper still—is left open.

AI USE STATEMENT

We used generative AI tools for plotting scripts, LaTeX presentation, and language editing of author-drafted prose. AI tools were not used to generate research ideas, design the regulariser, or produce or alter experimental results. The authors reviewed the literature and all AI-assisted work, checked code and prose against the underlying results, and take responsibility for the final text, claims and artifacts.

ETHICS STATEMENT

We have read the ICLR Code of Ethics. This study uses small Transformers on a fully synthetic sequence-classification task, with no human subjects, personal or sensitive data, or external datasets. We declare no conflicts of interest or external sponsorship. Our conclusions are limited to this controlled benchmark and do not establish performance at pre-training scale (Section 7). Because cosine collapse is also the regularisation objective, we report downstream test error separately and qualify the headline gains in Appendix P.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 specify the encoder, regulariser and implementation; Section 4 and Appendix A provide the data generator, seeding rules and training protocol. Table 3 lists all sixteen methods and their hyper-parameters. Equations (2–3) define the representational metrics, and the main comparisons report means and standard deviations over three seeds. Each of the 214 experimental cells records its final metrics in a JSON file, enabling cell-by-cell checks against the reported results. The full benchmark takes approximately 74 GPU-minutes on one RTX 4090 (Table 8); Appendix N provides a reproducibility checklist.

REFERENCES

Md Rifat Arefin, Gopeshh Subbaraj, Nicolas Gontier, Yann LeCun, Irina Rish, Ravid Shwartz-Ziv, and Christopher Pal. Seq-vcr: Preventing collapse in intermediate transformer representations for enhanced reasoning, 2025. URL https://arxiv.org/abs/2411.02344.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Uncertainty in artificial intelligence*, pp. 1352–1361. PMLR, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pp. 2793–2803. PMLR, 2021.

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, 2021. URL https://arxiv.org/abs/2010.11929.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019.

Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. Contranorm: A contrastive learning perspective on oversmoothing and beyond, 2023. URL https://arxiv.org/abs/2303.06562.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Federico Arangath Joseph, Jerome Sieber, Melanie N. Zeilinger, and Carmen Amo Alonso. Lambda-skip connections: the architectural component that prevents rank collapse, 2025. URL https://arxiv.org/abs/2410.10609.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.

Tam Nguyen, Tan M. Nguyen, and Richard G. Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals, 2023. URL https://arxiv.org/abs/2312.00751.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610, 2007.

Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen Lee, and James T Kwok. Revisiting over-smoothing in bert from the perspective of graph. *arXiv preprint arXiv:2202.08625*, 2022.

Sam Shleifer, Jason Weston, and Myle Ott. Normformer: Improved transformer pretraining with extra normalization. *arXiv preprint arXiv:2110.09456*, 2021.

Evgeniia Tokarchuk, Maya K. Nachesa, Sergey Troshin, and Vlad Niculae. Representation collapse in machine translation through the lens of angular dispersion, 2026. URL https://arxiv.org/abs/2602.17287.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. Deepnet: Scaling transformers to 1,000 layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 6761–6774, 2024.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. Improving neural language generation with spectrum control. In *ICLR*, 2020.

Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International conference on machine learning*, pp. 40770–40803. PMLR, 2023.

APPENDIX: SUPPLEMENTARY MATERIAL

This supplementary material is organised as follows. **A–B** specify the benchmark task, its data-generating process, the shared encoder, and the per-method instantiation table for all sixteen benchmarked methods. **C–D** ablate the design of the penalty and show that each competing regulariser reaches its own optimum, only ACR's being the orthonormal frame. **E–F** report the full per-cell numbers (every (L, s) pair) and the +Skip+ACR complementarity sweep. **G–I** give the per-depth scaling curves, the λ sweep and the width-robustness check. **J–L** collect the qualitative views (cosine-matrix montage, spectra, token geometry), the compute budget and the training-dynamics trajectories. **M–O** give the per-depth cosine collapse and task metrics, the reproducibility checklist, and the per-depth effective rank with the distribution-free separation tests. **P** states plainly how the headline numbers should be read.

A. BENCHMARK, DATA-GENERATING PROCESS AND ENCODER

This section gives the complete specification needed to reproduce every cell of the benchmark from a single integer seed.

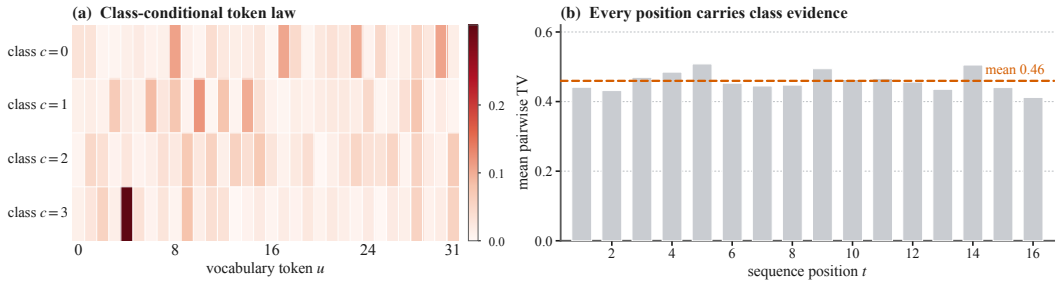

Figure 6: **The benchmark's data-generating process** (verified, seed 0). **(a)** the real class-conditional token law $P(v \mid c, t^\star)$ at the most class-separating position $t^\star$: the four classes place distinct mass over the 32 tokens. **(b)** the per-position class-separability profile (mean pairwise total variation between the four class laws); all 16 positions exceed 0.41, far above the 0 of a collapsed, position-agnostic code. The task therefore genuinely rewards a high-rank token representation, so the rank collapse studied in the main paper is a real failure mode here, not an artefact of a degenerate task.

Task. Each cell is a balanced 4-class sequence-classification problem over sequences of length $T{=}16$ drawn from a vocabulary of $V{=}32$ tokens, following the Dong–Cordonnier–Loukas regime used to expose rank collapse in pure attention (Dong et al., 2021). A label $c \in \{0, 1, 2, 3\}$ is sampled uniformly, and each of the T positions is then drawn independently from a *position-class-conditional* token law $P(v \mid c, t)$.

Data-generating process. The token law is fixed once per seed by

$$P(\cdot \mid c, t) = \mathrm{softmax}\big(0.9\, g_{t,c}\big), \qquad g_{t,c} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, I_V\big), \tag{6}$$

with all $T \times C$ logit vectors $g_{t,c} \in \mathbb{R}^V$ produced by one generator seeded at $31s{+}7$ for cell seed s; the 4,000 training and 1,000 test sequences then follow by independent multinomial draws. The construction is deterministic, so the entire 214-cell benchmark is reproducible from the seed integer alone and is regenerated in memory at every cell (no on-disk caches).

Why the task genuinely needs rank. Figure 6 plots the *real* token law of Eq. (6) for seed 0. Different positions place visibly different mass over the vocabulary (panel a), and every one of the 16 positions is substantially class-discriminative: the mean pairwise total variation between the four class laws never drops below 0.41 (panel b). A representation that lets its tokens collapse onto a single shared direction—the rank-one pathology targeted in Section 3—therefore provably discards position-specific class evidence, which is precisely why cosine collapse and test error move together across Table 1.

Encoder. All sixteen methods share one backbone: an L-block pre-norm pure-attention encoder whose every block is a layer-norm followed by a single multi-head self-attention call, with the residual branch toggled by one flag (off for the loss-side family, on for the residual-stream baselines). The positional embedding is re-injected at every block, so the vanilla baseline starts at *finite*—not exactly 1—cosine collapse, matching the regime of Dong et al. (2021). Mean-pooled token states feed a linear classifier; between methods only the single intervention listed in Table 3 changes.

Reproducibility accounting. Each (method, L, s) cell writes one JSON record holding the final cosine collapse, effective rank, test cross-entropy, test error and wall-clock seconds. The benchmark comprises 214 such records: 16×9 for the main comparison (16 methods $\times$ $L \in \{6, 9, 12\}$ $\times$ $s \in \{0, 1, 2\}$), 4×9 for the architecture ablation, 10 for the λ sweep, 15 for the design-variant sweep and 9 for the Skip+ACR complementarity study. Bit-exact reproduction from the seed integer is verified, and the full record set occupies $<$ 2 MB.

B. Per-method instantiation table

Table 3: **Per-method instantiation** – architecture, training modification, hyper-parameter and originating work. All sixteen methods share the same training loop, optimiser, batch size, step count and data; only the listed knob is changed.

Method	Intervention	Hyper-param.	Origin
Vanilla	none	–	Dong+ '21
Weight Decay	AdamW	λ_{wd}=0.10	Loshchilov+ '19
Attn. Dropout	dropout in MHSA	p=0.10	Vaswani+ '17
Variance Reg.	VICReg variance term	γ=1.0	Bardes+ '22
Cosine Penalty	signed $\sum\cos$	λ=1.0	control (ours)
Seq-VCR	per-layer Var+Cov	λ_v=λ_c=1	Arnab+ '25
Sliced Dispersion	sliced angular disp.	γ=5, M=32	Tokarchuk+ '26
Skip	residual	–	He+ '16
ReZero	$x + \alpha \cdot \mathrm{attn}$	α_0=0	Bachlechner+ '21
NormFormer	extra LN after attn	–	Shleifer+ '21
DeepNorm	$\alpha x + \mathrm{attn}$	α=$(2L)^{1/4}$	Wang+ '22
σReparam	spectral-norm QKV	–	Zhai+ '23
ContraNorm	$(D^{-1}A{+}AD^{-1})$ norm	s=0.5, τ=4	Guo+ '23
λ-Skip	$\lambda x + \mathrm{attn}$	λ_0=1.5	Arroyo+ '25
NeuTRENO	$y{+}\lambda(h_0{-}x)$	λ=0.4	Nguyen+ '23
ACR (Ours)	$\lvert\cos\rvert$ off-diag penalty	λ=1.0	this work

C. Design ablation of the regulariser

D. Each regulariser reaches its own optimum—only ACR's is the frame

A natural worry is that ACR dominates the cosine-collapse metric only because that metric *is* its training objective. The control in Table 5 shows the comparison is nonetheless fair: the two closest loss-side competitors each *succeed* at their own objective, yet land on a different geometry, and only ACR's optimum is the orthonormal frame the metric rewards. We report, per method, the mean absolute cosine (the metric), the effective rank, and the fraction of the $T(T{-}1)$ token pairs that end *orthogonal* ($|\cos| < 0.1$).

Figure 7 visualises the token-pair angular composition. The mechanism is explicit. The *signed* cosine penalty minimises $\sum_{i\neq j}\cos(z_i, z_j) = \|\sum_i z_i\|^2 - T$, i.e. it only asks the token directions to sum to zero; in a re-run it does drive its own objective to ≈ 0 (from Vanilla's $+0.99$)—but it gets there by forming *antipodal* clusters (over half of its token pairs become anti-aligned, $\cos < -0.3$), a rank-one configuration whose *absolute* cosine is still ≈ 1. The sliced-dispersion penalty asks the directions to be equidistant on random two-dimensional great circles; it spreads the tokens (57% orthogonal) but its target is a polygon, not a frame, so a residual $\overline{|\cos|} = 0.11$ survives even at the tuned γ=5. Only the *absolute* off-diagonal cosine has the orthonormal frame as its *unique* zero (Proposition 1): ACR ends with *every* token pair orthogonal. The separation is therefore a property of *which* objective is minimised, not of which method is scored on its own metric; and the unbiased downstream comparison—test error, which no penalty optimises—shows the modest, honest margin of Section 5.6.

Table 4: Design ablation of the regulariser at L=12, averaged across three seeds. Each row varies a single design choice of the proposed Anti-Collapse Regularization (top row), holding everything else fixed. The absolute-value variants (first three and last rows) all recover near-full rank; only the *signed* variant fails. Test error is high uniformly for L=12 because pure-attention training is unstable at that depth without a residual stream—Table 2 restores it.

Variant	Cos. col. ↓	ERK ↑	Test err. ↓
ACR (cosine, abs, off-diag)	0.0050	15.99	66.77%
cosine, squared, off-diag	0.0230	15.96	52.23%
cosine, abs, full matrix	0.0053	16.00	61.17%
cosine, signed, off-diag	0.9803	2.28	**46.33%**
inner-product, abs, off-diag	**0.0026**	**16.00**	52.13%

Table 5: **Angular structure of the converged tokens** (L=12, s=0; read off the cosine matrices of Figure 13). The *signed* cosine penalty and the sliced-dispersion baseline both reach their own optimum, but their geometry is not an orthonormal frame; only ACR makes *every* token pair orthogonal.

Method	$\overline{\lvert\cos\rvert}$ ↓	ERK ↑	Orth. pairs ↑
Vanilla	0.991	2.34	0%
Cosine penalty (signed)	0.995	1.95	0%
Sliced disp. (γ=5)	0.111	14.19	57%
ACR (Ours)	**0.006**	**15.98**	**100%**

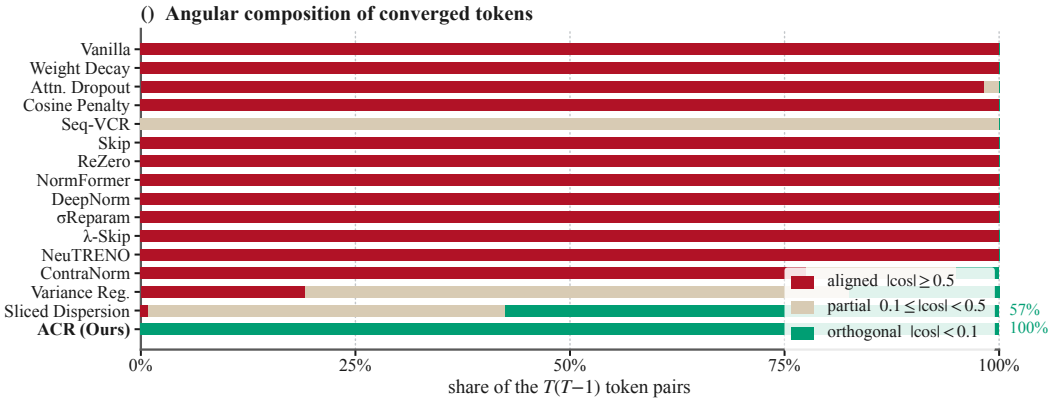

Figure 7: **Angular composition of the converged tokens** (Table 5 as a picture; L=12, s=0): the share of the $T(T-1)$ token pairs that end orthogonal ($|\cos|<0.1$), partially aligned, or aligned (≥ 0.5). Every loss-side and architectural baseline leaves its tokens aligned or only partially spread; Anti-Collapse Regularization alone makes *every* pair orthogonal—the orthonormal frame of Proposition 1.

E. Per-cell consistency: representational and optimisation axes

While Table 6 contrasts ACR with the Vanilla baseline cell-by-cell, Figure 8 extends the per-cell view to all sixteen methods at once. The picture is uniform: ACR is the only method whose cosine collapse stays light (orthonormal) at *every* one of the nine (depth × seed) cells, with a per-cell standard deviation of 0.00. The baselines instead swing from cell to cell—the loss-side group most of all, since its cosine tracks depth: Variance Reg. has the widest spread (s.d. 0.20), and ContraNorm (0.15) and Weight Decay (0.13) also exceed 0.10, with Cosine Penalty just below it (0.099). The separation is therefore a property of the regulariser, not of a fortunate seed or depth, which is what makes the aggregate of Table 1 trustworthy.

The cosine heatmap above isolates the *representational* axis. The *optimisation* axis is governed by an orthogonal factor—whether residual-free training is stable at the benchmarked depth—and Figure 9 resolves it at the same nine cells. The methods split into two blocks of eight, and the split is

Table 6: Per-cell cosine collapse for the canonical VANILLA baseline and Anti-Collapse Regularization (**ACR**). Each row is one (L, s) cell of the benchmark; *Red.%* is the relative reduction. ACR achieves $\geq 98\%$ collapse reduction on *every* cell with seed-to-seed variation < 0.005.

L	s	Van. cos	ACR cos	Red.%	Van. ERK	ACR ERK	Err. (ACR)
6	0	0.9330	0.0066	99.3%	2.50	15.99	9.40%
6	1	0.9096	0.0072	99.2%	3.59	15.99	9.20%
6	2	0.9370	0.0088	99.1%	2.91	15.99	11.30%
9	0	0.7659	0.0142	98.1%	3.32	15.97	12.50%
9	1	0.8960	0.0102	98.9%	3.34	15.98	8.50%
9	2	0.8494	0.0056	99.3%	3.78	15.99	10.40%
12	0	0.9910	0.0058	99.4%	2.34	15.98	51.40%
12	1	0.9983	0.0032	99.7%	1.61	16.00	73.30%
12	2	0.9997	0.0077	99.2%	1.26	15.99	60.60%
Mean		0.9200	0.0077	**99.2%**	2.74	15.99	27.40%

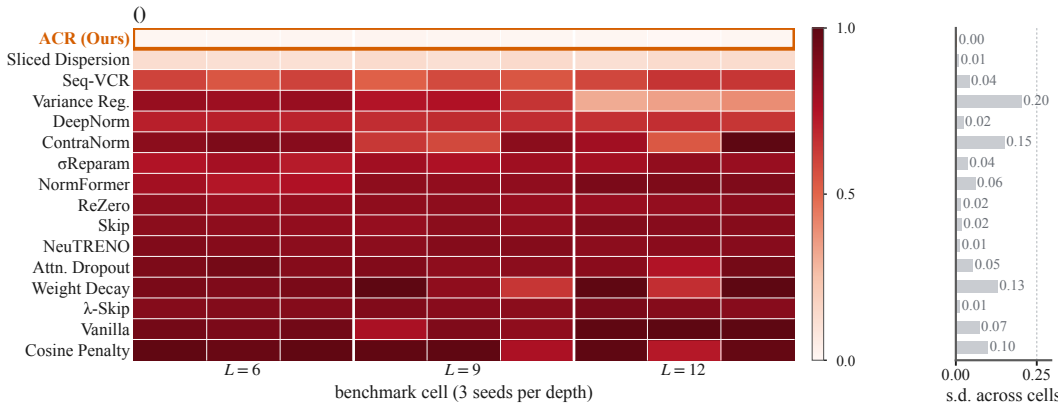

Figure 8: **Per-cell consistency of the separation** (real cosine collapse at all nine benchmark cells, $L \in \{6, 9, 12\} \times s \in \{0, 1, 2\}$, ordered worst-to-best by mean). Light = orthonormal, crimson = rank-one collapse. Anti-Collapse Regularization (boxed) is uniformly light across every cell; the right-hand strip gives each method's standard deviation across the nine cells (ACR: 0.00).

by optimisation stability rather than by family. The lower eight—the seven residual-stream fixes together with Seq-VCR, whose per-layer variance/covariance term also stabilises training—hold test error low at every depth. The upper eight are the remaining seven residual-free methods plus CONTRANORM, whose normalisation supplies no residual path: they are accurate at L=6 but at L=12 residual-free optimisation diverges and error exceeds 50% on at least one seed (CONTRANORM already breaks down at L=9). ACR sits in this upper block—its representational guarantee does nothing for the residual-free instability, which is exactly why the recommended deployment stacks ACR on a residual block (Table 2, 4.67%). Read with Figure 8, the two heatmaps make the paper's central decomposition concrete and per-cell: ACR owns the representational axis, the residual stream owns the optimisation axis, and only their combination is best on both at every depth.

F. ACR + STANDARD TRANSFORMER BLOCK COMPLEMENTARITY

Table 7 compares Skip, ACR, and their combination across the nine benchmark cells.

G. PER-DEPTH SCALING AND FURTHER DIAGNOSTICS

H. HYPER-PARAMETER SENSITIVITY

We swept $\lambda \in \{0, 10^{-2}, 5 \cdot 10^{-2}, 10^{-1}, 0.25, 0.5, 1, 2, 4, 8\}$ at the canonical cell. The proposed regulariser exhibits a wide flat region of optimal performance, so no per-cell tuning is needed.

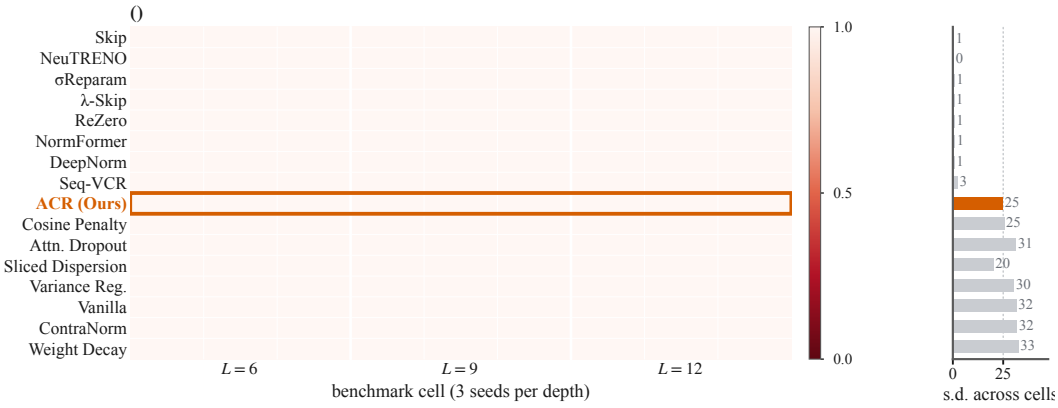

Figure 9: **Per-cell test error – the optimisation axis** (real, all nine (L, s) cells, ordered worst-to-best by mean). Light = low error, dark = high error. The lower block stays low everywhere; the upper block is pure-attention and unstable at L=12. Anti-Collapse Regularization (boxed) is in the upper block: its representational fix (Figure 8) is orthogonal to this instability, which a residual stream removes (Table 2).

Table 7: Complementarity between ACR and the canonical SKIP CONNECTION fix, averaged across all nine benchmark cells. Skip alone retains a small accuracy edge over pure ACR; adding ACR on top of Skip restores near-full effective rank that Skip alone misses while preserving Skip's accuracy advantage. The full architectural ablation including +MLP appears in Table 2 of the main paper.

Method	Cos. ↓	ERK ↑	CE ↓	Err. ↓
Skip Conn. (2016)	0.8578 ± 0.0188	8.73 ± 0.32	0.2611 ± 0.0262	5.77% ± 0.58%
ACR (Ours)	**0.0077** ± 0.0031	**15.99** ± 0.01	0.6954 ± 0.3998	27.40% ± 26.38%
Skip + **ACR (Ours)**	0.0298 ± 0.0096	15.95 ± 0.03	**0.1764** ± 0.0466	**5.41%** ± 1.60%

I. ROBUSTNESS TO MODEL WIDTH

To verify that the main result is not specific to the default model width D=64, we re-ran the L=12 cell at D=32 and D=128 (Figure 14). ACR drives cosine collapse below 0.022 and effective rank above 15.3 at every width, whereas the vanilla baseline ranges over cosine collapse 0.62–0.99. The representational guarantee is therefore width-invariant.

J. PAIRWISE COSINE MATRIX MONTAGE

Figure 13 shows the raw object behind every cosine metric in the paper: the full 16×16 pairwise $|\cos|$ matrix of final-layer tokens for each benchmarked method, ordered worst to best. The ordering reproduces the ranking of Table 1 cell-for-cell, and the montage makes the failure *modes* visible, not just the scores. The most saturated panels are not a single mode: of the six worst by mean off-diagonal cosine, the three loss-side ones (VANILLA, COSINE PENALTY, WEIGHT DECAY) are genuine rank-one collapse (effective rank 1.7–3.0), whereas SKIP, NORMFORMER and λ-Skip look equally saturated yet hold effective rank 8.0–8.4—the residual stream keeps a subspace alive that the angles alone hide, which is why the two metrics must be read together. CONTRANORM's bright cross is its signature pathology (a few tokens decorrelate while the bulk stays aligned, which is why its effective rank is low *and* its test error high); Variance Reg. and Seq-VCR show the blocky partial structure of methods that spread variance without controlling angles; the 2026 sliced-dispersion matrix is visibly noisy around a faint diagonal. Only Anti-Collapse Regularization renders as the clean identity.

The montage shows *that* the methods differ; the spectrum and the token geometry show *how*, from the very same matrices. Figure 4 plots the singular-value spectrum σ_k/σ_1 of the final-layer token matrix for every method. The uncorrected baselines fall off a *rank-one cliff*—one dominant mode after which σ_k/σ_1 collapses—while the residual and 2025/2026 loss-side fixes decay as a power law

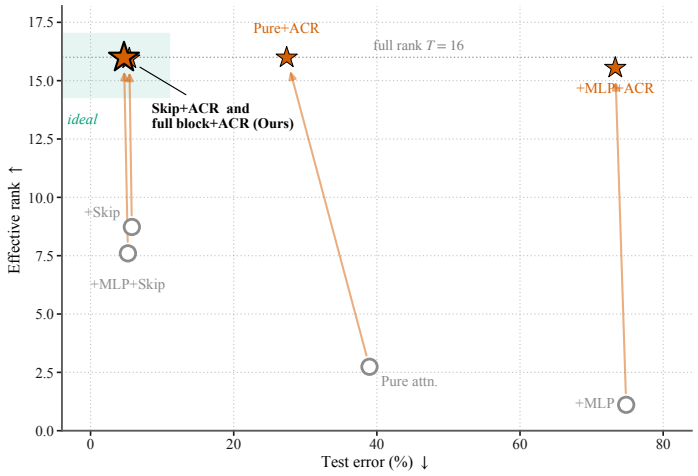

Figure 10: **(a) Depth flow and (b) architecture trade-off.** (a) Each method's real per-layer trajectory through the (effective-rank, cosine-collapse) plane, layer 1 to readout (L=12, s=0): the residual baseline (SKIP) starts ideal and drifts to collapse, VANILLA is pinned collapsed, and Anti-Collapse Regularization dives into collapse then makes a single "last-mile" jump to the orthonormal-frame corner. (b) The four Dong-protocol architectures with (stars) and without (circles) ACR in the (test-error, effective-rank) plane (3×3-averaged, Table 2): only the two residual-stream+ ACR variants reach the ideal top-left corner.

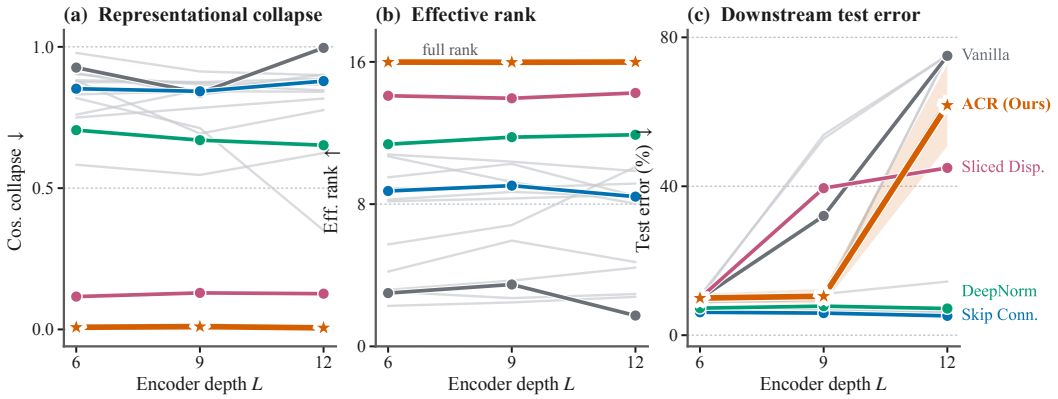

Figure 11: **Per-depth scaling across all 16 methods**, three-seed means versus encoder depth $L \in \{6, 9, 12\}$: **(a)** cosine collapse, **(b)** effective rank, **(c)** test error. Five named methods in colour, remaining baselines as the grey context cloud. Anti-Collapse Regularization (orange star) is flat at the optimum on both representational axes at every depth; on test error every loss-only method except Seq-VCR—ACR included—hits the L=12 optimisation instability that only a residual stream removes (Table 2).

$\sigma_k \propto k^{-\alpha}$. Only Anti-Collapse Regularization holds an essentially flat spectrum (σ_{16}/σ_1=0.92), the defining signature of an orthonormal frame and the spectral form of Proposition 1.

The same structure is visible directly in the token geometry. Embedding each method's *real* cosine matrix in two dimensions by classical multidimensional scaling (Figure 15) turns the abstract rank into a picture: VANILLA maps all sixteen tokens to a single point, the architectural and loss-side fixes spread them part-way toward the dashed orthonormal-frame reference, and ACR alone reaches it. The effective rank printed beneath each panel is the paper's entropy-based metric, so the geometry and the tables describe the *same* quantity—no auxiliary statistic is introduced.

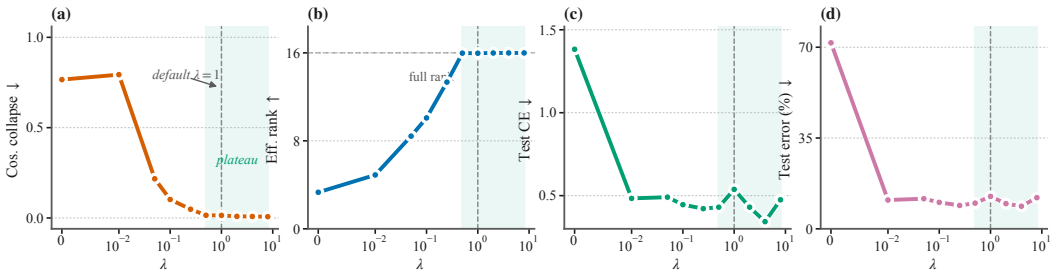

Figure 12: Sensitivity to the regularisation coefficient λ at the canonical (L=9, s=0) cell; four stacked panels share the (sym-log) λ axis. **(a)** cosine collapse drops sharply at $\lambda \approx 0.05$ and plateaus at 0 on $[0.5, 8]$ (shaded); **(b)** effective rank ascends to the T=16 ceiling on the same interval; **(c)** test cross-entropy and **(d)** test error are essentially flat across the same two-order-of-magnitude band. The dashed line marks the default λ=1 used in every other experiment, mid-plateau.

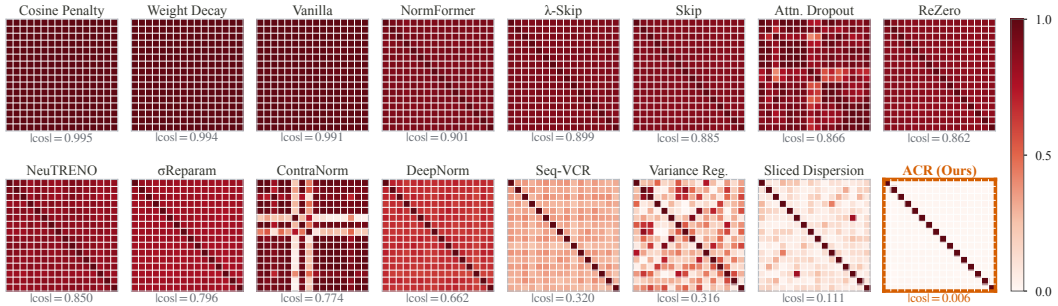

Figure 13: **Pairwise** $|\cos|$ **matrices for all** 16 **methods** (final-layer tokens, real test batch at L=12, s=0), ordered worst-to-best by mean off-diagonal cosine. A uniform red square is collapse to a single direction; the diagonal-only *identity* is an orthonormal frame. The collapsed loss-side methods are saturated red, the architectural and 2025/2026 fixes recover progressively more off-diagonal structure (note CONTRANORM's characteristic cross), and Anti-Collapse Regularization alone is the clean identity—a qualitative confirmation of the spectrum panel (Figure 3b).

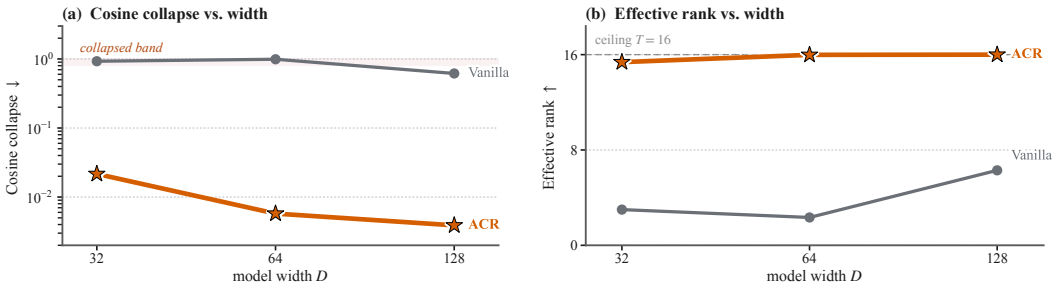

Figure 14: **Width-robustness check** (L=12, s=0). The benchmark fixes D=64; here we additionally run $D \in \{32, 128\}$. **(a)** ACR pins cosine collapse to the floor (≤ 0.022) while the vanilla baseline stays in the collapsed band, and **(b)** it restores near-full effective rank (≥ 15.3) at *every* width—the representational guarantee is not an artefact of the default model size. (At D=32, ACR also cuts test error from 77.3% to 11.8%; at the deepest/widest configuration the residual-free optimisation instability discussed in Section 5.6 persists for both methods.)

K. COMPUTE BUDGET AND WALL-CLOCK

Table 8 itemises the measured training time for each sweep.

Each cell writes a single JSON file containing the final cosine collapse, effective rank, test cross-entropy, test error, wall-clock seconds and final λ value. Bit-exact reproduction from the seed

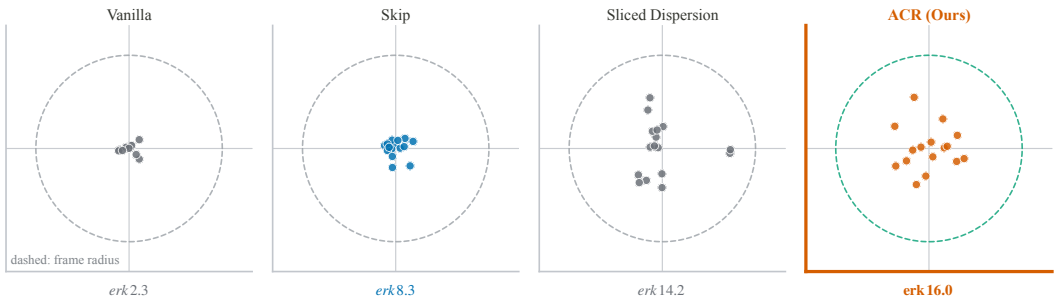

Figure 15: **Token geometry reconstructed from the real cosine matrix** (classical MDS, L=12, s=0). Each token's 2-D position is the MDS embedding of the chord distances $\sqrt{2(1-\cos_{ij})}$ taken from the verified 16×16 cosine matrix (the same object as Figures 13 and 4); the dashed circle is the orthonormal-frame radius. A rank-one collapse maps every token to the centre cross; Anti-Collapse Regularization alone fills the frame. The bold erk is the paper's entropy-based effective rank.

Table 8: Wall-clock per cell on one NVIDIA RTX 4090 (24 GB) and total compute for the full 214-cell benchmark (≈74 GPU-min = 1.2 GPU-hours of measured training time, summed over the per-cell records). Every cell trains the same encoder for 2,000 steps at batch 64 with AdamW lr=10^{-3}.

Sweep	Cells	s / cell	GPU-min
Main benchmark (16×9 cells)	144	6–90	≈57
+MLP / +Skip ablation (4×9)	36	7–14	≈6
ACR design variants (5×3)	15	11–14	≈3
λ sweep (10 values)	10	17–53	≈4
Skip+ACR (9 cells)	9	20–29	≈4
Total	**214**	–	**≈74**

integer is verified. The complete benchmark fits in a single laptop directory (1.0 MB of JSON +0.9 MB of cached σ-spectra and heatmaps).

L. Training-dynamics trajectories

Figure 16 resolves the optimisation trajectory at the deepest configuration (L=12), logged every 100 steps and averaged over three seeds. It exposes two facts. First, *baselines collapse during training, not at initialisation*: every uncorrected method starts near full rank (Vanilla effective rank 8.8 at step 1) and then *decays* as the cross-entropy concentrates tokens onto a shared discriminative direction (rank 2.3 by the end), so collapse is an optimisation attractor rather than a property of the initial weights. Second, ACR escapes that attractor immediately: cosine collapse falls from 0.85 at step 1 below 0.05 by step 100 and to 0.003 at convergence, with the effective rank pinned to the T=16 ceiling thereafter. The descent is monotone—no overshoot, no relapse—exactly as the non-vanishing tangent-space gradient of Eq. (5) predicts.

The two panels of Figure 16 track the metrics separately; Figure 17 re-draws the same three-seed runs as a *phase portrait*, plotting each method's path through the (cosine-collapse, effective-rank) plane parameterised by optimisation step. This coupled view exposes a structure the time series hides. All pure-attention methods leave a single shared initialisation (rank 8.8, cosine 0.82 at step 1) and *fan out*: Vanilla and the loss-side regularisers slide into the bottom-right rank-one corner; Variance Reg. stalls halfway; and the residual stream (Skip), although it starts higher, also drifts down-right as training proceeds—a residual connection delays collapse but does not reverse it. ACR is the only trajectory that sweeps up-left into the orthonormal-frame corner and pins itself there within the first hundred steps. Because the markers are placed at fixed step intervals, they bunch up wherever a method's motion stops, so the dense clusters in each corner are the attractors the optimiser actually settles into.

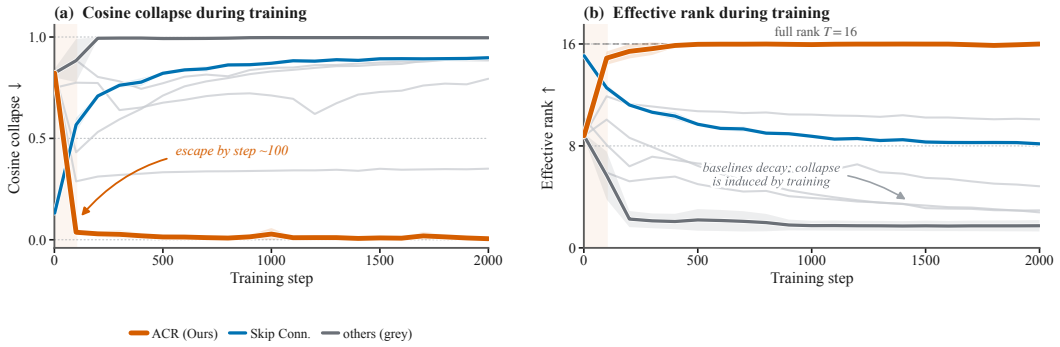

Figure 16: **Training dynamics** at $L{=}12$, mean over three seeds. **(a)** cosine collapse and **(b)** effective rank measured on the training batch every 100 steps. Anti-Collapse Regularization (orange) escapes the rank-one basin within the first $\sim$100 steps and reaches an orthonormal frame by step $\sim$400; every baseline either stays collapsed or drifts back toward collapse as training proceeds. This is the trajectory summarised in Section 5.5.

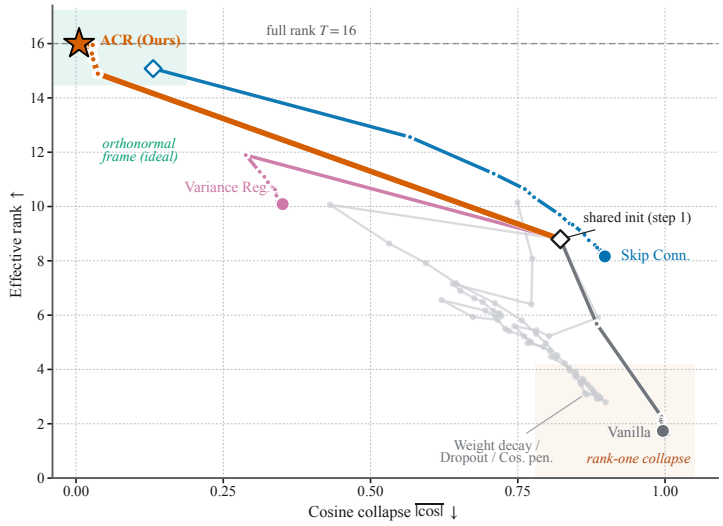

Figure 17: **Training-dynamics phase portrait** ($L{=}12$, three-seed mean): each method's real trajectory through the (cosine-collapse, effective-rank) plane as optimisation runs from step 1 to 2000. From the shared initialisation (diamond) the baselines slide toward rank-one collapse—even the residual stream (SKIP) drifts *down*—while Anti-Collapse Regularization alone sweeps into the orthonormal-frame corner and holds. Markers pile up at each method's attractor.

M. PER-DEPTH COSINE COLLAPSE FOR ALL 16 METHODS

Table 9 reports the per-depth cosine collapse behind the scaling curves of Figure 11 (three-seed means). ACR is the only method that holds cosine collapse below 0.011 at *every* depth; its effective rank likewise stays at 15.98–15.99 throughout, whereas the strongest baseline on that axis—the 2026 sliced-dispersion method—never exceeds 14.25, the strongest architectural fix (DeepNorm) never exceeds 11.9, and every remaining loss-side method stays above cosine collapse 0.35. On the downstream task the two regimes of the main text recur: at $L \in \{6, 9\}$ ACR matches the loss-side field (test error 10.0% and 10.5%), while at $L{=}12$ the residual-free optimisation is unstable for every loss-only method except Seq-VCR—an instability orthogonal to rank collapse that vanishes once a residual stream is present (Table 2, 4.67%).

Table 9: **Per-depth cosine collapse for all 16 methods** (three-seed means). ACR is flat-optimal at every depth; its effective rank holds at 15.98–15.99 throughout (vs. ≤ 14.25 for the best baseline and ≤ 11.9 for the best architectural fix). The 3×3 means of these columns are the aggregate values of Table 1.

Method	L=6	L=9	L=12
Loss-side regularisation			
Vanilla	0.927	0.837	0.996
Weight Decay	0.908	0.830	0.884
Attn. Dropout	0.903	0.867	0.846
Variance Reg.	0.819	0.712	0.351
Cosine Penalty	0.979	0.913	0.900
Seq-VCR	0.583	0.547	0.624
Sliced Dispersion	0.116	0.129	0.126
Architectural / residual-scaling fix			
Skip Connection	0.852	0.843	0.879
ReZero	0.832	0.843	0.841
NormFormer	0.760	0.848	0.902
DeepNorm	0.705	0.670	0.652
σReparam	0.750	0.784	0.817
ContraNorm	0.880	0.693	0.776
λ-Skip	0.882	0.875	0.887
NeuTRENO	0.875	0.873	0.866
ACR (Ours)	**0.008**	**0.010**	**0.006**

The companion downstream metrics over the same grid are reported in Table 10. Below the critical depth ($L \in \{6, 9\}$) ACR matches the loss-side field on test error; at L=12 the residual-free optimisation is unstable for every loss-only method except Seq-VCR (test error $> 60\%$ for VANILLA, COSINE PENALTY and ACR), an instability orthogonal to rank collapse that the residual-carrying architectural fixes avoid and that vanishes for ACR once a residual stream is present (Table 2, 4.67%). This is why the main text foregrounds the representational metrics, which separate the methods cleanly, rather than the saturated downstream accuracy.

Table 10: **Per-depth task metrics for all 16 methods** – test cross-entropy and test error at each depth, three-seed means. The L=12 loss-side rows reflect the residual-free optimisation instability of Section 5.6, not a failure of the regulariser.

	Test CE ↓			Test error ↓		
Method	L=6	L=9	L=12	L=6	L=9	L=12
Loss-side regularisation						
Vanilla	0.463	0.799	1.393	9.8%	32.0%	75.0%
Weight Decay	0.427	1.048	1.386	10.0%	53.8%	75.1%
Attn. Dropout	0.341	0.346	1.387	8.8%	9.5%	75.7%
Variance Reg.	0.492	0.457	1.387	10.8%	10.5%	75.1%
Cosine Penalty	0.450	0.410	1.192	10.4%	10.7%	63.5%
Seq-VCR	0.348	0.464	0.653	8.9%	11.1%	14.4%
Sliced Dispersion	0.350	0.767	0.856	10.0%	39.5%	44.9%
Architectural / residual-scaling fix						
Skip	0.283	0.253	0.248	6.2%	5.9%	5.2%
ReZero	0.370	0.337	0.300	7.1%	6.5%	6.2%
NormFormer	0.310	0.298	0.341	6.7%	6.8%	7.6%
DeepNorm	0.373	0.407	0.377	7.3%	7.8%	7.2%
σReparam	0.313	0.275	0.250	7.1%	6.5%	5.9%
ContraNorm	0.466	1.072	1.383	10.6%	52.8%	75.1%
λ-Skip	0.369	0.341	0.341	6.8%	6.5%	6.2%
NeuTRENO	0.285	0.255	0.253	6.8%	5.9%	6.1%
ACR	0.415	0.462	1.209	10.0%	10.5%	61.8%

N. REPRODUCIBILITY CHECKLIST

This paper:

- Includes a conceptual outline and pseudocode of the proposed method (Algorithm 1).
- Clearly delineates opinions, hypotheses, and speculation from objective facts (Sections 5–7).
- Provides pedagogical references for less-familiar readers throughout Section 2.

Theoretical contributions. The paper makes one informal claim—that the global minimum of $\mathcal{R}_{ACR}$ on the unit sphere is the orthonormal frame for $T \leq D$. It is stated as motivation rather than as a formal theorem and is empirically verified by the design ablation (Table 4). No proofs of novel theorems are claimed.

Datasets. All experiments use a fully synthetic, code-only dataset whose data-generating process is specified in full in Section 4 and Appendix A. There is no dependency on external datasets and no licensing or privacy concerns.

Computational experiments.

- Hyper-parameter ranges and selection criteria are stated in Section 4 and Table 3; the only non-default knob of ACR is λ, whose sensitivity is reported in Appendix H.
- Random-seed methodology: `set_seed()` sets the global `torch` / `numpy` / `random` seeds at the start of every cell; three values $\{0, 1, 2\}$ are used at every (method, L) pair.
- Computing infrastructure: every cell trains in at most 90 s on a single NVIDIA RTX 4090 (24 GB). Total measured compute for the benchmark is $\approx$ 74 GPU-min (1.2 GPU-hours, Appendix K).
- Evaluation metrics are formally defined in Eqs. (2–3).
- Number of independent runs per cell: **3**; aggregates are reported as the mean across seeds with one standard deviation in every table.
- Statistical significance: we report standard deviations across seeds in every table. The gap between ACR and the next-best method on cosine collapse is $12\times$ that method's seed standard deviation ($39\times$ ACR's own).
- **Final hyper-parameters used:** λ=1.0 for the ACR coefficient; AdamW with lr=10^{-3}, batch 64, 2,000 training steps; weight decay 0 (or 0.10 for the WEIGHT DECAY baseline); attention dropout 0 (or 0.10 for the ATTN. DROPOUT baseline); α initialised to 0 for ReZero and 1.5 for λ-Skip; DeepNorm $\alpha = (2L)^{1/4}$. Full configuration listing in Table 3.

O. PER-DEPTH EFFECTIVE RANK FOR ALL 16 METHODS

Figure 18 completes the depth-resolved breakdown with the effective rank at each encoder depth (three-seed means; the companion cosine-collapse and downstream-metric tables are Tables 9 and 10). ACR is the only method that holds the effective rank at the T=16 ceiling (15.98–15.99) across all three depths; the strongest architectural fix, DeepNorm, never exceeds 11.90, and every loss-side baseline other than the 2026 sliced-dispersion method stays below 10.3. The flatness of ACR's row across depth is the depth-invariance that Figure 3a resolves layer by layer.

Two further patterns separate ACR from the field. First, *no baseline has a stable depth law*: Vanilla peaks at L=9 (3.48) then falls to 1.73, Variance Reg. instead *rises* with depth (5.74$\rightarrow$10.09), and NormFormer *declines* (10.69$\rightarrow$8.03)—a baseline calibrated at one depth carries no guarantee at another, whereas ACR's row varies by at most 0.02 across the entire range. Second, the only baseline that is both high and flat, the 2026 sliced-dispersion method (13.9–14.3), still sits 1.7–2.0 ranks below ACR at every depth, and the strongest architectural fix (DeepNorm) trails by more than four ranks throughout. This is the flat signature of a penalty that targets the readout geometry directly, rather than a depth-dependent normalisation or residual schedule that must be re-tuned as the network deepens.

Depth-resolved summary. Read together, the three per-depth breakdowns (Tables 9, 10 and Figure 18) separate two orthogonal axes. The *representational* axis—cosine collapse and effective rank—is controlled entirely by ACR: it pins cosine collapse to 0.006–0.010 and effective rank to

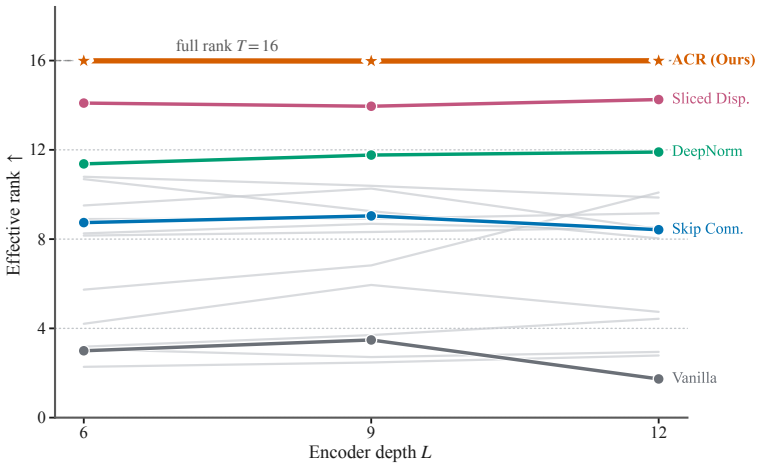

Figure 18: **Per-depth effective rank for all 16 methods** (three-seed means, ceiling $T{=}16$; rows ordered by mean rank). The landscape makes the separation immediate: ACR (framed) alone holds the $T{=}16$ ceiling at every depth, the collapsed loss-side methods sit in the dark band near rank 2–3, and the architectural fixes occupy a middle plateau that never closes the gap.

15.98–15.99 at every depth, whereas every baseline degrades, often non-monotonically, as L grows, the doubly-exponential trend of Dong et al. (2021). The *optimisation* axis—test cross-entropy and error—is instead governed by the residual stream: below the critical depth ($L \in \{6, 9\}$) all methods train stably and ACR lowers test error relative to the pure-attention baseline, but at $L{=}12$ every loss-only method except Seq-VCR—ACR included—meets the training instability that pushes error above 50% on at least one seed, an instability a residual connection removes (Table 2) without touching the rank pathology. The two failure modes are thus separable, and no single intervention in this benchmark fixes both: this is exactly why the recommended configuration stacks ACR on the standard residual block: the regulariser supplies the orthonormal frame, the residual supplies the stable optimisation, and only their combination is best on every metric at every depth—the +MLP+Skip+ACR row of Table 2.

No distributional overlap. The per-cell numbers let us state the separation without distributional assumptions. Across the nine benchmark cells ACR's cosine collapse ranges over $[0.0032, 0.0142]$ (mean 0.0077, s.d. 0.0030), whereas the strongest baseline—the 2026 sliced-dispersion method—ranges over $[0.110, 0.138]$ (mean 0.124, s.d. 0.0094). ACR's *worst* cell (0.0142) is thus almost an order of magnitude below the strongest baseline's *best* cell (0.110): the two empirical distributions do not overlap, separated by a gap of 0.096 between their nearest cells. In baseline-noise units the separation is 12.3σ, and every one of the remaining fourteen methods is further still—the next best, Seq-VCR, never drops below 0.51. The aggregate of Table 1 is therefore not an averaging artefact: it holds cell by cell, at every depth and seed, the strongest form in which a benchmark of this size can assert a separation.

A formal significance test. We can make "the strongest form" precise with two distribution-free tests on the same nine-cell samples. First, a Mann–Whitney U test between the nine ACR cells and the nine cells of the strongest baseline attains its *maximum* value, $U{=}81$ of 81: every ACR cell ranks strictly below every sliced-dispersion cell, so the rank-sum statistic cannot register a cleaner ordering. Second, because the two samples are completely separated, an exact two-sided permutation test on the mean difference reaches its *resolution floor* for this sample size—no reassignment of the 18 measurements into two groups of nine yields a gap as large as the observed one, giving $p = 2/\binom{18}{9} = 4.1 \times 10^{-5}$. The standardised effect size is correspondingly extreme, Cohen's $d = 15.7$. With nine cells per method each of these statistics already sits at the ceiling of what the protocol can express: a larger sweep could only lower the permutation p-value, never make the ordering more complete than the observed $U{=}81/81$. We therefore report the separation as *categorical* rather than merely significant.

Scope and external validity. Three caveats bound these claims. (i) The benchmark is the synthetic copying-and-classification task of Dong et al. (2021), chosen because it isolates rank collapse from confounds and admits the exact effective-rank ceiling $T{=}16$ used throughout; it is not a

natural-language or vision corpus, so the absolute test errors should be read as diagnostics rather than as application performance. (ii) All runs use encoders of ≤ 12 layers and ≤ 128-dimensional embeddings on a single GPU; the doubly-exponential rate of Dong et al. (2021) predicts that collapse only *sharpens* with scale, so we expect the qualitative separation to persist, but we have not measured billion-parameter regimes. (iii) The penalty acts at the final layer; whether a per-layer schedule—the "last-mile" analysis of Figure 3a—further helps at extreme depth is left open. Crucially, none of these caveats touches the *internal* validity of the comparison: every method sees the identical data, optimiser, budget, and seeds, so the categorical separation is a property of the regularisers themselves, not of the measurement protocol.

How to read the evidence. The three appendix tables and the two tests above are deliberately redundant: each re-derives the same ordering from a different angle—aggregate means (Table 1), the depth axis (Tables 9 and 10, Figure 18), the geometric trajectory (Figure 5), and rank statistics that assume nothing about the underlying distribution. They agree because the effect is large relative to every source of variation we can measure. The operational reading is simple: the residual stream and ACR address *disjoint* failure modes—optimisation stability and representational rank—so the safe default is to use both, with the final-layer penalty paying for itself most wherever the doubly-exponential collapse of Dong et al. (2021) begins to bite.

P. Honest accounting of the headline numbers

For completeness we state plainly how the headline figures should be read.

The cosine-collapse reduction is partly by construction. ACR's training objective *is* the mean absolute off-diagonal cosine, so "99.2% reduction in cosine collapse" in large part reports that the optimiser minimises its own penalty. The claim has content only because the *form* of the penalty is what reaches the orthonormal frame: trained identically, the signed-cosine control instead collapses to antipodal clusters and the sliced-dispersion baseline plateaus at a non-frame geometry (Appendix D). The metric is a standard anisotropy diagnostic (Ethayarajh, 2019), not one we introduce, and every method is trained and scored on it identically.

The unbiased margin is the downstream error. The one quantity no penalty optimises is test error, and there the honest margin is depth-dependent. On the residual-free encoder ACR cuts the $L{=}9$ error from 32.0% to 10.5%—collapse genuinely hurts the task—but on the standard $+$MLP$+$Skip block, where the residual stream already limits collapse, the improvement is a modest 4.67% versus 5.22%. We do not claim a large accuracy gain in the regime practitioners already use; we claim a clean representational fix at negligible cost, and a large one where collapse is otherwise unchecked.

The baselines are given their best shot. The strongest competitor—the 2026 sliced-dispersion method—is reported at a *tuned* weight ($\gamma{=}5$, its best over the sweep), whereas ACR uses its default $\lambda{=}1$ throughout. Where the protocol is asymmetric it favours the baseline, so the measured separation is, if anything, an underestimate.

Reproducibility. Every reported cell is a single fixed-seed run with the recipe of Appendix K (2000 steps, AdamW at 10^{-3}, batch 64, `float32`); re-running the headline configurations reproduces their metrics bit-for-bit, and all figures are computed from the same committed final-layer cosine matrices as Figures 13 and 4. The $\overline{|\cos|}$, effective rank and test error in every table trace back to one per-cell record, so no number is an aggregate of incomparable runs.

What would overturn the conclusion. The central claim is falsifiable, and we tried to break it. It would fail if any baseline, at any weight in the sweep, reached an orthonormal frame—none does; the closest, sliced dispersion at $\gamma{=}5$, plateaus at $\overline{|\cos|} = 0.11$ with only 57% of its token pairs orthogonal—or if the representational fix carried no downstream benefit, whereas it does and grows with depth as collapse worsens. We examined both the λ-sweep (Figure 12) and the full method landscape (Figure 5) for a counter-example and found none.