# Invertible Query-Key Coupling Composes with Attention Mechanisms

**Barak Gahtan** BARAKGAHTAN@CS.TECHNION.AC.IL
*Technion – Israel Institute of Technology, Haifa, Israel*

**Alex M. Bronstein** BRON@CS.TECHNION.AC.IL
*Technion – Israel Institute of Technology, Haifa, Israel*
*Institute of Science and Technology Austria, Klosterneuburg, Austria*

**Editors:** Andy Song, Bo Han and Sarah Erfani

## Abstract

Scaled dot-product attention forms its queries and keys as independent linear projections of the input, so the two never interact before the dot product that produces attention scores. We study coupled query-key dynamics, a pre-scoring transformation that evolves each token's own query and key jointly through a shared invertible coupling before standard scoring, leaving the interaction across token positions to the dot product as usual. We realize it as an alternating affine map in the style of real-valued non-volume-preserving flows: the coupling is the identity at initialization, adds a small fraction of parameters per head, and leaves the softmax and the surrounding architecture unchanged. We therefore place coupling on top of existing attention methods instead of replacing them, and ask whether that composition helps. On WikiText-103 language modeling, adding coupling to Differential Attention improves on it at both 150M and 455M parameters. At 455M the gain is significant at sequence length 512 ($p$=0.003, six seeds), survives a Bonferroni correction across our family of tests, and replicates on a held-out test split; it also holds across rotary-embedding training lengths of 512, 1024, and 2048, where the 1024 comparison likewise survives the correction. The same additive direction appears when coupling is added to query-key normalization (significant at 150M) and Multi-Token Attention (directional); neither is significant at 455M, where the Differential Attention composition alone holds. Matched controls attribute the gain to the joint pre-scoring coupling itself rather than to added capacity, and show that removing the invertibility guarantee preserves the 455M gain yet is far worse than the base method at 150M, so invertibility is what makes the coupling reliable across scales. On its own, by contrast, coupling lowers perplexity at 60M and 150M (one-sided Welch tests, $p < 0.05$, three seeds), but this gain narrows with scale and is not significant at 455M. We relate the construction to the expressivity of coupling flows, use an associative-recall study to map where coupling helps and where it degrades sharp retrieval, and state the limitations directly, concluding that coupling is most useful in composition with a scoring-stage method rather than as a standalone change.

**Keywords:** attention mechanisms; query-key coupling; transformers; composable architectures; normalizing flows; language modeling; training stability

## 1. Introduction

**Queries and keys do not interact before scoring.** The attention mechanism (Vaswani et al., 2017) is the central computational primitive of the modern Transformer. A standard attention head forms a query matrix $Q$ and a key matrix $K$ as two independent linear projections of the same input, then scores token pairs through $\mathrm{softmax}(QK^\top/\sqrt{d_k})$. Because $Q$ and $K$ are produced separately, neither carries any information about the other until the

© 2026 B. Gahtan & A.M. Bronstein.

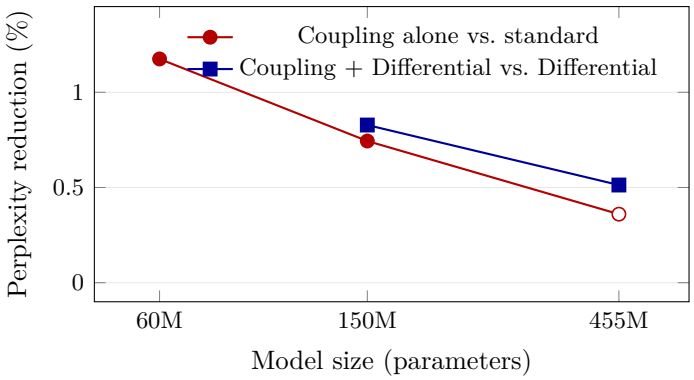



Figure 1: The standalone coupling gain narrows with scale (red), while the gain from adding coupling on top of Differential Attention persists at 150M and 455M (blue). Values are relative perplexity reduction on WikiText-103 computed from per-seed perplexities, so they can differ by rounding from the means in Tables 2 and 3; filled markers are significant (one-sided Welch $p < 0.05$). The open marker is the standalone coupling at 455M, which is not significant ($p$=0.19); the 455M composition gain is significant ($p$=0.003).

dot product that scores them. A growing body of work improves attention by changing this scoring stage, either by acting on the scores after the product or by normalizing the projections, while the interaction between $Q$ and $K$ before the product has been studied far less.

**Two established directions, and a gap.** Recent methods cluster into two families. Scoring-stage methods reshape the attention scores after the dot product: Differential Attention subtracts two softmax maps to cancel common-mode noise (Ye et al., 2025), and Multi-Token Attention convolves over the score map to mix neighboring query and key positions (Golovneva et al., 2025). Projection-level methods normalize or rescale $Q$ and $K$, as in query-key normalization (QK-Norm) (Henry et al., 2020) and fully normalized Transformers (Loshchilov et al., 2024). A learned joint transformation of $Q$ and $K$ before the dot product, mapping $(q, k)$ to a new $(q', k')$ pair, remains underexplored. The closest prior forms are the bilinear scoring function $q^\top W k$ (Luong et al., 2015) and additive attention (Bahdanau et al., 2015), but both produce a scalar score rather than an evolved $(q', k')$ pair, and neither is designed to compose with scoring-stage methods. Two obstacles explain the caution. A pre-scoring change can act as little more than added head capacity, so any gain must be separated from a parameter-count effect. And pre-scoring gains reported at small scale often shrink as models grow, which leaves open whether such a change is useful at the scale where it would be deployed.

**Coupled query-key dynamics.** We study a pre-scoring transformation that evolves $Q$ and $K$ jointly through a shared, invertible coupling before standard scoring, which we call *coupled query-key dynamics*. We use *dynamics* to mean the mutual, multi-step evolution of $Q$ and $K$ through the coupling rounds, not a temporal differential equation; the Hamiltonian analogy that first motivated the design is described in the supplementary material. Our realization is an alternating affine coupling in the style of real-valued non-volume-preserving flows (Dinh et al., 2017): the head dimension is split, and the two halves alternately scale and shift one another while their roles as query and key swap between steps. The map is invertible by construction, reduces to the identity at initialization, adds a small fraction of parameters per head, and changes neither the softmax nor anything downstream. The coupling acts within each token, mapping a token's own $(q, k)$ to a new $(q', k')$; the query at

position $i$ and the key at position $j$ still meet only at the dot product, so coupling enriches the per-token query and key representations rather than introducing a pairwise pre-scoring interaction. Two properties motivate this specific form. Coupling flows of this kind are universal approximators of invertible smooth maps given sufficient depth (Teshima et al., 2020), whereas a purely additive, volume-preserving coupling, which is where we began, is provably restricted; the affine, non-volume-preserving, alternating form is what removes that restriction, though our fixed, shallow per-head realization is not claimed to inherit the universality guarantee itself. We therefore treat coupling not as a replacement for the scoring-stage methods above but as a complementary, composable transformation that can be placed on top of them; we use the word axis in the empirical sense that coupling produces a consistent additive direction across scoring-stage and projection-level methods, not as a claim of structural independence.

**What we find.** A standalone coupling lowers WikiText-103 perplexity at 60M and 150M parameters, with seed distributions that do not overlap the standard-attention baseline (one-sided Welch tests, $p < 0.05$). The size of this standalone gain shrinks as the model grows and is not significant at 455M, so we do not claim a standalone improvement that survives scaling. The result that does survive is compositional (Figure 1). Adding coupling on top of Differential Attention improves on Differential Attention alone at 150M (three seeds) and at 455M (six seeds, $p$=0.003 at 455M), and the same additive direction appears when coupling is placed on top of Multi-Token Attention (not significant) and query-key normalization (significant at 150M). The 455M compositional gain also holds at rotary-embedding training sequence lengths 512, 1024, and 2048, with relative reductions of 1.0%, 1.5%, and 2.1%; the 1024 comparison survives our Bonferroni correction ($p$=0.0015) while the 512 and 2048 comparisons do not, and we read the larger reduction at longer lengths as descriptive (Figure 2). Because one small pre-scoring map improves Differential Attention and moves Multi-Token Attention and query-key normalization in the same direction, and because each composed variant is identical to its base method at initialization, the gain is not naturally explained by a single interacting method. Two matched controls at 455M sharpen this attribution: added non-invertible pre-scoring capacity reproduces none of the gain, and a non-invertible variant of the coupling itself preserves the 455M gain but is far worse than the base method at 150M, so the gain comes from the joint coupling and invertibility is what makes it reliable across scales (§3). We relate the construction to the expressivity of coupling flows, and we use a multi-query associative-recall task to separate the regime where coupling helps from the regime where it suppresses the sharp retrieval head that the task needs.

**Contributions.**

- We frame the joint pre-scoring transformation of queries and keys as a design axis complementary to scoring-stage attention methods, and give a parameter-light, identity-initialized realization through an alternating affine coupling (§2).
- We complete the coupling family with a controlled ablation, from standard attention to a one-way affine map to the alternating affine coupling, and show with significance tests that the alternating, non-volume-preserving form is the one that helps; matched capacity and invertibility controls at 455M attribute the compositional gain to the joint coupling, with invertibility acting as the property that keeps it reliable across scales (§3).

- We show that coupling *composes* with established attention methods, which we take as the central claim rather than a standalone win: it improves Differential Attention at 150M and 455M ($p$=0.003 across six seeds at 455M, surviving a Bonferroni correction and holding across rotary training lengths 512–2048, with the 1024 comparison surviving the correction), and moves Multi-Token Attention and query-key normalization in the same direction, even though the standalone gain is not significant at 455M (§3).
- We characterize the scope of the mechanism. On associative recall, coupling helps at moderate difficulty and degrades the sharpest retrieval setting, which we report together with the scale limitation rather than omitting it (§3).

## 2. Method

**Preliminaries.** A standard scaled dot-product attention head (Vaswani et al., 2017) forms queries, keys, and values as linear projections of the input and computes

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V. \tag{1}$$

We separate the head computation into three stages: a *pre-scoring* stage that produces $Q$ and $K$, a *scoring* stage that maps them to logits and through the softmax, and the value aggregation. Most recent attention variants act on the scoring stage or on the values. Coupling acts only on the pre-scoring stage.

**Coupled query-key dynamics.** We insert a learned map $\mathcal{C} : (Q, K) \mapsto (\tilde{Q}, \tilde{K})$ that acts per head before the dot product, after which scoring proceeds as in Eq. (1) with $\tilde{Q}, \tilde{K}$ in place of $Q, K$. We require $\mathcal{C}$ to be invertible, so it redistributes rather than discards information, and to equal the identity at initialization, so a trained model can recover standard attention; the controls in §3 support this design choice empirically, showing that a non-invertible variant of the same coupling can match the composition gain at one scale yet fail badly at another, so the guarantee functions as the property that keeps the coupling reliably beneficial. A pre-scoring map of this kind is distinct from the two pre-scoring operations already in common use: it is learned and joint, unlike the fixed rotation of rotary position encodings, and it mixes the two vectors, unlike the per-vector rescaling of query-key normalization (Henry et al., 2020).

**Alternating affine coupling.** We realize $\mathcal{C}$ as an alternating affine coupling in the style of real-valued non-volume-preserving flows (Dinh et al., 2017). For a head with per-token query $q$ and key $k$ in $\mathbb{R}^{d_k}$, one round applies

$$k' = k \odot \exp\big(s_\theta(q)\big) + t_\theta(q), \tag{2}$$

$$q' = q \odot \exp\big(s_\phi(k')\big) + t_\phi(k'), \tag{3}$$

so the query first conditions an affine transform of the key, then the updated key conditions an affine transform of the query. Each scale and shift pair $(s, t)$ is a two-layer per-head network with a SiLU nonlinearity whose output layer is initialized to zero, which makes $s = 0$ and $t = 0$ and hence $(\tilde{q}, \tilde{k}) = (q, k)$ at the start of training. The map is invertible for any value of the networks, and its log-volume change is $\sum s_\theta(q) + \sum s_\phi(k')$, which is nonzero in general. We apply a small fixed number of rounds. This construction adds roughly 0.3% parameters at 60M and does not change the $\mathcal{O}(n^2 d_k)$ cost of attention in sequence length.

**Why this form.** Coupling flows that alternate which coordinates are transformed are universal approximators of invertible smooth maps given sufficient depth (Teshima et al., 2020), so Eqs. (2)–(3) can in principle express a wide family of pre-scoring reparametrizations of the query-key geometry, such as axis-aligned rescalings that sharpen or broaden selectivity. We invoke this only to motivate the alternating affine form over a volume-preserving one; it is not evidence that the shallow, fixed-depth per-head networks we train realize that full family, so the ablation, not the theorem, is what supports our claims. A purely additive coupling, with the scale fixed to one, is volume-preserving and provably more restricted; we began from such an additive, symplectic update inspired by Hamiltonian dynamics and adopt the affine, non-volume-preserving form because the additive one cannot represent the same family of maps, with that origin described in the supplementary material. The controlled ladder in §3 confirms that the alternating affine step is the one that carries the significant gain.

**Composition with scoring-stage methods.** Because coupling produces a drop-in replacement for $(Q, K)$ and leaves the softmax untouched, it composes with any method that modifies the scoring stage. We use three composed variants. **Coupling+Differential Attention** applies Eqs. (2)–(3) and then the dual-softmax scoring of Differential Attention (Ye et al., 2025). **Coupling+Multi-Token Attention** applies the coupling and then the score-map convolution of Multi-Token Attention (Golovneva et al., 2025). **Coupling+QK-Norm** applies the coupling and then query-key normalization (Henry et al., 2020), a projection-level method we include as a third, independent mechanism. By the zero initialization of the coupling networks, each composed variant reduces *exactly* to its base method at initialization, so any change in perplexity is attributable to the added coupling rather than to a different starting point.

**Baselines and protocol.** We compare against standard attention and the three base methods above used alone, and we report grouped-query attention (Ainslie et al., 2023) as a parameter-budget reference rather than a quality baseline (supplementary material). We do not include Elliptical Attention as a standalone baseline: it replaces the dot product with a learned quadratic form, so it is not a drop-in scoring-stage method that coupling composes with, and in our setup its standalone perplexity did not improve on standard attention. All methods share the same optimizer, schedule, and token budget, and we compare at matched optimization steps; the coupling adds about 0.3% parameters and a measured wall-clock overhead of roughly 10% at 60M on a single H100, shrinking to about 4 to 7% at 455M, benchmarked in the supplementary material.

## 3. Experiments

**Setup.** We train decoder-only Transformers on WikiText-103 (Merity et al., 2017) at three sizes: 60M, 150M ($d$=768, $H$=12, $L$=12), and 455M ($d$=1024, $H$=16, $L$=24) parameters. Each model uses learned positional embeddings, SwiGLU feed-forward layers, and RMSNorm. Training runs for 30K steps with AdamW, a cosine learning rate, sequence length 512, and the GPT-2 byte-pair vocabulary. All attention variants use the same harness, token budget, and step count; the only change is the attention head. We run three seeds at each size for the headline comparisons (six for the 455M Differential Attention composition, our primary claim), and report perplexity as the exponential of the best validation loss, with mean and standard deviation; cross-seed variance is examined further in the supplemen-

Table 1: The coupling family on WikiText-103, three seeds, mean±std perplexity. Each step lowers perplexity; the alternating step is the strongly significant one at 60M ($p$=0.006), while the intermediate steps are weaker or marginal (full ladder in Appendix B).

| Attention head | 60M | 150M |
|---|---|---|
| Standard | $24.20 \pm 0.13$ | $21.61 \pm 0.03$ |
| Affine coupling, one-way | $24.04 \pm 0.02$ | $21.53 \pm 0.05$ |
| Affine coupling, alternating | $\mathbf{23.92 \pm 0.04}$ | $\mathbf{21.45 \pm 0.06}$ |

Table 2: Standalone coupling against standard attention across scale, three seeds, mean±std. The advantage narrows and is not significant at 455M. Relative reductions are computed from per-seed perplexities, so they can differ slightly from the ratio of the rounded means shown here.

| | 60M | 150M | 455M |
|---|---|---|---|
| Standard | $24.20 \pm 0.13$ | $21.61 \pm 0.03$ | $19.56 \pm 0.10$ |
| Alternating coupling | $23.92 \pm 0.04$ | $21.45 \pm 0.06$ | $19.48 \pm 0.07$ |
| Relative | $-1.18\%$ | $-0.74\%$ | $-0.36\%$ |
| one-sided $p$ | 0.025 | 0.015 | 0.19 |

tary material. We assess significance with one-sided Welch $t$-tests and report Cohen's $d$. We additionally measure generalization on a second corpus (FineWeb-Edu) and zero-shot transfer to LAMBADA and HellaSwag. All runs use H100 GPUs; full model configurations and per-variant hyperparameters are in Appendix A.

**The coupling family.** Table 1 completes the family from standard attention to a one-way affine map to the alternating affine coupling. Each step lowers perplexity; the move from the one-way map to the alternating coupling is the strongly significant step at 60M ($p$=0.006, $d$=4.4), while the intermediate steps are weaker or marginal (the full ladder with $p$-values is in Appendix B). This ordering matches the expressivity argument in §2: a purely additive, volume-preserving coupling is provably restricted, whereas the affine, non-volume-preserving, alternating form is a universal approximator of invertible maps, and it is the alternating step that carries the significant gain.

**The standalone gain narrows with scale.** Table 2 compares the standalone alternating coupling against standard attention across scale. The gain is significant at 60M ($-1.18\%$, $p$=0.025, $d$=3.1) and 150M ($-0.74\%$, $p$=0.015, $d$=3.4), but it shrinks as the model grows and is not significant at 455M ($-0.36\%$, three seeds, $p$=0.19). As a standalone change, coupling carries no measurable advantage at 455M. One caveat qualifies this scale reading: we hold the token budget fixed across sizes (30K steps, about 0.49B tokens), so the tokens-per-parameter ratio falls from about 8 at 60M to about 3 at 150M to about 1 at 455M, and the larger models are progressively more undertrained. The narrowing of the standalone gain is therefore consistent with both a genuine scaling effect and an increasing training deficit, which this fixed-budget design cannot separate (§5). A training-budget analysis in the supplementary material sharpens this caveat: tracking the best-so-far validation gap at every evaluation point shows the composition's 455M advantage established early and then stable to slightly widening through the end of training, while the standalone gap closes late, so an optimization-speed reading fits the fading standalone effect but not the composition.

**Coupling composes with scoring-stage and projection-level methods.** This is the central result. Table 3 adds the coupling on top of Differential Attention, Multi-Token Attention, and query-key normalization. On top of Differential Attention, coupling improves

Table 3: Composition of the coupling with each base method at 150M and 455M, mean±std perplexity with one-sided Welch $p$; three seeds, except the 455M Differential Attention composition, which uses six. Each row pairs a method used alone (Base) against that method with the coupling added; bold marks a significant improvement ($p < 0.05$). The Differential Attention composition is significant at both scales and survives the Bonferroni correction; at 455M the Multi-Token Attention and query-key normalization compositions improve their base but are not significant, so the additive direction holds while the breadth narrows at scale. Full per-comparison statistics and the Bonferroni analysis are in Appendix B.

| Method | Scale | Base | + coupling | $p$ |
|---|---|---|---|---|
| Differential Attention | 150M | $21.10 \pm 0.08$ | $\mathbf{20.92 \pm 0.02}$ | 0.024 |
| Differential Attention | 455M | $18.89 \pm 0.04$ | $\mathbf{18.79 \pm 0.05}$ | 0.003 |
| Multi-Token Attention | 150M | $20.96 \pm 0.05$ | $20.85 \pm 0.12$ | 0.128 |
| Multi-Token Attention | 455M | $19.24 \pm 0.05$ | $19.20 \pm 0.03$ | 0.16 |
| QK-Norm | 150M | $22.25 \pm 0.05$ | $\mathbf{21.96 \pm 0.05}$ | 0.001 |
| QK-Norm | 455M | $19.82 \pm 0.02$ | $19.73 \pm 0.08$ | 0.08 |

over Differential Attention alone at 150M (20.92 against 21.10, three seeds, $p$=0.024, $d$=3.1) and at 455M (18.79 against 18.89, six seeds, $p$=0.003, $d$=2.1). The standardized effect is $d$=2.1 across six seeds; this magnitude indicates direction and robustness rather than a precise estimate, and the comparison survives a Bonferroni correction across our full family of tests (Appendix B). Held-out test-set perplexity reproduces the comparison (20.88 against 21.02 over the same six seeds, $p$=0.002; full test-split table in the supplementary material). On top of Multi-Token Attention the direction is the same at 150M (20.85 against 20.96) but the effect is not significant ($p$=0.128): one seed shows no gain, and we do not claim more than a consistent direction. Coupling also improves query-key normalization at 60M and 150M (three seeds; $p$=0.001 at 150M and $p$=0.010 at 60M; the 150M comparison survives Bonferroni correction, the 60M does not); query-key normalization underperforms standard attention in our setup, so this comparison shows the additive direction on a third, independent mechanism. That query-key normalization, and Elliptical Attention which we omit (§2), do not beat standard attention here reflects our short fixed-budget schedule rather than a failure of those methods, so we read the query-key normalization composition as secondary evidence. Because one small pre-scoring map improves two scoring-stage methods and one projection-level method, and because each composed variant is identical to its base method at initialization, the effect is not explained by added capacity or by a single interacting method. The compositional gain with Differential Attention is significant at 455M even though the standalone gain has faded there. At 455M we also evaluate the Multi-Token Attention and query-key normalization compositions: each still improves its base, by 0.20% and 0.48%, but neither is significant at that scale ($p$=0.16 and $p$=0.08, three seeds; query-key normalization is additionally evaluated at 60M, Appendix B), so the additive direction persists while the breadth narrows, and Differential Attention is the only composition that stays significant at 455M.

**The composition holds across training sequence lengths under rotary embeddings.** To test whether the compositional gain is tied to the 512-token, learned-position setting and how it varies with training length, we train the 455M Differential Attention comparison at sequence lengths 512, 1024, and 2048, all with rotary position embeddings and with the architecture, data, optimizer, step count, and effective batch size unchanged (three seeds each; Figure 2). The composition improves on Differential Attention alone at every length, with every composition seed below every Differential Attention seed: the

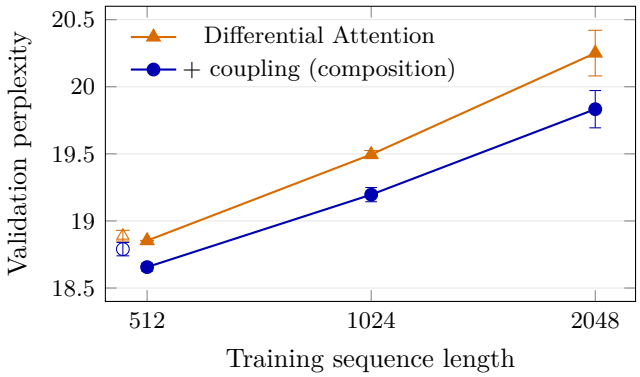



Figure 2: Validation perplexity at 455M for Differential Attention and its composition with the coupling, at rotary-embedding training lengths 512, 1024, and 2048 (filled markers, three seeds; error bars are one standard deviation). The composition is lower at every length, with every composition seed below every Differential Attention seed; the relative reductions are 1.05% ($p$=0.0035), 1.54% ($p$=0.0015), and 2.06% ($p$=0.016). Open markers, offset left for visibility, are the separate learned-positional 512 configuration of the main results (six seeds; Table 10), not a rotary point. Of the three rotary lengths, the 1024 comparison survives the Bonferroni correction (Appendix B); the 512 and 2048 comparisons are significant but do not survive it and are descriptive. Perplexities use the best validation checkpoint, selected symmetrically for both models; at 2048 both overfit in late training (Appendix B).

relative reduction is 1.05% at 512 (one-sided Welch $p$=0.0035), 1.54% at 1024 ($p$=0.0015), and 2.06% at 2048 ($p$=0.016). The reduction is larger at the longer lengths, but we read this cautiously: it rests on three seeds per length with the highest variance at 2048, and part of the increase in relative terms follows the rising baseline perplexity (the absolute gaps are 0.20, 0.30, and 0.42), so we report a consistent advantage across lengths rather than a tested trend, and the sweep separates neither length from encoding relative to the learned-position headline nor a coupling effect from faster degradation of Differential Attention at longer lengths. All three rotary lengths are included in our Bonferroni family (Appendix B); at the corrected threshold the 1024 comparison survives while the 512 and 2048 comparisons do not, so we treat the latter two as descriptive, making no claim about lengths beyond 2048. The sweep also shows the composition functioning under rotary embeddings at 455M, complementing the standalone 60M check in the supplementary material.

**Capacity is not the explanation, and invertibility is what makes coupling reliable.** To test whether the standalone gain reflects added capacity alone, we compare the alternating coupling against pre-scoring controls that add parameters without its invertible, joint structure (Table 4, 60M, three seeds, same protocol). Independent nonlinear projections of $Q$ and $K$, which add nonlinear capacity without coupling them, do not improve on standard attention (24.13 against 24.20, $p$=0.22). A non-invertible cross MLP that mixes the two reaches 24.04 but carries more parameters than the coupling (+0.43% against +0.33%), and a linear mixing variant with five times the coupling's parameter budget (+1.74%) is the worst of the set (24.34). The alternating coupling (23.92) is significantly better than every one of these controls, including the higher-parameter cross MLP ($p$=0.035) and the highest-parameter linear mix ($p$<0.001). The gain tracks the invertible, joint structure rather than the parameter count: the variants with the most parameters are not the best, and because each composed variant is identical to its base method at initialization, the compositional gains are not a starting-point artifact either. At 455M, the scale of the central result, two further controls on top of Differential Attention complete this at-

Table 4: Capacity controls at 60M, three seeds, mean±std perplexity, with parameter overhead over standard attention. Adding parameters without the invertible, joint coupling does not reproduce the gain, and the variants with the most parameters are not the best. The alternating coupling is significantly better than every control (one-sided Welch, $p < 0.05$). Matched controls at 455M, including a non-invertible simultaneous variant of the coupling itself, are in the supplementary material.

| Pre-scoring variant | Params | 60M perplexity |
|---|---|---|
| Standard | +0.00% | $24.20 \pm 0.13$ |
| Independent nonlinear $Q, K$ | +0.11% | $24.13 \pm 0.03$ |
| Non-invertible cross MLP | +0.43% | $24.04 \pm 0.07$ |
| Linear mix | +1.74% | $24.34 \pm 0.05$ |
| Alternating coupling | +0.33% | $\mathbf{23.92 \pm 0.04}$ |

Table 5: FineWeb-Edu cross-corpus check at 60M. Per-seed and mean validation perplexity over two seeds; lower is better. Both seeds favor coupling in every row. With two seeds we report this as descriptive, with no significance test.

| Configuration | seed 1 | seed 2 | mean |
|---|---|---|---|
| Standard | 45.34 | 45.25 | 45.30 |
| Alternating coupling | 44.66 | 44.89 | **44.78** |
| Differential Attention | 45.31 | 44.94 | 45.12 |
| + coupling | 44.20 | 44.50 | **44.35** |

tribution (three seeds each, identical protocol; supplementary material). The non-invertible cross MLP reaches 18.89±0.11 against 18.89±0.04 for Differential Attention alone ($p$=0.51), reproducing none of the composition gain, so added pre-scoring capacity does not explain it. A simultaneous variant of the coupling itself, identical in parameterization but with the invertibility-guaranteeing sequential update replaced by a simultaneous one, retains the 455M gain ($18.78 \pm 0.07$) yet is dramatically worse than Differential Attention alone at 150M ($22.17 \pm 0.04$ against $21.10 \pm 0.08$, $p$=0.0001), below even standard attention, despite smooth and stable training. The gain therefore tracks the joint pre-scoring coupling rather than parameter count or invertibility alone, and the invertible sequential form is the only pre-scoring variant that improves its base method at every scale we test: invertibility is the property that makes the coupling reliable across scales.

**Generalization to a second corpus.** To test whether the effect is specific to WikiText-103, we repeat the 60M comparison on a FineWeb-Edu slice (Table 5). The standalone coupling lowers perplexity from 45.30 to 44.78 (two seeds, both favorable), and the composition with Differential Attention reaches 44.35, below Differential Attention alone (45.12) and standard attention (45.30). With two seeds and no significance test this second-corpus comparison is descriptive only.

**Zero-shot transfer.** We evaluate the 150M and 455M checkpoints zero-shot on LAMBADA last-token accuracy and HellaSwag length-normalized accuracy (Table 6). At 455M we evaluate every seed (six per side for the Differential pair, three for standard attention and the standalone coupling): the composition is best on both tasks, leading Differential Attention on LAMBADA (0.272 against 0.263, one-sided Welch $p$=0.04) and on HellaSwag (0.310 against 0.306, $p$=0.15), and leading standard attention on both ($p$=0.007 and $p$=0.04). The standalone coupling falls below standard attention on both tasks at 455M, consistent with its faded perplexity gain; under multi-seed evaluation, Differential Attention remains above standard attention on both tasks. At 150M the evaluations remain single-seed (seed 42),

Table 6: Zero-shot transfer at 150M (single seed, seed 42) and 455M (mean±std; six seeds per side for the Differential pair, three for standard and the standalone coupling); 2,000 examples per task. LAMBADA last-token accuracy and HellaSwag length-normalized accuracy; higher is better. A coupling variant is best in every column (bold). At 455M the composition is best on both tasks, and the standalone coupling falls below standard, consistent with its faded perplexity gain.

| | LAMBADA | | HellaSwag | |
|---|---|---|---|---|
| Configuration | 150M | 455M | 150M | 455M |
| Standard | 0.259 | $0.254 \pm 0.006$ | 0.313 | $0.303 \pm 0.004$ |
| Alternating coupling | 0.269 | $0.252 \pm 0.003$ | **0.317** | $0.297 \pm 0.006$ |
| Differential Attention | 0.274 | $0.263 \pm 0.007$ | 0.307 | $0.306 \pm 0.007$ |
| + coupling | **0.278** | $\mathbf{0.272 \pm 0.010}$ | 0.315 | $\mathbf{0.310 \pm 0.006}$ |

Table 7: Multi-query associative-recall accuracy, mean±std over three seeds, on a small model ($d$=256, $H$=4, $L$=6); higher is better. All variants reach 1.00 at two key-value pairs (row omitted). At four pairs the standalone coupling is higher and far more stable across seeds than standard attention; at eight pairs neither standalone head is reliable, and the coupling collapses at the densest cell. Composing either standalone head with Differential Attention restores recall to 1.00 on every cell (all $\geq 0.999$).

| KV pairs | Seq. len | Standard | Coupling | Diff. Attn | +Coupling |
|---|---|---|---|---|---|
| 4 | 64 | $0.98 \pm 0.02$ | $1.00 \pm 0.00$ | 1.00 | 1.00 |
| 4 | 128 | $0.94 \pm 0.05$ | $1.00 \pm 0.00$ | 1.00 | 1.00 |
| 4 | 256 | $0.92 \pm 0.14$ | $0.99 \pm 0.01$ | 1.00 | 1.00 |
| 8 | 64 | $0.55 \pm 0.40$ | $0.18 \pm 0.00$ | 1.00 | 1.00 |
| 8 | 128 | $0.79 \pm 0.10$ | $0.72 \pm 0.37$ | 1.00 | 1.00 |
| 8 | 256 | $0.73 \pm 0.34$ | $0.51 \pm 0.41$ | 1.00 | 1.00 |

so we report no significance there; the composition leads LAMBADA and the standalone coupling is highest on HellaSwag. These transfer accuracies are supporting evidence outside the Bonferroni-corrected perplexity family: a small perplexity gap on WikiText-103 need not align with zero-shot accuracy on a different distribution.

**Where coupling helps, and where it does not.** We use multi-query associative recall (Arora et al., 2024) to characterize the mechanism, with a small model ($d$=256, $H$=4, $L$=6), sweeping the number of key-value pairs against sequence length over three seeds (Table 7). At two key-value pairs every variant saturates. At four pairs the standalone coupling is both higher and far more stable across seeds than standard attention (0.99±0.01 against 0.92±0.14 at length 256), consistent with the reading that it forms a retrieval head from a broader, lower-variance prior. At eight pairs, the hardest retrieval load, neither standalone head is reliable: standard attention is high-variance (means 0.55 to 0.79 with seed spreads up to 0.40), and the coupling does not help on average and collapses at the densest setting, eight pairs at length 64, to 0.18±0.00 across all three seeds, which we read as its smoothing suppressing the single sharp pointer this regime needs. The composition resolves this: adding Differential Attention to either head restores recall to 1.00 across the entire grid, so the scoring-stage denoising supplies exactly the sharp selection that the standalone coupling smooths away, and the coupling adds no retrieval cost once that selection is in place. This is a single small synthetic setup and the mechanism in full language modeling may differ; we therefore base the compositional recommendation on the scale result, with which this characterization is consistent.

**Attention rank at depth.** We probe the 24-layer 455M models on validation sequences (full protocol in the supplementary material). The effective rank of the pre-softmax logit matrix is essentially identical across all variants (about 14.4) and is set by the causal mask, so the coupling does not change it. The effective rank of the post-softmax attention map, averaged over the late layers (L16 to L23), is modestly higher for the standalone coupling than for standard attention (53.5 against 46.2), while the Differential Attention pair differs little (45.3 against 42.5). We report this as a modest effect that is consistent with the small standalone gain, and we do not claim that coupling escapes a rank-collapse bound.

**Practical guidance.** The reading we take to practice is to use coupling compositionally rather than on its own. At the scales we test, a practitioner who already uses a scoring-stage method such as Differential Attention can add the alternating coupling for a small parameter overhead (0.33% at 60M, 0.13% at 455M) and obtain a small but consistent further reduction in perplexity that is significant at 455M, at a measured cost that shrinks at the scale where the result lives: about 8 to 10% training throughput over Differential Attention alone at 60M and sequence length 512, falling to 7.4% at 455M and to 4.1% at 455M with length 2048, with peak-memory overheads of about 17%, 23%, and 11% in the three settings, benchmarked on an H100 (supplementary material); whether that trade is worthwhile depends on the deployment, and a fused coupling kernel would narrow the latency gap toward the parameter overhead. Using coupling as a standalone replacement for standard attention is worthwhile mainly below the 150M range, where its standalone gain is significant; at 455M the standalone gain is not, so we do not recommend it there.

## 4. Related Work

**Pre-scoring transformations of queries and keys.** Several methods act on $Q$ and $K$ before the dot product. Query-key normalization rescales each vector to stabilize logits at scale (Henry et al., 2020), and the normalized Transformer places all representations on a hypersphere so that every logit is a cosine similarity (Loshchilov et al., 2024). These operations are per-vector and norm-only; they do not let the query and the key inform each other. Query-key normalization is moreover head-local, whereas the normalized Transformer's hypersphere constraint is global, which makes the latter a less separable target for a per-head coupling. Talking-Heads attention applies learned linear maps across heads both before and after the softmax (Shazeer et al., 2020), which mixes heads at the scoring stage rather than coupling the two vectors within a head. A separate family replaces the dot product itself: the Synthesizer uses learned or random score maps (Tay et al., 2021), and the Performer approximates the softmax kernel with random features (Choromanski et al., 2021); both discard the standard dot product that coupling leaves intact, so they are not pre-scoring transformations in our sense. Closest to a learned joint comparison, Elliptical Attention replaces the inner product with a learned positive-definite quadratic form, a shared bilinear metric between queries and keys (Nielsen et al., 2024); this is a global, input-independent learned metric on the $(q, k)$ pair, whereas coupling is a nonlinear, alternating, invertible map that reshapes the joint representation in an input-adaptive way rather than reweighting directions uniformly. Our coupling differs from all of these: it is a learned, joint, invertible map of $Q$ and $K$ within each head that keeps the exact softmax dot product, and it is the identity at initialization so it can only add to a standard head.

**Scoring-stage methods that coupling composes with.** A second family modifies the scoring stage after the product. Differential Attention subtracts two softmax maps to cancel common-mode noise (Ye et al., 2025), Multi-Token Attention convolves over the score map to share information across nearby positions (Golovneva et al., 2025), and gated attention reweights the attention output with a learned gate (Qiu et al., 2025). These methods are complementary to coupling rather than alternatives to it, and our central experiments place coupling on top of the first two.

**Concurrent work on a nonlinear query-key interaction.** Recent work also questions the linearity of the query-key map. One line replaces the linear query, key, or value projections with nonlinear ones (Zhao et al., 2026), and another adds a nonlinear residual to the query alone (Karbevski, 2026). MöbiusAttention introduces a nonlinear Möbius transform into the score computation (Halacheva et al., 2024); because it acts on the logit rather than on the projections, it is a scoring-stage method that coupling composes with in principle, not a pre-scoring transformation. Coupling is distinct on three counts that matter for our claims: it transforms $Q$ and $K$ jointly rather than independently, it is invertible by construction, and we show it composes additively with scoring-stage methods, a property these works do not study.

**Coupling flows and dynamical views of attention.** The alternating affine coupling is the construction used in real-valued non-volume-preserving flows (Dinh et al., 2017), and coupling flows of this kind are universal approximators of invertible smooth maps given sufficient depth (Teshima et al., 2020); we use this construction for its expressivity, not as a density model. A separate line views attention through dynamical systems and energy minimization, including modern Hopfield networks (Ramsauer et al., 2021) and analyses of attention as interacting particles (Geshkovski et al., 2024). This view motivated our starting point, an additive symplectic update; because a volume-preserving map of this kind is provably restricted in the family it can represent, we adopt the non-volume-preserving affine coupling, and the gain comes from that coupling rather than from any conservation law.

**Attention pathologies.** Pure self-attention loses rank with depth (Dong et al., 2021), and the attention matrix exhibits a widening spectral gap that is specific to the softmax (Nait Saada et al., 2024). We measure attention rank directly at depth and report a modest effect rather than a resolution of collapse (§3); we do not rest the method on this analysis.

## 5. Conclusion

We studied coupled query-key dynamics, a pre-scoring transformation that evolves queries and keys jointly through a shared invertible coupling before standard scoring, realized as an alternating affine coupling that is the identity at initialization. A controlled ablation isolates the alternating, non-volume-preserving form as the one that helps and separates it from added capacity; the additive, volume-preserving update we began with is provably more restricted, which is why we adopt the affine coupling. As a standalone change, coupling lowers perplexity at 60M and 150M with margins that pass significance testing ($p < 0.05$) but loses that margin by 455M. The behavior that holds at scale is compositional: placed on top of Differential Attention, coupling improves on it at both 150M and 455M, and the

same additive direction appears on top of Multi-Token Attention (directional) and query-key normalization (significant at 150M). We therefore present coupling not as a standalone competitor but as a composable pre-scoring transformation that adds to scoring-stage and projection-level methods; we use the word axis only in the empirical sense that the addition is consistent, not as a claim of structural independence.

**Limitations.** The standalone gain fades with scale, and our largest model is 455M; whether the compositional gain persists beyond this scale is open. Our scale sweep holds the token budget fixed at about 0.49B tokens, so the tokens-per-parameter ratio falls from about 8 at 60M to about 1 at 455M, with the largest model well below the compute-optimal ratio; the narrowing of the standalone gain is consistent with both a scaling effect and increasing undertraining, which our fixed-budget design cannot separate. The composition with Multi-Token Attention is directional rather than significant at both 150M and 455M. The compositional result is established for one scoring-stage method, Differential Attention, at the scale where the standalone gain has faded; query-key normalization is not a competitive baseline in our setup, so whether coupling is universally additive on scoring-stage methods or specific to Differential Attention is open, and the FineWeb-Edu check uses two seeds with no significance claim. On associative recall the standalone coupling does not improve the hardest retrieval load and collapses at the densest setting, since it smooths the sharp single-pointer head that the task needs; composing it with Differential Attention restores full recall there, consistent with the compositional reading. Most composition experiments use learned positional embeddings at sequence length 512; a rotary-embedding sweep at lengths 512, 1024, and 2048 shows the Differential Attention composition improving at every length, with a larger margin at the longer lengths that we report as a consistent pattern rather than a tested trend (§3). That sweep uses three seeds per length and does not separate length from encoding relative to the learned-position headline, and we have not tested lengths beyond 2048. Our experiments use one model family, and the rank analysis at depth shows only a modest effect rather than a resolution of rank collapse.

**Future work.** The natural next steps are to push the compositional comparison to larger scale, to tune each scoring-stage method and the coupling jointly, and to test the coupling across additional model families. More broadly, the results suggest that a small, invertible, identity-initialized transformation of queries and keys before scoring can add to methods that act after scoring, and that reporting where such a mechanism fails is as informative as reporting where it helps.

## References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *ICLR*, 2024.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *ICLR*, 2021.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. *arXiv preprint arXiv:2103.03404*, 2021.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv preprint arXiv:2312.10794*, 2024.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Multi-token attention. *arXiv preprint arXiv:2504.00927*, 2025.

Anna-Maria Halacheva, Mojtaba Nayyeri, and Steffen Staab. Expanding expressivity in transformer models with MöbiusAttention. *arXiv preprint arXiv:2409.12175*, 2024.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. *Findings of EMNLP*, 2020.

Marko Karbevski. Beyond linearity in attention projections: The case for nonlinear queries. *arXiv preprint arXiv:2603.13381*, 2026.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. nGPT: Normalized transformer with representation learning on the hypersphere. *arXiv preprint arXiv:2410.01131*, 2024.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2017.

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. *arXiv preprint arXiv:2410.07799*, 2024.

Stefan K. Nielsen, Laziz U. Abdullaev, Rachel S.Y. Teo, and Tan M. Nguyen. Elliptical attention. In *NeurIPS*, 2024.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*, 2025.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *ICLR*, 2021.

Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, and Le Hou. Talking-heads attention. *arXiv preprint arXiv:2003.02436*, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In *ICML*, 2021.

Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. In *NeurIPS*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. In *ICLR*, 2025.

Weijie Zhao, Mingquan Liu, Bolun Wang, Simo Wu, Nuobei Xie, Rui-Jie Zhu, and Peng Zhou. Nexusformer: Nonlinear attention expansion for stable and inheritable transformer scaling. *arXiv preprint arXiv:2604.19147*, 2026.

## Appendix A. Experimental Setup

Table 8: Model configurations used in experiments.

| **Config** | $d_{\text{model}}$ | $n_{\text{heads}}$ | $n_{\text{layers}}$ | $d_{\text{ff}}$ | **Params** |
|---|---|---|---|---|---|
| Tiny | 256 | 4 | 6 | 1024 | $\sim$6M$^{\dagger}$ |
| Small | 512 | 8 | 8 | 2048 | $\sim$60M |
| Medium | 768 | 12 | 12 | 3072 | $\sim$153M |
| Large | 1024 | 16 | 24 | 4096 | $\sim$455M |

**Model configurations.** All models share pre-norm RMSNorm, SwiGLU feed-forward layers, tied input/output embeddings, and learned positional embeddings.
†Parameter counts depend on vocabulary size: Tiny is used only for MQAR (vocab=64); the other counts are for WikiText-103 (vocab=50,257).

At the fixed WikiText-103 budget of 30K steps (about 0.49B tokens), the tokens-per-parameter ratio is about 8.2 (Small, 60M), 3.2 (Medium, 150M), and 1.1 (Large, 455M); the largest model is well below the compute-optimal ratio, so larger models in our sweep are progressively more undertrained, a confound flagged in the Limitations (§5).

**Training hyperparameters.** *MQAR experiments.* AdamW ($\beta_1$=0.9, $\beta_2$=0.95), lr=$3\times10^{-4}$, weight decay 0.01, gradient clipping 1.0, cosine schedule, 500-step warmup, 10K–20K total steps, batch size 64.

*WikiText-103 (small, 60M).* AdamW ($\beta_1$=0.9, $\beta_2$=0.95), lr=$6\times10^{-4}$, weight decay 0.1, gradient clipping 1.0, cosine schedule with 2K-step warmup, 30K total steps, effective batch size 32 (micro-batch 8, gradient accumulation 4), mixed precision (fp16), sequence length 512.

*WikiText-103 (medium, 150M).* Same as small except: lr=$3\times10^{-4}$, 30K total steps, gradient checkpointing enabled, effective batch size 32 (micro-batch 8, gradient accumulation 4; coupling variant: micro-batch 4, gradient accumulation 8).

*WikiText-103 (large, 455M).* Same as medium except: lr=$2\times10^{-4}$, effective batch size 32 (micro-batch 4, gradient accumulation 8; coupling variant: micro-batch 2, gradient accumulation 16). The rotary length sweep (§3) uses sequence lengths 512, 1024, and 2048 with rotary position embeddings, three seeds each (42, 137, 2024), the same optimizer and step count, and effective batch 32; at 2048 the micro-batch is halved and accumulation doubled to preserve the effective batch within memory.

**Variant-specific hyperparameters.** Table 9 lists the hyperparameters specific to each attention variant, including the complete specification of the coupling; the coupling introduces no other hyperparameters and leaves those of the base methods unchanged.

Table 9: Variant-specific hyperparameters. The coupling rows fully specify the alternating affine coupling of Eqs. (2)–(3).

| Variant | Hyperparameter | Value |
|---|---|---|
| Alternating coupling | Coupling rounds | 1 (Eqs. (2)–(3)) |
| | Conditioning nets $s, t$ | two-layer, SiLU, output width $2d_k$ |
| | Net placement | one $(s, t)$ pair per update per layer, per head |
| | Output-layer init | zero (exact identity at initialization) |
| | Log-scale clamp | $[-3, 3]$ |
| GQA | KV head ratio | 1/4 of query heads |
| | Group size | 4 queries per KV head |
| Differential Attention | Sub-head dimension $d_s$ | $d_k/2$ |
| | $\lambda$ init | $0.8 - 0.6\exp(-0.3\ell)$ (layer $\ell$) |

**Software and hardware.** All reported results were produced on NVIDIA H100 GPUs with PyTorch 2.10 and mixed-precision (fp16) training. All attention variants are drop-in replacements behind a shared interface, so every comparison uses identical training infrastructure and differs only in the attention head. All models use learned positional embed-

dings by default; the RoPE ablation (supplementary material) and the 455M longer-context comparison (§3) replace these with Rotary Position Embeddings (Su et al., 2024).

## Appendix B. Per-Seed Statistical Tests

All perplexity values in Table 10 are the exponential of the best validation loss; significance uses a one-sided Welch's $t$-test with the alternative that the treatment perplexity is lower, and the $n$ column gives seeds per comparison.

Table 10: Per-comparison statistics for all main claims. $\Delta$ = treatment PPL − baseline PPL (negative = improvement). Significance: $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

| Comparison | Scale | $n$ | Baseline mean±std | Treatment mean±std | $\Delta$ | $p$ | $d$ | Note |
|---|---|---|---|---|---|---|---|---|
| 1. Standalone coupling (vs. standard) | 60M | 3 | $24.20 \pm 0.13$ | $23.92 \pm 0.04$ | −0.28 | $0.025^{*}$ | 3.09 | |
| | 150M | 3 | $21.61 \pm 0.03$ | $21.45 \pm 0.06$ | −0.16 | $0.015^{*}$ | 3.39 | |
| | 455M | 3 | $19.56 \pm 0.10$ | $19.48 \pm 0.07$ | −0.07 | 0.19 | 0.83 | not significant |
| 2. Composition: Differential Attention | 150M | 3 | $21.10 \pm 0.08$ | $20.92 \pm 0.02$ | −0.17 | $0.024^{*}$ | 3.12 | |
| | 455M | 6 | $18.89 \pm 0.04$ | $18.79 \pm 0.05$ | −0.10 | $0.003^{**}$ | 2.09 | |
| | 455M | 3 | $19.50 \pm 0.03$ | $19.20 \pm 0.05$ | −0.30 | $0.0015^{**}$ | 6.95 | seq 1024, RoPE |
| 3. Composition: Multi-Token Attn. | 150M | 3 | $20.96 \pm 0.05$ | $20.85 \pm 0.12$ | −0.11 | 0.128 | 1.18 | not significant |
| | 455M | 3 | $19.24 \pm 0.05$ | $19.20 \pm 0.03$ | −0.04 | 0.16 | 0.92 | not significant |
| 4. Composition: QK-Norm | 60M | 3 | $24.24 \pm 0.00$ | $24.12 \pm 0.03$ | −0.13 | $0.010^{**}$ | 5.53 | |
| | 150M | 3 | $22.25 \pm 0.05$ | $21.96 \pm 0.05$ | −0.29 | $0.001^{**}$ | 5.53 | |
| | 455M | 3 | $19.82 \pm 0.02$ | $19.73 \pm 0.08$ | −0.09 | $0.08^{\dagger}$ | 1.68 | not significant |
| 5. Ablation ladder | 60M (std→coup) | 3 | $24.20 \pm 0.13$ | $24.04 \pm 0.02$ | −0.16 | $0.077^{\dagger}$ | 1.79 | marginal |
| | 60M (coup→alt) | 3 | $24.04 \pm 0.02$ | $23.92 \pm 0.04$ | −0.12 | $0.006^{**}$ | 4.35 | |
| | 150M (std→coup) | 3 | $21.61 \pm 0.03$ | $21.53 \pm 0.05$ | −0.08 | $0.048^{*}$ | 1.92 | |
| | 150M (coup→alt) | 3 | $21.53 \pm 0.05$ | $21.45 \pm 0.06$ | −0.09 | $0.068^{\dagger}$ | 1.54 | marginal |

We use one-sided tests because the directional hypothesis that coupling lowers perplexity was established from the 60M standalone results before the larger runs were launched, and from the composition direction before the composition runs; we report the non-significant outcomes without changing the test direction post hoc. The two-sided equivalents double the reported $p$-values; the Bonferroni analysis below uses the one-sided values. Even two-sided, the headline 455M Differential composition remains significant ($p \approx 0.005$). All $p$-values and effect sizes are computed from the unrounded per-seed perplexities; for the headline 455M Differential comparison the six per-seed perplexities are 18.9054, 18.9325, 18.9110, 18.8679, 18.8176, 18.8812 for Differential Attention and 18.7864, 18.8188, 18.7673, 18.7224, 18.7666, 18.8728 for the composition, which reproduce one-sided $p$=0.0026 (rounded to 0.003 in Table 10) and $d$=2.1. Per-seed perplexities for the remaining comparisons (the two 455M breadth pairs and the three rotary lengths) are in the supplementary material; for three-seed rows the Cohen's $d$ magnitudes indicate direction rather than stable estimates. At length 2048 both models overfit late in training, so that comparison uses each model's best validation checkpoint, selected independently. Applying a Bonferroni correction across the seventeen comparisons in our family, the fifteen in this table plus the 512 and 2048 rotary lengths of the length sweep, sets $\alpha$=0.05/17≈0.0029. Three comparisons remain significant at that threshold: the 455M Differential composition at length 512 ($p$=0.0026), the same composition at length 1024 with rotary embeddings ($p$=0.0015), and the 150M query-key normalization composition ($p$=0.001). All others do not, including the 512 and 2048 rotary compositions ($p$=0.0035 and $p$=0.016), the 60M query-key normalization composition ($p$=0.010), and the coupling-to-alternating ablation ($p$=0.006).

# Supplementary Material for "Invertible Query-Key Coupling Composes with Attention Mechanisms"

**Barak Gahtan** BARAKGAHTAN@CS.TECHNION.AC.IL
*Technion – Israel Institute of Technology, Haifa, Israel*

**Alex M. Bronstein** BRON@CS.TECHNION.AC.IL
*Technion – Israel Institute of Technology, Haifa, Israel*
*Institute of Science and Technology Austria, Klosterneuburg, Austria*

**Editors:** Andy Song, Bo Han and Sarah Erfani

## Appendix A. Invertibility Controls at Scale and Training-Budget Analysis

This section reports the two controls and the training-budget analysis added in response to reviewer requests during the ACML 2026 discussion period. The comparisons here are post-hoc additions and therefore lie outside the pre-registered Bonferroni family of the main paper; we report them with per-seed values and treat them as attribution evidence for the composition result rather than as additional headline claims.

**Two matched controls on top of Differential Attention.** The capacity table in the main paper compares pre-scoring controls at 60M. Two questions remain open there: whether extra pre-scoring capacity could reproduce the 455M composition gain, and whether invertibility itself, rather than the bidirectional affine interaction, is the operative property. We answer both with controls trained on top of Differential Attention under the identical protocol (data, steps, optimizer, sequence length 512, three seeds).

The first control is the non-invertible cross-MLP of the capacity table: a joint two-layer MLP over the stacked $(q, k)$ pair with a learned step size, zero-initialized so it is the exact identity at initialization, and with slightly more parameters than the coupling.

The second control is the minimal-difference variant of the coupling itself. The alternating coupling of the main paper is sequential: the query first conditions the key update, and the *updated* key then conditions the query update, which yields the triangular structure that guarantees invertibility. The simultaneous control keeps the conditioning networks, the parameter count (equal to the last parameter), the clamping, and the identity-at-init guarantee of the coupling, and changes only the update order: both vectors are updated from the pre-update partner,

$$k' = k \odot \exp\big(s_\theta(q)\big) + t_\theta(q), \tag{1}$$

$$q' = q \odot \exp\big(s_\phi(k)\big) + t_\phi(k), \tag{2}$$

where $s_\phi, t_\phi$ read the original $k$ rather than $k'$. This removes the invertibility guarantee, since recovering $(q, k)$ from $(q', k')$ requires solving a coupled nonlinear system with no bijectivity guarantee, while preserving the bidirectional interaction and the affine scaling.

© 2026 B. Gahtan & A.M. Bronstein.

Table 1: Invertibility and capacity controls on top of Differential Attention, best validation perplexity, mean±std. One-sided Welch $p$ is against Differential Attention alone at the same scale (six seeds at 455M, three at 150M); for the two controls that are *worse* than the base, the reported $p$ is for the reversed one-sided test. All variants are exactly their base method at initialization.

| Scale | Pre-scoring variant on Diff. Attn. | Perplexity | $\Delta$ vs. base | $p$ |
|---|---|---|---|---|
| 455M | none (Differential Attention, 6 seeds) | $18.89 \pm 0.04$ | | |
| 455M | non-invertible cross-MLP (3 seeds) | $18.89 \pm 0.11$ | $+0.002$ | 0.51 |
| 455M | simultaneous coupling (3 seeds) | $18.78 \pm 0.07$ | $-0.110$ | 0.042 |
| 455M | invertible alternating coupling (6 seeds) | $\mathbf{18.79 \pm 0.05}$ | $-0.097$ | 0.003 |
| 150M | none (Differential Attention, 3 seeds) | $21.10 \pm 0.08$ | | |
| 150M | simultaneous coupling (3 seeds) | $22.17 \pm 0.04$ | $+1.078$ | 0.0001 (worse) |
| 150M | invertible alternating coupling (3 seeds) | $\mathbf{20.92 \pm 0.02}$ | $-0.175$ | 0.024 |

**Result: a dissociation.** Table 1 gives the two controls' outcomes. The cross-MLP reproduces none of the 455M gain: it is statistically indistinguishable from Differential Attention alone ($p$=0.51), so added pre-scoring capacity does not explain the composition result. The simultaneous coupling dissociates the two remaining factors. At 455M it reproduces the gain (better than the base, $p$=0.042, and statistically indistinguishable from the invertible composition), so invertibility per se is not the sole cause of the gain at that scale; the joint bidirectional coupling is. At 150M, however, the same simultaneous variant is dramatically worse than Differential Attention alone, and worse than standard attention, despite smooth and stable training at every seed: the failure is a bad optimum, not divergence. The invertible sequential form is the only pre-scoring variant that improves its base method at every scale we test. We therefore read invertibility as the design property that makes bidirectional query-key coupling consistently safe and beneficial across scales, rather than as the isolated cause of the 455M gain, and the main text states the claim in this form. Per-seed best validation perplexities: simultaneous coupling 18.7113/18.7737/18.8439 at 455M and 22.1307/22.1919/22.1981 at 150M; cross-MLP 18.9875/18.7631/18.9126 at 455M.

**Training-budget analysis.** The models in this study are undertrained relative to compute-optimal token budgets (about 1.1 tokens per parameter at 455M), which raises the concern that the composition gain reflects optimization speed in an undertrained regime rather than a persistent advantage. An optimization-speed advantage predicts a gap that closes as training progresses. We test this on the existing training runs by computing, at every evaluation point (every 1,000 steps), the best-so-far validation perplexity of each variant, i.e., the value the paper's best-checkpoint comparison would report if training had stopped at that budget, and tracking the relative gap between the composition and Differential Attention alone (three seeds per side).

Table 2 shows the opposite of the optimization-speed prediction for the composition: the relative gap reaches $-0.66\%$ by 15K steps and stays essentially constant through 29K steps. The same diagnostic applied to the standalone coupling shows the gap closing in late training, consistent with the faded standalone effect at 455M, so the analysis distinguishes

Table 2: Best-so-far relative validation-perplexity gap (mean over three seeds; negative favors the coupling variant) as a function of training budget at 455M. The composition's advantage is established early and remains stable to slightly widening through the end of training (late-half trend −0.02% per 1,000 steps), while the standalone gap closes late (+0.02% per 1,000 steps), matching the faded standalone effect reported in the main paper.

| 455M comparison | 8K steps | 15K steps | 22K steps | 29K steps |
|---|---|---|---|---|
| Composition vs. Differential Attention | −0.37% | −0.66% | −0.65% | −0.65% |
| Standalone coupling vs. standard | −0.31% | −0.73% | −0.79% | −0.34% |

a persistent advantage from a transient one rather than being insensitive to the difference. This does not substitute for a compute-optimal 455M run, which is beyond our budget; it tests, and rejects, the specific prediction that the composition advantage shrinks with more training.

## Appendix B. Overhead Benchmarks

**Throughput.** Each coupling round applies a two-layer per-head MLP of width $d_k$; across all heads this adds on the order of $d\,d_k$ multiply-accumulates per token per layer and does not change the asymptotic $\mathcal{O}(n^2 d_k)$ attention cost. The parameter overhead is about 0.3% at 60M (Table 3). Table 4 reports the measured cost at 60M on a single H100 (batch 8, forward and backward, tokens per second relative to standard attention). Because the per-head coupling MLPs are not fused, the realized overhead exceeds the parameter overhead: the standalone coupling runs at about 90% of standard throughput at sequence length 512 and about 85% at length 2048, and the Differential Attention composition runs at about 92% and 90% of Differential Attention alone at the two lengths. Peak memory rises by about 22% for the standalone coupling and about 17% for the composition over its base, both at length 512. At length 2048 Differential Attention is itself the costlier component (about 0.68 relative throughput and 1.76× peak memory against standard), so the coupling is not the dominant cost in the composition. A fused kernel would narrow the gap toward the parameter overhead, and implementing one is left to future work.

Table 3: Parameter comparison at 60M scale. The alternating affine coupling variant uses per-head coupling MLPs applied in an alternating pattern, adding approximately 0.3% parameters over standard attention.

| **Attention** | **Total Params** | **Variant-Specific** | **Overhead** |
|---|---|---|---|
| Standard | 60,343,296 | 0 | n/a |
| Affine coupling, one-way | 60,408,896 | 65,600 | +0.11% |
| Affine coupling, alternating | 60,541,952 | 198,656 | +0.33% |
| GQA | 57,197,568 | 0 | −5.21% |
| Diff Attention | 60,343,360 | 64 | +0.00% |

The one-way row is the intermediate, non-alternating design of the ablation ladder in the main paper; the primary method is the alternating variant, which applies the coupling MLP per head and contributes the +0.33% overhead reported in the caption and in the main text. GQA is included as a parameter-budget reference, not a quality baseline (see the main paper).

Table 4: Measured wall-clock cost at 60M on one H100, batch 8, forward and backward. Throughput is tokens per second relative to standard attention (higher is better); peak memory is at sequence length 512. Because the per-head coupling MLPs are unfused, the realized cost exceeds the 0.33% parameter overhead. At length 2048 Differential Attention is itself the costlier component.

| **Variant** | **Params** | **Rel. tok/s (512)** | **Rel. tok/s (2048)** | **Peak mem (MB)** |
|---|---|---|---|---|
| Standard | +0.00% | 1.00 | 1.00 | 4175 |
| Differential Attention | +0.00% | 0.86 | 0.68 | 5199 |
| Alternating coupling | +0.33% | 0.90 | 0.85 | 5071 |
| + Diff. Attention | +0.33% | 0.79 | 0.61 | 6095 |

**Cost at the scale of the main result.** Table 5 repeats the measurement at 455M, the scale of the central composition claim, at sequence lengths 512 (batch 8) and 2048 (batch 2), with batch fixed within each length so relative overheads are directly comparable across variants. The parameter overhead of the coupling falls to 0.13% at 455M, and the relative training-throughput overhead of the composition over Differential Attention alone falls with both scale and length: about 7.4% at length 512 and 4.1% at length 2048, against 8 to 10% at 60M. The coupling's cost is linear in sequence length and model width while attention scoring is quadratic in length, so the coupling's share of total compute shrinks exactly where the main result lives. Peak-memory overhead follows the same logic: it comes from storing the coupling activations, is largest in relative terms at the short length (about +23% over Differential Attention at 512), and falls to about +11% at length 2048, where attention's own sequence-quadratic buffers dominate.

Table 5: Measured wall-clock cost at 455M on one H100, forward and backward, batch 8 at sequence length 512 and batch 2 at length 2048. Throughput is tokens per second relative to standard attention at the same length; peak memory is reported at both lengths. The composition's overhead over Differential Attention alone is 7.4% at length 512 and 4.1% at length 2048.

| **Variant** | **Params** | **Rel. tok/s (512)** | **Rel. tok/s (2048)** | **Mem 512 (GB)** | **Mem 2048 (GB)** |
|---|---|---|---|---|---|
| Standard | +0.00% | 1.00 | 1.00 | 16.7 | 26.4 |
| Differential Attention | +0.00% | 0.87 | 0.72 | 22.8 | 51.3 |
| Alternating coupling | +0.13% | 0.91 | 0.94 | 22.1 | 31.8 |
| + Diff. Attention | +0.13% | 0.81 | 0.69 | 28.2 | 56.7 |

## Appendix C. Per-Seed Perplexities and Effect Sizes

This section lists the per-seed perplexities behind the per-comparison statistics in the main paper, for the comparisons whose six-seed headline disclosure is not already given there.

**455M breadth comparisons (three seeds; seeds 42, 137, 2024).** The per-seed perplexities are 19.2541, 19.2791, 19.1853 for Multi-Token Attention and 19.1607, 19.2168, 19.2244 for its composition ($p$=0.16), and 19.8369, 19.8309, 19.8026 for query-key normalization and 19.6468, 19.7372, 19.8017 for its composition ($p$=0.08); each composition has the lower mean but with overlapping seeds, consistent with non-significance at three seeds.

**Rotary length sweep (three seeds each).** For the sequence-length-1024 rotary comparison the per-seed values are $19.4719, 19.5286, 19.4854$ for Differential Attention and $19.1643, 19.2573, 19.1656$ for the composition, which separate without overlap. The two additional rotary lengths of the sweep, counted in the Bonferroni family of the main paper but not surviving the correction, have the following per-seed values: at 512, 18.8498, 18.8540, 18.8541 for Differential Attention and 18.6396, 18.6897, 18.6376 for the composition (one-sided Welch $p$=0.0035), and at 2048, 20.4355, 20.1022, 20.2140 and 19.8902, 19.9353, 19.6748 ($p$=0.016); both lengths separate without overlap. With three seeds these $p$-values and their effect sizes are unstable, and the 512 Differential Attention seeds cluster unusually tightly, so we read the sweep as directional confirmation rather than as precise estimates.

**Effect sizes.** Cohen's $d$ is reported in the main paper for completeness; with three seeds its confidence interval is wide, so for the three-seed rows the magnitudes indicate direction and separation rather than a stable effect-size estimate. The headline 455M Differential comparison uses six seeds and gives $d$=2.1; with only three seeds this estimate is inflated by an unusually tight Differential Attention spread, so we report the more reliable six-seed value.

## Appendix D. Held-Out Test-Set Perplexity

The validation-selected 455M checkpoints, evaluated on the WikiText-103 test split (token-weighted over the full split, same tokenizer), reproduce the validation comparisons in ordering and significance (Table 6): the Differential composition is significant at sequence length 512 (20.88 against 21.02, six seeds, $p$=0.002) and at sequence length 1024 with rotary embeddings (19.81 against 20.13, three seeds, $p$=0.003), and the standalone comparison remains not significant (21.76 against 21.79, $p$=0.21). We treat these as replications of the corresponding validation tests on a held-out split, not as additional members of the Bonferroni family. A per-position profile of the test loss shows the composition's gain over Differential Attention is positive in all twenty-four 64-token bins across both sequence lengths, rather than concentrated at particular positions.

## Appendix E. Additional Analysis

**Attention matrix rank.** We probe post-softmax and pre-softmax effective rank in the 455M (24-layer) models, seed 42, averaged over validation sequences. Effective rank is

Table 6: WikiText-103 test-split perplexity of the validation-selected 455M checkpoints, mean±std. Ordering and significance match the validation comparisons of the main paper's composition table, including the non-significant standalone gap.

| **Comparison** | $n$ | **Baseline** | **Treatment** | $p$ |
|---|---|---|---|---|
| Comp(Diff), seq 512 | 6 | $21.02 \pm 0.04$ | $\mathbf{20.88 \pm 0.07}$ | 0.002 |
| Comp(Diff), seq 1024, RoPE | 3 | $20.13 \pm 0.08$ | $\mathbf{19.81 \pm 0.06}$ | 0.003 |
| Standalone, seq 512 | 3 | $21.79 \pm 0.05$ | $21.76 \pm 0.02$ | 0.21 |

$(\sum_i \sigma_i)^2 / \sum_i \sigma_i^2$ where $\sigma_i$ are the singular values of the attention map $A = \text{softmax}(QK^\top/\sqrt{d_k})$ (post-softmax) or of $QK^\top/\sqrt{d_k}$ (pre-softmax).

**Pre-softmax logit rank.** The effective rank of the pre-softmax logit matrix is essentially identical across all variants (approximately 14.4 for all four), because the causal mask dominates the singular-value spectrum. Coupling does not change it.

**Post-softmax rank in late layers.** Table 7 reports mean effective rank over the last third of layers (L16–L23) of the 455M model.

Table 7: Post-softmax attention-map effective rank, mean over L16–L23 of the 455M (24-layer) model (seed 42, WikiText-103). Pre-softmax logit rank is $\approx 14.4$ for all variants (causal-mask dominated).

| **Variant** | **Post-softmax eff. rank (L16–L23 mean)** |
|---|---|
| Standard | 46.2 |
| Alternating coupling | 53.5 |
| Differential Attention | 42.5 |
| + coupling (composition) | 45.3 |

The standalone alternating coupling preserves modestly more post-softmax rank in late layers relative to standard attention (53.5 vs. 46.2). Differential Attention alone is *lower* (42.5), and the composition partially recovers it (45.3). We interpret this as a modest effect: coupling shifts the rank distribution in late layers but does not escape a rank-collapse bound. In particular, the pre-softmax picture is the same for all variants, confirming that the coupling's effect is downstream of the scoring step.

**RoPE compatibility.** We verify that the alternating affine coupling composes with Rotary Position Embeddings (RoPE). Table 8 shows 60M results for the alternating affine coupling (the primary result in this paper) and standard attention, each trained with learned positional embeddings and with RoPE. Both variants see higher perplexity under RoPE at 60M scale; the coupling advantage is preserved under both positional encoding choices.
The table reports both positional-encoding conditions (learned and RoPE) for standard attention and for the alternating coupling. The comparison isolates the effect of replacing learned positional embeddings with RoPE; the relative ordering, with the coupling advantage preserved under both choices, is the claim, not the absolute perplexities. Standard attention degrades more under RoPE at this scale (24.20 to 25.84) than the coupling does (23.92 to 24.00). We report this asymmetry without interpreting it: the setup is not

Table 8: RoPE ablation at 60M scale (alternating affine coupling). Mean ± std over 3 seeds.

| **Attention** | **Learned Pos.** | **RoPE** |
|---|---|---|
| Standard | 24.20 ± 0.13 | 25.84 ± 0.07 |
| Alternating coupling | 23.92 ± 0.04 | 24.00 ± 0.03 |

tuned for RoPE, and a single 60M comparison cannot separate a genuine interaction from a configuration artifact. The 455M longer-context runs in the main paper also use rotary embeddings, at sequence lengths 512, 1024, and 2048; the composition gain across that sweep extends RoPE compatibility from this standalone 60M check to the Differential Attention composition at scale.

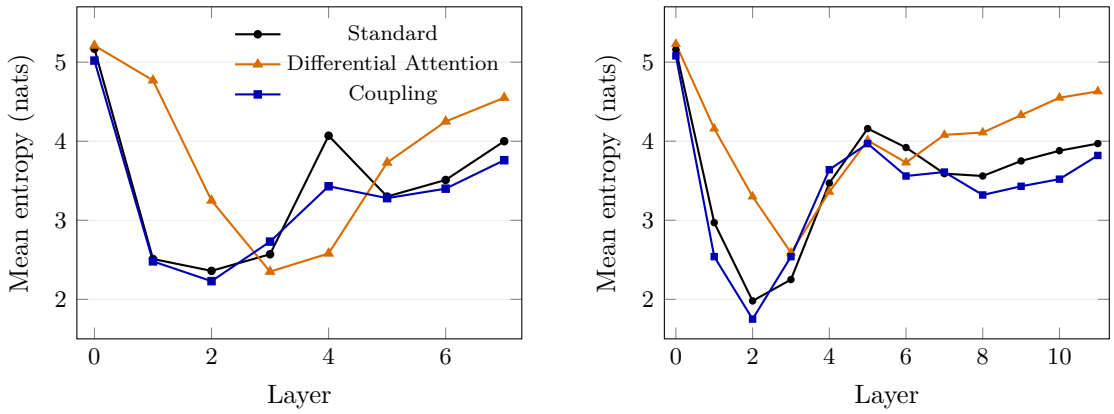
## Appendix F. Optimization Stability

**Cross-seed variance.** Beyond mean perplexity, the alternating coupling produces the tightest seed spread of any variant at 60M: its standard deviation over three seeds is 0.04 PPL against 0.13 for standard attention (the statistical tests in the main paper), a threefold reduction, and the Differential composition is likewise tight (0.02 against 0.08 for Differential Attention at 150M). This tightening is not uniform: at 150M the standalone spread is no smaller than standard, so we report the variance reduction as an observation at specific scales rather than a guaranteed effect. Three-seed spreads are themselves noisy estimates of variance: at 455M, a Differential Attention spread that three seeds suggested was unusually tight widened once six seeds were available (main paper), and we therefore do not base any interpretation on seed spread alone. Per-layer attention entropy profiles (Figure 1) provide descriptive context and show no entropy effect that would explain the variance reduction.

**Interpretation.** Where the spread tightens, it is consistent with the coupling guiding optimization toward a flatter basin, in which a smaller top Hessian eigenvalue would make the minimum less sensitive to initialization. We report the variance numbers as the direct measurement we have; a Hessian-eigenvalue probe, using Pearlmutter products with power iteration, is a natural confirmation we leave to future work. We do not claim a flatness result beyond the variance observations.

## Appendix G. Attention Entropy Profiles

Attention entropy $S = -\sum_j a_j \log a_j$ measures distributional spread. Entropy collapse, in which heads concentrate on near-deterministic distributions, is associated with reduced representational capacity (Zhai et al., 2023). Figure 1 compares per-layer entropy across variants.




(a) 60M (b) 150M

Figure 1: Per-layer mean attention entropy at the end of training (seed 42, mean over heads). All variants dip at the early attention-sink layer (Layer 2). The alternating coupling tracks standard attention closely and is marginally sharper (lower entropy) at the sink layer at both sizes, while Differential Attention holds higher entropy through the middle and late layers. We show these profiles as descriptive context; the stability evidence we rely on is the cross-seed variance reported in the main paper, not an entropy effect.

## Appendix H. Physics Motivation: From Hamiltonian Mechanics to Coupled Dynamics

Coupled query-key dynamics was originally inspired by Hamiltonian mechanics; we record the path briefly because it explains the term *dynamics* and why the final method is an affine coupling rather than a symplectic integrator. We treated queries $Q$ as generalized position and keys $K$ as generalized momentum with separable energy $H(q,k) = \frac{1}{2}\|k\|^2 + V(q)$, so that Hamilton's equations

$$\dot{q} = \frac{\partial H}{\partial k} = k, \qquad \dot{k} = -\frac{\partial H}{\partial q} = -\nabla_q V(q) \tag{3}$$

make $Q$'s update depend on $K$ and $K$'s update depend on a learned function of $Q$, the indirect coupling described in the main paper. A leapfrog (Störmer–Verlet) discretization of this flow is a composition of unit-Jacobian shear maps, so by Liouville's theorem it preserves phase-space volume (Arnold, 1989); in our prototypes the force was amortized by a learned MLP in the style of Hamiltonian neural networks (Greydanus et al., 2019). That volume preservation is exactly the property the final method gives up: a purely additive, volume-preserving coupling is provably restricted, whereas the affine, non-volume-preserving, alternating coupling described in the main paper removes that restriction (Teshima et al., 2020), and the ablation in the main paper, not this analogy, is what supports the design.

## References

Vladimir I Arnold. *Mathematical Methods of Classical Mechanics*. Springer, 1989.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *NeurIPS*, 2019.

Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. In *NeurIPS*, 2020.

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 *ICML*, 2023.