# 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.

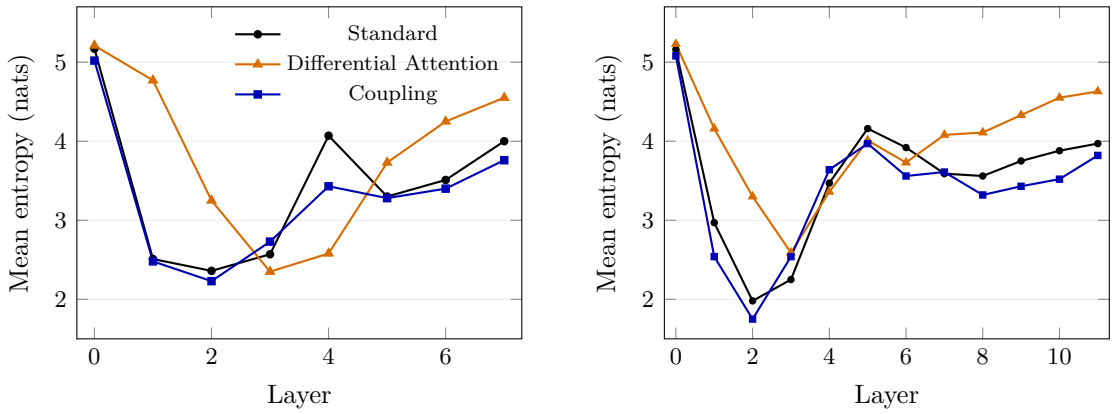
## Appendix F. Optimization Stability

**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.