When Can Attention Heads Be Statically Defined?

Selective Attention Freezing (SAF): fixed attention weights, live value mixing

Paper authors: Weixian Waylon Li, Yintao Tai, Marcio Fonseca, Shay B. Cohen

arXiv:2609.34650v1 10 November 2026

Publication not found as of 2026-11-10 | Official code | Versioned paper

A subset of heads may repeat the same positional pattern

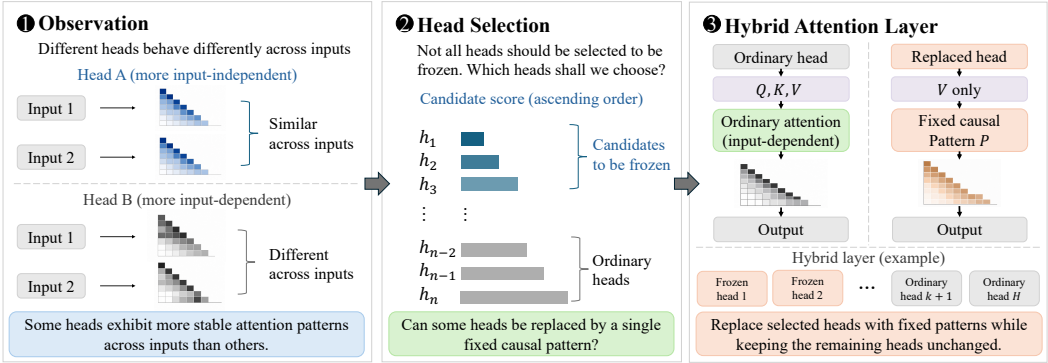
- Ordinary attention recomputes input-dependent $QK^\top$ scores and softmax for every head and sequence.
- Some heads' attention maps change little across different inputs.
- If a map is nearly content-free, repeating its score and softmax work may be unnecessary.

Research question

Which heads can use a fixed causal pattern, and can a mixed layer turn that change into real training and inference savings?

Paper, Sec. 1, PDF pp. 1–2.

SAF freezes the weights, not the token mixing



- Ordinary heads retain Q, K, V and input-dependent attention.
- Replaced heads use a fixed causal P but still mix the current input's V vectors.
- Value and output projections keep training.

Paper, Fig. 1, PDF p. 2; Sec. 3, PDF p. 3.

One calibration event defines the frozen heads

1. Train with ordinary attention to a chosen checkpoint, usually the midpoint.
2. Run a small calibration set without updating weights or advancing the training sampler.
3. Rank heads by attention-map variance; propose the lowest-variance heads.
4. Fit each mean map to a compact positional pattern and reject fits above 0.2 nats per row.
5. Install the fixed patterns and resume training with the same model and optimizer state.

Paper, Sec. 3.1–3.3 and Algorithm 1, PDF pp. 3–5, 16.

Variance ranks how well one fixed map reconstructs a head

For head (ℓ, h), calibration estimates

$$\widehat{\mathbf{A}}_{\ell h} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{A}_{\ell h}^{(n)}, \qquad s_{\ell h}^{\mathrm{var}} = \frac{1}{(N-1)\,T^2}\sum_{n=1}^{N} \left\| \mathbf{A}_{\ell h}^{(n)} - \widehat{\mathbf{A}}_{\ell h} \right\|_F^2 .$$

- Choose the $k = \mathrm{round}(rLH)$ lowest-scoring heads across layers.
- This measures attention-map variation, not whether the head is functionally disposable.
- At 124M and 4K, 30–31 of 36 quarter-training selections remain selected at the midpoint.

Paper, Eq. (1), Sec. 3.1, PDF p. 3; Appendix B.2, PDF p. 20.

The post-softmax mean is the best fixed map for squared error

$$\mathcal{R}(\mathbf{P}) = \frac{1}{T^2}\mathbb{E}\|\mathbf{A} - \mathbf{P}\|_F^2 = \mathcal{R}(\mathbb{E}[\mathbf{A}]) + \frac{1}{T^2}\|\mathbf{P} - \mathbb{E}[\mathbf{A}]\|_F^2.$$

- The probability mean is causal and row-normalised, so it is a valid pattern.
- Attention variance estimates the minimum reconstruction error; this links the choice of pattern and selector.
- This argument concerns attention reconstruction, not language-model loss.
- Empirically, the mean is best or tied at low replacement rates; sampled Gaussian and Dirichlet patterns do not improve on it.

Paper, Appendix B.1, PDF p. 18; Fig. 3, PDF p. 6.

Two length-T vectors encode a causal pattern

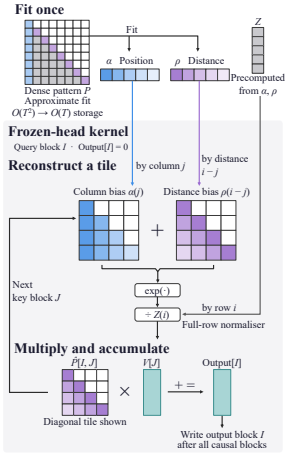
Let $\alpha(j)$ encode absolute key position and $\rho(i-j)$ encode relative distance:

$$\widehat{P}(i,j) = \begin{cases} \exp(\alpha(j) + \rho(i-j))/Z(i), & j \le i, \\ 0, & j > i, \end{cases} \qquad Z(i) = \sum_{k=1}^{i} \exp(\alpha(k) + \rho(i-k)).$$

- $Z(i)$ normalises each causal query row.
- Fit by row-averaged cross-entropy to the empirical mean pattern.
- Storage falls from $O(T^2)$ to $O(T)$ per replaced head.

Paper, Eqs. (3)–(4), PDF p. 4; fit tolerance and settings, Appendix A.2, p. 16.

The fused kernel reconstructs tiles and keeps head order



- One layer launch handles ordinary and fixed-pattern heads.
- Ordinary heads use FlashAttention-2.
- Fixed heads reconstruct causal tiles in registers, then multiply by live V tiles.
- Training skips replaced-head Q, K projections, score computation and softmax.

Paper, Fig. 2 and Sec. 3.3, PDF pp. 5, 36–37.

The evaluation separates model quality from kernel speed

Study	Model and training	Main question
Controlled pretraining	124M; 12 layers, 12 heads; FineWeb-Edu; 2.4576B tokens	Pattern, selector, rate, schedule, pruning controls
Longer context	124M at 8K and 16K	PPL and systems behavior across context
Scale-up	1B; 8K; 19.667B tokens; four GH200 GPUs	PPL and distributed update time
Pretrained transfer	Qwen3-4B; no additional training	Zero-shot tasks and GQA prefill

- Calibration is separated from final PPL reporting.
- Main speedups are paired per-update measurements; 124M quality and runtime use separate runs.

Paper, Sec. 4, PDF pp. 5–6; Appendix C, PDF pp. 22–35.

The mean pattern edges out alternatives at low replacement rates

PPL increase over ordinary attention (%)

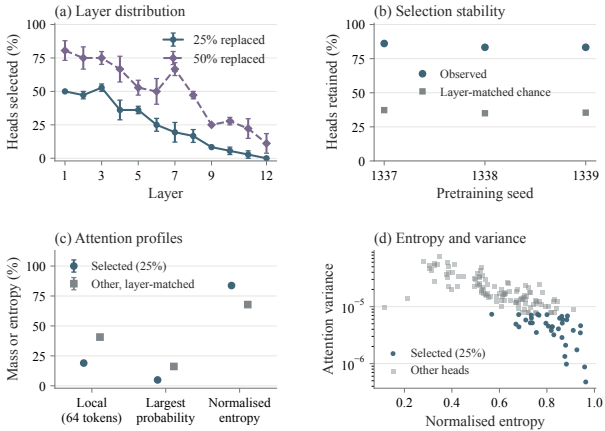
	10	25	50	75
Post-softmax mean	**+0.177**	**+0.662**	**+2.649**	**+7.821**
Sharp mean	+0.174	+0.665	+2.780	+7.870
Dirichlet	+0.186	+0.687	+2.825	+8.083
Structured random	+0.225	+0.773	+3.214	+9.210
Gaussian	+0.226	+0.775	+3.254	+9.129

Replaced heads (%)

- At 25% replacement: mean $+0.662\%$, sharp mean $+0.665\%$ PPL.
- At 50%: mean $+2.649\%$, sharp mean $+2.780\%$.
- Gaussian, Dirichlet and structured-random controls are not better than the mean.
- These figure values are a pattern comparison, distinct from the repeated three-seed main results.

Paper, Fig. 3, Sec. 5.1, PDF p. 6.

Low-variance heads are often early-layer and diffuse



- At 25%, selected heads cluster in earlier layers.
- Layer-matched normalized entropy: 0.837 selected vs. 0.678 other.
- Recent-64-token mass: 19.0% vs. 40.7%.
- Variance still predicts local replacement cost after controlling for layer and entropy (partial Spearman 0.543).

Paper, Appendix B.2, Fig. 7, PDF p. 20.

At 124M, 25% replacement is the quality-oriented setting

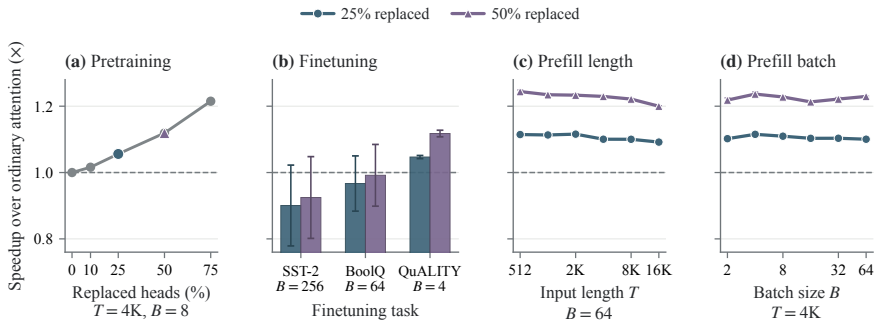
Context	PPL 25%	PPL 50%	Speed 25%	Speed 50%
4K	$0.768 \pm 0.067\%$	$2.492 \pm 0.083\%$	$1.056\times$	$1.119\times$
8K	$0.578 \pm 0.011\%$	$2.058 \pm 0.066\%$	$1.030\times$	$1.101\times$
16K	$0.450 \pm 0.012\%$	$1.792 \pm 0.112\%$	$1.025\times$	$1.070\times$

Interpretation

More replacement usually saves more update time while increasing PPL cost. Peak memory falls by 0.82–1.96% at 25% and 1.57–3.04% at 50%. Each context uses its own trained model and fitted patterns.

Paper, Table 2a, PDF p. 7. PPL changes are relative percentages; errors are SE over three seeds.

Measured speedups depend on the workload



- More replaced heads speed pretraining; short-task fine-tuning may slow.
- At batch 64, causal prefill gains are 1.09–1.12× at 25% and 1.20–1.24× at 50% across 512–16K tokens.

Paper, Fig. 4 and Table 2b, pp. 7–8; Appendix D.3, p. 38.

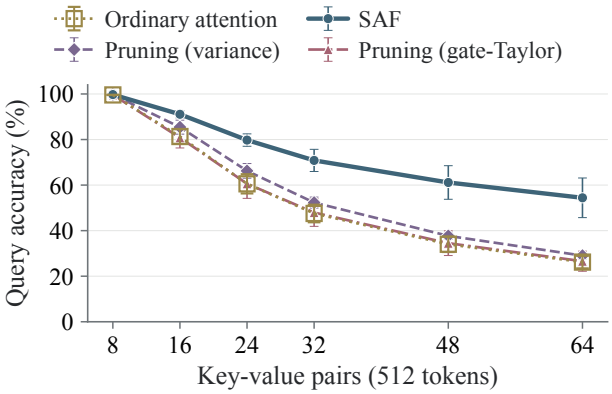
Token and time budgets favor different methods

Operation (25% heads)	Matched tokens	Matched time
Mean + variance selection	$0.663 \pm 0.018\%$	$0.456 \pm 0.108\%$
Prune the same heads	$0.806 \pm 0.018\%$	$0.113 \pm 0.169\%$
Mean + random heads	$1.282 \pm 0.056\%$	$1.243 \pm 0.112\%$

- Same-head SAF beats deletion at equal tokens.
- At equal time, pruning completes more updates and reaches lower PPL.
- Gate-Taylor pruning is stronger than same-head pruning in the later-intervention study.

Paper, Table 1, PDF p. 7; Appendix C.3.4, pp. 28–29. Mean $\pm$ SE over three seeds.

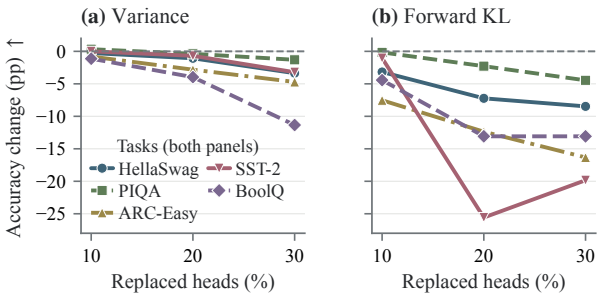
After MQAR adaptation, SAF handles more pairs at fixed length



- All models adapt on 8 pairs in 512-token sequences.
- Test increases pair count while keeping length at 512 tokens.
- At 64 pairs: SAF $54.42 \pm 8.70\%$; ordinary $26.19 \pm 2.75\%$; variance-pruned $29.03 \pm 1.42\%$; Gate-Taylor $26.59 \pm 4.41\%$.
- Length-only tests favor pruning; immediate replacement is less consistent.

Paper, Fig. 5, PDF p. 8; Appendix C.3.6, PDF pp. 31–32.

On Qwen3-4B, variance selection preserves more zero-shot accuracy



- No additional training; five zero-shot tasks.
- Variance selection beats forward-KL selection in all 15 task-rate comparisons.
- Mean accuracy change at 10% replacement: -0.37 pp with variance, -3.27 pp with forward KL.
- At 30%, variance selection still averages -4.79 pp.

Paper, Fig. 6 and Appendix C.6, PDF pp. 9, 35.

1B: the update trade-off extends to four GPUs

Quality and update speed	Scale-up caveat
Ordinary attention: PPL 12.446. SAF 25%: PPL 12.509 (+0.507%), update speedup 1.068×. SAF 50%: PPL 12.698 (+2.027%), update speedup 1.167×. Gate-Taylor pruning 25%: PPL 12.529.	One pretraining seed. Three fine-tuning seeds measure task-training variation, not pretraining variation.

- 8K context; 19.667B training tokens; four GH200 GPUs.
- Distributed timing includes gradient communication.
- 25% calibration and fitting took about 24 minutes.

Paper, Table 3, PDF p. 9; Appendix C.3.1/C.5, pp. 27, 34.

The speed claim has concrete boundaries

- The fixed path still computes $\widehat{P}V$ at $O(T^2 d_h)$; it removes score/softmax work and lowers pattern storage.
- Main speedups are per-update measurements and exclude calibration/fitting. For the reported 124M B=8 control, intervention took about 55 s; the full matched-token path was 184.7 min for SAF vs. 188.7 min for ordinary attention.
- Batch-one causal prefill at 4K was near break-even: median 0.969× at 25% and 0.972× at 50% due to launch overhead.
- Short-task fine-tuning did not consistently speed up; Qwen's 10% replacement still reduced mean accuracy by 0.37 pp.

Paper, Sec. 3.3 and Appendix C.3.1, D.3, PDF pp. 5, 27, 38; Table 2b, p. 7; Appendix C.6, p. 35.

Take-home: freeze stable routing, keep useful value mixing

1. Low attention variance is a practical signal for heads whose routing can be approximated by a fixed causal map.
2. The post-softmax mean plus a compact position-distance fit gives a simple static pattern.
3. SAF trades some quality for lower update cost; 25% is the more quality-oriented tested point.
4. Retained token mixing can beat deleting the same heads at matched tokens and in fixed-length associative recall, but matched-time pruning can be better.

Synthesis of Paper, Secs. 3–6, PDF pp. 3–9.

References and links

- Li, Tai, Fonseca, and Cohen. *When Can Attention Heads Be Statically Defined?* arXiv:2609.34650v1, 2026. Paper.
- Author-maintained implementation: Selective-Attention-Freezing on GitHub.
- Related execution baselines discussed by the paper: FlashAttention-2; same-head and Gate-Taylor pruning.

Publication status

No formal publication record found as of 2026-11-10; the author page labels this work "Preprint 2026".

The arXiv version, author page and code repository were checked on 2026-11-10.

Backup: replacement timing exposes the value of adaptation

- Replacing 25% of heads at 25%, 40%, 50%, 60%, 75%, and 100% of training gives PPL increases of 0.724, 0.679, 0.662, 0.711, 0.936, and 8.168%.
- Mid-training replacement leaves room to adapt. Replacing after training leaves no recovery updates.
- One-time midpoint replacement matches or beats the tested gradual and adaptive schedules at comparable final rates.

Paper, Appendix C.1, PDF pp. 22–23; Appendix B.3, p. 21.

Backup: the static matrix saves storage, not all attention arithmetic

- For the 4K, 36-head, 10-layer representation comparison, compact buffers use 0.133 GB vs. 8.060 GB in the dense implementation (60.6× less allocated pattern storage).
- An ideally packed dense allocation for only those 36 selected matrices is 2.416 GB.
- Dense vs. compact mean patterns change PPL by less than 0.014% in the tested absolute-position and RoPE settings.
- Fixed-pattern token mixing still costs $O(T^2 d_h)$.

Paper, Appendix D.1–D.2, Tables 24–25, PDF pp. 36–37.