Virtual Head Attention

A structured training parameterization that deploys as standard GQA-2

Paper by Xibo Ding, Guoxia Wang, Jinle Zeng, Jiabin Yang, Sirui Tao, Chao Yang, and Dianhai Yu

Baidu Inc. and Peking University

Independent research reading presentation

arXiv:2610.17959v1 | 19 October 2026

Main discussion: approximately 25 minutes; technical backups follow

The target is quality at a small KV budget

Research question

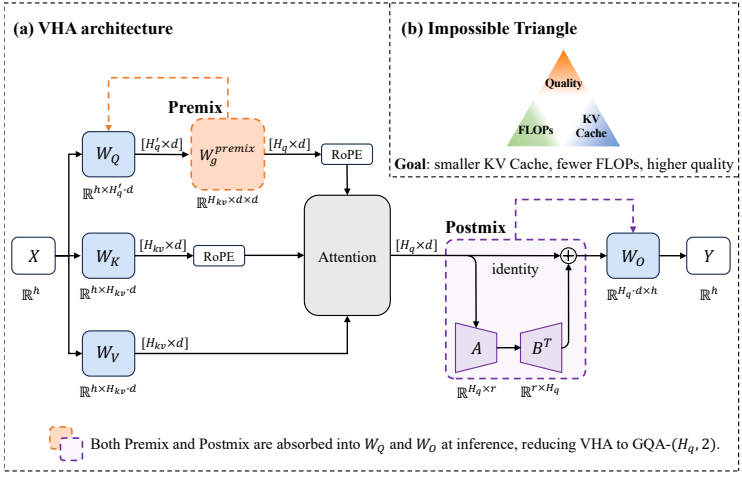
Can two KV heads retain GQA quality, reduce training arithmetic, and use standard serving kernels?

1.7B model	KV KB/layer/token	Train G/token	Avg. accuracy
GQA-8	4.00	10.61	54.09
MLA	1.12	10.42	52.49
MFA	1.00	15.90	54.06
VHA	**1.00**	**9.78**	**54.50**

50B training tokens; 4K context; eight-task zero-shot average (%).

Paper, Tables 1 and 5, PDF pp. 6 and 8. KV is BF16; KB uses 1024 bytes.

Two linear modules surround ordinary attention



Paper, Fig. 1(a,b), PDF p. 4; cropped from the original vector figure.

Eight physical heads become sixteen virtual heads

$$H'_q = H_q/2, \qquad H_{kv} = 2, \qquad H_v = H_{kv} H'_q = H_q.$$

1.7B configuration	Physical Q	Attention Q	KV heads
GQA baseline	16	16	8
VHA during training	8	16	2
VHA after absorption	16	16	2

- Each of two KV groups receives a transformed copy of all eight physical query heads.
- The attention operation still computes sixteen query heads.
- The paper's VHA 8-2-2 labels the training projections.

Paper, Sec. 3.2, PDF p. 4; Algorithm 1, PDF p. 20.

Premix specializes a shared query basis by KV group

$$Q_i = XW_Q^{(i)}, \qquad \widetilde{Q}_{i,g} = Q_i W_g^{\text{premix}}, \qquad W_g^{\text{premix}} \in \mathbb{R}^{d \times d}.$$

- Transform the feature dimension within each physical head.
- Share one transform across all physical heads in the same KV group.
- Initialize with an identity anchor: $W_g^{\text{premix}} = I + \alpha\epsilon$.

1.7B dimensions

$X : S \times 2048$; $W_Q^{(i)} : 2048 \times 128$;
two 128×128 group transforms; eight physical heads.

Paper, Sec. 3.3, Eqs. (1)–(3), PDF pp. 4–5; Table 12, p. 18.

Postmix adds a low-rank residual on the head axis

For one token, $Y \in \mathbb{R}^{H_v \times d}$ contains all attention-head outputs.

$$Y' = Y + B(A^\top Y) = (I + BA^\top)Y, \qquad A, B \in \mathbb{R}^{H_v \times r}.$$

- $A^\top$: compress head information into r channels.
- B: distribute it back across heads, including other KV groups.
- $B = 0$ at initialization, so the initial map is the identity.

The low-rank object is the residual

$\mathrm{rank}(BA^\top) \le r$; the full $I + BA^\top$ can be full rank.
At 1.7B: $H_v = 16$, $r = 4$.

Paper, Sec. 3.4, Eq. (4), PDF p. 5.

Both modules merge exactly into standard projections

$$\widetilde{W}_Q^{(i,g)} = W_Q^{(i)} W_g^{\mathrm{premix}}, \qquad W_Q^{\mathrm{abs}} = \mathrm{concat}_{g,i}\big(\widetilde{W}_Q^{(i,g)}\big).$$

$$W_O^{\mathrm{abs}} = \big((I + AB^\top) \otimes I_d\big) W_O.$$

- Premix is before RoPE; retain that order after merging.
- $BA^\top$ becomes $AB^\top$ because flattening transposes the head mixing.
- Discard the factor matrices and run ordinary GQA with two KV heads.

Meaning of zero extra deployment overhead

No additional Premix/Postmix operators relative to the same-shape GQA-2.

Paper, Eq. (5), p. 5; Algorithm 1 and Eqs. (14)–(18), PDF p. 20.

Forward equivalence does not imply identical training

Premix

Shared physical heads and group transforms couple gradients.

$$\widetilde{W}_Q^{(i,g)} = W_Q^{(i)} W_g^{\mathrm{premix}}$$

The query family is structurally constrained.

Postmix

The ordinary output projection already combines all heads.

$$W_O^{\mathrm{abs}} = ((I + AB^\top) \otimes I_d) W_O$$

Extra factors change the optimization path.

Interpretation: beneficial parameter sharing and optimization bias, rather than a larger deployed attention function class.

Paper, Sec. 2.3, p. 3; Appendices G–H, PDF pp. 17–21.

Factorized compute savings change after absorption

1.7B Q path, per layer	Parameters	Forward FLOPs/token
Ordinary 16-head Q	4.194M	8.389M
VHA physical Q + Premix	2.130M	4.719M
Absorbed VHA Q	4.194M	8.389M

- Q parameters fall by 49.2%; factorized Q FLOPs fall by 43.75%.
- H_v is unchanged: the quadratic attention term remains.
- Table 5's 7.9% reduction concerns total training FLOPs.
- Fully merged GQA prefill expands Q again; the KV savings remain.

Reader calculation from Sec. 3.5 and Appendix G dimensions, pp. 5–6, 18.
Reported total training cost: Table 5, PDF p. 8.

Three evaluation tracks test different claims

Track	Training setup	What it tests
0.6B / 1.7B / 4B / 8B-A1B	50B FineWeb-Edu tokens; 4K; BF16	Eight-task quality and arithmetic across scale
30B-A3B MoE	500B internal-corpus tokens; 8K; FP8	Four-category macro-average over 12 benchmarks
10B-A1B MoE	500B at 8K + 100B at 32K + 100B at 128K	Long-context behavior in a matched hybrid backbone

The 30B and long-context results use separate model families and recipes.

Paper, Sec. 4.1, PDF p. 6; Appendices A and D, pp. 14, 18.

VHA has the highest mean at all four tested scales

Scale	MHA	GQA	MLA	MFA	VHA	Δ vs. GQA
0.6B	50.22	50.42	49.52	50.66	**51.04**	+0.62
1.7B	53.24	54.09	52.49	54.06	**54.50**	+0.41
4B	56.68	56.90	56.38	56.94	**57.71**	+0.81
8B-A1B	54.60	54.86	53.79	54.81	**54.91**	+0.05

- Eight-task accuracy (%); differences are percentage points.
- VHA cache: 1 KB per layer/token at all four scales.
- GQA cache: 4 KB for dense models, 2 KB for the MoE.
- The 8B-A1B gain is small; no repeated-run uncertainty is given at that scale.

Paper, Table 1, PDF p. 6. Parameter-matched configurations use different head counts by method.

The compressed baseline isolates the recovery

1.7B configuration	Actual attention Q	Avg.	Δ vs. GQA-8
GQA 16-8-8	16	54.09	—
GQA 16-2-2	16	53.09	−1.00
GQA 8-2-2	8	52.38	−1.71
Premix only	16	53.31	−0.78
Postmix only	—	53.95	−0.14
Full VHA	**16**	**54.50**	**+0.41**

Same deployed shape: VHA beats GQA 16-2-2 by 1.41 points.

At 4B, Postmix-only already beats GQA by 0.32 points; the full model is best at +0.81.

Paper, Tables 2 and 4, PDF pp. 7–8. Postmix-only attention-head expansion is not specified precisely.

The 30B result is a four-category macro-average

Category / metric	GQA	MLA	VHA
General	52.86	55.57	**56.25**
Math & STEM	43.09	43.49	**46.40**
Coding	51.85	**54.37**	53.23
Multilingual	43.26	45.20	**45.56**
Four-category mean	47.76	49.66	**50.36**
KV KB/layer/token	2.00	1.12	**1.00**

VHA: +2.60 points vs. GQA; +0.70 vs. MLA.

HumanEval+ is a negative case: 47.56 vs. 50.00 / 52.44.

Paper, Table 3, PDF p. 7; Table 11, p. 17. 500B tokens; 8K; FP8 training; BF16 cache accounting.

Long-context validation includes 200B extra tokens

Matched 10B-A1B hybrid backbone

Full attention : sliding-window attention = 1 : 3.
Only full-attention layers change: GQA-4 $\rightarrow$ VHA-2.
Sliding-window layers keep eight KV heads in both models.

Stage	Tokens in stage	Cumulative steps	RoPE base
8K pretraining	500B	60K	10^4
32K mid-training	100B	72K	6.4×10^5
128K mid-training	100B	96K	5×10^6

KV halves in full-attention layers; the total hybrid cache ratio also depends on the unchanged sliding-window component.

Paper, Appendix A.1, PDF p. 14.

RULER gains are largest in variable tracking

	RULER Overall		Variable Tracking	
Length	GQA	VHA	GQA	VHA
8K	79.1	**90.1**	44.6	94.0
16K	78.2	**88.5**	28.4	93.4
32K	72.6	**85.3**	3.2	92.6
64K	69.5	**81.3**	4.2	87.2
128K	65.7	**76.8**	3.0	83.0
Length mean	73.0	**84.4**		

At 128K, variable tracking contributes about $80/13 = 6.15$ of the 11.1-point Overall gap. Overall weights all 13 tasks equally.

Paper, Table 6, PDF p. 15; final 128K checkpoint. Contribution is a reader calculation.

Long-context gains vary by task and training stage

	32K checkpoint		128K checkpoint	
Score ↑	GQA	VHA	GQA	VHA
LongBench-v2	**28.0**	22.3	18.9	**24.2**
InfiniteBench	**29.9**	27.3	**31.8**	31.5
BABILong	36.5	**37.4**	31.1	**38.2**
Short eight-task mean	62.7	**63.7**	62.6	**64.1**

- LongBench-v2 ranking reverses between checkpoints.
- Final HELMET length means: 3 wins / 1 tie / 1 loss.
- HELMET ICL favors GQA at all five tested lengths.

Paper, Tables 6–7, PDF p. 15. BPB and HELMET details are in the backup slides.

Small linear factors work best in the ablations

Postmix width, 1.7B

r	Avg.
4	**54.50**
32	54.24
128	54.15

Training loss decreases with larger r, but downstream accuracy falls.

SiLU insertion, 1.7B

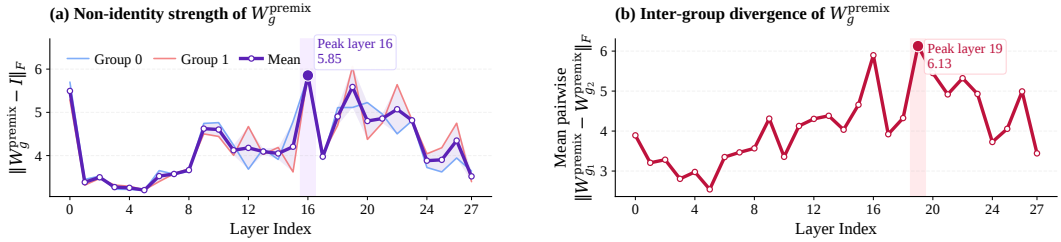
Variant	Avg.
Premix + SiLU	53.83
Postmix + SiLU	53.79
Both + SiLU	53.67
Linear default	**54.50**

Nonlinearity also prevents exact projection absorption.

At $H_v = 16$, factor widths 32 and 128 both allow rank at most 16.

Paper, Sec. 4.4 and Table 4, PDF p. 8. Rank cap follows from matrix dimensions.

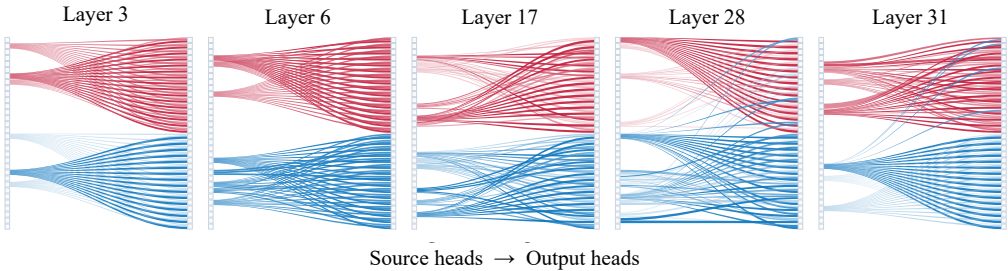
Premix learns different transforms for the two groups



- Left: distance from identity across layers; peak group mean 5.85.
- Right: distance between the two group transforms; peak 6.13.
- The trained factors are nontrivial and group-specific.

Paper, Fig. 2, Sec. 5, PDF p. 9. Trained 1.7B model; the two peaks occur at different layers.

Postmix mainly redistributes within KV groups



- Source heads flow to recombined heads; color denotes KV group.
- Strong off-diagonal mixing, mainly within groups, with cross-group bridges.
- Absolute residual weights omit signs and the identity path.

Paper, Fig. 3, Sec. 5, PDF pp. 9–10. Trained 4B model. This is a weight-structure probe.

Repeated runs and initialization support the design

Repeated 0.6B training

Model	Runs	Mean ± SD
VHA	4	50.40 ± 0.07
GQA	3	50.00 ± 0.34

SD is the sample standard deviation across run averages.

Premix initialization at 1.7B

Initialization	Avg.
$I + 0.1\epsilon$	**54.04**
$I + 1.0\epsilon$	53.45
Fully random	53.11

All three runs were stable; the default gave the highest score.

These studies use a different training implementation from the main tables; compare within each study.

Paper, Appendix B, Tables 8–10, PDF pp. 14–17.

The evidence supports quality and arithmetic savings

Directly supported

Exact algebraic absorption; smaller KV cache; reported FLOPs; quality gains under the evaluated training recipes.

- The claimed 2–4× decoding speedup is a memory-bound estimate. No end-to-end serving benchmark is reported.
- Standard-GQA absorption expands Q; factorized prefill savings require a compatible execution path.
- Weight probes suggest specialization and communication; targeted interventions would test their causal roles.

Paper, Sec. 3.5, p. 6; Sec. 5, pp. 9–10; inference analysis based on Algorithm 1, p. 20.

Train with structure; deploy with ordinary GQA

1. **Share and specialize queries.**
   Two group transforms expand a compact physical query basis.
2. **Change the output parameterization.**
   Low-rank residual mixing improves the tested training outcomes.
3. **Keep deployment simple.**
   Both factors merge into standard projections with two KV heads.

Discussion: Does the benefit persist at equal training FLOPs, and which learned connections actually cause the gain?

Xibo Ding et al., *Virtual Head Attention*, arXiv:2610.17959v1, 2026.
https://arxiv.org/abs/2610.17959v1. Next: technical backup slides.

Backup: Premix couples gradients across heads and groups

Let $D_{i,g} = \partial \mathcal{L} / \partial \widetilde{Q}_{i,g}$. Then

$$\frac{\partial \mathcal{L}}{\partial W_Q^{(i)}} = \sum_{g=1}^{H_{kv}} X^\top D_{i,g} \big(W_g^{\text{premix}}\big)^\top,$$

$$\frac{\partial \mathcal{L}}{\partial W_g^{\text{premix}}} = \sum_{i=1}^{H_q'} (X W_Q^{(i)})^\top D_{i,g}.$$

- A physical query basis aggregates signals from both KV groups.
- A group transform aggregates signals from all physical heads.
- This establishes different update dynamics; the downstream benefit is empirical.

Paper, Appendix G, Eqs. (12)–(13), PDF p. 18.

Backup: the transpose comes from row-major flattening

For $Y \in \mathbb{R}^{H_v \times d}$ and $M = I + BA^\top$,

$$\mathrm{flat}(MY) = \mathrm{flat}(Y)(M^\top \otimes I_d).$$

Thus

$$\begin{aligned}\mathrm{flat}(Y')W_O &= \mathrm{flat}(Y)\big((I + AB^\top) \otimes I_d\big)W_O \\ &= \mathrm{flat}(Y)W_O^{\mathrm{abs}}.\end{aligned}$$

What is required?

Consistent head ordering and adjacent linear operations.
M need not be invertible for this absorption identity.

This preserves real-arithmetic outputs; the deployed Q projection contains all virtual heads.

Paper, Appendix H, Eqs. (14)–(18), PDF p. 20. Invertibility observation is a reader clarification.

Backup: applying a shared matrix still costs FLOPs

Per token and layer, counting a multiply-add as two FLOPs:

$$
\begin{aligned}
F_Q^{\mathrm{GQA}} &= 2hH_v d, \\
F_Q^{\mathrm{VHA}} &= 2hH'_q d + 2H_{kv} H'_q d^2, \\
F_{\mathrm{Postmix}} &= 4H_v r d.
\end{aligned}
$$

- At 1.7B: $h = 2048, d = 128, H'_q = 8, H_v = 16, r = 4$.
- Q FLOPs: 8,388,608 $\rightarrow$ 4,194,304 + 524,288.
- Postmix adds 32,768 FLOPs.
- Shared Premix parameters are counted once, but applied to every physical head.

Reader calculation from Sec. 3.5, PDF p. 5, and Appendix G, p. 18. These are forward costs, not total training FLOPs.

Backup: joint resource trade-offs across all scales

Scale	Training G/token, $S = 4096$ GQA	MFA	VHA	KV KB/layer/token GQA	VHA
0.6B	7.57	11.54	6.96	4.00	1.00
1.7B	10.61	15.90	9.78	4.00	1.00
4B	24.38	36.04	22.52	4.00	1.00
8B-A1B	10.27	17.82	9.60	2.00	1.00

- Reported VHA training reductions: 8.1 / 7.9 / 7.6 / 6.5%.
- MFA matches the 1 KB cache but uses larger head dimension.
- MHA uses fewer query heads in the matched configurations; it can have fewer FLOPs but a much larger cache.

Paper, Table 5, PDF p. 8. Relative percentages are copied from the paper; displayed FLOPs are rounded.

Backup: final HELMET and language-modeling results

HELMET length means ↑

Category	GQA	VHA
RAG	28.14	**37.48**
LongQA	15.70	**16.10**
Summarization	6.94	**7.20**
ICL	**72.68**	68.20
Reranking	0.84	0.84

ProLong bits per byte ↓

Domain	GQA	VHA
arXiv	0.582	**0.564**
Books	0.819	**0.804**
Code	0.311	**0.296**
Web	0.768	**0.760**

HELMET means are reader calculations over 8K, 16K, 32K, 64K, and 128K, using the final checkpoint.

Paper, Tables 6–7, PDF p. 15. BPB uses the final 128K checkpoint.

Backup: category gains hide meaningful task differences

30B-A3B benchmark	GQA	MLA	VHA
MMLU	55.97	59.34	**60.24**
MMLU-Pro	36.20	**41.12**	40.03
GPQA	**34.62**	31.62	34.11
GSM8K	**56.48**	55.42	56.25
MATH	38.18	43.42	**48.84**
HumanEval+	50.00	**52.44**	47.56

Four categories contain 4 / 3 / 3 / 2 benchmarks. The reported overall score is

$$\frac{1}{4}\left(\overline{s}_{\mathrm{General}} + \overline{s}_{\mathrm{Math}} + \overline{s}_{\mathrm{Coding}} + \overline{s}_{\mathrm{Multilingual}}\right).$$

Paper, Tables 3 and 11, PDF pp. 7 and 17. Each benchmark retains the paper's evaluation protocol.

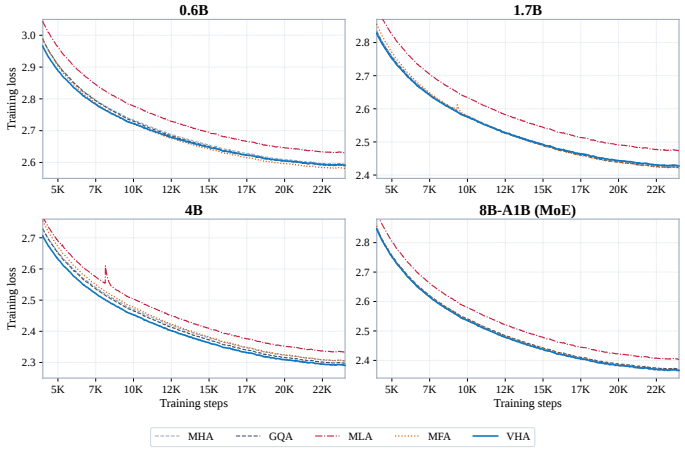
Backup: model sizes share a consistent optimizer recipe

	0.6B	1.7B	4B	8B-A1B	30B-A3B
Layers	28	28	36	24	46
Hidden width	1024	2048	2560	2048	1536
VHA physical Q	8	8	16	16	24
VHA virtual Q	16	16	32	32	48
Postmix r	4	4	8	8	12
Training tokens	50B	50B	50B	50B	500B
Context	4K	4K	4K	4K	8K

AdamW: $(\beta_1, \beta_2) = (0.9, 0.95)$; weight decay 0.1;
gradient clipping 1.0; warmup 2000 steps.
Premix: $\alpha = 0.1$, $\sigma = 1$; two KV heads throughout.

Paper, Tables 1, 3 and 12, PDF pp. 6–7, 18. The separate 10B long-context recipe is on slide 14.

Backup: VHA loss is competitive across four scales



Paper, Fig. 4, PDF p. 19. 50B tokens; 4K context; EMA smoothing $\alpha = 0.93$; curves start at step 4K. VHA (solid blue) is lowest at 4B; the paper reports a gap within 0.01 nats from the best at the other scales.