TransMLA: MLA is ALL You Need

Fanxu Meng*, Pingzhi Tang*, Xiaojuan Tang, Zengwei Yao, Xing Sun, Muhan Zhang

Peking University · Tencent Youtu Lab · Xiaomi Corp. | *arXiv 2502.07864*

Presented by ***Kangqi (Fred) Yang***, ***Jessica Hernandez***

The KV Cache Bottleneck

- Autoregressive generation caches K and V for every past token
- Cache size grows **linearly** with context length
- At long contexts: **memory bandwidth**, not compute, is the bottleneck

Approach	Drawback
MQA / GQA	Accuracy Loss
Post-training compression	Non-standard kernels; hard to deploy
MLA (DeepSeek V2/V3/R1)	Requires **retraining from scratch**

> Most Providers are locked into GQA — billions of tokens already invested.

Can We Convert GQA → MLA Without Retraining?

LLaMA, Qwen, Gemma, Mistral → MLA
Inheriting pretrained weights, with only light fine-tuning

Two obstacles:

1. **Is it worth it?** MLA must be provably better than GQA
2. **RoPE blocks the key MLA inference optimization** — must be decoupled

Related Work

Attention variants (same KV budget, increasing expressiveness):

$$\text{MHA} \rightarrow \text{GQA} \rightarrow \text{MQA} \rightarrow \text{MLA}$$

Closest prior works:

- **Palu** — low-rank K/V compression, but **ignores RoPE** → Absorb blocked → no real speedup
- **MHA2MLA** (concurrent) — norm-based RoPE pruning → sparse, uneven → **no speedup demonstrated**

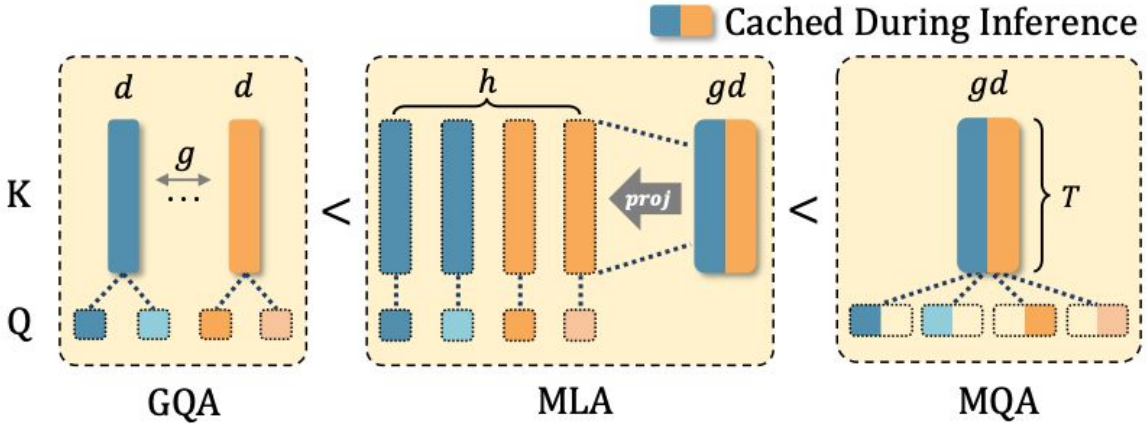
TransMLA: proper RoPE decoupling → Absorb works → **10.6× speedup**

Background: Group Query Attention (GQA)

h query heads, g shared K/V heads ($g < h$)

$$\text{KV cached per token} = 2gd \text{ scalars}$$

Expressiveness limit: only g distinct K/V vectors — all queries in the same group see identical keys and values



Background: MLA & the Absorb Operation

Cache one small latent $\mathbf{c}_t^{KV} \in \mathbb{R}^{r_{kv}}$ instead of g full K/V heads

Absorb — fold W^{UK} into the query at inference:

$$\mathbf{q}^\top \underbrace{W^{UK}\mathbf{c}^{KV}}_{\mathbf{k}} = \underbrace{(W^{UK\top}\mathbf{q})}_{\hat{\mathbf{q}}}{}^\top \mathbf{c}^{KV}$$

→ Attend directly over the tiny latent — behaves like MQA at inference

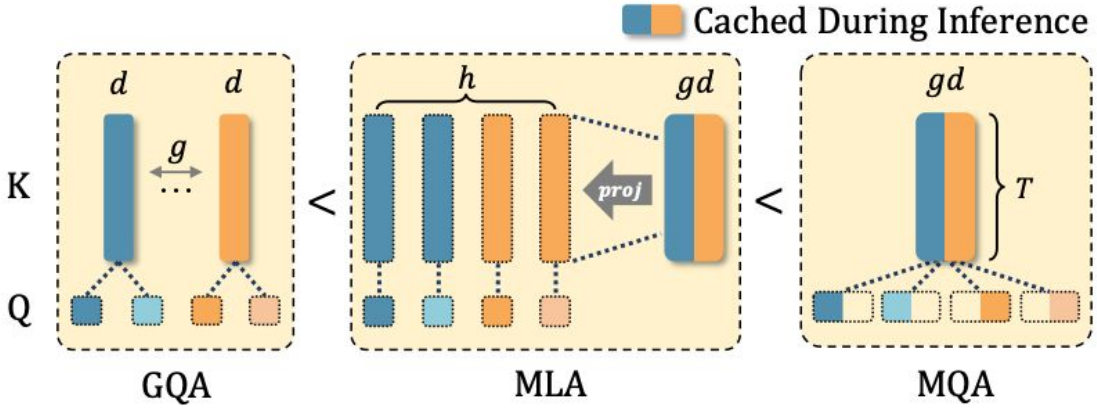
Requirement: $\mathbf{c}^{KV}$ must be **RoPE-free**

MLA is Provably More Expressive Than GQA

$$\text{GQA} < \text{MLA} < \text{MQA}$$

(same KV cache budget)

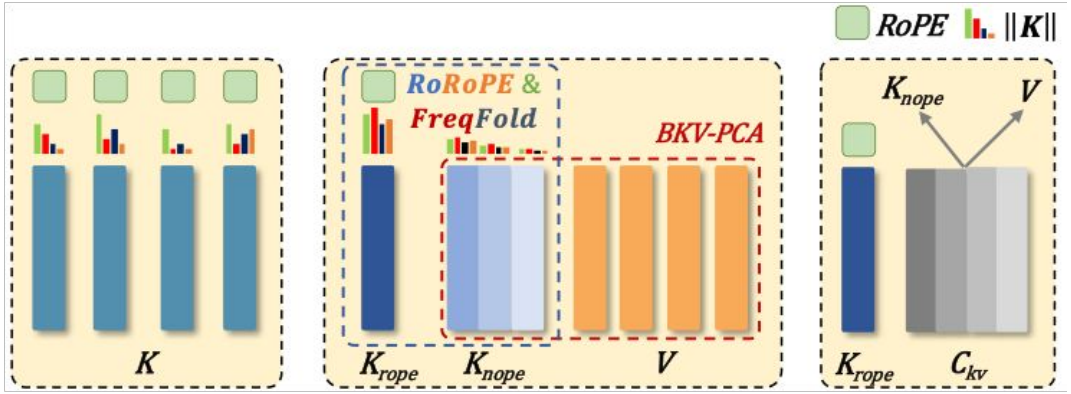
- **GQA < MLA:** GQA restricted to g shared K/V vectors; MLA's dense W^{UK} creates h distinct ones
- **MLA < MQA:** absorbed MQA allows higher-rank query–key interaction



TransMLA: Three-Step Pipeline

GQA model → Step 1: Merge → Step 2: RoRoPE + FreqFold → Step 3: BKV-PCA → MLA model

Step	What it does	Loss
Merge	g K/V groups → one shared latent	Zero
RoRoPE + FreqFold	Concentrate positional info → decouple RoPE	Minimal
BKV-PCA	Compress NoPE-K and V jointly	Tunable



Step 1: Merge K/V Groups into One Latent

Initialize W_i^{UK} as a **block identity selector** for each head

$$\mathbf{c}_t^{KV} = \begin{pmatrix} W^K \\ W^V \end{pmatrix} \mathbf{x}_t \in \mathbb{R}^{2gd} \qquad \longleftarrow \text{exact, zero loss}$$

Why merge first?

- MLA requires a single shared latent
- Enables cross-head PCA in Step 3
- Required for RoPE decoupling in Step 2

Step 2: Why RoPE Blocks Absorb

RoPE encodes position by rotating each K/Q vector:

$$\mathbf{q}_t^\top \mathbf{k}_j^R = f(\mathbf{q}_t, \mathbf{k}_j, \underbrace{t-j}_{\text{relative pos}})$$

After merging: **every dimension** of the latent carries positional info

→ Cannot separate position-dependent rotation from static $\mathbf{c}^{KV}$

→ **Absorb operation breaks**

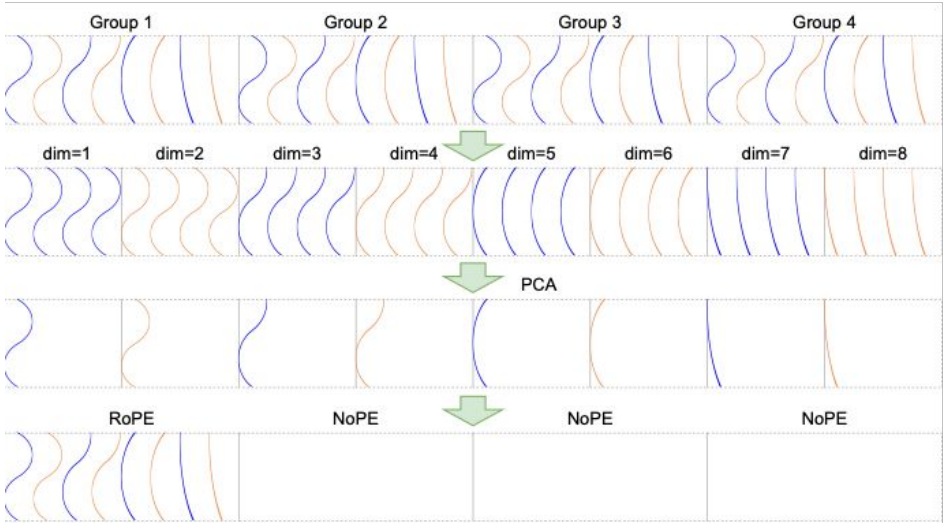
Step 2: RoRoPE — Rotate to Concentrate, Then Remove

Key insight: orthogonal rotation U_l within each RoPE frequency subspace preserves attention scores

→ Use PCA to find that U_l concentrates positional energy into **head 1**

→ Remove RoPE from heads 2… g — they now carry near-zero positional content

FreqFold: group adjacent frequencies (nearly identical $\theta_l \approx \theta_{l+1}$) for joint PCA → more capacity, provably better variance retention.



Step 3: Balanced KV-PCA

Goal: compress $[K_{\text{nope}}; V]$ jointly into low-rank latent r_{kv}

Problem: $\|K_{\text{nope}}\|_2 \gg \|V\|_2$

→ native PCA ignores values entirely → accuracy drop

Fix — scale by α before PCA:

$$\alpha = \frac{\mathbb{E}[\|K_{\text{nope}}\|_2]}{\mathbb{E}[\|V\|_2]}, \qquad K_{\text{nope}} \leftarrow \frac{K_{\text{nope}}}{\alpha} \quad \text{(mathematically equivalent)}$$

Evaluation

Experimental Setup

- **Models:** Models SmolLM-1.7B and LLaMa-2-7B
- **Benchmarks:** 6 standard tasks (MMLU, ARC, PIQA, HellaSwag, OBQA, Winogrande)
- **Baseline:** Compares directly against MHA2MLA
- **Hardware:** 8-GPU cluster (A100/H100 class) for training; consumer GPUs (24GB/40GB/64GB) for inference speed tests

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

Model	Tokens	KV Mem.	Avg.	MMLU	ARC	PIQA	HS	OBQA	WG
SmolLM-1.7B	1T	–	55.90	39.27	59.87	75.73	62.93	42.80	54.85
- MHA2MLA	0	-68.75%	40.97	27.73	41.96	63.00	29.19	34.40	49.53
		-81.25%	37.14	26.57	32.73	55.77	26.90	31.40	49.49
		-87.50%	34.01	25.32	27.15	51.36	25.47	26.20	48.54
	6B	-68.75%	54.76	38.11	57.13	76.12	61.35	42.00	53.83
		-81.25%	54.65	37.87	56.81	75.84	60.41	42.60	54.38
		-87.50%	53.61	37.17	55.50	74.86	58.55	41.20	54.38
- TransMLA	0	-68.75%	51.95	35.70	55.68	73.94	53.04	39.80	53.51
		-81.25%	47.73	32.87	47.89	69.75	48.16	36.20	51.46
		-87.50%	44.12	29.97	41.72	66.87	41.15	34.80	50.28
	300M	-68.75%	55.24	38.60	58.95	74.97	61.52	43.00	54.38
	700M	-81.25%	54.78	37.79	57.53	75.52	59.88	42.80	55.17
	1B	-87.50%	54.01	37.24	56.32	74.81	60.08	42.40	53.20
LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- MHA2MLA	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- TransMLA	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

SmolLM 1.7B

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

Model	Tokens	KV Mem.	Avg.	MMLU	ARC	PIQA	HS	OBQA	WG
SmolLM-1.7B	1T	–	55.90	39.27	59.87	75.73	62.93	42.80	54.85
- MHA2MLA	0	-68.75%	40.97	27.73	41.96	63.00	29.19	34.40	49.53
		-81.25%	37.14	26.57	32.73	55.77	26.90	31.40	49.49
		-87.50%	34.01	25.32	27.15	51.36	25.47	26.20	48.54
	6B	-68.75%	54.76	38.11	57.13	76.12	61.35	42.00	53.83
		-81.25%	54.65	37.87	56.81	75.84	60.41	42.60	54.38
		-87.50%	53.61	37.17	55.50	74.86	58.55	41.20	54.38
- TransMLA	0	-68.75%	51.95	35.70	55.68	73.94	53.04	39.80	53.51
		-81.25%	47.73	32.87	47.89	69.75	48.16	36.20	51.46
		-87.50%	44.12	29.97	41.72	66.87	41.15	34.80	50.28
	300M	-68.75%	55.24	38.60	58.95	74.97	61.52	43.00	54.38
	700M	-81.25%	54.78	37.79	57.53	75.52	59.88	42.80	55.17
	1B	-87.50%	54.01	37.24	56.32	74.81	60.08	42.40	53.20
LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- MHA2MLA	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- TransMLA	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

LLaMA-2-7B

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

Model	Tokens	KV Mem.	Avg.	MMLU	ARC	PIQA	HS	OBQA	WG
SmolLM-1.7B	1T	–	55.90	39.27	59.87	75.73	62.93	42.80	54.85
- MHA2MLA	0	-68.75%	40.97	27.73	41.96	63.00	29.19	34.40	49.53
		-81.25%	37.14	26.57	32.73	55.77	26.90	31.40	49.49
		-87.50%	34.01	25.32	27.15	51.36	25.47	26.20	48.54
	6B	-68.75%	54.76	38.11	57.13	76.12	61.35	42.00	53.83
		-81.25%	54.65	37.87	56.81	75.84	60.41	42.60	54.38
		-87.50%	53.61	37.17	55.50	74.86	58.55	41.20	54.38
- TransMLA	0	-68.75%	51.95	35.70	55.68	73.94	53.04	39.80	53.51
		-81.25%	47.73	32.87	47.89	69.75	48.16	36.20	51.46
		-87.50%	44.12	29.97	41.72	66.87	41.15	34.80	50.28
	300M	-68.75%	55.24	38.60	58.95	74.97	61.52	43.00	54.38
	700M	-81.25%	54.78	37.79	57.53	75.52	59.88	42.80	55.17
	1B	-87.50%	54.01	37.24	56.32	74.81	60.08	42.40	53.20
LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- MHA2MLA	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- TransMLA	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

LLaMA-2-7B

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

Model	Tokens	KV Mem.	Avg.	MMLU	ARC	PIQA	HS	OBQA	WG
SmolLM-1.7B	1T	–	55.90	39.27	59.87	75.73	62.93	42.80	54.85
- MHA2MLA	0	-68.75%	40.97	27.73	41.96	63.00	29.19	34.40	49.53
		-81.25%	37.14	26.57	32.73	55.77	26.90	31.40	49.49
		-87.50%	34.01	25.32	27.15	51.36	25.47	26.20	48.54
	6B	-68.75%	54.76	38.11	57.13	76.12	61.35	42.00	53.83
		-81.25%	54.65	37.87	56.81	75.84	60.41	42.60	54.38
		-87.50%	53.61	37.17	55.50	74.86	58.55	41.20	54.38
- TransMLA	0	-68.75%	51.95	35.70	55.68	73.94	53.04	39.80	53.51
		-81.25%	47.73	32.87	47.89	69.75	48.16	36.20	51.46
		-87.50%	44.12	29.97	41.72	66.87	41.15	34.80	50.28
	300M	-68.75%	55.24	38.60	58.95	74.97	61.52	43.00	54.38
	700M	-81.25%	54.78	37.79	57.53	75.52	59.88	42.80	55.17
	1B	-87.50%	54.01	37.24	56.32	74.81	60.08	42.40	53.20
LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- MHA2MLA	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- TransMLA	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

LLaMA-2-7B

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- MHA2MLA	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- TransMLA	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- MHA2MLA	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- TransMLA	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- *MHA2MLA*	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- *TransMLA*	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- MHA2MLA	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- TransMLA	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- MHA2MLA	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- TransMLA	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- MHA2MLA	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- TransMLA	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

TransMLA vs MHA2MLA Baselines

- Comparison of SmolLM and Llama-2 performance across 6 benchmarks

Model	Tokens	KV Mem.	Avg.	MMLU	ARC	PIQA	HS	OBQA	WG
SmolLM-1.7B	1T	–	55.90	39.27	59.87	75.73	62.93	42.80	54.85
- MHA2MLA	0	-68.75%	40.97	27.73	41.96	63.00	29.19	34.40	49.53
		-81.25%	37.14	26.57	32.73	55.77	26.90	31.40	49.49
		-87.50%	34.01	25.32	27.15	51.36	25.47	26.20	48.54
	6B	-68.75%	54.76	38.11	57.13	76.12	61.35	42.00	53.83
		-81.25%	54.65	37.87	56.81	75.84	60.41	42.60	54.38
		-87.50%	53.61	37.17	55.50	74.86	58.55	41.20	54.38
- TransMLA	0	-68.75%	51.95	35.70	55.68	73.94	53.04	39.80	53.51
		-81.25%	47.73	32.87	47.89	69.75	48.16	36.20	51.46
		-87.50%	44.12	29.97	41.72	66.87	41.15	34.80	50.28
	300M	-68.75%	55.24	38.60	58.95	74.97	61.52	43.00	54.38
	700M	-81.25%	54.78	37.79	57.53	75.52	59.88	42.80	55.17
	1B	-87.50%	54.01	37.24	56.32	74.81	60.08	42.40	53.20
LLaMA-2-7B	2T	–	59.85	41.43	59.24	78.40	73.29	41.80	64.96
- MHA2MLA	0	-68.75%	37.90	25.74	32.87	59.41	28.68	28.60	52.09
		-81.25%	34.02	25.50	26.44	53.43	27.19	22.60	49.01
		-87.50%	32.70	25.41	25.79	50.60	26.52	19.40	48.46
	6B	-68.75%	59.51	41.36	59.51	77.37	71.72	44.20	62.90
		-81.25%	59.61	40.86	59.74	77.75	70.75	45.60	62.98
		-87.50%	58.96	40.39	59.29	77.75	69.70	43.40	63.22
- TransMLA	0	-68.75%	58.20	39.90	57.66	77.48	70.22	41.00	62.90
		-87.50%	51.19	34.39	45.38	71.27	60.73	37.40	57.93
		-92.97%	43.26	28.93	36.32	63.38	45.87	31.60	53.43
	500M	-68.75%	59.82	40.87	59.18	77.91	71.82	45.20	63.93
	3B	-87.50%	59.36	40.77	58.84	78.18	71.28	43.60	63.46
	6B	-92.97%	58.68	40.82	59.72	76.55	69.97	43.60	61.40

Result

TransMLA achieves 98.35% of the original model's performance without any training

6B Tokens for pre-training

- Apply fine-tuning phase with only **6B tokens** (~0.3% of Llama-2's original training)

> **Result**
> 98% Recovery after fine tuning

6B Tokens for pre-training - Example

- LLaMa-2-7B, 7% KV cache)

w/o Training	The capital of France is Paris. Its geographical position in the Iberian Plain of France, Spain, Spain, and Morocco are the four largest cities. This region is located in Asia, Spain and Morocco.
w/o Training	The president of the United States is elected by the legislature. The legislature controls the national armed forces, but only provides the funds to establishing a national guard.

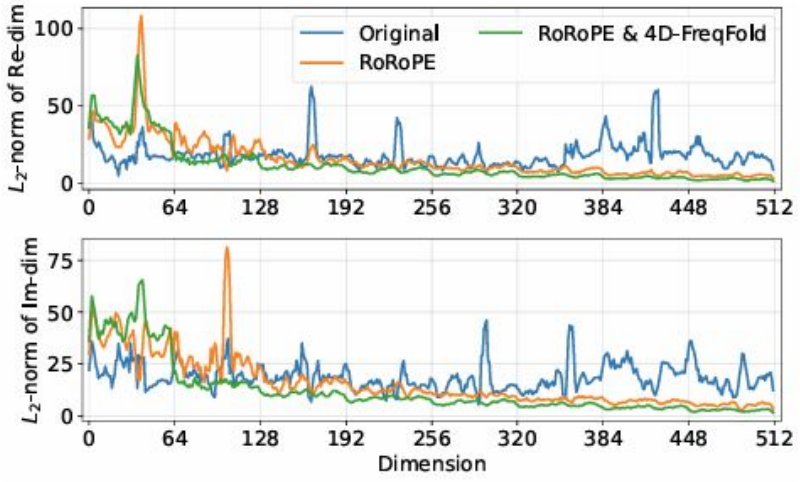
6B Tokens for pre-training - Example (cont.)

- After pre-training phase…

Pre-Training	The capital of France is a major business city and it is a favorite destination for businesses from all over the world. It has a strategic location in the heart of the European Union, which makes it one of the most popular cities in Europe.
Pre-Training	The president of the United States is elected to a four-year term by the people of each state in a general election held every four years on the Tuesday following the first Monday in November.

RoRoPE Effect

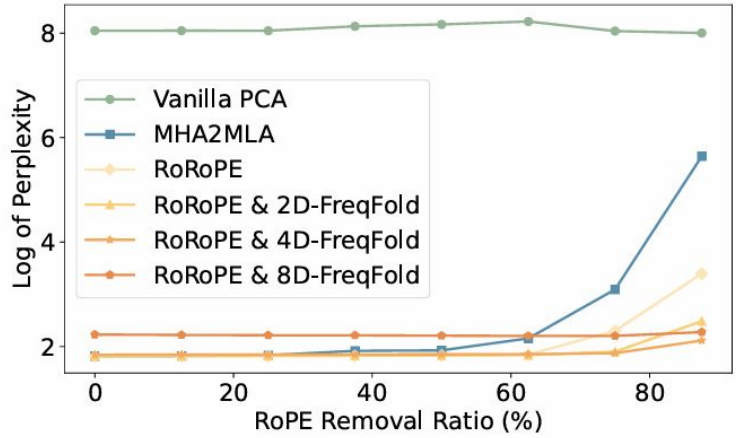
- In the original model, the L2 Norm is scattered across all heads.
  - After RoRoPE, almost all energy is spiked in the first 128 dimensions



(a) Key norms of the first layer.

Perplexity

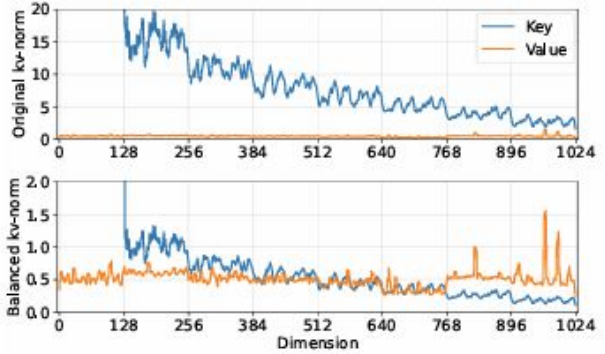
- Tested limits of RoPe removal to see when model's logic collapses
  - Increased RoPe removal ratio from 0% to 90%



(b) RoPE removal results, evaluated on WikiText-2.

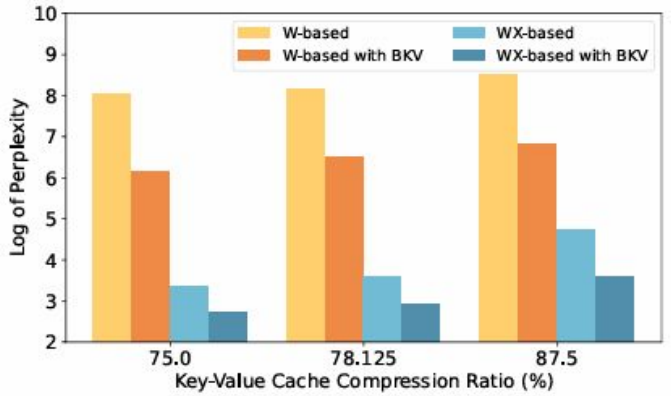
BKV - Balancing Keys and Values

- **Problem:** Key (K) activations often have much larger magnitudes than Value (V) activations
- **Risk:** Applying naive joint-PCA for compression may cause the model to forget the content (V) to save the position (K)
- **Solution:** “Balanced Key-Value” normalization step



(a) Norm magnitude of keys and values before and after KV balancing is applied.

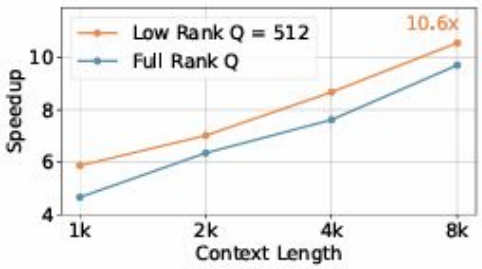
BKV - Balancing Keys and Values



(b) Comparison of different methods of KV cache compression, with and without KV balancing.

10.6x Speedup Inference

- Findings: 10.6x speedup at 8K context
  - Speed gains grow exponentially with context length



(a) 165.2 TFLOPS | 24GB

Conclusion

- TransMLA provides a bridge from GQA to MLA architecture
  - Pros: 10x speedup, min performance hit, compatible with DeepSeek ecosystem
  - Future works:
    - Validation across more models
    - Pruning, quantization, token selection to explore upper bounds of inference acceleration

Thank you!