DGIST Daegu Gyeongbuk Institute of Science & Technology

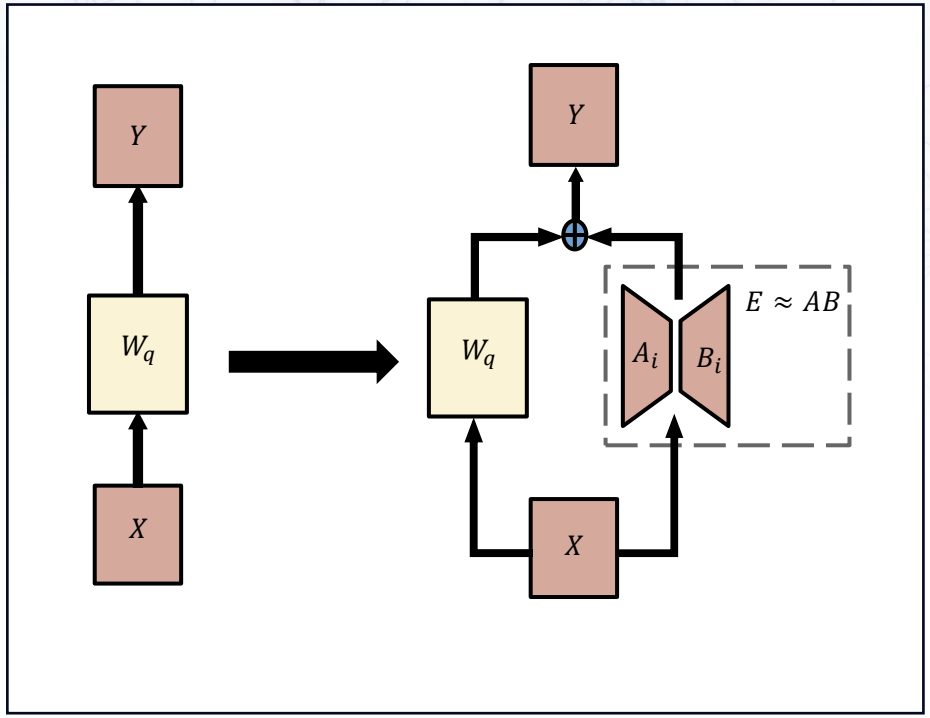
GlowQ: Group-Shared LOw-Rank Approximation for Quantized LLMs

Authors : Selim An, Ilhong Suh, Yeseong Kim

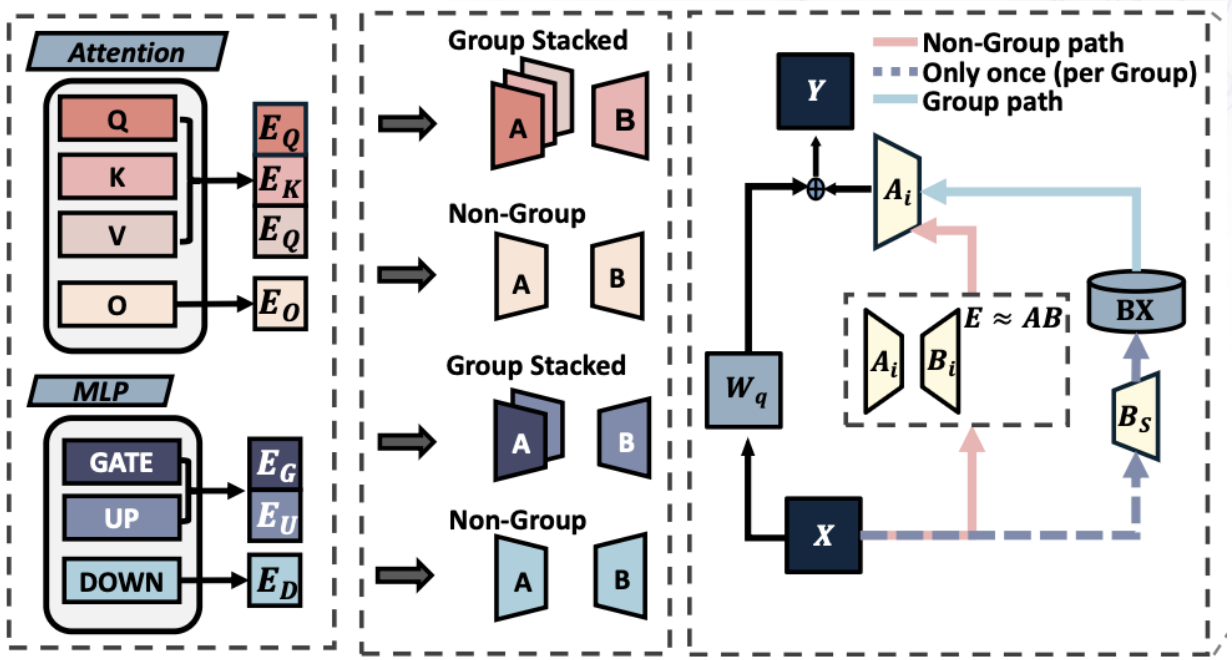
ICLR 2026

Motivation

- Post-training quantization reduces memory and bandwidth cost
- But low-bit quantization still hurts model quality
- A common fix is to add a low rank correction term AB which is factorized error ($E = W - W_q$)
- But attaching a separate AB to each layer/module can cause memory traffic.

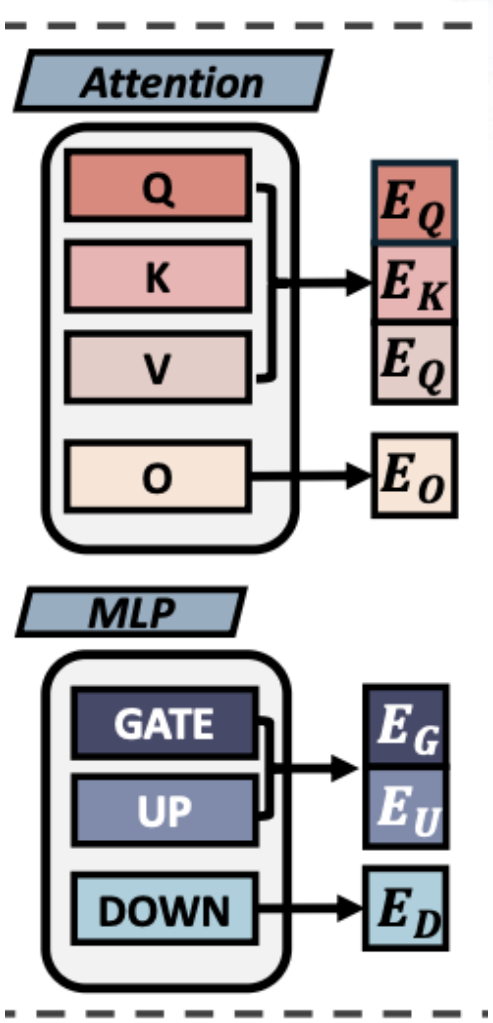


Overview of GlowQ



- Group error modules that share the same input.
- For each group, it learns :
  - one shared right factor B_{shared} per group
  - keep module-specific left factors A_i.
- At inference:
  - compute $R = B_{shared}X$ once
  - reuse it for all modules in the group via A_iR.

Grouping strategy 1



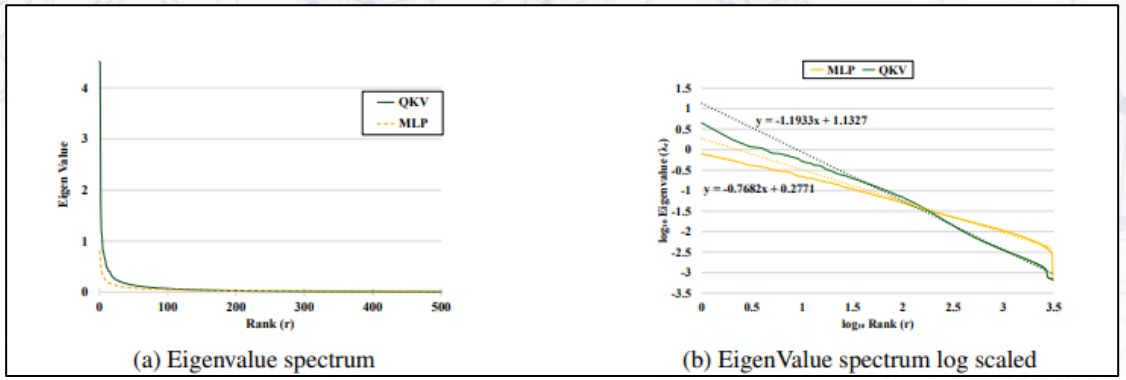
- In transformer blocks, several modules consume the same input activation.
- GlowQ forms input-sharing groups such as:
- Attention group : q, k, v
- MLP group : $gate, up$
- For grouped modules, GlowQ vertically stacks their quantization errors :
- $E_{cat} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_m \end{bmatrix}$
- Then it solves a joint low-rank approximation : $E_i \approx A_i B, E_{cat} \approx AB$

Grouping strategy2

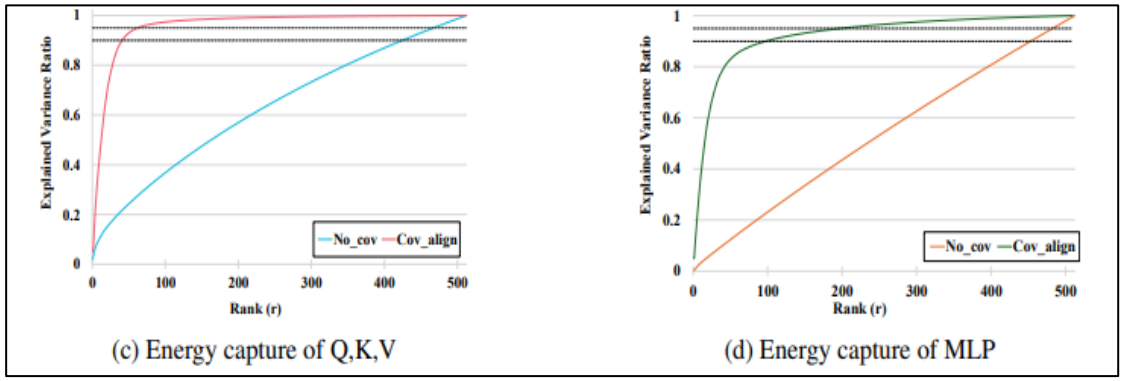
- A plain stacked SVD only minimizes reconstruction error in an unweighted Frobenius sense.
- Real activations are highly anisotropic : some directions are used much more frequently than others.
- GlowQ introduces a covariance-aligned objective :

$$\min_{A,B} \left\| (E_{cat} - AB)\, \Sigma_x^{\frac{1}{2}} \right\|_F^2$$

- This makes the learned right subspace focus on data-preferred directions rather than rarely used directions.



(a) Eigenvalue spectrum (b) EigenValue spectrum log scaled



(c) Energy capture of Q,K,V (d) Energy capture of MLP

Randomized SVD + QR Decomposition

This yields practical advantages such as avoiding materialization of huge matrices, lower compute/memory cost.

Algorithm 1 Covariance-aligned QR reduction and randomized SVD on the core

Require: Stacked error $\mathbf{E}_{\text{cat}} \in \mathbb{R}^{m \times d}$, covariance $\mathbf{\Sigma}_{\mathbf{x}} \succeq 0$, target rank r, oversampling p, power iters q

Ensure: Low-rank factors $(\mathbf{A}^\star, \mathbf{B}^\star)$ for the covariance-aligned objective

1: **Thin QR of $\mathbf{E}_{\text{cat}}$**: compute $\mathbf{Q}_e \mathbf{R}_e = \mathbf{E}_{\text{cat}}$ with $\mathbf{Q}_e^\top \mathbf{Q}_e = \mathbf{I}_d$

2: **Core construction**: set $\mathbf{M} \leftarrow \mathbf{R}_e \, \mathbf{\Sigma}_{\mathbf{x}}^{1/2} \in \mathbb{R}^{d \times d}$

3: **Random sketch / range finding**: draw $\mathbf{\Omega} \sim \mathcal{N}(0,1)^{d \times (r+p)}$, set $\mathbf{Y} \leftarrow \mathbf{M}\mathbf{\Omega}$; optionally do q power steps $\mathbf{Y} \leftarrow \mathbf{M}(\mathbf{M}^\top \mathbf{Y})$

4: **Orthonormalize**: $\mathbf{Q} \leftarrow \text{orth}(\mathbf{Y}) \in \mathbb{R}^{d \times (r+p)}$

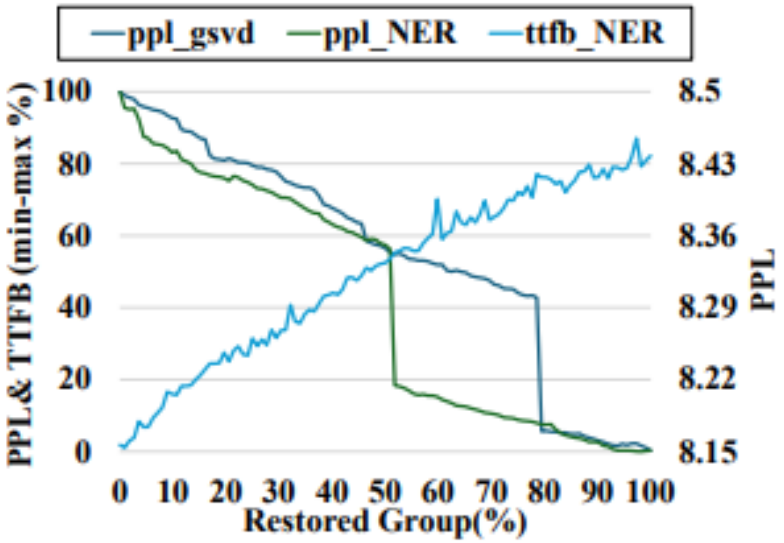
5: **Compressed SVD**: $\mathbf{B}_{\text{small}} \leftarrow \mathbf{Q}^\top \mathbf{M}$; compute $\mathbf{B}_{\text{small}} = \tilde{\mathbf{U}} \, \mathbf{\Sigma} \, \mathbf{V}^\top$

6: **Lift left factor**: $\mathbf{U} \leftarrow \mathbf{Q} \, \tilde{\mathbf{U}}$

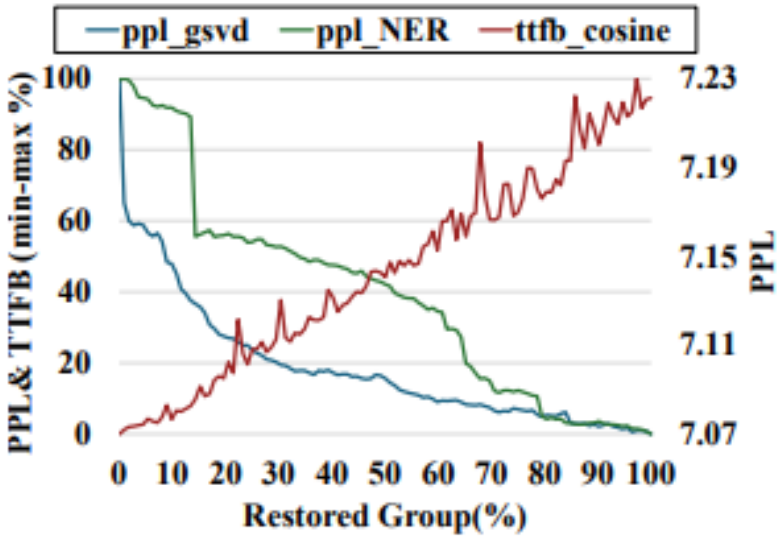
7: **Truncate (top-r) & balance**: keep $(\mathbf{U}_r, \mathbf{\Sigma}_r, \mathbf{V}_r)$ and set $\widehat{\mathbf{A}}^\star \leftarrow \mathbf{U}_r \, \mathbf{\Sigma}_r^{1/2}$, $\widehat{\mathbf{B}}^\star \leftarrow \mathbf{\Sigma}_r^{1/2} \mathbf{V}_r^\top$

8: **Lift to original variables**: $\mathbf{A}^\star \leftarrow \mathbf{Q}_e \, \widehat{\mathbf{A}}^\star$, $\mathbf{B}^\star \leftarrow \widehat{\mathbf{B}}^\star \, \mathbf{\Sigma}_{\mathbf{x}}^{-1/2}$ ▷ use a pseudoinverse if $\mathbf{\Sigma}_{\mathbf{x}}$ is singular

GlowQ-S Strategy



(a) LLaMA 3.2-3B



(b) Qwen 2.5-7B

- Not all groups or layers need restoration.
- GlowQ-S activates only the most beneficial units under a latency or memory budget.
- LLaMA models rank groups with normalized error ratio; Qwen models rank groups with SVD energy-capture score.

Experiments 1

Method	Q config	LLaMA 2		LLaMA 3		Qwen 2.5		Qwen 3		Mistral	OPT	
		7B	13B	3.2-3B	3.1-8B	7B	14B	8B	14B	7B	1.3B	6.7B
FP16	-	5.48	4.90	7.81	6.24	6.86	5.29	9.73	8.64	5.32	14.62	10.85
BnB	NF4	5.64	4.97	8.29	6.66	7.10	**5.64**	9.97	8.88	5.51	15.16	**10.94**
AWQ	INT4, g128	5.61	4.97	8.24	6.64	7.11	6.17	10.19	9.00	5.51	15.22	11.23
GPTQ	INT4, g128	5.65	5.35	9.46	6.63	7.11	5.75	9.98	8.90	5.51	15.00	11.07
ZeroQuant-V2	INT4, g128	5.72	4.99	8.44	6.79	8.41	5.75	10.19	9.04	5.53	15.10	11.14
QERA	INT4, g128	5.61	4.98	8.22	6.64	8.09	**5.64**	10.07	8.85	5.48	14.85	11.00
L2QER	INT4, g128	5.68	**4.94**	8.30	6.75	8.14	5.66	10.07	8.85	5.46	15.30	11.16
GlowQ	INT4, g128	**5.58**	4.96	**8.16**	**6.59**	**7.07**	**5.64**	**9.90**	**8.80**	**5.42**	**14.84**	11.00
GlowQ-S	INT4, g128	5.60	4.96	8.22	6.62	7.09	5.68	9.97	8.89	5.45	15.00	11.00
L2QER	W4A4	**5.90**	5.18	9.42	7.65	9.11	**6.52**	10.76	9.36	**5.73**	27.40	11.32
L2QER	W4A8	5.69	4.95	8.31	6.76	8.15	**5.67**	10.11	8.86	5.47	14.90	11.00
GlowQ	W4A4	**5.90**	**5.20**	**9.21**	**7.42**	**8.03**	6.55	**10.66**	**9.33**	5.74	**26.35**	**11.31**
GlowQ-S	W4A4	5.92	5.20	9.25	7.45	8.05	6.61	10.72	9.37	5.79	27.42	11.33
GlowQ	W4A8	**5.59**	**4.97**	**8.20**	**6.63**	**7.12**	5.71	**10.08**	**8.85**	**5.43**	**14.85**	**10.97**
GlowQ-S	W4A8	5.60	4.97	8.24	6.64	7.13	5.77	10.10	8.92	5.48	14.99	10.99

Method	Rank	LLaMA 3.2-3B		LLaMA 3.1-8B		Qwen 3-8B		Qwen 3-14B	
		Acc (↑)	C4 (↓)	Acc (↑)	C4 (↓)	Acc (↑)	C4 (↓)	Acc (↑)	C4 (↓)
FP16	-	67.14	10.30	73.29	9.00	71.48	14.52	74.10	13.08
ZeroQuant-V2	64	65.38	11.45	**73.48**	9.87	70.19	15.00	72.62	13.79
QERA		65.48	11.04	72.86	9.68	69.86	14.78	73.14	13.29
L2QER		66.19	11.04	72.43	9.63	69.52	14.82	73.24	13.80
GlowQ		**66.90**	**10.98**	73.33	**9.59**	**70.71**	**14.60**	**73.84**	**13.26**
GlowQ-S		66.33	11.07	72.62	9.78	70.29	14.77	73.24	13.48

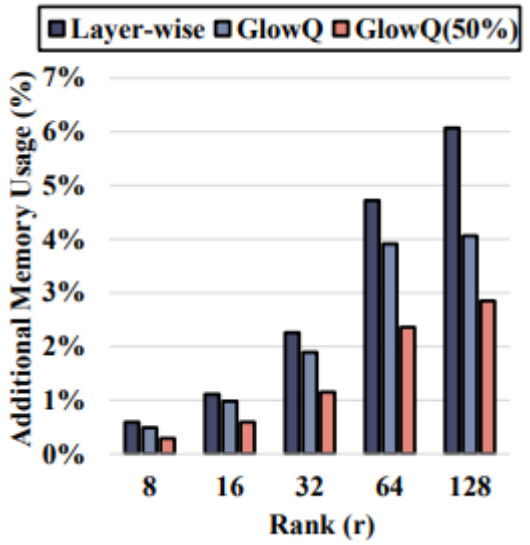
- Main metrics :
  - Wikitext-2 perplexity
  - C4 perplexity
  - Downstream accuracy
- The paper presents these gains as average improvements over strong baselines.

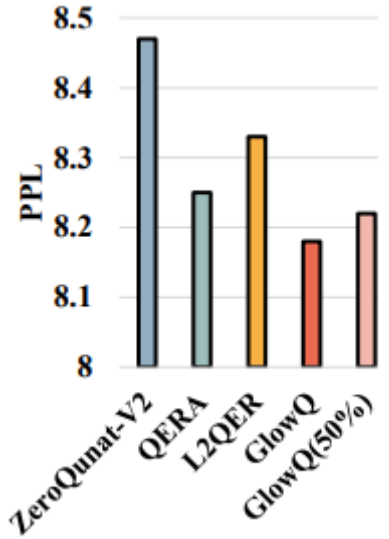
Experiments 2

Models		Setting	TTFB(ms) ↓	tok/s ↑	Prefill(ms) ↓	Dec(ms/tok) ↓
LLaMA 2	7B	Layerwise	88.45	15.66	95.13	63.17
		GlowQ	82.66	17.12	92.23	58.32
		GlowQ-S	66.68	21.16	72.35	45.90
	13B	Layerwise	128.70	11.22	141.76	85.91
		GlowQ	122.78	12.33	136.53	81.15
		GlowQ-S	100.17	15.68	112.09	62.98
		Avg. Δ BX (%)	**-5.57**	**+9.61**	**-3.37**	**-6.61**
		Avg. Δ R50 (%)	**-23.39**	**+37.44**	**-22.44**	**-27.01**

- It improves not only quality but also inference efficiency.
- GlowQ-S maintains accuracy within 0.2 percentage points on average while achieving much larger runtime gains.

- GlowQ reduces memory overhead and delivers better performance than competing baselines under the same memory budget.





Conclusion

- GlowQ introduces a group-shared low-rank approximation for quantized LLMs.
- It shares one right factor across input-sharing modules, reducing redundant computation and memory overhead.
- A covariance-aligned objective improves restoration quality by focusing on data-preferred directions.
- GlowQ-S further improves efficiency by selectively restoring only the most beneficial groups.
- Overall, GlowQ achieves a better trade-off between accuracy, latency, throughput, and memory usage.


Thank you