# LatentLLM: Attention-Aware Joint Tensor Compression

**Research paper discussion | 22 main frames + 6 backup frames**

Reading presentation by Codex

Paper: Toshiaki Koike-Akino, Xiangyu Chen, Jing Liu,
Ye Wang, Pu (Perry) Wang, Matthew Brand (MERL)

arXiv:2505.18413v1 | Read on 24 September 2026

Preserve useful computations through statistical weighting, compact factors, and joint objectives.

# The reported gain is a better quality–compression tradeoff

| OPT-6.7B: WT2 perplexity ↓ | |
|---|---|
| Original | 10.9 |
| Local RootCov | 13.5 |
| **LatentLLM** | **11.5** |

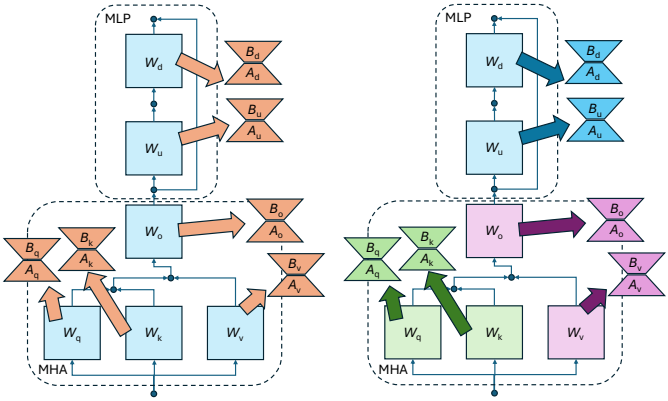
| LLaVA: ScienceQA accuracy ↑ | |
|---|---|
| Original | 70.03% |
| Local RootCov | 57.53% |
| **LatentLLM** | **63.85%** |

Both compressed rows use the paper's 20% compression setting.

- Calibration and factorization; no end-to-end fine-tuning.
- Improvement over the tested local low-rank baselines; some quality loss remains.

Paper, Table 2, PDF p.6; Table 4, PDF p.10. "20%" is the nominal setting.

# Compress interactions between adjacent projections




- Q/K: preserve attention scores.
- V/O: approximate the composed output map.
- U/D: coordinate both MLP layers.

"Global" here concerns coupled modules; no whole-network optimum is established.

Paper, Fig.1, Secs.3.1 and 4, PDF pp.2–5. Author figure reused.

# RootCov solves the local activation reconstruction problem

$$W \in \mathbb{R}^{d' \times d}, \quad X \in \mathbb{R}^{d \times l}, \quad \widehat{W} = BA, \quad \operatorname{rank}(BA) \le r$$

$$\mathcal{L}_1 = \mathbb{E}\,\|(W - BA)X\|_F^2 = \left\|(W - BA)C^{1/2}\right\|_F^2, \quad C = \mathbb{E}[XX^\top].$$

**A weighted SVD is enough for this objective**

$$BAC^{1/2} = \operatorname{svd}_r(WC^{1/2}).$$

- $C$ is an uncentered second moment; finite data use damping.
- $P = C$ weights the error by $C^2$; $P = C^{1/2}$ matches the stated loss.
- Optimality is local, with fixed rank and fixed statistics.

Paper, Eqs.(1)–(5), Table 1, PDF pp.2–3; Appendix B, pp.13–15.

# An identity block removes redundant factor parameters

$$USV = \mathrm{svd}_r(WP), \qquad B = USJ, \quad A = J^{+}VP^{+}.$$

$$VP^{+} = [\, V_1 \; V_2 \,], \quad J = V_1 \quad \Longrightarrow \quad A = [\, I_r \; V_1^{-1} V_2 \,].$$

## Same low-rank product; fewer stored coefficients

$$r(d + d') \quad \longrightarrow \quad r(d + d') - r^2.$$

- Requires an invertible block; pivot columns if needed.
- At fixed rank, the product and its error are unchanged.
- Practical savings require an implementation that omits the identity block.

Paper, Eqs.(6)–(9), Remarks 4–5, PDF pp.3–4. Numerical conditioning is a reader concern.

# Latent dimension, parameter budget, and FLOPs differ

For a single square $d \times d$ layer, let $\rho = r/d$.

$$\text{Dense-factor retention} = 2\rho, \qquad \text{Identity-block retention} = 2\rho - \rho^2.$$

| Latent retention | Dense-factor parameters | Identity-block parameters |
|---|---|---|
| 75% | 150% | 93.75% |
| 50% | 100% | 75% |

- A 25% latent reduction yields only a 6.25% parameter reduction in the first example.
- OPT-6.7B at nominal 20%: $6.66\mathrm{B} \to 5.53\mathrm{B}$ parameters, a 16.97% reduction.

Single-layer illustration from Sec.3.3, PDF p.3. Whole-model counts: Table 3, p.7. Percentages calculated.

# Joint QK preserves the scores that determine attention

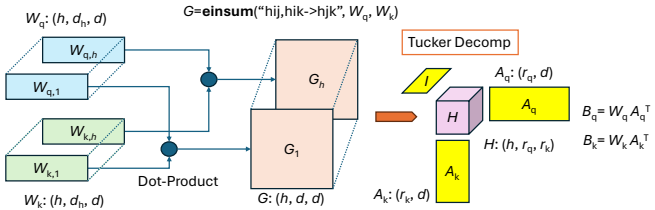
$$M_i = X^\top W_{q,i}^\top W_{k,i} X \in \mathbb{R}^{l \times l}, \qquad W_{q,i}, W_{k,i} \in \mathbb{R}^{d_h \times d}.$$

$$\widehat{M}_i = X^\top A_q^\top B_{q,i}^\top B_{k,i} A_k X, \qquad \mathcal{L}_2 = \sum_{i=1}^{h} \left\| M_i - \widehat{M}_i \right\|_F^2.$$

- $A_q \in \mathbb{R}^{r_q \times d}$ and $A_k \in \mathbb{R}^{r_k \times d}$ are shared across heads.
- $B_{q,i}, B_{k,i}$ are head-specific decompression factors.
- The target is pre-softmax scores; it does not directly optimize task loss.

Paper, Eqs.(10)–(12), PDF p.4. The paper omits the usual score scale in this derivation.

# The score objective becomes a Tucker approximation




$$G_i = C^{1/2} W_{q,i}^\top W_{k,i} C^{1/2}, \quad Q = A_q C^{1/2}, \quad K = A_k C^{1/2}$$

$$\mathcal{L}_2 = \sum_i \left\| G_i - Q^\top H_i K \right\|_F^2, \qquad H_i = B_{q,i}^\top B_{k,i}.$$

Two input modes are compressed; the head mode stays intact.

$Q, K$ denote weighted coordinates on this slide; the author diagram omits preconditioning.

Paper, Fig.3 and Eq.(13), PDF p.4. Author diagram reused.

# Alternating subspace updates give a practical solver

With row-orthonormal $Q, K$, the optimal core is $H_i = QG_iK^\top$.

$$K \leftarrow \mathrm{TopEigRows}_{r_k}\left(\sum_i G_i^\top Q^\top Q G_i\right),$$

$$Q \leftarrow \mathrm{TopEigRows}_{r_q}\left(\sum_i G_i K^\top K G_i^\top\right).$$

- Initialize from a marginal Gram matrix; alternate, then undo weighting.
- Each step solves a conditional problem; a joint global optimum is not guaranteed.
- Experiments use 8 QK iterations.

Paper, Algorithm 1, p.5; Appendix E, pp.18–20; iteration count, p.8.

Reader notation: TopEigRows returns leading eigenvectors as rows. Q update follows the dimensionally consistent Appendix E; Algorithm 1 has a transpose error.

# The simple joint VO objective misses attention weighting

Let $S_i$ be the normalized attention weights, with a consistent orientation.

$$Y' = \sum_i W_{o,i} W_{v,i} X S_i.$$

The simple proxy instead minimizes

$$\mathcal{L}_3 = \sum_i \| W_{o,i} W_{v,i} X - B_o A_{o,i} B_{v,i} A_v X \|_F^2 .$$

## The paper reports a negative result

This joint VO approximation was not effective relative to separate V/O compression (Remark 11).

Appendix G explores actual attention-weighted statistics, but provides no separate end-to-end gain table for those alternatives.

Paper, Sec.4.2, Eqs.(14)–(18), Remark 11, PDF p.5; Appendix G, pp.28–31.

# Auxiliary variables coordinate both MLP projections

$$Z = W_u X, \quad Z' = \sigma(Z), \quad Y = W_d Z'$$

$$\mathcal{L}_4 = \alpha \left\| W_u X - Z \right\|_F^2 + \beta \left\| Z' - \sigma(Z) \right\|_F^2 + \gamma \left\| W_d Z' - Y \right\|_F^2 .$$

1. Update auxiliary activations using the fixed calibration pair $(X, Y)$.
2. Fit effective maps $ZX^+$ and $YZ'^+$ with weighted low-rank SVD.
3. Alternate the updates; the experiments use 4 UD iterations.

The closed-form nonlinear update is derived for ReLU. Handling SiLU, GELU, or gated MLPs needs additional detail.

Paper, Sec.4.3, Eqs.(19)–(22), PDF pp.5–7; Appendix H, pp.31–32. Iterations: p.8.

# Extensions have different levels of validation

| Extension | Mechanism | Evidence boundary |
|---|---|---|
| Bias | Center activations; update bias | Derivations for linear, QK, VO |
| GQA | Sum scores over KV groups and query heads | Derivation; no main GQA benchmark |
| RoPE | Include relative-position rotations | 10-token proxy-loss experiment |
| Quantization | STE on low-rank factors | Proposal; no main bit-width table |

A listed architecture or derived extension is not an end-to-end validation result.

Paper, Appendices B.2, E.2–E.3, F, I.1, J; PDF pp.14–15, 21–28, 34–37.

# Calibration is small, but its data source matters

## Language evaluation

- OPT: 125M, 350M, 1.3B, 2.7B, 6.7B, 13B.
- 64 C4 segments $\times$ 2048 tokens for calibration.
- WT2, PTB, C4 perplexity.

## Multimodal evaluation

- LLaVA-v1.6-Vicuna-7B.
- 64 ScienceQA **training** examples for calibration.
- Overall and subgroup accuracy.

- A40 GPUs; PyTorch and HuggingFace Transformers.
- "No fine-tuning" does not imply "no task-specific calibration."
- FLOPs are profiled separately at 128 tokens.

Paper, Sec.5, PDF pp.7–10. Evaluation details and run-to-run variation are incompletely reported.

# RootCov is the most informative nearby baseline

OPT-6.7B, nominal 20% compression, WT2 perplexity ↓:

| Method | Perplexity |
|---|---|
| Plain SVD | 67517.7 |
| ASVD (Hessian) | 17.3 |
| ASVD ($\ell_2$) | 14.6 |
| ASVD (Cov) | 9842.6 |
| ASVD (RootCov) | 13.5 |
| **LatentLLM (RootCov)** | **11.5** |

The final gap also includes structural savings and joint updates.
It cannot be assigned entirely to the QK objective from this table.

Paper, Table 2, PDF p.6. Baselines are local low-rank variants, not full quantization/pruning systems.

# Quality degrades as compression becomes aggressive

OPT-6.7B: WT2 / PTB / C4 perplexity ↓

| Nominal compression | Local RootCov | LatentLLM |
|---|---|---|
| Original | — | 10.9 / 15.8 / 12.7 |
| 10% | 11.8 / 17.7 / 14.2 | 10.7 / 16.1 / 13.0 |
| 20% | 13.5 / 19.5 / 15.4 | 11.5 / 17.4 / 13.7 |
| 30% | 17.0 / 23.9 / 17.8 | 13.5 / 19.2 / 15.3 |
| 40% | 27.2 / 36.1 / 24.0 | 18.0 / 24.2 / 18.4 |

- At 20%: WT2 is 14.81% lower than local RootCov, but 5.50% above the original.
- At 10%: the small WT2 improvement does not extend to PTB or C4.

Paper, Table 2, PDF p.6; percentages calculated from displayed values. No significance claim.

# The OPT trend spans six evaluated model sizes

WT2 perplexity at nominal 30% compression ↓:

| OPT model | Original | Local RootCov | LatentLLM |
|---|---|---|---|
| 125M | 27.7 | 88.8 | 43.4 |
| 350M | 22.0 | 71.6 | 32.9 |
| 1.3B | 14.6 | 33.9 | 20.2 |
| 2.7B | 12.5 | 24.1 | 16.5 |
| 6.7B | 10.9 | 17.0 | 13.5 |
| 13B | 10.1 | 14.3 | 12.0 |

Larger OPT models tolerate this setting better in these data.
The configuration tables include larger models; the main results stop at 13B.

Paper, Table 2, PDF p.6; Figs.4–5, pp.8–9. This is an empirical family-specific trend.

# Fewer counted FLOPs do not establish runtime speedup

OPT-6.7B, `calflops`, 128-token input:

| Nominal compression | FLOPs | MACs | Parameters |
|---|---|---|---|
| 0% | 1.70T | 851G | 6.66B |
| 20% | 1.36T | 681G | 5.53B |
| 30% | 1.19T | 596G | 4.87B |
| 40% | 1.02T | 511G | 4.20B |
| 50% | 851G | 425G | 3.54B |

- At 50%: about half the FLOPs and 46.85% fewer parameters.
- No reported latency, throughput, peak-memory, or energy benchmark.
- Real savings depend on factor layout, contraction order, and kernels.

Paper, Table 3 and computational-complexity paragraph, PDF p.7. Parameter reduction calculated.

# ScienceQA gains over RootCov coexist with absolute losses

LLaVA-v1.6-Vicuna-7B: overall accuracy ↑

| Compression | RootCov | LatentLLM | Gain (pp) | Loss vs. original (pp) |
|---|---|---|---|---|
| 0% | — | 70.03% | — | — |
| 10% | 60.67% | 65.76% | +5.09 | 4.27 |
| 20% | 57.53% | 63.85% | +6.32 | 6.18 |
| 30% | 54.37% | 60.13% | +5.76 | 9.90 |
| 40% | 52.23% | 54.59% | +2.36 | 15.44 |
| 50% | 49.30% | 52.25% | +2.95 | 17.78 |

pp = percentage points. The 20% gain is 6.32 pp, or 10.99% relative to RootCov. Calibration uses ScienceQA training examples.

Paper, Table 4, PDF p.10; calibration, p.7. Differences calculated.

# Subgroup accuracy clarifies what remains after compression

ScienceQA at nominal 20% compression:

| Subgroup | Original | LatentLLM | Loss (pp) |
|---|---|---|---|
| Natural science | 72.47 | 66.39 | 6.08 |
| Social science | 69.18 | 61.19 | 7.99 |
| Language science | 65.73 | 60.82 | 4.91 |
| Image context | 68.82 | 63.41 | 5.41 |
| No context | 65.99 | 60.62 | 5.37 |
| Grades 1–6 | 72.72 | 66.41 | 6.31 |
| Grades 7–12 | 65.19 | 59.26 | 5.93 |

Accuracy is in percent. Groups describe different, potentially overlapping attributes; their simple mean is not the overall score.

Paper, Table 4 and Fig.6, PDF p.10. Subgroup counts and full aggregation details are not reported.

# Mechanism evidence is useful, but not a full ablation

- **RootCov:** random-matrix experiments support the statistical weighting (Fig.7).
- **Joint QK:** attention-score error improves over activation-only fitting on synthetic inputs (Fig.10).
- **Joint QKV and split heads:** sharing and objective choice have different tradeoffs (Figs.8–9).
- **Joint VO:** the simple proxy has no reported advantage (Remark 11).

**Missing separation**

The main tables do not isolate RootCov, identity blocks, QK, and UD under fully matched parameter/rank budgets.

Paper, Remark 11, p.5; Tables 2 and 4, pp.6 and 10; Figs.7–10, pp.15–17 and 21.

# The next validation should target deployment and robustness

1. **Optimization:** quantify the gap from score reconstruction to full-model quality.
2. **Robustness:** vary calibration data, damping, conditioning, and seeds.
3. **Architecture:** verify RoPE, GQA, and non-ReLU/gated MLP handling on real models.
4. **Systems:** measure KV bytes, prefill/decode latency, throughput, and compression cost.

Reader assessment: the current evidence is strongest for the reported quality–compression tradeoffs.

Evidence boundaries from Secs.4–5 and Appendices E–J. These validation priorities are this reading's analysis.

# Three ideas to retain from LatentLLM

## Metric

Weight errors by the activation statistics that determine their effect.

## Representation

Use factorization freedom to remove redundant parameters.

## Objective

Approximate coupled computations when local projections are insufficient.

Primary source: Koike-Akino et al., LatentLLM: Attention-Aware Joint Tensor Compression, arXiv:2505.18413v1, 23 May 2025.

arxiv.org/abs/2505.18413v1. Reading uses the matching author TeX and 37-page PDF; no model experiments were independently reproduced.

# Backup: why the alternating QK updates follow

For row-orthonormal $Q, K$, define $H_i = QG_iK^\top$.

$$\begin{aligned}\mathcal{L}_2 &= \sum_i \left\| G_i - Q^\top H_i K \right\|_F^2 \\ &= \sum_i \left( \|G_i\|_F^2 - \left\| QG_iK^\top \right\|_F^2 \right).\end{aligned}$$

With $K$ fixed, minimizing the loss means maximizing

$$\operatorname{tr}\left[ Q \left( \sum_i G_i K^\top K G_i^\top \right) Q^\top \right], \qquad QQ^\top = I_{r_q}.$$

The leading eigenspace solves this conditional problem. The symmetric argument updates $K$; coupling still makes the joint problem nonconvex.

Paper, Appendix E, PDF pp.18–20. Q and K are this presentation's weighted-coordinate notation.

# Backup: bias changes the statistics to be preserved

$$Y = WX + b\mathbf{1}^\top, \qquad \mu = X\mathbf{1}/I.$$

For fixed $BA$, the optimal bias is

$$\widehat{b} = b + (W - BA)\mu.$$

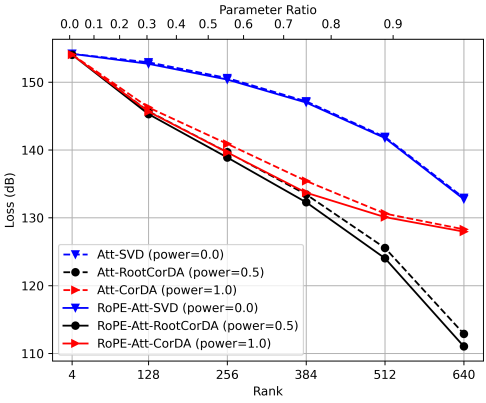
Substitution leaves a centered reconstruction objective:

$$\left\|(W - BA)(X - \mu\mathbf{1}^\top)\right\|_F^2 = \left\|(W - BA)C_0^{1/2}\right\|_F^2.$$

- $C_0 = (X - \mu\mathbf{1}^\top)(X - \mu\mathbf{1}^\top)^\top$ here is unnormalized.
- Joint QK bias requires an augmented, structurally constrained formulation.

Paper, Appendix B.2, Eqs.(45)–(49), PDF p.15; Appendix E.2, pp.21–25.

# Backup: RoPE evidence uses a short synthetic window




- Rank on the x-axis; loss in dB on the y-axis, lower is better.
- 10-token window, $d = 768$, RoPE base $10^4$.
- Authors report roughly 1–2 dB additional improvement.
- Derivation assumes independent tokens; no long-context benchmark follows.

Paper, Appendix F.3, Fig.12, PDF pp.27–28. Author figure reused; dB is not percent accuracy.

# Backup: UD updates combine regression and weighted SVD

Fixed $Z, W_d$ give

$$Z' = (\gamma W_d^\top W_d + \beta I)^+ \left(\beta\sigma(Z) + \gamma W_d^\top Y\right).$$

Then approximate the effective maps using

$$\mathrm{svd}_{r_u}\left(ZX^+ C^{1/2}\right), \qquad \mathrm{svd}_{r_d}\left(YZ'^+ C_d^{1/2}\right), \quad C_d = Z'Z'^\top.$$

- ReLU permits piecewise candidates for $Z$; branch constraints must be respected.
- $X, Y$ remain the calibration pair; auxiliary variables and factors are updated.
- All penalty weights and non-ReLU implementation details are not specified.

Paper, Sec.4.3, PDF pp.5–7; Appendix H, pp.31–32. Bias-free notation for clarity.

# Backup: sparse and quantized extensions remain exploratory

$$\widehat{W} = BA + D, \qquad \|D\|_0 \le \kappa.$$

- Alternate a low-rank fit with sparse approximation of the residual.
- Synthetic tests favor some sparse-only solutions over low-rank plus sparse ones (Figs.13–16).
- Nonzero indices and execution kernels still determine storage and speed.
- STE can put quantized $A, B$ in the forward pass; no real-model bit-width sweep is reported.

These are extension directions and limited proxy-loss results, not a validated combined compression system.

Paper, Appendix I–I.1, PDF pp.32–36. Here $\|D\|_0$ counts nonzeros, rather than a matrix norm.

# Backup: reproducibility requires resolving notation issues

- **Algorithm 1:** the Q update needs $K^\top K$ for row-shaped $K$; the printed $KK^\top$ is dimensionally inconsistent.
- **Appendix B, Eq.(42):** an extra singular-value factor conflicts with the main-text factorization.
- **VO:** fix the softmax orientation and re-count head-dependent attention multiplications.
- **Table 2:** OPT-125M repeats PTB/C4 values at 30% and 40%; the source does not explain the repetition.

The reading retains reported data, labels its corrections, and uses the consistent main/appendix derivations. It is not an author erratum or an independent reproduction.

Paper, Algorithm 1, p.5; Eq.(42), p.14; Eqs.(14)–(18), p.5; Table 2, p.6.