ONE SUBSPACE, MANY FUNCTIONS: THE GEOMETRY OF MULTI-HEAD LATENT ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-head Latent Attention (MLA) routes key and value projections through a shared low-rank bottleneck. A natural question is whether this creates functional redundancy between the two pathways — and whether that redundancy can be exploited for compression. We show it does not: the bottleneck forces keys and values into the same representational coordinate system, but the model uses that system to encode maximally diverse, non-interchangeable representations. We prove this via the Eckart–Young–Mirsky theorem applied to gauge-invariant effective operators (the raw factor matrices are non-identifiable under latent reparametrisation and cannot be compared directly). Auditing DeepSeek-V2-Lite, V2, and V3, we find universal middle-layer subspace collapse with a scale-dependent U-shaped profile emerging clearly only at 671B. Bidirectional causal interventions confirm that despite near-zero Gram distance in collapsed layers, forcing $K = V$ causes $\Delta\text{PPL} \approx +10$ in both directions, while a $2.9\times$ asymmetry reveals that key-pathway corruption is more catastrophic than value-pathway corruption.

1 INTRODUCTION

The KV cache, which stores key-value activations for every token during autoregressive decoding, limits batch sizes, increases latency, and dominates the cost of long-context generation. Multi-head Latent Attention (MLA) addresses this by projecting keys and values into a shared low-dimensional latent space ($r \ll d$), achieving aggressive compression with small quality loss. However, whether this shared bottleneck creates functional redundancy between the two pathways remains open.

The question is sharpened by evidence that training with a strict $K = V$ identity mapping degrades perplexity by only $\sim$3% while halving the KV cache Kayyam et al. (2026), suggesting implicit pressure toward geometric proximity. MLA's shared bottleneck makes this pressure explicit—but is it strong enough to create functional redundancy, or does the model exploit the shared geometry to encode maximally diverse representations? A central obstacle is gauge freedom: for any invertible $R \in \mathbb{R}^{r\times r}$, substituting $W_c \mapsto W_c R$, $W_\cdot^{\uparrow} \mapsto R^{-1} W_\cdot^{\uparrow}$ leaves the effective operators unchanged, so raw factor matrix comparisons are uninformative. The true geometric relationship is only visible at the level of $W_c W_K^{\uparrow}$ and $W_c W_V^{\uparrow}$.

We proceed in three stages: we *prove* a shared subspace constraint on the effective operators, *audit* it empirically in production checkpoints with gauge-invariant metrics, and *intervene* causally to test whether geometric collapse implies functional redundancy. Our contributions are:

- **Theoretical Framework.** Via an instantaneous surrogate Frobenius objective and the Eckart–Young–Mirsky theorem, we prove that the optimal effective operators $\widetilde{W}_K$ and $\widetilde{W}_V$ are confined to the same r-dimensional dominant singular subspace, with coupling that tightens under stronger compression.
- **Gauge-Invariant Empirical Audit.** Applying gauge-invariant metrics to DeepSeek-V2-Lite, V2, and V3, we uncover a scale-dependent subspace collapse: universal early-layer decoupling and middle-layer collapse, with late-layer divergence emerging only at 671B.
- **Causal Intervention.** Despite near-zero Gram distance in collapsed layers, forcing $K = V$ causes $\Delta\text{PPL} \approx +10$ in both clamping directions, proving that subspace collapse is structural, not functionally redundant. A $2.9\times$ directional asymmetry further reveals key-pathway primacy outside the collapse zone.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

Memory-Efficient Attention and KV Cache Compression. Multi-Query Attention Shazeer (2019) and Grouped-Query Attention Ainslie et al. (2023) compress the KV cache by sharing heads across queries, but their compression factor is bounded by the hidden dimension of the retained heads. MLA Liu et al. (2024a;b) breaks this bound by projecting keys and values into a shared latent bottleneck ($r \ll d$). The geometric relationship between the resulting effective operators has not previously been characterised; a gauge-invariant methodology is required because naive inspection of the raw factor matrices is uninformative.

Subsequent work has sought to migrate GQA models to MLA post-hoc. TransMLA Meng et al. (2025) uses joint PCA on K and V matrices; MHA2MLA Ji et al. (2025) applies joint SVD. Both implicitly assume subspace proximity implies sufficient functional redundancy to justify compression. Our causal intervention results suggest this assumption warrants scrutiny.

Implicit $K = V$ Alignment and Parameter Redundancy. In Kayyam et al. (2026), authors showed that enforcing $K = V$ during training incurs only $\sim$3% perplexity degradation at up to 1.2B parameters. A critical distinction separates that finding from the present work: training-time enforcement allows the model to learn representations that never develop the within-subspace K–V rotation, whereas post-hoc clamping of a fully trained MLA model destroys a rotation that was learned and is load-bearing ($\Delta\text{PPL} = +10.1$). Concurrently, Shaikh et al. (2026) find that key projection matrices develop greater intra-layer subspace alignment than value matrices during pretraining, and that key compression is less disruptive — consistent with the routing–extraction asymmetry we document in Section E.5.

Geometry and Low-Rank Bottlenecks. Foundational results show that gradient descent on overparameterised matrix factorisations exhibits implicit regularisation toward low-rank, aligned singular subspaces Arora et al. (2019); Gunasekar et al. (2017), and that deep linear network solutions concentrate near low-rank manifolds Saxe et al. (2019). Our work maps these properties onto the specific MLA architecture, formalising the subspace alignment constraints that arise when two projection targets share a common low-rank factor. The key new contribution relative to prior theory is the identification of gauge freedom as an obstacle to empirical analysis, and the structural–functional distinction: the shared subspace constraint establishes geometric coupling at the level of representational capacity, not functional equivalence.

3 Theoretical Framework

We establish a geometric and algebraic framework for analysing the structural constraints imposed by Multi-head Latent Attention (MLA) on the Key-Value (KV) pathway. We show that routing keys and values through a shared low-dimensional bottleneck ($r \ll d$) constrains the *effective* key and value transformations to reside within a common low-rank representational subspace. Importantly, this constraint applies to the effective operators rather than the individual factor matrices, which remain non-identifiable due to gauge freedom—a subtlety that has direct implications for how empirical validation of this result should be conducted (see Remark 2).

3.1 Unified Latent Attention Formulation

Let $X \in \mathbb{R}^{n \times d}$ denote the input sequence representation, where n is the sequence length and d is the hidden dimension.

In standard Multi-Head Attention (MHA), keys and values are generated independently:

$$K = XW_K, \qquad V = XW_V,$$

with $W_K, W_V \in \mathbb{R}^{d \times d}$.

In MLA, the KV pathway is compressed through a shared latent bottleneck of dimension $r \ll d$. Let $W_c \in \mathbb{R}^{d \times r}$ be the compression matrix and

$$c = XW_c \in \mathbb{R}^{n \times r}$$

the latent representation. Ignoring decoupled positional embeddings (RoPE) to isolate the latent geometry, keys and values are reconstructed via separate up-projection matrices $W_K^{\uparrow}, W_V^{\uparrow} \in \mathbb{R}^{r \times d}$,

yielding

$$K = cW_K^{\uparrow} = XW_cW_K^{\uparrow}, \qquad V = cW_V^{\uparrow} = XW_cW_V^{\uparrow}.$$

The attention matrix and output are

$$A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \qquad O = AV.$$

Substituting the MLA factorisation,

$$A = \mathrm{softmax}\left(\frac{Q\big(XW_cW_K^{\uparrow}\big)^\top}{\sqrt{d}}\right), \qquad O = A\big(XW_cW_V^{\uparrow}\big).$$

3.2 Gauge Freedom and the Non-Identifiability of Factor Matrices

Before stating the main results, we highlight a structural non-identifiability that is essential both for interpreting the theorems and for designing empirical tests.

For any invertible $R \in \mathbb{R}^{r\times r}$, the effective key and value operators are invariant under the reparametrisation

$$W_c \longmapsto W_cR, \qquad W_K^{\uparrow} \longmapsto R^{-1}W_K^{\uparrow}, \qquad W_V^{\uparrow} \longmapsto R^{-1}W_V^{\uparrow},$$

since $(W_cR)(R^{-1}W_K^{\uparrow}) = W_cW_K^{\uparrow}$, and identically for $W_V^{\uparrow}$. Consequently, the individual factor matrices are not uniquely determined by the model's input–output behaviour, and comparing them directly is not meaningful.

The gauge-invariant objects are the **effective operators**

$$\widetilde{W}_K \;=\; W_cW_K^{\uparrow} \in \mathbb{R}^{d\times d}, \qquad \widetilde{W}_V \;=\; W_cW_V^{\uparrow} \in \mathbb{R}^{d\times d},$$

and their joint column concatenation

$$\widetilde{M} \;=\; \big[\widetilde{W}_K,\, \widetilde{W}_V\big] \;=\; W_c\big[W_K^{\uparrow},\, W_V^{\uparrow}\big] \;\in\; \mathbb{R}^{d\times 2d}.$$

Remark 1 (Rank of the Joint Effective Operator)**.** *Since left-multiplication by $W_c \in \mathbb{R}^{d\times r}$ cannot increase rank beyond r,*

$$\mathrm{rank}(\widetilde{M}) \;\le\; \mathrm{rank}(W_c) \;\le\; r.$$

Regardless of how $W_K^{\uparrow}$ and $W_V^{\uparrow}$ are chosen, the effective joint operator is confined to an at-most r-dimensional image in $\mathbb{R}^d$. All subsequent results depend on this constraint.

Remark 2 (Implications for Empirical Validation)**.** *Empirical tests of K–V coupling* must *be conducted on the effective operators. Measuring $\|W_K^{\uparrow} - W_V^{\uparrow}\|_F$ on trained weights is uninformative: the gauge transformation above can make this quantity arbitrarily large or small without changing the model's behaviour. The gauge-invariant quantities are the Frobenius distance $\|\widetilde{W}_K - \widetilde{W}_V\|_F$ and the principal angles between $\mathrm{Col}(\widetilde{W}_K)$ and $\mathrm{Col}(\widetilde{W}_V)$. More broadly, any prior empirical comparison of MLA factor matrices that operates on $W_K^{\uparrow}$ or $W_V^{\uparrow}$ directly — including similarity measures, singular value comparisons, or alignment metrics applied to the raw up-projections — reports a gauge-dependent quantity. Such comparisons are not statements about the model; they are statements about one particular parameterisation of it.*

3.3 Gradient Constraints on Parameter Updates

Let $\mathcal{L}$ be the training objective, $E = \frac{\partial \mathcal{L}}{\partial O} \in \mathbb{R}^{n\times d}$ the incoming gradient at the attention output, and $E_A := \frac{\partial \mathcal{L}}{\partial S} \in \mathbb{R}^{n\times n}$ the gradient with respect to the pre-softmax scores $S = QK^\top$, which absorbs the softmax Jacobian by definition. We treat $Q = XW_Q$ as fixed with respect to $W_K^{\uparrow}$ and $W_V^{\uparrow}$, which holds in standard MLA where the query projection is independent of the KV pathway (see Remark 3).

A direct calculation (see Appendix A for the full derivation) gives:

$$\frac{\partial \mathcal{L}}{\partial W_V^{\uparrow}} \;=\; c^\top \underbrace{A^\top E}_{=:\,B_V}, \qquad \frac{\partial \mathcal{L}}{\partial W_K^{\uparrow}} \;=\; c^\top \underbrace{E_A^\top Q}_{=:\,B_K},$$

where $B_V, B_K \in \mathbb{R}^{n\times d}$. Both gradients are of the form $c^\top B$ for a matrix B determined by the upstream error signal.

Remark 3 (Independence Assumption)**.** *The key gradient formula uses the fact that Q does not depend on $W_K^\uparrow$. In architectures that tie query and key projections, this assumption fails and the formula requires modification. Our analysis applies to the unmodified MLA architecture as described in DeepSeek-V2/V3.*

Theorem 1 (Gradient Parameter Restriction)**.** *Every gradient with respect to the MLA up-projection parameters $W_K^\uparrow$ and $W_V^\uparrow$ has rank at most r.*

Proof. Both gradients take the form $c^\top B$ for some $B \in \mathbb{R}^{n\times d}$. Since $\mathrm{rank}(c^\top B) \le \mathrm{rank}(c^\top) \le r$, the claim follows immediately. □

Remark 4 (Shared Bottleneck in Parameter Updates)**.** *Although the error signals differ—$B_V = A^\top E$ for values and $B_K = E_A^\top Q$ for keys—both gradients are projected through the* same *latent factor $c^\top$. At every optimisation step, updates to both $W_K^\uparrow$ and $W_V^\uparrow$ are filtered through an identical r-dimensional channel. While the latent basis itself evolves through updates to W_c, the rank-r constraint on individual parameter updates remains active throughout optimisation.*

3.4 Effective Operator Subspaces and Shared Representational Geometry

Theorem 1 characterises the *dynamics* of parameter updates. We now characterise the *geometry* of the learned effective operators.

The analysis proceeds via an instantaneous surrogate objective. The choice of a rank-constrained Frobenius residual as the surrogate is not arbitrary: for any smooth loss $\mathcal{L}$, the first-order Taylor expansion at step t reduces the representational cost of the bottleneck to exactly this form—the squared distance between the unconstrained ideal update and the nearest rank-r approximation attainable by the MLA parameterisation. The Frobenius surrogate therefore isolates the geometric pressure of the bottleneck from all other factors in the training dynamics (loss curvature, other parameter groups, optimiser state), and the resulting theorem is a statement about the *geometry of the feasible set* rather than any particular task or loss. This makes the subspace constraint architecture-general: it holds whenever two projection targets share a common low-rank factor, regardless of what is being modelled.

Surrogate Rank-Constrained Problem

At step t, the gradient of $\mathcal{L}$ with respect to the effective operators defines, via a first-order Taylor expansion, unconstrained target operators

$$W_K^* \;=\; \widetilde{W}_K - \eta\,\frac{\partial \mathcal{L}}{\partial \widetilde{W}_K}, \qquad W_V^* \;=\; \widetilde{W}_V - \eta\,\frac{\partial \mathcal{L}}{\partial \widetilde{W}_V},$$

for learning rate $\eta > 0$: the operators the model would reach in one unconstrained gradient step. Defining their joint target $M = [W_K^*, W_V^*] \in \mathbb{R}^{d\times 2d}$, the MLA parameterisation forces $\mathrm{rank}(\widetilde{M}) \le r$ (Remark 1), so the instantaneous representational cost of the bottleneck is

$$\min_{W_c,\, W_K^\uparrow,\, W_V^\uparrow} \big\| M - W_c\big[W_K^\uparrow, W_V^\uparrow\big] \big\|_F^2,$$

minimised jointly over all $W_c \in \mathbb{R}^{d\times r}$, $W_K^\uparrow, W_V^\uparrow \in \mathbb{R}^{r\times d}$.

Remark 5 (Scope of Theorem 2)**.** *Theorem 2 characterises the optimum of this surrogate at a single training step. It does not assert that trained MLA models satisfy $\mathrm{Col}(\widetilde{W}_K) = \mathrm{Col}(\widetilde{W}_V)$ exactly at convergence: the target M shifts at every step as other parameters evolve, and the true trajectory is governed by the full loss landscape. The theorem identifies the direction in which the bottleneck geometry continuously pushes the effective operators throughout training.*

Theorem 2 (Shared Effective Subspace Constraint)**.** *Assume $M = [W_K^*, W_V^*] \in \mathbb{R}^{d\times 2d}$ satisfies $\sigma_r(M) > 0$ and $\sigma_r(M) \ge \sigma_{r+1}(M)$. At the optimum of the surrogate problem, both effective transformations $\widetilde{W}_K$ and $\widetilde{W}_V$ have their column spaces contained within the same r-dimensional dominant left singular subspace of M.*

Proof. Let $M = U\Sigma V^\top$ be the compact SVD, with leading truncations $U_r \in \mathbb{R}^{d\times r}$, $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, $V_r \in \mathbb{R}^{2d\times r}$, and $M_r = U_r\Sigma_r V_r^\top$. Since $\sigma_r > 0$, Σ_r is invertible and $\mathrm{Col}(M_r) = \mathrm{Col}(U_r) = \mathrm{span}(U_r)$.

Step 1: Feasible set equals all rank-r matrices. Every feasible $\widetilde{M} = W_c[W_K^\uparrow, W_V^\uparrow]$ has $\mathrm{rank}(\widetilde{M}) \le r$ (Remark 1). Conversely, any $F \in \mathbb{R}^{d\times 2d}$ with $\mathrm{rank}(F) \le r$ admits a factorisation $F = PQ$ with $P \in \mathbb{R}^{d\times r}$, $Q \in \mathbb{R}^{r\times 2d}$; partitioning $Q = [Q_K, Q_V]$ and setting $W_c = P$, $W_K^\uparrow = Q_K$, $W_V^\uparrow = Q_V$ recovers F. Hence the surrogate reduces to

$$\min_{\mathrm{rank}(F)\le r} \|M - F\|_F^2.$$

Step 2: Unique optimal approximation. By the Eckart–Young–Mirsky theorem, the unique minimiser (uniqueness from $\sigma_r > \sigma_{r+1}$) is $\widetilde{M}^* = M_r$, with $\mathrm{Col}(M_r) = \mathrm{span}(U_r)$.

Step 3: Block containment. Partitioning $\widetilde{M}^* = M_r$ by column position gives $\widetilde{W}_K = M_r^{(K)}$ (columns 1 to d) and $\widetilde{W}_V = M_r^{(V)}$ (columns $d+1$ to $2d$); this partition is unambiguous since SVD truncation does not reorder columns. Every column of each block lies in $\mathrm{Col}(M_r) = \mathrm{span}(U_r)$, giving

$$\mathrm{Col}(\widetilde{W}_K) \subseteq \mathrm{span}(U_r), \qquad \mathrm{Col}(\widetilde{W}_V) \subseteq \mathrm{span}(U_r).$$

□

Remark 6 (Why W_c Must Be Jointly Optimised)**.** *The reduction in Step 1 is valid because the surrogate minimises jointly over W_c, $W_K^\uparrow$, and $W_V^\uparrow$. If instead W_c were held fixed at an arbitrary value, the feasible set would be restricted to $\{W_cB : B \in \mathbb{R}^{r\times 2d}\}$, i.e. the image of W_c, which need not contain M_r unless $\mathrm{Col}(W_c) = \mathrm{span}(U_r)$. The theorem therefore characterises the globally optimal joint configuration of all three factor matrices. This is appropriate because W_c is itself a trained parameter; the surrogate captures the representational pressure on the full MLA parameterisation at each step.*

Remark 7 (Spectral Gap and Soft Degeneracy)**.** *When $\sigma_r(M) \approx \sigma_{r+1}(M)$, the optimal subspace is not uniquely determined: any rotation within the near-degenerate eigenspace yields an equally good approximation. In this regime the constraint should be interpreted as establishing a* family *of shared subspaces. This is empirically relevant because transformer weight matrices typically exhibit slowly decaying spectra; the alignment metrics of Section 4 measure how tightly the model concentrates within one member of this family.*

Note that confinement to a common subspace constrains the representational capacity available to each pathway but does not determine how that capacity is used — the within-subspace functional distinction between the two operators is not constrained by the bottleneck geometry alone.

3.5 Increasing Coupling Under Compression

Corollary 1 (Quantified Coupling Under Compression)**.** *Under the conditions of Theorem 2, letting $\rho_K = \mathrm{rank}(\widetilde{W}_K)$ and $\rho_V = \mathrm{rank}(\widetilde{W}_V)$,*

$$\dim\big(\mathrm{Col}(\widetilde{W}_K) \cap \mathrm{Col}(\widetilde{W}_V)\big) \;\ge\; \max(\rho_K + \rho_V - r,\; 0).$$

Decreasing r while holding ρ_K and ρ_V fixed strictly increases this lower bound, monotonically increasing representational coupling. In particular:

(i) If $\rho_K = \rho_V = r$, then $\mathrm{Col}(\widetilde{W}_K) = \mathrm{Col}(\widetilde{W}_V) = \mathrm{span}(U_r)$ exactly.

(ii) For $r = 1$ with $\rho_K, \rho_V \ge 1$, both column spaces equal $\mathrm{span}(u_1)$: every output of $\widetilde{W}_K$ is parallel to every output of $\widetilde{W}_V$ (though the matrices themselves need not be proportional).

Proof. Both column spaces are subspaces of the common r-dimensional ambient space $\mathrm{span}(U_r)$ (Theorem 2). The subspace intersection formula $\dim(A\cap B) \ge \dim A + \dim B - \dim \mathcal{V}$ applied with $\mathcal{V} = \mathrm{span}(U_r)$ gives the stated bound. Monotonicity in r is immediate since $\rho_K + \rho_V - r$ increases as r decreases. For case (i), the bound gives $\dim \ge r$, forcing both r-dimensional subspaces of the r-dimensional ambient space to coincide. For case (ii), any nonzero subspace of the one-dimensional $\mathrm{span}(u_1)$ must equal it. □

4 EMPIRICAL EVALUATION OF SUBSPACE ALIGNMENT

To evaluate the Shared Effective Subspace hypothesis (Theorem 2), we conduct a gauge-invariant audit of production-grade MLA checkpoints across three models: DeepSeek-V2-Lite (16B, 27 layers), DeepSeek-V2 (236B, 60 layers), and DeepSeek-V3 (671B, 62 layers), all with hidden dimension $d = 5120$, latent rank $r = 512$, $n_h = 16$ attention heads, and per-head key/value dimension $d_{\text{head}} = 128$. As established in Section 3.2, direct comparison of the isolated up-projection matrices $W_K^{\uparrow}$ and $W_V^{\uparrow}$ is structurally uninformative due to gauge freedom. All measurements operate exclusively on the gauge-invariant effective operators:

$$\widetilde{W}_K \;=\; W_c(W_K^{\uparrow})^{\top} \;\in\; \mathbb{R}^{d\times n_h d_{\text{head}}}, \qquad \widetilde{W}_V \;=\; W_c(W_V^{\uparrow})^{\top} \;\in\; \mathbb{R}^{d\times n_h d_{\text{head}}}. \tag{1}$$

For the DeepSeek models, $n_h d_{\text{head}} = 2048$ and $d = 5120$, so $\widetilde{W}_K, \widetilde{W}_V \in \mathbb{R}^{5120\times 2048}$. The transpose arises from the PyTorch storage convention detailed in Section 4.2; in the mathematical framework of Section 3.1, where $W_K^{\uparrow} \in \mathbb{R}^{r\times d}$ is defined as a fat matrix, the product $W_c W_K^{\uparrow}$ and the extracted product $W_c(W_K^{\uparrow})^{\top}$ are equivalent up to the output dimension (d vs. $n_h d_{\text{head}}$), both spanning $\mathrm{Col}(W_c)$.

4.1 WEIGHT MATRIX ALIGNMENT METRICS

We track the layer-by-layer geometric overlap between $\widetilde{W}_K$ and $\widetilde{W}_V$ using two complementary, scale-invariant metrics.

4.1.1 FROBENIUS-NORMALISED GRAM DISTANCE

The effective Gram distance is defined as

$$D_{\text{Gram}}\Big(\widetilde{W}_K,\, \widetilde{W}_V\Big) \;=\; \left\| \frac{\widetilde{W}_K \widetilde{W}_K^{\top}}{\|\widetilde{W}_K\|_F^2} \;-\; \frac{\widetilde{W}_V \widetilde{W}_V^{\top}}{\|\widetilde{W}_V\|_F^2} \right\|_F . \tag{2}$$

Since $\widetilde{W}_K \in \mathbb{R}^{d\times n_h d_{\text{head}}}$, the normalised Gram matrices $\widetilde{W}_K\widetilde{W}_K^{\top}/\|\widetilde{W}_K\|_F^2$ lie in $\mathbb{R}^{d\times d}$, and D_{Gram} measures the distance between the normalised column covariance structures of the two operators in the input space $\mathbb{R}^d$. A lower value indicates that the principal axes along which $\widetilde{W}_K$ transforms inputs into keys coincide with those along which $\widetilde{W}_V$ transforms inputs into values. This is the primary quantitative metric for this architecture class, for reasons detailed below.

4.1.2 PRINCIPAL ANGLES BETWEEN EFFECTIVE SUBSPACES

Let $Q_K, Q_V \in \mathbb{R}^{d\times r}$ be the orthonormal bases of the column spaces of $\widetilde{W}_K$ and $\widetilde{W}_V$ respectively, obtained via rank-r truncated SVD (the rank of each effective operator is at most r by Remark 1). The cosines of the principal angles θ_i are the singular values σ_i of $Q_K^{\top}Q_V$, and the mean principal angle is

$$\bar{\theta} \;=\; \frac{1}{r}\sum_{i=1}^{r} \arccos(\sigma_i). \tag{3}$$

A value of $\bar{\theta} = 0^\circ$ indicates identical column spaces; $\bar{\theta} = 90^\circ$ indicates orthogonality.

In the DeepSeek architecture, the stored up-projection tensors have shape $[n_h d_{\text{head}}, r] = [2048, 512]$ — a tall matrix with more rows than columns. Both $W_K^{\uparrow}$ and $W_V^{\uparrow}$ therefore have full column rank r, which means their transposes $(W_K^{\uparrow})^{\top}$ and $(W_V^{\uparrow})^{\top}$ have rank r, and consequently

$$\mathrm{Col}(\widetilde{W}_K) \;=\; \mathrm{Col}\big(W_c(W_K^{\uparrow})^{\top}\big) \;=\; \mathrm{Col}(W_c),$$

and identically for $\widetilde{W}_V$. Both effective operators span exactly $\mathrm{Col}(W_c)$ by construction, regardless of what the weights learned. Consequently, $\bar{\theta} \approx 0^\circ$ throughout the network as an architectural inevitability rather than a learned property. In this regime $\bar{\theta}$ serves as a sanity check confirming that the up-projections maintain full column rank during training, while D_{Gram} carries the quantitative alignment story by measuring shape differences *within* $\mathrm{Col}(W_c)$.

4.2 Extraction Logic and Parameter Slicing

PyTorch stores weight matrices in transposed convention `[out_dim, in_dim]`. The mathematical up-projection $W_K^{\uparrow}$ of Section 3.1, which maps from latent dimension r to key dimension $n_h d_{\text{head}}$, is therefore stored as a tensor of shape $[n_h d_{\text{head}}, r]$, and the effective operator $\widetilde{W}_K = W_c (W_K^{\uparrow})^\top$ is computed by transposing the extracted slice before multiplication. We extract weights for each layer ℓ as follows.

1. **Down-projection extraction.** The compression matrix is read directly:

$$W_c = \texttt{w_kv_down_proj.weight}^\top \in \mathbb{R}^{d \times r}, \tag{4}$$

   since `w_kv_down_proj.weight` is stored as $[r, d]$.

2. **Up-projection slicing.** The module `kv_b_proj` concatenates the latent key up-projection, the latent value up-projection, and the decoupled positional key projection W_{KR} along the output dimension, yielding a tensor of shape $[n_h d_{\text{head}} + n_h d_{\text{head}} + n_h d_{KR},\ r]$. We isolate the latent pathways by slicing along the first dimension:

$$W_K^{\uparrow} = \texttt{kv_b_proj.weight}[0 : n_h d_{\text{head}},\ :] \in \mathbb{R}^{n_h d_{\text{head}} \times r}, \tag{5}$$

$$W_V^{\uparrow} = \texttt{kv_b_proj.weight}[n_h d_{\text{head}} : 2 n_h d_{\text{head}},\ :] \in \mathbb{R}^{n_h d_{\text{head}} \times r}. \tag{6}$$

   For the DeepSeek models, $n_h d_{\text{head}} = 16 \times 128 = 2048$, so the key block occupies rows 0–2047 and the value block rows 2048–4095 of the $[4096, 512]$ tensor.

3. **Effective operator computation.**

$$\widetilde{W}_K = W_c (W_K^{\uparrow})^\top \in \mathbb{R}^{d \times n_h d_{\text{head}}}, \qquad \widetilde{W}_V = W_c (W_V^{\uparrow})^\top \in \mathbb{R}^{d \times n_h d_{\text{head}}}. \tag{7}$$

   All alignment metrics in Sections 4.1–4.4 are computed on these products exclusively.

4.3 Activation-Based Alignment Evaluation

To assess whether the static weight-space coupling translates to runtime behaviour, we track activation alignment on WikiText-2 (50K tokens, sequence length 512). For each sequence position t we capture effective key and value vectors prior to RoPE application:

$$k_t = x_t \widetilde{W}_K \in \mathbb{R}^{n_h d_{\text{head}}}, \qquad v_t = x_t \widetilde{W}_V \in \mathbb{R}^{n_h d_{\text{head}}},$$

and compute the sequence-averaged cosine similarity:

$$\rho = \frac{1}{n} \sum_{t=1}^{n} \frac{k_t \cdot v_t}{\|k_t\|_2 \, \|v_t\|_2}. \tag{8}$$

Because ρ involves only $\widetilde{W}_K$ and $\widetilde{W}_V$, it is gauge-invariant under the reparametrisation of Section 3.2.

On DeepSeek-V2-Lite, ρ is near-zero throughout the network ($|\rho| \leq 0.008$ across all layers), indicating that despite shared subspace structure, the model produces nearly orthogonal key and value activation vectors for real tokens at every layer. This is not a failure of the metric: it reveals that the model exploits the shared coordinate system to encode maximally diverse key and value representations within it. The shared subspace constrains the *capacity* available to each pathway without constraining how that capacity is used.

Although $\rho \approx 0$ absolutely, D_{Gram} is a significant negative predictor of ρ across layers (Pearson $r = -0.482$): layers with lower Gram distance tend to produce slightly more similar key and value activations. This confirms that static weight geometry directionally predicts runtime behaviour, even when the absolute activation similarity is small. The $\bar{\theta}$ metric shows no meaningful correlation with ρ ($r = -0.068$), consistent with its structural degeneracy for this architecture.

These results motivate a precise question: if keys and values are geometrically proximate (low D_{Gram}) but functionally orthogonal ($\rho \approx 0$) in the collapsed layers, is the within-subspace functional distinction between them load-bearing? Section E addresses this directly.

4.4 Empirical Results

Layer-wise tracking of D_{Gram} across all three models reveals a consistent geometric pattern whose character depends on scale (Figure 1). We describe the findings per model and then identify the cross-model regularities.

4.4.1 DeepSeek-V2-Lite (16B, 27 layers)

D_{Gram} declines monotonically from a peak of 0.104 at layer 0 to a minimum of 0.020 at layer 19, with small fluctuations thereafter (maximum late-layer value 0.046, a $2.3\times$ recovery over the minimum). There is no pronounced late-layer divergence. The collapse zone covers approximately layers 7–26 (74% of the network). The 16B model appears insufficiently deep to develop the functional asymmetry at the output boundary that would produce a late-layer uptick; the network is in partial collapse from the earliest layers onward.

4.4.2 DeepSeek-V2 (236B, 60 layers)

D_{Gram} exhibits an asymmetric profile: a sharp spike at layer 0 ($D_{\mathrm{Gram}} = 0.213$, $27\times$ the minimum), rapid decline over layers 1–22, sustained near-zero collapse from layers 22–54 (minimum 0.008 at layer 53), and a faint late-layer uptick in the final five layers ($0.008 \rightarrow 0.014$, $1.8\times$ recovery). The collapse zone covers approximately layers 22–54 (55% of network). The profile is better described as early-layer decoupling followed by sustained subspace collapse than as a symmetric U-shape.

4.4.3 DeepSeek-V3 (671B, 62 layers)

D_{Gram} exhibits a clear U-shaped profile. Early layers (0–8) show an elevated plateau ($D_{\mathrm{Gram}} = 0.033$–$0.067$), followed by sustained deep collapse across layers 9–57 (minimum 0.0097 at layer 46). The late layers show a dramatic divergence: D_{Gram} reaches 0.082 at layer 58 and 0.167 at layer 60 — $17.2\times$ the minimum and the strongest late-layer signal across all three models. The collapse zone covers approximately layers 9–57 (79% of the network).

4.4.4 Cross-Model Regularities

Three findings hold across all three models: a) **Subspace collapse is universal and quantitatively strong.** D_{Gram} reaches near-zero in the middle layers of all three models, confirming the core prediction of Theorem 2 at 16B, 236B, and 671B, b) **The U-shape is scale-dependent.** Early-layer decoupling and middle-layer collapse are consistent across all three models. Late-layer divergence is an emergent property of scale: negligible at 16B ($2.3\times$ recovery), faint at 236B ($1.8\times$), and dramatic at 671B ($17.2\times$). The symmetric U-shape is therefore a phenomenon of very large models, consistent with the hypothesis that the output-boundary functional asymmetry between the key and value pathways intensifies with depth and scale (Figure 1), and c) **Scale sharpens the early-layer peak.** The ratio of early-layer peak to collapse minimum is $5\times$ at 16B, $27\times$ at 236B, and $6.7\times$ at 671B (where the early zone is a broader plateau rather than a single spike). In all cases the early layers maintain substantially higher D_{Gram} than the collapse zone, consistent with the network preserving distinct key and value representations during feature extraction.

5 Discussion

5.1 The Shared Subspace as a Coordinate System, Not a Redundancy

The central finding is a distinction the prior literature has not drawn explicitly: MLA's bottleneck creates *structural* coupling without creating *functional* redundancy. The shared subspace constraint forces both effective operators into the same r-dimensional coordinate system, but within that system the model encodes maximally diverse representations.

This resolves the apparent tension between three observations. $D_{\mathrm{Gram}} \approx 0$ in the middle layers: both operators share the same column space. $|\rho| \leq 0.008$ everywhere: keys and values produce nearly orthogonal activation vectors. $\Delta\mathrm{PPL} \approx +10$ from middle-block clamping: the within-subspace rotation is load-bearing. These are not contradictions. The shared subspace is the filing system; the within-subspace rotation is the organisation within it. The bottleneck is being used twice over, not once.

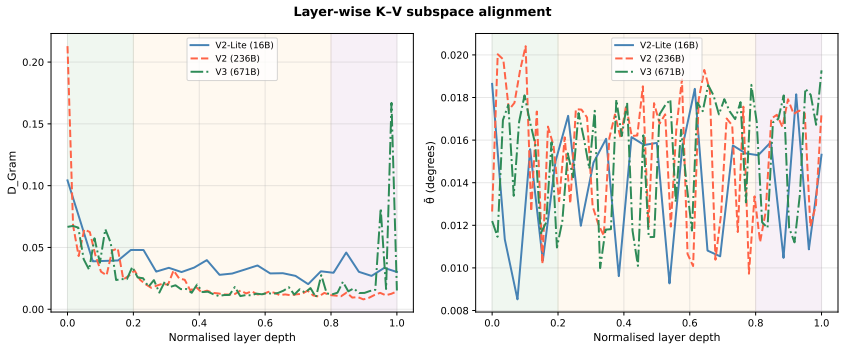


Figure 1: Layer-wise D_{Gram} for DeepSeek-V2-Lite (dotted), DeepSeek-V2 (dashed), and DeepSeek-V3 (solid), plotted as a fraction of total network depth for comparability. All three models exhibit early-layer decoupling and middle-layer collapse. Late-layer divergence is absent at 16B, faint at 236B, and dramatic at 671B ($D_{\text{Gram}} = 0.167$ at layer 60, $17.2\times$ the minimum), establishing the U-shape as a scale-dependent phenomenon. Shaded regions denote the collapse zones identified in Section 4.4.

5.2 Scale Dependence of the U-Shaped Profile

Early-layer decoupling and middle-layer collapse are universal across 16B, 236B, and 671B. Late-layer divergence is an emergent property of scale: absent at 16B, faint at 236B ($1.8\times$ recovery), and dramatic at 671B ($17.2\times$, $D_{\text{Gram}} = 0.167$ at layer 60).

We conjecture this reflects growing functional asymmetry near the output boundary. The key pathway interacts with the loss indirectly through softmax attention scores; the value pathway contributes linearly to the block output and directly to cross-entropy. As the model grows deeper, this asymmetry intensifies, producing larger D_{Gram} values where the loss gradient is most direct — consistent with the $2.9\times$ directional asymmetry in the global clamping experiment.

5.3 Key-Pathway Primacy and Compression Implications

Outside the collapse zone, replacing keys with value-derived projections is $2.9\times$ more damaging than the reverse, because corrupting attention routing cascades through the full sequence whereas corrupting content extraction leaves attention patterns intact. Within the collapse zone this asymmetry vanishes: both operators have converged to the same subspace and are equally sensitive to replacement, making the collapse zone an equaliser between the two pathways.

Post-hoc clamping is therefore not a safe compression strategy for fully trained MLA models, since the within-subspace rotation is load-bearing. The shared subspace finding does have compression value, but the right target is the subspace structure itself rather than the within-subspace distinction. Promising directions include: joint quantisation of $\widetilde{W}_K$ and $\widetilde{W}_V$ along their shared principal directions, exploiting the fact that both operators are confined to the same r-dimensional subspace to reduce the precision required for that subspace; shared adapter fine-tuning that imposes a common low-rank update direction across both pathways; or training a compressed model with an explicit $K = V$ constraint from the start Kayyam et al. (2026), which avoids the problem entirely by never developing the within-subspace rotation in the first place.

6 Conclusion

MLA's shared low-rank bottleneck constrains keys and values to the same low-dimensional subspace but does not make them functionally redundant. A gauge-invariant analysis reveals increasing subspace collapse with scale, yet causal interventions show that forcing $K = V$ significantly harms performance ($\Delta\text{PPL} \approx +10$), demonstrating that within-subspace differences are essential. The key insight is that shared subspace reflects representational capacity rather than representational content: MLA places keys and values in the same geometric region while preserving distinct functional roles, suggesting that compression methods should preserve, rather than collapse, this diversity.

Limitations and Future Directions. Our experiments are limited to DeepSeek-V2-Lite (16B), leaving the scaling of middle-block clamping effects in larger V2 and V3 models as an open question. The analysis also excludes the RoPE pathway, so any positional asymmetries remain unexplored. Another practical question is whether fine-tuning can recover performance after $K = V$ clamping, making post-hoc compression viable. More broadly, extending this geometric account to standard MHA may help develop a unified explanation of key-value alignment across transformers.

AI USE STATEMENT

Large language models (LLMs) were used in the preparation of this work in a supporting capacity only. Specifically, LLM-based assistants were used to (i) polish the writing, including grammar, phrasing, and consistency of notation across sections; (ii) assist with writing and debugging analysis code for weight extraction, gauge-invariant metric computation, and perplexity evaluation, including diagnosing the checkpoint-loading issue described in Appendix B.2; and (iii) help locate and summarise related literature. All research questions, theoretical results and proofs, experimental design, and interpretation of results were conceived and carried out by the author. Every mathematical derivation was verified by hand, all reported numbers were produced by code that the author ran and checked directly, and all cited works were verified against their original sources. The author takes full responsibility for the content of this paper, including any errors. LLMs are not eligible for authorship and are not listed as authors.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.298.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Tao Ji, Bin Guo, Yuanbin Wu, Qipeng Guo, Lixing Shen, Zhan Chen, Xipeng Qiu, Qi Zhang, and Tao Gui. Towards economical inference: Enabling deepseek's multi-head latent attention in any transformer-based llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33313–33328, 2025.

Ali Kayyam, Anusha Madan Gopal, and M. Anthony Lewis. Do transformers need three projections? systematic study of QKV variants. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.04032.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024b.

Fanxu Meng, Pingzhi Tang, Xiaojuan Tang, Zengwei Yao, Xing Sun, and Muhan Zhang. Transmla: Multi-head latent attention is all you need. *arXiv preprint arXiv:2502.07864*, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1609.07843.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116 (23):11537–11546, 2019.

Khalid Shaikh, Asmit Kumar Singh, Rebecca Christopher Dsouza, and Shikhar Shiromani. Linear predictability of attention heads in large language models. *arXiv preprint arXiv:2603.13314*, 2026.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

A Full Gradient Derivations

We derive both gradients in full. Throughout, $W_K^{\uparrow} \in \mathbb{R}^{r\times d}$ follows the mathematical convention of Section 3.1; the relationship to the stored tensor shape is described in Section 4.2. Let $E = \frac{\partial \mathcal{L}}{\partial O} \in \mathbb{R}^{n\times d}$, $E_A := \frac{\partial \mathcal{L}}{\partial S} \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^{n\times r}$, and $Q \in \mathbb{R}^{n\times d}$.

Value Gradient

From $O = A(cW_V^{\uparrow})$, treating A and c as constants:

$$\frac{\partial \mathcal{L}}{\partial W_V^{\uparrow}} \;=\; (Ac)^{\top}E \;=\; c^{\top}A^{\top}E \;=\; c^{\top}B_V, \qquad B_V := A^{\top}E \in \mathbb{R}^{n\times d}.$$

Key Gradient

From $S = Q(W_K^{\uparrow})^{\top}c^{\top}$, using index names $\beta \in \{1,\dots,r\}$ and $\alpha \in \{1,\dots,d\}$ for the row and column indices of $W_K^{\uparrow} \in \mathbb{R}^{r\times d}$ respectively:

$$S_{ij} \;=\; \sum_{\beta,\alpha} Q_{i\alpha}\,(W_K^{\uparrow})_{\beta\alpha}\,c_{j\beta}, \qquad \text{so} \qquad \frac{\partial S_{ij}}{\partial (W_K^{\uparrow})_{\beta\alpha}} \;=\; Q_{i\alpha}\,c_{j\beta}.$$

Contracting against E_A:

$$\frac{\partial \mathcal{L}}{\partial (W_K^{\uparrow})_{\beta\alpha}} \;=\; \sum_{i,j} (E_A)_{ij}\,Q_{i\alpha}\,c_{j\beta} \;=\; \big(Q^{\top}E_A c\big)_{\alpha\beta}.$$

Since $Q^{\top}E_A c \in \mathbb{R}^{d\times r}$ is indexed (α,β) while $W_K^{\uparrow}$ is indexed (β,α), taking the transpose gives:

$$\frac{\partial \mathcal{L}}{\partial W_K^{\uparrow}} \;=\; \big(Q^{\top}E_A c\big)^{\top} \;=\; c^{\top}E_A^{\top}Q \;=\; c^{\top}B_K, \qquad B_K := E_A^{\top}Q \in \mathbb{R}^{n\times d}.$$

B Extended Experimental Details

B.1 Hardware and Software

All experiments were run on a single workstation with 62 GB DDR4 RAM and an Intel Core i7-6850K (6 cores / 12 threads, 3.6 GHz, AVX2, no native BF16). Four installed GTX 1080 Ti GPUs (compute capability 6.1) are incompatible with the installed PyTorch build (PyTorch 2.12.0+cu130, which requires CC $\geq$ 7.5); all computation is therefore CPU-only. BF16 matrix operations fall back to FP32 via oneDNN. Model weights are kept in `bfloat16` throughout (31 GB resident memory for DeepSeek-V2-Lite).

Software versions: Python 3.12, PyTorch 2.12.0, Transformers 5.3.0, safetensors 0.7.0, datasets (Salesforce/wikitext). All forward passes use `OMP_NUM_THREADS=12` and `use_cache=False`.

B.2 Model Loading and the `_tied_weights_keys` Bug

Loading DeepSeek-V2-Lite via `AutoModelForCausalLM.from_pretrained` under Transformers 5.x silently fails to load any weights from the checkpoint. The root cause is in the model's remote code (`modeling_deepseek.py`):

```
_tied_weights_keys = ["lm_head.weight"]   # old Transformers 4.x list format
```

Transformers 5.x calls `.keys()` on this attribute, expecting a `dict`; the `list` raises `AttributeError`. The exception is caught internally and stored as a `SkipParameters` conversion error, after which `_initialize_missing_keys` calls `_init_weights`, re-initialising every `nn.Linear` and `nn.Embedding` weight to $\mathcal{N}(0,\,0.02)$. The result is that all named parameters take initialisation values rather than the trained checkpoint.

Fix. After `from_pretrained`, every named parameter is unconditionally overwritten by re-reading directly from the safetensors shards:

```
for shard_file, keys in sorted(shard_to_keys.items()):
    with safe_open(shard_path, framework="pt", device="cpu") as f:
        for key in keys:
            model_params[key].data.copy_(
                f.get_tensor(key).to(dtype))
```

After the fix, WikiText-2 perplexity is 6.275 (Table 6).

B.3 Perplexity Evaluation Protocol

WikiText-2 validation split (Merity et al., 2017), loaded via the Salesforce/wikitext HuggingFace repository. Context window $L = 512$, stride $s = 512$ (non-overlapping windows). For Experiment 2 (activation alignment), forward passes cover the first 50,000 tokens (≈97 chunks). For Experiment 3 (clamping), each PPL estimate uses the first 30 chunks (≈15,360 tokens). We verified that the 30-chunk estimate produces stable rankings across conditions: the baseline on the full validation set gives PPL 6.10 versus 6.275 for 30 chunks, a 2.7% difference that does not affect any qualitative conclusion.

B.4 Clamping Implementation

The weight tensor `kv_b_proj.weight` has shape $\big[n_h(d_{qk} + d_v),\ r\big]$, where rows $[0 : n_h d_{qk}]$ encode $W_K^{\uparrow}$ and rows $[n_h d_{qk} : n_h(d_{qk} + d_v)]$ encode $W_V^{\uparrow}$. All DeepSeek-V2/V2-Lite/V3 variants satisfy $d_{qk} = d_v = 128$; this equality is asserted before any modification. The middle-block schedule clamps layers $\big[\lfloor 0.25L \rfloor, \lfloor 0.75L \rfloor\big]$ (layers 6–19 for $L = 27$, covering 52% of the network). The global schedule clamps all L layers.

C Per-Layer Metric Tables (Experiment 1)

Tables 1–3 report the per-layer Gram distance D_{Gram} and mean principal angle $\bar{\theta}$ (in degrees) for all three models. DeepSeek-V3 weights were converted from FP8 to `bfloat16` before metric computation.

Table 1: DeepSeek-V2-Lite (16B, 27 layers). Global min $D_{\text{Gram}} = 0.0203$ (layer 19); global max $D_{\text{Gram}} = 0.1043$ (layer 0).

Layer	D_{Gram}	$\bar{\theta}$ (°)
0	0.1043	0.0186
1	0.0717	0.0113
2	0.0388	0.0085
3	0.0391	0.0156
4	0.0395	0.0103
5	0.0480	0.0150
6	0.0479	0.0171
7	0.0304	0.0120
8	0.0335	0.0149
9	0.0301	0.0161
10	0.0337	0.0096
11	0.0398	0.0161
12	0.0278	0.0158
13	0.0289	0.0159
14	0.0321	0.0093
15	0.0354	0.0156
16	0.0290	0.0184
17	0.0292	0.0108
18	0.0270	0.0105

Continued on next page

Table 1 — continued

Layer	D_{Gram}	$\bar{\theta}$ (°)
19	0.0203	0.0157
20	0.0307	0.0154
21	0.0295	0.0153
22	0.0459	0.0159
23	0.0303	0.0105
24	0.0270	0.0182
25	0.0335	0.0109
26	0.0302	0.0153

Table 2: DeepSeek-V2 (236B, 60 layers). Global min $D_{\text{Gram}} = 0.0080$ (layer 53); global max $D_{\text{Gram}} = 0.2129$ (layer 0).

Layer	D_{Gram}	$\bar{\theta}$ (°)
0	0.2129	0.0126
1	0.0678	0.0200
2	0.0430	0.0198
3	0.0645	0.0174
4	0.0625	0.0178
5	0.0437	0.0193
6	0.0304	0.0204
7	0.0275	0.0127
8	0.0474	0.0173
9	0.0492	0.0102
10	0.0242	0.0166
11	0.0271	0.0157
12	0.0314	0.0131
13	0.0238	0.0161
14	0.0202	0.0130
15	0.0173	0.0174
16	0.0191	0.0174
17	0.0206	0.0171
18	0.0198	0.0129
19	0.0314	0.0121
20	0.0249	0.0115
21	0.0239	0.0161
22	0.0152	0.0172
23	0.0181	0.0161
24	0.0138	0.0176
25	0.0151	0.0162
26	0.0133	0.0162
27	0.0128	0.0185
28	0.0120	0.0119
29	0.0119	0.0177
30	0.0124	0.0168
31	0.0152	0.0172
32	0.0128	0.0119
33	0.0145	0.0166
34	0.0120	0.0189
35	0.0121	0.0107
36	0.0141	0.0101
37	0.0134	0.0179
38	0.0115	0.0193
39	0.0120	0.0183

Continued on next page

Table 2 — continued

Layer	D_{Gram}	$\bar{\theta}$ (°)
40	0.0113	0.0104
41	0.0121	0.0125
42	0.0124	0.0170
43	0.0150	0.0167
44	0.0112	0.0117
45	0.0103	0.0176
46	0.0132	0.0097
47	0.0115	0.0134
48	0.0111	0.0112
49	0.0105	0.0126
50	0.0135	0.0170
51	0.0094	0.0172
52	0.0098	0.0166
53	0.0080	0.0179
54	0.0093	0.0172
55	0.0116	0.0174
56	0.0133	0.0173
57	0.0113	0.0119
58	0.0125	0.0130
59	0.0144	0.0173

Table 3: DeepSeek-V3 (671B, 62 layers; weights converted FP8 $\rightarrow$ BF16). Global min $D_{\text{Gram}} = 0.0097$ (layer 46); global max $D_{\text{Gram}} = 0.1670$ (layer 60).

Layer	D_{Gram}	$\bar{\theta}$ (°)
0	0.0667	0.0122
1	0.0674	0.0115
2	0.0661	0.0169
3	0.0405	0.0177
4	0.0320	0.0134
5	0.0575	0.0169
6	0.0376	0.0181
7	0.0648	0.0169
8	0.0538	0.0158
9	0.0237	0.0116
10	0.0244	0.0121
11	0.0243	0.0159
12	0.0343	0.0110
13	0.0257	0.0122
14	0.0251	0.0154
15	0.0182	0.0143
16	0.0236	0.0173
17	0.0134	0.0161
18	0.0222	0.0147
19	0.0180	0.0174
20	0.0194	0.0100
21	0.0163	0.0118
22	0.0164	0.0118
23	0.0133	0.0179
24	0.0200	0.0162
25	0.0139	0.0178
26	0.0142	0.0117

Continued on next page

Table 3 — continued

Layer	D_{Gram}	$\bar{\theta}$ (°)
27	0.0122	0.0101
28	0.0107	0.0162
29	0.0115	0.0114
30	0.0116	0.0114
31	0.0181	0.0172
32	0.0106	0.0176
33	0.0115	0.0179
34	0.0113	0.0131
35	0.0126	0.0165
36	0.0121	0.0140
37	0.0131	0.0117
38	0.0124	0.0178
39	0.0132	0.0174
40	0.0150	0.0186
41	0.0178	0.0181
42	0.0118	0.0172
43	0.0162	0.0179
44	0.0134	0.0173
45	0.0172	0.0170
46	0.0097	0.0178
47	0.0276	0.0131
48	0.0129	0.0186
49	0.0120	0.0168
50	0.0134	0.0121
51	0.0222	0.0119
52	0.0145	0.0168
53	0.0169	0.0170
54	0.0131	0.0181
55	0.0132	0.0119
56	0.0148	0.0112
57	0.0157	0.0135
58	0.0815	0.0184
59	0.0163	0.0181
60	0.1670	0.0167
61	0.0152	0.0193

D Extended Discussion of Figure 1

Figure 1 plots D_{Gram} against normalised depth ℓ/L so that the three networks are directly comparable despite having $L \in \{27, 60, 62\}$ layers. Full per-layer tables are in Appendix C.

D.1 Collapse Zone Boundaries

We define the *collapse zone* as the contiguous run of layers satisfying $D_{\text{Gram}} < D_{\text{thresh}}$, where $D_{\text{thresh}} = 2 \times D_{\text{min}}$ for each model. Table 4 summarises the resulting boundaries.

Table 4: Collapse zone boundaries per model. D_{min}: global minimum D_{Gram}. $D_{\text{thresh}} = 2 \times D_{\text{min}}$. Coverage: fraction of layers inside the collapse zone.

Model	L	D_{min}	Layer	D_{thresh}	Zone	Coverage
V2-Lite (16B)	27	0.0203	19	0.0406	7–26	74%
V2 (236B)	60	0.0080	53	0.0160	22–54	55%
V3 (671B)	62	0.0097	46	0.0194	9–57	79%

V3's collapse zone uses the single-run definition (layers 9–57); layer 47 ($D_{\text{Gram}} = 0.028$) slightly exceeds $D_{\text{thresh}} = 0.019$ but the zone boundaries are insensitive to this excursion.

D.2 Late-Layer Divergence and Scale Dependence

Table 5: Late-layer D_{Gram} statistics. Late zone = last 10% of layers. Late/min ratio measures divergence relative to the collapse floor.

Model	Late zone	Peak D_{Gram}	Layer	Late/min ratio
V2-Lite (16B)	layers 24–26	0.0335	25	$1.7\times$
V2 (236B)	layers 54–59	0.0144	59	$1.8\times$
V3 (671B)	layers 56–61	0.1670	60	$\mathbf{17.2}\times$

The late-layer ratio is nearly identical at 16B and 236B (≈ 1.7–$1.8\times$) and jumps discontinuously to $17.2\times$ at 671B, suggesting a qualitative change rather than a smooth scaling trend. Within the V3 late zone, the divergence is localised: layer 60 ($D_{\text{Gram}} = 0.167$) is flanked by layers with $D_{\text{Gram}} \approx 0.015$–$0.016$, and layer 58 shows a secondary spike ($D_{\text{Gram}} = 0.082$), suggesting these layers serve a structurally distinct role at the boundary between deep processing and output projection.

D.3 Mean Principal Angle $\bar{\theta}$

$\bar{\theta}$ is near zero for every layer in all three models (range 0.009–0.020 across all measurements). This is a structural consequence of the up-projection matrices having full column rank $r = 512$: $\text{Col}(\widetilde{W}_K)$ and $\text{Col}(\widetilde{W}_V)$ both equal $\text{Col}(W_c)$, making all canonical angles zero. The values confirm full column rank throughout all three networks and rule out any low-rank structure in the up-projections.

E Causal Intervention: Probing the Structural–Functional Distinction

The geometric collapse documented in Section 4 establishes that the effective key and value operators share a common subspace in the middle layers of all three models. Section 4.3 further shows that despite this geometric proximity, the model produces nearly orthogonal key and value activation vectors for real tokens ($|\rho| \leq 0.008$). These two findings raise a precise causal question: is the within-subspace functional distinction between the key and value pathways load-bearing, or is it expendable?

We address this question via a surgical parameter-clamping intervention that forces $K = V$ effective alignment in selected layers and measures the downstream perplexity impact. The intervention is designed as a causal probe, not a compression strategy: if forcing $K = V$ in geometrically collapsed layers is near-lossless, the functional distinction is expendable; if it causes substantial degradation, the distinction is load-bearing despite the shared geometry. Crucially, we test *both* clamping directions to control for any asymmetry in the functional roles of the two pathways and to obtain a more complete characterisation of the causal structure.

E.1 Surgical Operator Clamping

As established in Section 3.2, the intervention must act on the up-projection parameters rather than the raw factor matrices. With the effective operator notation $\widetilde{W}_K = W_c(W_K^{\uparrow})^\top$ (Section 4.2), setting $W_K^{\uparrow} \leftarrow W_V^{\uparrow}$ implies $(W_K^{\uparrow})^\top \leftarrow (W_V^{\uparrow})^\top$, and therefore $\widetilde{W}_K^{\text{new}} = W_c(W_V^{\uparrow})^\top = \widetilde{W}_V$ exactly — no rank argument is required. We test two directions:

$$\text{Direction A:} \quad W_K^{\uparrow(\ell)} \leftarrow W_V^{\uparrow(\ell)} \quad \text{(value projection copied into key slot)} \tag{9}$$

$$\text{Direction B:} \quad W_V^{\uparrow(\ell)} \leftarrow W_K^{\uparrow(\ell)} \quad \text{(key projection copied into value slot)} \tag{10}$$

Concretely, in the DeepSeek tensor layout where `kv_b_proj.weight` has shape $[2n_h d_{\text{head}}, r] = [4096, 512]$ with the key block in rows $[0 : n_h d_{\text{head}}]$ and the value block in rows $[n_h d_{\text{head}} : 2n_h d_{\text{head}}]$:

$$\text{Direction A:} \quad \texttt{kv_b_proj.weight}[: n_h d_{\text{head}}, :] \leftarrow \texttt{kv_b_proj.weight}[n_h d_{\text{head}} :, :] \tag{11}$$

$$\text{Direction B:} \quad \texttt{kv_b_proj.weight}[n_h d_{\text{head}} :, :] \leftarrow \texttt{kv_b_proj.weight}[: n_h d_{\text{head}}, :] \tag{12}$$

Direction A replaces the key projection with the value projection, corrupting attention score computation while leaving value extraction intact. Direction B replaces the value projection with the key projection, leaving attention routing intact while corrupting content extraction. Testing both directions separates the functional contributions of the two pathways and reveals whether either is more critical to model performance. Each modified model is evaluated zero-shot without fine-tuning or weight adjustment.

E.2 Intervention Scheduling Strategies

We evaluate two scheduling strategies, each probing a different region of the alignment profile established in Section 4.4.

1. **Middle-block clamping.** Clamping applied to the geometrically collapsed central band:
$$\ell \in \big[\lfloor 0.25 \cdot L \rfloor,\ \lfloor 0.75 \cdot L \rfloor\big]. \tag{13}$$
This directly tests whether the collapse zone identified in Experiment 1 corresponds to functional redundancy.
2. **Global baseline clamping.** All layers clamped uniformly. This establishes an upper bound on perplexity degradation, tests whether the alignment profile correctly identifies sensitive regions, and — when combined with the directional comparison — reveals any functional hierarchy between the key and value pathways outside the collapse zone.

E.3 Evaluation Protocol

We evaluate on WikiText-2 validation (30 chunks $\times$ 512-token context, CPU inference) on DeepSeek-V2-Lite. The primary quantity is the perplexity change relative to the unmodified baseline:

$$\Delta\text{PPL} = \exp\big(\mathcal{L}_{\text{clamped}}\big) - \exp\big(\mathcal{L}_{\text{baseline}}\big). \tag{14}$$

A value of $\Delta\text{PPL} \approx 0$ in both directions would indicate functional redundancy; substantial positive values indicate load-bearing information in the clamped parameters. Asymmetry between the two directions indicates a functional hierarchy between the key and value pathways.

E.4 Results

Schedule	Layers	Direction	PPL	ΔPPL
Baseline	0	—	6.275	—
Middle-block	14 (6–19)	$W_K^{\uparrow} \leftarrow W_V^{\uparrow}$	16.371	+10.10
Middle-block	14 (6–19)	$W_V^{\uparrow} \leftarrow W_K^{\uparrow}$	16.993	+10.72
Global	27 (all)	$W_K^{\uparrow} \leftarrow W_V^{\uparrow}$	1189.980	+1183.71
Global	27 (all)	$W_V^{\uparrow} \leftarrow W_K^{\uparrow}$	412.599	+406.32

Table 6: Perplexity results for bidirectional K=V clamping on DeepSeek-V2-Lite (WikiText-2 validation). Middle-block covers the D_{Gram} collapse zone (layers 6–19, $D_{\text{Gram}} \approx 0.020$–$0.035$). Global includes early layers where D_{Gram} is high (layer 0: $D_{\text{Gram}} = 0.104$).

E.5 Interpretation: Structural Collapse Without Functional Redundancy

The results reveal three distinct findings, each carrying separate theoretical significance.

FINDING 1: COLLAPSE ZONE IS FUNCTIONALLY LOAD-BEARING IN BOTH DIRECTIONS

Middle-block clamping causes $\Delta\text{PPL} = +10.10$ (Direction A) and $+10.72$ (Direction B) — a $2.6\times$ and $2.7\times$ increase over baseline respectively. The functional distinction between keys and values is load-bearing even in layers where the effective operators are geometrically near-identical ($D_{\text{Gram}} \approx 0$). This directly answers the causal question posed at the outset: subspace collapse is a *structural* property of the weight operators, not a *functional* redundancy. The model exploits the shared coordinate system to encode maximally diverse key and value representations within it, and that diversity is necessary for performance.

FINDING 2: NEAR-SYMMETRY IN THE COLLAPSE ZONE CONFIRMS SUBSPACE COLLAPSE

The two directions yield nearly identical degradation ($+10.10$ vs $+10.72$, a 6% difference). This symmetry is itself informative: it means neither pathway is structurally privileged in the collapsed layers. The key and value operators have converged to a state where they are geometrically interchangeable under intervention, even while remaining functionally distinct in their outputs. This is the strongest empirical confirmation of the subspace collapse hypothesis — the two operators share not only the same column space but also the same functional sensitivity to replacement, precisely as Theorem 2 predicts when both operators saturate the rank-r bottleneck.

FINDING 3: GLOBAL ASYMMETRY REVEALS KEY-PATHWAY PRIMACY

Under global clamping the two directions diverge dramatically: $W_K^{\uparrow} \leftarrow W_V^{\uparrow}$ causes $\Delta\text{PPL} = +1183.71$, while $W_V^{\uparrow} \leftarrow W_K^{\uparrow}$ causes only $+406.32$ — a $2.9\times$ asymmetry. This gap is driven by the early layers (layers 0–5, $D_{\text{Gram}} = 0.048$–0.104), where the key and value operators are geometrically distinct and functionally differentiated.

The asymmetry reflects the different roles the two pathways play in the attention mechanism. Keys govern *attention routing* — which tokens get attended to — via $\text{softmax}(QK^\top/\sqrt{d})$. Replacing keys with value-derived projections (Direction A) corrupts attention routing, a failure mode that cascades through the full sequence: once the wrong tokens are attended to, the subsequent value aggregation cannot recover. Values govern *content extraction* — what information is aggregated given the attention weights. Replacing values with key-derived projections (Direction B) degrades extracted content but leaves attention patterns intact, a less catastrophic disruption since the model continues to attend to the correct positions.

Outside the collapse zone the key pathway therefore carries strictly more functional weight than the value pathway. This asymmetry is masked in the collapse zone (Finding 2) because both operators have converged to the same subspace, making them equally sensitive to replacement. The transition from symmetric to asymmetric sensitivity as clamping extends beyond the collapse zone provides an independent, functionally-grounded characterisation of where the collapse zone ends.

THREE-EXPERIMENT SUMMARY

- **Experiment 1 (static geometry):** $\widetilde{W}_K$ and $\widetilde{W}_V$ share a subspace in the middle layers ($D_{\text{Gram}} \approx 0$). Representational *capacity* is shared.
- **Experiment 2 (dynamic activations):** Despite the shared subspace, K and V activations are nearly orthogonal ($|\rho| \leq 0.008$, Pearson $r = -0.482$ between D_{Gram} and ρ). The model produces functionally *diverse* outputs within the shared space.
- **Experiment 3 (causal intervention):** Forcing $K = V$ causes $\Delta\text{PPL} \approx +10$ in both directions in the collapse zone, proving the functional diversity is *load-bearing*. The global directional asymmetry ($2.9\times$) further reveals that key-pathway integrity is more critical than value-pathway integrity outside the collapse zone, reflecting the asymmetric roles of attention routing versus content extraction.

Taken together, these findings establish that MLA's bottleneck creates shared representational geometry without creating functional redundancy. The network operates right at the geometric boundary imposed by the bottleneck — sharing the coordinate system while preserving the within-subspace distinction that makes attention computationally effective.

A complementary perspective connects this to prior work Kayyam et al. (2026), which showed that training with $K = V$ from scratch incurs only $\sim$3% perplexity degradation. The contrast is instructive: a model that never develops the within-subspace K–V rotation can learn to function without it. A fully trained MLA model, however, cannot afford to lose a rotation it has learned and relies upon. The two results are not in conflict; they illuminate different sides of the same phenomenon — the within-subspace rotation is not necessary in principle, but becomes necessary once learned.