RoFold: Reconstruction-Free Low-Rank KV Caching under RoPE

Anonymous authors
Paper under double-blind review

Abstract

Low-rank KV caching reduces the storage and bandwidth costs of long-context inference, but pre-RoPE methods reconstruct compressed historical keys to full width before rotation and scoring. We ask whether keys can instead remain latent throughout the read path without giving up a data-adaptive basis. We show that exact latent execution, RoPE invariance of the decoder subspace, and commutation of its orthogonal projector with RoPE are equivalent, and that the execution error of a general subspace decomposes orthogonally into subspace escape and latent-dynamics mismatch. Based on this characterization, we introduce RoFold, which combines (i) a data-adaptive construction of same-frequency cross-head modes and an attention-aware left-inverse encoder fitted by a convex ridge problem without increasing cache width or retained rank, and (ii) a reconstruction-free key path that never rebuilds full-width keys. This path is exactly equivalent, in exact arithmetic, to reconstruct-then-rotate for the same compressed representation, and a frozen unit-level hybrid applies it to 41 of 64 units. On Llama-3.1-8B, at a matched 40.1% KV-reduction budget, RoFold's autoregressive decode advantage over the reconstruct-then-rotate PaLU baseline grows from 9.1% at 8K to 56.4% at 64K, while the relative PPL increase remains below 10% across all evaluated lengths.

1 Introduction

The KV cache has become one of the dominant memory and bandwidth costs of long-context large language model inference. Token eviction (Zhang et al., 2023), quantization (Liu et al., 2024), and feature-dimension compression reduce this cost from different directions. Low-rank compression stores the key and value of each historical token in a smaller latent space, but common pre-RoPE schemes reduce only the persistent storage: before attention can be computed, the cached key is reconstructed to the full head dimension and only then rotated by rotary position embeddings (RoPE) (Su et al., 2024). This reconstruction is repeated across historical tokens and layers at every decode step, so its cost grows with context length and offsets part of the execution benefit of compression. Figure 1 contrasts this reconstruct-then-rotate path with the latent read path we study.

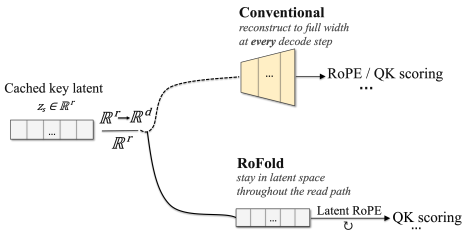


Figure 1: Conventional pre-RoPE caching reconstructs each cached key latent to full width before RoPE/QK at every decode step. RoFold instead keeps historical keys latent throughout the read path.

Keeping the read path latent. We ask whether a cached key can instead remain low-dimensional throughout the read path, with relative RoPE and query–key scoring executed directly in the latent coordinates. This creates a subspace dilemma: conventional low-rank compression is free to choose a basis for representation fidelity, whereas exact latent RoPE requires that the same decoder subspace remain invariant under every RoPE rotation. These

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.

objectives need not select the same subspace; a basis that reconstructs keys well can therefore still fail to support exact latent RoPE. The central question is thus: *can an exactly RoPE-compatible low-rank representation retain enough data-adaptive freedom to avoid full-width reconstruction?*

Existing reconstruction-free designs show that exact compatibility is possible, but through more restricted structures. RAP (Xin et al., 2026) retains whole original RoPE coordinate pairs through structured selection, while CommVQ (Li et al., 2025) imposes a commuting structure in its quantization codebook. These approaches leave open whether the compatible family itself contains a continuous, data-adaptive set of low-rank directions rather than only fixed coordinate choices.

RoFold. We answer this question through a structural property of grouped RoPE: each RoPE frequency is repeated across KV heads. At a fixed frequency, the corresponding coordinates across heads undergo the same rotation, making the exactly compatible family richer than selecting individual coordinate pairs. In particular, continuous same-frequency cross-head mixtures can remain RoPE-compatible. RoFold uses calibration key statistics to select such modes, forming a data-adaptive compatible decoder basis, and then fits an attention-aware left-inverse encoder while keeping that decoder subspace fixed. The resulting compressed representation supports relative RoPE and QK scoring directly in latent space, without reconstructing full-width historical keys.

We study both the all-unit reconstruction-free endpoint, RoFold-Pure, and a frozen unit-level hybrid in which selected units fall back to conventional reconstructed low-rank execution. The two paths use the same prescribed per-unit ranks and persistent-KV payload, exposing a quality–decode-speed trade-off without token eviction or dynamic routing.

Setting and scope. RoFold is a post-hoc, key-side modification of a frozen pretrained RoPE model. It uses only offline calibration and requires no additional LLM pretraining or gradient-based uptraining. Values do not undergo RoPE, so RoFold leaves the shared value-side compression unchanged; our reconstruction-free claim concerns only the historical-key path from cache read through relative RoPE and QK scoring.

The main contributions are:

- We characterize exact latent RoPE execution through decoder-subspace invariance and projector commutation, and decompose the execution error of a general subspace orthogonally into subspace-escape and latent-dynamics-mismatch terms.
- We construct data-adaptive RoPE-compatible bases from continuous same-frequency cross-head modes and fit an attention-aware left-inverse encoder without changing the retained rank or compatible latent dynamics.
- We realize reconstruction-free execution as both an all-unit endpoint and a frozen unit-level hybrid, preserving the same persistent-KV payload while exposing a quality–decode-speed trade-off whose decode advantage increases with context length and improves both quality and speed over a matched truncated-SVD reconstruction control.

2 Preliminaries

2.1 RoPE and Relative-Position Attention

Rotary position embedding (RoPE) (Su et al., 2024) encodes position through a block-diagonal rotation. For a head dimension $d = 2m$, RoPE partitions the coordinates into m two-dimensional planes, each rotated at a frequency θ_j, and writes the rotation at position t as $R_t = \bigoplus_{j=1}^{m} R(t\theta_j)$, where $R(\phi)$ is a 2×2 rotation and $\oplus$ denotes the direct sum of blocks. The rotations are orthogonal, so $R_t^\top R_s = R_{s-t}$; consequently the query–key logit of a token at position t attending to a token at position s depends on position only through the relative offset $\Delta = s - t$,

$$q_t^\top R_t^\top R_s \, k_s = q_t^\top R_{s-t} \, k_s = q_t^\top R_\Delta k_s, \tag{1}$$

where $q_t, k_s \in \mathbb{R}^d$ are the query and key of one head and R_Δ is the relative rotation.

Grouped-query attention (GQA) (Ainslie et al., 2023) shares each key/value head among a group of query heads. We let H_{kv} denote the number of KV heads jointly served by one decoder basis within a group, so that the concatenated group key has width $D_{\text{group}} = H_{\text{kv}} d$. Throughout, d is the per-head dimension, $m = d/2$ the number of RoPE planes, and $r < D_{\text{group}}$ the retained rank.

2.2 Low-Rank KV Caching under RoPE

Low-rank KV caching compresses the per-token key into a smaller latent space. For clarity, we first write the conventional pre-RoPE path for one head: given $k_s \in \mathbb{R}^d$, it caches the latent $z_s = Ek_s$ produced by $E \in \mathbb{R}^{r \times d}$ and reconstructs an approximate full-width key

$$z_s = Ek_s, \qquad \widehat{k}_s = Uz_s, \tag{2}$$

from an orthonormal decoder basis $U \in \mathbb{R}^{d \times r}$ ($U^\top U = I_r$) before RoPE applies; the query is then scored in the full head dimension as $q_t^\top R_{s-t} U z_s$. The deployed GQA construction uses the same notation for a concatenated KV group after replacing d by D_{group}. Because compression is applied before the position-dependent RoPE rotation, the same encoder E and decoder U can be used for keys at every position. At read time, however, the cached latent must still be decoded into a full-width historical key before applying the appropriate relative rotation and QK inner product. Post-RoPE compression instead acts on the already-rotated keys $R_s k_s$, for which a fixed low-rank basis is harder to share across positions (Singhania et al., 2024).

Key-side compression for RoPE models falls into three classes: reconstruction-based low-rank caching, which rebuilds a full-width key before rotation and scoring (Chang et al., 2025; Sun et al., 2025; Lin et al., 2025; Zhu et al., 2026); coordinate-preserving or commuting designs that remove reconstruction but restrict the representable directions, by selecting whole RoPE coordinate pairs (Xin et al., 2026) or constraining the codebook to commute with the rotation (Li et al., 2025); and approaches that alter the positional representation itself (Shi et al., 2025; Ji et al., 2025; Zhou et al., 2025; Wong et al., 2026). Each class resolves only part of the problem: reconstruction-based methods permit general low-rank factors but restore full width; coordinate-preserving or commuting methods stay latent by restricting representable directions; and position-altering methods change the positional mechanism itself. This leaves the central question unresolved: can a frozen RoPE model retain a data-adaptive low-rank key representation while executing relative rotation and QK scoring entirely in latent space? See Appendix A for a fuller review of related key-side compression designs.

Method	No uptrain.	Comp. KV	Adapt. basis	RF latent
Full-Width KV	✓	×	×	n/a
PaLU	✓	✓	×	×
EliteKV	×	✓	×	×
RAP	×*	✓	×	✓
RoFold-Pure	✓	✓	✓	✓

Table 1: Comparison of key-side KV-cache compression designs for RoPE models. Adapt. basis: the decoder basis is selected from calibration key statistics, not fixed or weight-only. RF: reconstruction-free latent RoPE/QK. PaLU (Chang et al., 2025); EliteKV (Zhou et al., 2025); RAP (Xin et al., 2026); *RAP's reported 30% operating point uses LoRA-KD recovery.

3 Motivation: When Can Historical Keys Remain Latent?

3.1 Why Reconstruction Error Is Not Enough

We ask which low-dimensional key subspaces let cached keys undergo RoPE and QK scoring without full-width reconstruction. For $\Delta = s - t$, this requires an $r \times r$ latent rotation $\widehat{R}_\Delta$ whose logit $(U^\top q_t)^\top \widehat{R}_\Delta z_s$ exactly matches the reconstruct-then-rotate reference $q_t^\top R_\Delta U z_s$ for all $q_t \in \mathbb{R}^d$ and $z_s \in \mathbb{R}^r$. Appendix B.1.1 proves that such exact matching is possible if and only if the decoder subspace carries the full-space rotation:

$$R_\Delta U = U \widehat{R}_\Delta, \qquad \forall\, \Delta \in \mathbb{Z}. \tag{3}$$

Exact execution as subspace invariance. Equation (3) says that the decoder subspace $\mathcal{S}_U = \mathrm{range}(U)$ is invariant under every RoPE rotation R_Δ: for any $z \in \mathbb{R}^r$ the rotated vector $R_\Delta U z = U\widehat{R}_\Delta z$ still lies in $\mathcal{S}_U$, so the rotation that originally acts in $\mathbb{R}^d$ is represented exactly by a low-dimensional operator inside the subspace. Equivalently, the orthogonal projector $P = UU^\top$ commutes with R_Δ.

Equation (3) is a structural condition, not a statement about reconstruction quality, and it does not hold automatically for an arbitrary low-rank basis U. With $P = UU^\top$, inserting $PR_\Delta U$ in the middle of the execution gap decomposes the total residual into two named terms,

$$R_\Delta U - U\widehat{R}_\Delta = \underbrace{(I-P)R_\Delta U}_{\text{subspace escape}} + \underbrace{U\big(U^\top R_\Delta U - \widehat{R}_\Delta\big)}_{\text{latent-dynamics mismatch}}. \tag{4}$$

The names follow from where each term lives. The first, left-multiplied by $I - P$, measures the part of $R_\Delta U$ that falls outside the current subspace range(U): this is *subspace escape*, removable only by changing the subspace. The second lies inside range(U) and compares the dynamics RoPE induces there, $U^\top R_\Delta U$, with the operator used: this is *latent-dynamics mismatch*, minimized by $\widehat{R}_\Delta = U^\top R_\Delta U$. The two terms live in range$(U)^\perp$ and range(U) and are therefore Frobenius-orthogonal, and both vanish exactly under Eq. (3). This decomposition reveals a limitation of conventional low-rank objectives: optimizing representation quality alone does not ensure that the learned subspace is closed under RoPE. Consequently, a basis may reconstruct keys accurately yet remain incompatible with exact latent-space RoPE execution. At the logit level, comparing the latent logit with the full-width logit $\ell^{\text{full}}_{s,t} = q_t^\top R_\Delta k_s$ separates these two sources of error,

$$\ell^{\text{full}}_{s,t} - \hat{\ell}_{s,t} = \underbrace{q_t^\top R_\Delta\,(I - UE)k_s}_{\text{representation error}} + \underbrace{q_t^\top\big(R_\Delta U - U\widehat{R}_\Delta\big)Ek_s}_{\text{execution error}}. \tag{5}$$

The first term is the compression *representation error*, jointly determined by U and E and nonzero even when latent execution is exact; the second is the reconstruction-free *execution error*, which vanishes exactly for the subspaces characterized above. See Appendix B.1 for the derivation.

Executability versus representation quality. Equation (5) motivates designing U and E in two stages rather than treating them as an unconstrained low-rank factorization. We first choose U from the RoPE-compatible family and set $\widehat{R}_\Delta = U^\top R_\Delta U$. Then $R_\Delta U - U\widehat{R}_\Delta = 0$, so the execution-error term vanishes identically for every encoder E and every encoded key. With the compatible subspace and its latent dynamics fixed, the remaining approximation is the representation-error term $q_t^\top R_\Delta(I - UE)k_s$. We therefore fit E to retain attention-relevant information in $z_s = Ek_s$, subject to the left-inverse constraint $EU = I_r$. Optimizing E can thus reduce representation error without altering range(U) or reintroducing execution error. Appendix B.1.1 formalizes this separation.

3.2 RoPE-Compatible Subspaces and the Cross-Head Opportunity

We now ask what a basis U satisfying Equation (3) can look like. Appendix B.2.1 answers this when the rotation rates are distinct: the family is rigid — the smallest exactly compatible unit is a complete two-dimensional RoPE plane, and every compatible U is a direct sum of whole planes. Coordinate-preserving designs such as RAP (Xin et al., 2026) stay inside that rigid family by selecting whole original coordinate pairs through a binary selection matrix; exactness then holds by construction, but the admissible bases are discrete, and U has little freedom beyond which pairs to keep.

Across KV heads, however, the family is not rigid. As introduced in Section 2.1, a RoPE frequency is shared by the two or more heads served by one decoder basis, and compatibility survives when those heads are mixed. The family is therefore a continuum rather than a discrete list of whole coordinate pairs (Appendix B.2.2 characterizes it exactly). RoFold uses this freedom, choosing continuous cross-head modes within each same-frequency space from calibration data and selecting among them under a global rank budget (Section 4).

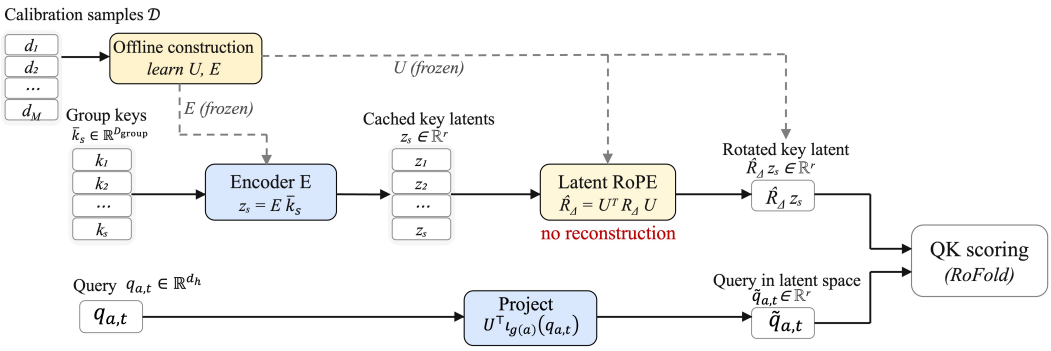


Figure 2: One-time offline construction uses calibration statistics to fix the RoPE-compatible decoder basis U and fit the encoder E. With both frozen, a cached key remains latent through relative RoPE and QK scoring, without full-width reconstruction.

Algorithm 1 RoFold Offline Construction

Require: calibration keys and causal query–key pairs; global rank budget K; ridge coefficient γ
Ensure: basis U and encoder E
1: **for** each frequency index $p = 1, \ldots, m$ **do**
2: **Estimate second moment:** $H_p = \frac{1}{N} \sum_n \xi_{p,n} \xi_{p,n}^{\mathrm{H}}$ over the same-frequency heads
3: **Eigendecompose:** every eigenvector of H_p is a candidate whole mode, scored by $\lambda_i(H_p)$
4: **end for**
5: **Select modes:** rank all modes across planes by score and retain the globally top-K ▷ no cross-plane renormalization
6: **Assemble basis:** place the retained modes at their RoPE frequency coordinates ▷ $U^\top U = I_r$, $r = 2K$
7: **Fit encoder:** solve the causal-QK ridge problem of Eq. (36) for C
8: **return** U and $E = U^\top + C\,(I - UU^\top)$

Figure 4 shows that cross-head energy is substantial at many (layer, KV-group)–frequency pairs, exceeding what the strongest single head pair provides, so this freedom is available in practice; the diagnostic motivates mixed-mode selection but is not itself a downstream quality result.

4 RoFold

RoFold is defined *per unit*, one (layer, KV-group) pair: the key block that shares a single low-rank basis. Units are compressed independently, but they need not be executed alike. Each unit is assigned offline, and frozen before evaluation, to the reconstruction-free path developed here or to a conventional reconstructed path, with the majority following the reconstruction-free one.

Building on the post-hoc, key-side setting of Section 1 and the two-stage separation of Section 3.1, Figure 2 summarizes the two offline stages and the resulting reconstruction-free key path. Offline, a RoPE-compatible decoder basis U is first selected from calibration key statistics; with U fixed, an attention-aware encoder E is then fitted inside it (Algorithm 1). After U and E are frozen, a cached key remains compressed while relative RoPE and QK scoring are performed in latent coordinates, without ever reconstructing a full-width historical key. The subsections below develop the basis construction, the encoder fit, and the reconstruction-free and hybrid execution in turn.

4.1 RoPE-Compatible Decoder Basis

Why selection is still needed. Equation (3) tells us which decoder subspaces are legal for exact latent RoPE, but not which legal subspace is useful for compression. An arbitrary compatible basis can eliminate execution error while discarding important key directions.

We therefore use frozen calibration keys to choose, among compatible rank-r subspaces, one that retains as much key energy as possible. This stage determines *where* keys will be represented; Section 4.2 later determines *how* keys should be mapped there for attention. Because compatibility permits mixing only among coordinates with the same frequency, the selection becomes one small problem per frequency followed by a global rank allocation.

What the spectral construction does. Fix a RoPE plane p. We gather its two pre-RoPE coordinates from every KV head that shares the decoder basis. Since all of these coordinate pairs rotate by the same angle, they may be mixed across heads without breaking compatibility. We write each pair as one complex number only to keep this shared two-dimensional rotation compact: for calibration sample n, the vectors $a_{p,n}, b_{p,n} \in \mathbb{R}^{H_{\mathrm{kv}}}$ contain the first and second coordinate from each head, and $\xi_{p,n} = a_{p,n} + \mathrm{i} b_{p,n}$. Their second moment is

$$H_p = \frac{1}{N} \sum_{n=1}^{N} \xi_{p,n} \xi_{p,n}^{\mathrm{H}} \in \mathbb{C}^{H_{\mathrm{kv}} \times H_{\mathrm{kv}}}, \tag{6}$$

The diagonal entries of H_p are the energies available from individual heads; its off-diagonal entries record cross-head correlations that a mixed mode can exploit. Consequently, the leading eigenvectors of H_p identify the most energetic cross-head mixtures, and their eigenvalues provide directly comparable scores. Each eigenvector produces a real two-dimensional mode consisting of one direction and its quarter-turn. RoPE rotates within this pair, so the mode remains compatible even when it combines several heads. The complex notation is used only for construction; the deployed basis and latent operator are real. Appendix B.3.1 gives the corresponding real identities.

What this produces. Every frequency supplies a short list of compatible candidate modes, each scored by how much calibration-key energy it retains. We pool these lists and keep the globally top K (Algorithm 1), so informative frequencies may receive several modes while weak frequencies receive none. Each candidate contributes its complete two-dimensional pair—a mode is never split—and the selected pairs are assembled into the decoder basis U. The result is a data-adaptive rank-r subspace that preserves exact latent RoPE while using the available rank where the calibration keys carry the most energy. The construction needs only one pass of second-order statistics and small eigendecompositions, with no model training or nonconvex optimization. Appendix B.3.3 proves compatibility and optimality, and Appendix C.5 verifies the deployed bases for all 64 units.

4.2 Attention-Aware Encoder Fitting

Why $U^\top$ is not enough. Because the decoder reconstructs a latent key as $\hat{k} = Uz$ and U has orthonormal columns, its transpose is the natural encoder: $z = U^\top k$, followed by the orthogonal reconstruction $\hat{k} = UU^\top k$. This map is exact for keys already in range(U), but it discards every component outside that subspace. Those discarded components can nevertheless affect the query–key score. Thus choosing U fixes where a key can be represented and rotated exactly, but $U^\top$ need not be the best way to preserve the information that particular queries rely on.

What the fitted encoder does. We improve the encoding without changing the compatible subspace selected above. Decompose a key into the part inside range(U) and its orthogonal complement. The first part already has exact latent coordinates $U^\top k$; the second is allowed to influence those same coordinates through a learned correction C:

$$E = \underbrace{U^\top}_{\text{orthogonal encoder}} + \underbrace{C\,(I - UU^\top)}_{\text{complement correction}}, \tag{7}$$

where $C \in \mathbb{R}^{r \times D_{\mathrm{group}}}$ is the only fitted part. The first term captures the component of k inside the selected subspace. The second captures information outside that subspace that would otherwise be discarded by the simple projection $U^\top k$, with the learnable map C deciding which of this information to retain in the latent coordinates. If k already lies in the selected subspace, then $(I - UU^\top)k = 0$ and the correction vanishes automatically.

How the correction is learned. Not all information outside the selected subspace is equally useful: we only want to retain the part that affects attention scores. We therefore freeze the model and U, collect causal query–key pairs, and fit C so that the resulting latent QK scores match their full-width counterparts. In effect, $U^\top$ provides the initial encoding and C corrects the score error caused by information that this projection would discard. This is a convex ridge-regression problem performed once before evaluation, with no backbone update. Because $E(Uz) = z$ (equivalently, $EU = I_r$), the fit leaves the decoder basis and its latent RoPE geometry unchanged. Appendices B.5.1 and B.5.2 give the formal characterization and objective. Once fitted, E changes only how each latent is formed; it adds neither persistent cache width nor online reconstruction.

4.3 Reconstruction-Free and Hybrid Execution

Pure endpoint. With U and E obtained offline, RoFold-Pure caches $z_s = Ek_s \in \mathbb{R}^r$ directly — never the full-width k_s — and executes relative rotation and scoring with the latent operator $\widehat{R}_\Delta = U^\top R_\Delta U$. By Eq. (3) this reproduces the reconstruct-then-rotate logit exactly in exact arithmetic under the same (U, E), so the all-unit reconstruction-free path is the theoretical endpoint that isolates the structural contribution; the finite-rank compression error itself still remains. See Appendix B.4 for the per-query-head bookkeeping.

Frozen hybrid routing. Compatibility makes the representation directly executable but can also restrict which directions are available at a fixed rank. RoFold-Hybrid therefore treats reconstruction-free execution as a frozen unit-level option. Using the unit definition at the start of Section 4, our Llama-3.1-8B instantiation has 64 units in total. Let $\mathcal{N}$ denote this set and $\mathcal{S} \subseteq \mathcal{N}$ those assigned to the conventional reconstructed low-rank fallback. Units in $\mathcal{N} \setminus \mathcal{S}$ use the compatible RoFold path, whereas units in $\mathcal{S}$ use an unconstrained low-rank basis and reconstruct their historical keys before RoPE. Both paths store low-dimensional persistent caches at the same prescribed rank and payload; the assignment is selected offline, fixed before evaluation, and shared by all tokens and contexts, and is neither token eviction nor dynamic routing. Writing $\mathcal{L}(\mathcal{S})$ for a calibration loss and c_u for a reconstruction-work proxy of fallback unit u, the routing problem can be expressed as

$$\min_{\mathcal{S} \subseteq \mathcal{N}} \ \mathcal{L}(\mathcal{S}) \quad \text{subject to} \quad \sum_{u \in \mathcal{S}} c_u \le B_{\text{rec}}, \tag{8}$$

where B_{rec} is the reconstruction-work budget. The pure endpoint is $\mathcal{S} = \varnothing$ and the fully reconstructed endpoint assigns every unit to the fallback path. The composition is dispatched statically; Section 5.4 reports the pure and deployed compositions in Figure 3b.

5 Experiments

5.1 Experimental Setup

Model and baselines. We evaluate RoFold on Llama-3.1-8B under a shared frozen-backbone harness. The quantitative baselines are the full-width KV cache, PaLU (Chang et al., 2025), and a matched-budget truncated-SVD control. The latter applies rank-r truncated SVD to the `k_proj` projection of each unit, using the same key-side rank table but no weighting, whitening, or calibration, and reconstructs historical keys to full width before RoPE and QK scoring. Because PaLU, Truncated SVD, and RoFold share the same rank budget and payload, they form the aligned quantitative comparison; the other RoPE-aware compressors in Table 1 are compared qualitatively.

Matched cache budget. All compressed arms use the same discrete per-unit rank table, realizing a 40.1% persistent-KV reduction; Appendix C.6 gives the exact byte count. The value path is fixed to the same PaLU compression for every compressed arm (65,536 → 57,344 B/token, uncoupled from RoPE). PaLU, Truncated SVD, and RoFold differ only on the key side (65,536 → 21,120 B/token), so their quality differences isolate the key-side method.

Table 2: Matched-budget quality and decode throughput on Llama-3.1-8B (protocol in Section 5.1). Among the compressed methods (Truncated SVD, PaLU, RoFold), bold marks the best and underline the second best at each context length.

Method	KV reduction	8K		16K		32K		64K	
		PPL↓	tok/s↑	PPL↓	tok/s↑	PPL↓	tok/s↑	PPL↓	tok/s↑
Full-Width KV	–	6.2439	6.456±0.193	6.5891	5.079±0.068	6.6725	3.169±0.137	6.9344	1.823±0.062
Truncated SVD	40.1%	9.4321	2.866±0.149	9.8172	2.001±0.098	10.0789	<u>1.355</u>±0.012	10.3868	<u>0.803</u>±0.003
PaLU	40.1%	**7.4946**	<u>3.020</u>±0.025	**7.8631**	<u>2.104</u>±0.028	**7.9028**	1.297±0.066	**8.0824**	0.737±0.022
RoFold	40.1%	<u>8.0025</u>	**3.294**±0.054	<u>8.4043</u>	**2.597**±0.021	<u>8.5717</u>	**1.715**±0.120	<u>8.8329</u>	**1.153**±0.050

Data and metrics. Calibration data is disjoint from the held-out evaluation corpus. The 8K, 16K, and 32K evaluation spans are mutually disjoint and disjoint from calibration; we additionally report 64K on a frozen held-out span. We measure held-out teacher-forced perplexity (PPL), negative log-likelihood per token, persistent KV bytes per token, and decode throughput on a single A100 80GB PCIe GPU. Throughout the paper, *throughput* is batch-1 steady-state autoregressive decode—the per-step rate once the cache is populated—averaged over three independent trials with the sample standard deviation. Appendix C summarizes the data splits, byte accounting, and runtime protocol.

Construction checks and ablations. The deployed RoFold basis follows the individual-column recipe of Appendix C.1 and is numerically certified equivalent to the canonical whole-mode construction of Section 4.1. Appendix C.4 separately isolates the contributions of the continuous cross-head compatible family and the attention-aware encoder through controlled mechanism ablations.

5.2 Matched-Budget Quality and Memory

Same memory footprint. Every compressed method reduces the persistent KV payload by 40.1%. Within RoFold, both execution paths use the same frozen per-unit rank table and cache only rank-dimensional latents. Routing a unit therefore changes its read-time computation, not its stored payload. The deployed 41/23 hybrid, PaLU, and Truncated SVD are consequently compared at the same memory footprint; Appendix C.6 gives the exact accounting.

Quality at this budget. Table 2 shows a consistent result at all four context lengths. Compared with the conventional truncated-SVD reconstruct-then-rotate path, RoFold achieves both lower PPL and higher decode throughput. PaLU remains the strongest quality-oriented compressed baseline; without LLM uptraining, RoFold has a 6.8–9.3% relative PPL gap, equivalent to only 0.066–0.089 additional nats per token, in exchange for the throughput gains studied next. Full-Width KV attains the lowest PPL but uses the uncompressed payload, so it serves as a quality reference rather than a matched-memory baseline; the same holds for throughput. Full-Width KV also decodes fastest in Table 2 because it never reconstructs a historical key: every reconstruction-based compressed arm (Truncated SVD, PaLU, and the 23 reconstructed units of the deployed hybrid) pays a per-step rebuild cost that a larger, but simpler, cache read does not, so compressing the persistent payload does not by itself guarantee a faster decode step. RoFold's throughput claims are therefore stated only relative to the other compressed, matched-memory arms, where the reconstruction-free path removes exactly this rebuild cost; Full-Width KV is excluded from that comparison for the same reason it is excluded from the quality one.

5.3 Decode Scaling with Context Length

Growth with context. The quality cost grows much more slowly than the speed benefit. Relative to PaLU, the PPL gap rises from 6.8% at 8K to 9.3% at 64K, while the autoregressive decode advantage expands from 9.1% to 56.4%; the intermediate throughput gains are 23.4% at 16K and 32.2% at 32K. Figure 3a visualizes this widening speed advantage. A reconstruction-free unit never rebuilds a full-width key before rotation and scoring, so the reconstruction work retained by PaLU becomes an increasingly large fraction of each

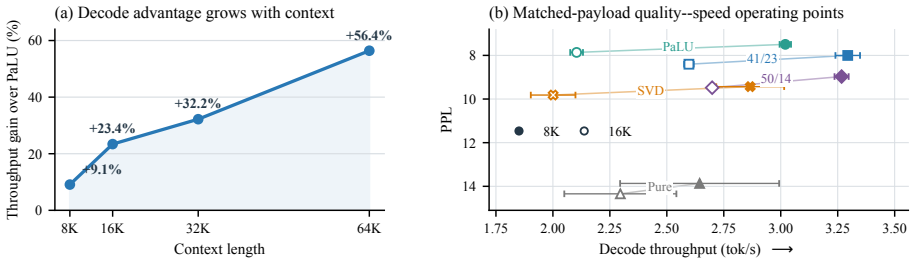


Figure 3: (a) Decode-throughput speedup of the deployed 41/23 hybrid over PaLU at the matched 40.1% budget; see Table 2. (b) Held-out PPL versus decode throughput at 8K (filled) and 16K (open) for Truncated SVD, PaLU, and three RoFold routing compositions with 23, 14, and 0 reconstructed units respectively (deployed 41/23; 50/14; all-unit RoFold-Pure); horizontal error bars show one standard deviation of decode throughput over three trials, PPL is deterministic.

decode step as context grows. Section 5.1 defines throughput as the held-out decode rate with a populated cache.

5.4 Deployed Quality–Speed Operating Point

Just as low-rank compression trades quality for memory, the frozen 41/23 hybrid trades quality for decode throughput at the same budget; the full-width comparison remains a memory comparison only, needing neither latent encoding nor reconstruction. Figure 3b situates this point among the alternatives available at 8K and 16K, all measured under the same frozen benchmark protocol: a 50/14 composition sits close to the deployed 41/23 point in throughput but gives up additional quality, while RoFold-Pure, the all-unit reconstruction-free endpoint of Section 4.3, reaches the highest decode rate at a substantially larger quality cost. Relative to the quality-oriented PaLU reconstruction baseline, RoFold provides a substantially more favorable systems operating point, trading a modest quality gap for an increasingly large decode-throughput advantage as context grows; the deployed 41/23 composition keeps most of RoFold-Pure's throughput benefit while keeping the quality gap to PaLU far smaller than the reconstruction-free endpoint alone would allow.

6 Conclusion

We asked whether a compressed key can stay latent under RoPE without giving up a data-adaptive basis. It can. Exact latent execution is exactly RoPE invariance of the decoder subspace, and frequencies repeated across KV heads make that family rich enough for calibration to choose data-adaptive cross-head modes rather than only fixed coordinate pairs, with an attention-aware encoder recovering a substantial fraction of the remaining quality. RoFold-Pure realizes this for every unit; RoFold-Hybrid reconstructs only where it helps. The deployed 41/23 hybrid trades a modest quality gap for a growing decode-throughput advantage, a genuine systems trade-off.

AI use statement

The authors used an LLM solely for grammar and clarity. All scientific claims were produced and verified by the authors, and the LLM did not generate or modify any experimental data.

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, 2023.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, et al. Palu: KV-cache compression with low-rank projection. In *International Conference on Learning Representations (ICLR)*, 2025.

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Junyan Li, Yang Zhang, Muhammad Yusuf Hassan, Talha Chafekar, Tianle Cai, Zhile Ren, et al. CommVQ: Commutative vector quantization for KV cache compression. In *International Conference on Machine Learning (ICML)*, 2025.

Bokai Lin, Zihao Zeng, Zipeng Xiao, Siqi Kou, TianQi Hou, Xiaofeng Gao, et al. MatryoshkaKV: Adaptive KV compression via trainable orthogonal projection. In *International Conference on Learning Representations (ICLR)*, 2025.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, et al. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *International Conference on Machine Learning (ICML)*, 2024.

Luohe Shi, Zuchao Li, Lefei Zhang, Baoyuan Qi, Guoming Liu, and Hai Zhao. KV-Latent: Dimensional-level KV cache reduction with frequency-aware rotary positional embedding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Prajwal Singhania, Siddharth Singh, Shwai He, Soheil Feizi, and Abhinav Bhatele. Loki: Low-rank keys for efficient sparse attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Hanshi Sun, Li-Wen Chang, Wenlei Bao, Size Zheng, Ningxin Zheng, Xin Liu, et al. ShadowKV: KV cache in shadows for high-throughput long-context LLM inference. In *International Conference on Machine Learning (ICML)*, 2025.

Jeffrey T. H. Wong, Cheng Zhang, Xinye Cao, Pedro Gimenes, Christos-Savvas Bouganis, and George Constantinides. A3: An analytical low-rank approximation framework for attention. In *International Conference on Machine Learning (ICML)*, 2026.

Jihao Xin, Tian Lyu, David Keyes, Hatem Ltaief, and Marco Canini. RAP: KV-cache compression via RoPE-aligned pruning. *arXiv preprint arXiv:2602.02599*, 2026.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, et al. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuhao Zhou, Sirui Song, Boyang Liu, Zhiheng Xi, Senjie Jin, Xiaoran Fan, Zhihao Zhang, Wei Li, and Xuanjing Huang. EliteKV: Scalable KV cache compression via RoPE frequency selection and joint low-rank projection. *arXiv preprint arXiv:2503.01586*, 2025.

Yuxuan Zhu, David H. Yang, Mohammad Mohammadi Amiri, Keerthiram Murugesan, Tejaswini Pedapati, and Pin-Yu Chen. OjaKV: Context-aware online low-rank KV cache compression. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 10161–10178, 2026.

Appendix

Overview

This appendix provides related work, full proofs, and experimental details supporting the main text. It is organized as follows:

- **Appendix A (Related Work):** reviews related key-side compression designs for RoPE models.
- **Appendix B (Theoretical Analysis):** proves the equivalence between exact latent execution, subspace invariance, and projector commutation; decomposes the execution error; characterizes RoPE-compatible subspaces at distinct and repeated frequencies; derives the data-adaptive, rank-budgeted whole-mode construction; extends the result to grouped-query attention; and parameterizes the fitted encoder with its convex ridge objective.
- **Appendix C (Experimental Details and Additional Results):** gives the construction provenance and models and baselines, the cross-head calibration diagnostic, position-conditioned stability of the compatible directions, the mechanism ablations isolating the cross-head basis family and the fitted encoder, the numerical verification of the deployed basis and encoder, persistent-KV byte accounting, the static routing protocol, and limitations.

A Related Work

Token eviction (Zhang et al., 2023) and quantization (Liu et al., 2024) are complementary to the feature-dimension compression studied here. PaLU (Chang et al., 2025), ShadowKV (Sun et al., 2025), MatryoshkaKV (Lin et al., 2025) and OjaKV (Zhu et al., 2026) demonstrate the promise of low-rank KV caching but mainly optimize reconstruction or task fidelity while retaining a full-width key on the read path. RAP (Xin et al., 2026) preserves the original RoPE coordinate pairs through structured, whole-pair selection with a binary selection matrix, and thereby also removes the reconstruction step. Other RoPE-aware approaches alter the positional representation itself: KV-Latent redesigns the RoPE frequencies in the compressed representation (Shi et al., 2025), MHA2MLA decouples a position-dependent component from the latent representation (Ji et al., 2025), and EliteKV and A3 exploit the frequency structure to allocate representational capacity (Zhou et al., 2025; Wong et al., 2026). CommVQ imposes a commuting constraint in its discrete quantization codebook (Li et al., 2025).

Relation to RAP and reconstruction-based low-rank caching. RAP removes the reconstruction step by selecting whole original RoPE coordinate pairs through a binary selection matrix, so its compatible family is discrete: each retained direction lies inside a single coordinate pair. RoFold keeps the same reconstruction-free execution but replaces that discrete selection with continuous same-frequency cross-head modes, whose coefficient vector is a general complex direction in the same-frequency space, and then fits an attention-aware left-inverse encoder inside the resulting basis. Relative to reconstruction-based low-rank caching (Chang et al., 2025; Sun et al., 2025; Lin et al., 2025; Zhu et al., 2026), the difference is where rotation acts: those methods rebuild a full-width key before RoPE, whereas RoFold rotates and scores the latent directly.

B Theoretical Analysis of Reconstruction-Free RoPE Execution

We first characterize when a compressed decoder subspace supports exact latent RoPE execution. We then classify compatible subspaces, derive the data-adaptive cross-head construction, extend the result to GQA, and finally analyze the attention-aware encoder fit.

B.1 Exact Latent Execution

B.1.1 Equivalence of Invariance, Commutation, and Path Equality

Conventional pre-RoPE compression caches $z_s = Ek_s$ for an encoder $E \in \mathbb{R}^{r\times d}$ and an orthonormal decoder basis $U \in \mathbb{R}^{d\times r}$, reconstructing the full-width key $\hat{k}_s = Uz_s$ before rotation, so the reconstructed key lies in $\mathcal{S}_U = \mathrm{range}(U)$, $\dim(\mathcal{S}_U) = r$. Since RoPE satisfies $R_t^\top R_s = R_{s-t}$ (Su et al., 2024), the logit depends on the two positions only through $\Delta = s - t$. The reconstruction-free path instead applies an $r \times r$ operator $\widehat{R}_\Delta$ to the cached latent, with the query entering through $U^\top q_t$,

$$\hat{\ell}_{s,t} = (U^\top q_t)^\top \widehat{R}_{s-t}\, z_s, \tag{9}$$

so that historical keys are never restored to width d. Henceforth

$$U^\top U = I_r, \qquad \mathcal{U} = \mathrm{range}(U), \qquad P = UU^\top,$$

so that P is the orthogonal projector onto $\mathcal{U}$ and $\mathcal{U}^\perp$ is its orthogonal complement; R_Δ is orthogonal.

Theorem 1 (Equivalence conditions for exact execution at a fixed offset)**.** *Fix $\Delta \in \mathbb{Z}$. The following four conditions are equivalent, and in conditions 1 and 4 the matrix $\widehat{R}_\Delta \in \mathbb{R}^{r\times r}$ is required to satisfy the stated identity at that* same *fixed offset:*

1. *there exists $\widehat{R}_\Delta \in \mathbb{R}^{r\times r}$ such that $R_\Delta U = U\widehat{R}_\Delta$;*
2. *$\mathcal{U}$ is invariant under R_Δ, i.e. $R_\Delta \mathcal{U} = \mathcal{U}$;*
3. *the orthogonal projection P commutes with R_Δ, i.e. $R_\Delta P = PR_\Delta$;*
4. *for any $q \in \mathbb{R}^d$ and $z \in \mathbb{R}^r$, $q^\top R_\Delta U z = (U^\top q)^\top \widehat{R}_\Delta z$.*

When the conditions hold, the latent-space operator is unique and equal to $\widehat{R}_\Delta = U^\top R_\Delta U$.

Proof of Theorem 1.

Proof. We prove the three equivalences $1 \Leftrightarrow 4$, $1 \Leftrightarrow 2$ and $2 \Leftrightarrow 3$; the four conditions are then mutually equivalent.

$1 \Leftrightarrow 4$. The right-hand side of condition 4 is $q^\top U\widehat{R}_\Delta z$, so the identity holds for all q, z if and only if

$$q^\top R_\Delta U z = q^\top U\widehat{R}_\Delta z \;\; \forall q, z \quad \Longleftrightarrow \quad R_\Delta U = U\widehat{R}_\Delta.$$

$1 \Leftrightarrow 2$, *with uniqueness.* If condition 1 holds, then for $x = Uz \in \mathcal{U}$,

$$R_\Delta x = R_\Delta U z = U\widehat{R}_\Delta z \in \mathcal{U},$$

so $R_\Delta \mathcal{U} \subseteq \mathcal{U}$; since R_Δ is invertible and preserves dimension, this inclusion is an equality and condition 2 holds. Conversely, condition 2 gives $R_\Delta U = US_\Delta$ for some S_Δ, and left-multiplying by $U^\top$ yields

$$S_\Delta = U^\top R_\Delta U,$$

which is condition 1 and proves uniqueness.

$2 \Leftrightarrow 3$. If $R_\Delta \mathcal{U} = \mathcal{U}$ then $R_\Delta^\top \mathcal{U} = \mathcal{U}$, because $R_\Delta^{-1} = R_\Delta^\top$; hence for $x \in \mathcal{U}^\perp$ and $u \in \mathcal{U}$,

$$\langle R_\Delta x, u\rangle = \langle x, R_\Delta^\top u\rangle = 0,$$

so $\mathcal{U}^\perp$ is invariant as well and R_Δ commutes with P. Conversely, $R_\Delta P = PR_\Delta$ gives $R_\Delta x = R_\Delta Px = PR_\Delta x \in \mathcal{U}$ for $x = Px \in \mathcal{U}$, which is condition 2. □

Theorem 1 immediately gives path equivalence: with $z = Ek$ and condition 1 at the same offset,

$$q^\top R_\Delta U E k = q^\top U\widehat{R}_\Delta E k = (U^\top q)^\top \widehat{R}_\Delta z, \tag{10}$$

logit by logit. When the conditions hold for every Δ, the latent operator also satisfies three identities — orthogonality, composition and inversion — each a one-line consequence of condition 1 at the relevant offsets. Orthogonality follows from $UU^\top R_\Delta U = U\widehat{R}_\Delta = R_\Delta U$ and

$$\widehat{R}_\Delta^\top \widehat{R}_\Delta = U^\top R_\Delta^\top U U^\top R_\Delta U = U^\top R_\Delta^\top R_\Delta U = I_r,$$

while applying condition 1 at Δ_1, Δ_2 and $\Delta_1 + \Delta_2$ gives $\widehat{R}_{\Delta_1+\Delta_2} = \widehat{R}_{\Delta_1}\widehat{R}_{\Delta_2}$, and setting $\Delta_2 = -\Delta$ gives $\widehat{R}_\Delta \widehat{R}_{-\Delta} = \widehat{R}_0 = U^\top U = I_r$, hence $\widehat{R}_{-\Delta} = \widehat{R}_\Delta^{-1} = \widehat{R}_\Delta^\top$. The latent operator is thus the rotation itself, seen inside the retained subspace, which is what licenses the term *latent RoPE*. The identity $UEk = k$ is nowhere used, so this is not equivalent to lossless compression.

B.1.2 Orthogonal Decomposition of Execution Error

Proposition 2 (Execution-error decomposition)**.** *Let U have orthonormal columns and let $\widehat{R}_\Delta$ be any $r \times r$ matrix.*

(i) Inserting $PR_\Delta U$ between the two execution paths gives

$$R_\Delta U - U\widehat{R}_\Delta = (I-P)R_\Delta U + U\big(U^\top R_\Delta U - \widehat{R}_\Delta\big), \tag{11}$$

whose two terms are Frobenius-orthogonal because $(I-P)U = 0$. Hence

$$\|R_\Delta U - U\widehat{R}_\Delta\|_F^2 = \|(I-P)R_\Delta U\|_F^2 + \|U^\top R_\Delta U - \widehat{R}_\Delta\|_F^2, \tag{12}$$

where left-multiplication by U preserves the Frobenius norm. The first term, the subspace-escape *term, can be removed only by changing the decoding subspace; the second, the* dynamics-mismatch *term, is minimized over $\widehat{R}_\Delta$ by $\widehat{R}_\Delta = U^\top R_\Delta U$. Both vanish simultaneously exactly when the conditions of Theorem 1 hold.*

(ii) Comparing the full-width logit $\ell_{s,t}^{\text{full}} = q_t^\top R_\Delta k_s$ with (9), and adding and subtracting $q_t^\top R_\Delta U E k_s$, gives

$$\ell_{s,t}^{\text{full}} - \hat{\ell}_{s,t} = q_t^\top R_\Delta (I - UE) k_s + q_t^\top \big(R_\Delta U - U\widehat{R}_\Delta\big) E k_s. \tag{13}$$

The first term is the compression representation error *from replacing k_s by UEk_s; it is jointly determined by U and E and persists even when latent execution is exact. The second is the reconstruction-free* execution error*.*

Proof of Proposition 2.

Proof. (i) Decomposition and orthogonality. Inserting $PR_\Delta U$ and using $(I-P)U = 0$ gives the displayed decomposition, whose cross term is

$$\operatorname{tr}\big[(R_\Delta U)^\top (I-P)U(U^\top R_\Delta U - \widehat{R}_\Delta)\big] = 0;$$

the two summands are therefore Frobenius-orthogonal, and with $\|UM\|_F^2 = \|M\|_F^2$ whenever $U^\top U = I_r$ the displayed Pythagorean identity for $\|R_\Delta U - U\widehat{R}_\Delta\|_F^2$ follows. The first term (subspace escape) depends on U only through range(U) and vanishes exactly when $R_\Delta U \subseteq$ range(U); the second (dynamics mismatch) is minimized by $\widehat{R}_\Delta = U^\top R_\Delta U$, and both vanish together exactly under Theorem 1.

(ii) Logit-level error. Substituting $\hat{\ell}_{s,t} = (U^\top q_t)^\top \widehat{R}_\Delta E k_s$ and $z_s = Ek_s$ and adding and subtracting $q_t^\top R_\Delta U E k_s$ groups the error by whether it carries $(I - UE)$ or $(R_\Delta U - U\widehat{R}_\Delta)$, yielding (13): the first term is the compression *representation error* (persisting even when latent execution is exact) and the second the reconstruction-free *execution error*. □

B.2 Characterization of RoPE-Compatible Subspaces

Theorem 1 leaves open which subspaces are compatible. This appendix characterizes them in the two cases that occur in practice — distinct rotation rates within a head, and rates repeated across the KV heads that share one decoder basis — and then constructs the decoder basis from calibration key statistics. Section B.2.1 treats the distinct-rate case, where compatibility is purely within a head, and Section B.2.2 treats the repeated-rate case, where a single rotation acts jointly on the same frequency plane of all shared KV heads and complex cross-head modes become available; together they characterize the compatible family under their stated frequency assumptions.

B.2.1 Distinct-Frequency RoPE Planes

Coordinate convention. Write $d = 2m$ and suppose first that the rotation rates are distinct,

$$R_1 = \bigoplus_{j=1}^{m} R(\theta_j), \qquad 0 < \theta_j < \pi, \qquad \theta_j \neq \theta_k \;\; (j \neq k), \tag{14}$$

where $R(\phi)$ is the two-dimensional rotation. Identifying the j-th coordinate pair with $z_j = x_{2j-1} + \mathrm{i}x_{2j}$, R_1 acts on $\mathbb{C}^m$ as $D = \mathrm{diag}(e^{\mathrm{i}\theta_1}, \ldots, e^{\mathrm{i}\theta_m})$, and every real linear T decomposes uniquely as $T(z) = Bz + G\bar{z}$. The commutant of R_1 is computed in the proof of Proposition 3 below and consists exactly of those maps that multiply an independent complex scalar on each plane, so that they can neither mix frequencies nor contain an antilinear part.

Proposition 3 (Complete planes are the exact-compatibility unit at distinct rates)**.** *A real linear map T that commutes with $R_1 = \bigoplus_{j=1}^{m} R(\theta_j)$ with distinct rates $0 < \theta_j < \pi$ multiplies an independent complex scalar on each plane: it can neither mix frequencies nor contain an antilinear part. Consequently, if the rotation rates are distinct, a real subspace is invariant under R_Δ for every $\Delta \in \mathbb{Z}$ if and only if it is a direct sum of complete planes.*

Proof of Proposition 3.

Proof. Commutant computation. The commutation $T(Dz) = DT(z)$ forces the entrywise conditions

$$(e^{\mathrm{i}\theta_k} - e^{\mathrm{i}\theta_j})B_{jk} = 0, \qquad (e^{-\mathrm{i}\theta_k} - e^{\mathrm{i}\theta_j})G_{jk} = 0. \tag{15}$$

Distinctness makes the first factor nonzero for $j \neq k$, so B is complex diagonal, while $e^{-\mathrm{i}\theta_k} = e^{\mathrm{i}\theta_j}$ would give $\theta_j + \theta_k \equiv 0 \pmod{2\pi}$, impossible since $0 < \theta_j + \theta_k < 2\pi$, so $G = 0$. A real linear map commuting with R_1 therefore multiplies an independent complex scalar on each plane, and can neither mix frequencies nor contain an antilinear part.

Compatible subspaces. Sufficiency is immediate, since a complete plane is rotated within itself. For necessity, a projector commuting with R_Δ for all Δ commutes with R_1, so the commutant computation above applies with $T = P$: $P = B = \mathrm{diag}(b_1, \ldots, b_m)$ with $b_j \in \mathbb{C}$. Self-adjointness forces $b_j \in \mathbb{R}$ and $P^2 = P$ forces $b_j^2 = b_j$, so $b_j \in \{0, 1\}$ and the projection is a direct sum of complete planes,

$$P = \bigoplus_{j=1}^{m} b_j I_2, \qquad b_j \in \{0, 1\}, \qquad \mathrm{range}(P) = \bigoplus_{j \in \mathcal{I}} \mathrm{span}(e_{2j-1}, e_{2j}), \tag{16}$$

for some index set $\mathcal{I} \subseteq \{1, \ldots, m\}$. □

B.2.2 Repeated Frequencies and Cross-Head Complex Modes

Now let one rotation rate recur across the H_{kv} KV heads served by a single decoder basis, and write $D_{\mathrm{group}} = H_{\mathrm{kv}} d$ for the concatenated group width. Plane p is shared by all heads, and in the concatenated key its coordinates form the ordered list

$$I_p = \big(p,\ d+p,\ \ldots,\ (H_{\mathrm{kv}}-1)d+p\big) \,\|\, \big(p+m,\ d+p+m,\ \ldots,\ (H_{\mathrm{kv}}-1)d+p+m\big), \tag{17}$$

the cos-halves of the H_{kv} heads first and their sin-halves second, the I_p pairwise disjoint and covering all D_{group} coordinates. The joint rotation on this block is generated by the quarter-turn

$$J_p = \begin{pmatrix} 0 & -I_{H_{\text{kv}}} \\ I_{H_{\text{kv}}} & 0 \end{pmatrix}, \qquad J_p^\top = -J_p, \qquad J_p^2 = -I_{2H_{\text{kv}}}, \\ R_\Delta^{(p)} = \cos(\Delta\theta_p) I + \sin(\Delta\theta_p) J_p, \tag{18}$$

written in this *half-split block ordering*, which we use throughout.

Coordinate convention. We write J for the block complex structure, i.e. the block-diagonal operator acting as (18) on each plane block, so that $J = \bigoplus_p J_p$, $J^\top = -J$ and $J^2 = -I_{D_{\text{group}}}$, and we call a subspace J*-closed* when it is invariant under J. Write $\Omega = \bigoplus_p \theta_p J_p$ for the group's RoPE generator in the block ordering $\{I_p\}$, so that $R_\Delta = \exp(\Delta\Omega)$. The adjacent-pair picture used for intuition in the main text is a different coordinate ordering of the same rotation: if Π carries the interleaved concatenated key to the half-split one, $\bar{k}_{\text{half}} = \Pi\, \bar{k}_{\text{interleaved}}$, then the assembled generator and its plane-local form read

$$\bigoplus_p J_p^{\text{int}} = \Pi^\top \Big(\bigoplus_p J_p \Big) \Pi, \qquad J_p^{\text{int}} = I_{H_{\text{kv}}} \otimes \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \pi^\top J_p\, \pi, \tag{19}$$

where π is the $2H_{\text{kv}} \times 2H_{\text{kv}}$ plane-local restriction of Π to the coordinates I_p (the same matrix for every p). Claims about range(U), J-closure and the latent operator are invariant under Π, but the *coordinate form* of the realification map is not, so (20), the columns (26) and every displayed J_p below are written in the half-split block ordering. Finally, identify $w \in \mathbb{R}^{2H_{\text{kv}}}$ with the complex vector

$$\zeta(w) = \big(w_0 + \mathrm{i} w_{H_{\text{kv}}},\ \ldots,\ w_{H_{\text{kv}}-1} + \mathrm{i} w_{2H_{\text{kv}}-1} \big) \in \mathbb{C}^{H_{\text{kv}}}, \tag{20}$$

an $\mathbb{R}$-linear isomorphism under which the block quarter-turn (18) acts as multiplication by i: splitting $w = (a; b)$ into its cos- and sin-halves gives $J_p w = (-b; a)$ and $\zeta(J_p w) = \mathrm{i}\,\zeta(w)$. Write $\mathcal{R}(M)$ for the realification of a complex matrix M under this identification, and let ϕ be the inverse of (20).

Proposition 4 (Whole same-frequency modes at a repeated rate)**.** *On a same-frequency block, a real subspace $\mathcal{W} \subseteq \mathbb{R}^{2H_{\text{kv}}}$ is invariant under $R_\Delta^{(p)}$ for every $\Delta \in \mathbb{Z}$ if and only if it is J_p-invariant, equivalently the realification of a $\mathbb{C}$-linear subspace of $\mathbb{C}^{H_{\text{kv}}}$, i.e. a direct sum of complex lines $\mathbb{C}\zeta$. In particular its real dimension is even, and it is $2k$ rather than invariably 2 when a complex eigenspace has dimension k.*

Proof of Proposition 4.

Proof. Generator equivalence. Since $R_1^{(p)} = \cos(\theta_p) I + \sin(\theta_p) J_p$ with $\sin(\theta_p) > 0$, invariance under all $R_\Delta^{(p)}$ implies invariance under

$$J_p = \frac{R_1^{(p)} - \cos(\theta_p) I}{\sin(\theta_p)},$$

and J_p-invariance conversely gives invariance under every $R_\Delta^{(p)}$.

Complex-line decomposition. Under (20), J_p acts as multiplication by i, so a J_p-invariant real subspace is the realification of the $\mathbb{C}$-linear subspace it spans under complex scaling, and conversely. A complex line $\mathbb{C}\zeta$ realifies to the real two-plane

$$\operatorname{span}\{\phi(\zeta),\ J_p \phi(\zeta)\},$$

and a complex subspace of dimension k to a direct sum of k such planes, so the real dimension is $2k$. □

Together, Proposition 3 and Proposition 4 characterize the RoPE-compatible family under their stated frequency assumptions. At distinct rates a real orthogonal projection P commuting with R_Δ for every $\Delta \in \mathbb{Z}$ is a direct sum of whole coordinate planes,

$$P = \bigoplus_{j=1}^{m} b_j I_2, \qquad b_j \in \{0,1\}, \qquad \mathrm{range}(P) = \bigoplus_{j\in\mathcal{I}} \mathrm{span}(e_{2j-1}, e_{2j}), \tag{21}$$

for some index set $\mathcal{I} \subseteq \{1,\dots,m\}$, so the complete two-dimensional plane is the exact-compatibility unit and not a heuristic constraint. At a repeated rate the family is strictly larger: on a same-frequency block a real subspace is invariant under $R_\Delta^{(p)}$ for every $\Delta \in \mathbb{Z}$ exactly when it is the realification of a $\mathbb{C}$-linear subspace of $\mathbb{C}^{H_{\mathrm{kv}}}$ — the same-frequency cross-head complex modes that a general complex coefficient direction can realize.

B.3 Data-Adaptive Construction

B.3.1 Compatible Second-Order Statistics

Let $\Sigma \in \mathbb{R}^{D_{\mathrm{group}}\times D_{\mathrm{group}}}$ be the second moment of the pre-RoPE concatenated keys over a frozen calibration set, symmetrized as $S = \frac{1}{2}(\Sigma + \Sigma^\top)$, and let $S_p = S[I_p, I_p] \in \mathbb{R}^{2H_{\mathrm{kv}}\times 2H_{\mathrm{kv}}}$ be the plane-p diagonal block. Define the *complex-linear part*

$$\Sigma_p^{\mathrm{comm}} = \tfrac{1}{2}\big(S_p - J_p S_p J_p\big) = \tfrac{1}{2}\big(S_p + J_p S_p J_p^\top\big). \tag{22}$$

Lemma 5 (Commutant projection)**.** *Σ_p^{comm} is the Frobenius-orthogonal projection of S_p onto the commutant $\{M : MJ_p = J_pM\}$. It is symmetric, satisfies $J_p\Sigma_p^{\mathrm{comm}} = \Sigma_p^{\mathrm{comm}}J_p$, and for all $u \in \mathbb{R}^{2H_{\mathrm{kv}}}$*

$$u^\top \Sigma_p^{\mathrm{comm}} u = \tfrac{1}{2}\big(u^\top S_p u + (J_p u)^\top S_p (J_p u)\big) \;\geq 0 \qquad \textit{whenever } S_p \succeq 0. \tag{23}$$

Proof of Lemma 5.

Proof. The maps $\{I, J_p, -I, -J_p\}$ form the cyclic group $\langle J_p\rangle$, whose averaging projection is exactly (22):

$$\frac{1}{4}\sum_{k=0}^{3} J_p^k S_p J_p^{-k}, \qquad J_p^{-1} = J_p^\top = -J_p.$$

Averaging over a group of orthogonal maps is the Frobenius-orthogonal projection onto its fixed-point set, the commutant of J_p, which gives the projection and commutation claims, and symmetry follows from $S_p^\top = S_p$ and $J_p^\top = -J_p$. For (23), $J_p^\top u = -J_p u$ turns $u^\top(-J_pS_pJ_p)u$ into $(J_pu)^\top S_p(J_pu) \geq 0$ whenever $S_p \succeq 0$. □

By Lemma 5, Σ_p^{comm} is the part of the plane-p key second moment consistent with the rotation structure: it discards exactly the component that is not $\mathbb{C}$-linear in the sense of (20).

B.3.2 Hermitian Equivalence and the Mode Family

Lemma 6 (Hermitian equivalence)**.** *Under (20), $\Sigma_p^{\mathrm{comm}} = \frac{1}{2}\mathcal{R}(H_p)$ is the realification of a Hermitian $M_p = \frac{1}{2}H_p \in \mathbb{C}^{H_{\mathrm{kv}}\times H_{\mathrm{kv}}}$, where H_p is the complex plane-local second moment of Eq. (6); it is positive semidefinite in this application because $H_p \succeq 0$. The eigenvalues of Σ_p^{comm} are real and equal to those of M_p, and a Hermitian eigenvalue of complex multiplicity k realifies to a real eigenvalue of real multiplicity $2k$: the real multiplicity of Σ_p^{comm} is always* even*, but it is $2k$ rather than invariably 2.*

Proof of Lemma 6.

Proof. A real matrix commuting with J_p is the realification of a unique $\mathbb{C}$-linear M_p; for real $w \mapsto \zeta$, $w' \mapsto \zeta'$,

$$\langle \Sigma_p^{\mathrm{comm}} w, w'\rangle = \mathrm{Re}\langle M_p\zeta, \zeta'\rangle, \qquad \langle w, \Sigma_p^{\mathrm{comm}} w'\rangle = \mathrm{Re}\langle \zeta, M_p\zeta'\rangle,$$

so symmetry of Σ_p^{comm} is equivalent to $M_p = M_p^{\text{H}}$, and $\Sigma_p^{\text{comm}} = \frac{1}{2}\mathcal{R}(H_p)$. An eigenpair $M_p\zeta = \lambda\zeta$ ($\lambda \in \mathbb{R}$) realifies to $\Sigma_p^{\text{comm}} w = \lambda w$, and $J_p w$, the realification of $\mathrm{i}\zeta$, is an orthogonal eigenvector with the same λ; hence real multiplicity $2k$ for complex multiplicity k. □

Energy convention. With $J_p^\top = -J_p$, $\Sigma_p^{\text{comm}} = \frac{1}{2}(S_p + J_p S_p J_p^\top) = \frac{1}{2}(S_p - J_p S_p J_p) = \frac{1}{2}\mathcal{R}(H_p)$, so $\mathcal{R}(H_p) = 2\Sigma_p^{\text{comm}}$: the raw real covariance realifies the complex covariance only through its J_p-commuting part. Scores are quoted in the H_p convention $\lambda_i(H_p)$, twice the legacy quantity $M_p = \frac{1}{2}H_p$, so a cutoff or tie tolerance stated in the legacy M-convention must be doubled here to be compared; a global factor leaves the ordering, hence the selected modes, unchanged.

Remark 7 (The real eigendecomposition and J_p-adapted selection)**.** By Lemma 6 the real and complex problems correspond at the matrix and eigenspace level: an eigenline $\mathbb{C}\zeta$ of H_p realifies to an eigen-two-plane of $\Sigma_p^{\text{comm}} = \frac{1}{2}\mathcal{R}(H_p)$ with eigenvalue $\lambda/2$, and conversely. The complex cross-head coefficients of a mode are the entries of ζ, a general complex vector, so the mode family is not confined to the real, phase-aligned subclass $\zeta \in \mathbb{R}^{H_{\text{kv}}}$ (the class $u \otimes I_2$) that coordinate-pair constructions retain. Bases, however, are not eigenspaces: inside a degenerate eigenspace a generic real eigensolver may return any real orthonormal basis, which need not be J_p-adapted, whereas the complex-Hermitian eigendecomposition produces eigenlines directly — precisely the freedom that Definition 11 removes. The construction thus reaches the complex mode family available in the achievable eigenspaces, and no claim is made that eigenvalue-ordered selection exhausts the class of all $\mathbb{C}$-linear-invariant subspaces.

B.3.3 Spectral Construction of Compatible Modes

Definition 8 (Candidate columns)**.** For each plane p compute one real symmetric eigendecomposition $\Sigma_p^{\text{comm}} = V_p \Lambda_p V_p^\top$ of the compatible second moment (22), with $V_p = [\, e_{p,1} \;\cdots\; e_{p,2H_{\text{kv}}} \,]$ orthonormal and $\Lambda_p = \operatorname{diag}(\lambda_{p,1}, \ldots, \lambda_{p,2H_{\text{kv}}})$ real. Let $\iota_p : \mathbb{R}^{2H_{\text{kv}}} \to \mathbb{R}^{D_{\text{group}}}$ be the coordinate embedding

$$(\iota_p w)_{I_p[\ell]} = w_\ell \quad (\ell = 0, \ldots, 2H_{\text{kv}} - 1), \qquad \text{and} \quad (\iota_p w)_c = 0 \text{ otherwise}, \tag{24}$$

and call $u_{p,j} := \iota_p e_{p,j}$ a *candidate column* with score $\text{score}(p,j) := 2\lambda_{p,j}$, the factor 2 placing the score in the H_p convention above; the $2H_{\text{kv}}m = D_{\text{group}}$ candidates form an orthonormal basis of $\mathbb{R}^{D_{\text{group}}}$. A subset of $\{u_{p,j}\}$ is an *individual-column selection*; it is the historical recipe that produced the reported bases; Appendix C.1 records this construction-provenance detail. It is **not** compatible by construction, because it may retain a lone member of a J_p-pair (Remark 10). The canonical algorithm is instead the *whole-mode selection* of Definition 11.

All candidates are eigenvector-derived columns of diagonal blocks of one and the same global second moment S, scored by their eigenvalues, so no cross-plane normalization is applied or needed: by (23) the score of $u_{p,j}$ is twice the average key energy along $e_{p,j}$ and along $J_p e_{p,j}$, i.e. the full energy of the mode $\operatorname{span}\{e_{p,j}, J_p e_{p,j}\}$; a mode identified by its score is written $\lambda_i(H_p) = \text{score}(p,j)$, so the raw real eigenvalue $\lambda_{p,j}$ is exactly half of it.

For any selection let $S_p \subseteq \{1, \ldots, 2H_{\text{kv}}\}$ be the retained index set of plane p and $U_p = \iota_p V_p[:, S_p]$, so that, after regrouping columns by plane, U is the column concatenation of the U_p over the planes with $S_p \neq \varnothing$. Let $\widetilde{J}_p = \iota_p J_p \iota_p^\top$ be the plane-p generator embedded in $\mathbb{R}^{D_{\text{group}}}$ (the embedded block-diagonal form of Ω); the plane-p generator in the retained local coordinates is $\widehat{J}_p = U_p^\top \widetilde{J}_p U_p = V_p[:, S_p]^\top J_p V_p[:, S_p]$, the two expressions being equal because $\iota_p^\top \iota_p = I_{2H_{\text{kv}}}$.

Theorem 9 (Orthonormality and conditional compatibility of a selection)**.** *Let U be any selection of candidate columns, with retained plane sets $S_p \subseteq \{1, \ldots, 2H_{\text{kv}}\}$ and local factors $U_p = \iota_p V_p[:, S_p]$.* (i) Orthonormality. *$U^\top U = I_r$, where r is the number of selected columns and there is no rank deficiency: the D_{group} candidates form an orthonormal basis, and eigendecomposition orthonormality holds for any real symmetric input, degenerate eigenvalues or not.* (ii) Compatibility. *If for every p the retained index set S_p is closed under J_p — i.e.*

$u_{p,j} \in \mathrm{range}(U)$ *implies* $\widetilde{J}_p u_{p,j} \in \mathrm{range}(U)$ *for the local direction of the same plane — then* $\mathrm{range}(U)$ *is invariant under every RoPE rotation* R_Δ*, and*

$$R_\Delta U = U \widehat{R}_\Delta, \qquad \widehat{R}_\Delta = U^\top R_\Delta U = \bigoplus_p \exp\big(\Delta\theta_p \, \widehat{J}_p\big), \qquad \widehat{J}_p = U_p^\top \widetilde{J}_p U_p, \tag{25}$$

a direct sum over the selected planes of the plane-p rotation restricted to that plane's retained modes. The displayed form holds in a J-adapted latent basis*, one that exhibits each retained mode as a* $\mathrm{span}\{w, J_p w\}$ *pair and in which each summand is a standard* 2×2 *rotation; in an arbitrary latent basis the same operator appears as* $U^\top \Omega U$*, still real antisymmetric, and only its spectrum and its* J_p*-compatibility are basis-independent.*

Proof of Theorem 9.

Proof. (i) Orthonormality. For fixed p, the eigendecomposition of a real symmetric matrix returns orthonormal columns for any real symmetric input, degenerate eigenvalues or not — a guarantee of the symmetric eigendecomposition itself, not of any pairing argument. As ι_p is an isometry by (24), the $u_{p,j}$ are orthonormal for fixed p; for $p \neq p'$ the supports $I_p, I_{p'}$ are disjoint by (17), so

$$\langle u_{p,j}, u_{p',j'} \rangle = 0$$

regardless of the embedded values; any subset therefore has orthonormal columns.

(ii) Compatibility. With $\widetilde{\Omega} = \sum_p \theta_p \widetilde{J}_p$ (the embedded plane blocks have disjoint support) and $R_\Delta = \exp(\Delta\widetilde{\Omega})$, the closure hypothesis gives, for each p,

$$\widetilde{\Omega}\, \iota_p V_p[:, S_p] = \theta_p \, \iota_p J_p V_p[:, S_p] = \iota_p V_p[:, S_p]\, \theta_p \widehat{J}_p,$$

because J_p maps the retained local eigenspace into itself and the discarded local directions are orthogonal to it. Summing over the disjoint planes gives $\widetilde{\Omega} U = U \widehat{\Omega}$ with $\widehat{\Omega} = U^\top \widetilde{\Omega} U = \bigoplus_p \theta_p \widehat{J}_p$, hence $\exp(\Delta\widetilde{\Omega}) U = U \exp(\Delta\widehat{\Omega})$, which is (25); each $\widehat{J}_p$ is real antisymmetric with $\widehat{J}_p^2 = -I$ on every retained J_p-pair, i.e. a direct sum of 2×2 rotation generators. □

B.3.4 Whole-Mode Selection under a Global Rank Budget

The results of this subsection form one logical chain. We begin from the candidate columns of Section B.3.3 and the failure mode of individual-column selection (Remark 10); whole-mode selection closes a retained span under J_p unconditionally (Proposition 12); Proposition 13 then records the sufficient construction fact that a direct sum of whole same-frequency modes across planes is RoPE-invariant; and Proposition 14 closes the chain by identifying the pooled top-K cutoff with the globally optimal allocation of a rank budget across planes.

Remark 10 (Why whole-mode selection is required)**.** Individual-column selection can violate the requirement of Theorem 9(ii) in two ways. An odd rank budget forces a retained span of odd real dimension, whereas every J_p-invariant subspace has even real dimension; the generic instance is a lone member of a real multiplicity-2 pair. And the cutoff may fall inside a degenerate eigenspace of real multiplicity at least four, where a generic real eigensolver need not return a J_p-adapted basis, so even an even rank budget does not by itself guarantee closure; an instance is a single plane with $H_{\mathrm{kv}} = 4$ whose plane second moment is a multiple of the identity, where every direction is eigenvalue-tied. Whole-mode selection removes both cases by selecting complete complex modes and forcing an even real rank. The decoder bases used in our experiments were produced by an earlier individual-column implementation, but all of their ranks are even and they satisfy the required subspace-closure conditions numerically. See Appendix C.5 for the numerical verification.

Definition 11 (Whole-mode selection)**.** Group the candidate columns of Definition 8 into *complex modes*: within each plane p, partition the $2H_{\mathrm{kv}}$ real eigenvectors into eigenvalue-tied clusters and, within each cluster of complex dimension c, choose *any* complex-orthonormal basis $\zeta_{p,1}, \ldots, \zeta_{p,c} \in \mathbb{C}^{H_{\mathrm{kv}}}$ of the corresponding eigenspace of H_p (Lemma 6); every such basis is valid because all directions in a genuinely tied eigenspace carry identical score (23). Each $\zeta_{p,i}$ defines one *complex mode*, realized as the pair of real columns

$$\big[\, \iota_p \phi(\zeta_{p,i}),\ \iota_p \phi(\mathrm{i}\zeta_{p,i}) \,\big] \in \mathbb{R}^{D_{\mathrm{group}} \times 2}, \qquad \phi(a + \mathrm{i}b) = (a_1, \ldots, a_{H_{\mathrm{kv}}}, b_1, \ldots, b_{H_{\mathrm{kv}}})^\top, \tag{26}$$

with shared score $\lambda_i(H_p)$.

The algorithm is: (i) compute $\Sigma_p^{\text{comm}} = \frac{1}{2}\mathcal{R}(H_p)$ per plane and its eigendecomposition, or solve the Hermitian eigenproblem for H_p directly, or convert the real eigenspace into a J_p-adapted complex basis as in Remark 7; (ii) organize the eigenvectors into complex modes as above; (iii) rank all modes across all planes by score $\lambda_i(H_p)$ and select the top K; (iv) require an even real rank budget $r = 2K$; and (v) never select a single column of (26) in isolation, and never combine within one mode directions drawn from different frequency planes.

B.3.5 Unconditional Compatibility of Whole-Mode Selection

Proposition 12 (Unconditional compatibility)**.** *For any $0 \le K \le mH_{\text{kv}}$, the decoder basis produced by Definition 11 satisfies $U^\top U = I_{2K}$ and* range(U) *is exactly invariant under every R_Δ — with no condition on the rank cutoff, including cutoffs that select only part of a tied cluster's complex dimensionality.*

Proof of Proposition 12.

Proof. Orthonormality. Two modes with complex-orthonormal coefficient vectors ζ, ζ' have mutually orthogonal realified two-planes, because

$$\langle \phi(\zeta), \phi(\zeta') \rangle = \text{Re}\langle \zeta, \zeta' \rangle = 0, \qquad \langle \phi(\zeta), \phi(\text{i}\zeta') \rangle = \text{Re}\langle \zeta, \text{i}\zeta' \rangle = 0;$$

modes from different planes are orthogonal by disjoint support, as in Theorem 9(i). Within one mode, $\phi(\zeta)$ is a unit real vector and J_p is real antisymmetric, so $\phi(\text{i}\zeta) = J_p\phi(\zeta)$ is a unit vector orthogonal to $\phi(\zeta)$; hence $U^\top U = I_{2K}$.

RoPE invariance. By (26) every *selected* column is one of a complete pair that spans the realification of a single J_p-invariant complex line, because $\phi(\text{i}\zeta) = J_p\phi(\zeta)$; the selected set is therefore always a union of complete J_p-invariant two-planes, hence trivially closed under J_p, and Theorem 9(ii) applies unconditionally, giving $R_\Delta U = U\widehat{R}_\Delta$ for every Δ. □

When the cutoff lands inside a degenerate complex eigenspace the top-K optimum is *set-valued* rather than unique: every mode above the cutoff must be retained, no mode below it may be retained, and within the cutoff eigenspace any complex-orthonormal subspace of the required complex dimension may be chosen, independently within each frequency plane and subject to the remaining budget. The theory therefore claims no unique canonical projector, and the numerical certificate records the cutoff as set-valued whenever the tie is non-trivial.

B.3.6 Compatibility Across Frequency Planes

Proposition 13 (Whole same-frequency modes across planes are compatible)**.** *When the plane blocks are indexed by rotation rate (all same-frequency multiplicities grouped into a single block), a direct sum of whole same-frequency modes drawn from many blocks side by side is J-closed and therefore RoPE-invariant, because each mode lies in exactly one block and the blocks are independent under $R_\Delta = \exp(\Delta\Omega)$.*

Proof of Proposition 13.

Proof. Each selected mode lies in a single block, in which Ω acts as $\theta_p J_p$, so a span that is the direct sum of per-block pieces each invariant under $\theta_p J_p$ is invariant under $\Omega = JB$ with $B = \sum_p \theta_p \Pi_p$ and Π_p the block projector, hence under every $R_\Delta = \exp(\Delta\Omega)$. □

B.3.7 Optimality of the Pooled Top-K Cutoff

Proposition 14 (The pooled top-K cutoff solves a rank-budgeted trace program)**.** *For each plane p let H_p be the Hermitian calibration second moment of Eq. (6), with eigenvalues $\lambda_1(H_p) \ge \cdots \ge \lambda_{H_{\text{kv}}}(H_p)$ and orthonormal eigenvectors, and let $k_p \in \{0, \ldots, H_{\text{kv}}\}$ be the number of complex modes retained in plane p. With $m = d/2$ planes the global budget*

satisfies $0 \le K \le mH_{\mathrm{kv}}$*, and the retained real rank is* $r = 2\sum_p k_p = 2K$*, hence even and at most* D_{group}*. Consider*

$$\max_{\{Z_p\}} \sum_p \mathrm{tr}\big(Z_p^{\mathrm{H}} H_p Z_p\big) \qquad \textit{subject to} \qquad Z_p^{\mathrm{H}} Z_p = I_{k_p}, \quad Z_p \in \mathbb{C}^{H_{\mathrm{kv}} \times k_p}, \quad \sum_p k_p = K. \tag{27}$$

(i) Fixed allocation. *For a fixed allocation* $(k_p)_p$ *the optimum equals* $\sum_p \sum_{j \le k_p} \lambda_j(H_p)$ *and is attained by any* Z_p *whose columns are orthonormal eigenvectors of* H_p *spanning its* k_p *largest eigenvalues.*

(ii) Optimal allocation. *The optimum over all* $(k_p)_p$ *with* $\sum_p k_p = K$ *equals the sum of the* K *largest values of the* pooled *multiset* $\{\lambda_j(H_p)\}_{p,j}$*. Consequently the whole-mode selection of Definition 11, which ranks modes across planes by* $\lambda_i(H_p)$ *with no per-plane renormalization and truncates at the top* K*, is optimal for* (27) *in the sense of (ii). This is the precise sense in which "top-*K*" names a solution of a stated objective rather than a heuristic; if the cutoff is not tied, the optimum as a multiset of scores is unique.*

Proof of Proposition 14.

Proof. (i). For each plane separately, the maximum of $\mathrm{tr}(Z_p^{\mathrm{H}} H_p Z_p)$ over complex matrices with orthonormal columns is the sum of the k_p largest eigenvalues of H_p by the Ky Fan maximum principle, attained exactly when the columns span the corresponding eigenspace; the planes are independent, so the optima add.

(ii). Substituting (i) leaves

$$\max_{\{k_p\}} \sum_p \sum_{j \le k_p} \lambda_j(H_p) \qquad \text{subject to} \qquad \sum_p k_p = K, \quad 0 \le k_p \le H_{\mathrm{kv}}.$$

The plane blocks are diagonal blocks of one global calibration second moment, so their eigenvalues are already comparable. If a feasible allocation left some selected value smaller than an unselected value of the pooled multiset, exchanging the two would keep the budget and not decrease the objective; repeating this exchange yields the global top K, precisely the multiset selected by Definition 11. □

B.4 Extension to Grouped-Query Attention

The decoder is built per KV group but attention is computed per query head. Let H_q denote the number of query heads served by one decoder basis, and let $g : \{1, \dots, H_q\} \to \{1, \dots, H_{\mathrm{kv}}\}$ map each of these H_q query heads to its KV head (the production model uses 4 query heads per KV head, i.e. $H_q = 4H_{\mathrm{kv}}$ and $g(a) = \lceil a/4 \rceil$), and let

$$\iota_h : \mathbb{R}^{d_h} \to \mathbb{R}^{D_{\mathrm{group}}}, \qquad \iota_h(x) = e_h \otimes x, \tag{28}$$

be the *head injection*, an isometry placing x in head h's block and zero elsewhere. With the concatenated group key $\bar{k}_s = (k_s^{(1)}; \dots; k_s^{(H_{\mathrm{kv}})})$ and the latent operator $\widehat{R}_\Delta = U^\top R_\Delta U$ of (25), the lifted latent query and per-query-head latent logit are

$$\tilde{q}_a = U^\top \iota_{g(a)}(q_a) \in \mathbb{R}^r, \qquad \hat{\ell}^a_{s,t} = \frac{1}{\sqrt{d_h}}\, \tilde{q}_a^\top \widehat{R}_{s-t}\, z_s. \tag{29}$$

Theorem 15 (MHA/GQA bridge)**.** *Assume* range(U) *is* R_Δ*-invariant, so that* (25) *holds. Then for every query head* a*, every query position* t*, and every key position* s*,*

$$\hat{\ell}^a_{s,t} = \frac{1}{\sqrt{d_h}}\, q_a^\top R_{s-t} \left(U E \bar{k}_s\right)^{(g(a))}, \tag{30}$$

where $(\cdot)^{(h)}$ *extracts the* h*-th head block; i.e. the latent logit equals the reconstruct-then-rotate logit of the same compressed representation* (U, E)*, read at the query head's KV slot, in exact arithmetic.*

Proof of Theorem 15.

Proof. RoPE acts on the concatenated key as the block-diagonal operator $R_\Delta = \bigoplus_{h=1}^{H_{\mathrm{kv}}}$(shared per-head rotation), so

$$\iota_{g(a)}(q_a)^\top R_\Delta \bar{k}_s = q_a^\top R_\Delta k_s^{(g(a))},$$

because every other block pairs with the zeros inserted by $\iota_{g(a)}$. Applying Theorem 1 with query $\iota_{g(a)}(q_a)$ and compressed key $z_s = E\bar{k}_s$ gives

$$\iota_{g(a)}(q_a)^\top R_\Delta U E \bar{k}_s = (U^\top \iota_{g(a)}(q_a))^\top \widehat{R}_\Delta E \bar{k}_s = \tilde{q}_a^\top \widehat{R}_\Delta z_s,$$

and the same block-diagonal argument applied to $R_\Delta(UE\bar{k}_s)$ in place of $R_\Delta \bar{k}_s$ gives

$$\iota_{g(a)}(q_a)^\top R_\Delta (UE\bar{k}_s) = q_a^\top R_\Delta (UE\bar{k}_s)^{(g(a))}.$$

Combining the three identities and restoring $1/\sqrt{d_h}$ yields (30). □

Because (30) reproduces the reconstruct-then-rotate logit of the same compressed representation (U, E) exactly, in exact arithmetic and for every rank r, the temperature must remain the model's own $1/\sqrt{d_h}$ (a rescaling such as $1/\sqrt{r}$ would rescale every logit and change the softmax); the residual gap to the true full-width logit is the representation error (13) controlled by (U, E) and is not removed by the bridge.

B.5 Attention-Aware Encoder Fitting

B.5.1 Parameterization of Left-Inverse Encoders

For an orthonormal decoder basis $U \in \mathbb{R}^{D_{\mathrm{group}} \times r}$, any k decomposes orthogonally as

$$k = UU^\top k + (I - UU^\top)k = k_\parallel + k_\perp, \qquad k_\parallel \in \mathcal{S}_U, \quad k_\perp \in \mathcal{S}_U^\perp; \tag{31}$$

the orthogonal encoder $U^\top$ returns the latent coordinates exactly on $\mathcal{S}_U$ but discards $k_\perp$, which may still affect the query–key interaction. Remapping that component into the existing r latent coordinates gives

$$E = U^\top + C(I - UU^\top), \tag{32}$$

which is Eq. (7) of the main text.

Proposition 16 (Complete parameterization of left-inverse encoders)**.** *For a fixed orthonormal U, the set of linear encoders satisfying $EU = I_r$ is exactly $\{U^\top + C(I - UU^\top) : C \in \mathbb{R}^{r \times D_{\mathrm{group}}}\}$.*

Proof of Proposition 16.

Proof. For any C,

$$EU = U^\top U + C(I - UU^\top)U = I_r + C(U - UU^\top U) = I_r.$$

Conversely, if $EU = I_r$, write $P = UU^\top$; then

$$E = EP + E(I - P) = EUU^\top + E(I - P) = U^\top + E(I - P),$$

so E has the stated form with $C = E$. Hence the parameterization is the general form of every linear left-inverse encoder for the fixed basis U, and the remaining degrees of freedom come exactly from $\mathcal{S}_U^\perp$. □

Because $EU = I_r$ holds for every C, the action of the encoder on $\mathcal{S}_U$ is already determined: learning C leaves the compatible decoding subspace and its latent-space RoPE unchanged, while letting information in $\mathcal{S}_U^\perp$ compensate part of the representation loss at no extra cache width or rank.

B.5.2 Convex Ridge Objective

The encoder is fitted on the *unnormalized* query–key score, whereas attention consumes $(1/\sqrt{d_h})$ times that score (29). For a causal pair $s < t$, write $\Delta = s - t$; query head a reads the concatenated key through the injection $\iota_{g(a)}$, so with $\bar{q}_{a,t} = \iota_{g(a)}(q_{a,t})$ and $E = U^\top + C(I - P)$ the compressed unnormalized score splits into a base term and a correction,

$$\tilde{\ell}_{a,s,t} = \bar{q}_{a,t}^\top R_\Delta U U^\top \bar{k}_s + \bar{q}_{a,t}^\top R_\Delta U C (I - P) \bar{k}_s, \tag{33}$$

and the logit entering the softmax is $\hat{\ell}_{a,s,t} = (1/\sqrt{d_h})\,\tilde{\ell}_{a,s,t}$. The positive factor $1/\sqrt{d_h}$ cancels from every exact-equivalence statement but not from the ridge problem, so the fit is stated for unnormalized scores and the reported γ belongs to that convention. The regression *target* is the full-width unnormalized score

$$\tilde{\ell}^{\text{full}}_{a,s,t} = \bar{q}_{a,t}^\top R_\Delta \bar{k}_s, \tag{34}$$

the score of the uncompressed key, and the fitted quantity is its residual over the orthogonal-encoder baseline

$$\tilde{\ell}^{\text{base}}_{a,s,t} = \bar{q}_{a,t}^\top R_\Delta U U^\top \bar{k}_s. \tag{35}$$

The *compressed* score cannot serve as that target: by (34)–(35) and the definition of E,

$$\tilde{\ell}_{a,s,t} - \tilde{\ell}^{\text{base}}_{a,s,t} = \bar{q}_{a,t}^\top R_\Delta U C (I - P) \bar{k}_s,$$

which is exactly the fitted term below, so regressing $\tilde{\ell} - \tilde{\ell}^{\text{base}}$ on C would make every residual vanish identically and drive $C \to 0$; the full-width score is what carries the information about the attention logit that C must reproduce. Every calibration row is indexed by a triple (a_i, s_i, t_i): the query heads assigned to one KV head are separate rows sharing the same x_{s_i} and the same C. With $g_{a,s,t} = U^\top R_\Delta^\top \bar{q}_{a,t}$ and $x_s = (I - P)\bar{k}_s$, so that $\bar{q}_{a,t}^\top R_\Delta U C x_s = g_{a,s,t}^\top C x_s$, the encoder is fitted by the ridge objective

$$\min_C \; \frac{1}{M} \sum_{i=1}^{M} \Big[\big(\tilde{\ell}^{\text{full}}_{a_i,s_i,t_i} - \tilde{\ell}^{\text{base}}_{a_i,s_i,t_i} \big) - g_{a_i,s_i,t_i}^\top C x_{s_i} \Big]^2 + \gamma \|C\|_F^2, \tag{36}$$

where M is the number of rows and $\gamma > 0$ is the regularization coefficient. The $1/M$ normalization is part of the definition: the deployed fit uses $\tau = 10^{-2}$ and

$$\gamma = \tau \, \frac{\operatorname{tr}(H)}{r\,(D_{\text{group}} - r)}, \qquad H = \frac{1}{M} \sum_{i=1}^{M} (x_i x_i^\top) \otimes (g_i g_i^\top), \tag{37}$$

the denominator $r(D_{\text{group}} - r)$ counting the degrees of freedom that actually act on $\text{range}(U)^\perp$, because the component of C inside $\text{range}(U)$ is annihilated by $(I - P)$ and driven to zero by the penalty. Dropping $1/M$ would require multiplying γ by M; the normalized form is the one that was run.

Let $y_i = \tilde{\ell}^{\text{full}}_{a_i,s_i,t_i} - \tilde{\ell}^{\text{base}}_{a_i,s_i,t_i}$, $g_i = g_{a_i,s_i,t_i}$, $x_i = x_{s_i}$ and $c = \text{vec}(C)$. Using $g_i^\top C x_i = (x_i \otimes g_i)^\top c$, (36) is the standard ridge normal form

$$\min_c \; M^{-1} \|y - Ac\|_2^2 + \gamma \|c\|_2^2, \tag{38}$$

where $y \in \mathbb{R}^M$, $c \in \mathbb{R}^{r D_{\text{group}}}$, and row i of $A \in \mathbb{R}^{M \times r D_{\text{group}}}$ satisfies $A_i^\top = x_i \otimes g_i$. Absorbing the normalization into $\bar{A} = A/\sqrt{M}$ and $\bar{y} = y/\sqrt{M}$, the Hessian is $2(\bar{A}^\top \bar{A} + \gamma I) \succ 0$ for $\gamma > 0$, so the C-subproblem is strictly convex and its unique minimizer satisfies $(\bar{A}^\top \bar{A} + \gamma I)c = \bar{A}^\top \bar{y} = M^{-1} A^\top y$.

Scope of the two-stage construction. The identity $EU = I_r$ is an *encoder design constraint*, not a prerequisite for RoPE compatibility or path equivalence: those properties depend only on $\text{range}(U)$ being R_Δ-invariant together with $U^\top U = I_r$, and involve E only through $z_s = Ek_s$, so Theorem 1 holds for *any* E. ($U^\top U = I_r$ is a left-inverse identity between $U^\top$ and U, distinct from the projector $UU^\top$, which equals I_r only in the trivial case $r = D_{\text{group}}$.) U is chosen first by the spectral-energy criterion of Appendix B.3.3 on

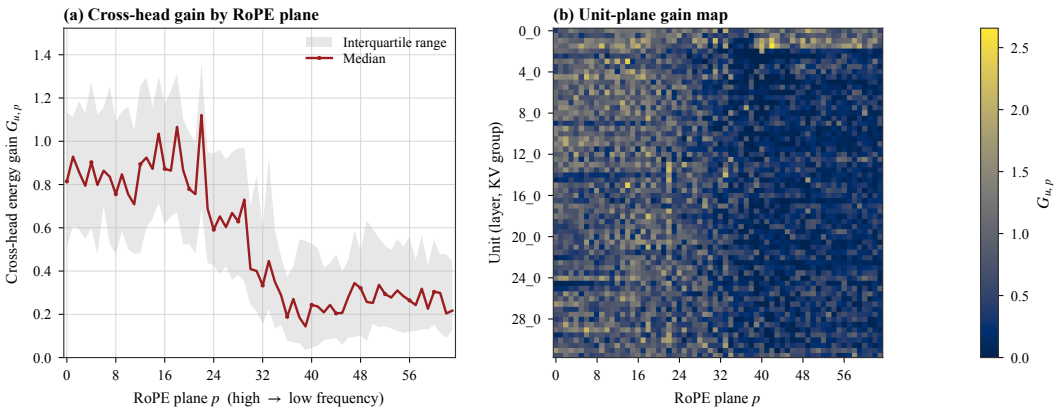


Figure 4: Cross-head structure in the frozen calibration statistics. (a) Median and interquartile range of $G_{u,p}$ over 64 units; (b) all unit–plane values in their original order. Plane index runs from high to low RoPE frequency. Larger values indicate more energy available to a same-frequency cross-head mode than to the strongest original head pair.

calibration key statistics alone, and C is then fitted to attention logits with U frozen, so range(U) stays exactly RoPE-invariant while the degrees of freedom affecting representation quality are concentrated in E. This is not a joint optimum — no single objective over (U, E) is optimized — and the optimality statements are confined to Proposition 14 and to convexity of the C-subproblem for fixed U.

C Experimental Details and Additional Results

C.1 Models, Baselines, and Implementation

Basis implementation. The deployed bases were generated by the individual-column implementation of Definition 8. For each plane p, it eigendecomposes $\Sigma_p^{\text{comm}} = V_p \Lambda_p V_p^\top$, scores each column $u_{p,j} = \iota_p e_{p,j}$ by $2\lambda_{p,j}$, retains the global top-r columns, and applies a thin QR. Appendix C.5 verifies that, for all 64 units, these bases span the same subspaces as the canonical whole-mode construction of Definition 11 and satisfy the identities required by Theorem 1. The experiments therefore instantiate Algorithm 1; compatibility of this implementation follows from the numerical verification rather than from its column-wise selection rule alone (Remark 10).

Evaluation protocol. The five operating points in Table 2 and Figure 3b—Truncated SVD, PaLU, the deployed 41/23 hybrid, the 50/14 composition, and RoFold-Pure—share the same frozen backbone, corpora, runtime, and repetition procedure. Calibration and evaluation data are disjoint. The 8K and 16K evaluations use one held-out corpus; 32K uses three non-overlapping held-out blocks; and 64K uses a separate held-out span. Quality and throughput are measured on an NVIDIA A100 80GB PCIe GPU, while persistent bytes are counted separately on the same device class.

C.2 Cross-Head Structure in Calibration Statistics

Cross-head energy gain. Figure 4 asks whether mixing KV heads at a *fixed* RoPE frequency exposes useful directions that no original head provides by itself. For each unit u and plane p, let $H_{u,p} \in \mathbb{C}^{4\times 4}$ be the frozen pre-RoPE covariance across the four KV heads. Its diagonal entry $(H_{u,p})_{hh}$ is the calibration energy of the original coordinate pair in head h, whereas $\lambda_{\max}(H_{u,p})$ is the maximum energy captured by any unit-norm complex combination of those four same-frequency pairs. We therefore report

$$G_{u,p} = \frac{\lambda_{\max}(H_{u,p})}{\max_h (H_{u,p})_{hh}} - 1. \tag{39}$$

Thus $G_{u,p} = 0$ means that cross-head mixing offers no advantage over the strongest original head, while $G_{u,p} = 1$ means that the best mixed mode captures twice as much calibration

energy. Because the mixture never crosses frequency planes, it remains inside the RoPE-compatible family characterized in Appendix B.2.

The two panels are complementary views of the same, complete 64×64 matrix. Panel (a) shows, for each frequency plane, the median and interquartile range over all 64 layer–KV-group units; panel (b) retains every unit–plane value in the original order, with brighter cells denoting larger gains. Plane index increases from left to right, so the left side contains the higher RoPE frequencies. No units or planes are reordered, selected, or clipped.

Frequency-dependent structure. The dominant pattern is frequency dependent rather than a handful of isolated bright cells. Across the first 24 (higher-frequency) planes, the median over all unit–plane values is 0.851, and 79.1% of values exceed 0.5. Across planes 32–63, the corresponding median is 0.265, and only 26.0% exceed 0.5. The median curve in panel (a) makes the same transition visible across units, while panel (b) shows that the effect recurs across layers and KV groups. This supports the specific motivation for a data-adaptive cross-head basis: the repeated-frequency structure contains substantial shared energy, especially at higher RoPE frequencies, that coordinate-pair selection cannot exploit. This diagnostic is silent on downstream quality, throughput, and the final rank allocation, which Section 5 evaluates separately.

C.3 Position-Conditioned Stability of the Compatible Directions

Position-bucket diagnostic. The decoder basis is calibrated once and then frozen across all evaluated context lengths. We therefore examine whether the compatible directions identified from pre-RoPE key statistics remain representative at different positions within a long sequence. Using a separate diagnostic corpus, disjoint from the calibration and held-out evaluation corpora, we analyze 16 document-disjoint 65,536-token sequences and divide each into four fixed position buckets: $B1 = [0, 8192)$, $B2 = [8192, 16384)$, $B3 = [16384, 32768)$, and $B4 = [32768, 65536)$. For every one of the 64 layer–KV-group units and 64 RoPE planes, we recompute the leading same-frequency cross-head mode in each bucket and measure its phase-invariant overlap, $|v_a^{\mathrm{H}} v_b|^2$, across bucket pairs. This diagnostic probes the position-conditioned statistics used to choose the decoder basis U; it makes no claim about downstream quality and no causal claim about absolute position.

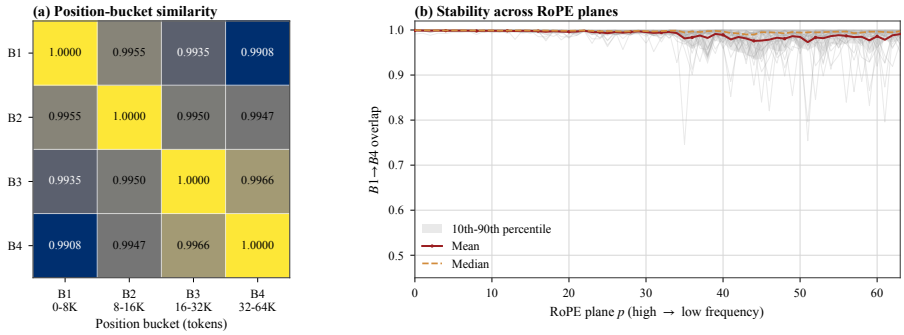


Figure 5: Position-conditioned stability of the RoPE-compatible directions over 64K tokens. (a) Mean leading-mode overlap between position buckets, averaged over all 64 units and 64 RoPE planes. (b) Overlap between the first and last buckets as a function of RoPE plane. The solid and dashed curves denote the mean and median across the 64 units, the shaded region spans the 10th–90th percentiles, and the faint gray curves show per-layer means across KV groups. Plane index increases from high to low RoPE frequency. A small tail of position-associated variation appears more prominently among later, lower-frequency planes.

Stability across context. Figure 5 shows a gradual change rather than an abrupt positional transition. Mean overlap is 0.996 between adjacent buckets ($B1$–$B2$) and remains 0.991 between the first and last buckets; the corresponding median is 0.998. Most compatible directions therefore remain close to their initial-position counterparts, with a wider tail of variation at lower RoPE frequencies. Two robustness checks give similar conclusions: document-level resampling yields a mean overlap of 0.989, and restricting to well-separated

leading modes (normalized eigengap $r_1 \geq 0.1$) yields 0.994. These results support using one frozen compatible basis over the evaluated context range. They concern the statistics used to select U and do not separately establish position invariance of the fitted encoder E.

C.4 Mechanism Ablations: Cross-Head Modes and Attention-Aware Encoding

We isolate the fitted encoder and the cross-head basis family in two controlled comparisons. The first holds the deployed basis fixed and removes only the learned encoder correction. The second recomputes two bases under a shared pipeline and varies only the compatible family. Because the comparisons use different basis instances, we interpret the change within each table rather than absolute PPL across tables.

C.4.1 Encoder Ablation

Removing the fitted correction. To isolate the encoder's contribution, we keep the deployed U, 40.1% budget, 41/23 routing, and execution path fixed, and set $C = 0$ so that $E = U^\top$. The only removed component is therefore the attention-aware correction derived in Appendix B.5; both variants use the held-out protocol of Table 2.

Table 3: Encoder ablation at the deployed 40.1% budget. Both RoFold variants share the same basis, routing, and execution; ΔPPL is relative to the fitted encoder.

Method	8K PPL	16K PPL
Truncated SVD	9.4321	9.8172
PaLU	7.4946	7.8631
RoFold (deployed)	8.0025	8.4043
RoFold ($E = U^\top$)	8.9785	9.3016
ΔPPL vs. deployed	+0.9760 (+12.2%)	+0.8973 (+10.7%)

Encoder contribution. Dropping the fitted encoder increases PPL by 0.976 at 8K and 0.897 at 16K, roughly twice the deployed method's gap to PaLU. The attention-aware fit therefore recovers a substantial fraction of the quality lost by the bare orthogonal projection. We report PPL only because changing E leaves the memory layout, routing, and reconstruction-free execution path unchanged.

C.4.2 Independent Basis-Family Ablation

Restricting the compatible family. To isolate the compatible-basis family, both variants use $E = U^\top$. We compare the continuous same-frequency cross-head construction of Section 4.1 with a restriction that retains only original per-head RoPE coordinate pairs, scored by their diagonal calibration energy $(H_p)_{hh}$. The two bases are recomputed with the same rank table, 40.1% budget, 41/23 routing, value compression, calibration data, and held-out 8K/16K protocol; only the admissible basis family changes. The restriction is conceptually related to pair-selection methods such as RAP (Xin et al., 2026), but does not reproduce RAP's full pipeline.

Table 4: Compatible-family ablation at the matched 40.1% budget with $E = U^\top$ fixed. Both variants share the same construction and evaluation protocol; comparisons are within this table.

Compatible basis ($E = U^\top$)	8K PPL	16K PPL
Cross-head modes	8.4742	8.9023
Whole-pair restriction	12.1667	12.8898

Cross-head contribution. Restricting the compatible family to original whole RoPE pairs increases PPL by 43.6% at 8K and 44.8% at 16K. This provides downstream evidence that the continuous same-frequency cross-head freedom identified by the theory is practically useful.

C.5 Numerical Verification of the Deployed Basis and Encoder

Finite-precision check. The exact equivalence is proved in Appendix B.1.1. Here we verify that the finite-precision bases and encoder used in the experiments preserve the required identities. Table 5 reports the worst residual over all 64 (layer, KV-group) units; every value lies within its prescribed tolerance. Writing P for the deployed individual-column projector and P_{can} for the canonical whole-mode projector of Definition 11: RoPE closure is $\|JP - PJ\|_F$ and $\|\Omega P - P\Omega\|_F$ (does the deployed subspace alone stay RoPE-invariant); spectral locality is $\|\Sigma_p^{\text{comm}} P - P\Sigma_p^{\text{comm}}\|_F$ (does the deployed span align with the compatible second moment's own eigenspaces); off-block is the Frobenius norm of P restricted to distinct-plane coordinate blocks (cross-frequency leakage); cutoff leakage is the residual energy from modes at the rank-r tie boundary; and, critically, *projector* is $\|P - P_{\text{can}}\|_F$, the check that verifies the deployed basis and the canonical whole-mode basis of Definition 11 span the same subspace, while *representation map* is the stronger $\|U_{\text{can}} - UQ\|_F$ for $Q = U^\top U_{\text{can}}$, confirming the two bases are related by an orthogonal change of basis rather than merely agreeing to first order.

Table 5: Finite-precision verification over all 64 deployed units. Residuals are Frobenius norms; the rank-cutoff row reports the minimum eigengap.

Check	Worst-case value	Tolerance
RoPE closure	J: 1.7×10^{-11}; Ω: 3.6×10^{-13}	$\leq 10^{-9}$
Subspace locality	spectral: 9.2×10^{-13}; off-block: 1.2×10^{-11}; cutoff leakage: 8.3×10^{-12}	$\leq 10^{-9}$
Basis consistency	orthonormality: 1.5×10^{-14}; projector: 1.2×10^{-11}; representation map: 1.3×10^{-11}	$\leq 10^{-11}$ (orth.); $\leq 10^{-9}$ (others)
FP64 GQA logits	1.4×10^{-11}	$\leq 10^{-9}$
Global rank cutoff	8.6×10^{-4}	$> 10^{-9}$
BF16 encoder consistency	normalized residual: 1.7×10^{-3}	$\leq 10^{-2}$

Numerical agreement. The closure and locality residuals show that the deployed bases retain the compatible subspaces characterized by the theory, and the projector residual confirms that the deployed individual-column basis and the canonical whole-mode basis of Definition 11 span the same subspace, the bridge between the theory and the deployed implementation. The FP64 logit discrepancy remains two orders of magnitude below tolerance. The rank cutoff is also well separated from numerical ties, and the stored BF16 encoder satisfies the normalized consistency criterion. These checks validate the finite-precision implementation; the exact statement itself follows from the symbolic proof.

C.6 Persistent-KV Accounting

Matched persistent payload. This accounting verifies that all compressed arms use the same stored KV capacity, so their quality and throughput differences are not caused by unequal cache budgets.

Table 6: Persistent-KV payload for Llama-3.1-8B. Counts include live K/V tensors and exclude temporary workspace and metadata.

Arm	Persistent KV (B/token)	KV reduction
Full-Width KV	131,072	–
Truncated SVD	78,464	40.1%
PaLU	78,464	40.1%
RoFold	78,464	40.1%

Realized budget. The discrete rank table realizes a 40.137% exact reduction, reported as 40.1% in the main text. Each compressed arm stores 21,120 B/token for keys and 57,344 B/token for values; the value path is identical across arms, so the comparison isolates the key-side method. These per-token quantities are fixed by the rank table and therefore do not depend on context length.

Accounting scope. We compute logical bytes from the live persistent K/V tensors after causal prefill and verify them against the analytical counts implied by the rank table. Logical and unique-storage counts coincide. Temporary workspace, allocator capacity, metadata, and CPU-side objects are excluded from all methods.

C.7 Static Routing of the Deployed Hybrid

Routing objective. The hybrid must decide which units benefit most from reconstructed execution under a fixed reconstruction-work budget. We optimize the calibration loss $\mathcal{L}(\mathcal{S})$ in Eq. (8) by local search over one-for-one unit reassignments, using only the calibration data of Section 5.1.

Deployed composition. The selected mask assigns 41 units to the reconstruction-free path and 23 to reconstructed execution. It is frozen before evaluation and shared by all tokens and context lengths. The held-out 8K–64K spans and the position-stability corpus of Appendix C.3 are not used for routing selection.

C.8 Limitations and Future Work

RoFold targets frozen RoPE models, and its compatibility result holds in exact arithmetic; Appendix C.5 quantifies the corresponding finite-precision residuals. The construction relies on frequencies shared across KV heads, so other positional encodings require a new analysis of their compatible subspaces. Generalization across model families, domains, and production serving kernels also remains future work. The ablations of Appendices C.4.1 and C.4.2 isolate two mechanisms within RoFold; the whole-pair control does not reproduce RAP's full pipeline, and a direct quantitative comparison with RAP remains future work.