KVFIXER: PREDICTIVE KV CACHE REPAIR FOR CONTEXT REVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Caching makes a model's past computation reusable, but not automatically revisable. Deleting an upstream span removes its key–value (KV) entries while leaving its influence in retained states. We introduce KVFixer, which predicts low-rank corrections from the original cache without replaying the retained suffix. It learns context-conditioned bases from paired executions and predicts edit-specific coordinates, refined with cache and output-distribution supervision. A fixed-endpoint refinement comparison lowers text and image KL, while paired projection diagnostics show concentrated rather than uniform gains across requests. On internal text and image panels, mean-token KL is 66.5% and 94.5% below reported deletion-only endpoints. External answer-choice KL is lower than the evaluated prefix controls and KVEraser adaptation. Replacing selected retained states with exact states further improves diagnostic fidelity, distinguishing compact representation, predicted correction and the output behavior of the revised cache.

1 INTRODUCTION

Efficient long-context inference depends on reusing computation, but that computation must remain consistent with the context a model is asked to use (Yao et al., 2025; Agarwal et al., 2025; Gim et al., 2024; Jin et al., 2025). Key–value (KV) caching supports reuse naturally during autoregressive decoding: new tokens use the cached states of an unchanged prefix. Context revision poses a different requirement. Previously processed evidence may be withdrawn before a subsequent query, while the remaining input is retained. Recomputing that input restores the appropriate state but repeats work on unchanged content. The challenge is therefore to make an existing computation revisable, rather than to choose between reusing it unchanged and computing it again.

We study this challenge for span deletion after prefill. In a Transformer, KV entries are contextualized representations, not independent records of input tokens. An upstream span influences the states of later tokens; deleting its own cache entries leaves those downstream states unchanged. The remaining cache can therefore continue to reflect information that is no longer in the input. We formulate **KV cache repair** as approximating the retained state of the revised execution from the original cache. Exact recomputation defines the target, but the repair should obtain the required update without replaying the retained suffix through the model.

Prior work provides several approaches to adapting cached computation. CacheBlend and CacheCraft selectively recompute token states to restore interactions across reused contexts (Yao et al., 2025; Agarwal et al., 2025). KVEraser learns steering states that compensate for a removed span while reusing the suffix (Li et al., 2026). Kamera represents conditioning differences with low-rank patches derived from paired caches (Ma et al., 2026). The latter suggests that repair need not reconstruct the full state: it can model the change instead. Yet a compact difference obtained from two executions is not an update available from one. A usable repair must identify directions that represent the correction and predict how much to change along each direction, without access to the recomputed cache. This is the step from representing context effects to predicting a state update.

We introduce **KVFixer**, which learns this step from paired training executions and applies it to new edits without suffix replay. KVFixer separates shared correction directions from edit-specific coordinates. A bank of bases is learned from training cache differences, and descriptors of the retained context select a basis for each request. The deleted-span KV summary and dirty retained states then supply the features for predicting coordinates. Decoding these coordinates produces a

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.

patch that is subtracted from the existing cache. Joint cache-reconstruction and output-distribution supervision trains the update toward the recomputed execution, with affine refinement folded into the predictor. We evaluate standalone prediction and, separately, offline combinations with exact retained states.

We examine both the repair and the structure it reuses. An orthogonal error decomposition separates representation loss from coordinate-prediction error. Controlled studies show sharing across removed spans and retained positions, with additional loss across retained contexts; independent text studies on Qwen3-VL, Llama-3.1 and Gemma-3 characterize the accompanying computations. A paired diagnostic on the deployed model shows lower mean output KL for learned updates than exact projection, even though exact projection performs better on most individual requests. This exposes a distributional trade-off between reconstruction and functional recovery. Fixed-endpoint refinement and shared-basis comparisons further probe these design choices without equating different training budgets.

On 1,200 text and 1,200 image sources with frozen Qwen3-VL-8B-Instruct, KVFixer reduces mean-token KL by 66.5% and 94.5% relative to the deletion-only point estimates. On MileBench and MM-NIAH, no-replay KVFixer has lower answer-choice KL than the evaluated controls and KVEraser adaptation. Offline combinations with exact states at 20% selected prefix positions yield full-vocabulary KL 0.1567 and 0.2305; these controls test complementary state updates, not an implemented partial-replay cost. Together with an operation-count analysis, these results connect conditional representations, prediction from existing computation, and fidelity to the revised execution.

2 Related work

2.1 Cache reuse and prediction

Cache reuse methods reduce repeated prompt computation by composing previously computed states. CacheBlend (Yao et al., 2025) and Cache-Craft (Agarwal et al., 2025) selectively recompute token states to restore interactions among cached segments. EPIC (Hu et al., 2025) and MPIC (Zhao et al., 2025) support position-independent reuse in text and multimodal settings, while KVLink (Yang et al., 2025) combines positional realignment with trainable linking tokens. TurboRAG (Lu et al., 2025) precomputes chunk-level caches for retrieval-augmented generation. KV Prediction (Horton et al., 2025) takes a learned approach to initial prefill: a smaller auxiliary model processes the prompt, and linear projections map its cache to the base model's cache for subsequent generation. These methods assemble or approximate the state needed for a prompt. KVFixer instead starts from a completed base-model execution and predicts how its retained states should change after deleting part of the input.

2.2 Low-rank representations and intervention responses

Low-rank attention and projection factorizations compress inference state (Saxena et al., 2024; Zhang et al., 2024); cross-layer sharing (Chang et al., 2025), online basis adaptation (Zhu et al., 2026) and residual coding against historical states (Hao et al., 2026) exploit further redundancy. Their goal is to preserve an execution with a smaller representation, rather than estimate the effect of changing its input. Intervention studies relate internal representations to behavior (Zhang & Nanda, 2023; Sharma et al., 2026). Sharma et al. (2026) model behaviors as low-rank subspaces and relate their overlap and decision coupling to cross-behavior effects. Our geometry experiments examine deletion-induced K/V differences: which directions generalize across removed spans and retained positions, and how that sharing depends on the retained context. This conditions the analysis on a fixed revised input rather than a behavioral category.

2.3 Context erasure and corrective cache editing

KVEraser (Li et al., 2026) learns a length-preserving steering block at the erased interval, leaving the suffix's KV states and positions unchanged. Trained to reproduce post-erasure behavior, the block compensates for the removed span through subsequent attention. KVFixer instead removes the interval, aligns retained rows to the revised positions, and predicts corrections to the suffix states

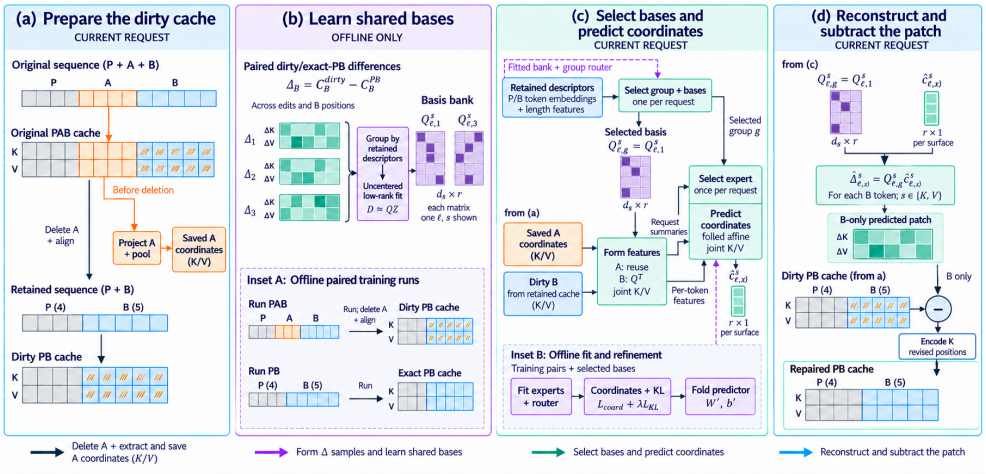

Figure 1: **Overview of KVFixer. (a)** Delete A, align retained B, and save A-side K/V coordinates. **(b)** Offline dirty/exact-PB differences are grouped by retained descriptors to learn low-rank bases. **(c)** Retained descriptors select a basis group, while saved A and dirty-B features select an expert and predict per-token coordinates. **(d)** Decode and subtract the B-only patch, then encode Keys at revised positions. Dashed arrows denote offline fitting or parameter transfer.

themselves. The distinction is the object being edited: a compensating block versus the retained state of the revised execution.

Kamera (Ma et al., 2026) separates positional relocation from low-rank conditioning patches. It constructs a patch from conditioned and canonical caches at compile time, then reuses it without a forward pass at serve time. Its analyses include shared directions and removal-induced residuals, while identifying inexpensive estimation of item-specific coefficients as an open problem. KVFixer studies this prediction problem for new deletions: paired executions provide training targets, but each runtime correction is inferred from the original execution without first constructing the exact revised cache. This extends the use of compact differences from compiled patches to predicted updates.

3 Geometry-Guided KV Cache Repair

KVFixer separates two requirements of repair: representing the change caused by deletion and predicting that change from an existing execution. Building on prior observations of low-dimensional conditioning residuals and intervention responses (Ma et al., 2026; Sharma et al., 2026), we first define the representation, then develop its runtime predictor and functional training objective.

3.1 Low-rank decomposition of context-induced changes

A frozen causal Transformer has processed a prefix P, a span A to be removed, and a retained suffix B. Let C^{PAB} be the original cache and C^{PB} the cache recomputed with A omitted and B's tokens held fixed. Deleting A's entries and aligning retained rows to native PB positions gives

$$C^{\text{dirty}} = \mathcal{T}(C^{PAB}), \qquad \Delta^{KV} = C^{\text{dirty}} - C^{PB}. \tag{1}$$

We model this **cache difference**, restricted to B; P is reused when its inputs and computation remain unchanged. The alignment changes stored coordinates without recomputing hidden states. Keys are represented before rotary encoding, retaining native Key normalization, and are re-encoded at revised positions after repair. Native positional and visual-preprocessing effects remain part of the target (Appendix A).

Fix a layer and a K/V surface of width d_s. For a retained computational context B, including P and the target position convention, collect change vectors across removed spans and retained positions into a weighted matrix $D_B \in \mathbb{R}^{d_s \times N}$. Truncated SVD gives

$$D_B = U\Sigma V^\top \approx Q_B^{\mathrm{ref}} Z_B, \qquad Q_B^{\mathrm{ref}} = U_{:,1:r}, \quad Z_B = (Q_B^{\mathrm{ref}})^\top D_B. \tag{2}$$

The columns of Q_B^{ref} are shared directions; each column of Z_B specifies a change's coordinates. The decomposition is uncentered because the mean change also requires repair. Appendix D.2 gives the weighting and equivalent second-moment fit.

This B-specific basis is a **representation reference** fitted from paired changes. Deployment instead selects from a finite training-fitted bank. The reference measures what directions can capture; the predictor must obtain a correction without the current request's C^{PB}.

3.2 From decomposition to predictive repair

Reusable geometry, edit-specific prediction. Retained-context descriptors select the correction space, while A-side and dirty-B features determine the update within it. We group paired training examples by retained descriptors and apply Equation (2) to each group's pooled differences. The bank stores an orthonormal basis $Q^s_{\ell,g} \in \mathbb{R}^{d_s \times r}$ for every layer ℓ, group g and surface $s \in \{K, V\}$.

At runtime, o_B is a 256-dimensional projection of a mean P embedding, four ordered B embedding bins and three length features. The descriptor $o_{A,B}$ additionally contains A-side and dirty-B cache summaries. Two fitted routers select a basis group and an expert:

$$g_B = \rho_B(o_B), \qquad h_{A,B} = \rho_{\mathrm{reg}}(o_{A,B}, g_B), \qquad \widehat{Q}^s_{\ell,B} = Q^s_{\ell,g_B}. \tag{2a}$$

These decisions are made once per request and shared across layers and positions; the bases and expert parameters remain layer-specific. Changing A while holding o_B fixed can change the expert but not the basis group. The implemented retained descriptors contain token embeddings, not retained-image content features.

Predicting joint K/V coordinates. Before deleting A, retain its K/V summaries. Let $a^s_{\ell,A} \in \mathbb{R}^r$ denote the pooled A-side coordinates in the selected basis, and let $k^{\mathrm{dirty}}_{\ell,x}, v^{\mathrm{dirty}}_{\ell,x}$ denote dirty B's states. With $g = g_B$ and $h = h_{A,B}$, the selected joint ridge uses

$$z_{\ell,x} = \begin{bmatrix} (\widehat{Q}^K_{\ell,B})^\top k^{\mathrm{dirty}}_{\ell,x} \\ (\widehat{Q}^V_{\ell,B})^\top v^{\mathrm{dirty}}_{\ell,x} \\ a^K_{\ell,A} \\ a^V_{\ell,A} \end{bmatrix}, \qquad c^0_{\ell,x} = W_{\ell,g,h} z_{\ell,x} + b_{\ell,g,h}. \tag{2b}$$

Here $z_{\ell,x} \in \mathbb{R}^{4r}$, $c^0_{\ell,x} = [c^{K,0}_{\ell,x}; c^{V,0}_{\ell,x}] \in \mathbb{R}^{2r}$, and $W_{\ell,g,h} \in \mathbb{R}^{2r \times 4r}$. The affine notation absorbs fitted input/output standardization, so predicted coordinates, reconstruction targets and losses use the original orthonormal-basis units. Experts are fitted in their selected group's coordinates on a source-disjoint partition, with out-of-fold targets for routing. The deployed configuration uses rank 256, two groups and two experts per group; Appendix D records the fitting sequence and specification details.

Applying the correction. Refinement in Section 3.3 produces $\widehat{c}_{\ell,x} = [\widehat{c}^K_{\ell,x}; \widehat{c}^V_{\ell,x}]$. Each surface is decoded separately:

$$\widehat{\Delta}^s_{\ell,x} = \widehat{Q}^s_{\ell,B} \widehat{c}^s_{\ell,x}, \qquad \widehat{C}^{PB}_B = C^{\mathrm{dirty}}_B - \widehat{\Delta}^{KV}. \tag{2c}$$

The learned update requires no suffix forward. Section 4 additionally studies offline combinations with exact retained states. For one layer, position and surface, write the selected basis as Q. Orthonormality separates the remaining Euclidean error into

$$\|\Delta - Q\widehat{c}\|_2^2 = \underbrace{\|(I - QQ^\top)\Delta\|_2^2}_{\text{representation error}} + \underbrace{\|Q^\top\Delta - \widehat{c}\|_2^2}_{\text{coordinate-prediction error}} . \tag{3}$$

The first term measures directions outside the selected space. The second measures coordinate deviation from the Euclidean projection, including expert selection, regression and refinement. Section 4.3 evaluates both errors and their output consequences; Appendix A.3 gives a conditional fixed-query bound.

3.3 Refining for functional fidelity

State reconstruction and output fidelity are distinct objectives. With the bases fixed, refinement changes coordinates but not the representation residual in Equation (3). We freeze the backbone, bases, routes and ridge, then apply a diagonal affine map followed by a scale-normalized rank-16 residual. Suppressing layer/group/expert indices and writing frozen output scales as $S = \mathrm{diag}(s)$, the column-vector form is

$$\begin{aligned} a &= \mathrm{diag}(1+\gamma)c^0 + S\beta, \\ \widehat{c} &= (I + SU_\uparrow^\top D_\downarrow^\top S^{-1})a = Fc^0 + \alpha. \end{aligned} \tag{4}$$

Here $D_\downarrow \in \mathbb{R}^{2r\times 16}$ and $U_\uparrow \in \mathbb{R}^{16\times 2r}$ are independent per layer/group/expert and shared across positions. The up factor is zero-initialized; the down factor has initialization scale $(2r)^{-1/2}$ (Appendix D.3).

The joint target $c^*_{\ell,x}$ concatenates $(\widehat{Q}^s_{\ell,B})^\top \Delta^s_{\ell,x}$ for K and V. We optimize

$$\mathcal{L} = \mathbb{E}_{\text{balanced sources}} \left[\frac{\sum_{\ell,x} \|\widehat{c}_{\ell,x} - c^*_{\ell,x}\|_2^2}{\max(\|\Delta^{KV}\|_F^2, \epsilon)} + \frac{\lambda}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}(p_{PB,t} \| p_{\mathrm{repair},t}) \right]. \tag{5}$$

Both distributions follow the same PB-teacher greedy prefix to EOS or eight prediction positions, with full-vocabulary KL and $\lambda = 1$. The rank-16 coordinate residual is distinct from the rank-256 correction bases.

For each layer/group/expert, folding $W' = FW$ and $b' = Fb + \alpha$ into Equation (2b) leaves only routing, projection, affine prediction and reconstruction online. No additional network or clean endpoint is required (Appendices D.3 and G).

Algorithm 1. Applying KVFixer after prefill.
Input: original cache, retained token/position metadata, deletion boundary, and fitted routers, bases and folded experts.

1. Select the basis group from retained descriptors; summarize A in that group's K/V bases before deleting it.
2. Delete A and align retained rows. Select the expert using A-side and dirty-B summaries.
3. For every layer and B position, form the four feature blocks in Equation (2b), predict coordinates with the folded expert, and subtract the decoded K/V patch. Encode Keys at the revised positions.

Output: a PB-shaped repaired cache ready for the query. Exact PB is not an input to the learned update; it is used separately by the offline replacement controls.

Online arithmetic. Counting a multiply-accumulate as two FLOPs, the direct evaluation of Equations (2b)–(2c), with equal K/V width d_{KV} and rank r, gives

$$C_{\mathrm{core}} = Ln_B(8d_{KV}r + 16r^2), \qquad C_{\mathrm{online}} = C_{\mathrm{core}} + C_{\mathrm{aux}}. \tag{5a}$$

The core counts dirty K/V projection, joint prediction and patch reconstruction. Auxiliary work includes descriptors, A summaries, routing, alignment, scalar operations and required retained-vision

computation. Appendix G derives repair/replay costs at a common post-prefill boundary and separates arithmetic from storage.

4 EXPERIMENTS

We evaluate predicted repair, controlled combinations with exact states (Table 1), and the relation between reconstruction and output fidelity (Table 2). Source-level reaggregation adds paired uncertainty and horizon diagnostics without refitting the model.

4.1 EXPERIMENTAL SETUP AND METRICS

Model and panels. We fit one repairer on frozen Qwen3-VL-8B-Instruct (Qwen, n.d.), with 24,000 basis-only and 96,000 expert/router/refinement sources, plus 1,600 calibration. Training uses HotpotQA (Yang et al., 2018)/MisBench (Peng et al., 2025), LLaVA-OneVision-Data (LMMs-Lab, n.d.a) and LLaVA-Video-178K (LMMs-Lab, n.d.b). The primary KVFixer results use the epoch-5 export; additional fixed-endpoint comparisons are reported separately. The internal panel contains 1,200 text sources and 1,200 image sources admitted with retained length $B \geq 32$; their source IDs are disjoint from the fitting and calibration partitions. External panels contain 300 MileBench (Song et al., 2024) and 300 MM-NIAH (Wang et al., 2024) sources and are excluded from fitting and checkpoint selection.

Comparators and information access. The learned update reads A-side summaries and dirty B, not exact PB. Prefix/uniform **exact-row replacement controls** copy selected rows from a precomputed exact-PB cache into dirty or repaired states; their percentages denote selected rows, not executed decoder FLOPs. These controls measure the value of partial exact state information rather than an online replay implementation. Internal combinations are evaluated in the same source-level runs as standalone repair. KVEraser is included as a separately trained adaptation.

Fixed-endpoint comparisons. A refinement pilot compares the KV-only export (zero recorded refinement epochs) with one epoch of KV+KL on both Text and Image. Shared-Q/fixed-routing is evaluated at epoch 2 on both subsets. The configurations retain the primary model's recorded request-level route decisions. Epoch counts and paired comparisons are given in Table 2(D) and Appendix F.5.

Output metrics. At common PB-teacher positions, with reached length $T_i \leq 64$ and equal source weights w_i,

$$\begin{aligned} \mathrm{KL}(m) &= \sum_i w_i \frac{1}{T_i} \sum_{t=1}^{T_i} D_{\mathrm{KL}}(p_{it}^{PB} \| p_{it}^{m}), \\ \mathrm{Agree}(m) &= 100 \sum_i w_i \frac{1}{T_i} \sum_{t=1}^{T_i} \mathbf{1}[\arg\max p_{it}^{m} = \arg\max p_{it}^{PB}]. \end{aligned} \tag{6}$$

External choice KL/agreement score choice-normalized probabilities at the registered answer position; task accuracy uses labels. KL is in nats. Early-EOS sources remain in each horizon mean with their available positions. KVEraser is compared in choice space; Table 1 leaves its continuation field blank rather than mixing a separately scored trace with the common teacher-path metric.

Geometry and statistics. For fixed B, rank-r capture is

$$\mathrm{Cap}_r(X; Q) = 100 \left[1 - \frac{\mathbb{E}_{A,x} \|(I - QQ^\top)X\|^2}{\mathbb{E}_{A,x} \|X\|^2} \right]. \tag{7}$$

Controlled studies average B-level energy ratios: eight text B contexts crossed with 32 fit/32 held-out image A, and two disjoint 40-B text streams per model (Qwen, n.d.; Meta, n.d.; Google, n.d.). Kc denotes normalized, pre-RoPE Key content; rank 256 is 25% of K/V width. The deployed-bank diagnostic measures all-layer realized patches and their projected coordinates, with source-normalized errors; finite-precision residuals are separate from the ideal identity (Appendix F.3).

Table 1: **Repair and exact-state replacement diagnostics.** Internal KL/agreement and external choice KL/full KL/choice agreement/accuracy; percentages are nominal selection settings, not executed replay. Bold marks the best displayed comparable mean, not an equal-cost ranking. Full sweeps and intervals: Appendices E–F.

(a) Internal: full-vocabulary fidelity

Method	Nominal selection	Text (1,200) KL ↓	Text (1,200) Agree. ↑	Image (1,200; $B \geq 32$) KL ↓	Image (1,200; $B \geq 32$) Agree. ↑
Original PAB	–	0.4738	89.12	0.9234	82.54
Exact PB	100%	0.0000	100.00	0.0000	100.00
Delete-and-shift	0%	0.0869	95.33	2.2449	62.25
Exact-row prefix control	20%	0.0375	96.94	0.3921	87.81
Exact-row prefix control	50%	0.0230	**97.83**	0.3045	89.17
Exact-row uniform control	20%	0.0382	96.92	0.1118	95.44
Exact-row uniform control	50%	0.0331	97.38	0.1059	**95.49**
KVFixer	0%	0.0291	96.92	0.1226	89.88
KVFixer + exact prefix	20%	**0.0173**	97.72	0.1049	89.82
KVFixer + exact uniform	20%	0.0219	97.14	**0.0941**	90.63
KVEraser (image adaptation)	0%	–	–	0.5979	77.67

(b) External: choice and teacher-path observations

Method	MileBench (300) C-KL ↓	MileBench (300) F-KL ↓	MileBench (300) C-Ag. ↑	MileBench (300) Acc. ↑	MM-NIAH (300) C-KL ↓	MM-NIAH (300) F-KL ↓	MM-NIAH (300) C-Ag. ↑	MM-NIAH (300) Acc. ↑
Original PAB	–	0.4017	–	67.67	–	0.4472	–	77.00
Exact PB	0.0000	0.0000	100.00	33.67	0.0000	0.0000	100.00	37.00
Dirty deletion	2.0261	1.3883	22.33	40.67	1.4615	1.6953	30.33	60.67
Exact-prefix control (20%)	2.0239	0.3431	43.33	**67.33**	1.4284	0.3096	52.67	77.67
Exact-prefix control (50%)	1.9817	0.2545	44.67	**67.33**	1.3648	0.2542	52.33	**78.00**
Unrefined reference	1.3761	0.9815	31.00	59.67	1.5449	1.9730	19.33	45.67
KVFixer	0.4731	0.2178	40.33	52.33	0.8366	1.3412	44.33	46.00
KVFixer + exact prefix (20%)	**0.4401**	0.1567	58.67	57.00	0.8304	0.2305	56.00	68.33
KVFixer + exact uniform (20%)	0.4563	**0.1470**	**59.67**	56.67	**0.3991**	**0.1930**	**65.67**	56.67
KVEraser adaptation	1.3382	–	33.67	59.67	1.6165	–	25.33	48.67

C-KL/C-Ag.: choice KL/agreement; F-KL: full-vocabulary teacher-path KL.

New 95% percentile intervals use 2,000 paired source draws, seed 20260925, conditional on fitted parameters. Subgroup intervals are exploratory and unadjusted.

4.2 Repair experiments

Table 1 distinguishes learned updates from target-assisted replacement controls. The selector does not inspect target values; the replacement operation nevertheless requires exact states.

Predicted updates improve mean fidelity. Standalone Text/Image KL is 0.029111/0.122648, 66.5%/94.5% below the earlier deletion-only endpoints. These endpoint comparisons are descriptive across runs. Within E1, adding prefix-20% exact rows lowers KL to 0.017305/0.104861. Paired differences from standalone are -0.0118 [-0.0155, -0.0085] and -0.0178 [-0.0253, -0.0111]; 59.75%/56.75% of sources improve. Thus the combination improves the mean without uniformly improving every request.

Exact-state replacement complements prediction externally. At 20% selected prefix rows, the combination yields full-vocabulary KL 0.156739 on MileBench and 0.230466 on MM-NIAH. Relative to the corresponding exact-prefix controls, the paired differences are -0.1863 [-0.2112, -0.1635] and -0.0791 [-0.1036, -0.0523]. Full sweeps and paired contrasts are in Appendix E. These are exact-state combination diagnostics, not measurements of partial-recomputation cost.

Table 2: **Reconstruction, geometry and fixed-endpoint comparisons. A:** deployed-bank E1 diagnostics. **B–C:** rank-256 geometry. **D:** H=64 fixed endpoints at their reported refinement epochs; not equal-budget ablations. Details: Appendix F.5.

A. Deployed bank: mean and request-level fidelity		
Diagnostic	Text	Image
Representation residual / delta energy (%)	20.48	23.12
Projected-coordinate residual / delta energy (%)	62.50	37.27
Realized KV residual / delta energy (%)	82.99	60.39
Predicted update: mean teacher-path KL	0.029111	0.122648
Exact projection: mean teacher-path KL	0.090327	0.207185
Predicted update lower KL: sources (%)	10.50	45.25
Median predicted-minus-projection KL	+0.00261	+0.00501
B. Controlled image-A/text-B; terminal capture		
Basis	Key-content	Value
B-specific	96.65	88.43
Other-B	88.36	60.83
Reported two-group diagnostic	89.61	62.95
Centered B-specific	85.27	70.70
C. Cross-model text; primary / replication		
Model	Key-content	Value
Qwen3-VL-8B	94.93 / 95.11	83.51 / 83.59
Llama-3.1-8B	94.13 / 94.14	82.83 / 83.50
Gemma-3-4B	90.38 / 90.31	81.07 / 81.27
D. Fixed endpoints: KL / agreement (%)		
Configuration (refinement epochs)	Text	Image
KV-only export (0)	0.082833 / 95.5402	0.704690 / 77.0922
KV+KL refinement (1)	0.032285 / 96.6760	0.133141 / 89.6395
Shared-Q, fixed routing (2)	0.031678 / 96.7778	0.125301 / 89.9692

External repair is stable under source-paired comparisons. Versus dirty deletion, answer-choice KL falls by 1.5529 [1.3739, 1.7286] nats on MileBench and 0.6249 [0.4289, 0.8378] on MM-NIAH; teacher-path KL falls by 1.1705 [1.0899, 1.2492] and 0.3541 [0.1833, 0.5387]. Choice agreement rises by 18.0/14.0 points, while accuracy changes by +11.7/-14.7 points, separating PB fidelity from preservation of the original answer. KVFixer also has lower choice KL than the reported exact-prefix controls and KVEraser adaptation.

Analytic cost at observed retained lengths. Under the supplied P-empty, A=256 dense-operation model, observed B-length p10–p90 ratios are 1.21–1.43% of exact-PB replay on MileBench and 0.42–0.80% on MM-NIAH (Appendix G). These are arithmetic, not wall-clock, ratios and exclude memory, kernel/communication and vision-front-end costs.

4.3 Geometry and repair diagnostics

The paired E1 records compare predicted and exact-projection updates in the same selected bank. They also reveal how aggregate KL differs from typical-request behavior.

Lower mean KL does not imply a typical-request advantage. Final-minus-projection KL is -0.0612 [-0.0864, -0.0387] on Text and -0.0845 [-0.1088, -0.0631] on Image. Yet only 126/1,200 Text and 543/1,200 Image sources improve; median differences are +0.00261 and +0.00501. Large improvements on a subset outweigh more frequent smaller losses. Text median/99th-percentile KL is 0.00559/0.33369 for prediction versus 0.00086/2.36745 for projection. The resulting mean-versus-typical trade-off is more informative than a single average ranking.

Horizon comparisons expose the same structure. KVFixer and Shared-Q track closely across horizons, whereas KV-only remains substantially higher, especially on Image; exact projection

is lower only at Image H=1 and higher thereafter. Figure 2 uses EOS-truncated teacher paths (35.33%/48.50% of Text/Image sources reach H=64). The fixed endpoints are not equal-budget ablations; task-level slices remain in Appendix F.4.

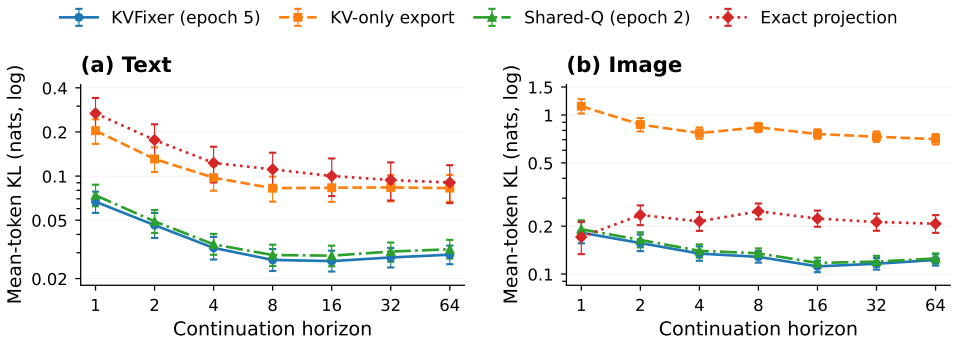

Figure 2: **Output fidelity over continuation horizon.** Final KVFixer is compared with the KV-only export, Shared-Q, and the exact-projection reference; bars are 95% source-bootstrap intervals. Fixed-endpoint variants use their reported refinement epochs and are not equal-update-budget ablations. Paths stop at PB EOS.

Fixed-endpoint variants. One KV+KL epoch lowers Text/Image KL from 0.082833/0.704690 to 0.032285/0.133141 (61.0%/81.1%); paired differences are −0.05055 [−0.06489, −0.03806] and −0.57155 [−0.61887, −0.52574]. Epoch-2 Shared-Q remains close to the epoch-5 primary (0.031678/0.125301 versus 0.029111/0.122648). Because update counts and endpoints differ, these comparisons do not isolate the KL term or basis conditioning (Appendix F.5).

5 Discussion and Limitations

State updates and output risk. Exact projection wins on most requests, while prediction avoids some much larger errors, so mean and typical-request behavior diverge. Refinement and Shared-Q remain informative fixed-endpoint comparisons, but their unequal update counts do not isolate individual design effects. Exact-state replacement further improves fidelity using precomputed clean states, diagnosing complementary information rather than online replay efficiency. **Scope.** Functional repair uses one backbone and fixed PB teacher paths; three-model evidence concerns geometry rather than repair.

6 Conclusion

KVFixer predicts edit-specific updates in reusable correction spaces without replaying the retained suffix. Its mean-KL gains over exact projection are concentrated rather than universal, while fixed-endpoint variants show that refinement and shared bases both support useful repair. Together, the results separate representational capacity, functional fidelity and the computation needed to revise cached state.

Reproducibility Statement

Appendices A–G specify training, evaluation, statistics, geometry diagnostics and symbolic operation counts; source-level scripts reproduce the paired statistics and horizon figures. Efficiency claims use arithmetic counts rather than wall-clock timing.

AI Use Statement

AI assistance supported manuscript restructuring, code writing, figure design, and portions of the mathematical exposition; the authors remain responsible for verifying the experiments, scientific claims and final text.

REFERENCES

Shubham Agarwal, Sai Sundaresan, Subrata Mitra, Debabrata Mahapatra, Archit Gupta, Rounak Sharma, Nirmal Joshua Kapu, Tong Yu, and Shiv Kumar Saini. Cache-Craft: Managing Chunk-Caches for Efficient Retrieval-Augmented Generation. *Proceedings of the ACM on Management of Data*, 3(3):136:1–136:28, 2025. doi: 10.1145/3725273. URL https://doi.org/10.1145/3725273.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Hung-Yueh Chiang, Yash Akhauri, Xilai Dai, Huiqiang Jiang, Yucheng Li, Luis Ceze, Kai-Chiang Wu, and Mohamed S. Abdelfattah. xKV: Cross-Layer KV-Cache Compression via Aligned Singular Vector Extraction. arXiv preprint arXiv:2503.18893v2, 2025. URL https://arxiv.org/abs/2503.18893v2. Version 2 revised in 2026; authors report acceptance at ICML 2026.

In Gim, Guojun Chen, Seung seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt Cache: Modular Attention Reuse for Low-Latency Inference. In *Proceedings of Machine Learning and Systems*, volume 6, pp. 325–338, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/a66caa1703fe34705a4368c3014c1966-Abstract-Conference.html.

Google. Gemma-3-4B-IT. Model card, n.d. URL https://huggingface.co/google/gemma-3-4b-it. Accessed September 26, 2026.

Jitai Hao, Qiang Huang, Yaowei Wang, Min Zhang, and Jun Yu. DeltaKV: Residual-Based KV Cache Compression via Long-Range Similarity. arXiv preprint arXiv:2602.08005, 2026. URL https://arxiv.org/abs/2602.08005.

Maxwell Horton, Qingqing Cao, Chenfan Sun, Yanzi Jin, Sachin Mehta, Mohammad Rastegari, and Moin Nabi. KV Prediction for Improved Time to First Token. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models (SCOPE) at ICLR 2025*, 2025. URL https://openreview.net/forum?id=Bni1wy49vY.

Junhao Hu, Wenrui Huang, Weidong Wang, Haoyi Wang, Tiancheng Hu, Qin, Zhang, Hao Feng, Xusheng Chen, Yizhou Shan, and Tao Xie. EPIC: Efficient Position-Independent Caching for Serving Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24391–24402. PMLR, 2025. URL https://proceedings.mlr.press/v267/hu25j.html.

Chao Jin, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Shufan Liu, Xuanzhe Liu, and Xin Jin. RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation. *ACM Transactions on Computer Systems*, 44(1):2:1–2:27, 2025. doi: 10.1145/3768628. URL https://doi.org/10.1145/3768628.

Mufei Li, Shikun Liu, Dongqi Fu, Haoyu Wang, Yinglong Xia, Hong Li, Hong Yan, and Pan Li. KVEraser: Learning to Steer KV Cache for Efficient Localized Context Erasing. arXiv preprint arXiv:2606.17034v2, 2026. URL https://arxiv.org/abs/2606.17034v2.

LMMs-Lab. LLaVA-OneVision-Data. Dataset card, n.d.a. URL https://huggingface.co/datasets/lmms-lab/LLaVA-OneVision-Data. Accessed September 26, 2026.

LMMs-Lab. LLaVA-Video-178K. Dataset card, n.d.b. URL https://huggingface.co/datasets/lmms-lab/LLaVA-Video-178K. Accessed September 26, 2026.

Songshuo Lu, Hua Wang, Yutian Rong, Zhi Chen, and Yaohua Tang. TurboRAG: Accelerating Retrieval-Augmented Generation with Precomputed KV Caches for Chunked Text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6588–6601. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.334. URL https://aclanthology.org/2025.emnlp-main.334/.

Bole Ma, Jan Eitzinger, Harald Koestler, and Gerhard Wellein. Kamera: Unified Position-Invariant Multimodal KV Cache for Training-Free Reuse. arXiv preprint arXiv:2606.23581v1, 2026. URL https://arxiv.org/abs/2606.23581v1.

Meta. Llama-3.1-8B-Instruct. Model card, n.d. URL https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Accessed September 26, 2026.

Miao Peng, Nuo Chen, Jianheng Tang, and Jia Li. How does Misinformation Affect Large Language Model Behaviors and Preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13711–13748. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.674. URL https://aclanthology.org/2025.acl-long.674/.

Qwen. Qwen3-VL-8B-Instruct. Model card, n.d. URL https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct. Accessed September 26, 2026.

Utkarsh Saxena, Gobinda Saha, Sakshi Choudhary, and Kaushik Roy. Eigen Attention: Attention in Low-Rank Space for KV Cache Compression. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15332–15344. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.899. URL https://aclanthology.org/2024.findings-emnlp.899/.

Angira Sharma, Christian Schroeder de Witt, Philip Torr, Anisoara Calinescu, and Jialin Yu. A Low-Rank Subspace Analysis of LLM Interventions. arXiv preprint arXiv:2606.14388v1, 2026. URL https://arxiv.org/abs/2606.14388v1.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. MileBench: Benchmarking MLLMs in Long Context. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Uhwze2LEwq.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. Needle In A Multimodal Haystack. In *Advances in Neural Information Processing Systems*, volume 37, pp. 20540–20565, 2024. doi: 10.52202/079017-0649. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/24a8968affe71ffe4067d022b9d16566-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Jingbo Yang, Bairu Hou, Wei Wei, Yujia Bao, and Shiyu Chang. KVLink: Accelerating Large Language Models via Efficient KV Cache Reuse. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4461. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c217d123885341853cecbdd2be809983-Abstract-Conference.html.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 94–109. Association for Computing Machinery, 2025. doi: 10.1145/3689031.3696098. URL https://doi.org/10.1145/3689031.3696098.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. arXiv preprint arXiv:2309.16042, 2023. URL https://arxiv.org/abs/2309.16042.

Rongzhi Zhang, Kuang Wang, Liyuan Liu, Shuohang Wang, Hao Cheng, Chao Zhang, and Yelong Shen. LoRC: Low-Rank Compression for LLMs KV Cache with a Progressive Compression Strategy. arXiv preprint arXiv:2410.03111, 2024. URL https://arxiv.org/abs/2410.03111.

Shiju Zhao, Junhao Hu, Rongxiao Huang, Jiaqi Zheng, and Guihai Chen. MPIC: Position-Independent Multimodal Context Caching System for Efficient MLLM Serving. arXiv preprint arXiv:2502.01960v2, 2025. URL https://arxiv.org/abs/2502.01960v2.

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 2026*, pp. 10161–10178. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.494. URL https://aclanthology.org/2026.findings-acl.494/.

A Deletion target and error identities

A.1 Retained-state coordinates

The target is native deletion with B's realized tokens held fixed. Alignment changes stored coordinates but does not recompute the hidden states from which the cache was written. P is reused when its input and computation are unchanged. For Qwen3-VL, Key-content is measured after native Key normalization and before RoPE; repaired Keys are encoded at native PB positions. Values use native decoder rows.

A.2 Orthogonal error decomposition

For an orthonormal selected basis Q,

$$\Delta - Q\widehat{c} = (I - QQ^\top)\Delta + Q(Q^\top\Delta - \widehat{c}).$$

The two terms are orthogonal, giving Equation (3) after summing across layers, B positions and K/V surfaces. With fixed bases, refinement changes only the coordinate term.

A.3 Fixed-query connection

For a fixed query or shared teacher continuation q, suppose the score map F_q is locally Lipschitz between C^{PB} and $\widehat{C}^{PB}$. Then

$$\|F_q(\widehat{C}^{PB}) - F_q(C^{PB})\| \le L_q\|\Delta^{KV} - \widehat{\Delta}^{KV}\|. \tag{A1}$$

This motivates separate measurements of representation, reconstruction and output fidelity; no numerical Lipschitz constant is estimated.

B Controlled image/text geometry

B.1 Rank and cross-context transfer

Eight fixed text B contexts (three AI2D, three ChartQA and two ScienceQA) are crossed with 32 fit-A and 32 held-out-A images. Qwen3-VL-8B-Instruct is frozen; K/V width is 1,024. A basis is fitted per B/layer/surface using fit-A only and shared across retained positions. Terminal capture is:

Rank	K: B-specific	K: other-B	V: B-specific	V: other-B
64	91.32%	75.56%	76.37%	36.89%
128	94.41%	82.07%	82.79%	47.04%
256	96.65%	88.36%	88.43%	60.83%
512	98.53%	95.07%	94.01%	79.57%

At rank 256, the two-group diagnostic bank captures 89.61% K and 62.95% V on this panel.

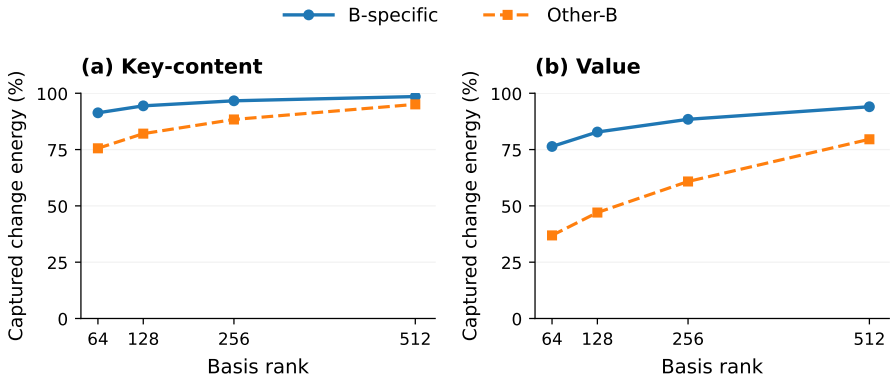

Figure 3: **Conditional rank–capture.** Terminal Kc and V at Qwen3-VL layer 35. Each projector serves all positions; other-B excludes the evaluated context. Lines connect measured ranks.

B.2 Position sharing and centered variation

Using one rank-256 projector across positions, relative positions 0/25/50/75/100% have capture 99.57/95.72/95.03/95.83/95.69% for K and 90.55/88.49/85.49/90.29/93.15% for V. Centering each split across A at fixed position gives terminal capture 85.27% K and 70.70% V, showing substantial structure beyond the conditional mean.

C Cross-model geometry

C.1 Terminal capture

Qwen3-VL-8B-Instruct, Llama-3.1-8B-Instruct and Gemma-3-4B-it use two source-disjoint streams of 40 B contexts per model, with 32 fit-A and 32 held-out-A examples per stream. Bases are fitted per layer/surface and shared across positions. Terminal rank-256 capture (% [95% CI]) is:

Model	K-content, primary	K-content, replication	V, primary	V, replication
Qwen3-VL	94.93 [94.58, 95.23]	95.11 [94.75, 95.44]	83.51 [82.42, 84.48]	83.59 [82.46, 84.73]
Llama-3.1	94.13 [93.80, 94.44]	94.14 [93.78, 94.51]	82.83 [81.96, 83.67]	83.50 [82.66, 84.39]
Gemma-3	90.38 [89.83, 90.86]	90.31 [89.86, 90.76]	81.07 [79.94, 82.05]	81.27 [80.43, 82.18]

C.2 Computational diagnostics

Concentration is already present after the first block and varies nonmonotonically with depth. Primary/replication hidden capture after block 0 is 90.31/90.91% for Qwen, 90.09/90.44% for Llama and 99.77/99.72% for Gemma. At the final block input it is 79.16/79.16%, 74.33/75.13% and 79.40/79.70%.

At the terminal block, the inside/outside signed-energy ratios $(R_{\text{in}}, R_{\text{out}})$ are Qwen (1.102/1.105, 1.036/1.038), Llama (0.916/0.914, 1.038/1.038) and Gemma (1.431/1.468, 1.037/1.036), primary/replication. Thus Qwen/Gemma's relative concentration is associated with stronger inside reinforcement, whereas Llama shows the opposite relative pattern; none exhibits absolute outside cancellation at the terminal block.

Fixed-weight readout directions capture less of the terminal geometry than learned bases. Key-linear learned/fixed capture is 92.59/68.83% for Qwen, 94.13/79.37% for Llama and 88.70/64.74% for Gemma on the primary stream; Value learned/fixed capture is 83.51/44.79%, 82.83/41.36% and 81.07/43.36%.

D TRAINING AND IMPLEMENTATION

D.1 DATA AND FEATURES

The backbone and vision weights are frozen. Fitting uses HotpotQA/MisBench, LLaVA-OneVision-Data and LLaVA-Video-178K:

Training modality	Fit sources	Basis-only	Expert/router/refinement
Text	78,500	15,700	62,800
Image	40,000	8,000	32,000
Video	1,500	300	1,200
Total	**120,000**	**24,000**	**96,000**

Calibration contains 1,600 additional sources. The reported functional panel contains 1,200 text sources and 1,200 image sources admitted with $B \geq 32$; their source IDs are disjoint from the fitting and calibration partitions.

The retained-context descriptor concatenates a mean P embedding, four ordered B embedding bins and three length features, followed by a fixed-width projection. Weighted k-means defines two basis groups. A-side and dirty-B summaries feed a request-level expert router. The joint predictor input is dirty K, dirty V, pooled A K and pooled A V coordinates (width 1,024), and predicts 512 K/V coordinates.

D.2 BASIS AND EXPERT FITTING

Basis-only and expert/router/refinement partitions are source-disjoint. Group/layer/surface second moments define the basis bank. Experts regress joint coordinate targets in their assigned bases, with source-level out-of-fold targets for routing. The deployed configuration uses rank 256, two basis groups and two experts per group.

For weighted native-deletion vectors δ_j,

$$M = \sum_j w_j \delta_j \delta_j^\top. \tag{D1}$$

The leading r eigenvectors of M minimize weighted squared reconstruction error and are equivalent to the left singular vectors of the weighted data matrix. No centering is applied.

D.3 FUNCTIONAL REFINEMENT AND FOLDING

The backbone, bank, routes and base ridge are frozen during refinement. Training uses the joint cache-coordinate loss plus PB-teacher KL in Equation (5), with teacher horizon up to eight positions. Evaluation uses horizons through 64.

For each group/expert/layer, the row-vector adapter is

$$\begin{aligned} a_{\mathrm{row}} &= c^0_{\mathrm{row}} \odot (1+\gamma) + \beta \odot s, \\ \widehat{c}_{\mathrm{row}} &= a_{\mathrm{row}} + [(a_{\mathrm{row}} \oslash s) D_\downarrow U_\uparrow] \odot s. \end{aligned} \tag{D2}$$

The gain/offset tensors have shape $2\times2\times36\times512$; down/up factors have shapes $2\times2\times36\times512\times16$ and $2\times2\times36\times16\times512$. Rank 16 applies to the coordinate-space residual, not the rank-256 cache basis. The affine composition is folded into the deployed ridge as $W' = FW$, $b' = Fb + \alpha$.

E EXTERNAL EVALUATION

E.1 PANELS AND METRICS

MileBench and MM-NIAH each contain 300 source groups with one registered request per source. MM-NIAH equally divides image reasoning and retrieval and retains query-side images in B. External targets are excluded from fitting and checkpoint selection.

Choice KL/agreement use probabilities normalized over registered answer choices at one answer position. Continuation KL/agreement use the common PB teacher path of Equation (6), truncated at EOS or 64 positions. Exact-prefix/uniform controls copy selected exact-PB rows and therefore measure partial exact-state correction rather than online recomputation. KVEraser is reported in choice space as a separately trained adaptation.

E.2 Standalone and exact-row results

Marginal 95% source-bootstrap intervals (2,000 draws) for the main standalone arms are:

Panel / arm	Choice KL [95% CI]	Full-vocabulary KL [95% CI]
MileBench / Dirty deletion	2.0261 [1.8434, 2.2088]	1.3883 [1.3078, 1.4760]
MileBench / Unrefined reference	1.3761 [1.2259, 1.5294]	0.9815 [0.9076, 1.0625]
MileBench / KVFixer	0.4731 [0.4186, 0.5340]	0.2178 [0.1989, 0.2404]
MM-NIAH / Dirty deletion	1.4615 [1.2798, 1.6533]	1.6953 [1.5554, 1.8479]
MM-NIAH / Unrefined reference	1.5449 [1.4151, 1.6811]	1.9730 [1.8536, 2.1045]
MM-NIAH / KVFixer	0.8366 [0.7354, 0.9432]	1.3412 [1.2406, 1.4407]

The complete exact-row sweep is:

Panel	Method	Choice KL	Full KL	Choice agree. (%)	Accuracy (%)
MileBench	Exact-prefix 5%	2.0491	0.3653	42.33	68.00
MileBench	KVFixer + exact prefix 5%	0.4450	0.1730	58.00	57.00
MileBench	KVFixer + exact prefix 20%	0.4401	0.1567	58.67	57.00
MileBench	KVFixer + exact prefix 50%	0.4923	0.1020	60.33	56.33
MileBench	KVFixer + exact uniform 20%	0.4563	0.1470	59.67	56.67
MileBench	KVFixer + exact uniform 50%	0.2791	0.0989	71.00	48.00
MileBench	KVEraser adaptation	1.3382	—	33.67	59.67
MM-NIAH	Exact-prefix 5%	1.4268	0.3466	53.00	78.33
MM-NIAH	KVFixer + exact prefix 5%	0.8269	0.2632	53.33	69.67
MM-NIAH	KVFixer + exact prefix 20%	0.8304	0.2305	56.00	68.33
MM-NIAH	KVFixer + exact prefix 50%	0.8382	0.1795	56.67	68.67
MM-NIAH	KVFixer + exact uniform 20%	0.3991	0.1930	65.67	56.67
MM-NIAH	KVFixer + exact uniform 50%	0.1474	0.0888	74.00	49.33
MM-NIAH	KVEraser adaptation	1.6165	—	25.33	48.67

E.3 PAIRED CONTRASTS

Differences are left minus right; negative KL favors the left. Intervals use 2,000 paired source-bootstrap draws.

Panel	Contrast	Metric	Mean difference	95% interval
MileBench	KVFixer − dirty	Choice KL	-1.5529	[-1.7286, -1.3739]
MileBench	KVFixer − dirty	Full KL	-1.1705	[-1.2492, -1.0899]
MileBench	KVFixer − dirty	Choice agreement	+18.00 pp	[+11.67, +24.67] pp
MileBench	KVFixer − dirty	Accuracy	+11.67 pp	[+4.67, +18.67] pp
MileBench	KVFixer + exact prefix 20% − exact-prefix control 20%	Full KL	-0.1863	[-0.2112, -0.1635]
MileBench	KVFixer + exact prefix 20% − exact-prefix control 20%	Choice KL	-1.5838	[-1.7848, -1.3816]
MileBench	KVFixer − KVEraser	Choice KL	-0.8651	[-0.9940, -0.7447]
MM-NIAH	KVFixer − dirty	Choice KL	-0.6249	[-0.8378, -0.4289]
MM-NIAH	KVFixer − dirty	Full KL	-0.3541	[-0.5387, -0.1833]
MM-NIAH	KVFixer − dirty	Choice agreement	+14.00 pp	[+7.33, +21.00] pp
MM-NIAH	KVFixer − dirty	Accuracy	-14.67 pp	[-22.00, -7.00] pp
MM-NIAH	KVFixer + exact prefix 20% − exact-prefix control 20%	Full KL	-0.0791	[-0.1036, -0.0523]
MM-NIAH	KVFixer + exact prefix 20% − exact-prefix control 20%	Choice KL	-0.5980	[-0.7332, -0.4689]
MM-NIAH	KVFixer − KVEraser	Choice KL	-0.7800	[-0.9059, -0.6553]

F INTERNAL EVALUATION AND DEPLOYED-BANK DIAGNOSTICS

F.1 PROTOCOL AND EXACT-ROW CONTROLS

The internal panel contains 1,200 Text and 1,200 Image sources. Evaluation follows native-PB greedy teacher prefixes to EOS or at most 64 positions; requests are averaged over valid positions and then weighted equally within modality. Paired bootstrap uses 2,000 source draws.

For nominal fraction $\rho > 0$, the exact-row selector chooses $m = \max(1, \min(n_B, \lfloor n_B \rho \rfloor))$ rows. Prefix and uniform controls select the same number of unique rows per request. The operation copies selected K/V rows from exact PB into dirty or repaired states; it is an offline state-replacement diagnostic, not decoder replay.

F.2 COMPLETE INTERNAL SWEEP

Standalone exact-row controls:

Subset / selector	5%	10%	20%	50%
Text / exact prefix	0.0469 / 96.55	0.0442 / 96.57	0.0375 / 96.94	0.0230 / 97.83
Text / exact uniform	0.0466 / 96.47	0.0434 / 96.64	0.0382 / 96.92	0.0331 / 97.38
Image / exact prefix	0.4229 / 87.13	0.4084 / 87.49	0.3921 / 87.81	0.3045 / 89.17

Subset / selector	5%	10%	20%	50%
Image / exact uniform	0.1151 / 95.51	0.1150 / 95.51	0.1118 / 95.44	0.1059 / 95.49

Same-run KVFixer combinations:

Arm	Text KL	Text agree. (%)	Image KL	Image agree. (%)
KVFixer	0.029111	96.9199	0.122648	89.8766
+ exact prefix 5%	0.021514	97.3450	0.112219	89.6512
+ exact prefix 20%	0.017305	97.7190	0.104861	89.8242
+ exact prefix 50%	0.013743	98.1513	0.084665	91.3990
+ exact uniform 5%	0.021333	97.3621	0.112891	89.5335
+ exact uniform 20%	0.021935	97.1430	0.094138	90.6305
+ exact uniform 50%	0.014976	98.0573	0.051503	93.5400

F.3 DEPLOYED-BANK DIAGNOSTICS

For each source, the evaluator records the realized patch $\widetilde{\Delta} = C^{\mathrm{dirty}} - C^{\mathrm{repaired}}$. Define

$$E_{\mathrm{rep}} = \sum \|\Delta - QQ^{\top}\Delta\|^2,$$
$$E_{\mathrm{coord}} = \sum \|Q^{\top}\Delta - Q^{\top}\widetilde{\Delta}\|^2, \tag{F1}$$
$$E_{\mathrm{total}} = \sum \|\Delta - \widetilde{\Delta}\|^2.$$

Table 2 averages each quantity after source-wise normalization by positive E_{Δ}.

The exact-projection reference applies $QQ^{\top}\Delta$ in the same selected bank. Learned-minus-projection KL is:

Panel	Mean difference	95% interval	Median difference	Learned lower (%)
Text	-0.061216	[-0.086402, -0.038676]	+0.002610	10.50
Image	-0.084537	[-0.108782, -0.063113]	+0.005013	45.25

Thus the lower mean KL is driven by larger improvements on a subset of sources rather than per-request dominance. At H=64, mean reached lengths are 29.28 (Text) and 36.48 (Image), with 35.33% and 48.50% of sources reaching the full horizon.

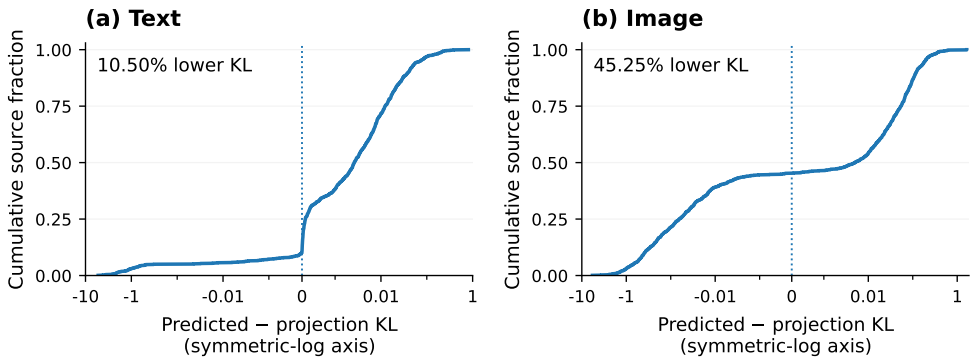

Figure 4: **Source-paired KL difference distributions.** Empirical CDFs of learned-minus-exact-projection H=64 KL. Negative differences favor the learned update.

F.4 Task strata

Differences are learned minus exact-projection H=64 KL. Intervals use paired source draws within each original dataset stratum.

Stratum	N	Difference	95% interval
HotpotQA	600	-0.136431	[-0.184441, -0.090814]
MisBench	600	+0.013999	[0.010314, 0.018374]
AI2D	200	-0.197072	[-0.266915, -0.131629]
ChartQA	700	+0.004718	[-0.024037, 0.028205]
ScienceQA	300	-0.217778	[-0.268811, -0.172839]

F.5 Fixed-endpoint refinement and shared-basis comparisons

Configurations. The refinement pilot uses the 96,000-source expert/refinement pool and a rank-16 coordinate adapter. The selection record assigns one refinement epoch to KV+KL and zero to the KV-only export; the latter is therefore a reference, not an equally trained KV-only adapter. Both exports are evaluated on the complete Text and Image panels. The Shared-Q/fixed-routing diagnostic is evaluated at a fixed epoch-2 endpoint selected on calibration, rather than the primary model's epoch-5 endpoint. These results test specific configurations, not matched-budget component effects.

Evaluation. Each completed modality/export has 1,200 unique source records matching the E1 panel. Recorded basis-group and expert-route decisions, input lengths, PB teacher lengths and per-horizon valid counts agree source-by-source with E1. Reports use H=1/2/4/8/16/32/64, native PB greedy teacher prefixes, EOS stopping and equal source weights. The new runs do not collect KV-error decomposition, so no cache-error change is inferred from their functional scores. Exact-prefix/uniform combinations still copy selected clean rows and are not online replay measurements.

H=64 results. All six modality/export evaluations contain 1,200 sources. Parentheses give recorded refinement epochs; agreement is in percent.

Configuration (epochs)	Update	Text KL	Text agree. (%)	Image KL	Image agree. (%)
KV-only (0)	Standalone	0.082833	95.5402	0.704690	77.0922
KV-only (0)	Exact prefix 20%	0.033927	97.2527	0.145096	89.4971
KV-only (0)	Exact uniform 20%	0.031847	97.1985	0.118314	90.9909
KV+KL (1)	Standalone	0.032285	96.6760	0.133141	89.6395
KV+KL (1)	Exact prefix 20%	0.018578	97.5241	0.105719	91.5274

Configuration (epochs)	Update	Text KL	Text agree. (%)	Image KL	Image agree. (%)
KV+KL (1)	Exact uniform 20%	0.021342	97.3482	0.091675	92.0173
Shared-Q (2)	Standalone	0.031678	96.7778	0.125301	89.9692
Shared-Q (2)	Exact prefix 20%	0.019374	97.6950	0.102331	90.7899
Shared-Q (2)	Exact uniform 20%	0.022880	97.1829	0.092259	91.0907

Source-paired standalone contrasts. Differences are left minus right. Negative KL and positive agreement favor the left. Percentile 95% intervals use 2,000 source-paired draws, seed 20260925, conditional on the fitted endpoints; agreement differences are percentage points.

Contrast (epochs)	Panel	KL difference [95% CI]	Agreement difference [95% CI], pp
KV+KL (1) − KV-only (0)	Text	-0.050548 [-0.064891, -0.038057]	+1.1358 [+0.8592, +1.4610]
KV+KL (1) − KV-only (0)	Image	-0.571549 [-0.618868, -0.525737]	+12.5473 [+11.5741, +13.4765]
Shared-Q (2) − primary (5)	Text	+0.002567 [+0.001133, +0.004043]	-0.1421 [-0.3684, +0.0808]
Shared-Q (2) − primary (5)	Image	+0.002653 [+0.000185, +0.005001]	+0.0926 [-0.3006, +0.5089]

The Text/Image refinement comparisons improve KL on 52.92%/93.00% of sources, with median differences -0.000140/-0.219808 and mean reductions of 61.0%/81.1%. These rates compare refinement with the KV-only export, not with the exact-projection reference in F.3. Shared-Q has slightly higher standalone mean KL than the primary model, but its Image exact-prefix and exact-uniform combination KL is lower (0.102331/0.092259 versus 0.104861/0.094138). These mixed rankings and unequal refinement endpoints prevent attributing the differences solely to the loss term or basis conditioning.

G Algorithmic computation and storage

G.1 Counting boundary

One request starts from the existing PAB cache and returns a PB-shaped repaired cache. The original prefill is shared work; training and target construction are offline. One multiply-accumulate counts as two FLOPs; storage is reported separately.

G.2 Predicted update

For L layers, retained length n_B, K/V width d_{KV} and rank r,

Operation	Leading FLOPs per layer/token
Project dirty K and V	$4d_{KV}r$
Joint coordinate prediction	$16r^2$
Reconstruct K and V patches	$4d_{KV}r$

Hence

$$C_B = Ln_B(8d_{KV}r + 16r^2), \qquad C_A = 4r\sum_{\ell} d_{KV,\ell} n_{A,\ell}^{\mathrm{proj}}. \tag{G1}$$

Descriptors, routing, positional alignment, elementwise operations and necessary visual processing are additional. For $L = 36, d_{KV} = 1024, r = 256$, the B-side core is 113,246,208 FLOPs per retained token and A projection is 37,748,736 FLOPs per projected A row across all layers.

To ground the symbolic count in the evaluated length regime, we substitute the observed retained-token p10/p50/p90 values from the two external panels into the supplied dense-operation calculator. The calculator uses its registered P-empty scope and an analytic A=256 instance; retained visual features are assumed equally reusable. These values are arithmetic counts rather than measured latency.

Panel	B quantile	B tokens	Repair GFLOPs	Exact-PB replay GFLOPs	Repair / replay
MileBench	p10	57	11.297	792.792	1.43%
MileBench	p50	67	12.430	932.076	1.33%
MileBench	p90	89	14.921	1,238.709	1.21%
MM-NIAH	p10	2,146	247.869	31,170.030	0.80%
MM-NIAH	p50	11,384	1,294.037	196,363.845	0.66%
MM-NIAH	p90	45,523	5,160.149	1,243,556.741	0.42%

The ratio decreases with retained length because the decoder replay model includes the transformer's per-token linear/MLP work and growing attention work, whereas KVFixer's leading token-local correction remains linear in n_B. Memory traffic, fused-kernel behavior, communication, normalization/softmax and vision-front-end work are outside this ledger, so the percentages are not wall-clock speedups.

G.3 Decoder recomputation and storage

For hidden/query width d and a gated MLP with three $d \leftrightarrow d_{\mathrm{ff}}$ projections, leading linear work per replayed token is $4d^2 + 4dd_{KV} + 6dd_{\mathrm{ff}}$. With full causal attention from B into cached P and previous B,

$$N_{\mathrm{pairs}} = n_B n_P + \frac{n_B(n_B+1)}{2}, \tag{G2}$$

with score/value work $4LdN_{\mathrm{pairs}}$. Local-attention or masked implementations use their actual executed pair counts. Exact-row controls in Table 1 already materialize PB and are therefore not assigned online replay costs.

The reported repair tensors occupy 598.54 MiB, with a separate 2.318 GiB CPU FP32 embedding copy, excluding the backbone and ordinary KV cache. Low rank describes the correction, not serving-cache compression.