RECEIVER-CONDITIONED STATE RECONSTRUCTION IN HYBRID RECURRENT LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model programs can compute several candidate branches before selecting one to continue. Saving a full internal state for each candidate increases storage and transfer costs. The retained state of one completed branch can instead serve as a reference for compact records of the other computed branches. The selected branch can then be restored without retaining their shared starting state. For fixed affine branch maps, the theory identifies missing information and corrections that limit error amplification. In short Gated DeltaNet branches, decay matching represents the recurrent state as a scaled sibling state plus a compact residual. A stored lossless bitwise repair corrects numerical discrepancies to recover the native state exactly. On Qwen3.5-4B, repaired records reproduce native cache bits, fresh logits and 128-token continuations at all 64 evaluated endpoints in the same runtime. Across 48 short endpoints, median paired byte ratios are 29.1% relative to canonical lossless Full and 30.1% relative to direct sibling exclusive OR (XOR).

1 INTRODUCTION

A branching language-model program can compute several candidates before an external choice determines which one to continue. Saving each candidate's state avoids repeating its computation, but requires storing or transferring several full states. Systems such as SGLang support branching and structured generation directly (Zheng et al., 2024; SGLang Contributors, 2026). If one candidate remains in memory, can it help restore another from a smaller record (Figure 1)?

Hybrid recurrent models combine attention key/value (KV) caches with recurrent matrix states (Yang et al., 2025; Zuo et al., 2025; Qwen Team, 2026). Their prefix caches must therefore manage recurrent checkpoints alongside token matches (Pan et al., 2025). The candidates share a starting state, called the *fork*, and form *sibling* branches. Retaining the fork allows either replaying a branch or reconstructing its state from a factor record. Once the fork is discarded, a completed sibling still contains shared history, but mixed with its own updates. Recovery must separate that shared history from the sibling's branch-specific changes.

The selected branch to restore is the *source*; the retained sibling is the *target*. The receiver holds the target but not the fork, and the sender encodes the already computed source relative to that target. Their decay-matched difference can be compact even when both endpoint states are dense. In Gated DeltaNet, matching cumulative decay cancels shared history outside the space spanned by both branches' keys. The source becomes a scaled target plus a residual,

$$\boxed{S_s = \lambda S_t + \Delta_{s|t}, \qquad \Delta_{s|t} = UV^\top.} \tag{1}$$

The residual's rank is at most the combined branch length. Direct factors collect the executed updates; shared coordinates express the same residual in a common basis. Neither construction needs to invert the target's transition matrix. A lossless repair record can also correct numerical discrepancies, recovering the exact floating-point state produced by the original execution.

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.

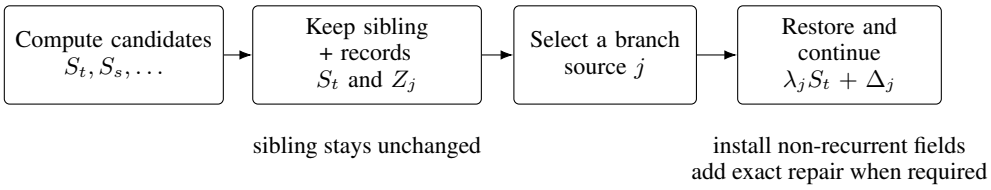

Figure 1: Sibling-relative branch recovery.

2 Receiver Information and Reconstruction

Recovery requires sufficient information and a stable way to use it. For one recurrent head, $X, S_s, S_t \in \mathbb{R}^{d_k \times d_v}$ are the fork, source and target states, with key/value dimensions d_k, d_v. Branch $b \in \{s, t\}$ executes n_b affine updates. Its net transition P_b and accumulated write Q_b give

$$S_{b,i} = A_{b,i} S_{b,i-1} + B_{b,i}, \qquad S_b = P_b X + Q_b. \tag{2}$$

The transitions P_s, P_t and offsets Q_s, Q_t are fixed and known from a completed execution. A new neural execution from a different base can produce different coefficients. A *shared-base* receiver retains X; a *sibling-only* receiver retains S_t and receives the additional information needed for recovery. Experimental records also call the latter *target-only*. Appendix B gives the full proofs.

2.1 Information sufficiency

Two fork states can produce the same target while producing different sources. Recovery from the target alone requires it to distinguish every base direction that matters to the source.

Theorem 1 (Target sufficiency and minimum supplementation)**.** *For fixed P_s, P_t, Q_s, Q_t, the target determines the source for every X if and only if*

$$\ker P_t \subseteq \ker P_s. \tag{3}$$

For supplementary observations FX, with F fixed independently of X, the minimum number of supplementary rows is

$$r_{\text{miss}} = \operatorname{rank} \begin{bmatrix} P_t \\ P_s \end{bmatrix} - \operatorname{rank}(P_t) = \operatorname{rank}\big(P_s(I - P_t^{\dagger} P_t)\big). \tag{4}$$

Proof idea. A base direction erased by P_t must also be irrelevant to P_s. Otherwise, it must be supplied. A basis for the source row-space component absent from the target supplies exactly the missing directions, attaining (4). □

This static recovery problem is conceptually related to functional observability (Fernando et al., 2010). For $P_s = I_2$ and $P_t = \operatorname{diag}(1, 0)$, the target retains the first base row but erases the second. Sending the second row supplies exactly the missing information.

2.2 Stable supplementation

An invertible target map preserves all base information, but recovering it may amplify error. The direct reconstruction operator is $M = P_s P_t^{-1}$. Replacing it with a bounded-gain operator A limits amplification, while a supplement $C_A X$, where $C_A = P_s - A P_t$, supplies the remainder.

Theorem 2 (Minimum correction rank at bounded gain)**.** *For $\gamma \geq 0$ and invertible P_t,*

$$r_\gamma := \min_{\|A\|_2 \leq \gamma} \operatorname{rank}(P_s - A P_t) = \#\{i : \sigma_i(M) > \gamma\}. \tag{5}$$

If $M = U_M \operatorname{diag}(\sigma_i) V_M^{\top}$, the clipped operator $A_\gamma = U_M \operatorname{diag}(\min(\sigma_i, \gamma)) V_M^{\top}$ attains the minimum.

Proof. Since $C_A = (M - A)P_t$ and P_t is invertible, $\operatorname{rank}(C_A) = \operatorname{rank}(M - A)$. If this rank is k, the spectral low-rank approximation bound gives $\|A\|_2 \geq \sigma_{k+1}(M)$ (Eckart & Young, 1936; Mirsky, 1960). Every singular value above γ requires a correction direction. Clipping exactly those singular values attains the bound. □

Factor $C_A = U_c V_c^\top$ with $U_c^\top U_c = I$, and supply $G = V_c^\top X$. Then

$$S_s = A(S_t - Q_t) + U_c G + Q_s, \qquad \left\| \widehat{S}_s - S_s \right\|_F \leq \gamma \|E_t\|_F + \|E_G\|_F . \tag{6}$$

Here E_t and E_G perturb the target and supplement; the operator and offsets remain fixed. For $P_t = \operatorname{diag}(1, \varepsilon)$ and $P_s = I_2$, the target preserves both directions, but recovering the second amplifies error by $1/\varepsilon$. Supplying that direction separately bounds this amplification. Thus r_{miss} counts lost information, whereas r_γ counts directions that need protection from amplification. Record size also depends on how the operator, affine writes and other cache fields are encoded. For Gated DeltaNet, the executed updates give a factor representation of the full residual.

3 Residual Encodings for Gated DeltaNet

Gated DeltaNet rescales each head and writes along key k_i, using value v_i and update strength β_i (Yang et al., 2025). In the transposed state convention, step i gives

$$S_i = \alpha_i (I - \beta_i k_i k_i^\top) S_{i-1} + \beta_i k_i v_i^\top . \tag{7}$$

Here α_i is scalar per head, as in the evaluated Qwen3.5-4B Gated DeltaNet layers. Decay matching then cancels shared history outside the joint key span. Channel-wise decay requires a different complement-space analysis (Appendix C).

3.1 Direct factor difference

Expanding (7) gives $S_i = \alpha_i S_{i-1} + k_i r_i^\top$, where

$$r_i = \beta_i (v_i - \alpha_i S_{i-1}^\top k_i). \tag{8}$$

The innovation r_i is the write vector computed from the state encountered during execution. Cumulative decay is $D_b = \prod_i \alpha_{b,i}$, and each update is weighted by the subsequent decay $\omega_{b,i} = \prod_{j=i+1}^{n_b} \alpha_{b,j}$. Collecting these weighted updates separates the decayed fork from the branch's own writes,

$$U_b = [k_{b,i}]_{i=1}^{n_b}, \quad V_b = [\omega_{b,i} r_{b,i}]_{i=1}^{n_b}, \qquad S_b = D_b X + U_b V_b^\top . \tag{9}$$

For $D_t \neq 0$, set $\lambda = D_s / D_t$. Subtracting the scaled target cancels X, giving the Direct Factor Difference (DFD) record,

$$S_s = \lambda S_t + U V^\top, \qquad U = [U_s, U_t], \quad V = [V_s, -\lambda V_t]. \tag{10}$$

The sender transmits λ, U, V, and the receiver adds the residual to its scaled target. The factor width is $p = n_s + n_t$, independent of the rank of X or either endpoint. DFD requires neither a target-transition inverse nor an explicit joint basis.

For example, suppose two-dimensional keys all lie along the first coordinate. The second row of each branch is then $D_b X_{2,:}$. Decay matching cancels that row, so the residual needs only the first-coordinate direction. The same cancellation holds outside the joint key span in higher dimensions.

3.2 Shared-subspace coordinates

The keys from both branches may span fewer directions than the direct factors contain. Let $W \in \mathbb{R}^{d_k \times m}$ be an orthonormal basis for this joint key span and $\Pi_\perp = I - W W^\top$. Scalar Subspace Residual (SSR) uses W as a shared coordinate system. Its innovation-based construction, SSR-I, projects the direct factors into those coordinates,

$$\Delta_{s|t} = U V^\top, \qquad R_I = (W^\top U) V^\top, \qquad U V^\top = W R_I . \tag{11}$$

The coefficient-based construction, SSR-C, instead tracks how each branch acts within the shared coordinates. The matrix T_b describes the transition after removing cumulative scalar decay, and J_b collects the affine writes. Outside this subspace, only the scalar decay remains, giving

$$P_b = D_b(\Pi_\perp + WT_bW^\top), \qquad Q_b = WJ_b, \qquad m \le \min(d_k, n_s + n_t). \tag{12}$$

Appendix C gives the coefficient recurrences and the scope beyond scalar-decay updates.

Proposition 3 (Sibling-relative subspace residual)**.** *Under* (12)*, suppose* $D_t \neq 0$ *and define* $\lambda = D_s/D_t$*. Then*

$$\boxed{S_s = \lambda S_t + WR_C, \qquad R_C = J_s - \lambda J_t + D_s(T_s - T_t)W^\top X.} \tag{13}$$

The identity holds for singular or invertible T_t*. With a fixed exact basis, target perturbation* E_t*, and residual perturbation* E_R*,*

$$\left\|\widehat{S}_s - S_s\right\|_F \le |\lambda| \, \|E_t\|_F + \|E_R\|_F \, . \tag{14}$$

Proof. Subtract λS_t from S_s in (12). The complementary base terms cancel because $D_s = \lambda D_t$, leaving WR_C. The perturbation is $\lambda E_t + WE_R$. Multiplication by W preserves the Frobenius norm because its columns are orthonormal. □

The sender computes SSR-C from the reduced branch maps and $W^\top X$. The receiver needs only λ, W, R_C and its target, not the fork. With an exact basis and the same executed coefficients, $R_I = R_C$ in exact arithmetic. Both constructions avoid inverting T_t, and the gate $|\lambda| \le \gamma$ bounds target amplification.

Table 1: Three ways to encode the branch difference.

Method	Inputs used to build the record	Transmitted recurrent fields
DFD	Executed innovations with suffix-decay weights	λ, U, V
SSR-I	The same U, V and joint basis W	$\lambda, W, (W^\top U)V^\top$
SSR-C	Aligned D_b, T_b, J_b, W, and $W^\top X$	λ, W, R_C from (13)

DFD: Direct Factor Difference. SSR: Scalar Subspace Residual, with innovation-based (I) or coefficient-based (C) construction. All paths carry the same source non-recurrent fields.

For a compact head with $p = n_s + n_t$, equal-precision records contain

$$N_{\mathrm{DFD}} = 1 + p(d_k + d_v), \qquad N_{\mathrm{SSR}} = 1 + m(d_k + d_v). \tag{15}$$

DFD applies the residual in $O(d_k p d_v)$ arithmetic; SSR uses $O(d_k m d_v)$, with additional basis construction and aggregation at the sender. When $m = p$, the two representations contain the same number of values. Shared coordinates reduce this count only when $m < p$.

3.3 Operator absorption

Shared coordinates also remove an intermediate matrix from the Key-Subspace Operator (KSO) construction. For invertible T_t and $|\lambda| \le \gamma$, KSO clips $\lambda T_s T_t^{-1}$ to A_W and supplies an affine correction H. This reconstructs the source as

$$S_s = \lambda S_t + W[(A_W - \lambda I_m)W^\top S_t + H]. \tag{16}$$

Substituting $W^\top S_t = D_t T_t W^\top X + J_t$ gives

$$(A_W - \lambda I_m)W^\top S_t + H = R_C. \tag{17}$$

Computing R_C directly removes the solve, spectral clipping, m^2 operator values and receiver target projection. The cancellation holds wherever KSO is defined, while SSR-C also covers singular T_t. Section 5 measures the savings; Appendix C gives H and the full cancellation.

3.4 HEADWISE RECOVERY

Each recurrent head is stored as either a compact record or a dense source matrix. The sender checks the decay ratio and construction-specific conditions; invalid ratios, including $D_t = 0$, select the dense path. The complete record also carries the source attention suffix, convolution buffers, positions and model/state binding. The receiver restores all fields at the same token boundary in separate storage, leaving the target unchanged. Algorithm 1 gives the procedure; Appendix D separates mapping, encoding and application errors against the native source.

Algorithm 1 Build and restore a sibling-conditioned hybrid endpoint

Require: Executed branches, path j, fixed target binding, headwise numerical policy
Sender
for each recurrent head h **do**
  Reject invalid decay, $D_{t,h} = 0$, or a non-finite ratio; otherwise form $\lambda_h = D_{s,h}/D_{t,h}$
  For valid ratios, evaluate gain and path-specific admissibility checks
  if the head is admissible **then**
    Encode (λ_h, U_h, V_h) for DFD, or (λ_h, W_h, R_h) for SSR
  else
    Encode the dense source head and a dense-field tag
  end if
end for
Encode source KV, convolution, positions, and source/target binding; transmit record
Receiver
Validate the record against the immutable resident target
for each recurrent head h **do**
  Write $\lambda_h S_{t,h} + U_h V_h^\top$, $\lambda_h S_{t,h} + W_h R_h$, or the dense field
end for
Install non-recurrent fields in independent source storage; continue from the source boundary

3.5 FROM A RESIDUAL PREDICTION TO NATIVE-BIT RECOVERY

Finite-precision reconstruction can differ from the native source state, even when the algebraic identity is exact. DFD with headwise dense fallback is denoted DFD-H. Its decoded state $\widehat{S}_s$ serves as a prediction; a separate lossless record corrects the remaining bit differences. Let $\mathcal{B}$ map 32-bit floating-point (FP32) recurrent tensors to raw bit strings in a fixed layout. The repair and its inverse are

$$E_{\mathrm{bit}} = \mathcal{B}(S_s) \oplus \mathcal{B}(\widehat{S}_s), \qquad \mathcal{B}(S_s) = \mathcal{B}(\widehat{S}_s) \oplus E_{\mathrm{bit}}. \tag{18}$$

Here $\oplus$ is bitwise exclusive OR (XOR). The sender losslessly compresses E_{bit} after a byte-plane transform. The receiver reconstructs the prediction from the transmitted DFD-H record, verifies the predictor hash, and reverses the XOR before installing the source's non-recurrent fields. Exactness requires the sender and receiver to reconstruct identical predictor bits. Section 4.2 measures all record bytes and local costs.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

The main experiments use Qwen3.5-4B on one RTX 5080. Full stores recurrent fields in FP32; DFD-H uses FP32 factors and gain threshold $\gamma = 2$. Unless labeled legacy, DFD-H uses the corrected capture described in Section 4.3. Complete records include source non-recurrent suffix fields and framing under the same prefix-reuse policy.

The exact-recovery and precision studies evaluate the same 64 endpoints. The 48 short endpoints span lengths 1, 2, 3 and 4 in 12 prompt clusters. The 16 boundary endpoints span lengths 8, 16, 32 and 64 in four overlapping clusters. The exact-recovery protocol was frozen after the precision results were known; it tests the same cohort under a stricter fidelity requirement.

Each endpoint and method has two warm-ups and five formal repeats. Timing comparisons use the median within each endpoint, followed by method-paired ratios summarized across endpoints.

For the short cohort, 10,000 whole-prompt-cluster resamples give pointwise 95% percentile intervals; boundary lengths are descriptive. State error is the relative Frobenius error of recurrent fields against the native source. The trajectory metric A_{128} counts exact agreement through 128 greedy predictions, not task correctness.

4.2 Compression at native-bit fidelity

Does predicting the source from a sibling still save bytes when the restored state must be bit-for-bit exact? Table 2 compares repaired DFD-H with canonical lossless Full, direct sibling XOR, zlib-compressed Full packets and raw Full packets. Canonical Full and both difference encodings use the same fixed byte-plane and zlib codec. Direct sibling XOR uses the unscaled sibling as its prediction, testing whether captured updates improve on a simple state difference.

All five methods reproduced native cache bits, fresh logits and 128 continuation tokens at all 64 endpoints in the measured runtime. Across the 48 short endpoints, repaired DFD-H had a median paired byte ratio of 0.291 to canonical lossless Full, with interval $[0.280, 0.324]$. The ratio to direct sibling XOR was 0.301, with interval $[0.288, 0.332]$. Repaired DFD-H was smaller than both baselines at every short endpoint, while restoring the same native state.

Table 2: Record size and cost with exact recovery.

Method	Record (MB) ↓	Local time (ms) ↓	A_{128}
Canonical lossless Full	44.523	2250.1	48/48
Zlib Full packet	48.883	1840.7	48/48
Raw Full packet	52.239	347.2	48/48
Exact sibling XOR	43.442	2285.1	48/48
DFD-H plus exact repair	**12.972**	2330.0	48/48

48 short endpoints; medians of five-repeat endpoint medians. All methods recover native cache bits and fresh logits. A_{128} counts exact 128-token continuations. DFD-H: Direct Factor Difference with headwise dense fallback; XOR: exclusive OR; MB: 10^6 bytes. Local time uses the lifecycle in Section 4.2.

Local timing includes required execution, capture, copies, coding, prediction, repair, restoration and fresh forward. Full executes only the source; sibling methods also execute a sender target. The clock assumes a supplied sender fork and receiver sibling, excluding their setup, network transfer and continuation. Appendix S also reports costs with receiver-target setup added to every method.

The smaller records required more local work. Paired time ratios were 1.044 to canonical lossless Full, 1.274 to zlib Full and 6.638 to raw Full. Against sibling XOR, the ratio was 1.035 with interval $[0.973, 1.080]$, so neither method had a clear timing advantage.

4.3 Capture accuracy without exact repair

Capturing executed updates more accurately improved the state prediction without adding bytes. The legacy path normalized raw keys in bfloat16 (BF16) when reconstructing innovations. Normalizing in FP32 and consistently recomputing the associated read and innovation reduced median state error from 1.53×10^{-3} to 1.30×10^{-7}. The model, native kernel, residual formula, gate, packet bytes and audited compact masks were unchanged (Appendix Q.4).

Without exact repair, corrected DFD-H used 14.0% of Full bytes and 27.0% of Dense BF16 bytes (Table 3). Dense FP16 and Dense BF16 transmit recurrent fields in 16-bit floating point and bfloat16, respectively, then restore them to FP32. DFD-H was smaller and had lower state error than both at every short endpoint. Despite its lower state error, it matched 25/48 complete continuations, compared with 29/48 for Dense FP16 and 26/48 for Dense BF16.

Unlike the exact-recovery protocol, this study charged both branch executions to every method. Timing also included required capture, construction, codecs, restoration and fresh forward; it excluded prefix execution, artifact I/O, audit hashes, transfer and continuation. The paired request-time ratio was 1.559 to Full, with interval $[1.487, 1.591]$, and 1.972 to Dense BF16, with interval $[1.891, 2.033]$.

Table 3: Record size and fidelity before exact repair.

Method	Record (MB) ↓	Request (ms) ↓	State error ↓	A_{128}
Full FP32	52.24	407.5	0	48/48
Dense FP16	27.08	324.7	2.09×10^{-4}	29/48
Dense BF16	27.09	329.5	1.67×10^{-3}	26/48
DFD-H (legacy)	7.32	635.4	1.53×10^{-3}	17/48
DFD-H (corrected)	7.32	612.5	1.30×10^{-7}	25/48

48 endpoints in 12 prompt clusters. Entries are medians; request times first use five-repeat medians within each endpoint. State error is recurrent relative L_2 error; A_{128} counts exact 128-token continuations. FP32/FP16: 32-/16-bit floating point; BF16: bfloat16; DFD-H: Direct Factor Difference with headwise dense fallback. Corrected capture normalizes keys in FP32. This request clock is defined in Section 4.3; MB: 10^6 bytes.

4.4 Effect of branch length

Residual size grows with branch length because each update contributes a factor column. For equal-length branches with 128-dimensional heads, DFD stores $1 + 512n$ values per compact head, compared with 16,384 dense values. In the earlier three-encoding short suite, SSR has $m = p = 2n$ and provides no further width reduction. An offline rank oracle likewise finds little short-branch factor redundancy at tight tolerances (Appendix Q.3).

At equal precision, a subspace record contains fewer recurrent values than a dense head when

$$1 + m(d_k + d_v) < d_k d_v. \tag{19}$$

With $m = p = 2n$, the equal-precision count favors residuals for $n < 31.998$. Complete records also include fallback, non-recurrent fields and framing. Approximate DFD-H is smaller than Full on 8/16 boundary endpoints and smaller than both low-precision baselines on 3/16. With exact repair, all four endpoints at each length 8 and 16 use fewer bytes than every lossless baseline; none do at lengths 32 and 64 (Appendices Q.2 and S).

4.5 Continuation after releasing the source

Approximate recovery was tested after discarding the source state, using constructed short candidate pairs on 17 previously exposed ALFWorld development tasks (Shridhar et al., 2021). A fixed rule selected each source before its runtime handle was released. Full restored an independent snapshot; corrected DFD-H restored from the sibling. Each method then continued in a separate environment instance, adding actual observations to its cache without replaying the source prompt.

Table 4: Continuing after the source state is released.

Method	Task success	Same outcome	Exact trajectory (actions + tokens)	State error (relative L_2)
Native	10/17	17/17	17/17	–
Full FP32	10/17	17/17	17/17	0
DFD-H (corrected)	10/17	17/17	13/17	1.67×10^{-7}
Non-recurrent-only	10/17	15/17	7/17	1.31×10^{-1}

17 development tasks. Outcome and trajectory matches are relative to Full. Native retains the source; non-recurrent-only installs source non-recurrent fields but keeps the sibling recurrent state. State error compares the restored recurrent state with Full. FP32: 32-bit floating point; DFD-H: Direct Factor Difference with headwise dense fallback.

Full reproduced uninterrupted Native execution on all 17 tasks, including complete trajectories and terminal caches. DFD-H preserved Full's success/failure outcome on every task and its action-and-token sequence on 13/17 (Table 4). A non-recurrent-only control instead installed source KV, convolution and position fields while retaining the sibling recurrent state. It matched seven complete trajectories and changed two outcomes, one in each direction. Recurrent-state recovery thus preserved more trajectories, although all four methods had ten environment-reported successes.

A post-hoc check found that one environment-scored success did not verify an additional placement action. Appendix Q.5 gives the case details, paired trajectory counts and selection procedure.

Controlling attention-prefix reuse. An offline comparison re-encoded dense baselines with the same common-KV-prefix reuse as DFD. DFD remained smaller and lower-error than both low-precision baselines on all 17 cases, with a paired byte ratio of 0.409 to Dense BF16. Dense paths require only the prefix, whereas DFD requires the complete sibling (Appendix Q.6).

Earlier register and route tests provide additional non-recurrent-only controls (Appendices F, G and N).

5 Construction Ablations

Matrix absorption removes work before encoding. In the earlier matched-basis ablation, SSR-C reduced KSO packet bytes by 3.8% and construction time by 82.2%, measured with operands already available. The complete-request ratio was 0.580, with a 95% prompt-cluster interval of $[0.573, 0.588]$. Removing the solve, clipping and target projection saved more construction work than bytes (Appendix J).

A common head set isolates re-encoding. Within each of 12 conditions, DFD and both SSR methods used the same compact heads. SSR packets were less than 0.14% larger than DFD, and all fresh top-1 readouts matched. The two 100%-compact conditions also had similar sizes, ruling out shared dense fallback as the sole explanation (Appendix J).

6 Receiver Conditions and Recovery Cost

6.1 Recovery from a resident sibling

In a two-candidate, 2-token workload, the receiver retained candidate A and restored candidate B. DFD-H used A directly; Source-Factor reconstructed $D_s X + U_s V_s^\top$ and needed the missing fork X. These earlier studies used legacy capture and paced Transmission Control Protocol (TCP) loopback, separately from the exact-recovery timings.

Setup-inclusive cost sums measured prefix and branch execution, required capture, preparation, base delivery, packet transfer, decoding, restoration and fresh forward. Full omitted capture and had the same preparation and reuse opportunities. Each condition had two warm-ups and five formal repeats. This calibrated sum of measured components is distinct from whole-wall deployment latency.

Table 5: Recovery cost under two receiver conditions.

	Cost (ms)			Paired ratio to Full	
Rate	Full	DFD-H	SF	DFD-H	SF
Resident sibling					
0.5 Gbit/s	1786.4	1451.8	2089.1	0.813	1.169
2 Gbit/s	1167.2	1370.5	1346.1	1.168	1.158
Unpaced	1131.7	1370.6	1338.1	1.201	1.174
Cold reference ($N = 2$)					
0.5 Gbit/s	2987.7	2545.2	2528.3	0.853	0.847
2 Gbit/s	1698.1	1859.1	1852.5	1.105	1.096
Unpaced	1644.3	1748.4	1809.9	1.068	1.097

One workload, five technical repeats and one request round ($q = 1$). Costs are median measured component sums; ratios are paired within repeats. Resident sibling restores B with A retained. Cold reference delivers a base for both candidates. DFD-H: Direct Factor Difference with headwise dense fallback; SF: Source-Factor, which additionally requires the fork in the resident-sibling case.

At 0.5 Gbit/s, DFD-H cost 18.7% less than Full, whereas Source-Factor cost 16.9% more after receiving its missing fork (Table 5). Retaining both references increased base storage from 55.181 to 110.297 MB. Without pacing, DFD-H cost 1.201 times Full.

6.2 When the reference can be chosen freely

Under *cold free choice*, neither reference is resident, so each layout delivers its base for both candidates. At 0.5 Gbit/s, DFD-H and Source-Factor cost 0.853 and 0.847 times the Full layout; both exceeded Full at higher tested rates.

A sibling layout retains one candidate and $N - 1$ residuals; a fork layout retains the base and N source-factor records. At both tested lengths (2 and 8), Source-Factor used fewer packet bytes for 4 and 8 candidates. For eight 2-token candidates, Source-Factor packets totaled 29.554 MB, compared with 42.558 MB for the sibling layout. Nearly equal base sizes preserved the fork layout's storage advantage (Appendix P).

6.3 Transfer cost on inventory tasks

On eight inventory endpoints, SSR-C-H had a component-cost ratio to Full of 0.799 at 0.5 Gbit/s, with interval $[0.684, 0.936]$. Without pacing, the ratio rose to 1.036; all continuations matched in both settings. This study charged local anchor setup but excluded its initial transmission (Appendix H).

Low-precision dense storage narrowed the benefit. Across 12 endpoints at 0.5 Gbit/s, legacy DFD-H had a task-cost ratio of 1.086 to Dense BF16, with only two faster endpoints. A bulk-codec variant gave 1.064, with five faster endpoints (Appendix R).

Appendix R summarizes the evidence across studies.

7 Related Work

SGLang exposes structured generation and branching (Zheng et al., 2024), while Marconi chooses and reuses hybrid-model prefix checkpoints (Pan et al., 2025). DASC stores selected state units in compressed checkpoints, with optional suffix refresh (Yu et al., 2026). Tail-Replay caches full-attention KV and output hidden states, then approximates recurrent state by replaying a recent suffix of those hidden states (Liu et al., 2026b). Sibling-relative recovery instead starts from a completed sibling and a residual record bound to that sibling.

Executed updates also provide compact state representations. DeltaNet and Gated DeltaNet use rank-one updates and compact transitions (Yang et al., 2024; 2025). SpecLA stores verification factors for accepted-state recovery (Wang et al., 2026), and HYPIC composes segment transitions and zero-start states (Liu et al., 2026a). Here, executed factors or shared coordinates encode the difference between two branches from a common fork.

8 Conclusion

A resident sibling can support recovery of another completed branch without the common fork. Decay matching gives a compact residual, encoded as direct factors or shared coordinates without a target-transition inverse. On tested short branches, repaired DFD-H reproduced native execution with fewer bytes than canonical lossless Full and direct sibling XOR. The method trades local computation for smaller records, with branch length and reference residency determining its useful range.

AI USE STATEMENT

We used generative AI to aid or polish writing, for retrieval and discovery of related work, and for research ideation or execution.

REPRODUCIBILITY STATEMENT

The source package contains proofs, historical experiment records, runtime snapshots and CPU-only table-regeneration scripts. The exact-recovery bundle also reproduces paired statistics and token comparisons from exported timing and continuation records. Appendices E, I, I.1, Q and S document measurement boundaries and provenance. Rebuilding tables and statistics is distinct from rerunning model experiments; this package does not certify a fresh-environment model reproduction.

REFERENCES

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

Tyrone Lucius Fernando, Hieu Minh Trinh, and Les Jennings. Functional observability and the design of minimum order linear functional observers. *IEEE Transactions on Automatic Control*, 55(5):1268–1273, 2010. doi: 10.1109/TAC.2010.2042761.

Ali Hatamizadeh, Yejin Choi, and Jan Kautz. Gated DeltaNet-2: Decoupling erase and write in linear attention. arXiv preprint arXiv:2605.22791, 2026. URL https://arxiv.org/abs/2605.22791.

Yifei Liu, Juntong Wu, Yang Liu, Junhao Hu, Minghao Li, Xiaoxu Chen, and Weihang Chen. HYPIC: Accelerating hybrid-attention LLM serving with position-independent caching. arXiv preprint arXiv:2607.01299, 2026a. URL https://arxiv.org/abs/2607.01299.

Yirui Liu, Ruoling Qi, Xuaner Wu, Penghang Liu, and Jian Chen. Tail-Replay: Escaping the curse of linear attention in prefix caching for hybrid LLMs. arXiv preprint arXiv:2608.30310, 2026b. URL https://arxiv.org/abs/2608.30310.

L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960. doi: 10.1093/qmath/11.1.50.

Rui Pan, Zhuang Wang, Zhen Jia, Can Karakus, Luca Zancato, Tri Dao, Yida Wang, and Ravi Netravali. Marconi: Prefix caching for the era of hybrid LLMs. In *Proceedings of Machine Learning and Systems*, volume 7, 2025. URL https://proceedings.mlsys.org/paper_files/paper/2025/hash/7c180af017258d239bac6248d1eb26ac-Abstract-Conference.html.

Qwen Team. Qwen3.5-4B model card. Hugging Face model repository, 2026. URL https://huggingface.co/Qwen/Qwen3.5-4B. Accessed September 17, 2026.

SGLang Contributors. SGLang frontend language. Official documentation, 2026. URL https://docs.sglang.io/docs/references/frontend/frontend_tutorial. Accessed September 19, 2026.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03768.

Zhibin Wang, Xuying Han, Zhaohua Yang, Fuliang Liu, Xue Li, Rong Gu, Sheng Zhong, and Chen Tian. SpecLA: Efficient speculative decoding for linear-attention models. arXiv preprint arXiv:2607.16673, 2026. URL https://arxiv.org/abs/2607.16673.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d13a3eae72366e61dfdc7eea82eeb685-Abstract-Conference.html.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.06464.

Yanqi Yu, Pingwei Sun, Jianchao Tan, Tao Zhang, Yuchen Xie, Xunliang Cai, and Yao Liu. DASC: Decay-aware state compression for hybrid linear-attention serving. arXiv preprint arXiv:2608.30386, 2026. URL https://arxiv.org/abs/2608.30386.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62557–62583, 2024. doi: 10.52202/079017-2000. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/724be4472168f31ba1c9ac630f15dec8-Abstract-Conference.html.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.971. URL https://aclanthology.org/2023.emnlp-main.971/.

Jingwei Zuo, Maksim Velikanov, Ilyas Chahed, Younes Belkada, Dhia Eddine Rhayem, Guillaume Kunsch, Hakim Hacid, Hamza Yous, Brahim Farhat, Ibrahim Khadraoui, Mugariya Farooq, Giulia Campesan, Ruxandra Cojocaru, Yasser Djilali, Shi Hu, Iheb Chaabane, Puneesh Khanna, Mohamed El Amine Seddik, Ngoc Dung Huynh, Phuc Le Khac, Leen AlQadi, Billel Mokeddem, Mohamed Chami, Abdalgader Abubaker, Mikhail Lubinets, Kacper Piskorski, and Slim Frikha. Falcon-H1: A family of hybrid-head language models redefining efficiency and performance. arXiv preprint arXiv:2507.22448, 2025. URL https://arxiv.org/abs/2507.22448.

A Notation and Table Abbreviations

Table 6 collects the method names, metrics, and abbreviated labels used throughout the tables.

Table 6: Notation and abbreviations used in the manuscript tables.

Label	Meaning
Full / FULL	Dense source-state reference under the specified receiver protocol
DFD	Direct Factor Difference
SSR	Scalar Subspace Residual
SSR-I / SSR-C	Innovation-based / coefficient-based SSR construction
H / E	Headwise / endpoint-wide dense fallback, appended to a method name
KSO	Key-Subspace Operator
SF	Source-Factor representation
KV / conv	Key/value attention cache / convolution buffers
JSON	JavaScript Object Notation, the structured task-output format
GDN	Gated DeltaNet
n_s, n_t; n	Source and target branch lengths; common length for equal-length branches
p; m	Concatenated factor width $n_s + n_t$; joint key-subspace dimension
A_h	Exact agreement with Full through h predicted tokens, including earlier termination; shown as a count or percentage
Coverage	Fraction of recurrent head instances stored compactly, weighted by dense-head bytes when applicable
Request	Complete local request under the timing boundary specified for that study
Source / Target	Source / target branch execution and capture in stage-time tables
Build / Fresh	Recorded descriptor construction / fresh forward readout
State error / rel. L_2	Recurrent-state relative Frobenius error against the matched Full state
Mismatch / First step	Endpoints with a top-1 mismatch / first such step under the shared Full history
a; $\|\Delta z\|_\infty$	Full top-1 logit margin; maximum absolute logit perturbation
Beh. Pareto	Fraction of endpoints on the recorded behavior-cost Pareto frontier
CI / IQR	Confidence interval / interquartile range
EOS / P95	End-of-sequence token / 95th percentile
MB / MiB / ms	10^6 bytes / 2^{20} bytes / milliseconds
Gbit/s / Mbit/s	10^9 / 10^6 bits per second
FP32 / FP64 / BF16	32-bit / 64-bit floating point / bfloat16
INT8 / P2P / TCP	8-bit integer / peer-to-peer / Transmission Control Protocol
Config. / n/a	Model configuration / not applicable

B Complete Reconstruction Arguments

B.1 Linear observations of an executed branch

For one recurrent head, $X, S_s, S_t \in \mathbb{R}^{d_k \times d_v}$, the branch transitions have shape $d_k \times d_k$, and the offsets have shape $d_k \times d_v$. Theorem 1 holds these transitions and offsets fixed while varying X. Subtracting the known offsets reduces the observations to $Y_t = P_t X$ and $Y_s = P_s X$. The argument applies independently to every value column. Matrix 2-norms below are induced spectral norms, and state errors use the Frobenius norm.

In the model, branch coefficients are obtained from actual execution. A new execution with a different base can produce different coefficients through preceding layers. Conditioning on the realized maps specifies the linear problem; native arithmetic is related to this problem through the residuals in Appendix D.

B.2 Target sufficiency and the minimum supplement

If $P_t u = 0$ and $P_s u \neq 0$, then X and $X + uw^\top$ are indistinguishable at the target for any nonzero $w \in \mathbb{R}^{d_v}$, while their sources differ. Thus target sufficiency requires $\ker P_t \subseteq \ker P_s$. Conversely, this condition is equivalent to $\mathrm{row}(P_s) \subseteq \mathrm{row}(P_t)$. With the orthogonal row-space projector $R_t =$

$P_t^\dagger P_t$,

$$P_s = P_s R_t, \qquad S_s = P_s P_t^\dagger (S_t - Q_t) + Q_s. \tag{20}$$

This proves sufficiency.

For supplementation define $C_0 = P_s(I - R_t)$. The row space of C_0 is the orthogonal projection of the source row space outside the target row space. The dimension formula for the sum of the two row spaces gives

$$\operatorname{rank}(C_0) = \operatorname{rank}\begin{bmatrix} P_t \\ P_s \end{bmatrix} - \operatorname{rank}(P_t). \tag{21}$$

If FX contains r supplementary rows, recovery for all X requires

$$\operatorname{row}(P_s) \subseteq \operatorname{row}\begin{bmatrix} P_t \\ F \end{bmatrix}. \tag{22}$$

Adding r rows increases the observation row-space dimension by at most r, proving the lower bound. For achievability, factor $C_0 = U_c V_c^\top$ with r_{miss} columns. Sending $G = V_c^\top X$ yields

$$S_s = P_s P_t^\dagger (S_t - Q_t) + U_c G + Q_s. \tag{23}$$

The supplement contains $r_{\text{miss}} d_v$ values when the operator and offset descriptions are already supplied.

B.3 Correction rank at bounded gain

For invertible P_t, let $M = P_s P_t^{-1}$. Right multiplication by P_t preserves rank, hence

$$\operatorname{rank}(P_s - A P_t) = \operatorname{rank}(M - A). \tag{24}$$

If $\operatorname{rank}(M - A) = k$, the spectral low-rank approximation bound implies $\|A\|_2 \geq \sigma_{k+1}(M)$. Therefore a feasible A with $\|A\|_2 \leq \gamma$ must modify every singular direction whose singular value exceeds γ. For $M = U_M \operatorname{diag}(\sigma_i) V_M^\top$, choosing

$$A_\gamma = U_M \operatorname{diag}(\min(\sigma_i, \gamma)) V_M^\top \tag{25}$$

achieves exactly that rank and has spectral norm at most γ. This proves Theorem 2.

Let $P_s - A_\gamma P_t = U_c V_c^\top$ with $U_c^\top U_c = I$ and $G = V_c^\top X$. With target perturbation E_t and supplement perturbation E_G, the reconstructed source error is

$$\widehat{S}_s - S_s = A_\gamma E_t + U_c E_G, \qquad \left\| \widehat{S}_s - S_s \right\|_F \leq \gamma \|E_t\|_F + \|E_G\|_F. \tag{26}$$

Different optimal operators can attain the same correction rank. Operator representation and affine-offset representation remain separate costs.

B.4 Two maps distinguish missing information from unstable information

Consider two illustrative heads with zero offsets and X with rows $x_1, x_2 \in \mathbb{R}^{1 \times d_v}$. First, let

$$P_s = I_2, \qquad P_t = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{27}$$

The target reveals x_1 and erases x_2, while the source needs both. Here $r_{\text{miss}} = 1$: sending the single row x_2 supplies exactly the missing source information. This example also shows why counting supplementary rows differs from counting complete packet bytes. The receiving protocol must already know the maps and the meaning of the transmitted row.

Now replace the zero by ε, with $0 < \varepsilon < 1/\gamma$ and $\gamma \geq 1$. Then $P_t = \operatorname{diag}(1, \varepsilon)$ is invertible and $r_{\text{miss}} = 0$, but $M = \operatorname{diag}(1, \varepsilon^{-1})$ amplifies error in the second target direction. The bounded-gain construction has $r_\gamma = 1$ and recovers

$$S_s = \operatorname{diag}(1, \gamma) S_t + \begin{bmatrix} 0 \\ 1 \end{bmatrix} (1 - \gamma\varepsilon) x_2. \tag{28}$$

The target supplies the stable part of that direction and the sender supplies the remainder. Thus r_{miss} identifies information that has disappeared, whereas r_γ identifies information that a bounded-gain receiver should obtain through supplementation. GDN's key span gives these supplementary directions a concrete structured representation.

C Executed Factors and Subspace Residuals

C.1 Gated DeltaNet invariant span

The affine sufficiency statements apply to fixed realized maps. Scalar matching additionally requires scalar action on the complement of the joint key span. KDA and Gated DeltaNet-2 use channel-wise decay, and Gated DeltaNet-2 separates erase and write gates (Hatamizadeh et al., 2026). These updates require a different invariant-space argument and are not evaluated here. For $k \in \mathrm{span}(W)$ with $u = W^\top k$,

$$(I - \beta k k^\top)\Pi_\perp = \Pi_\perp, \qquad (I - \beta k k^\top)W = W(I_m - \beta u u^\top). \tag{29}$$

The additive update $\beta k v^\top$ also has columns in $\mathrm{span}(W)$. Induction over the executed branch therefore gives

$$P_b = D_b(\Pi_\perp + W T_b W^\top), \qquad Q_b = W J_b. \tag{30}$$

Writing $F_i = I_m - \beta_i u_i u_i^\top$, the normalized recurrences start from $(D_0, T_0, J_0) = (1, I_m, 0)$ and are

$$D_i = \alpha_i D_{i-1}, \qquad T_i = F_i T_{i-1}, \qquad J_i = \alpha_i F_i J_{i-1} + \beta_i u_i v_i^\top. \tag{31}$$

A runtime may instead keep $\bar{T}_b = D_b T_b$. The coefficient residual then takes the equivalent form

$$R_C = J_s - \lambda J_t + (\bar{T}_s - \lambda \bar{T}_t) W^\top X. \tag{32}$$

The normalized and unnormalized forms use the scalar decay exactly once.

If the basis is expanded by a direction orthogonal to previous keys, the preceding normalized transition extends by an identity block and J by a zero row. An orthogonal rotation $W' = WO$ changes coordinates to $T' = O^\top T O$ and $J' = O^\top J$. The source and target must be aligned to the same basis before residual construction.

C.2 Innovation factors are tied to native execution

Expanding the update yields

$$S_i = \alpha_i S_{i-1} + k_i r_i^\top, \qquad r_i = \beta_i (v_i - \alpha_i S_{i-1}^\top k_i). \tag{33}$$

The innovation is a state-dependent residual, not the uncorrected value vector. With $\omega_{b,i} = \prod_{j=i+1}^{n_b} \alpha_{b,j}$, define

$$U_b = [k_{b,1}, \ldots, k_{b,n_b}], \qquad V_b = [\omega_{b,1} r_{b,1}, \ldots, \omega_{b,n_b} r_{b,n_b}]. \tag{34}$$

Unrolling gives $S_b = D_b X + U_b V_b^\top$. For $D_t \neq 0$, set $\lambda = D_s / D_t$ and concatenate

$$U = [U_s, U_t], \qquad V = [V_s, -\lambda V_t]. \tag{35}$$

Then $S_s = \lambda S_t + U V^\top$. Once innovations have been captured, this packet can be formed without reading the base again. The capture operation has already incorporated the effect of the actual base into the innovations. Holding these factors fixed while varying X is therefore distinct from holding the affine maps fixed in Theorem 1.

C.3 Direct factors and the two SSR construction paths

If W covers the columns of U,

$$U V^\top = W(W^\top U) V^\top = W R_I. \tag{36}$$

Independently, subtracting the affine branch decompositions with $D_s = \lambda D_t$ gives

$$S_s - \lambda S_t = W[J_s - \lambda J_t + D_s(T_s - T_t) W^\top X] = W R_C. \tag{37}$$

Since W is orthonormal, $R_I = R_C$ in the matching exact calculation. SSR-I and SSR-C obtain the same residual through different inputs and arithmetic paths.

For an approximate basis, the projection-only source error is

$$E_{\mathrm{proj}} = -\big(I - W W^\top\big) U V^\top. \tag{38}$$

With $p = n_s + n_t$, direct factors have $1 + p(d_k + d_v)$ recurrent values, while SSR has $1 + m(d_k + d_v)$. Thus column-space redundancy $m < p$ provides a count reduction. At $m = p$, reencoding changes coordinates without reducing this count. Dtypes and field framing determine the final serialized size.

C.4 Eliminating the matrix term

For invertible T_t, KSO uses $M_W = \lambda T_s T_t^{-1}$ and $A_W = \mathrm{svclip}_\gamma(M_W)$. In the orthogonal decomposition its target operator is

$$A = \lambda \Pi_\perp + W A_W W^\top = \lambda I + W(A_W - \lambda I)W^\top. \tag{39}$$

When the complement is nonempty, the condition $|\lambda| \le \gamma$ bounds its complementary gain. KSO's correction is

$$H = J_s - A_W J_t + (D_s T_s - D_t A_W T_t) W^\top X. \tag{40}$$

Using $W^\top S_t = D_t T_t W^\top X + J_t$,

$$\begin{aligned} &(A_W - \lambda I)W^\top S_t + H \\ &= (A_W - \lambda I)(D_t T_t W^\top X + J_t) + J_s - A_W J_t + (D_s T_s - D_t A_W T_t) W^\top X \\ &= J_s - \lambda J_t + D_s(T_s - T_t) W^\top X = R_C. \end{aligned} \tag{41}$$

The matrix terms cancel before transmission. Direct coefficient construction avoids calculating them first.

When $m < d_k$, every operator in this structured class has $\|A\|_2 = \max(|\lambda|, \|A_W\|_2)$. The scalar choice $A = \lambda I$ attains the minimum worst-direction gain within that class. The clipped operator instead optimizes base-correction rank. On the common domain of invertible P_t and $|\lambda| \le \gamma$,

$$r_\gamma \le \mathrm{rank}\big(D_s(T_s - T_t)\big) \le m. \tag{42}$$

The spectrum contains the singular values of M_W and $d_k - m$ copies of $|\lambda|$. The supplied experiments do not retain per-head spectra for an empirical evaluation of r_γ.

C.5 Affine information, large gain, and fallback

If $P_s = P_t = I$ and $\gamma \ge 1$, then $r_\gamma = 0$, while $Q_s - Q_t$ can still be nonzero. The complete residual must encode this affine information. A count such as $r_\gamma d_v$ describes only the supplementary projection results, with operators and offsets supplied separately.

The SSR gain is $|\lambda|$ when the packet is fixed. Let $L_b = U_b V_b^\top$ denote the weighted writes accumulated along branch b. Substituting a different scalar a gives

$$S_s - aS_t = (D_s - aD_t)X + L_s - aL_t. \tag{43}$$

Its complementary contribution $(D_s - aD_t)\Pi_\perp X$ must be supplied unless it vanishes. Clipping the scalar alone does not preserve the original residual formula. The implementation therefore uses explicit admissibility checks and dense fallback. For positive decay, log-domain ratio calculation can avoid underflow of separately multiplied scalars; it does not remove amplification intrinsic to the target state.

D Numerical State and Continuation

D.1 Separating native mapping from packet error

Define native affine residuals

$$\eta_b = S_b^{\mathrm{native}} - (P_b X + Q_b). \tag{44}$$

For exact coefficient fields and a receiver perturbation E_t,

$$\widehat{S}_s^{\mathrm{SSR-C}} - S_s^{\mathrm{native}} = \lambda E_t + \lambda \eta_t - \eta_s, \tag{45}$$

while KSO contributes $AE_t + A\eta_t - \eta_s$. These expressions describe different propagation operators, not a universal ordering of native endpoint error.

For direct factors define separately $\eta_b^{\mathrm{factor}} = S_b^{\mathrm{native}} - (D_b X + U_b V_b^\top)$. DFD has the same scalar form with η_b^{factor} replacing η_b. This definition groups innovation-capture and accumulation errors separately from the affine-map residual. SSR-I also introduces projection and residual-encoding arithmetic.

With field errors $\delta\lambda, E_W, E_R$ and application error E_{arith}, expand

$$\widehat{S}_s = (\lambda + \delta\lambda)(S_t^{\text{native}} + E_t) + (W + E_W)(R_C + E_R) + E_{\text{arith}}. \tag{46}$$

For orthonormal nominal W,

$$\begin{aligned}\left\|\widehat{S}_s - S_s^{\text{native}}\right\|_F \le\ & |\lambda| \left\|E_t\right\|_F + |\lambda| \left\|\eta_t\right\|_F + \left\|\eta_s\right\|_F \\ & + |\delta\lambda| \left\|S_t^{\text{native}} + E_t\right\|_F + \left\|E_R\right\|_F \\ & + \left\|E_W\right\|_2 \left(\left\|R_C\right\|_F + \left\|E_R\right\|_F\right) + \left\|E_{\text{arith}}\right\|_F .\end{aligned} \tag{47}$$

The residual term includes coefficient conversion, projection, and accumulation. If both native endpoints are available for an audit, the identity

$$\lambda S_t^{\text{native}} + W(W^\top S_s^{\text{native}} - \lambda W^\top S_t^{\text{native}}) - S_s^{\text{native}} = -\Pi_\perp(S_s^{\text{native}} - \lambda S_t^{\text{native}}) \tag{48}$$

measures the native difference outside the supplied key span.

D.2 Decision margins

Let i be the unique top-1 token of reference logits z under a common token history, and let $a = z_i - \max_{j \neq i} z_j > 0$. Then

$$\widehat{z}_i - \widehat{z}_j \ge a - 2\left\|\widehat{z} - z\right\|_\infty . \tag{49}$$

Thus $2\left\|\widehat{z} - z\right\|_\infty < a$ preserves the current greedy decision. An identical continuation requires all decisions to agree through termination. After the first token difference, freely generated trajectories condition on different histories.

$$2\left\|\widehat{z} - z\right\|_\infty < a \quad \Longrightarrow \quad \arg\max \widehat{z} = \arg\max z. \tag{50}$$

The reference-based diagnostic $\rho = 2\left\|\Delta z\right\|_\infty / \max(a, 10^{-6})$ is used in the shared-base analysis below. The floored score is descriptive: when $a < 10^{-6}$, $\rho < 1$ need not imply the unfloored sufficient condition.

E Unified Headwise Comparison

S4.11 evaluates the original capture implementation on 54 endpoints; Appendix Q.1 evaluates corrected capture. Appendix R summarizes the evidence boundaries across studies.

E.1 Conditions, controls, and reported outputs

The unified headwise study S4.11 contains 40 short conditions from an existing 10-cluster input suite and 14 boundary conditions spanning five of those clusters. Short lengths are 1 to 4. The boundary suite has four conditions each at lengths 8, 16, and 32, and two at length 64. The methods are Full Snapshot, DFD-H, SSR-I-H, and SSR-C-H. All compact paths use per-head dense fallback, and endpoint-atomic checkpoints bind each result to its inputs and configuration.

All 40/40 short and 14/14 boundary conditions completed, with matching Full-to-Full controls on all 54. The boundary delivery contains 56 method and 56 continuation records. Each method and condition uses five warm-ups, 20 timing repeats, and greedy continuation capped at 128 new tokens.

The artifact contains 216 endpoint-method records and 4,320 request measurements. For endpoint e and method j, let t_{ej} be the median of its 20 formal repetitions. The request ratio is $\text{median}_e(t_{ej}/t_{eF})$; packet ratios use the same endpoint pairing. Absolute columns summarize each method separately. Short-suite intervals resample 10 prompt clusters with replacement, retaining all endpoints within each selected cluster, for 10,000 replicates (seed 4112026). Marginal percentile bounds use sorted bootstrap indices 250 and 9750 (zero-based). Analysis files include raw records, direct paired SSR/DFD comparisons and joint A_{128} counts. Boundary results are descriptive, with five clusters and smaller per-length groups.

Table 7: Historical short-branch comparison (40 endpoints, 10 prompt clusters). Packet sizes and request times are endpoint medians; ratios use endpoint pairing. Brackets give marginal 95% confidence intervals (CI). Coverage pools compact heads; A_{128} counts exact continuation matches. Appendix E defines this study's estimator and timing boundary; Table 1 defines DFD and SSR. H denotes headwise fallback; MB is 10^6 bytes and ms milliseconds.

Method	Packet MB	/ Full	Request ms	/ Full [95% CI]	Coverage	A_{128}
Full	52.241	1.000	499.0	1.000	n/a	40/40
DFD-H	8.114	0.155	717.7	1.416 [1.402, 1.443]	95.44%	17/40
SSR-I-H	8.120	0.155	955.5	1.890 [1.817, 1.963]	95.44%	18/40
SSR-C-H	8.117	0.155	676.4	1.347 [1.336, 1.382]	95.44%	20/40

Table 8: Aggregate headwise coverage for the short and boundary suites. Each row gives the head counts and compact fraction for each of the three compact methods. Abbreviations and metrics are defined in Table 6.

Suite	Conditions	Head instances	Dense fallback	Compact coverage
Short	40	30,720	1,402	95.44%
Boundary	14	10,752	1,862	82.68%

E.2 Head coverage

For endpoint e, head h, and method j, let b_{eh} denote dense recurrent bytes and z_{ejh} indicate actual compact use. Bytes-weighted coverage is

$$c_{ej} = \frac{\sum_h b_{eh} z_{ejh}}{\sum_h b_{eh}}. \tag{51}$$

The endpoint caches contain FP32 recurrent matrices with $d_k = d_v = 128$, giving $b_{eh} = 65{,}536$ bytes. Short aggregate coverage is $1 - 1402/30720 = 0.954362$, and boundary coverage is $1 - 1862/10752 = 0.826823$. Comparing every saved layer/head key gives zero mask disagreements for every method pair on all 54 endpoints. These masks come from each method's final pipeline call; timing-repeat and quality-reference masks were not stored.

DFD-H and SSR-I-H each record 1,402 high-gain rejections in the short suite. Boundary rejections comprise 1,785 high-gain and 77 zero-target-decay cases, with no recorded nonfinite factor/basis rejection. SSR-C-H logs only `SSR_C_UNCERTIFIED` for its 1,402 and 1,862 rejected heads; the specific causes remain unknown. The artifact retains per-length coverage and raw head records, including dense fallback.

The DFD/SSR-I finite λ distributions have short-suite quartiles $(0.9002, 0.9976, 1.0421)$ and boundary quartiles $(0.7869, 1.0047, 1.3147)$. Invalid decay ratios leave 77 boundary values missing; SSR-C logs λ only for admitted compact heads. The frozen concatenation code gives factor width $p = n_s + n_t$. The log key `m_nominal` denotes padded basis width for SSR-I and effective rank for SSR-C. On short compact heads SSR-C has $m/p = 1$ throughout; boundary ratios range from 0.875 to 1 with median 1. Both serializers retain padded width $\min(d_k, p)$, so a lower effective rank need not reduce packet size.

E.3 Boundary table and aggregation

Table 9 includes matched Full bytes and times from the same checkpoints. The pooled boundary packet ratio is approximately 0.687; request ratios are 1.407, 1.744 and 1.557 for DFD-H, SSR-I-H and SSR-C-H. Packets slightly exceed Full at length 32 and reach approximately 1.73 times Full at length 64 (Figure 2).

The boundary continuation totals are 8/14 for DFD-H, 6/14 for SSR-I-H, and 7/14 for SSR-C-H.

Table 9: Boundary results from S4.11. Equal source and target lengths are denoted by n. Absolute columns are endpoint medians; ratios summarize paired endpoint comparisons. Groups contain 4, 4, 4, and 2 endpoints and are descriptive. Coverage is the compact-head fraction; A_{128} counts exact continuation matches. Abbreviations and metrics are defined in Table 6.

n	Method	Packet MB	/ Full	Request ms	/ Full	Coverage	A_{128}
8	Full	52.420	1.000	920.0	1.000	n/a	4/4
8	DFD-H	17.440	0.333	1223.7	1.352	92.90%	3/4
8	SSR-I-H	17.446	0.333	1543.8	1.675	92.90%	2/4
8	SSR-C-H	17.443	0.333	1306.8	1.441	92.90%	2/4
16	Full	52.682	1.000	1377.6	1.000	n/a	4/4
16	DFD-H	32.212	0.611	2038.9	1.471	77.08%	2/4
16	SSR-I-H	32.217	0.612	2576.0	1.819	77.08%	2/4
16	SSR-C-H	32.215	0.611	2173.6	1.550	77.08%	2/4
32	Full	53.206	1.000	2482.0	1.000	n/a	4/4
32	DFD-H	53.396	1.004	3584.3	1.444	79.98%	2/4
32	SSR-I-H	53.401	1.004	4355.0	1.744	79.98%	2/4
32	SSR-C-H	53.399	1.004	4025.8	1.652	79.98%	3/4
64	Full	54.255	1.000	4660.0	1.000	n/a	2/2
64	DFD-H	94.127	1.735	6526.2	1.400	78.84%	1/2
64	SSR-I-H	94.132	1.735	7955.8	1.707	78.84%	0/2
64	SSR-C-H	94.129	1.735	7901.4	1.696	78.84%	0/2

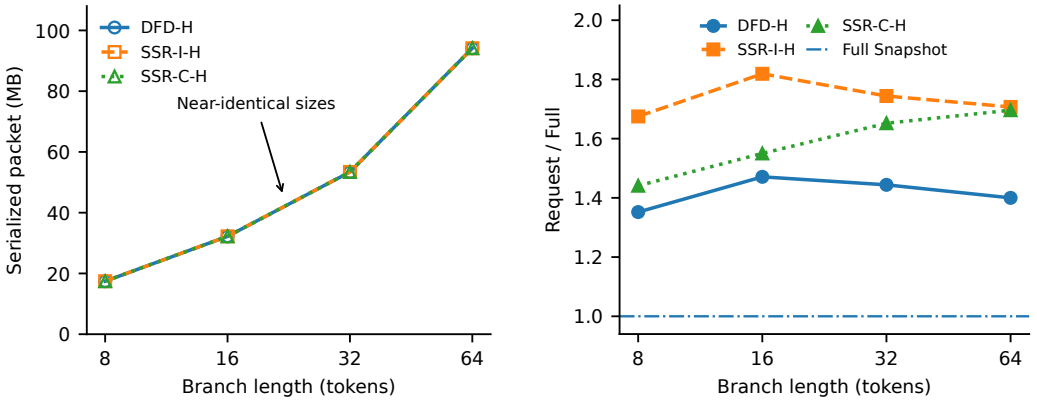

Figure 2: Historical boundary comparison (S4.11). Left: serialized packet size. Right: complete local request ratios. Lines connect observed length groups with 4, 4, 4 and 2 endpoints. DFD and SSR are defined in Table 1; H denotes headwise fallback.

E.4 Behavior, stopping, and matched-history diagnostics

The saved token lists reproduce every A_1, A_{10}, A_{32}, A_{64} and A_{128} flag and first-divergence index. The frozen readout helper appends token ID 198 before the reported predictions. Greedy decoding includes EOS and stops on tokenizer EOS or the 128-prediction cap. Token IDs and lengths were saved, but explicit per-record termination flags were not. Free-running divergence indices are zero-based; matched-history steps are one-based.

The diagnostic roster uses the 12 lowest hashes of a fixed salt plus endpoint ID from the 40 primary conditions, without outcome filtering. It spans seven clusters, with 128 teacher-forced predictions per endpoint/method and 1,536 steps per method. Table 11 reports first top-1 disagreement, reference margin, logit perturbation and terminal recurrent error. Full has no disagreement; DFD-H, SSR-I-H and SSR-C-H have 10, 10 and 13 mismatching steps. Equation (50) certifies 1,387, 1,385 and 1,386 steps, respectively, all of which agree. Another 139, 141 and 137 uncertified steps also agree. A zero top-2 margin supplies no unique-top-1 certificate; median first-mismatch margins are zero for DFD-H and SSR-C-H.

Table 10: Exact greedy continuation agreement with Full through h predictions (A_h). Token-prefix comparisons include the end-of-sequence (EOS) token and earlier termination. Entries count matching endpoints rather than task successes. Abbreviations and metrics are defined in Table 6.

Suite	Method	A_1	A_{10}	A_{32}	A_{64}	A_{128}
Primary	Full	40/40	40/40	40/40	40/40	40/40
Primary	DFD-H	39/40	37/40	34/40	28/40	17/40
Primary	SSR-I-H	40/40	37/40	32/40	26/40	18/40
Primary	SSR-C-H	40/40	37/40	35/40	28/40	20/40
Boundary	Full	14/14	14/14	14/14	14/14	14/14
Boundary	DFD-H	14/14	14/14	13/14	9/14	8/14
Boundary	SSR-I-H	14/14	13/14	12/14	9/14	6/14
Boundary	SSR-C-H	14/14	14/14	13/14	10/14	7/14

Table 11: Matched-history diagnostics on 12 hash-selected endpoints from seven clusters. All methods follow 128 predictions along the saved Full history. First step, margin a, and maximum logit perturbation $\|\Delta z\|_\infty$ are separate medians over endpoints with a mismatch. Steps are one-based; n/a indicates no mismatch. Terminal relative error is the median recurrent-state Frobenius error over all 12 endpoints. Abbreviations and metrics are defined in Table 6.

Method	Mismatch	First step	Margin a	$\|\Delta z\|_\infty$	Terminal rel. L_2
Full	0/12	n/a	n/a	n/a	0.00000
DFD-H	7/12	72	0.0000	0.1562	0.00680
SSR-I-H	6/12	36.5	0.0625	0.1406	0.00672
SSR-C-H	6/12	39	0.0000	0.1328	0.00676

Restoration-time median recurrent relative errors are 0.001673, 0.001673 and 0.001476; terminal medians rise to approximately 0.0068. All compact terminal KV and convolution equality flags are false, whereas all Full flags are true. These post-forward comparisons do not establish an initial assembly defect. Restoration-boundary equality, full logits, intermediate field trajectories and precision-isolation controls were not recorded, leaving the numerical source unresolved. The terminal position flag uses `getattr(layer, cache_position, None)`, which can compare equal for absent attributes. Forward calls share explicit position IDs derived from the endpoint offset and token schedule, but these tensors were not separately logged.

E.5 Complete request and packet boundaries

Complete request cost is synchronized wall time around `endpoint_pipeline`. It includes both branch executions, capture, descriptor and packet construction, serialization, decoding, restoration, fresh readout, source-state error calculation and wrapper bookkeeping. Prefix prefill and common-fork creation precede this interval. An endpoint-seeded schedule shuffles methods within each warm-up/repetition. Table 12 records stages from single reference calls, with path-specific boundaries; their medians do not sum to the complete-request median. SSR-I basis construction and residual aggregation enter request wall time but are excluded from its recorded descriptor-build field.

A complete serialized record has components

$$B_j = B_{j,\mathrm{recurrent}} + B_{\mathrm{KV}} + B_{\mathrm{conv}} + B_{\mathrm{position}} + B_{\mathrm{metadata}} + B_{\mathrm{framing}}. \tag{52}$$

All methods reuse the same resident prefix and send the same non-recurrent source suffix. Full sends dense recurrent heads; compact paths include dense fallback in $B_{j,\mathrm{recurrent}}$. Total serialized bytes were saved for the reference packet and every timing repeat, but serialized blobs and a byte-by-field ledger were not retained.

Receiver-resident storage, serialized bytes, and simultaneous live memory are separate quantities. For a device d, a peak requires

$$M_d^{\mathrm{peak}} = \max_t \sum_{o \in \mathcal{L}_d(t)} \mathrm{bytes}(o), \tag{53}$$

Table 12: Recorded stage times in milliseconds, summarized as endpoint medians of single reference calls. Source and Target include branch execution and capture; Build is descriptor construction; Fresh is fresh forward readout. SSR-I Build excludes basis construction and residual aggregation. SSR-C Build sums its instrumented construction stages. Stages have different boundaries and do not sum to the complete-request median. Abbreviations and metrics are defined in Table 6.

Suite	Method	Source	Target	Build	Serialize	Restore	Fresh
Primary	Full	80.9	79.7	0.0	63.4	32.4	39.5
Primary	DFD-H	99.7	101.3	18.4	36.4	95.3	39.2
Primary	SSR-I-H	105.0	103.1	18.0	35.9	95.0	38.9
Primary	SSR-C-H	100.7	97.6	74.3	36.0	94.2	39.1
Boundary	Full	592.4	619.3	0.0	66.0	32.8	39.1
Boundary	DFD-H	752.8	762.6	64.8	67.7	90.1	39.4
Boundary	SSR-I-H	785.6	703.3	46.8	67.3	95.4	39.4
Boundary	SSR-C-H	711.6	778.5	464.8	61.0	89.2	36.4

with storage aliases counted once. The available ledgers do not establish this sender/receiver/workspace peak.

E.6 Execution configuration and source binding

SHA-256 inventories bind the S4.11 manifest, configurations and checkpoints; configuration hashes and roster membership agree for all 54 checkpoints. Eleven recorded source/input bindings match the recovered files. Unhashed modules remain recovered dependencies rather than verified historical process versions. Table 13 distinguishes declared model/tokenizer revisions from imported software versions and lists the missing execution fields.

Table 13: S4.11 execution metadata. Missing fields retain their historical provenance status. Data types and cache abbreviations are defined in Table 6.

Field	Recorded value or provenance
Model	`Qwen/Qwen3.5-4B`
Model snapshot	`851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a`
Weights / KV / conv	BF16 / BF16 / BF16
Recurrent state	FP32; 24 GDN layers, 32 value heads/layer, $d_k = d_v = 128$
Captured fields	FP32 key, value, beta, decay, and `native_error`
Descriptor compute / packet	FP32 / FP32
Gain / rank threshold	$\gamma = 2$; rank threshold 10^{-6}
Capture path	State-based innovation reconstruction at the actual recurrent callable boundary; one native call per step
Recurrent backend	Transformers Qwen3.5 callable; exact selected kernel and kernel commit not independently logged
Recorded software	Python 3.12.13, PyTorch 2.10.0+cu130, CUDA 13.0, Transformers 5.14.1
Runtime revision	`a39334701154b05400b09609f5adb708777b6ae0`; recorded in manifest, separate runtime Git verification not replayed
Tokenizer / driver	Tokenizer revision declared equal to model snapshot; driver version not recorded
Chunk / positions	Token-wise branch and continuation forwards; explicit positions from endpoint offset; common prompt prefilled before timed interval
Kernel / experiment commit	Not recorded; experiment source and five runtime-module hashes recorded separately

The manifest's cache dtypes came from a separate S4.11 smoke run. Formal endpoint records independently confirm FP32 recurrent state and BF16 KV/convolution fields.

E.7 KV SUFFIX CORRECTION AND COMPARISON VALIDITY

The legacy equal-tail replacement helper attempted to place a 64-token source suffix into eight target positions when $(n_s, n_t) = (64, 8)$. The correction removes exactly n_t target KV positions before concatenating the source suffix, leaving recurrent residual equations and field types unchanged.

The equal-length 40-condition primary and 14-condition boundary suites do not exercise this defect. Initial unequal-length failures are invalid method comparisons; all 12 asymmetric conditions were rerun with corrected assembly and passing Full controls.

F SIBLING-ONLY TASK QUALITY AND BOUNDARY AUDITS

These studies test continuation and restoration boundaries with the historical capture paths. Appendix Q reports corrected capture and persistent interaction.

Table 14: Task success after sibling-only recovery.

Method	Inventory	Registers	Routes
Full	128/128	116/128	59/60
DFD-H	128/128	117/128	60/60
SSR-I-H	128/128	116/128	n/r
SSR-C-H	128/128	118/128	60/60
Sibling alone	n/r	0/128	n/r
Non-recurrent only	n/r	123/128	60/60

Inventory and registers grade fields in an extracted JSON (JavaScript Object Notation) object. Routes grade all three decisions and edge costs with strict JSON, using the 60-case confirmation. DFD and SSR follow Table 1; H denotes headwise fallback; n/r means not run. Non-recurrent only installs source non-recurrent fields while retaining the sibling recurrent state.

F.1 CONTROLLED TASKS AND GRADING

The inventory study (S4.12) contains 128 formal cases, with 64 each at pause lengths 2 and 8. A prompt specifies two warehouses, one item, prices and requested quantity. The source selects one warehouse; the sibling selects the other. Continuation must return the selected warehouse, item identifier, quantity and total in an extracted JSON object. All four methods pass this frozen field grader on all cases. Whole-output JSON parseability is graded separately: Full and DFD-H pass on 64 cases, and SSR-I-H and SSR-C-H on 63.

The register study (S4.14) uses two four-instruction integer programs containing addition, multiplication, subtraction and a clamp to zero. The source selects the program to execute. Output fields are `branch,s1,s2,s3,s4`; each numbered field records the state after that instruction, including unchanged states. The first extracted object must contain exactly these fields, with integer states. Boolean values, duplicate keys, extra fields and incorrect intermediate states fail. Secondary grading checks the complete supplied-prefill-plus-generated output and gives the same success counts. Greedy decoding uses a 112-token limit, the tokenizer's chat template with thinking disabled, and a supplied opening brace. All paths consume the same neutral continuation suffix after restoration.

The register roster balances selected branch, pause length and whether the clamp changes the state, with 16 cases in each of eight cells. The 128 full source/sibling program pairs are unique and disjoint from development pairs; every formal case remains in the denominator. These controlled local families share implementation and template structure. Results are reported separately, rather than pooled as 256 independent deployment tests; Appendix R summarizes their statistical and application scope.

F.2 PAIRED OUTCOMES AND FIELD CONTROLS

The non-recurrent-only control succeeds on more cases overall but fails on three that Full solves. Its strong result leaves the contribution of recurrent repair unresolved. Sibling alone follows the

Table 15: Exact complete-continuation token agreement with Full on the inventory and register studies, distinct from functional correctness. DFD and SSR names follow Table 1; H denotes headwise fallback.

Method	Inventory	Registers
DFD-H	122/128	104/128
SSR-I-H	124/128	101/128
SSR-C-H	124/128	108/128

Table 16: Register outcomes paired with Full across all 128 formal cases. The four outcome columns report task success, not token equality. DFD and SSR constructions follow Table 1; H denotes headwise fallback. Sibling alone uses no source record; non-recurrent only replaces source non-recurrent fields but keeps sibling recurrent state.

Method	Both correct	Full only	Method only	Both wrong
DFD-H	115	1	2	10
SSR-I-H	114	2	2	10
SSR-C-H	116	0	2	10
Sibling alone	0	116	0	12
Non-recurrent only	113	3	10	2

wrong branch on every case and produces the sibling's complete correct state sequence on 119/128. Restored non-recurrent fields match the native source in all formal cases. Full restoration matches the complete native source cache; captured and uncaptured source caches also match. Audits verify unchanged sibling recurrent fields and immutable resident anchors.

Register packet medians are 52.323 MB for Full and 10.898, 10.904 and 10.901 MB for DFD-H, SSR-I-H and SSR-C-H. Each compact path uses 96,891 compact heads out of 98,304. The 96 cases flagged for dense heads use headwise, not endpoint-wide, fallback. Median recurrent relative errors are 0.001235, 0.001235 and 0.001074.

For inventory, all compact/Full pairs occupy the both-success cell. The conservative 95% accuracy-difference interval is $[-0.0284, 0.0284]$ per comparison. It uses an exact zero-cell bound at level $1 - 0.05/2$ for each discordant direction, correcting within each comparison rather than across all three methods. A separate 24-case timing subset contains 480 formal rows after two warm-ups and five repeats per method. Pooled lifecycle medians are 1661.35 ms for Full and 1933.87, 2163.42 and 1935.55 ms for the compact paths. These times include diagnostic work and remain separate from the deployment-oriented S4.13 timing boundary.

F.3 Development history and input overlap

The S4.13 register pilot requested a trace list. Full passed 5/12 initially and 2/12 after one chat-interface revision, below the frozen 10/12 launch threshold; the formal study was not run. Eight revised-pilot outputs prepended the initial value to an otherwise correct four-state sequence. The frozen grader was retained, with success at 2/12.

S4.14 uses numbered fields under a new frozen protocol. Full solves 11/12 reused development programs and passes the same launch threshold. Formal program pairs are new, but an audit during the frozen run found two source programs and initial values previously seen with different siblings. The primary denominator remains 128. Excluding those inputs gives a secondary subset of 126, with correct counts 114, 115, 114 and 116 for Full and the compact paths. Sibling alone solves zero and non-recurrent only 121. The subset was identified before final outcome analysis but was not a frozen primary analysis. The artifact retains the overlap identifiers and development outputs.

F.4 Restoration boundary and numerical stages

The S4.12 follow-up examines 12 matched-history and four unequal-length endpoints. All pass the recorded non-recurrent-field, token-boundary and anchor checks. The stage audit samples the first,

middle and last recurrent layers and heads 0, 7, 15 and 31, producing 192 rows. It compares captured-coefficient recurrence in FP64, native states, and application of the same DFD/dense packet in FP32 and FP64. Packet round trips preserve sampled fields exactly, and audited fork/source/target caches have no cross-cache storage aliases. Cache-resident position attributes are absent on all 16 endpoints and marked missing; explicit position inputs and token lengths are recorded separately.

Across sampled heads, the FP32-versus-FP64 application relative difference has median 4.06×10^{-8} and maximum 1.23×10^{-7}. FP32 application versus native source has median 1.48×10^{-3} and maximum 8.05×10^{-3}, using the recorded reference denominators. Applying the same packet in FP64 rather than FP32 changes little relative to its discrepancy from the native source. This sampled DFD/dense audit leaves upstream causes unresolved: model execution and packet construction were not rerun in FP64, and SSR-I/SSR-C precision was not isolated.

G Multi-Turn Route Recovery

G.1 Task, receiver and frozen confirmation

The route task requires three decisions on a directed, layered graph. A shared prompt specifies ten roads, each with a base cost and a per-load coefficient. Executed source and sibling branches supply different starting nodes and vehicle loads. Each branch has seven tokens. At every turn, the model selects the first edge of the cheapest remaining route and reports that edge's cost. Ties are broken by alphabetical next node. After the first move, a tool updates four road costs, so later decisions depend on both the recovered vehicle state and new observations.

Success requires all three actions and reported costs to be correct. The strict JSON object contains only `next` and `cost`, with a string node and integer cost. Duplicate keys, Boolean costs, malformed output and illegal moves fail. A legal but incorrect move continues but fails the task, even if the model later reaches the destination.

The initial held-out study (S4.16) has 24 tasks balanced over six starting-node/load cells. A separate 60-task confirmation (S4.17), with ten new tasks per cell, was frozen after inspecting the initial results. Confirmation tasks and source contexts are disjoint from the 12 development tasks and initial 24 tasks. The constructor, receiver, model revision, gain gate and decoding budgets remain fixed. Confirmation stops at 60 tasks regardless of outcomes, with no quality-based exclusions or retries. Table 14 uses these 60 only; Table 17 reports the earlier study separately.

Full, DFD-H and SSR-C-H restore the source cache under the same resident-sibling condition. Full installs dense recurrent fields; DFD-H and SSR-C-H reconstruct them from the sibling and their decoded records. The non-recurrent-only control installs source non-recurrent fields but keeps sibling recurrent state. SSR-I-H was not run in this study. The shared decoding engine retains all generated reasoning in the live cache across tool turns. Each turn allows 2,048 reasoning tokens followed by up to 128 answer tokens, using greedy decoding. At the reasoning limit, the interface supplies a closing thought delimiter and an opening brace. A consumed closing brace ends the answer before appending the next tool response. No oracle answer is supplied to the model, and no source history is rebuilt between turns.

G.2 Paired quality, precision and recorded bytes

In the 24-task study, Full, DFD-H, SSR-C-H and non-recurrent only solve 23, 24, 24 and 23 tasks. In the separate confirmation, they solve 59, 60, 60 and 60. All recorded outputs satisfy strict JSON: 72 per method initially and 180 per method in confirmation. Full makes one incorrect decision in each study; each compact path makes none. The initial non-recurrent-only control fails on a different task from Full.

The task is the analysis unit; its three decisions and four method evaluations are repeated measurements. The frozen primary contrast is SSR-C-H versus Full; DFD-H is secondary and non-recurrent only a field control. Confirmation has one compact-only success and no Full-only success, an accuracy difference of 1.67 percentage points.

A post-run precision calculation bounds undetected losses under an independent-task sampling model. A loss means Full succeeds and the compared method fails. With zero losses in n tasks,

Table 17: Route outcomes paired with Full, separately for the initial study and the new confirmation. Each row counts tasks, not individual decisions. Both refers to both methods; wrong means task failure. Non-recurrent only retains sibling recurrent state. DFD and SSR names follow Table 1; H denotes headwise fallback.

n	Method	Both correct	Full only	Method only	Both wrong
24	DFD-H	23	0	1	0
24	SSR-C-H	23	0	1	0
24	Non-recurrent only	22	1	1	0
60	DFD-H	59	0	1	0
60	SSR-C-H	59	0	1	0
60	Non-recurrent only	59	0	1	0

a one-sided 95% upper bound on the mean loss probability is $1 - 0.05^{1/n}$. This follows from $\Pr(\text{no losses}) = \prod_i (1 - p_i) \leq (1 - \bar{p})^n$, allowing different loss probabilities across tasks. The per-comparison bounds are 4.87% for the 60-task confirmation and 11.73% for the initial 24 tasks. They concern the tested sampling design and are not simultaneous across methods.

The confirmation's single Full failure occurs on `route-formal-072`, at the second decision. Full selects the correct next node F but reports cost 2 instead of 3; all other methods answer correctly. The cause of this difference remains unresolved, and the perfect non-recurrent-only score does not establish a contribution from recurrent repair.

Both compact paths use 45,347 compact heads out of 46,080 and select identical masks on all 60 confirmation tasks. Actual serialized packet medians are about 26.5% of Full (Table 18). Episode wall times include diagnostics and reasoning, rather than isolating recovery cost. Transfer-cost measurements use the separate inventory and layout workloads.

G.3 Execution controls and development record

Before held-out inference, two hash-selected development tasks compare native continuation with Full restoration under the live-cache interface. All 8,913 paired forward calls have identical recorded logits; audited cache fields, explicit positions and attention lengths also match. Both tasks complete correctly. A separate trace check reproduces one audited task with the formal-study engine. These checks are engineering controls, not additional quality samples.

All 336 method/task records pass the recorded boundary and field checks. Captured source fields match native source fields, restored non-recurrent fields match the source, and the anchor stays unchanged. Scalar metadata and explicit positions are checked separately; absent cache-resident position attributes remain marked missing. CPU-only analysis recomputes strict parsing, optimal actions, tool observations, positions, outcomes and paired counts from frozen rosters and saved outputs.

Interface development reused 12 tasks. Direct answering solved 0/12, including full-context checks; thinking-enabled decoding with a 2,048-token cap produced no final answers. An early bounded-answer summary is invalid because its end-token grader was incorrect; the raw record and invalidation notice are retained. Corrected grading with shorter later-turn budgets gave 6/12. An interface with revised all-turn budgets reached 11/12 but rebuilt context without prior reasoning, unlike the formal live-cache engine. These inspected development configurations remain separate from held-out results and are not isolated causal ablations.

H Calibrated Transfer Measurements

H.1 Frozen comparison and timing boundaries

S4.13 selects four inventory endpoints per pause length by salted hash before observing transfer outcomes. All eight endpoints run four methods, three rate conditions, two warm-ups and five measured repeats, giving 672 records and 480 formal measurements. All formal measurements pass task, digest, restoration and anchor-integrity checks. Transfers use persistent IPv4 TCP loopback

Table 18: Recorded route-confirmation packets and head decisions ($n = 60$ tasks). MB means 10^6 bytes. Ratios are medians of task-paired byte ratios; coverage counts compact recurrent head instances. Non-recurrent only is a field control, not a serialized recovery method. DFD and SSR names follow Table 1; H denotes headwise fallback; n/a means not applicable.

Method	Packet MB	Paired / Full	Compact heads
Full	52.387	1.000	n/a
DFD-H	13.904	0.265	45,347/46,080
SSR-C-H	13.907	0.265	45,347/46,080

with application pacing and 262,144-byte chunks. Restoration deserializes received bytes, checked by SHA-256 digest and byte equality.

Full sends dense recurrent fields and the source non-recurrent suffix, using the resident target's common KV prefix. Its source and target execution omit affine capture; compact paths charge the capture required by their frozen constructors. The existing-anchor lifecycle measures source execution, record preparation, serialization, transfer, decoding, restoration and continuation. Setup-inclusive cost adds separately measured prefix prefill and local target setup, excluding initial target-anchor network transmission. This component sum is not an independently timed whole-wall interval. Extra state-error and diagnostic readout calls are excluded; integrity checks outside the lifecycle are labeled separately in the raw rows.

H.2 Pairing, uncertainty and observed variation

Five technical repeats are reduced to a median within each endpoint, rate and method, then paired across methods at each endpoint. The frozen primary comparison is SSR-C-H versus Full at 0.5 Gbit/s. Its 95% percentile interval resamples the eight endpoints 100,000 times with seed 41320260920. The paired ratio is 0.7988 with interval $[0.6837, 0.9363]$; all eight ratios are below one. A secondary two-sided exact sign test gives $p = 0.0078125$; other method and rate comparisons are descriptive. Uncertainty covers this templated endpoint selection, not variation across networks, machines or task families.

SSR-C-H and Full generate the same tokens at all eight endpoints, across all rates and repeats. DFD-H and SSR-I-H match Full on six and seven endpoints; complete output records retain the length differences. Directly paired SSR-C-H/DFD-H setup-cost ratios are 0.9865, 0.9414 and 0.9955 at 0.5 Gbit/s, nominal 2 Gbit/s and unlimited loopback.

An exploratory split of SSR-C-H/Full gives median ratios 0.6872 and 0.9175 for the 2- and 8-token groups at 0.5 Gbit/s. At nominal 2 Gbit/s they are 0.8214 and 1.0850, and under unlimited loopback 0.9761 and 1.1246. Each group contains four different tasks, rather than within-task length interventions. Nominal 2 Gbit/s achieves 1.71 to 1.84 Gbit/s; unlimited transfers achieve 2.12 to 2.23 Gbit/s, with the throughput bottleneck unresolved.

H.3 Anchor residency and scope

Median logical cache sizes are 55.116 MB for the fork and 55.280 MB each for source and target. Compact capture operands occupy 5.929 MB per branch; Full has none. Packet tensors occupy 52.068 MB for Full and 11.192 MB per compact method before framing. These are logical object sizes, not allocator peaks. S4.13 charges local setup and retains the anchor; source-factor layouts and repeated recovery with $q = 4$ were not run. Appendix P examines layouts and repeated delivery on a separate constructed workload.

I Runtime and Reproduction

S4.12, S4.13 and S4.14 use Qwen3.5-4B revision `851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a`. Runtime records identify Python 3.12.13, PyTorch 2.10.0+cu130, Transformers 5.14.1, CUDA 13.0 and RTX 5080 hardware; S4.12 and S4.13 also record driver 591.86. Provenance

Table 19: Calibrated TCP (Transmission Control Protocol) loopback at 0.5 Gbit/s, eight endpoints and five measured repeats per cell. Setup cost sums prefix, local anchor setup and measured lifecycle for one recovery; it excludes initial anchor transmission. Values are endpoint medians; ratios pair methods before aggregation. MB denotes 10^6 bytes and ms milliseconds. Method names follow Table 1; H denotes headwise fallback.

Method	Packet MB	Setup cost ms	Paired / Full
Full	52.321	2573.1	1.000
DFD-H	11.665	2064.4	0.792
SSR-I-H	11.671	2227.1	0.873
SSR-C-H	11.668	2020.4	0.799

Table 20: All calibrated transfer conditions. Setup cost and paired ratios use the same endpoint-first estimator as Table 19. Goodput is measured payload throughput, in Gbit/s. Unlimited means no application pacing. DFD and SSR names follow Table 1; H denotes headwise fallback.

Rate	Method	Setup ms	Paired / Full	Goodput
0.5 Gbit/s	Full	2573.1	1.000	0.481
0.5 Gbit/s	DFD-H	2064.4	0.792	0.493
0.5 Gbit/s	SSR-I-H	2227.1	0.873	0.493
0.5 Gbit/s	SSR-C-H	2020.4	0.799	0.493
2 Gbit/s	Full	1893.4	1.000	1.711
2 Gbit/s	DFD-H	2066.3	1.036	1.838
2 Gbit/s	SSR-I-H	2270.3	1.115	1.835
2 Gbit/s	SSR-C-H	1895.4	0.985	1.840
unlimited	Full	1836.7	1.000	2.115
unlimited	DFD-H	1916.4	1.001	2.228
unlimited	SSR-I-H	2260.9	1.146	2.231
unlimited	SSR-C-H	1994.6	1.036	2.225

remains study-specific, with missing S4.11 fields unfilled. The earlier S4.12 wrong-environment run is excluded.

The artifact contains anonymized outcomes, continuations, timings, head decisions, pilot records, numerical tables and runtime manifests. Historical checksums identify original bytes; exported checksums identify the path-redacted files. S4.14 uses a frozen wrapper around the S4.13 receiver. Its schema amendment maps retained S4.13 names to the numbered-field protocol. Complete snapshots are missing for the initial S4.13 pilot runner and an overwritten early transport smoke test.

The route studies use the model and runtime versions above. Their decoding engine has the same historical hash across the 24-task and 60-task freezes. The export contains all 336 method/task outputs, head decisions, frozen rosters, engine and runner snapshots, independent checks and development summaries. Original inventories verify 109 files for the initial study and 254 for confirmation before path redaction.

CPU-only integration scripts verify exported input hashes and rebuild tables; route grading is recomputed independently from graph definitions and saved outputs. Runner snapshots document execution but require path adaptation for model reproduction. The S4.11 model-reproduction entry point is separate from these analysis pipelines.

I.1 Precision and layout study provenance

S4.18 to S4.20 use Qwen3.5-4B revision `851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a`. Records include frozen rosters or queues, executed source snapshots, runtime manifests, analysis receipts and file hashes. The recorded environment is Python 3.12.13, PyTorch 2.10.0+cu130, Transformers 5.14.1, CUDA 13.0 and an NVIDIA GeForce RTX 5080, with driver 591.86 where recorded.

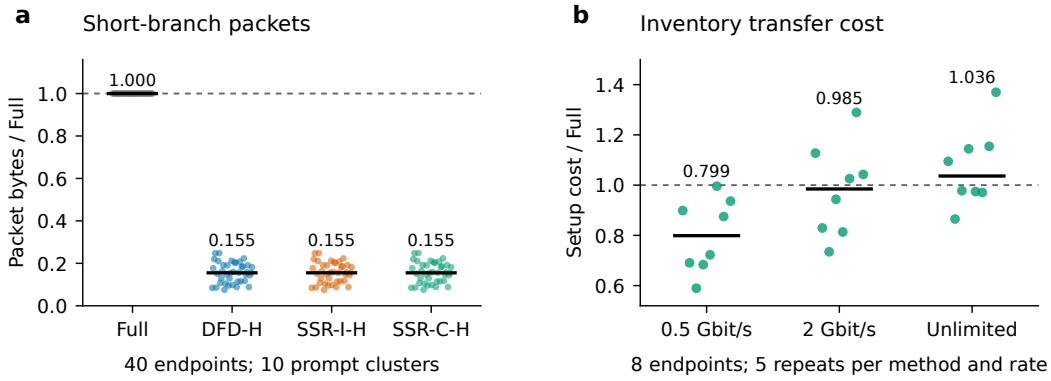

Figure 3: Short-branch packets and inventory transfer cost. **a**, Packet ratios from S4.11 (40 endpoints, 10 prompt clusters). **b**, SSR-C-H/Full component-cost ratios from S4.13 (eight inventory endpoints; five-repeat medians). Dots denote endpoints, bars and labels medians, and dashed lines Full parity. The panels use separate studies. Method names follow Table 1; H denotes headwise fallback.

J Supporting Encoding and Matrix Ablations

J.1 Matched operator study

The matrix ablation compares Full Snapshot, KSO, SSR-C and Clipped-Scalar on 40 short conditions from 10 prompt clusters and a 14-condition boundary set. KSO and SSR-C share the source, target and joint basis within each condition. Five warm-ups precede 20 repetitions; Tables 21 and 22 report absolute and paired results.

Table 21: Operator-ablation study on Qwen3.5-4B. Packet sizes, complete request times, and recurrent-state errors are medians. A_1 and A_{128} are exact agreements with the matched Full reference. The same 40 short conditions are used for every method. Abbreviations and metrics are defined in Table 6.

Method	Packet (MB)	Request (ms)	State error	A_1	A_{128}
Full Snapshot	52.24	391	0	100%	100.0%
KSO	8.44	1125	1.970×10^{-3}	100%	52.5%
SSR-C	8.12	642	1.505×10^{-3}	100%	50.0%
Clipped-Scalar	52.25	602	6.190×10^{-9}	100%	47.5%

The complete request includes preparation. These absolute medians are distinct from the paired ratio medians in Figure 4.

The largest relative reduction in Table 22 occurs in operand-ready construction.

In this matched study, the median endpoint ratio differs from the ratio of method medians:

$$\mathrm{median}_e(T_{R,e}/T_{K,e}) \neq \mathrm{median}_e(T_{R,e})/\,\mathrm{median}_e(T_{K,e}). \tag{54}$$

All four methods match all 40 first readouts. Exact 128-token counts are 40/40 for Full, 21/40 for KSO, 20/40 for SSR-C, and 19/40 for Clipped-Scalar. KSO/SSR joint counts are 14 both match, seven KSO-only, six SSR-only, and 13 both differ. The difference interval $[-0.20, 0.15]$ leaves their long-horizon ordering unresolved. Full-to-Full controls pass on all 54 short and boundary conditions.

Table 22: Reported paired median ratios and 95% prompt-cluster bootstrap intervals. Each ratio is formed under the same endpoint protocol before aggregation. Abbreviations and metrics are defined in Table 6.

Comparison	Metric	Median	95% interval
SSR-C / KSO	Serialized packet	0.962	[0.961, 0.967]
SSR-C / KSO	Operand-ready construction	0.178	[0.145, 0.225]
SSR-C / KSO	Restore + fresh readout	0.679	[0.675, 0.682]
SSR-C / KSO	Complete request	0.580	[0.573, 0.588]
KSO / Full	Complete request	2.94	[2.92, 3.02]
SSR-C / Full	Complete request	1.72	[1.64, 1.74]

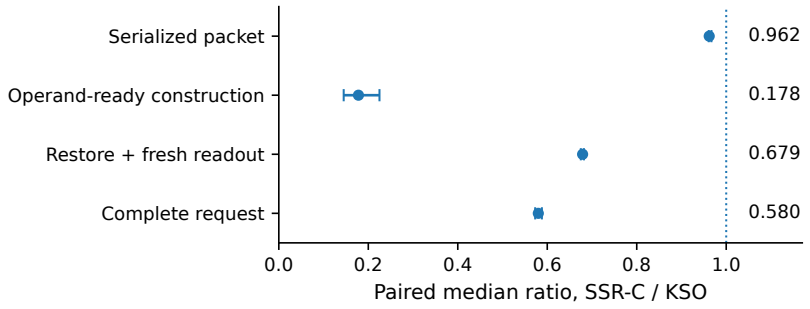

Figure 4: Earlier matrix-operator ablation: SSR-C/KSO paired median ratios with 95% prompt-cluster intervals. This is a separate matched experiment from the unified headwise study.

J.2 Earlier matrix packet scaling

Joint key dimensions are 16, 32, 64, and 128 at lengths 8, 16, 32, and 64, following the matrix field-count analysis. Table 9 reports the separate unified headwise comparison and its matched Full measurements.

J.3 Common-head residual-encoding diagnostic

Table 25 reports 12 conditions with a common admissible head set for DFD, SSR-I and SSR-C. Other heads use identical dense source fields. Coverage uses the supplied percentage, without reconstructing per-head weights or counts. Endpoint and source/target lengths identify each condition; reused endpoint labels do not add independent prompt clusters.

All three methods have 12/12 matching fresh top-1 readouts, and Full-to-Full controls are 12/12 exact. Complete-request medians include branch capture and prompt-prefix reconstruction. The B0-P manifest lacks an exact PyTorch binding; long-continuation outcomes and Full packet/time values are unavailable. The available timings are descriptive summaries, reported without inferred confidence intervals.

Table 23: Paired 128-token agreement with Full in the operator-ablation study. Abbreviations and metrics are defined in Table 6.

	SSR-C matches	SSR-C differs
KSO matches	14	7
KSO differs	6	13

SSR-C minus KSO agreement: -2.5 percentage points, with reported 95% prompt-cluster interval $[-20, 15]$ points.

Table 24: Separate branch-length boundary study. Packet columns are per-length medians in MB. The m column is the reported median common subspace dimension for KSO and SSR. Abbreviations and metrics are defined in Table 6.

Length	Conditions	m	Full (MB)	KSO (MB)	SSR-C (MB)
8	4	16	52.42	18.39	17.44
16	4	32	52.68	35.02	32.26
32	4	64	53.21	63.42	53.40
64	2	128	54.25	131.17	94.16

Table 25: All common-head diagnostic conditions. Lengths are source $\times$ target. Each method cell reports serialized MB / median diagnostic ms. Coverage is the reported common-head percentage, not a packet compression ratio. Abbreviations and metrics are defined in Table 6.

Input label	Lengths	Coverage	DFD	SSR-I	SSR-C
asym-03	4×1	100.0%	6.119 / 785.9	6.125 / 776.4	6.122 / 973.3
asym-07	16×4	100.0%	18.310 / 1532.7	18.317 / 1742.7	18.314 / 1551.8
asym-07	8×64	27.7%	54.231 / 3825.4	54.233 / 4449.9	54.232 / 4720.7
test-00	1×1	98.6%	4.357 / 1434.0	4.363 / 1648.0	4.360 / 1271.4
test-00	8×8	94.3%	17.052 / 1903.4	17.058 / 2048.8	17.055 / 1968.6
test-01	32×32	89.7%	53.419 / 4026.0	53.424 / 4304.4	53.421 / 4002.3
test-03	2×2	99.6%	5.451 / 830.3	5.457 / 865.7	5.454 / 739.7
test-04	1×1	97.4%	4.925 / 1390.5	4.931 / 1392.3	4.928 / 1492.2
test-06	4×4	94.7%	10.818 / 843.3	10.824 / 928.7	10.821 / 886.6
test-07	16×16	88.4%	30.639 / 1712.1	30.644 / 2840.0	30.642 / 2823.0
test-07	3×3	93.8%	9.709 / 989.2	9.715 / 874.8	9.712 / 1028.6
test-09	4×4	97.7%	9.506 / 963.9	9.512 / 1113.8	9.509 / 1058.3

All methods: fresh top-1 12/12. Full-to-Full: 12/12 exact. Times include the full diagnostic path, not only branch execution or restoration.

For method j and condition e, the displayed summaries are

$$\widetilde{b}_j = \operatorname{median}_e \frac{B_{j,e}}{B_{\mathrm{DFD,e}}}, \qquad \widetilde{t}_j = \operatorname{median}_e \frac{T_{j,e}}{T_{\mathrm{DFD,e}}}. \tag{55}$$

The inputs are rounded to 0.001 MB and 0.1 ms. Median packet increments are approximately 0.006 MB for SSR-I and 0.003 MB for SSR-C. The maximum relative increment is below 0.14% for both paths.

Size agreement persists in both 100%-compact conditions, where dense recurrent fallback cannot explain it. Complete packets still include non-recurrent fields; establishing $m = p$ requires actual factor counts, basis columns, dtypes and framing.

J.4 Complete-cache factor policies

The corrected cache-interface comparison uses 66 conditions: 40 short, 14 boundary and 12 asymmetric-length cases, all with passing Full-to-Full controls. Policies are Full Snapshot, DFD with endpoint-wide fallback (DFD-E), SSR-I with endpoint-wide fallback (SSR-I-E), and SSR-C with headwise fallback (SSR-C-H).

DFD-E and SSR-I-E use Full fallback on all 40 short endpoints. SSR-C-H uses headwise mixtures: median packet 8.114 MB, approximately 15.5% of Full, median request ratio 1.343, and 20/40 exact 128-token matches. These whole-policy measurements use the corrected cache interface, distinct from the operator-ablation request ratio of 1.72.

Endpoint-wide fallback can select Full despite local head admissibility. The unified study in Appendix E instead applies a common headwise policy.

Table 26: Complete-cache short-branch policy comparison. Counts refer to the 40 short conditions. E denotes endpoint-wide fallback and H headwise fallback. Abbreviations and metrics are defined in Table 6.

Policy	Fallback granularity	Observed execution
DFD-E	Endpoint	40/40 Full endpoints
SSR-I-E	Endpoint	40/40 Full endpoints
SSR-C-H	Head	Mixed encoding; 8.114 MB median packet; request ratio 1.343

K Branch-Record Cost Accounting

K.1 Residency, bindings, and alternatives

A record binds an immutable sibling, model/state format, source boundary, encoding and non-recurrent fields. An additional anchor consumes storage; changing or evicting it requires re-encoding dependent records or using another recovery path. Restoring an independent dense source also requires a runtime allocation.

A staging system can instead retain the common fork and store source factors. For N candidates, the two layouts have storage

$$B_{\text{sibling}} = B_{\text{anchor}} + \sum_{j=1}^{N-1} B_{\text{relative},j} + B_{\text{index}}, \tag{56}$$

$$B_{\text{fork}} = B_{\text{fork state}} + \sum_{j=1}^{N} B_{\text{source factors},j} + B_{\text{index}}. \tag{57}$$

Equation (57) accounts for shared fields consistently across both layouts; their choice depends on residency and recovery costs.

K.2 Single transfer and repeated recovery

For local cost C_j, serialized bytes B_j, and effective bandwidth w in bytes per second, the sequential model and crossing are

$$T_j(w) = C_j + B_j/w, \qquad w_j^* = \frac{B_F - B_j}{C_j - C_F} \quad (C_j > C_F,\ B_j < B_F). \tag{58}$$

The difference from Full is

$$T_j(w) - T_F(w) = (C_j - C_F) - (B_F - B_j)/w. \tag{59}$$

The crossing in (58) applies when $C_j > C_F$ and $B_j < B_F$. A larger, slower record never wins; a smaller, faster record needs no finite threshold. Startup latency and computation/transfer overlap require additional terms or pipeline measurements.

For unchanged bindings, let P_j be the one-time preparation cost and A_j the local cost per recovery. Then q recoveries give

$$T_j(q, w) = P_j + q(A_j + B_j/w), \qquad w_j^*(q) = \frac{q(B_F - B_j)}{(P_j - P_F) + q(A_j - A_F)} \tag{60}$$

when numerator and denominator are positive. Full Snapshot can also cache its prepared representation in this accounting model.

K.3 Earlier same-protocol numerical crossings

Projecting the matrix ablation's request costs and bytes gives median crossings of 0.474 Gbit/s for KSO and 1.326 Gbit/s for SSR-C. Their IQRs are $[0.392, 0.558]$ and $[1.010, 2.067]$; Clipped-Scalar has no favorable crossing on the 40 short conditions. These are operator-protocol projections; network latency was not measured.

L Interactive Cache Integration

L.1 Cache lineage and transport

The ALFWorld harness (Shridhar et al., 2021) checks cache continuity across model requests. The receiver decodes received bytes, applies a nonzero SSR-C residual and continues through two tool-observation appends without rebuilding the prefix. A perturbed-residual control confirms that later inference consumes the modified state. Independent caches using identical token chunks match across 64 tensors and logits; whole-prefix and incremental execution use different numerical schedules.

Transport uses user-space rate-limited TCP loopback. Table 27 gives the corrected 16 MiB calibration, excluding the initial rate-unit-error measurements. The 5 ms delay is simulated; the offered 10 Gbit/s rate is host-limited.

L.2 Interactive outcomes and reference replay

The integration suite runs 12 previously used tasks under four policies, giving 48 normally closed episodes. All methods share pre-handoff actions and source/target candidate tokens. Table 28 reports policy outcomes and actual compact use.

DFD-E and SSR-I-E succeed on the same task as Full, both through endpoint-wide Full fallback. SSR-C-H uses compact heads on all 12 tasks. A token-level top-2 rule selects source candidates; the parser accepts arbitrary single-line text, including observation echoes, reasoning text and control-style strings.

The unmodified policy and in-place continuation from newly generated candidates each succeed on 0/12 tasks; these candidates differ from the initial run. A stricter replay fixes the initial two actions and source tokens. All 12 boundary observations match, but every run reaches the 50-step cap without success. Seven trajectories match the original Full run; five diverge despite matching visible actions and feedback. Missing cross-run logits and state snapshots prevent numerical attribution.

In one initial Pencil-task pair, Full and SSR-C-H share five requests of actions and feedback, then diverge on the sixth. The unstable Full reference limits this study to cache integration, without evidence of task-quality or latency gains. No official `valid_unseen` or separate study-heldout evaluation was run.

M Complementary Shared-Base Evidence

M.1 Representation frontier

The Qwen3.5-4B shared-base comparison uses BF16 weights and compute on one RTX 5080. Its 240-condition aggregate includes 40 implementation-design and 200 assessment conditions. Seven methods span lengths 1, 2, 4, 8, 16, and 32, producing 1,680 rows (Table 29). Source factors, token replay and KV-plus-convolution lie on the state-cost frontier at every endpoint. KV-plus-convolution omits recurrent correction and has lower readout agreement, so frontier membership reflects a cost/fidelity tradeoff. Figure 5 shows the fixed-rank error trend.

M.2 Decoded packets, natural branches, and factual updates

The decoded-packet handoff comparison uses 30 endpoints, three warm-ups and ten synchronized repetitions. Full Snapshot, source factors and target-difference factors have packet/time medians of 52.12 MB/119.5 ms, 4.16 MB/94.6 ms and 6.52 MB/105.4 ms. Token replay takes 231.2 ms. Their P95 times are 143.0, 151.0, 193.1 and 1,174.9 ms, respectively, measured without an external arrival process.

The shared-base transfer model gives source-factor and Full costs of 125.6 and 530.9 ms at 1 Gbit/s. At 10 Gbit/s, the corresponding costs are 95.6 and 155.5 ms. Measured target-only transfers and layout comparisons appear in Appendices H and P.

Table 27: User-space loopback TCP calibration with a 16 MiB payload. Goodput is measured; the offered rate is a configuration value. Abbreviations and metrics are defined in Table 6.

Offered (Gbit/s)	Measured goodput (Gbit/s)
0.25	0.251
1	0.999
10	2.261

The 10 Gbit/s offer is host-limited. The 5 ms added delay is simulated, not a physical round-trip measurement.

Table 28: Initial interactive integration outcomes on the same 12 tasks. These policy outcomes are followed by the reference-replay diagnosis in the text. Abbreviations and metrics are defined in Table 6.

Policy	Task successes	Actual representation
Full Snapshot	1/12	Full Snapshot
DFD-E	1/12	3/12 whole-endpoint compact; 9/12 Full fallback
SSR-I-E	1/12	3/12 whole-endpoint compact; 9/12 Full fallback
SSR-C-H	0/12	12/12 with compact heads

The sole DFD-E and SSR-I-E success used Full fallback. The original Full success was not reproduced by the fixed-source in-place replay.

At 100 model-proposed endpoints immediately after an alternative token, source and target-difference factors preserve all first readouts. Their packet medians are 2.00 and 2.40 MB, versus 90.27 MB for Full.

The shared-base factual-update study uses 900 subject-unique MQuAKE-derived cases (Zhong et al., 2023) and Qwen3.5-9B correction branches with target-relative rank-2 records. These records and Full each answer 894/900 correctly, with identical generated outputs; KV-plus-convolution answers 889/900 correctly. Correction-state tensor payload is 4.62% of the dense reference. Serializer, host-transfer and receiver-application checks preserve the output matches.

M.3 Readout diagnostics and cross-model state sizes

On 200 Qwen endpoints, the reference-based margin diagnostic has AUC 0.9223 with cluster interval $[0.8957, 0.9454]$, versus 0.7937 for relative state error. Computing this local diagnostic requires access to the reference output.

The Falcon-H1-1.5B diagnostic completes 183 of 300 conditions; 117 fail an unequal-span structural check. Complete-case AUC is 0.9354 with a reported lower bound of 0.9135. A separate 96-condition study uses 64-token continuations and a fixed 5% wire cap; source and target-difference factors match Full on 48 and 51.

Factor count scales as $O(n(d_k + d_v))$ per head, versus $O(d_k d_v)$ for dense state. Table 30 computes sizes from recorded dimensions; configuration-only rows involve no model execution. Short Falcon branches report 2.08% to 4.23% tensor-payload ratios, maximum source-state error below 0.278%, and matching probes.

M.4 Matrix reconstruction on two GPUs

The matrix-KSO platform comparison uses 24 conditions on two T4 GPUs, with descriptors prepared before the measured transfer. At lengths 1 to 3, packet, host-staged handoff and CUDA P2P ratios to matched Full paths are 0.1583, 0.8813 and 0.9025. All short fresh readouts match, with maximum relative state error below 5.2×10^{-9}. At lengths 59 to 64, packet and handoff ratios reach approximately 1.93 and 2.07 to 2.08.

Unbounded scalar-ratio replay produces non-finite logits on eight length-63/64 T4 conditions and two long conditions in a separate confirmation record. The clipped-scalar path has no reported instances of this failure. CPU FP64 descriptor construction has median cost 733.0 ms; later GPU-reduced, incremental and paired operator-ablation timings use different protocols.

Table 29: Shared-base aggregate over 240 conditions: 40 for implementation design and 200 for assessment. Packet sizes are serialized medians in MB. Beh. Pareto denotes the fraction of endpoints on the recorded behavior-cost Pareto frontier; A_1 is first-token agreement. Abbreviations and metrics are defined in Table 6.

Representation	Packet (MB)	State error	A_1	Beh. Pareto
Full Snapshot	52.60	0	100.0%	0.967
Source-Factor	4.13	5.260×10^{-5}	97.5%	1.000
Target-Difference	6.49	7.540×10^{-5}	95.4%	0.438
Source-Factor rank 2	2.56	2.200×10^{-3}	95.0%	0.558
Target-Difference rank 2	2.56	3.970×10^{-3}	93.8%	0.138
KV + convolution only	1.77	1.150×10^{-2}	80.8%	1.000

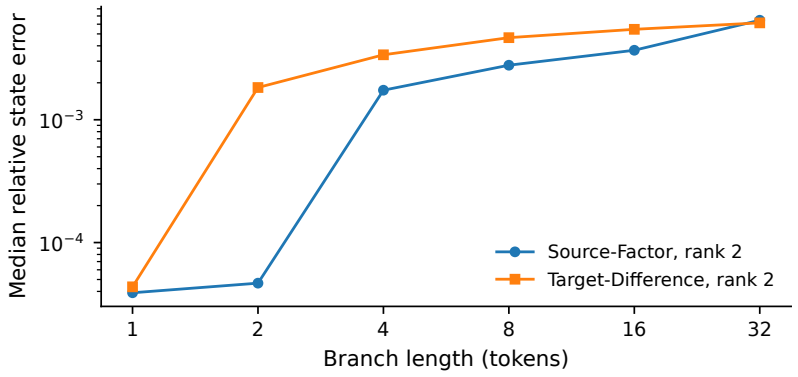

Figure 5: Shared-base rank-2 state error versus branch length. These are the reported full-suite medians, separate from the target-only operator ablation.

KSO retained-memory and post-restore ratios are 0.583 and 1.000, measured at different boundaries rather than over one matched lifecycle. An incomplete allocation ledger gives 1.502, without measuring a system peak. The 39-condition KSO continuation observation of approximately 59% also predates the paired counts in Table 23. A separate fixed per-head INT8 Full ablation gives approximately 7.9% state error, 60% first-readout agreement, 1.3% at 32 tokens, and 0% at 64 and 128 tokens. This INT8 result is unpaired with the operator ablation.

Table 30: Recurrent-component size accounting. Dense state is in MB; SF denotes Source-Factor, and n is branch length. Ratios exclude non-recurrent fields. Config. identifies configuration-derived sizes; Records identifies supporting execution records. Abbreviations and metrics are defined in Table 6.

Model	Dense state	SF $n = 1$	SF $n = 4$	Source
Falcon-H1-0.5B	14.16	2.356%	9.387%	Records + config.
Falcon-H1-1.5B	37.75	1.959%	7.819%	Records + config.
Falcon-H1-3B	67.11	1.175%	4.691%	Config.
Qwen3.5-4B	50.33	0.784%	3.128%	Records + config.
Qwen3.5-9B	50.33	0.784%	3.128%	Config.

N Dense Precision Frontier

This study compares record size, request cost and quality under the original capture implementation. Corrected-capture measurements are reported separately in Appendix Q.1.

S4.18 revisits 40 existing short endpoints in ten prompt clusters, comparing Full, dense FP16, dense BF16, DFD-H and SSR-C-H. Local request time includes branch execution, capture, construction, serialization, decoding, restoration, fresh readout and common diagnostic bookkeeping. It excludes network transfer and a timed 128-token continuation. Each method uses five warm-ups and 20 formal timing repetitions. These are technical repeats; analysis uses endpoints with cluster-preserving resampling.

The frozen primary contrast, SSR-C-H versus dense BF16, gives a median paired request-time ratio of 1.743 and cluster-bootstrap 95% interval $[1.695, 1.774]$. The median packet ratio is 0.299, or 29.9%: compact records are smaller but slower under this local request boundary. At 128 generated tokens, SSR-C-H and dense BF16 match the Full trajectory on 20 and 24 of 40 endpoints. The paired agreement-rate difference is -0.10, with descriptive cluster-bootstrap interval $[-0.25, 0.075]$. Table 31 reports all five methods at this horizon; other recorded horizons remain in the artifact.

The existing 128 register cases were regraded with the frozen S4.14 extracted-field grader as primary and strict whole-output grading as secondary. Both give the same counts: Full 116, dense FP16 117, dense BF16 116, DFD-H 117 and SSR-C-H 118. These descriptive results revisit an exposed roster, rather than a new held-out sample. The 128 cases are the independent units; horizons and method evaluations within a case are repeated measurements.

Table 31: Complete packets and local requests on 40 short endpoints.

Method	Packet (MB)	/Full	Request (ms)	Request/Full [95% interval]	A_{128}
Full FP32	52.241	1.000	476.9	1.000 (reference)	40/40
Dense FP16	27.085	0.518	372.9	0.787 [0.768, 0.809]	20/40
Dense BF16	27.089	0.519	394.2	0.816 [0.799, 0.825]	24/40
DFD-H	8.114	0.155	701.9	1.468 [1.442, 1.493]	17/40
SSR-C-H	8.117	0.155	674.8	1.398 [1.386, 1.411]	20/40

MB is 10^6 bytes; ms is milliseconds. Absolute values are medians across endpoints. Ratios pair methods within endpoints; brackets are 95% prompt-cluster bootstrap intervals (10 clusters, 10,000 resamples). A_{128} counts exact agreement through 128 predictions, including stopping. FP32, FP16 and BF16 are defined in Section 4; DFD and SSR in Table 1. H denotes headwise fallback.

Table 32: Exact task quality on 128 frozen S418 register-field tests.

Method	Primary exact	Whole-output exact	Loss among Full successes	Median packet (MB)
Full FP32	116/128	116/128	0/116	52.32
Dense FP16	117/128	117/128	1/116	27.17
Dense BF16	116/128	116/128	1/116	27.17
DFD-H	117/128	117/128	1/116	10.90
SSR-C-H	118/128	118/128	0/116	10.90

Counts are exact outcomes, not an equivalence test. "Loss among Full successes" uses the 116 cases solved by Full FP32 as its fixed denominator; MB is 10^6 bytes. Full FP32, Dense FP16, Dense BF16, DFD-H and SSR-C-H are defined in Table 31.

O Bounded Recurrent-Sensitivity Development

S4.19 tested whether retaining conflicting sibling recurrent state affected task outputs after all source non-recurrent fields had been restored. Its frozen 12-case development roster crossed selected branch, zero, two or four archived-card distractors, and two replicates. Full, DFD-H, SSR-C-H, non-recurrent-only and sibling-only were evaluated on every case; all recorded runtime audits passed.

Full and non-recurrent-only each solved 6/12 tasks, with one Full-only and one non-recurrent-only success. Full therefore failed the prespecified 11/12 stability gate, and the separately frozen 64-case formal roster was not run.

Post-hoc inspection identified outputs `Branch A` or `Branch B` where the grader expects `A` or `B`. The frozen scores and failed gate were retained.

P Reference-Layout Comparison

S4.20 compares sibling-referenced DFD-H, dense Full and fork-referenced source factors on one constructed workload. Stage A crosses branch lengths 2 and 8 with candidate counts $N \in \{2, 4, 8\}$, giving six dependent conditions. It records 936 rows, including 702 measurements from actual paced TCP transfers. Initial base-state transfer is unrecorded, so Stage A conditions on that base already being resident.

Under the external-sibling contract, sibling A is resident and candidate B is selected. Source factors require an additional fork allocation, exceeding the same one-base budget. Under the cold contract, no base is initially resident and the complete $N = 2$ candidate collection is charged. These residency contracts are evaluated separately.

Stage B adds complete cache codecs and base-state wire costs: 441 rows include 315 formal measurements of the same workload, not independent workload samples. For $q \in \{1, 4, 8\}$, each required record is prepared once and the same prepared packet set is delivered every round, without receiver-side packet reuse. The set contains B's packet under the external-sibling contract and the packets for A and B under the cold contract. Full reuses capture-free prepared records; DFD-H charges capture and preparation for both branches when required. Full restores the boundary cache exactly, while DFD-H and source factors approximate the native state before fresh readout.

Table 33: Static S420 packet totals and resident full-state bases.

n	N	Layout	Packet-only total (MB)	Fork base (MB)	Sibling base (MB)
2	2	fork_FULL	104.45	55.12	55.18
2	2	fork_SOURCE-FACTOR	7.39	55.12	55.18
2	2	sibling_DFD-H	6.12	55.12	55.18
2	2	sibling_FULL	52.22	55.12	55.18
2	4	fork_FULL	208.90	55.12	55.18
2	4	fork_SOURCE-FACTOR	14.78	55.12	55.18
2	4	sibling_DFD-H	18.00	55.12	55.18
2	4	sibling_FULL	156.67	55.12	55.18
2	8	fork_FULL	417.80	55.12	55.18
2	8	fork_SOURCE-FACTOR	29.55	55.12	55.18
2	8	sibling_DFD-H	42.56	55.12	55.18
2	8	sibling_FULL	365.57	55.12	55.18
8	2	fork_FULL	104.84	55.12	55.38
8	2	fork_SOURCE-FACTOR	17.22	55.12	55.38
8	2	sibling_DFD-H	15.68	55.12	55.38
8	2	sibling_FULL	52.42	55.12	55.38
8	4	fork_FULL	209.68	55.12	55.38
8	4	fork_SOURCE-FACTOR	34.44	55.12	55.38
8	4	sibling_DFD-H	46.91	55.12	55.38
8	4	sibling_FULL	157.26	55.12	55.38
8	8	fork_FULL	419.37	55.12	55.38
8	8	fork_SOURCE-FACTOR	68.88	55.12	55.38
8	8	sibling_DFD-H	110.18	55.12	55.38
8	8	sibling_FULL	366.95	55.12	55.38

$n \in \{2, 8\}$ is actual branch length and $N \in \{2, 4, 8\}$ is actual candidate count. Packet-only totals exclude the resident base; the final two columns report full base-state memory separately. FULL is full-state transfer, SOURCE-FACTOR (SF) is Source-Factor, and DFD-H is headwise Direct Factor Difference.

Table 34: All Stage-B paired cost ratios by contract, round count and rate.

Population	q	Rate	DFD/Full	SF/Full
Resident A / restored B	1	0.5 Gbit/s	0.813	1.169
Resident A / restored B	1	2 Gbit/s	1.168	1.158
Resident A / restored B	1	unlimited	1.201	1.174
Resident A / restored B	4	0.5 Gbit/s	0.480	0.572
Resident A / restored B	4	2 Gbit/s	0.851	0.845
Resident A / restored B	4	unlimited	0.946	0.865
Resident A / restored B	8	0.5 Gbit/s	0.385	0.415
Resident A / restored B	8	2 Gbit/s	0.735	0.687
Resident A / restored B	8	unlimited	0.744	0.753
Cold $N = 2$	1	0.5 Gbit/s	0.853	0.847
Cold $N = 2$	1	2 Gbit/s	1.105	1.096
Cold $N = 2$	1	unlimited	1.068	1.097
Cold $N = 2$	4	0.5 Gbit/s	0.571	0.651
Cold $N = 2$	4	2 Gbit/s	0.828	1.036
Cold $N = 2$	4	unlimited	0.850	1.035
Cold $N = 2$	8	0.5 Gbit/s	0.451	0.554
Cold $N = 2$	8	2 Gbit/s	0.752	0.977
Cold $N = 2$	8	unlimited	0.811	1.022

Each entry is the median of five matched per-repeat setup-inclusive component-sum ratios from one constructed workload. Technical repeats are not independent and no inferential interval is reported. Under the resident-sibling contract, A is resident and B is restored from the source; Cold $N = 2$ is the cold candidate collection. DFD, SF and Full are defined in Table 5.

For one recovery at 0.5 Gbit/s under the external-sibling contract, the DFD-H/Full and source-factor/Full component-sum ratios are 0.813 and 1.169. Under unlimited loopback, the DFD-H/Full ratio is 1.201. Under the cold $N = 2$ contract at 0.5 Gbit/s, the corresponding DFD-H and source-factor ratios are 0.853 and 0.847. The all-rate and all-q results are reported in Table 34.

Table 35 reports measured components whose sum defines total cost, rather than an independently timed whole-wall interval. The q-round sum reuses the single preparation measurement from the actual run.

Table 35: Stage-B Preparation/base components: Resident A / restored B.

q	Rate	Path	Prefix	Branch prep.	Base enc.	Base wire	Base dec.
1	0.5 Gbit/s	Full	541.9	77.0	0.0	0.0	0.0
1	0.5 Gbit/s	DFD	541.9	198.9	0.0	0.0	0.0
1	0.5 Gbit/s	SF	541.9	73.2	107.6	920.4	77.2
1	2 Gbit/s	Full	541.9	77.0	0.0	0.0	0.0
1	2 Gbit/s	DFD	541.9	198.9	0.0	0.0	0.0
1	2 Gbit/s	SF	541.9	73.2	107.6	262.5	77.9
1	unlimited	Full	541.9	77.0	0.0	0.0	0.0
1	unlimited	DFD	541.9	198.9	0.0	0.0	0.0
1	unlimited	SF	541.9	73.2	107.6	222.8	77.5
4	0.5 Gbit/s	Full	541.9	77.0	0.0	0.0	0.0
4	0.5 Gbit/s	DFD	541.9	198.9	0.0	0.0	0.0
4	0.5 Gbit/s	SF	541.9	73.2	107.6	919.6	77.2
4	2 Gbit/s	Full	541.9	77.0	0.0	0.0	0.0
4	2 Gbit/s	DFD	541.9	198.9	0.0	0.0	0.0
4	2 Gbit/s	SF	541.9	73.2	107.6	262.7	78.3
4	unlimited	Full	541.9	77.0	0.0	0.0	0.0
4	unlimited	DFD	541.9	198.9	0.0	0.0	0.0
4	unlimited	SF	541.9	73.2	107.6	218.8	79.6
8	0.5 Gbit/s	Full	541.9	77.0	0.0	0.0	0.0
8	0.5 Gbit/s	DFD	541.9	198.9	0.0	0.0	0.0
8	0.5 Gbit/s	SF	541.9	73.2	107.6	921.1	76.7
8	2 Gbit/s	Full	541.9	77.0	0.0	0.0	0.0
8	2 Gbit/s	DFD	541.9	198.9	0.0	0.0	0.0
8	2 Gbit/s	SF	541.9	73.2	107.6	259.1	78.1
8	unlimited	Full	541.9	77.0	0.0	0.0	0.0
8	unlimited	DFD	541.9	198.9	0.0	0.0	0.0
8	unlimited	SF	541.9	73.2	107.6	221.5	77.5

All entries are median milliseconds over five technical repeats. Full is sibling full-state transfer, Fork Full is fork-layout full-state transfer, DFD is Direct Factor Difference, and SF is Source-Factor.

Table 36: Stage-B Preparation/base components: Cold $N = 2$.

q	Rate	Path	Prefix	Branch prep.	Base enc.	Base wire	Base dec.
1	0.5 Gbit/s	Full	541.9	144.4	102.8	918.4	77.8
1	0.5 Gbit/s	Fork Full	541.9	144.4	107.6	920.7	79.7
1	0.5 Gbit/s	DFD	541.9	198.9	102.8	920.2	77.9
1	0.5 Gbit/s	SF	541.9	198.9	107.6	918.3	79.1
1	2 Gbit/s	Full	541.9	144.4	102.8	259.9	76.8
1	2 Gbit/s	Fork Full	541.9	144.4	107.6	258.3	78.7
1	2 Gbit/s	DFD	541.9	198.9	102.8	259.3	76.6
1	2 Gbit/s	SF	541.9	198.9	107.6	259.2	78.2
1	unlimited	Full	541.9	144.4	102.8	223.5	82.8
1	unlimited	Fork Full	541.9	144.4	107.6	223.5	79.0
1	unlimited	DFD	541.9	198.9	102.8	218.7	48.3
1	unlimited	SF	541.9	198.9	107.6	221.2	78.2
4	0.5 Gbit/s	Full	541.9	144.4	102.8	923.1	79.1
4	0.5 Gbit/s	Fork Full	541.9	144.4	107.6	918.5	80.3
4	0.5 Gbit/s	DFD	541.9	198.9	102.8	918.5	78.4
4	0.5 Gbit/s	SF	541.9	198.9	107.6	918.5	82.3
4	2 Gbit/s	Full	541.9	144.4	102.8	261.7	78.6
4	2 Gbit/s	Fork Full	541.9	144.4	107.6	259.5	78.9
4	2 Gbit/s	DFD	541.9	198.9	102.8	259.9	78.5
4	2 Gbit/s	SF	541.9	198.9	107.6	259.9	78.6
4	unlimited	Full	541.9	144.4	102.8	221.1	78.5
4	unlimited	Fork Full	541.9	144.4	107.6	222.0	76.9
4	unlimited	DFD	541.9	198.9	102.8	217.7	48.1
4	unlimited	SF	541.9	198.9	107.6	217.0	78.6
8	0.5 Gbit/s	Full	541.9	144.4	102.8	921.2	81.9
8	0.5 Gbit/s	Fork Full	541.9	144.4	107.6	919.8	80.0
8	0.5 Gbit/s	DFD	541.9	198.9	102.8	918.8	80.6
8	0.5 Gbit/s	SF	541.9	198.9	107.6	918.8	77.8
8	2 Gbit/s	Full	541.9	144.4	102.8	260.2	77.0
8	2 Gbit/s	Fork Full	541.9	144.4	107.6	261.9	80.0
8	2 Gbit/s	DFD	541.9	198.9	102.8	260.2	77.4
8	2 Gbit/s	SF	541.9	198.9	107.6	258.2	81.6
8	unlimited	Full	541.9	144.4	102.8	223.1	76.7
8	unlimited	Fork Full	541.9	144.4	107.6	220.1	78.5
8	unlimited	DFD	541.9	198.9	102.8	218.0	77.1
8	unlimited	SF	541.9	198.9	107.6	218.1	84.2

All entries are median milliseconds over five technical repeats. Full is sibling full-state transfer, Fork Full is fork-layout full-state transfer, DFD is Direct Factor Difference, and SF is Source-Factor.

Table 37: Stage-B Packet/restoration components: Resident A / restored B.

q	Rate	Path	Build	Serialize	Wire	Decode	Restore	Fresh fwd.	Total
1	0.5 Gbit/s	Full	56.7	109.9	868.5	62.9	32.3	32.2	1786.4
1	0.5 Gbit/s	DFD	272.5	155.1	99.3	47.2	98.5	37.9	1451.8
1	0.5 Gbit/s	SF	97.3	33.5	60.3	46.6	94.6	36.5	2089.1
1	2 Gbit/s	Full	56.7	109.9	246.0	63.0	32.4	39.0	1167.2
1	2 Gbit/s	DFD	272.5	155.1	26.6	47.7	94.7	38.7	1370.5
1	2 Gbit/s	SF	97.3	33.5	16.2	45.6	57.3	32.3	1346.1
1	unlimited	Full	56.7	109.9	213.4	63.6	30.5	36.0	1131.7
1	unlimited	DFD	272.5	155.1	23.3	47.0	92.2	38.0	1370.6
1	unlimited	SF	97.3	33.5	14.0	46.3	95.8	36.8	1338.1
4	0.5 Gbit/s	Full	56.7	109.9	3490.0	256.5	120.6	143.1	4791.5
4	0.5 Gbit/s	DFD	272.5	155.1	397.1	189.7	350.4	141.0	2300.1
4	0.5 Gbit/s	SF	97.3	33.5	241.8	185.5	336.5	129.4	2738.6
4	2 Gbit/s	Full	56.7	109.9	994.7	252.1	125.2	131.6	2290.5
4	2 Gbit/s	DFD	272.5	155.1	106.7	192.1	315.1	130.5	1984.1
4	2 Gbit/s	SF	97.3	33.5	64.9	185.2	324.0	129.8	1935.5
4	unlimited	Full	56.7	109.9	829.5	252.9	113.9	134.8	2152.9
4	unlimited	DFD	272.5	155.1	92.4	189.5	389.3	154.0	1992.2
4	unlimited	SF	97.3	33.5	56.4	185.9	276.8	130.9	1821.3
8	0.5 Gbit/s	Full	56.7	109.9	6972.9	512.0	221.7	271.4	8786.7
8	0.5 Gbit/s	DFD	272.5	155.1	793.4	383.2	738.2	281.1	3364.3
8	0.5 Gbit/s	SF	97.3	33.5	483.3	487.6	648.7	253.0	3630.0
8	2 Gbit/s	Full	56.7	109.9	1964.7	507.8	232.1	270.9	3804.4
8	2 Gbit/s	DFD	272.5	155.1	213.0	500.5	734.8	294.6	2796.0
8	2 Gbit/s	SF	97.3	33.5	129.4	375.7	558.3	249.2	2622.7
8	unlimited	Full	56.7	109.9	1666.0	519.0	286.2	289.1	3558.9
8	unlimited	DFD	272.5	155.1	182.2	381.5	706.7	283.2	2723.0
8	unlimited	SF	97.3	33.5	111.9	490.1	713.7	284.7	2670.0

All entries are median milliseconds over five technical repeats. Total is the median protocol-defined component sum; it need not equal the sum of the separately displayed component medians and is not a whole-wall timer. DFD is Direct Factor Difference and SF is Source-Factor.

Table 38: Stage-B Packet/restoration components: Cold $N = 2$.

q	Rate	Path	Build	Serialize	Wire	Decode	Restore	Fresh fwd.	Total
1	0.5 Gbit/s	Full	56.7	109.9	869.4	64.1	32.4	71.9	2987.7
1	0.5 Gbit/s	Fork Full	127.3	205.2	1745.4	131.4	56.1	70.1	4237.5
1	0.5 Gbit/s	DFD	272.5	155.1	99.0	46.9	56.4	64.6	2545.2
1	0.5 Gbit/s	SF	203.0	67.2	120.9	92.1	131.0	65.8	2528.3
1	2 Gbit/s	Full	56.7	109.9	244.5	63.1	27.3	73.8	1698.1
1	2 Gbit/s	Fork Full	127.3	205.2	488.6	127.4	71.0	74.5	2228.6
1	2 Gbit/s	DFD	272.5	155.1	26.6	46.5	98.8	75.9	1859.1
1	2 Gbit/s	SF	203.0	67.2	32.3	91.2	197.4	76.7	1852.5
1	unlimited	Full	56.7	109.9	211.9	62.9	39.7	73.2	1644.3
1	unlimited	Fork Full	127.3	205.2	418.2	127.2	65.0	73.6	2115.0
1	unlimited	DFD	272.5	155.1	22.9	47.0	63.7	68.4	1748.4
1	unlimited	SF	203.0	67.2	28.1	91.8	194.6	76.2	1809.9
4	0.5 Gbit/s	Full	56.7	109.9	3486.6	253.1	110.1	257.3	6060.9
4	0.5 Gbit/s	Fork Full	127.3	205.2	6961.5	518.4	243.5	289.4	10249.7
4	0.5 Gbit/s	DFD	272.5	155.1	396.6	189.3	336.0	281.5	3482.5
4	0.5 Gbit/s	SF	203.0	67.2	483.9	484.9	606.9	259.4	3952.3
4	2 Gbit/s	Full	56.7	109.9	984.8	256.6	127.9	289.9	2964.6
4	2 Gbit/s	Fork Full	127.3	205.2	1973.0	511.5	260.4	273.8	4497.9
4	2 Gbit/s	DFD	272.5	155.1	106.6	189.4	318.9	231.1	2493.4
4	2 Gbit/s	SF	203.0	67.2	129.7	485.0	678.9	271.1	3009.0
4	unlimited	Full	56.7	109.9	836.3	257.2	127.9	255.3	2753.5
4	unlimited	Fork Full	127.3	205.2	1664.8	511.3	238.5	272.4	4158.1
4	unlimited	DFD	272.5	155.1	90.8	189.9	284.9	246.5	2342.2
4	unlimited	SF	203.0	67.2	112.5	371.3	606.6	246.2	2836.1
8	0.5 Gbit/s	Full	56.7	109.9	6975.8	510.8	230.3	531.1	10215.7
8	0.5 Gbit/s	Fork Full	127.3	205.2	13945.6	1024.4	475.0	594.0	18228.8
8	0.5 Gbit/s	DFD	272.5	155.1	793.3	379.5	625.1	525.7	4600.5
8	0.5 Gbit/s	SF	203.0	67.2	967.4	852.8	1197.8	527.5	5648.2
8	2 Gbit/s	Full	56.7	109.9	1970.4	511.6	257.2	493.8	4526.3
8	2 Gbit/s	Fork Full	127.3	205.2	3930.6	1024.8	541.7	573.5	7643.2
8	2 Gbit/s	DFD	272.5	155.1	213.2	501.0	659.1	514.1	3452.6
8	2 Gbit/s	SF	203.0	67.2	259.2	854.9	1282.2	533.7	4419.4
8	unlimited	Full	56.7	109.9	1665.9	509.6	277.9	570.4	4276.3
8	unlimited	Fork Full	127.3	205.2	3329.5	1141.0	507.1	557.1	6933.4
8	unlimited	DFD	272.5	155.1	181.4	382.2	736.3	563.9	3446.8
8	unlimited	SF	203.0	67.2	225.2	852.2	1527.8	602.4	4517.3

All entries are median milliseconds over five technical repeats. Total is the median protocol-defined component sum; it need not equal the sum of the separately displayed component medians and is not a whole-wall timer. DFD is Direct Factor Difference and SF is Source-Factor.

Q CORRECTED CAPTURE AND CONTINUATION

All studies used Qwen3.5-4B and headwise fallback with $\gamma = 2$.

Q.1 CORRECTED CAPTURE FRONTIER

The historical corrected frontier promoted the key to FP32 before native normalization and consistently recomputed the associated read and innovation. The query, gates, values, packet format and head gate were unchanged. The interrupted run completed 52 of 54 planned endpoints: all 40 short endpoints from ten prompt clusters and 12 boundary endpoints from five overlapping clusters. The two missing length-64 endpoints were rerun separately and excluded from this timing cohort. Each method used two warm-ups and five formal timing repeats, with endpoints grouped by prompt cluster.

Table 39: **Corrected short-branch frontier.** Medians cover 40 historical endpoints in ten prompt clusters. MB denotes 10^6 bytes; request ms excludes network transfer; state error is relative L_2; A_{128} is exact agreement with Full through 128 tokens. H denotes headwise fallback.

Method	Packet (MB)	Request (ms)	State rel. L_2	A_{128}
Full FP32	52.239	404.5	0	40/40
Dense FP16	27.083	332.2	2.07×10^{-4}	20/40
Dense BF16	27.087	348.3	1.66×10^{-3}	24/40
DFD-H (legacy)	8.111	674.2	1.69×10^{-3}	17/40
DFD-H (corrected)	8.111	650.3	1.27×10^{-7}	20/40

Table 39 reports the complete short cohort. Corrected DFD retained the 8.111-MB median packet of legacy DFD while reducing median recurrent-state relative error from 1.69×10^{-3} to 1.27×10^{-7}. Its paired request-time ratios were 1.862 versus BF16 (cluster-bootstrap interval $[1.816, 1.956]$) and 1.920 versus FP16 ($[1.889, 2.092]$), with 0/40 faster endpoints in either comparison. Both intervals use 10,000 resamples of the ten short-cohort prompt clusters with seed 43220260923. At A_{128}, corrected DFD matched Full on 20/40 endpoints, compared with 24/40 for BF16 and 20/40 for FP16.

Across the 12 available boundary endpoints, the corresponding request ratios were 1.483 and 1.579, again with 0/12 faster endpoints. Corrected DFD had median packet size 32.212 MB, request time 1750.9 ms, state error 2.99×10^{-7} and 8/12 exact A_{128} trajectories. At lengths 8, 16 and 32 (four endpoints each), its BF16-paired request ratios were 1.592, 1.512 and 1.407, while packet ratios were 0.640, 1.170 and 1.903. These descriptive results cover 12 of the planned 14 boundary endpoints, without an interval or length-64 estimate.

Q.2 FRESH CORRECTED-CAPTURE COHORT

The fresh cohort contains 48 short endpoints from 12 prompt clusters and 16 boundary endpoints from four overlapping clusters. All 64 endpoints completed under a frozen model, kernel and execution configuration. Short branch lengths are 1, 2, 3 and 4; boundary lengths are 8, 16, 32 and 64. Table 3 reports the short group. Each method used two warm-ups and five formal repeats. Short-cohort comparisons pair the within-endpoint timing medians, then resample whole prompt clusters 10,000 times with seed 43220260923. The resulting 95% percentile intervals are pointwise.

Corrected DFD-H reduced packet bytes and recurrent-state error relative to both dense low-precision baselines on all 48 short endpoints. Its median endpoint-paired packet ratios were 0.140006 to Full, 0.269965 to FP16 and 0.269926 to BF16. Corresponding median paired whole-request wall-time ratios were 1.558669 ($[1.487124, 1.590794]$), 1.924118 ($[1.848096, 2.039364]$) and 1.972212 ($[1.890533, 2.032685]$). Table entries summarize absolute endpoint medians, whereas ratios summarize endpoint pairs. Requests charge source and target execution, method-required capture, construction, codecs, restoration and fresh forward. Prefix execution, audit work, file I/O, network transfer and the separate continuation are excluded.

Every short endpoint was eligible for A_{128}. Corrected and legacy DFD matched on 25/48 and 17/48 endpoints, respectively; FP16 and BF16 matched on 29/48 and 26/48. The corrected-versus-legacy paired counts were 11 matching under both, 14 matching only after correction, six matching only under legacy capture, and 17 matching under neither.

The boundary results are descriptive and analyzed separately from the short group and historical sessions. Only 3/16 records were smaller than both low-precision dense baselines and 8/16 were smaller than Full. The archive retains runtime and per-endpoint records with an independent post-outcome statistical recomputation. The portable table bundle checks its compact-input hash and generated tables; its maintainer import also checks the five upstream source hashes.

Q.3 Rank reduction within captured residuals

The offline rank analysis tests for redundancy in corrected DFD factors from the same 64 fresh endpoints. For each compact head, reduced QR factorizations of U and V followed by a small-core singular-value decomposition in 64-bit floating point (FP64) give the spectrum of $UV^\top$. Rank and error refer to this captured factor product, not the native-source residual or the bounded-gain rank r_γ.

For each of five fixed relative Frobenius tolerances, the analysis finds the smallest rank k whose ideal truncated approximation satisfies the error bound. Each endpoint ratio divides the sum of $k(d_k + d_v)$ over compact heads by their original factor-value count, $\sum_h p_h(d_k + d_v)$. Table 41 reports medians of these ratios, excluding dense fallback, scalar decay, non-recurrent fields and framing. Ideal FP64 errors omit rounding new factors to FP32, making this an optimistic factor-count reference rather than an implemented codec.

At tolerance 10^{-6}, the median short-endpoint ratio is 0.9993; at 10^{-2}, it is 0.9614. High-fidelity recompression therefore retains nearly all short-branch factor values, while longer residuals offer more rank reduction. These tolerances describe factor error, not native-state or continuation accuracy. All 64 endpoints have defined ratios at every tolerance; heads are diagnostic observations, not independent samples. The tolerance grid was frozen after model execution started but before the spectral analysis. The portable artifact binds the group CSV and protocol; the archive retains full per-head spectra.

Q.4 Capture mechanism

A one-endpoint native-call audit found BF16 raw keys with normalization enabled inside the fused recurrent call. Across 24 layers and 768 heads, FP32-before-normalization reduced aggregate absolute state error to 7.51×10^{-5} of its legacy value. Every head improved, while an unhooked control preserved logits and all cache fields exactly.

A 12-endpoint, seven-cluster paired diagnostic then constructed legacy and corrected records from the same captured native calls. Packet bytes and compact masks were unchanged. Median recurrent-state relative error fell from 1.67×10^{-3} to 1.16×10^{-7}, but both versions accumulated ten teacher-forced top-1 disagreements against Full across the fixed histories.

Q.5 Persistent continuation through tool observations

The cohort used 17 exposed ALFWorld development tasks selected by a fixed minimum-hash rule before this run's outcomes were available. Results from an earlier stopped run prompted a syntax correction in the public demonstrations. A deterministic adapter changed exactly 12 action lines from unsupported `put ... in/on ...` to installed TextWorld `move ... to ...` syntax, without rewriting runtime outputs. Native execution consumed three common greedy tokens, then completed candidates branching on the top-1 and top-2 next tokens. A fixed external rule selected the top-1 source after both candidates committed. These constructed branches tested the recovery interface. The selected source was released, restored and continued with appended action and observation tokens, without re-encoding the full conversation.

All four methods received the environment's original success score on 10/17 tasks. DFD reproduced Full's complete action history on 14/17 and token history on 13/17; non-recurrent-only restoration did so on 8/17 and 7/17.

Table 40: Branch-length boundary for corrected DFD-H. Each length has four endpoints. Packet size and state error are medians across endpoints; ratios are descriptive medians of endpoint-paired packet-size ratios.

Length	Packet (MB)	/Full FP32	/Dense FP16	/Dense BF16	State error	A_{128}
8	15.59	0.297	0.572	0.572	2.10×10^{-7}	2/4
16	29.63	0.562	1.076	1.076	2.87×10^{-7}	2/4
32	53.40	1.004	1.904	1.903	4.33×10^{-7}	2/4
64	88.47	1.631	3.040	3.040	7.22×10^{-7}	1/4

DFD-H: Direct Factor Difference with headwise dense fallback. FP32 and FP16: 32- and 16-bit floating point; BF16: bfloat16. Length is tokens per branch. State error is recurrent relative L_2 error; A_{128} is exact 128-token agreement / eligible endpoints. MB denotes 10^6 bytes.

Table 41: Captured-factor rank oracle. Ratio columns report the median ideal factor-value ratio at each relative Frobenius tolerance; the defined/total denominator applies to every displayed tolerance.

Group	Endpoints	Defined/total	10^{-8}	10^{-6}	10^{-4}	10^{-2}	10^{-1}
Short	48	48/48	0.9997	0.9993	0.9979	0.9614	0.6816
Boundary	16	16/16	0.9664	0.9419	0.8778	0.6018	0.2350
Length 1	12	12/12	1.0000	1.0000	1.0000	0.9953	0.9155
Length 2	12	12/12	1.0000	0.9998	0.9993	0.9708	0.7261
Length 3	12	12/12	0.9994	0.9990	0.9970	0.9496	0.6732
Length 4	12	12/12	0.9991	0.9980	0.9933	0.9247	0.5948
Length 8	4	4/4	0.9935	0.9858	0.9607	0.8145	0.4243
Length 16	4	4/4	0.9754	0.9583	0.9200	0.6832	0.2973
Length 32	4	4/4	0.9441	0.9213	0.8728	0.5627	0.2107
Length 64	4	4/4	0.9166	0.8773	0.7738	0.4218	0.1444

Ratios count ideal factor values for the captured compact-head residual $UV^\top$ only; they are neither packet-byte ratios nor native-source fidelity. Errors use an ideal 64-bit floating-point (FP64) singular-value decomposition and exclude 32-bit floating-point (FP32) rounding. Summaries weight endpoints equally rather than pooling heads; fallback state, λ, framing, and non-recurrent fields are excluded.

For joint action-and-token exactness relative to Full, six tasks matched under both DFD-H and the non-recurrent-only control, seven only under DFD-H, one only under the control, and three under neither.

Each arm was capped at 50 executed actions, 2,048 newly generated action tokens, 64 tokens per command and 8,192 context tokens. Captured and independent native branches matched exactly for all 34 candidates, and Full matched Native exactly on all 17 complete trajectories and terminal caches.

The runtime receipt records an NVIDIA GeForce RTX 5080, driver 591.86 and the FLA fused recurrent kernel SHA-256 `9c15dc68...5889b` across 24 recurrent layers. Package-version strings were not recorded.

In case 6, the Planning Domain Definition Language (PDDL) initial state already placed the same pan in a CounterTop and two StoveBurners. Pickup removed only the named StoveBurner relation, and cleaning supplied the remaining goal conjunct. That win does not verify that the policy executed an additional placement. This post-completion finding leaves the recorded trajectories and original 10/17 scores unchanged.

Q.6 Same-prefix offline packet accounting

The same 17 source caches supported a CPU-only packet comparison. Prefix-Full, prefix-FP16 and prefix-BF16 omit the common attention-KV prefix under an explicit receiver contract; corrected DFD uses the recorded sibling-bound packet. The dense codecs restore recurrent fields to FP32, so their state errors measure only wire conversion.

Table 42: **Same-prefix offline packets.** Medians cover 17 fixed development cases. DFD error is the original GPU restoration measurement; dense errors are CPU wire-cast errors and produced no model outputs. Sharing a prefix boundary does not equalize minimum receiver memory.

Method	Packet (MB)	State rel. L_2	Required receiver state
Prefix Full FP32	52.321	0	common KV prefix
Prefix Dense FP16	27.165	2.07×10^{-4}	common KV prefix
Prefix Dense BF16	27.169	1.66×10^{-3}	common KV prefix
DFD-H (corrected)	11.120	1.67×10^{-7}	common prefix + sibling

Corrected DFD used a median 11.120 MB, compared with 27.165 MB for prefix-FP16, 27.169 MB for prefix-BF16 and 52.321 MB for prefix-Full. DFD was smaller in all 17 paired cases. Its receiver required the complete sibling, with a median logical footprint of 94.437 MB, whereas the dense paths required only the common prefix.

The accompanying artifact contains displayed CSV rows, bound source summaries and a CPU-only display-regeneration script. The script checks bundled hashes and derived values, with optional verification against the original files. Model reruns require the separate execution packages and dependencies.

R Evidence Scope and Cost Follow-ups

R.1 What each comparison establishes

Table 43: Evidence and interpretation across the studies.

Comparison	Supported result	Scope
Exact recovery	Native bits, logits and continuations; smaller short-branch records	Measured runtime; higher median local cost and length-dependent savings
Approximate recovery	Lower recurrent-state error with corrected capture	State error, trajectory agreement and task success are separate outcomes
Paced transfer	Cost savings relative to Full in tested communication-limited conditions	Legacy capture and loopback component costs, not exact-repair deployment latency
ALFWorld continuation	Recurrent repair preserves more complete trajectories	Exposed development tasks and constructed forks; equal aggregate success counts
Reference layouts	Costs under resident-sibling and freely chosen reference contracts	Constructed workloads; the preferable layout depends on residency and candidate count

Statistical interpretation. Timing repeats measure within-case variation; tasks or endpoints are the analysis units, with prompt-cluster resampling where specified. Revisited inputs and overlapping short/boundary clusters do not add independent samples. Reported bootstrap intervals are pointwise; small boundary groups and exploratory comparisons are descriptive. No task-quality equivalence or non-inferiority margin was prespecified. Zero observed losses or intervals containing parity do not establish equivalence.

Application scope. Local timings, paced-transfer component sums and modeled bandwidth crossings use different cost definitions. Exact-repair compression has not been tested in a deployment transfer pipeline. Packet bytes exclude reference residency and temporary workspace, so they are not total-memory measurements. The frequency and reusable lifetime of suitable resident siblings in deployment remain unmeasured. Offline packet and rank analyses evaluate representation costs, not new task or latency outcomes.

R.2 DENSE-PRECISION COST COMPARISONS

S421 measured complete task lifecycles for 12 revisited endpoints, four at each span 2, 4 and 8, with five technical repeats per cell. Legacy-capture DFD-H had a median paired whole-task ratio of 1.0865 to Dense BF16 at 0.5 Gbit/s, with 2/12 endpoints faster. Ratios were 1.2076 at 2 Gbit/s (1/12 faster) and 1.2577 without pacing (0/12 faster).

S426 used bulk packet materialization on the same 12 endpoints and legacy capture. At 0.5 Gbit/s, its median paired task-cost ratio to Dense BF16 was 1.0638, with 5/12 endpoints faster. Ratios at spans two, four and eight were 0.8961, 1.0251 and 1.1326.

S432 measured corrected capture locally, without network transfer. Its complete 40-endpoint short subset had a median whole-wall ratio of 1.8616 to Dense BF16, with 0/40 endpoints faster. The historical run completed 52 of 54 planned endpoints; Appendix Q.1 reports that cohort separately.

S420's layout measurements and repeated deliveries are described in Appendix P. S424's $K = 2, 4, 7$ shared-anchor figures are projections from S421 components; S433 was a feasibility design that was not run. The recorded summaries and SHA-256 bindings remain in `current_cost_artifact`.

S EXACT RECOVERY WITH LOSSLESS REPAIR

S.1 CONSTRUCTION AND COMPARISON CONTRACT

The five-arm exact-recovery study reuses the 64 endpoints in Appendix Q.2: 48 short endpoints in 12 prompt clusters and 16 boundary endpoints in four overlapping clusters. Its protocol was frozen after the precision and offline-repair outcomes were known. All 64 completed without outcome-based exclusions.

Decoding the corrected DFD-H packet yields the recurrent-state prediction $\widehat{S}_s$. The sender XORs the prediction's bytes with native FP32 source bytes, applies a byte-plane transform and compresses the repair. The receiver reconstructs this prediction from transmitted bytes, verifies its hash and reverses the XOR. Both ends must produce identical predictor bits, checked separately from repair inversion. Records include the DFD packet, repair streams, source non-recurrent fields, metadata, framing and hashes.

The primary attribution control directly XORs native source and resident sibling, without scaling or capture. Canonical lossless Full uses the same byte-plane codec for native recurrent bytes and a typed non-recurrent skeleton. Secondary controls send complete historical Full packets, outer-zlib-compressed or raw. Compression uses fixed zlib level 6, window 15, memory level 8 and the default strategy. Raw Full performs no unused canonicalization; canonical Full does not first serialize a dense packet.

The primary clock is a synchronized same-host lifecycle with the sender fork and receiver sibling supplied as input resources. It starts before source execution and ends after restored fresh forward and synchronization. DFD additionally executes and captures a sender target; direct sibling XOR executes that target without capture. Full executes only the source, unlike Appendix Q.2, which charges both branches to every method. All required copies, extraction, construction, framing, hashing, coding, sender prediction and receiver restoration are charged. Every repeat constructs fresh streams. Prefix/fork creation, receiver-target generation, model loading, artifact I/O, separate audits and continuation are outside this primary clock. No network transfer is performed. The setup-inclusive sensitivity adds the separately measured receiver-target generation cost to every method within each repeat, before endpoint aggregation; it still excludes prefix/fork creation.

S.2 PAIRED MEASUREMENTS

Each method uses two warm-ups and five formal repeats per endpoint, giving 640 warm-up and 1,600 formal invocations. Method order rotates deterministically as $(\text{endpoint index} + \text{global repeat}) \bmod 5$; position balance is approximate. Comparisons use the median of five formal measurements within each endpoint and method, then form direct paired ratios and differences. Absolute table entries are medians across endpoint medians; paired ratios are summarized separately.

For the short group, 10,000 whole-prompt-cluster resamples with seed 44420260924 produce pointwise 95% percentile intervals. Each sampled cluster retains all its endpoints. The two primary contrasts compare DFD-H plus repair with exact sibling XOR and canonical lossless Full; both secondary Full contrasts are also reported. Boundary groups are descriptive.

Table 44: Absolute short-branch costs with and without receiver-target setup.

Method	Local time (ms)	Record (MB)	With setup (ms)
Canonical lossless Full	2250.122	44.523	2326.730
Zlib Full packet	1840.685	48.883	1916.636
Raw Full packet	347.244	52.239	426.539
Exact sibling XOR	2285.085	43.442	2363.608
DFD-H plus exact repair	2330.013	12.972	2405.035

DFD-H: Direct Factor Difference with headwise dense fallback; XOR: bitwise exclusive OR. MB = 10^6 bytes. Values are cohort medians of five-repeat endpoint medians. Local time is the synchronized same-host hot-target lifecycle; the setup column additionally includes receiver-target setup. Short branches: 48 endpoints in 12 prompt clusters.

Table 45: Paired short-branch contrasts for exact recovery.

Baseline	Metric	Ratio	Win/tie/loss	95% CI
Exact sibling XOR	Local time	1.035	20/0/28	[0.973, 1.080]
Exact sibling XOR	Record size	0.301	48/0/0	[0.288, 0.332]
Exact sibling XOR	With setup	1.034	20/0/28	[0.973, 1.078]
Canonical lossless Full	Local time	1.044	8/0/40	[1.025, 1.055]
Canonical lossless Full	Record size	0.291	48/0/0	[0.280, 0.324]
Canonical lossless Full	With setup	1.043	8/0/40	[1.025, 1.054]
Zlib Full packet	Local time	1.274	0/0/48	[1.260, 1.294]
Zlib Full packet	Record size	0.265	48/0/0	[0.256, 0.295]
Zlib Full packet	With setup	1.268	0/0/48	[1.250, 1.279]
Raw Full packet	Local time	6.638	0/0/48	[6.500, 6.960]
Raw Full packet	Record size	0.248	48/0/0	[0.239, 0.276]
Raw Full packet	With setup	5.635	0/0/48	[5.553, 5.717]

DFD-H: Direct Factor Difference with headwise dense fallback; XOR: bitwise exclusive OR. The numerator is DFD-H plus exact repair. Ratios pair endpoints before aggregation; a win means lower cost than the baseline. With setup adds receiver-target setup. CI: confidence interval. The pointwise 95% percentile intervals use 10,000 whole-prompt-cluster resamples.

All 48 short endpoints favor DFD-H plus repair in bytes against each baseline. Median paired byte differences are -30.278 MB against sibling XOR, -31.533 MB against canonical Full, -35.900 MB against zlib Full and -39.251 MB against raw Full. Corresponding local-time differences are $+78.8$, $+99.0$, $+493.7$ and $+1982.2$ ms. Adding the same receiver-target setup within a pair leaves its time difference unchanged, although the ratio changes. The sibling-XOR time interval crosses one, leaving that timing comparison unresolved.

S.3 Length dependence

Table 46 separates the four boundary lengths. At lengths 8 and 16, all four endpoints per length use fewer bytes than every baseline. At lengths 32 and 64, all four use more. Local time is higher in every method/length group; only one individual length-8 endpoint is faster than sibling XOR.

S.4 Exactness controls and reproduction

All five arms passed raw cache-bit equality, scalar/position checks, fresh-logit equality and fork/target immutability at every endpoint. Continuations use retained formal-repeat-zero streams, one run per endpoint and method plus the native reference. Every native run reached 128 predictions without an end-of-sequence (EOS) token, including prediction 128. Each arm matched tokens,

Table 46: Exact-recovery ratios by boundary length.

Length	Baseline	Record ratio	Local-time ratio	With-setup ratio
8	Canonical lossless Full	0.442	1.148	1.135
8	Zlib Full packet	0.403	1.390	1.349
8	Raw Full packet	0.376	5.674	4.173
8	Exact sibling XOR	0.440	1.030	1.027
16	Canonical lossless Full	0.797	1.411	1.342
16	Zlib Full packet	0.728	1.697	1.573
16	Raw Full packet	0.680	5.105	3.545
16	Exact sibling XOR	0.790	1.159	1.139
32	Canonical lossless Full	1.274	1.811	1.643
32	Zlib Full packet	1.165	2.117	1.829
32	Raw Full packet	1.087	4.699	3.055
32	Exact sibling XOR	1.264	1.364	1.296
64	Canonical lossless Full	1.937	2.319	1.923
64	Zlib Full packet	1.774	2.603	2.076
64	Raw Full packet	1.650	4.517	2.909
64	Exact sibling XOR	1.921	1.527	1.410

Ratios are DFD-H plus exact repair divided by the named baseline and are descriptive medians over four endpoints at each length. DFD-H, Direct Factor Difference with headwise dense fallback; XOR, bitwise exclusive OR.

Table 47: Absolute costs across all 16 boundary endpoints.

Method	Local time (ms)	Record (MB)	With setup (ms)
Canonical lossless Full	2985.606	45.143	3695.732
Zlib Full packet	2503.546	49.442	3211.955
Raw Full packet	986.601	52.944	1668.630
Exact sibling XOR	3832.575	45.487	4501.914
DFD-H plus exact repair	4837.461	48.403	5544.183

DFD-H: Direct Factor Difference with headwise dense fallback; XOR: bitwise exclusive OR. MB = 10^6 bytes. Values are cohort medians of five-repeat endpoint medians. Local time is the synchronized same-host hot-target lifecycle; the setup column additionally includes receiver-target setup. Boundary cases: 16 endpoints in four overlapping prompt clusters; descriptive results.

stopping behavior and terminal caches: 48/48 short, 16/16 boundary and 64/64 overall. Exactness was verified in the measured runtime; early EOS was not exercised.

An independent CPU validator rehashed and decoded one retained stream set per method/endpoint pair: 320 sets from formal repeat zero. Sibling-XOR inversion uses saved bytes of the independently generated receiver target; DFD repair inversion uses the validation-bound parent predictor. The validator checks decoding, not model or predictor re-execution; live receiver checks and predictor hashes cover those paths separately.

Measurements from the failed eight-endpoint input-guard run and the separate three-endpoint preflight are excluded. The revised protocol, code and bindings were frozen before preflight and formal execution.

The arithmetic bundle contains all 2,240 timing rows, 384 native/method continuations, the endpoint roster and frozen analyzer. Package-relative checks reproduce the 10,000-resample statistics and token classifications; an independent implementation reproduces the paired summaries. GPU execution and retained-stream decoding use separate artifacts containing execution entry points, provenance and streams.