DENSE EVIDENCE, SPARSE RECURRENCE: HOW MUCH RECURRENT COMPUTATION STATE-SPACE MRI RECONSTRUCTION NEEDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision state-space models inherit from sequence modeling the convention that every token is a recurrent step. For dense reconstruction this ties the cost of recurrence to the image grid rather than to what the reconstruction needs, and whether the convention is necessary has not been established. We ask how much recurrence a dense reconstruction actually needs. Answering the question requires separating two decisions that an ordinary scan conflates: what the scan carries, and which positions execute it. We build an unrolled MRI reconstruction architecture, **ResiMamba**, in which both are explicit and independently controllable. A smooth content carrier supplies the recurrent updates, while the complementary evidence conditions the state write/read interfaces and reaches the output through a dense path; within the carrier, a coverage-constrained budget selects the positions that execute a short packed scan, and unselected content is routed back into the evidence path rather than discarded. Each budget is trained as its own design point. The resulting frontier shows that recurrence is a sparse resource: a few percent of spatially distributed positions recover most of the quality that dense recurrence adds over an evidence-only model, with sharply diminishing returns beyond. The budgeted configurations remain close to dense quality across five MRI datasets, cut measured latency by a factor that grows with image size, and are both faster and better than the depth-reduced alternatives. Ablations attribute the retained quality to preserving the unselected content and spreading the sparse recurrence across space, and show that a simple selector suffices. Recurrent participation is then a design axis for state-space reconstruction alongside depth and width. Nothing in the mechanism is specific to MRI; the study uses MRI reconstruction because it offers data-consistent unrolling, five public benchmarks, and standard metrics.

1 INTRODUCTION

Vision state-space models turn a spatial feature grid into an ordered sequence and execute every position as a recurrent step. The convention is inherited from language modeling, where every token is an input that must be read; for dense image reconstruction it is a default rather than a requirement, and its cost is set by the image grid. Accelerated MRI reconstruction makes the tension concrete: the learned prior must supply long-range structural context and, at the same time, the local evidence that resolves anatomical boundaries and texture, and every stage of the unrolled solver must remain consistent with the acquired k-space. Unrolled solvers with convolutional regularizers (Hammernik et al., 2018; Aggarwal et al., 2019), Transformers (Huang et al., 2022; Guo et al., 2024), and state-space architectures (Chen et al., 2026; Li et al., 2025) differ mainly in how they provide the context; the state-space ones provide it with a selective scan whose input-dependent parameters decide how each token interacts with the state, while the length of the recurrent trajectory remains that of the grid.

This paper asks: *how much of that recurrence does a dense reconstruction actually need?* The question cannot be answered by dropping tokens from an existing scan, because a scan conflates two decisions. One is *what the scan carries*: the features useful for propagating context are not necessarily the features useful for conditioning the state or correcting the output. The other is *which*

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.

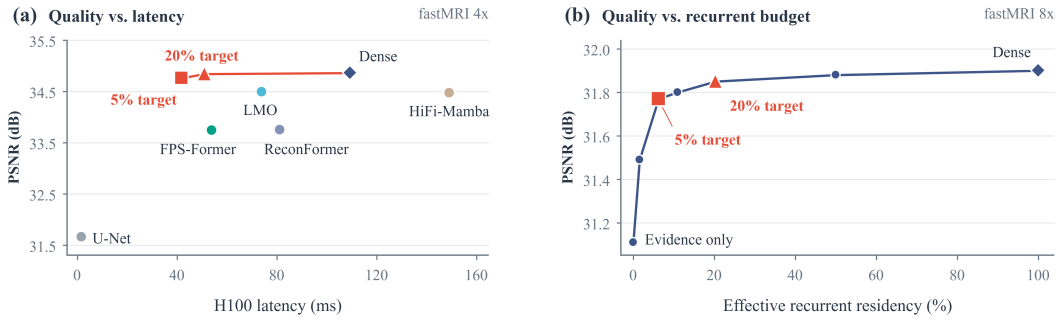


Figure 1: PSNR versus (a) H100 latency on fastMRI $4\times$ (320×320, batch size one) and (b) effective recurrent residency on fastMRI $8\times$. Dense is ResiMamba with all positions resident; labels are target budgets (5% and 20% correspond to 6.25% and 20.31% effective), each trained separately.

positions execute it: removing a position from the sequence also removes its evidence from the reconstruction, so quality loss from token reduction cannot be attributed to the missing recurrence alone. Answering the question requires an architecture in which both decisions are explicit and independently controllable.

We build such an architecture, **ResiMamba**, and use it to answer the question. A learned projection followed by fixed spatial filtering constructs a smooth content *carrier* and its complementary residual; the carrier supplies the recurrent content, while the residual and the bypassed features form a dense *evidence* field that conditions the state write/read interfaces and reaches the output directly. Within the carrier, a coverage-constrained scorer selects a budgeted subset of positions, packs them into one short ordered sequence, and runs an evidence-conditioned Mamba-3 scan on it. Unselected carrier content is not discarded but routed into the evidence field, so that a position can contribute to the reconstruction without becoming a step in the recurrent trajectory. Dense recurrence corresponds to every carrier position being selected. Each budget below it is trained as its own configuration: we treat the recurrent budget as a design choice made before training, in the same way that depth and width are, and report the quality–latency trade-off across budgets.

Our contributions are:

- **A controlled study of recurrent participation** in state-space reconstruction, separating what the scan carries from which positions execute it, with a residency frontier showing that a few percent of spatially distributed positions recover most of the benefit of dense recurrence (Sections 4.2–4.3).
- **ResiMamba**, the architecture that makes the study possible: carrier/evidence role separation with evidence acting at the state interfaces and at the output, and budgeted execution with coverage-constrained selection, order-preserving packing, and routing of unselected content back into the evidence path (Section 3).
- **Evidence that recurrent participation is a practical design axis**: budgeted configurations stay near-dense on five datasets, cut measured latency by two to four times, beat depth reduction, and owe their quality to routing and coverage rather than to a sophisticated selector (Sections 4.4–4.7).

2 Related Work

Learned MRI reconstruction. Unrolled methods interleave learned regularization with data-consistency updates; variational networks and MoDL established the approach with convolutional regularizers (Hammernik et al., 2018; Aggarwal et al., 2019), and Transformer-based models such as SwinMR, ReconFormer, and FPS-Former widened the range of feature interactions through hierarchical, recurrent, or frequency-aware attention (Huang et al., 2022; Guo et al., 2024; Meng et al., 2025). State-space regularizers, including the linear Mamba operator LMO (Li et al., 2025) and the dual-stream HiFi-Mamba (Chen et al., 2026), provide long-range context at linear cost and are the

strongest baselines evaluated here. All of them execute the scan over every feature position; our question is whether they need to.

Selective state-space models. Mamba introduced input-dependent selective state-space updates (Gu & Dao, 2024); we build on its successor Mamba-3 (Lahoti et al., 2026). Vision variants serialize the feature grid in one or several scan orders and execute every position; the selective parameters decide how each token interacts with the state, not whether it is executed at all, which is the quantity this paper controls.

Token reduction. DynamicViT and ToMe shorten Transformer sequences by learned sparsification and by merging similar tokens (Rao et al., 2021; Bolya et al., 2023), S5 selects inputs before state-space processing for long video (Wang et al., 2023), ToP-ViM prunes vision-Mamba tokens with hidden-state alignment and evaluates dense prediction as well (Zhan et al., 2024a), and ToR-SSM combines importance and similarity criteria for language state-space models (Zhan et al., 2024b). In all of these the removed or merged tokens are no longer individually available to the output. ResiMamba reduces only the recurrent sequence: unselected content stays in a dense path that also supplies the scan's conditioning, and local quotas keep the sparse recurrence spatially distributed. Section 4.7 compares pruning, similarity merging, and evidence routing under a common window quota and packing.

3 METHOD

3.1 UNROLLED RECONSTRUCTION AND COMPUTATIONAL ROLES

Let $y = \mathcal{A}x + \epsilon$ denote undersampled MRI measurements of an image x, with encoding operator $\mathcal{A}$ (Fourier sampling and, for multi-coil data, coil encoding) and noise ϵ. A stride-two 2×2 patch embedding of the zero-filled reconstruction forms a feature grid F_1 with $N = HW$ positions; a 320×320 image gives a 160×160 grid of $25{,}600$ candidate positions per scan. At stage t, two reconstruction units $\mathcal{R}_{\theta,t}$ are applied in sequence, each with its own internal residual, and data consistency (DC) is applied to the result and added back:

$$Z_t = \mathcal{R}_{\theta,t}(F_t), \qquad F_{t+1} = Z_t + \mathcal{E}_t\big(\mathcal{D}_t(Q_t(Z_t), y, \mathcal{A})\big), \tag{1}$$

where Q_t decodes the features to an image, $\mathcal{D}_t$ enforces agreement with the acquired measurements in image space, and $\mathcal{E}_t$ re-embeds the result. A terminal step decodes and enforces consistency once more, $\widehat{x} = \mathcal{D}_{\text{out}}(Q_{\text{out}}(F_{T+1}), y, \mathcal{A})$, without re-embedding. The default network has $T = 6$ stages with two separately parameterized units per stage and a 128-channel feature width.

Within a unit, we distinguish three roles. The *carrier* supplies the content/value tokens propagated by the recurrent computation. The *evidence* modulates how that content is written into and read from the state (the *Access* interface), and also contributes to the output through a dense path (the *Outlet*). *Residency* determines which spatial positions of the carrier actually become recurrent steps.

3.2 CONSTRUCTING CONTENT AND COMPLEMENTARY EVIDENCE

For a unit input $\mathbf{X} \in \mathbb{R}^{H \times W \times 2d}$, we omit the batch and unit indices and split the normalized features:

$$[\mathbf{X}^h, \mathbf{X}^\ell] = \text{Split}(\text{LN}(\mathbf{X})), \qquad \mathbf{U} = \phi(\mathbf{X}^\ell), \tag{2}$$

where both branches have $d = 64$ channels and ϕ is a learned 3×3 convolution. A fixed spatial operator constructs the carrier from this learned representation. With the normalized binomial kernel $\mathbf{b} = [1, 4, 6, 4, 1]/16$, define

$$\mathcal{P}_{\mathbf{b}}(\mathbf{U}) = \mathbf{b}^\top *_2 \big(\mathbf{b} *_1 \text{Pad}_{\text{ref}}(\mathbf{U})\big), \qquad \mathbf{L} = \text{Up}_2(\mathcal{P}_{\mathbf{b}}(\mathbf{U})). \tag{3}$$

The separable depthwise filtering applies a horizontal unit-stride pass and a second pass with spatial stride two; bilinear upsampling returns to the original feature grid, with reflection padding at the boundaries.

The complementary residual and initial evidence are

$$\mathbf{E} = \mathbf{U} - \mathbf{L}, \qquad \mathbf{E}_0 = \mathbf{X}^h + \mathbf{E}. \tag{4}$$

The learned projection decides what is decomposed and the fixed operator makes the carrier spatially smooth, leaving the difference in the evidence route. The carrier construction thus fixes *what* the scan processes, and residency (Section 3.3) fixes *how many* positions execute it.

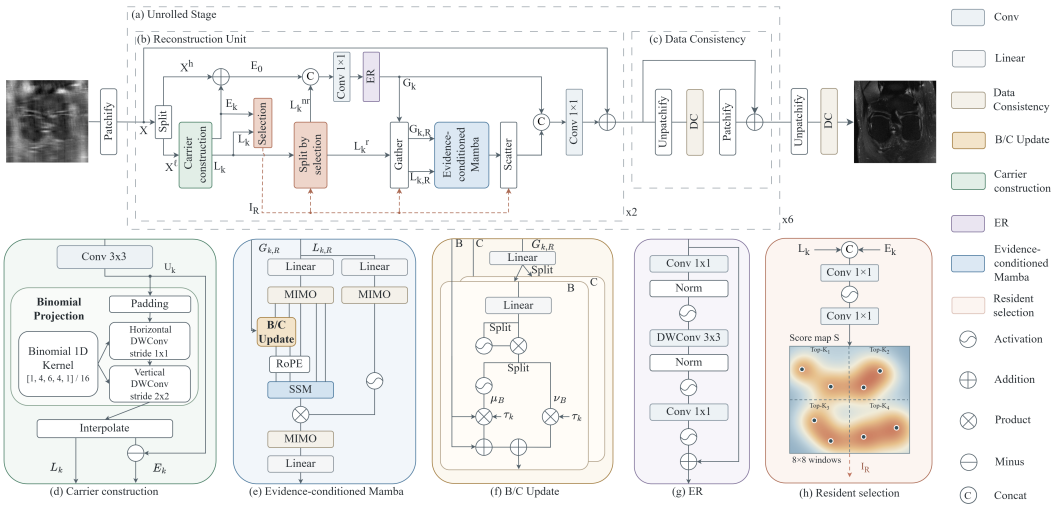


Figure 2: ResiMamba. (a) Each unrolled stage applies two reconstruction units and then (c) data consistency. (b) Inside a unit, one half of the features is smoothed into the content carrier $\mathbf{L}$ by (d) a learned projection and a fixed binomial filter; everything the carrier leaves out, together with the other half of the features and the carrier positions that are not selected, is refined by (g) the evidence refinement (ER) into the dense evidence $\mathbf{G}$. (h) A scorer picks a budgeted, window-distributed subset of carrier positions. Carrier and evidence at those positions are gathered into (e) one short scan, where (f) the evidence modulates the state write/read interfaces (Access). The scan output is scattered back and fused with the dense evidence into the unit output (Outlet). The subscript k in the panels indexes the unit.

3.3 Coverage-Constrained Resident Selection

For target residency $\rho \in (0, 1]$, a pointwise scorer predicts

$$\mathbf{S} = g_\psi(\mathbf{L}, \mathbf{E}) = \mathrm{Conv}_2(\mathrm{GELU}(\mathrm{Conv}_1([\mathbf{L} \| \mathbf{E}]))), \tag{5}$$

using two 1×1 convolutions and one output score per location (panel (h) of Figure 2). The feature grid is partitioned into non-overlapping windows $\{\Omega_j\}_{j=1}^{J}$. For $n_j = |\Omega_j|$, each window receives $K_j = \max(1, \lceil \rho n_j \rceil)$ recurrent positions, selected by

$$\mathbf{m}_j^\star = \arg \max_{\mathbf{z} \in \{0,1\}^{n_j}} \sum_{i \in \Omega_j} z_i S_i \quad \text{subject to} \quad \sum_{i \in \Omega_j} z_i = K_j. \tag{6}$$

The quota ensures local coverage, while the scores adapt the selected positions within each window to the input.

Assembling the local selections yields $\mathbf{M}^{\mathrm{hard}} \in \{0, 1\}^{H \times W}$. With $N = HW$ and $K = \sum_j K_j$, the effective residency is $\rho_{\mathrm{eff}} = K/N$. For the default 8×8 windows, a 5% target retains four positions per window, or 6.25% effective residency; a 20% target retains thirteen, or 20.31%; on a 320×320 input the 5% target thus executes $1{,}600$ of $25{,}600$ positions per scan.

3.4 Preserving Unselected Content

The discrete mask fixes which indices are gathered. During training, a straight-through gate supplies a surrogate gradient to the scorer:

$$\mathbf{M}^{\mathrm{gate}} = \mathbf{M}^{\mathrm{hard}} + \sigma(\mathbf{S}) - \mathrm{sg}[\sigma(\mathbf{S})]. \tag{7}$$

Here sg stops the gradient: the forward pass uses the hard selection and the backward pass the sigmoid derivative, while the gathered index set stays discrete.

Broadcasting the mask across channels gives complementary components

$$\mathbf{L}^{\mathrm{r}} = \mathbf{M}^{\mathrm{gate}} \odot \mathbf{L}, \qquad \mathbf{L}^{\mathrm{nr}} = (1 - \mathbf{M}^{\mathrm{gate}}) \odot \mathbf{L}. \tag{8}$$

Thus $\mathbf{L}^{\mathrm{r}} + \mathbf{L}^{\mathrm{nr}} = \mathbf{L}$. Unselected content is routed into the evidence representation:

$$\mathbf{G}_0 = \mathcal{P}_E([\mathbf{E}_0 \| \mathbf{L}^{\mathrm{nr}}]), \qquad \mathbf{G} = \Phi_E(\mathbf{G}_0). \tag{9}$$

The bias-free 1×1 projection $\mathcal{P}_E : \mathbb{R}^{2d} \to \mathbb{R}^d$ is initialized with $[\mathbf{I}_d \ \mathbf{0}]$. The evidence refinement Φ_E is a depthwise inverted bottleneck (panel (g) of Figure 2). Both operate on the full grid, so the residual, the bypassed branch, and the unselected carrier reach the reconstruction even where no recurrent step is executed.

3.5 Evidence-Conditioned Packed Recurrence

Let π be the row-major raster order and $\mathcal{I}_R$ the resident indices sorted in that order. Content and conditioning use exactly the same gather indices:

$$\mathbf{L}_R = \mathrm{gather}_{\mathcal{I}_R}(\mathrm{Flatten}_\pi(\mathbf{L}^{\mathrm{r}})), \qquad \mathbf{G}_R = \mathrm{gather}_{\mathcal{I}_R}(\mathrm{Flatten}_\pi(\mathbf{G})). \tag{10}$$

Both sequences have length K.

The recurrent operator $\mathcal{M}_\theta(\cdot;\cdot)$ is a selective state-space unit (panel (e) of Figure 2), instantiated with Mamba-3 (Lahoti et al., 2026). The carrier tokens determine its content projections and scan parameters; denote the unmodulated state write/read interfaces by $\mathbf{B}_s$ and $\mathbf{C}_s$. Token-wise conditioning maps of $\mathbf{G}_R$ produce affine parameters:

$$\begin{aligned}(\boldsymbol{\mu}_B, \boldsymbol{\nu}_B, \boldsymbol{\mu}_C, \boldsymbol{\nu}_C) &= P(\mathbf{G}_R),\\ \mathbf{B}'_s &= \mathbf{B}_s \odot (1 + \tau \tanh \boldsymbol{\mu}_B) + \tau \boldsymbol{\nu}_B,\\ \mathbf{C}'_s &= \mathbf{C}_s \odot (1 + \tau \tanh \boldsymbol{\mu}_C) + \tau \boldsymbol{\nu}_C.\end{aligned} \tag{11}$$

The learned scalar τ is initialized to zero. This FiLM-style conditioning (Perez et al., 2018) is the Access interface: evidence changes how carrier content is written to and read from the state without supplying a second content sequence, while transition and discretization parameters stay carrier-side; the scan update itself is given in Appendix C.

The unit assembles the state reads into the pre-output-projection readout $\{s_n\}_{n=0}^{K-1}$. After the output projection, the length-K output $\mathbf{R}_R$ is scattered back to its original raster positions to form $\mathbf{R}$, which is zero at unselected positions, whereas $\mathbf{G}$ remains dense. The unit output is $\mathbf{Y} = \mathbf{X} + \mathrm{DropPath}(\mathcal{P}_F([\mathbf{G} \| \mathbf{R}]))$, where the fusion projection maps $2d$ channels to $2d$ channels; the direct entry of $\mathbf{G}$ into this fusion is the Outlet. Algorithm 1 (Appendix C) lists the complete unit.

3.6 Dense Configuration and Computational Cost

With all carrier positions resident the scan has length N; the dense reference bypasses selection, uses $\mathbf{G} = \Phi_E(\mathbf{E}_0)$, and therefore omits $\mathcal{P}_E$ and the scorer (Appendix D). The residency budget is a design choice fixed before training, like depth and width: each reported budget is trained as its own configuration and gives a deterministic executed sequence length and latency for a given input size. Budgets are trained rather than imposed afterwards: the tested post-hoc reduction of the trained dense model costs 1.8–2.1 dB at 5–20% keeping (Appendix D.4) and does not reproduce this frontier, whereas the budget-trained models stay within 0.13 dB of dense.

At fixed feature and state dimensions, the sequence-dependent recurrent cost falls from $\mathcal{O}(N)$ to $\mathcal{O}(K)$ per unit. Across units u, the relative executed sequence length is $\sum_u K_u / \sum_u N_u$, equal to ρ_{eff} for a common grid and selection rule. Carrier construction, evidence refinement, fusion, and DC remain dense, and scoring, top-K, gathering, and scattering add overhead, so we report synchronized end-to-end latency rather than equating sequence reduction with runtime reduction.

4 Experiments

4.1 Experimental Setup

The budget study uses fastMRI (Zbontar et al., 2018) single-coil knee data at 320×320 with $8\times$ equispaced Cartesian undersampling of k-space synthesized from the magnitude reconstructions, following Chen et al. (2026); the operating points found there are then evaluated on fastMRI,

Variant	Content	Access	Outlet	PSNR ↑	SSIM ↑
Mamba-3	$\mathrm{LN}(\mathbf{X})$	–	–	34.13	0.802
Carrier only	$\mathbf{L}$	–	–	34.59	0.852
w/o Access	$\mathbf{L}$	–	✓	34.68	0.849
w/o Outlet	$\mathbf{L}$	✓	–	34.55	0.854
Evidence as content	$\{\mathbf{L}, \mathbf{G}\}$	✓	✓	31.71	0.762
Full role design (dense)	$\mathbf{L}$	✓	✓	**34.86**	**0.865**

Table 1: Content/evidence-role ablation on fastMRI $4\times$ with the same 6×2 unrolled network. Content lists the streams that supply the scan content tokens: Mamba-3 uses a plain Mamba-3 unit on $\mathrm{LN}(\mathbf{X})$ without the split or the evidence branch; Carrier only uses $\mathbf{L}$ with no evidence interface; Evidence as content concatenates $\mathbf{G}$ with $\mathbf{L}$ along channels and projects the result back to d channels as the scan content.

Method	PSNR $4\times$	PSNR $8\times$	Latency (ms)	Speedup
U-Net	31.66	28.60	1.69	–
FPS-Former	33.74	30.63	53.89	–
LMO	34.49	31.10	73.92	–
ReconFormer	33.75	30.42	81.18	–
HiFi-Mamba	34.47	31.38	149.12	–
ResiMamba				
Dense, 6×2	34.86	31.90	109.34	$1.00\times$
Dense, 5×2	–	31.72	91.10	$1.20\times$
Dense, 4×2	–	31.63	72.90	$1.50\times$
20%, 6×2	34.84	31.85	51.02	$2.14\times$
5%, 6×2	34.76	31.77	41.78	$2.62\times$

Table 2: Quality and H100 forward latency on fastMRI (320×320, batch size one). $s\times u$: s stages with u units each; speedup is relative to the 6×2 dense model.

CC359 (Souza et al., 2018), Prostate158 (Adams et al., 2022), ACDC (Bernard et al., 2018), and M4Raw (Lyu et al., 2023) at $4\times$ and $8\times$, covering knee, brain, prostate, and cardiac anatomy, Cartesian, random, and radial sampling, and single- and multi-coil acquisition. Baselines are U-Net, ISTA-Net, TransUNet, Mamba-UNet, ReconFormer, FPS-Former, LMO, and HiFi-Mamba (Ronneberger et al., 2015; Zhang & Ghanem, 2018; Chen et al., 2021; Wang et al., 2024; Guo et al., 2024; Meng et al., 2025; Li et al., 2025; Chen et al., 2026) under the same preprocessing and masks. PSNR and SSIM are averaged per case and then across cases, with case-level standard deviations where shown; NMSE is reported for the budget study and the ablations. All models use six stages of two units (6×2; $s\times u$ denotes s stages of u units). Budgeted models use 8×8 windows and one raster-ordered scan per unit; all ResiMamba variants are trained with the same recipe and seed, and baselines follow their published settings. Latency is the synchronized end-to-end forward time on one H100 at batch size one. Dataset splits, masks, metric aggregation, training hyperparameters, and the timing protocol are given in Appendices A and B.

4.2 Role Design at Full Residency

Before restricting recurrent execution, we verify the division of roles at full residency. Table 1 separates carrier construction, evidence-conditioned state access, and the dense output outlet on fastMRI $4\times$. The Mamba-3 baseline, a plain Mamba-3 unit on $\mathrm{LN}(\mathbf{X})$ in the same 6×2 unrolled network, reaches 34.13 dB. Constructing the carrier alone increases PSNR to 34.59 dB, 0.46 dB of the total gain; the two evidence interfaces add the remaining 0.27 dB. Adding the evidence outlet without interface conditioning gives 34.68 dB; conditioning without the outlet gives 34.55 dB. Combining carrier construction, conditioning, and the outlet achieves 34.86 dB and SSIM 0.865, the highest values in the comparison; this full design is the dense configuration used in all following experiments.

The two evidence interfaces contribute differently: the outlet acts on the reconstruction outside the scan, while conditioning changes how carrier content interacts with the state. Their complete combination improves PSNR beyond either partial configuration. Concatenating $\mathbf{G}$ with $\mathbf{L}$ along channels and projecting back to d channels as the scan content, with Access and Outlet retained, gives 31.71 dB, well below every single-content variant. The results are consistent with keeping the roles distinct: evidence influences the state and the reconstruction without becoming a second content source.

4.3 Recurrent Residency Frontier

With the roles fixed, we vary how many carrier positions execute the scan (Table 3, Figure 1(b)). A separately trained evidence-only configuration, in which all carrier content is routed to the evidence path and no scan is executed, reaches 31.11 dB; dense recurrence adds 0.79 dB over it. One executed position per 8×8 window (1.56% effective) recovers about half of that gap, four positions (6.25%) recover 83.5% of it, and every budget from 10% up to dense adds at most a further 0.13 dB. The scan is worth keeping, since its 0.79 dB exceeds the 0.52 dB margin between the dense model and

Target	Effective	Quality, fastMRI 8×			Latency in ms (speedup), by input size			
		PSNR ↑	SSIM ↑	NMSE ↓	256×256	320×320	384×384	512×512
0%	0%	31.11	0.752	0.0204	–	15.14 (7.22×)	–	–
1%	1.56%	31.49	0.759	0.0201	37.01 (2.15×)	39.30 (2.78×)	38.96 (3.82×)	44.97 (5.63×)
5%	6.25%	31.77	0.771	0.0197	37.42 (2.13×)	41.78 (2.62×)	43.46 (3.42×)	54.22 (4.67×)
10%	10.94%	31.80	0.772	0.0196	39.74 (2.00×)	44.58 (2.45×)	48.53 (3.06×)	61.95 (4.09×)
20%	20.31%	31.85	0.773	0.0195	44.53 (1.79×)	51.02 (2.14×)	56.81 (2.62×)	80.09 (3.16×)
50%	50.00%	31.88	0.773	0.0195	57.52 (1.38×)	73.05 (1.50×)	90.13 (1.65×)	140.60 (1.80×)
Dense	100%	31.90	0.773	0.0195	79.54 (1.00×)	109.34 (1.00×)	148.69 (1.00×)	253.23 (1.00×)

Table 3: Recurrent residency frontier on fastMRI 8×: quality of the trained budgets and H100 latency (speedup versus dense at the same input size). Effective residency follows the per-window quota. The 0% row is a separately trained evidence-only configuration (all carrier content routed to the evidence path, no scan); Dense is ResiMamba with all positions resident. The 256 and 320 sizes are the CC359 and fastMRI inputs; 384 and 512 are resized inputs profiled for latency only.

the strongest external baseline on this benchmark (Table 2), but it is worth keeping only sparsely: most of its contribution is obtained with a few spatially distributed positions.

4.4 Quality–Latency Trade-off

Sparse execution is only useful if the shorter scan is cheaper end to end. On H100 at 320×320 (Table 2, Figure 1(a)), the 5% and 20% configurations run $2.62\times$ and $2.14\times$ faster than dense and reach higher PSNR at lower latency than every external method except U-Net, which is much faster but well below in quality; HiFi-Mamba, the strongest external baseline, is $3.6\times$ slower than the 5% model and 0.39 dB lower at $8\times$. The budgeted models have slightly *more* parameters than dense (Table 9), so the reduction comes from shorter execution alone. The evidence-only model runs in 15 ms, so the 0.79 dB of the scan adds 94 ms of latency when executed densely and 27 ms at 5% residency for 83.5% of the gain. The measured advantage grows with input size: the same configurations give $2.1\times$ and $1.8\times$ speedups at 256×256 and $4.7\times$ and $3.2\times$ at 512×512 (the two largest sizes are resized inputs profiled for latency only).

4.5 Residency versus Depth Reduction

Removing unrolled stages is the standard alternative way to cut latency. Dense 5×2 and 4×2 models (Table 2) lose 0.18 and 0.27 dB for $1.2\times$ and $1.5\times$ speedups, whereas the 5% model keeps all six stages, loses 0.13 dB, and is $2.6\times$ faster. On the measured points, shortening the recurrence retains more quality per unit of latency saved than removing reconstruction stages.

4.6 Generalization Across Datasets

The operating points were chosen on fastMRI $8\times$; Tables 4 and 5 evaluate them against external baselines on five datasets at $4\times$ and $8\times$. The dense model is ahead of every external baseline in every setting; over HiFi-Mamba, the strongest of them, by 0.2–1.0 dB on fastMRI and CC359 and by 0.3–1.1 dB on Prostate158, ACDC, and M4Raw. The budgeted models stay close to it: the largest loss anywhere is 0.23 dB (Prostate158, 5%, $4\times$), the 20% model never loses more than 0.14 dB, both budgets exceed dense on CC359, and on M4Raw the 20% model exceeds dense while the 5% model is within 0.06 dB. The two operating points chosen on one benchmark retain near-dense quality across anatomies, sampling patterns, and coil settings. Figure 3 shows reconstructions and error maps of the dense model against the baselines; Appendix E shows the budgeted models against dense.

4.7 Ablations of Budgeted Execution

Section 4.2 fixed what the scan carries. The ablations in Table 6 keep that architecture and vary how budgeted execution treats unselected content, where the resident positions are placed, and how they are scored, all on fastMRI $8\times$ at the 5% and 20% targets.

Method	fastMRI				CC359			
	4×		8×		4×		8×	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Zero-Filling	$29.25^{\pm 1.20}$	$0.723^{\pm 0.049}$	$25.95^{\pm 1.18}$	$0.620^{\pm 0.063}$	$24.79^{\pm 0.46}$	$0.725^{\pm 0.013}$	$21.27^{\pm 0.38}$	$0.576^{\pm 0.010}$
U-Net	$31.66^{\pm 1.60}$	$0.798^{\pm 0.052}$	$28.60^{\pm 1.45}$	$0.697^{\pm 0.070}$	$28.27^{\pm 0.60}$	$0.847^{\pm 0.010}$	$24.28^{\pm 0.62}$	$0.720^{\pm 0.014}$
ISTA-Net	$33.27^{\pm 1.85}$	$0.832^{\pm 0.049}$	$29.44^{\pm 1.55}$	$0.714^{\pm 0.071}$	$32.03^{\pm 0.70}$	$0.902^{\pm 0.009}$	$25.44^{\pm 0.65}$	$0.744^{\pm 0.014}$
TransUNet	$31.84^{\pm 1.65}$	$0.810^{\pm 0.052}$	$29.01^{\pm 1.50}$	$0.711^{\pm 0.076}$	$30.20^{\pm 0.70}$	$0.890^{\pm 0.007}$	$25.69^{\pm 0.72}$	$0.770^{\pm 0.015}$
Mamba-UNet	$33.68^{\pm 1.97}$	$0.839^{\pm 0.047}$	$30.57^{\pm 1.62}$	$0.742^{\pm 0.070}$	$31.08^{\pm 0.38}$	$0.890^{\pm 0.006}$	$25.80^{\pm 0.57}$	$0.761^{\pm 0.014}$
ReconFormer	$33.75^{\pm 1.95}$	$0.837^{\pm 0.049}$	$30.42^{\pm 1.70}$	$0.728^{\pm 0.071}$	$32.46^{\pm 0.70}$	$0.906^{\pm 0.009}$	$26.47^{\pm 0.68}$	$0.766^{\pm 0.015}$
FPS-Former	$33.74^{\pm 1.95}$	$0.841^{\pm 0.049}$	$30.63^{\pm 1.70}$	$0.732^{\pm 0.071}$	$32.35^{\pm 0.70}$	$0.897^{\pm 0.009}$	$26.65^{\pm 0.68}$	$0.768^{\pm 0.015}$
LMO	$34.49^{\pm 2.05}$	$0.846^{\pm 0.048}$	$31.10^{\pm 1.75}$	$0.744^{\pm 0.071}$	$35.35^{\pm 0.75}$	$0.921^{\pm 0.008}$	$27.99^{\pm 0.72}$	$0.787^{\pm 0.015}$
HiFi-Mamba	$34.47^{\pm 2.24}$	$0.853^{\pm 0.048}$	$31.38^{\pm 1.85}$	$0.758^{\pm 0.072}$	$35.74^{\pm 0.75}$	$0.935^{\pm 0.009}$	$28.08^{\pm 0.81}$	$0.802^{\pm 0.015}$
ResiMamba (dense)	$\mathbf{34.86^{\pm 1.91}}$	$\mathbf{0.865^{\pm 0.048}}$	$\mathbf{31.90^{\pm 1.43}}$	$\mathbf{0.773^{\pm 0.072}}$	$36.73^{\pm 0.76}$	$0.954^{\pm 0.007}$	$28.30^{\pm 0.74}$	$0.830^{\pm 0.016}$
ResiMamba (5%)	$34.76^{\pm 1.95}$	$0.861^{\pm 0.050}$	$31.77^{\pm 1.47}$	$0.771^{\pm 0.073}$	$36.77^{\pm 0.79}$	$0.954^{\pm 0.008}$	$28.32^{\pm 0.76}$	$0.831^{\pm 0.016}$
ResiMamba (20%)	$34.84^{\pm 1.93}$	$0.863^{\pm 0.048}$	$31.85^{\pm 1.45}$	$\mathbf{0.773^{\pm 0.072}}$	$\mathbf{36.88^{\pm 0.78}}$	$\mathbf{0.955^{\pm 0.007}}$	$\mathbf{28.37^{\pm 0.75}}$	$\mathbf{0.832^{\pm 0.016}}$

Table 4: Quantitative comparison on fastMRI and CC359 under equispaced Cartesian sampling at $4\times$ and $8\times$ acceleration. PSNR is in dB. Dense is the full-recurrence configuration; 5% and 20% are target residency (6.25% and 20.31% effective).

AF	Method	Prostate158		ACDC		M4Raw	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
4×	U-Net	$25.96^{\pm 0.63}$	$0.772^{\pm 0.020}$	$25.19^{\pm 1.20}$	$0.796^{\pm 0.031}$	$30.02^{\pm 1.10}$	$0.770^{\pm 0.030}$
	TransUNet	$25.45^{\pm 0.57}$	$0.788^{\pm 0.027}$	$26.18^{\pm 1.25}$	$0.813^{\pm 0.031}$	$30.59^{\pm 1.08}$	$0.762^{\pm 0.031}$
	Mamba-UNet	$27.05^{\pm 0.67}$	$0.776^{\pm 0.025}$	$27.58^{\pm 1.15}$	$0.812^{\pm 0.029}$	$31.07^{\pm 0.93}$	$0.777^{\pm 0.029}$
	LMO	$28.15^{\pm 0.80}$	$0.808^{\pm 0.025}$	$28.74^{\pm 1.35}$	$0.833^{\pm 0.030}$	$29.86^{\pm 1.57}$	$0.769^{\pm 0.022}$
	HiFi-Mamba	$28.58^{\pm 0.86}$	$0.827^{\pm 0.023}$	$32.41^{\pm 1.91}$	$0.916^{\pm 0.027}$	$31.13^{\pm 0.82}$	$0.778^{\pm 0.031}$
	ResiMamba (dense)	$\mathbf{29.20^{\pm 0.89}}$	$\mathbf{0.840^{\pm 0.023}}$	$\mathbf{33.49^{\pm 1.96}}$	$\mathbf{0.926^{\pm 0.026}}$	$31.90^{\pm 1.08}$	$0.794^{\pm 0.033}$
	ResiMamba (5%)	$28.97^{\pm 0.88}$	$0.836^{\pm 0.027}$	$33.34^{\pm 1.94}$	$0.923^{\pm 0.033}$	$31.89^{\pm 1.07}$	$0.794^{\pm 0.022}$
	ResiMamba (20%)	$29.06^{\pm 0.78}$	$0.838^{\pm 0.031}$	$33.44^{\pm 1.95}$	$0.925^{\pm 0.026}$	$\mathbf{32.05^{\pm 1.08}}$	$\mathbf{0.797^{\pm 0.028}}$
8×	U-Net	$21.70^{\pm 0.48}$	$0.584^{\pm 0.028}$	$22.21^{\pm 1.10}$	$0.640^{\pm 0.041}$	$28.14^{\pm 0.90}$	$0.724^{\pm 0.034}$
	TransUNet	$21.47^{\pm 0.44}$	$0.565^{\pm 0.032}$	$22.69^{\pm 1.11}$	$0.663^{\pm 0.041}$	$28.95^{\pm 0.90}$	$0.728^{\pm 0.034}$
	Mamba-UNet	$22.35^{\pm 0.46}$	$0.596^{\pm 0.028}$	$23.78^{\pm 1.09}$	$0.660^{\pm 0.037}$	$29.43^{\pm 0.84}$	$0.738^{\pm 0.033}$
	LMO	$22.68^{\pm 0.47}$	$0.600^{\pm 0.030}$	$24.18^{\pm 1.18}$	$0.677^{\pm 0.039}$	$28.35^{\pm 1.19}$	$0.728^{\pm 0.026}$
	HiFi-Mamba	$23.49^{\pm 0.60}$	$0.647^{\pm 0.036}$	$27.06^{\pm 1.49}$	$0.792^{\pm 0.042}$	$29.65^{\pm 0.82}$	$0.741^{\pm 0.032}$
	ResiMamba (dense)	$\mathbf{23.90^{\pm 0.65}}$	$\mathbf{0.659^{\pm 0.035}}$	$\mathbf{27.81^{\pm 1.52}}$	$0.807^{\pm 0.041}$	$29.99^{\pm 0.81}$	$\mathbf{0.745^{\pm 0.032}}$
	ResiMamba (5%)	$23.75^{\pm 0.63}$	$0.657^{\pm 0.033}$	$27.67^{\pm 1.44}$	$0.806^{\pm 0.045}$	$29.93^{\pm 0.85}$	$0.744^{\pm 0.041}$
	ResiMamba (20%)	$23.83^{\pm 0.39}$	$\mathbf{0.659^{\pm 0.041}}$	$27.78^{\pm 1.51}$	$\mathbf{0.808^{\pm 0.041}}$	$\mathbf{30.06^{\pm 0.81}}$	$\mathbf{0.745^{\pm 0.034}}$

Table 5: Quantitative comparison on Prostate158, ACDC, and M4Raw under $4\times$ and $8\times$ acceleration. Zero filling, ISTA-Net, ReconFormer, and FPS-Former are reported on fastMRI and CC359 only.

Preserving non-resident content. Two controls keep the window quota and the packed scan and change what happens to unselected carrier content: pruning discards it, and a ToR-SSM-style control ranks positions by their feature magnitude, merges the unselected ones into the retained tokens by similarity, and propagates the retained outputs back to the merged positions (Zhan et al., 2024b). Routing the unselected content into the dense evidence path is better than both, by 0.16–0.19 dB over pruning and 0.14 dB over merging, with the best SSIM and NMSE in each group. Pruning differs from ResiMamba only in the fate of the unselected content, so its gap isolates the effect of preserving that content.

Spatial coverage under a fixed budget. With the target budget and the scorer fixed, changing only the selection region orders the variants identically at both budgets: 8×8 windows are best, 4×4 and 16×16 windows follow within 0.06 dB, and global top-K over the whole grid is last, 0.11 dB and 0.07 dB below the default. Each window size follows its own quota rounding and global selection uses the target fraction; on the two pairs with identical executed counts, 8×8 still leads by 0.03–0.04 dB, so placement matters at equal executed count.

What learned selection adds. With windows, quotas, evidence path, and packing fixed, the learned scorer beats uniform, feature-magnitude, and random selection by 0.09–0.13, 0.07, and 0.17–0.23 dB respectively. Random placement within the windows is the worst policy, but the best fixed

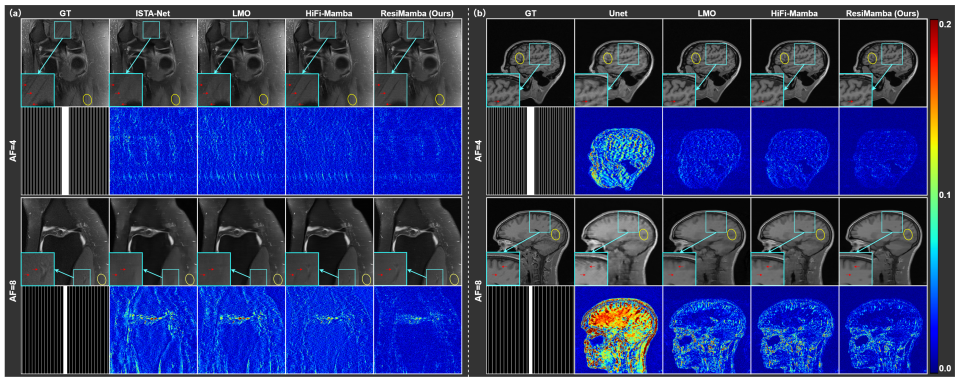


Figure 3: Dense ResiMamba versus external baselines on (a) fastMRI and (b) CC359 at $4\times$ and $8\times$, with sampling masks and absolute-error maps on a shared color range.

Variant	5% target			20% target		
	PSNR ↑	SSIM ↑	NMSE ↓	PSNR ↑	SSIM ↑	NMSE ↓
Non-resident content (same window quota and packed scan)						
Evidence route (ResiMamba)	**31.77**	**0.771**	**0.0197**	**31.85**	**0.773**	**0.0195**
Similarity merging (ToR-SSM-style)	31.63	0.766	0.0198	31.71	0.769	0.0196
Token pruning	31.58	0.764	0.0200	31.69	0.769	0.0197
Selection region (same target budget and scorer; effective residency follows the quota rule)						
Global	31.66	0.762	0.0199	31.78	0.768	0.0197
16×16 windows	31.71	0.767	0.0198	31.82	0.771	0.0196
8×8 windows (default)	**31.77**	**0.771**	**0.0197**	**31.85**	**0.773**	**0.0195**
4×4 windows	31.73	0.768	0.0198	31.83	0.772	0.0196
Selection policy (same window quota, evidence pathway, and packed execution)						
Learned (default)	**31.77**	**0.771**	**0.0197**	**31.85**	**0.773**	**0.0195**
Uniform	31.68	0.765	0.0199	31.72	0.768	0.0198
Feature magnitude	31.70	0.765	0.0199	31.78	0.771	0.0196
Random	31.60	0.762	0.0200	31.62	0.761	0.0200

Table 6: Residency ablations on fastMRI $8\times$ at the 5% and 20% targets (6.25% and 20.31% effective with 8×8 windows; the region block follows each window size.s quota). The default row is the same model in all three blocks.

policy, feature magnitude, comes within 0.07 dB of the learned scorer at both budgets: near-dense quality does not require a sophisticated selector. The two principal budgets, retrained with a second seed, reproduce their PSNR to within 0.008 dB (Appendix D.5).

Implementation studies. Appendix D reports the remaining comparisons: hard straight-through routing is within 0.02 dB of a Gumbel-perturbed variant and 1.9 dB better than a full-sequence soft-mask relaxation; a gap embedding and pre-construction scoring change PSNR by $+0.002$ and -0.095 dB. Profiling assigns about half of the budgeted runtime to the packed scan and 17–19% to scoring, packing, and scattering; the dense evidence path, fusion, and DC take the rest, which is why speedup grows more slowly than the inverse residency ratio.

5 Conclusion

We asked how much recurrence a state-space MRI reconstruction needs and built an architecture in which the content the scan carries and the positions that execute it are separate, controllable choices, with unselected content preserved in a dense evidence path. Recurrence turns out to be a sparse resource: a spatially distributed 6.25% of positions recovers 83.5% of what dense recurrence adds on fastMRI $8\times$, and the budgeted configurations stay near-dense across five datasets while cutting latency by two to four times, more than depth reduction achieves. Recurrent participation is then a design axis alongside depth and width; nothing in the mechanism is specific to MRI; other dense prediction tasks and budget-conditioned single models are the natural next steps.

Ethics Statement

This work uses publicly released, de-identified MRI datasets under their original licences and access conditions and collects no new human-subject data. The reported reconstructions are evaluated as a methodological comparison and are not intended for clinical diagnosis without further validation.

Reproducibility Statement

Section 3 defines the carrier/evidence construction, resident selection, evidence routing, and conditional recurrence; Algorithm 1 (Appendix C) gives the complete budgeted unit. Section 4.1 specifies the principal training and evaluation settings, and Appendices A–D provide dataset configurations, metric aggregation, architectural settings, additional ablations, component profiling, and independent-seed comparisons. The quality–latency results report input dimensions, hardware, and a synchronized timing protocol that excludes loading and warm-up. All datasets are public.

AI Use Statement

Large language models were used as general-purpose writing and coding assistants. Their use included drafting and editing prose, organizing and checking manuscript sections, formatting LaTeX content, assisting with code development and debugging, and supporting experiment orchestration and result aggregation. The authors determined the research questions, method, experimental design, and interpretation of the results. They verified the reported results and take full responsibility for the entire paper, including text and code produced with AI assistance.

References

Lisa C. Adams, Marcus R. Makowski, Günther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M. Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, and Keno K. Bressem. Prostate158 - an expert-annotated 3t mri dataset and algorithm for prostate cancer detection. *Computers in Biology and Medicine*, 148:105817, 2022. doi: 10.1016/j.compbiomed.2022.105817.

Hemant Kumar Aggarwal, Merry P. Mani, and Mathews Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 38(2):394–405, 2019. doi: 10.1109/TMI.2018.2865356.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohe, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Isgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018. doi: 10.1109/TMI.2018.2837502.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *International Conference on Learning Representations*, 2023.

Hongli Chen, Pengcheng Fang, Yuxia Chen, Yingxuan Ren, Jing Hao, Fangfang Tang, Xiaohao Cai, Shanshan Shan, and Feng Liu. Hifi-mamba: Dual-stream w-laplacian enhanced mamba for high-fidelity mri reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2868–2876, 2026.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*, 2024.

Pengfei Guo, Yiqun Mei, Jinyuan Zhou, Shanshan Jiang, and Vishal M. Patel. Reconformer: Accelerated mri reconstruction using recurrent transformer. *IEEE Transactions on Medical Imaging*, 43(1):582–593, 2024. doi: 10.1109/TMI.2023.3314747.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018. doi: 10.1002/mrm.26977.

Jiahao Huang, Yingying Fang, Yinzhe Wu, Huanjun Wu, Zhifan Gao, Yang Li, Javier Del Ser, Jun Xia, and Guang Yang. Swin transformer for fast mri. *Neurocomputing*, 493:281–304, 2022. doi: 10.1016/j.neucom.2022.04.051.

Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=HwCvaJOiCj`.

Wei Li, Jiawei Jiang, Jie Wu, Kaihao Yu, and Jianwei Zheng. Lmo: Linear mamba operator for mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5112–5122, 2025.

Mengye Lyu, Lifeng Mei, Shoujin Huang, Sixing Liu, Yi Li, Kexin Yang, Yilong Liu, Yu Dong, Linzheng Dong, and Ed X. Wu. M4raw: A multi-contrast, multi-repetition, multi-channel mri k-space dataset for low-field mri research. *Scientific Data*, 10(1):264, 2023. doi: 10.1038/s41597-023-02181-4.

Yucong Meng, Zhiwei Yang, Yonghong Shi, and Zhijian Song. Boosting vit-based mri reconstruction from the perspectives of frequency modulation, spatial purification, and scale diversification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6135–6143, 2025. doi: 10.1609/aaai.v39i6.32656.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Leticia Rittner, Richard Frayne, and Roberto Lotufo. An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, 170:482–494, 2018. doi: 10.1016/j.neuroimage.2017.08.021.

Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6387–6397, 2023.

Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018. doi: 10.48550/arXiv.1811.08839.

Zheng Zhan, Zhenglun Kong, Yifan Gong, Yushu Wu, Zichong Meng, Hangyu Zheng, Xuan Shen, Stratis Ioannidis, Wei Niu, Pu Zhao, and Yanzhi Wang. Exploring token pruning in vision state space models. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Zheng Zhan, Yushu Wu, Zhenglun Kong, Changdi Yang, Yifan Gong, Xuan Shen, Xue Lin, Pu Zhao, and Yanzhi Wang. Rethinking token reduction for state space models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1686–1697, Miami, Florida, USA, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.100.

Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1828–1837, 2018.

A Dataset and Evaluation Details

Dataset composition. The fastMRI knee collection contains $1{,}172$ complex-valued single-coil examinations; the experiments use its proton-density fat-suppressed subset at 320×320. The fastMRI $8\times$ budget evaluation comprises 99 volumes and $1{,}980$ slices. CC359 uses 25 training and 10 evaluation subjects in the reconstruction benchmark, with 256×256 slices. ACDC uses 100 training and 50 evaluation cases and 128×128 center crops. The M4Raw configuration uses 384 training and 150 evaluation T1-weighted volumes from its four-coil, $0.3\,\mathrm{T}$ acquisition. Prostate158 uses 139 training and 19 evaluation cases. These are the dataset-specific benchmark configurations used for the comparisons.

Sampling and representation. fastMRI uses single-coil proton-density fat-suppressed knee data at 320×320 and CC359 brain images at 256×256. Following the pipeline of Chen et al. (2026), the magnitude reconstruction of each slice is min–max normalized to $[0, 1]$, its k-space is synthesized by a Fourier transform, and equispaced Cartesian undersampling with a fully sampled center of $0.32/R$ of the lines at acceleration R is applied retrospectively; the central 20 slices of each fastMRI volume are used. Prostate158 uses random masks. ACDC uses golden-angle radial masks with 32 and 16 spokes for the $4\times$ and $8\times$ settings and 128×128 center crops. M4Raw supplies four-channel multi-coil brain k-space acquired at $0.3\,\mathrm{T}$ with equispaced Cartesian sampling; its four complex coil channels form eight real-valued input channels. For complex single-coil inputs, real and imaginary components are represented separately. Each dataset uses its official split. Zero filling, ISTA-Net, ReconFormer, and FPS-Former are reported on fastMRI and CC359 only; the other baselines follow their published configurations under the common data and mask settings.

Metric aggregation. Reconstructed magnitude images are clipped to $[0, 1]$ before scoring. PSNR uses the ground-truth dynamic range of the evaluated slice; SSIM uses Gaussian weights with $\sigma = 1.5$ and data range 1. Slice NMSE is $\|\widehat{x} - x\|_2^2/\|x\|_2^2$. For each metric, slice values are averaged within a case and case values are averaged over cases; the $\pm$ values in the cross-dataset tables are case-level standard deviations. Appendix ablations reported as differences use the same evaluation and aggregation for the variant and its matched reference.

B Architecture and Training Configuration

The real and imaginary parts of the zero-filled image are embedded with a 2×2, stride-two convolution and LayerNorm, so a 320×320 image produces a 160×160 feature grid. Each of the six stages applies two units with separate parameters in sequence, each unit being internally residual ($\mathbf{Y} = \mathbf{X} + \mathrm{DropPath}(\cdot)$), then decodes the features to an image, enforces data consistency by replacing the acquired k-space samples (the image is multiplied by the coil maps, transformed to k-space, the sampled locations are overwritten with the acquired data, and the result is transformed back and coil-combined with the conjugate maps; single-coil data use a unit map, and all masks, including the radial-shaped ACDC masks, act on the Cartesian k-space grid), re-embeds the result through a SiLU activation and the same patch embedding, and adds it back to the stage features; a final LayerNorm precedes the terminal decoding and DC. The evidence refinement Φ_E is a depthwise inverted bottleneck (1×1 expansion by a factor of three, 3×3 depthwise convolution, 1×1 projection, with normalization and activation). The conditioning map P is a token-wise linear layer from the d evidence channels to the four affine parameter groups of the state write/read interfaces. The state is local to each unit.s scan. Table 7 summarizes the shared architecture and the principal training configuration.

The main configurations train for 100 epochs with a magnitude-domain ℓ_1 objective and AdamW (weight decay 0.01, none for biases and one-dimensional parameters), a cosine schedule with five warm-up epochs from 5×10^{-4} to 10^{-6}, and seed 1337; the two principal budgets are retrained with seed 2027 in Appendix D.5. Budgeted models use the learned pointwise scorer, non-overlapping 8×8 windows, and one raster-ordered conditional scan per unit; the 5% and 20% targets correspond to 6.25% and 20.31% effective residency.

The main configurations use the checkpoint selected by reconstruction SSIM. The Gumbel, gap-conditioned, and soft-mask variants use the best reconstruction checkpoint within their training runs. Baselines use the common data and mask settings, with architectural and optimization settings following their published configurations or implementations.

Architecture	Value
Unrolled stages / units per stage	$6/2$
Feature width / branch width	$128/64$
Carrier projection	Learned 3×3 convolution
Carrier filter	Fixed separable $[1, 4, 6, 4, 1]/16$
Evidence refinement	Depthwise inverted bottleneck
Scorer	1×1 conv, GELU, 1×1 conv
Window / scan order	8×8 / raster
Mamba core / state dimension	Mamba-3 / 16
Head dimension / MIMO rank	$32/4$
Scan chunk size	16
Scan output-projection normalization	Disabled
Evidence-conditioning scale init.	$\tau = 0$
Non-resident projection init.	$[\mathbf{I}_d \ \mathbf{0}]$

Training	Value
Optimizer / main epochs	AdamW / 100
Initial / minimum learning rate	5×10^{-4} / 10^{-6}
Schedule / warm-up	Cosine / 5 epochs
Weight decay	0.01; 0 for biases, 1D
Objective	Magnitude-domain ℓ_1
Main seed / additional seed	$1337/2027$
Training mode	Single GPU, FP32

Table 7: Architecture (left) and principal training configuration (right).

C Budgeted Unit Algorithm and Scan Update

Algorithm 1 lists the complete computation of one budgeted reconstruction unit defined in Section 3.

Algorithm 1 Evidence-preserving budgeted recurrence in one unit

Require: Feature grid $\mathbf{X}$, target $\rho \in (0, 1]$, windows $\{\Omega_j\}$, raster order π

Split $\mathrm{LN}(\mathbf{X})$ into $\mathbf{X}^h, \mathbf{X}^\ell$
$\mathbf{U} \leftarrow \phi(\mathbf{X}^\ell)$; $\mathbf{L} \leftarrow \mathrm{Up}_2(\mathcal{P}_\mathbf{b}(\mathbf{U}))$
$\mathbf{E} \leftarrow \mathbf{U} - \mathbf{L}$; $\mathbf{E}_0 \leftarrow \mathbf{X}^h + \mathbf{E}$
Score $[\mathbf{L} \| \mathbf{E}]$ and select top-K_j in each window
Form $\mathbf{M}^{\mathrm{hard}}$ and gate $\mathbf{M}^{\mathrm{gate}}$ ▷ equal to $\mathbf{M}^{\mathrm{hard}}$ at inference
$\mathbf{L}^{\mathrm{r}} \leftarrow \mathbf{M}^{\mathrm{gate}} \odot \mathbf{L}$; $\mathbf{L}^{\mathrm{nr}} \leftarrow (1 - \mathbf{M}^{\mathrm{gate}}) \odot \mathbf{L}$
$\mathbf{G} \leftarrow \Phi_E(\mathcal{P}_E([\mathbf{E}_0 \| \mathbf{L}^{\mathrm{nr}}]))$
$\mathcal{I}_R \leftarrow$ selected indices sorted along π
Gather $\mathbf{L}^{\mathrm{r}}$ and $\mathbf{G}$ with the same $\mathcal{I}_R$ to form $\mathbf{L}_R, \mathbf{G}_R$
$\mathbf{R}_R \leftarrow \mathcal{M}_\theta(\mathbf{L}_R; \mathbf{G}_R)$ ▷ conditioned scan of length K
Scatter $\mathbf{R}_R$ into a zero grid at $\mathcal{I}_R$ to form $\mathbf{R}$
return $\mathbf{X} + \mathrm{DropPath}(\mathcal{P}_F([\mathbf{G} \| \mathbf{R}]))$

Scan update. Writing V_n for the native MIMO write formed from carrier token n and its conditioned write interface, the exponential-trapezoidal scan has the form

$$\begin{aligned} h_0 &= \gamma_0 V_0, \\ h_n &= \alpha_n h_{n-1} + \beta_n V_{n-1} + \gamma_n V_n, \qquad n = 1, \ldots, K-1, \\ \alpha_n &= \exp(\Delta_n A_n), \qquad \beta_n = (1 - \lambda_n)\Delta_n \alpha_n, \qquad \gamma_n = \lambda_n \Delta_n. \end{aligned} \tag{12}$$

Here A_n is the content-side transition generator, Δ_n its discretization step, and λ_n the trapezoidal interpolation coefficient at packed position n; A_n is distinct from the MRI encoding operator $\mathcal{A}$. These are the underlying Mamba-3 updates; rotations, skip/gating operations, and MIMO aggregation retain their native form. Only $\mathbf{L}_R$ supplies content/value factors in V_n. Evidence affects state writes through $\mathbf{B}'_s$ and reads through $\mathbf{C}'_s$, without supplying an additional content sequence. The scan uses state dimension 16, head dimension 32, MIMO rank 4, and chunk size 16.

D Additional Implementation Analysis

D.1 Component Profiling

End-to-end latency is measured on an NVIDIA H100 80 GB with batch size one over a 100-image loop that excludes model/data loading and host–device transfers; the first 20 forwards are warm-up and the remaining 80 are timed with CUDA events and synchronization around each forward, and the arithmetic mean is reported. The compared implementations use the same batch size, warm-up, timed count, and synchronization, and the input size is stated with every timing result. In the

Region	5%	20%
Scoring and top-K	12.9%	10.3%
Gather / pack	5.9%	6.5%
Scatter / unpack	0.4%	0.4%
Packed conditional scan	47.4%	57.6%
Evidence, fusion, DC, embedding	33.5%	25.2%

Table 8: Exclusive-region latency decomposition on fastMRI $8\times$ at 320×320, batch size one, H100, as a fraction of end-to-end latency. Under the same instrumentation the recurrent scan accounts for 86.3% of dense end-to-end latency.

Target	Parameters (M)	Recurrent steps	Evidence-conv. GFLOPs
1%	2.273	4,800	37.709
5%	2.273	19,200	37.709
10%	2.273	33,600	37.709
20%	2.273	62,400	37.709
50%	2.273	153,600	37.709
Dense	2.174	307,200	37.709

Table 9: Parameter and workload accounting for 320×320 image inputs. Recurrent steps are summed over the twelve units on the 160×160 feature grid. The last column is the analytically counted fixed dense evidence-convolution branch, using two FLOPs per multiply–accumulate; it is not a total-network FLOP count.

Estimator	ΔPSNR	ΔSSIM	NMSE
Straight-through	ref.	ref.	0.0197
Gumbel top-K	$+0.02$	$+0.0001$	0.0196
Soft mask (full sequence)	-1.93	-0.0430	0.0296

Table 10: Training relaxations of the residency gate, fastMRI $8\times$ at a 5% target budget, seed 1337. Straight-through is the default; the reference is the 5% operating point of Table 3; alternatives are measured differences from a matched-protocol training run of that operating point. NMSE uses the same slice-to-case aggregation as the reference. The soft-mask row is its best epoch (epoch 7).

Keep ratio	ΔPSNR vs dense (dB)
5%	-1.77
20%	-2.05
50%	-1.38
80%	-0.72

Table 11: Training-free token reduction applied to the trained dense model, fastMRI $8\times$, 99 cases, reported as the paired difference to the dense model under the same evaluation. The weights are fixed in this post-hoc comparison.

Budget	$\lvert\Delta\mathrm{PSNR}\rvert$ (dB)
5%	0.008
20%	0.001

Table 12: Independent retrainings of the two reported operating points on fastMRI $8\times$ with seeds 1337 and 2027, evaluated under a single protocol.

dedicated component profile, mutually exclusive CUDA-event regions measure scoring/selection, packing, scanning, scattering, and the remaining dense work. Instrumentation adds approximately 3% to total forward time; Table 8 reports within-run fractions, while the main latency tables use uninstrumented forwards. Rounded fractions may not sum to exactly 100%.

Scoring/selection, gather/pack, and scatter/unpack together account for 19.2% at the 5% budget and 17.2% at 20%. Scoring and top-K are measured together, so this region includes both the learned predictor and the selection operation. The scan accounts for 86.3% of dense runtime, but 47.4% and 57.6% in the two budgeted configurations. The non-scan components consequently become important at low residency. This decomposition is consistent with the smaller difference between the lowest-budget latencies at 256×256 than at 512×512.

Table 9 separates parameters, executed recurrent positions, and the fixed convolutional work in the evidence branch. The recurrent count changes with the selected positions, whereas the parameter count is shared by the sparse configurations and the evidence branch remains dense.

D.2 Residency-Gate Gradient Estimators

The default estimator uses a hard top-K forward assignment and a sigmoid surrogate derivative. The Gumbel variant perturbs the training logits with independent $\mathrm{Gumbel}(0, 1)$ noise and uses an annealed sigmoid surrogate. The soft-mask variant multiplies carrier features by continuous sigmoid weights during training and hard window-wise masks at evaluation, executing the full sequence in both cases; it therefore changes the executed recurrence as well as the gradient. Table 10 gives the matched 5% comparison.

Gumbel perturbation gives a small positive change, while the continuous relaxation has a substantially larger discretized quality gap. After its best epoch, the latter loses approximately a further 2.4 dB by epoch 32 even as its training objective improves. The default avoids this particular train–inference discrepancy by using hard routing in both forward computations.

D.3 Gap Encoding and Selection Placement

The gap variant adds a zero-initialized embedding of the $\log_2$-binned raster-index distance to the preceding resident position to each packed token. At a matched 5% budget, the changes are $+0.002$ dB PSNR, less than 10^{-4} SSIM, and unchanged NMSE of 0.0197. For selection placement, the pre-construction alternative scores the unit input rather than $[\mathbf{L}\|\mathbf{E}]$. At a matched 50% target, its PSNR is 0.095 dB lower than the post-construction configuration.

D.4 Post-Hoc Reduction of a Fixed Dense Model

Table 11 applies importance-ranked keeping and window-local similarity merging to the trained dense model without retraining. Removing 80–95% of the recurrent steps this way costs 1.77–2.05 dB (non-monotone in the keep ratio, since the dense scan was never trained to skip positions), and even 80% keeping costs 0.72 dB, whereas the budget-trained 5% and 20% models are within 0.13 dB of dense (Table 3). The tested post-hoc procedure does not reproduce the frontier; the frontier characterizes models optimized at their budget, which is why each budget is trained as its own configuration.

D.5 Independent Seeds

The 5% and 20% configurations were each retrained with seed 2027 and compared with seed 1337 under the same evaluation. Table 12 reports the absolute PSNR difference for each pair, separately from the case-level standard deviations in the main tables.

E Qualitative Results

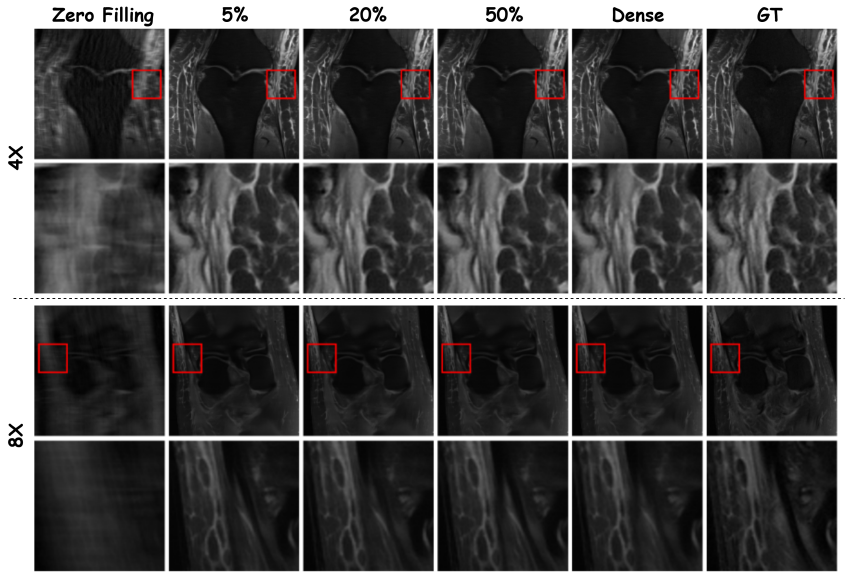


Figure 4: Budgeted versus dense recurrence on fastMRI at $4\times$ and $8\times$. Columns: zero filling, ResiMamba at 5%, 20%, 50%, dense, and ground truth; the second row of each block enlarges the boxed region.

Figure 4 compares the budgeted models with dense recurrence and the reference image; the 5% and 20% configurations remain close to dense at edges and fine texture; the enlarged regions illustrate the remaining differences.