RESFUSER: DECOMPOSING RESIDUAL FUSION IN AGGRESSIVELY TOKEN-DROPPED VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Transformers operating on dense scientific and medical inputs face a sequence-length bottleneck, motivating aggressive token dropping while preserving the information needed for restoration. We introduce **ResFuser**, a unified investigation framework that compares two existing residual topologies and four token-sampling rules with a shared backbone and controlled token and training budgets. Across five empirical cells spanning climate, medical, and natural-image restoration in deterministic and diffusion regimes, both full-resolution residual routes substantially improve over the no-residual ablation, while their relative ranking changes across cells. Within the evaluated 7.4M-parameter backbone, the better topology in each cell gives validation RMSE within $\pm 7\%$ of a dense ViT/DiT baseline at 42–59% fewer analytical forward FLOPs. The tested sampling rules produce smaller changes than residual design; masked-region analysis also shows that input salience need not identify the regions where restoration quality improves. These findings motivate a staged design procedure: establish a residual route, compare topologies under the target budget, and then evaluate task-aligned sampling. ResFuser decomposes these choices for dense restoration.

1 INTRODUCTION

Vision Transformers (Dosovitskiy et al., 2020) support discriminative and generative vision, including diffusion models for natural images (Ma et al., 2024; Peebles & Xie, 2023; Rombach et al., 2022), text-to-image generation, and video synthesis (Esser et al., 2024; OpenAI, 2024). Their adoption in medical imaging (Chung & Ye, 2022; Chung et al., 2022; Hatamizadeh et al., 2021), weather forecasting (Bi et al., 2023; Lam et al., 2023; Nguyen et al., 2023), and downscaling (Wang et al., 2025) exposes a sequence-length bottleneck: a 256^3 MRI volume with patch size 8 produces roughly 32K tokens, and a 128×256 ERA5 field with patch size 2 produces roughly 8K. Both exceed typical natural-image sequences of 256–1024 tokens. Since attention scales quadratically with sequence length (Tay et al., 2022), token reduction can lower computation while preserving quality.

Sequence-length reduction includes learned pruning in DynamicViT (Rao et al., 2021) and token merging in ToMe (Bolya et al., 2022) and ToMe-SD (Bolya & Hoffman, 2023). MaskDiT (Zheng et al., 2023), MicroDiT (Sehwag et al., 2025), MDTv2 (Gao et al., 2023), TREAD (Krause et al., 2025), and SPRINT (Park et al., 2025) combine sparse computation with training or structural recovery. SPRINT shows that 75% token dropping is feasible in Diffusion Transformer (DiT) models on ImageNet when a sequential residual skip preserves full-resolution information and structured-uniform sampling maintains spatial coverage. This motivates examining how the residual route interacts with token selection in dense restoration.

We examine two design questions. **First, the architectural locus of the residual.** How does SPRINT's encoder-feature bypass compare with an independent raw-input route? ORBIT-2 (Wang et al., 2025; LeCun et al., 1989; Ronneberger et al., 2015) uses a parallel-CNN residual for climate downscaling without token dropping, providing an alternative route to adapt into the same sparse transformer. **Second, the token-sampling rule.** At fixed token budget and residual topology, do image gradients or learned feature magnitudes improve on random or structured-uniform coverage? Dense restoration tests this assumption directly: salient input patches need not coincide with output

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.

Table 1: Five empirical cells used throughout. Token counts and dense no-drop GFLOPs are computed at patch size $p = 4$ for the dense ViT/DiT baseline with $L_M = 8$ middle blocks and $d = 256$.

Cell	Domain	Task	Regime	Tokens	GFLOPs
ERA5-det	Climate	Downscaling	Deterministic	2,048	90.4
ERA5-diff	Climate	Downscaling	EDM	2,048	90.5
fastMRI	Medical	Denoising	Deterministic	4,096	283.5
BraTS-GLI	Medical	Restoration	Deterministic	4,096	283.7
DIV2K	Natural images	RGB denoising	Deterministic	4,096	283.7

regions requiring improvement. In brain MRI, for example, strong boundaries may receive high importance scores while the desired improvement concerns interior tissue.

We develop **ResFuser**, a unified dual-path transformer framework for comparing residual designs and token-sampling rules under aggressive token dropping. ResFuser runs a sparse main transformer path alongside a residual route that retains full-resolution patch-grid features. The framework supports two implementations drawn from these designs: (i) an ORBIT-2-style parallel CNN (Wang et al., 2025; LeCun et al., 1989; Ronneberger et al., 2015) that processes the raw input, and (ii) a SPRINT-style sequential transformer-skip (Park et al., 2025) that originates from the encoder output. We study random, structured-uniform, gradient-based, and activation-based sampling across the available ablations. We focus on dense image restoration, where every output location has a target and dropped positions must be reconstructed from the retained tokens and any residual features. We evaluate five empirical cells (Table 1) spanning deterministic regression and EDM diffusion (Karras et al., 2022): ERA5 climate downscaling (Hersbach et al., 2020; Muñoz-Sabater et al., 2021; Wang et al., 2026); fastMRI brain image Rician denoising (Knoll et al., 2020; Zbontar et al., 2018); BraTS-GLI multi-contrast brain MRI (Baid et al., 2021; Bakas et al., 2017; Correia De Verdier et al., 2024); and DIV2K natural-image RGB Gaussian denoising (Timofte et al., 2018).

Contribution and scope. Our contribution is an empirical decomposition of three decisions: retaining a dense information route, choosing its source and transformation, and selecting tokens for the sparse stack. We standardize established residual components within a shared small backbone. The dense baseline anchors the cost of sparsifying that backbone, while ResFuser comparisons examine architecture selection under controlled token and training budgets.

Our contributions are threefold. First, ResFuser standardizes two existing residual topologies within one training and evaluation pipeline. The better topology per cell achieves validation RMSE within $\pm 7\%$ of the corresponding dense baseline at 42–59% fewer analytical forward FLOPs. Second, the residual ablations separate a consistent benefit from preserving a dense route from a cell-dependent preference for its implementation: no-residual RMSE is $2.2\times$–$9.4\times$ that of the corresponding residual variants across the five cells. Mechanism diagnostics further show that neither channel count nor channel redundancy alone predicts the topology ranking. Third, sampler ablations at fixed topology and token budget show smaller observed changes than residual design, with gradient-based selection improving ERA5-det but underperforming uniform coverage on several fastMRI evaluation masks. These results favor testing residual routes before token selection.

2 Methods

ResFuser aligns a token-dropped main transformer and a full-resolution residual on the patch grid, then fuses them before decoding (Figure 1).

2.1 Architecture Overview

Notation. Let $u \in \mathbb{R}^{C_{\text{in}} \times H \times W}$ denote the tensor presented to the network. In deterministic restoration, $u = x_{\text{deg}}$ is the degraded input. In conditional diffusion, $u = concat(y_t, x_{\text{LR}})$ is the channel-wise concatenation of the noisy high-resolution sample and the upsampled low-resolution condition. A patch embedding with patch size p maps u to tokens $z \in \mathbb{R}^{N \times d}$, where $N = \frac{HW}{p^2}$ and d is the token dimension. Unless otherwise stated, we use $p = 4$ throughout the paper, giving

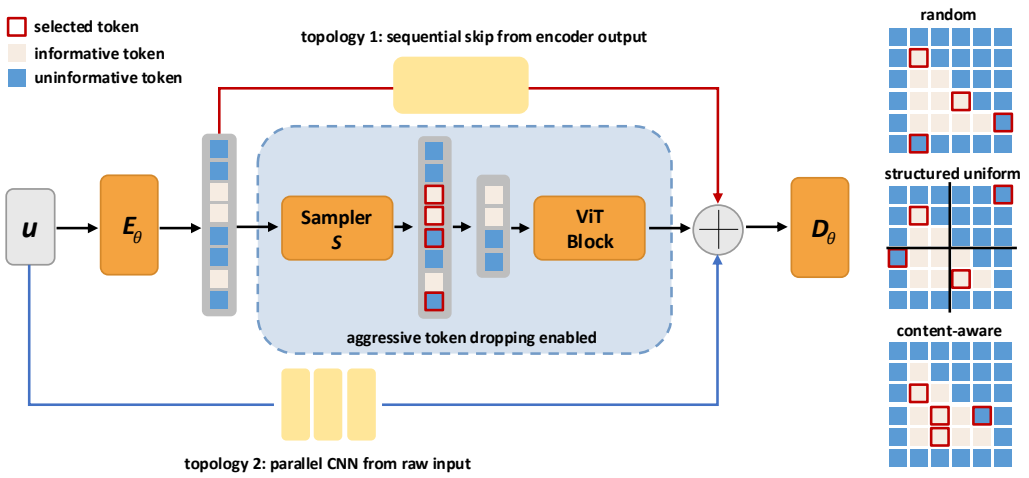


Figure 1: **ResFuser.** Comparing residual routes and token samplers under aggressive dropping.

$N = 4096$ for $H = W = 256$ and $N = 2048$ for $H \times W = 128 \times 256$. Let $r \in [0, 1)$ be the drop ratio. The sparse main path therefore operates on $N' = \lceil (1 - r)N \rceil$ kept tokens.

Forward pass. Figure 1 summarizes the architecture. An encoder E produces $z^{\text{enc}} = E(z)$; a sampler S selects $|\mathcal{K}| = N'$ indices, and the middle stack M processes $z^{\text{sparse}} = z^{\text{enc}}_{\mathcal{K}}$. A residual module R supplies full-resolution features $z^{\text{res}} \in \mathbb{R}^{N \times d}$ from the raw input or encoder output. We restore the sparse output to length N with learned mask tokens, fuse it with z^{res}, and decode it into an image. Only S and R vary; the remaining components stay fixed.

2.2 Main Path: Token Dropping and Sampling

The main-path cost is dominated by the middle stack M, where attention scales quadratically in sequence length N'. We evaluate aggressive dropping in $r \in \{0.5, 0.75\}$ for the main comparisons using the following token sampling strategies, with stress tests at $r = 0.875$.

Group-coverage constraint. For structured samplers, we partition the patch grid into non-overlapping spatial groups G_q of size $g \times g$ and keep n_{keep} tokens per group. When $(1 - r)g^2$ is an integer, we set $n_{\text{keep}} = (1 - r)g^2$ exactly; otherwise keep counts are distributed across groups to match the global budget N'. The main 50% and 75% drop experiments use $g = 2$. Higher-drop stress tests use larger groups under the same rule.

Sampling rules. We compare four samplers during training and validation, fixing validation seeds for stochastic variants. *Random* samples $\mathcal{K}$ uniformly from all N token indices. *Structured-uniform* follows the SPRINT coverage rule (Park et al., 2025), sampling n_{keep} tokens per group G_q. *Gradient* sampling scores tokens by input Laplacian energy, $s_t = \sum_c \sum_{(i,j) \in \text{patch}(t)} |(\Delta u_c)_{i,j}|$, and keeps the top tokens per group. *Activation* sampling scores tokens by first-block feature norm, $h^{(1)} = B_1(z + e_{\text{pos}})$, $s_t = \|h^{(1)}_t\|_2$, then keeps the top tokens per group.

Restoring dropped positions. After the sparse middle stack, $\tilde{z}^{\text{sparse}} \in \mathbb{R}^{N' \times d}$ is restored to length N by inserting a learned mask token $m \in \mathbb{R}^d$ at dropped positions. This restoration is required because the residual path remains full-resolution on the patch grid.

2.3 Residual Path: Two Topologies

The residual path provides full-resolution structural information that bypasses the sparse middle stack. Both topologies produce $z^{\text{res}} \in \mathbb{R}^{N \times d}$ aligned with $\tilde{z}^{\text{full}}$, and both fuse into the main path at the

same point, immediately before the decoder, which we hold constant for controlled comparison. The variants differ in (i) where the residual originates and (ii) what encoding pathway it uses.

Parallel CNN. Following ORBIT-2 (Wang et al., 2025), the CNN consumes u independently of the ViT encoder and produces token-aligned features through three 3×3 Conv2D–GroupNorm–GELU blocks (hidden width $C_{\text{hid}} = 64$), a 1×1 projection to d channels, and average pooling with stride p. We standardize its fusion location immediately before the decoder. This route provides translation-equivariant local processing with parameters independent of the transformer.

Sequential transformer-skip. Adapted from SPRINT (Park et al., 2025), this route projects encoder features per token: $z^{\text{res}} = z^{\text{enc}} W^{\text{skip}}$, with $W^{\text{skip}} \in \mathbb{R}^{d \times d}$. SPRINT forwards encoder features directly to fusion; our intermediate projection separates residual transformation from fusion in both variants. Parameter counts remain different. The projection reweights features, with cross-token mixing supplied by the shared encoder. This route shares encoder parameters with the main path and requires much less residual computation than the parallel CNN.

After restoration, the main-path tensor $\tilde{z}^{\text{full}}$ and residual tensor z^{res} have the same shape and are fused by channel-wise concatenation followed by a linear projection: $z^{\text{fuse}} = \text{concat}(\tilde{z}^{\text{full}}, z^{\text{res}}) W^{\text{fuse}}$ with $W^{\text{fuse}} \in \mathbb{R}^{2d \times d}$. A decoder with the same block type as the encoder and middle stack refines z^{fuse}, and the output head maps each token to $C_{\text{out}} p^2$ values, reshaped to image space. Transformer blocks follow DiT (Peebles & Xie, 2023), with adaptive layer normalization for timestep injection in the diffusion regime and standard pre-normalization otherwise.

2.4 Training and Inference Regimes

Both regimes share the architecture but differ in inputs, loss, and inference.

Deterministic restoration. Given degraded input x_{deg} and clean target y, the model predicts y with ℓ_1 loss: $\mathcal{L}_{\text{det}} = \mathbb{E}_{(x_{\text{deg}}, y)} \| f_\theta(x_{\text{deg}}) - y \|_1$. For ERA5, x_{deg} is a bilinearly upsampled low-resolution field; denoising uses Rician corruption (MRI) or additive Gaussian noise (natural images), with SNR sampled from $\{10, 15, 20\}$ dB during training and fixed at 15 dB for validation.

Conditional EDM diffusion. For conditional downscaling we use the standard EDM formulation (Karras et al., 2022) with log-normal σ sampling and weighting $\lambda(\sigma)$. The denoiser $D_\theta(y_\sigma, \sigma; x_{\text{LR}}) = f_\theta(\text{concat}(y_\sigma, x_{\text{LR}}); \sigma)$ predicts the clean high-resolution target y conditioned on the upsampled low-resolution input x_{LR}. Validation runs a 20-step Heun ODE sampler and reports RMSE between the sampled prediction and the target.

Training protocol. All ResFuser variants are trained end-to-end at the target drop ratio, with no sparse-pretraining or dense-finetuning curriculum. Backbone dimensions and optimizer settings are shared across cells, while input/output channels, normalization, corruption, conditioning, and training budget follow the task. Deterministic cells use 5,000 optimizer steps; ERA5-diff uses 30,000 steps and 20-step Heun validation. Within each cell, compared variants receive the same step budget. These are fixed-budget comparisons, and equal steps do not establish convergence for every architecture. Hyperparameters are listed in Appendix A.2.

2.5 Ablations and Compute Accounting

Ablation protocol. We compare three design factors under shared backbone and within-cell training budgets: residual presence, residual topology, and token-sampling strategy. The no-residual ablation removes R while keeping the sparse main path, mask restoration, fusion projection, and decoder dimensions unchanged. Topology comparisons fix the sampler and drop ratio while swapping R; sampler comparisons fix the topology and drop ratio while swapping S. Residual variants differ in parameter count and residual computation, so their comparison is matched in token and step budgets, not exactly in total FLOPs. Removing R removes both its information route and parameters; this ablation measures their combined benefit without separating flow from capacity.

Parameter and FLOPs budget. The default backbone has ≈ 7.4M parameters; the residual path adds ≈ 0.3M (parallel CNN) or ≈ 0.07M (sequential skip). We count analytical forward FLOPs per

Table 2: Residual-path ablation across five cells. Bold marks the winning topology per cell. Collapse ratios divide no-residual validation RMSE by the corresponding residual baseline.

		Validation RMSE			No-residual / residual	
Cell	**Drop**	**CNN**	**Skip**	**No res.**	**vs. CNN**	**vs. skip**
ERA5-det	50%	0.0769	**0.0650**	0.6123	8.0×	9.4×
fastMRI-det†	75%	**0.0426**	0.0467	0.2193	5.2×	4.7×
BraTS-det	75%	**0.0277**	0.0313	0.2105	7.6×	6.7×
DIV2K-det	75%	0.0443	**0.0379**	0.1933	4.4×	5.1×
ERA5-diff	50%	0.1936	**0.0710**	0.4315	2.2×	6.1×

† fastMRI-det values are means over seeds 42, 43, and 44. Other cells are single-seed results using seed 42.

sample using the 2× multiply–accumulate convention. Reducing the middle-stack sequence from N to N' lowers its quadratic attention terms and its linear projection/MLP terms; the encoder, restored decoder, and residual still process full-resolution representations. Thus, the drop ratio alone does not specify the end-to-end reduction. Table 3 reports 90–284 GFLOPs for dense baselines and 42–59% reductions for the selected residual variants.

Arithmetic cost and execution time. Latency, throughput, and peak memory also depend on scoring, gather/scatter, restoration, memory traffic, kernel launches, and synchronization. Our reference implementation executes the independent CNN and transformer branches sequentially; we report no measured speedup or memory reduction. Diffusion FLOPs count one forward pass, while complete sampling requires repeated denoiser evaluations under the 20-step Heun protocol. Appendix E describes a deployment assessment including these costs.

3 Experiments

Experiments examine residual presence (Sec. 3.2), topology preference (Secs. 3.4, 3.5), and sampling at fixed topology (Sec. 3.6). We also compare the better ResFuser topology in each cell with a no-drop dense ViT/DiT baseline (Sec. 3.3). Architecture comparisons use the default configuration and training protocol (Sec. 2) with seed 42; fastMRI is repeated at seeds 43 and 44.

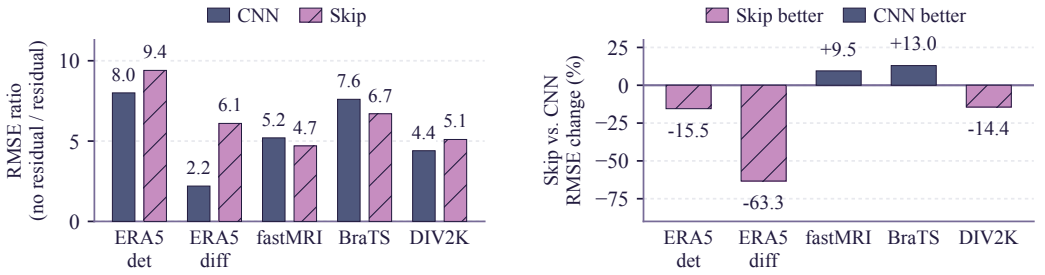
3.1 Experimental setup

The five cells use ERA5 (Hersbach et al., 2020), fastMRI brain DICOM data (Zbontar et al., 2018; Knoll et al., 2020), BraTS-GLI (Baid et al., 2021; Bakas et al., 2017), and DIV2K (Timofte et al., 2018); preprocessing and noise simulation are in Appendix A.3. We report best whole-image validation RMSE over training. Diffusion RMSE uses 20-step Heun sampling (Karras et al., 2022) and is comparable within ERA5-diff. Architecture comparisons fix structured-uniform sampling and the cell's default drop ratio, evaluating no residual, CNN, and skip, alongside a separately trained dense baseline. Sampler ablations fix topology; fastMRI additionally reports brain, high-variance, and high-Laplacian mask RMSE. The mechanism diagnostic adds CNN-wide and kernel-1 MLP variants on ERA5-det and fastMRI-det.

Model scale and uncertainty. All cells use $d = 256$, four heads, and 12 blocks (two encoder, eight middle, two decoder), with $p = 4$ and 2,048–4,096 tokens per image. Volumetric examples motivate the problem but are not experimental cells. fastMRI's CNN outperforms skip in all three seeds (Appendix Figure 5); other cells and sampler ablations lack repeated-seed uncertainty estimates. Four decimal places identify recorded estimates, not statistical precision.

3.2 A dense residual route improves all tested cells

Removing the full-resolution residual path substantially increases validation RMSE in every tested cell (Table 2). Relative to the corresponding residual baseline, the no-residual/residual ratio spans 2.2×–9.4× across both topology columns. Both routes improve on no residual, while their relative



(a) Benefit of each residual route **(b) Topology preference**

Figure 2: **Both residual routes help, but their ranking changes across cells. (a)** Grouped bars show no-residual RMSE divided by each residual baseline; solid blue bars denote CNN and hatched pink bars denote skip. **(b)** Signed bars show the RMSE change of skip relative to CNN; negative values favor skip. fastMRI values use the three-seed means; other cells are single-seed point estimates. Published ratios and percentages are from Tables 2 and 5.

Table 3: Dense ViT/DiT baseline versus the winning ResFuser topology per cell. FLOPs are per-sample forward, $2\times$ MACs convention, computed analytically from the YAML configs. The winning topology is sequential transformer-skip on ERA5-det, ERA5-diff, and DIV2K-det, and parallel CNN on fastMRI-det and BraTS-det. ERA5-diff numbers come from a 30,000-step longrun for both ResFuser and the dense DiT baseline; all other cells use 5,000 steps.

			Forward GFLOPs			Validation RMSE		
Cell	**Drop**	**Route**	**Dense**	**Sparse**	**Change**	**Dense**	**Sparse**	**Change**
ERA5-det	50%	Skip	90.4	52.5	$\mathbf{-41.9\%}$	0.0646	0.0650	$+0.6\%$
ERA5-diff[†]	50%	Skip	90.5	52.6	$\mathbf{-41.8\%}$	0.0666	0.0710	$+6.6\%$
fastMRI-det	75%	CNN	283.5	129.0	$\mathbf{-54.5\%}$	0.0446	**0.0426**	$\mathbf{-4.5\%}$
BraTS-det	75%	CNN	283.7	129.4	$\mathbf{-54.4\%}$	0.0268	0.0277	$+3.4\%$
DIV2K-det	75%	Skip	283.7	117.8	$\mathbf{-58.5\%}$	0.0368	0.0379	$+3.0\%$

[†] Both ERA5-diff entries trained for 30,000 steps with 20-step Heun validation. The 5,000-step Reslim ERA5-diff smoke runs predate validation logging, so no 5K-vs-5K RMSE comparison is available.

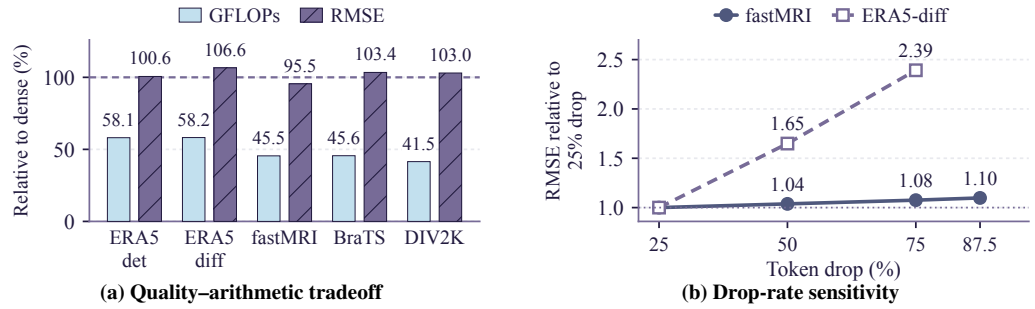
ranking changes across cells (Figure 2). The denominator matters most on ERA5-diff: no residual is $2.2\times$ worse than the parallel CNN and $6.1\times$ worse than the sequential skip. Thus, a dense route consistently helps, and its implementation matters.

Full-resolution residuals supply information for predicting dropped locations before decoding. The ablation quantifies this benefit across two implementations and five cells in the shared small backbone; its magnitude can depend on main-path capacity, restoration mechanism, and training budget.

3.3 Dense-baseline comparison under fixed training budgets

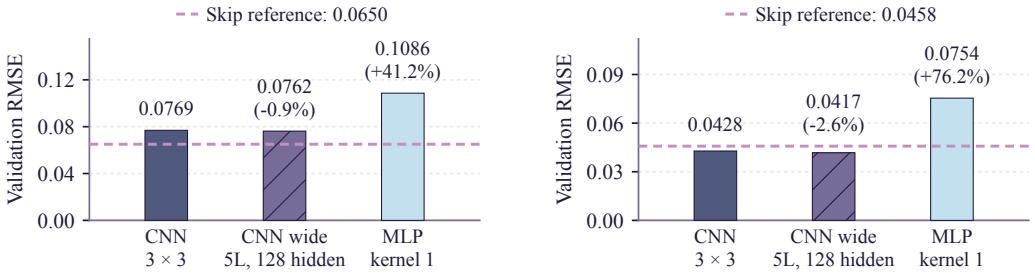
The dense baseline asks how much arithmetic work can be removed from the shared backbone at the observed restoration quality. We train a plain no-drop ViT/DiT in every cell with the same encoder/middle/decoder block budget, optimizer, and step count, then compare it with the better ResFuser topology by validation RMSE. The observed RMSE differences range from -4.5% to $+6.6\%$, with 42–59% fewer analytical forward FLOPs (Figure 3, left). On fastMRI, the selected CNN variant has lower reported RMSE and fewer FLOPs than this dense baseline; the other four cells trade a small RMSE increase for reduced arithmetic cost.

The $\pm 7\%$ range describes point estimates; topology and checkpoint selection both use validation RMSE. The dense reference isolates sparsification; competitive benchmarking and held-out evaluation remain separate tests. Training budgets match per cell: 30,000 steps for ERA5-diff and 5,000 for deterministic cells. No diffusion 5,000-step validation comparison is available.



(a) Quality–arithmetic tradeoff **(b) Drop-rate sensitivity**

Figure 3: **Lower arithmetic cost accompanies cell-dependent quality and drop sensitivity. (a)** Grouped bars show forward GFLOPs (light blue) and RMSE (hatched purple) for the selected sparse topology, each normalized to its corresponding dense baseline (100%, dashed line). Values retain the published percentage changes in Table 3; RMSE is within ±7% at 42–59% fewer FLOPs. **(b)** Parallel-CNN sweeps normalized to each cell's 25%-drop RMSE. Solid circles denote fastMRI; dashed hollow squares denote ERA5-diff. All points are shown, including fastMRI at 87.5% drop. ERA5-diff was not run at 87.5%, and CNN is its weaker topology. Values: Tables 3 and 7.



(a) ERA5-det, 50% drop **(b) fastMRI-det, 75% drop**

Figure 4: **Tested CNN widening does not close the ERA5 gap; removing spatial mixing hurts both cells.** Bars show the 3 × 3 CNN (blue), wider CNN (hatched purple), and kernel-1 MLP (light blue). Pink horizontal dashed lines mark sequential-skip RMSE. Labels give exact recorded RMSE and percentage change from the CNN reference. Panels use different RMSE scales. Numbers are from Table 6; diagnostic fastMRI values are separate from the three-seed architecture means.

3.4 Topology preference is dataset-dependent

At matched sampler and matched training budget, the two residual topologies disagree on which is better, as shown by the residual columns of Table 2: sequential skip wins ERA5-det (−15.5%), ERA5-diff (−63.3%), and DIV2K-det (−14.4%); parallel CNN wins fastMRI-det (+9.5%) and BraTS-det (+13.0%). The split is not explained by a simple domain label, so we treat topology preference as an empirical property of the cell and turn to mechanism in Sec. 3.5. The observed ERA5 gap is larger in the diffusion cell (63.3%) than in the deterministic cell (15.5%). This comparison changes input construction, objective, inference, and training duration together; it does not isolate a causal effect of diffusion. Medical diffusion cells were not evaluated. Topology preference is also largely sampler-independent on ERA5-det: sequential skip beats parallel CNN by 15.5% under structured-uniform sampling and by 12.5% under gradient-based content-aware sampling, so topology and sampler are approximately separable in this cell.

3.5 Mechanism diagnostic for the topology flip

The topology flip in Sec. 3.4 motivated us to ask what governs the preference. We examine a residual-width diagnostic and two proposed cross-domain explanations. The latter make predictions contradicted by the observed BraTS and DIV2K results.

Capacity and inductive-bias diagnostic. We first ask whether a wider and deeper residual CNN closes the ERA5 gap, and whether replacing local spatial kernels with a kernel-1 MLP reproduces the sequential-skip advantage. Figure 4 reports these diagnostic runs; Appendix Table 6 gives their point estimates, separately from the three-seed fastMRI architecture means. On ERA5, the wider CNN changes RMSE from 0.0769 to 0.0762, leaving the gap to the sequential skip (0.0650). On fastMRI, it improves the diagnostic CNN reference by 2.6%. This rules out the tested CNN enlargement as an explanation for the ERA5 ranking; it does not rule out changes in main-backbone capacity or optimization budget. The kernel-1 MLP is worse than both references on ERA5 and incurs a 76.2% increase over the diagnostic CNN reference on fastMRI. Local spatial mixing therefore matters for the tested raw-input branch, but removing it does not turn that branch into an encoder-feature skip.

Two falsified cross-domain predictions. We then proposed a channel-count discriminator (multi-channel inputs favor skip, single-channel favors CNN), which BraTS-GLI falsified: despite having four MRI modalities, BraTS favors the parallel CNN by 13.0%. We refined to a complementary-vs-redundant cross-channel hypothesis (skip when channels carry distinct information, CNN when redundant), which DIV2K falsified: despite redundant RGB channels, DIV2K favors sequential skip by 14.4%. We therefore stop short of claiming a tested mechanism. One descriptive pattern in the current five-cell data is field-of-view density (skip wins on dense-content cells, CNN on compact-object-on-background cells), which we treat as a candidate explanation rather than a conclusion since two earlier mechanism stories also fit the evidence before being falsified.

3.6 Sampling strategy is a small, characterized knob

We next compare token samplers at fixed residual topology, token count, and within-cell training budget. The detailed ablations in ERA5-det, fastMRI-det, and ERA5-diff use the parallel-CNN topology; the figures report the available sampler subsets rather than a complete four-sampler-by-five-cell grid. Their observed effects are smaller than the corresponding residual changes. These runs lack repeated-seed estimates and an ERA5-diff skip sweep, so they characterize the tested rules and configurations, not the full potential of token selection.

ERA5-det. At 50% drop with the parallel-CNN topology fixed, gradient-based sampling gives the largest observed sampler improvement, reducing RMSE by 7.4% relative to structured-uniform sampling; random sampling is 1.3% lower (Appendix Figure 6a). These are single-run differences, with no variance estimate for deciding whether the smaller gap is reproducible. The gradient improvement is smaller than the topology difference on the same cell.

fastMRI-det. On fastMRI-det at 75% drop with the parallel CNN fixed, both content-aware samplers underperform structured-uniform sampling on whole-image, brain, and high-variance masks (Appendix Figure 7). On the high-variance mask, gradient and activation importance give 2.2% and 2.8% higher RMSE, respectively. Gradient sampling is slightly better on the high-Laplacian mask, which targets the same type of signal as its score. This pattern is consistent with a mismatch between boundary salience and interior restoration, but does not by itself establish where the sampler allocates tokens. The high-variance mask is a clean-image intensity statistic, not an annotated lesion mask; the results measure regional denoising, not lesion detection or clinical performance.

ERA5-diff. After 30,000 training steps with 20-step Heun validation, gradient sampling gives 1.8% lower RMSE than structured-uniform, whereas activation sampling gives 5.9% higher RMSE (Appendix Figure 6b). These differences are smaller than the observed residual-topology gap in this cell. Earlier 5,000-step training-loss traces do not supply a matched validation comparison and cannot establish a convergence trend. Across the reported ablations, the useful design lesson is to verify an importance score against the target metric: a score derived from input edges or feature magnitude need not identify tokens that most reduce output error.

3.7 Summary of experimental findings

Across the reported configurations, preserving a residual route gives the largest observed quality benefit, topology preference changes across cells, and the tested samplers provide smaller refinements

(Appendix Table 8). The fastMRI topology ranking repeats across three seeds; the other cells remain point estimates. These findings motivate the staged selection procedure below.

Interpreting drop-rate sensitivity. Both sweeps use the parallel CNN with structured-uniform sampling (Appendix Table 7). fastMRI RMSE rises from 0.0398 at 25% drop to 0.0437 at 87.5%, while ERA5-diff rises from 0.1174 at 25% to 0.2808 at 75% (Figure 3, right). Over the common 25–75% range, fastMRI rises about 8%. The sweeps differ in training budget, and CNN is ERA5-diff's weaker topology. Skip reaches 0.0710 at 50% drop but lacks a sweep, leaving the preferred topology's curve and any intrinsic diffusion drop ceiling unresolved.

3.8 Practical selection under a fixed budget

For a new task, first fix the backbone, training budget, metric, and drop ratio, then compare residual routes with structured-uniform sampling under the intended quality–cost constraint. Second, fix the chosen topology and compare samplers, checking both whole-image and task-relevant region metrics. Third, sweep the drop ratio to characterize the selected configuration's quality–cost tradeoff. This order follows the effect hierarchy and narrows the search over topology, sampler, and drop ratio.

Revisit topology if changing the sampler reverses its ranking. Field-of-view density offers a hypothesis for prioritizing trials, without a validated selection threshold. Use repeated seeds for close choices and held-out data for final assessment; this staged procedure remains a recommendation to validate on new tasks. Appendix E adds runtime assessment.

4 Related Work

Pruning, merging, and sparse training. DynamicViT (Rao et al., 2021) learns importance predictors for progressive pruning; ToMe (Bolya et al., 2022) reduces sequence length by merging similar tokens; and ToMe-SD (Bolya & Hoffman, 2023) adapts merging to diffusion inference. Merging aggregates information from multiple tokens, whereas our samplers retain a subset and restore missing positions before fusion. MaskDiT (Zheng et al., 2023), MicroDiT (Sehwag et al., 2025), MDTv2 (Gao et al., 2023), TREAD (Krause et al., 2025), and SPRINT (Park et al., 2025) connect sparse computation with training or reconstruction mechanisms. Our four sampler rules require separate, matched restoration comparisons with these pruning, merging, and sparse-training methods.

Residual routes and alternative backbones. SPRINT is the source of our encoder-feature skip, while ORBIT-2 (Wang et al., 2025) motivates the independent CNN route. More broadly, long skip connections in U-Net (Ronneberger et al., 2015) and U-ViT (Bao et al., 2023) preserve information across processing stages; their usefulness for dense prediction is established. Swin Transformer (Liu et al., 2021) instead organizes attention through shifted local windows and a hierarchical representation. These are relevant architectural alternatives, but U-ViT's long skips and Swin's windowed hierarchy change different parts of the computation. ResFuser keeps the main backbone fixed to examine where a full-resolution route originates and how it interacts with aggressive dropping.

Position of this study. ResFuser tests residual presence, topology, and sampling in five restoration cells, revealing topology reversals and regional failures of salience-based selection. Its dense baseline is an internal quality–FLOPs reference. Testing pruning, merging, masked training, and other backbones requires matched data and training budgets, plus deployment measurements.

5 Conclusion

ResFuser decomposes residual fusion and token sampling under aggressive token dropping. Across five restoration cells, both routes outperform no residual, with cell-dependent rankings. Selected variants attain validation RMSE within $\pm 7\%$ of dense baselines at 42–59% fewer analytical forward FLOPs. Diagnostics constrain explanations for topology preference, while sampler ablations favor checking input salience against restoration metrics. These findings support evaluating residual topology before sampling and drop ratio. Larger backbones, repeated seeds across all cells, competitive baselines, and deployment measurements will test this ordering beyond this setting.

AI USE STATEMENT

We used a Large Language Model to assist with language polishing and improving the readability of this manuscript. The authors are fully responsible for all research ideas, experimental results, and claims presented in this paper.

ETHICS STATEMENT

We adhere to the ICLR Code of Ethics. This study uses public benchmarks under their terms of use and involves no new human-subject data collection. The medical-image experiments evaluate restoration on existing benchmarks; clinical use would require separate validation.

REPRODUCIBILITY STATEMENT

Section 2 and the appendix specify the architecture, token-sampling rules, training and inference settings, dataset configurations, and evaluation protocol. The appendix reports task-level results, mechanism and sampler ablations, drop-rate sweeps, and the three-seed fastMRI comparison.

REFERENCES

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1): 170117, 2017.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2209.12152.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4599–4603, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80:102479, 2022.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Maria Correia De Verdier, Jeffrey D. Rudie, and Rachit Saluja. The 2024 brain tumor segmentation challenge post-treatment glioma dataset (brats-ptg), 2024. URL https://repo-prod.prod.sagebase.org/repo/v1/doi/locate?id=syn59059776&type=ENTITY.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23164–23173, 2023.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020.

Felix Krause, Timy Phan, Ming Gui, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Tread: Token routing for efficient architecture-agnostic diffusion training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15703–15713, 2025.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL https://arxiv.org/abs/2103.14030.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, et al. Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data*, 13(9):4349–4383, 2021.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, February 2024.

Dogyun Park, Moayed Haji-Ali, Yanyu Li, Willi Menapace, Sergey Tulyakov, Hyunwoo J Kim, Aliaksandr Siarohin, and Anil Kag. Sprint: Sparse-dense residual fusion for efficient diffusion transformers. *arXiv preprint arXiv:2510.21986*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Vikash Sehwag, Xianghao Kong, Jingtao Li, Michael Spranger, and Lingjuan Lyu. Stretching each dollar: Diffusion training from scratch on a micro-budget. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28596–28608, 2025.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

Radu Timofte, Shuhang Gu, Jiqing Wu, Luc Van Gool, Lei Zhang, Ming-Hsuan Yang, Muhammad Haris, et al. Ntire 2018 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

Xiao Wang, Jong-Youl Choi, Takuya Kurihaya, Isaac Lyngaas, Hong-Jun Yoon, Xi Xiao, David Pugmire, Ming Fan, Nasik Muhammad Nafi, Aristeidis Tsaris, et al. Orbit-2: Scaling exascale vision foundation models for weather and climate downscaling. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 86–98, 2025.

Xiao Wang, Zezhong Zhang, Isaac Lyngaas, Hong-Jun Yoon, Jong-Youl Choi, Siming Liang, Janet Wang, Hristo G Chipilski, Ashwin M Aji, Feng Bao, et al. Global attention with linear complexity for exascale generative data assimilation in earth system prediction. *arXiv preprint arXiv:2604.16590*, 2026.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023.

Appendix

The appendix follows the main paper's evidence chain. Appendix A documents the shared implementation and evaluation settings; Appendix B expands the residual-topology comparisons and controls. Appendix C examines token sampling, and Appendix D reports drop-rate sensitivity and qualitative outputs. Appendix E connects the findings to model selection and runtime assessment, followed by the scope and limitations in Appendix F.

A Implementation and Evaluation Settings

A.1 Forward Pass

Following Figure 1, a strided convolution with stride p maps input u to tokens $z \in \mathbb{R}^{N\times d}$, and the full-token encoder produces $z^{\text{enc}} = E(z)$. The sampler selects $\mathcal{K} \subset \{1, \ldots, N\}$ with $|\mathcal{K}| = N'$. Only $z^{\text{enc}}_{\mathcal{K}}$ enters the sparse middle stack, reducing its sequence-length-dependent cost.

The residual module supplies $z^{\text{res}} \in \mathbb{R}^{N\times d}$ from the raw input or encoder output. Learned mask tokens restore dropped positions before channel-wise concatenation with the residual and projection to d channels. The decoder refines the fused tokens; the head predicts $C_{\text{out}}p^2$ values per token and reshapes them into the output image. Only the sampler and residual module vary; embedding, encoder, middle stack, fusion, and decoder are shared.

A.2 Training Defaults

Architecture. The backbone uses embedding dimension $d = 256$, four attention heads, MLP ratio 4, and patch size $p = 4$. Its DiT-style blocks (Peebles & Xie, 2023) are split into $L_E = 2$ encoder, $L_M = 8$ middle, and $L_D = 2$ decoder blocks. The parallel CNN has three convolutional blocks with hidden width $C_{\text{hid}} = 64$; the sequential skip applies a $d \times d$ linear projection to encoder features.

Optimization and seeds. AdamW uses learning rate 5×10^{-4}, weight decay 10^{-5}, and gradient clipping at norm 1.0. Training uses bfloat16 on eight MI250X GCDs, with batch size 4 per rank and effective batch size 32. Deterministic cells receive 5,000 steps; ERA5-diff long runs receive 30,000. Architecture comparisons use seed 42, with fastMRI repeated at seeds 43 and 44. Their aggregate results are distinguished from the separate mechanism, sampler, and drop-sweep point estimates.

A.3 Datasets and Evaluation

Table 4 lists the input, target, corruption, and default drop ratio for each cell. ERA5 evaluates conditional downscaling; fastMRI and BraTS use image-domain MRI restoration, and DIV2K uses RGB denoising. The main experiments report best whole-image validation RMSE over training. Conditional diffusion uses 20-step Heun sampling, so its validation RMSE measures the sampled prediction under that inference protocol. The high-variance and high-Laplacian fastMRI masks are specified in Appendix C.2.

Table 4: **Five restoration cells share the same patch-based backbone.** Input and target configurations, corruption type, and default token drop complement the domains and token counts in Table 1.

Cell	Input	Target	Noise	Drop
ERA5-det	LR↑, 128×256, 7 vars	HR, 128×256, 4 vars	None	50%
ERA5-diff	LR + noisy HR	Clean HR	EDM	50%
fastMRI-det	Magnitude, 256^2, 1 ch	Clean magnitude	Rician	75%
BraTS-det	MRI, 256^2, 4 ch	Clean 4-modality MRI	Rician	75%
DIV2K-det	RGB crop, 256^2, 3 ch	Clean RGB crop	Gaussian	75%

LR↑ denotes bilinearly upsampled low-resolution input. ERA5-diff uses LR-conditioned noisy HR and EDM log-normal σ. Rician and Gaussian noise use training SNR $\in \{10, 15, 20\}$ dB; validation SNR is 15 dB. Domains and regimes are listed in Table 1.

B Residual Topology and Controlled Comparisons

B.1 Cell-Dependent Topology Preference

The no-residual comparison in Table 2 establishes the benefit of retaining a dense route. Table 5 then isolates the choice between the two implemented routes: skip has lower RMSE in both ERA5 cells and DIV2K, while CNN has lower RMSE in fastMRI and BraTS. The signed changes use CNN as the denominator. Token counts and training budgets are fixed within each cell; the routes have different parameter counts and arithmetic costs.

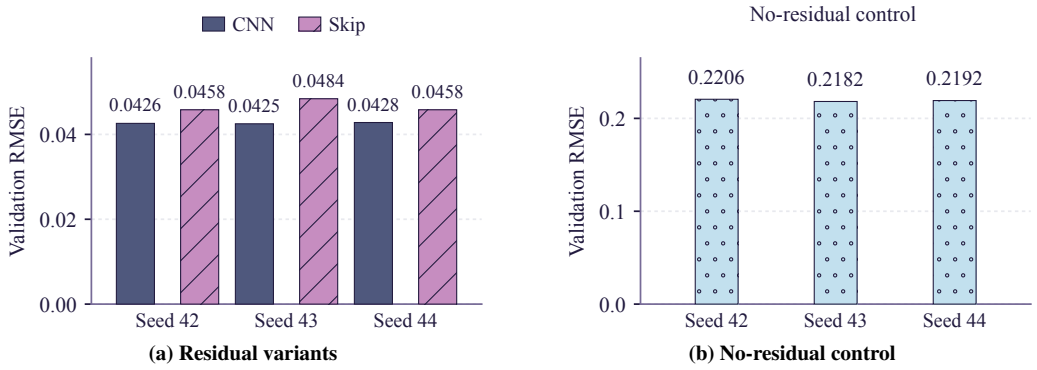
Table 5: **Residual-topology preference changes across cells.** Δ is the reported RMSE change of skip relative to CNN; negative values favor skip. Exact values support Figure 2.

Cell	CNN RMSE	Skip RMSE	Δ skip vs. CNN	Winner
ERA5-det	0.0769	0.0650	-15.5%	Skip
ERA5-diff	0.1936	0.0710	-63.3%	Skip
fastMRI-det†	0.0426	0.0467	$+9.5\%$	CNN
BraTS-det	0.0277	0.0313	$+13.0\%$	CNN
DIV2K-det	0.0443	0.0379	-14.4%	Skip

† fastMRI-det values are means over three seeds; the direction is unchanged at every individual seed.

B.2 Consistency Across fastMRI Seeds

Figure 5 shows every recorded architecture run on fastMRI at 75% drop. CNN outperforms skip at each seed, and both routes remain well below the no-residual control. This repeatability check supports the fastMRI ranking in Table 5; it does not supply uncertainty estimates for the other cells or the separate sampler diagnostics.



(a) Residual variants **(b) No-residual control**

Figure 5: **The fastMRI topology ranking repeats at all three seeds.** Each bar is one validation RMSE at 75% drop; lower is better. CNN is blue and skip is hatched pink. Panels use different vertical scales to show the within-residual comparison and the larger no-residual error. Reported mean $\pm$ standard deviation across seeds 42, 43, and 44 is 0.0426 ± 0.0002 for CNN, 0.0467 ± 0.0015 for skip, and 0.2193 ± 0.0012 for no residual.

B.3 Capacity and Spatial-Mixing Controls

Table 6 gives the exact values behind Figure 4. Widening the CNN leaves the ERA5 skip advantage intact and modestly improves fastMRI. Replacing local convolutions with a kernel-1 MLP raises error in both cells, supporting the role of spatial mixing in the raw-input branch. These controls test the residual implementation; they leave main-backbone scaling and the proposed field-of-view explanation unresolved. The fastMRI diagnostic references are separate runs from the architecture averages above.

Table 6: **CNN widening and spatial-mixing controls constrain the topology explanation.** The wider CNN has five layers and 128 hidden channels; the kernel-1 MLP removes spatial mixing. Δ values compare RMSE with each reference. Figure 4 visualizes these diagnostic runs.

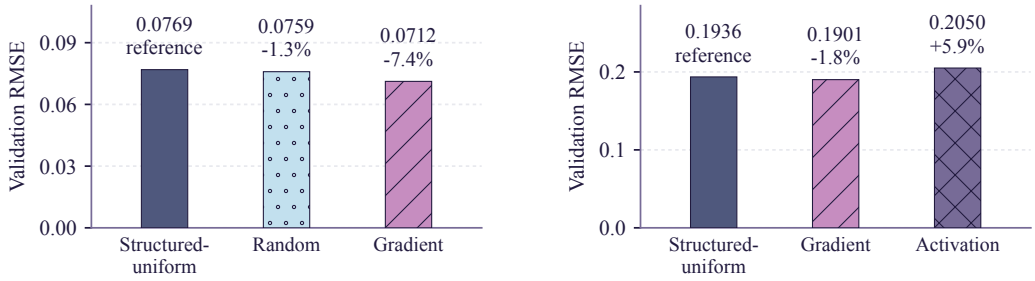
Variant	Val RMSE	Δ vs. CNN	Δ vs. skip
ERA5-det			
CNN, 3×3	0.0769	–	$+18.3\%$
CNN wide, 5 layers, 128 hidden	0.0762	-0.9%	$+17.2\%$
MLP, kernel 1, no spatial mixing	0.1086	$+41.2\%$	$+67.1\%$
Sequential skip	**0.0650**	-15.5%	–
fastMRI-det			
CNN, 3×3	**0.0428**	–	-6.6%
CNN wide, 5 layers, 128 hidden	**0.0417**	-2.6%	-9.0%
MLP, kernel 1, no spatial mixing	0.0754	$+76.2\%$	$+64.6\%$
Sequential skip	0.0458	$+7.0\%$	–

C Token Sampling at Fixed Residual Topology

These comparisons hold the parallel-CNN topology, token count, and training budget fixed within each cell. Figures 6 and 7 show the available sampler subsets, with all recorded values and reported percentage changes. They are single-run comparisons.

C.1 ERA5 Sampler Comparisons

Gradient sampling improves the ERA5-det point estimate more than random sampling does. On ERA5-diff, its improvement is smaller, while activation sampling raises error (Figure 6). The diffusion panel uses 30,000 training steps and 20-step Heun validation. Earlier 5,000-step training-loss traces lack comparable validation RMSE and therefore do not establish a convergence trend. These results support checking an importance score against the target restoration metric.



(a) ERA5-det, 5,000 steps **(b) ERA5-diff, 30,000 steps**

Figure 6: **Gradient sampling improves both ERA5 point estimates, with different effect sizes.** All runs use parallel CNN at 50% drop. Labels give validation RMSE and its reported percentage change from structured-uniform; negative change is better. Samplers are identified by both labels and fill patterns. Panels have different RMSE scales and training budgets; only the recorded sampler subsets are shown.

C.2 Regional Errors on fastMRI

The regional analysis uses clean-image masks: brain covers intensity > 0.05; high variance selects the top 10% of local standard deviation within the brain; high Laplacian selects the top 10% of Laplacian magnitude. High variance is an image-statistics proxy, not a lesion annotation. Whole-image and regional RMSE in Figure 7 show that structured-uniform sampling performs best on the whole image, brain, and high-variance masks. Gradient sampling improves the high-Laplacian mask, which targets

the same signal as its score. This pattern connects the usefulness of salience to the evaluation region without identifying the sampler's token allocation mechanism.

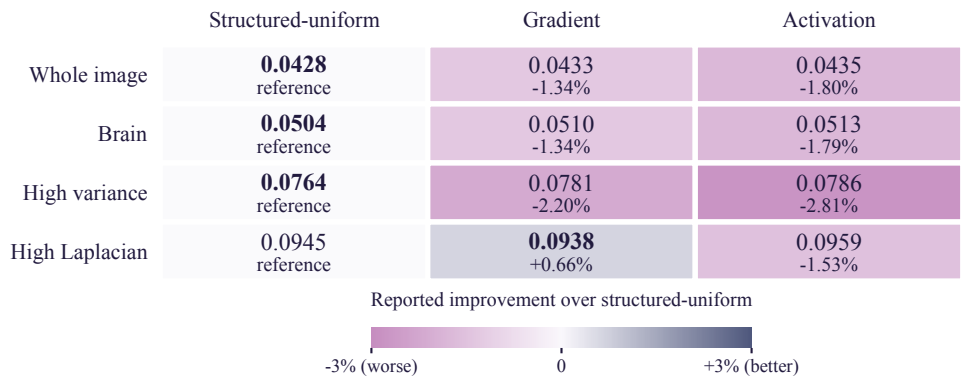


Figure 7: **Input salience does not consistently improve regional restoration.** Each cell shows fastMRI validation RMSE at 75% drop with parallel CNN; bold marks the lowest RMSE in each row. Percentages and shading show the reported improvement over structured-uniform: positive is better, negative is worse. Exact labels preserve the comparison without color. Percentages are the recorded improvements; RMSE is displayed to four decimals.

D Drop Sensitivity and Qualitative Restoration

D.1 Drop-Rate Sweeps

Table 7 supplies the absolute RMSE behind the normalized curves in Figure 3b. The fastMRI sweep extends to 87.5% drop, whereas ERA5-diff stops at 75%. Both use parallel CNN, which is the weaker ERA5-diff topology. Their different training budgets and incomplete topology coverage make these configuration-specific sensitivity curves, rather than an estimate of an intrinsic diffusion drop ceiling.

Table 7: **Drop sensitivity depends on the cell and tested topology.** Both sweeps use parallel-CNN residuals and structured-uniform sampling. fastMRI-det uses 5,000 training steps; ERA5-diff uses 30,000 steps and 20-step Heun validation. Figure 3b shows normalized curves.

Drop ratio	fastMRI-det RMSE	ERA5-diff RMSE
25%	0.0398	0.1174
50%	0.0413	0.1936
75%	0.0428	0.2808
87.5%	0.0437	Not run

D.2 Qualitative Comparison

Figure 8 shows the degraded input, both residual predictions, and target for each empirical cell. The shared row and column layout supports direct inspection of the restored structure. These representative examples complement the aggregate RMSE comparisons.

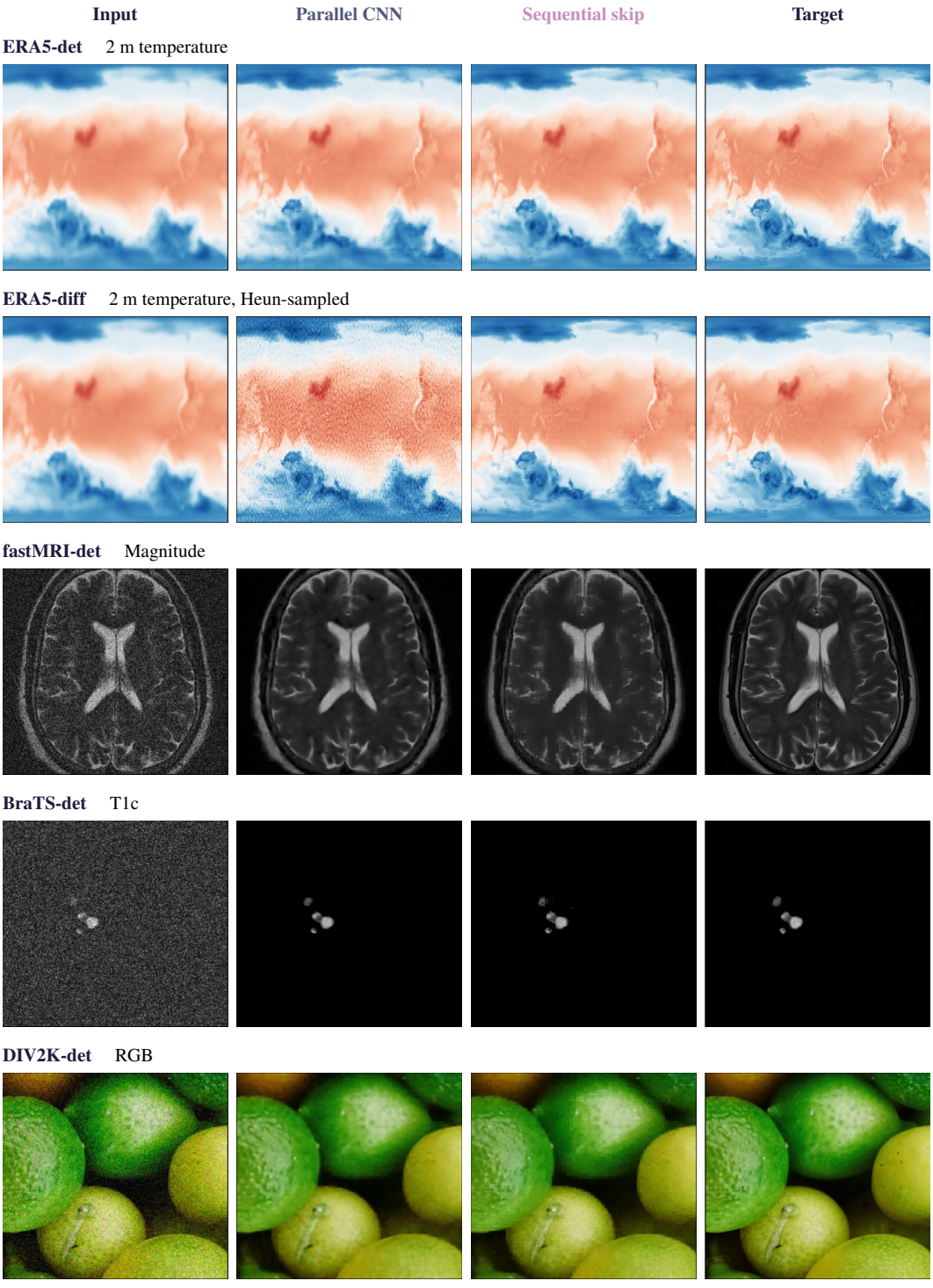


Figure 8: **Qualitative restoration examples** across the five empirical cells. Columns show the degraded input, parallel-CNN prediction, sequential-skip prediction, and target.

E SELECTION AND DEPLOYMENT ASSESSMENT

Table 8: **Evidence for the staged design procedure.** Each row links a finding to its measured comparison or diagnostic; magnitudes retain the reported settings and denominators.

Finding	Evidence	Observed magnitude
Residual helps across cells	Sec. 3.2	2.2–9.4× no-residual/residual RMSE
Dense-reference tradeoff	Sec. 3.3	±7% RMSE; 42–59% fewer FLOPs
Topology ranking changes	Sec. 3.4	9.5–63% gap in either direction
Mechanism unresolved	Sec. 3.5	Two proposed predictions contradicted
Sampler effects are smaller	Sec. 3.6	Best observed improvement: 7.4%
CNN sweep sensitivity varies	Table 7	+10% vs. +139%; different endpoints
fastMRI ranking repeats	Figure 5	9.5% mean gap; zero flips in three seeds

Runtime assessment is prospective: the present study reports analytical FLOPs.

Select and validate. Fix the backbone, training budget, metric, and initial drop ratio, and compare residual routes with structured-uniform sampling. Then compare samplers and sweep the drop ratio, revisiting topology if its ranking changes. Select on validation data and assess on held-out data, with normalization, corruption, and inference fixed within comparisons. Repeat close cases across seeds and define evaluation regions independently of sampling scores.

Measure the complete execution. Match device, precision, batch size, input shape, and inference protocol. Time scoring, token selection, gather/scatter, restoration, fusion, and residual computation together. After warm-up, synchronize device work and measure repeated latency, throughput, and peak allocated memory. For diffusion, report denoiser and complete 20-step Heun costs under the same solver settings.

Extend the comparison. Scale backbone width or depth while holding the task and residual comparison fixed, and assess convergence using validation curves over longer budgets. Evaluate pruning, merging, and alternative backbones on matched restoration data and training budgets, reporting both quality and execution cost.

F LIMITATIONS

Model scale and optimization. Evidence covers a $d = 256$, 12-block, $p = 4$ backbone trained for 5,000 deterministic or 30,000 diffusion steps; convergence and main-backbone scaling remain untested. Residual removal changes information flow and capacity together, while CNN widening tests one branch enlargement. Parameter-matched controls could alter the interpretation.

Statistical and selection scope. Only fastMRI architecture comparisons include three seeds; other cells and samplers remain point estimates. Validation-based topology and checkpoint selection require held-out assessment before the staged rule is applied predictively.

Mechanisms and coverage. Channel count and redundancy fail to predict all rankings; field-of-view density remains a hypothesis. ERA5 comparisons change the objective, inputs, inference, and training budget together. Its drop sweep uses the weaker CNN, and sampler coverage is partial. These tests leave causal topology mechanisms and diffusion-specific drop limits unresolved.

External baselines and deployment. The dense ViT/DiT is an internal reference; comparisons to ToMe, DynamicViT, MaskDiT, Swin, and U-ViT remain open. The skip adapts SPRINT through a projection. Runtime gains require measurement beyond analytical FLOPs. MRI results concern image-domain denoising and statistical masks; acquisition-level reconstruction and clinical utility require separate evaluation.