GROWING IN THE RIGHT COORDINATES: SCHUR-GATED FUNCTION-PRESERVING NETWORK EXPANSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Expanding a trained network without changing its current predictions is useful when additional capacity is needed but restarting training is undesirable. In this setting, however, adding new parameters is not enough: the added directions should complement the inherited model, and they should be parameterized in coordinates that the optimizer can use effectively. Existing growth and conditioning methods typically address these two issues separately. We introduce SKETCH-SCHUR-OST, a function-preserving network expansion method that treats representation and optimization geometry jointly. The key idea is to use a single conditional Schur metric to determine which candidate directions add information beyond the inherited tangent space, how they should be transported relative to the existing parameters, and how the resulting coordinates should be conditioned through a full SPD architectural gate. This produces a unified construction in which selection, redundancy removal, and conditioning are different parts of the same local geometric factorization. We show that the construction preserves the inherited function at insertion, separates support creation from conditioning, and yields an exact factorization of the corresponding damped calibration metric. We further develop a shared-sketch formulation and a row-space dual that make the method practical without materializing the full joint Jacobian–candidate design, while making explicit the conditions under which the sketch guarantees apply. Across controlled growth experiments, the unified construction consistently improves finite-budget optimization relative to matched ungated and sequential alternatives, with gains that persist across coordinate budgets, longer training horizons, adaptive optimization, heterogeneous candidate banks, and larger inherited models. The results also clarify the limits of the approach: aggressive undersketching can fall outside the certified regime, and stronger architectures may remain preferable when retraining from scratch is allowed. Overall, the paper argues that function-preserving growth is best viewed not only as a problem of adding capacity, but as a problem of adding useful directions in the right coordinates.

1 INTRODUCTION

When a trained network must grow without changing its current predictions, restarting with a stronger architecture is not an available intervention. The expansion must create useful functional directions, avoid duplicating the inherited tangent, and expose coordinates that the optimizer can use. These requirements are often handled by separate heuristics.

Spectral bias illustrates why this separation matters. A target direction can be absent from the local functional tangent space, making every first-order update orthogonal to it, or present with weak or correlated curvature, making learning slow (Rahaman et al., 2019; Basri et al., 2020). Architecture determines support; coordinates determine scaling. An SPD optimizer can reshape existing tangent directions but cannot create a missing one.

Recent scientific-machine-learning evidence reinforces that representation and conditioning cannot be assessed independently. Spectral learning in PINNs, PIKANs, and neural operators depends jointly on architecture, the physics-informed objective, and optimizer dynamics, while curvature-aware

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.

solvers can materially change high-frequency recovery (Khodakarami et al., 2026; Jnini et al., 2026); multiscale PINN training also exhibits parameter-wise scale imbalance (Yao et al., 2023). In gray-box inverse problems, architecture, precision, warm-up, and optimizer likewise interact in accuracy and stability (Ahmadi Daryakenari et al., 2026). These results motivate treating growth and optimization geometry together; our setting is the distinct intervention in which capacity is added without changing the inherited function.

Global parameter-space condition numbers are a poor diagnostic: they contain redundant directions and change under harmless reparameterizations (Dinh et al., 2017). We instead project the functional Jacobian onto task modes, producing a falsifiable chain

$$\text{parameterization symmetry} \rightarrow \text{tangent support} \rightarrow \text{modal curvature} \rightarrow \text{learning time}. \tag{1}$$

The operative question is therefore which proposed directions reduce the loss *after conditioning on the inherited tangent*, and which coordinates let gradient descent exploit them. A plausible sequential pipeline is GradMax-style selection, followed by projection against the old tangent and then conventional whitening. Its stages optimize different quantities: a large raw gradient may be removed by the later projection, while a whitening chosen after selection cannot revise that choice. It also supplies no single augmented-metric factorization for the deployed coordinates.

SKETCH-SCHUR-OST uses the same Schur complement for all three decisions. It first scores candidate directions after accounting for motion already available to the inherited network. The corresponding ridge solution determines the compensating transport of the base parameters, and the inverse square root of the remaining conditional block defines the SPD gate. In this way, selection, transport, and conditioning are derived from the same local geometry rather than designed as separate stages. To avoid materializing the full $N \times (P + q)$ Jacobian–candidate design, we also derive a shared-sketch implementation and a row-space dual. Our contributions are:

1. a function-preserving construction in which conditional selection, ridge transport, and full-matrix gating are the score, triangular factor, and diagonal factor of one damped augmented tangent metric;
2. conditional shared-sketch guarantees for both deployed geometry and gradient-dependent greedy selection, plus a row-space dual that avoids a full $P \times q$ transport; and
3. a controlled empirical evaluation that isolates the contribution of the full SPD gate and the unified construction, and tests robustness across coordinate budgets, optimization horizons, adaptive optimization, heterogeneous candidate banks, and larger inherited models.

2 RELATED WORK

Fourier features and periodic activations modify spectral support or kernel bandwidth (Tancik et al., 2020; Sitzmann et al., 2020); activation choice can also reverse-engineer kernels (Simon et al., 2022). Function-preserving growth spans GradMax, Net2Net, network morphism, and recent masked, Transformer-scale, and lifelong variants (Evci et al., 2022; Chen et al., 2016; Wei et al., 2016; Yao et al., 2024; Du et al., 2024; Han et al., 2025). Adjacent work studies adaptive tangent features, kernel spectra, projected augmentation, residual-depth sufficiency, and zero-gated activation (LeJeune & Alemohammad, 2024; Geifman et al., 2024; Györök et al., 2025; Liu et al., 2026; Lok, 2026). Our rule instead conditions the gradient on the inherited tangent and derives transport and the full SPD gate from that block; sequential GradMax–projection–whitening scores before those changes.

Whitening and reparameterization are established tools (Desjardins et al., 2015; Salimans & Kingma, 2016; Amid & Warmuth, 2020; Kristiadi et al., 2023), as are natural-gradient, full-matrix, and neural-field preconditioners (Martens & Grosse, 2015; Dangel et al., 2020; Gupta et al., 2018; Chng et al., 2025). The contribution is not whitening itself or the standard chain-rule equivalence. SCHUR-OST acts only on the small new-coordinate Schur block: its forward gate is tied to the selected directions, but an invertible map cannot add support.

Randomized sketching compresses least-squares and approximate-Hessian systems (Woodruff, 2014; Lacotte & Pilanci, 2020). We use one sketch for the joint old-tangent/candidate design so that selection, transport, and the forward gate remain factors of one approximated metric. Exact insertion and a separate out-of-sketch audit distinguish this setting from solving a fixed linear problem with a sketched Hessian.

3 TANGENT SUPPORT AND ARCHITECTURAL COORDINATES

For samples $x_1, \ldots, x_N$, let $f_\theta = (f_\theta(x_i))_{i=1}^N$, $J_\theta \in \mathbb{R}^{N\times P}$ be its parameter Jacobian, and $K_\theta = J_\theta J_\theta^\top$ the empirical NTK (Jacot et al., 2018). For normalized task mode q_k, define

$$\lambda_k(\theta) = q_k^\top K_\theta q_k = \|J_\theta^\top q_k\|_2^2. \tag{2}$$

A mode is *tangent-inaccessible* when $\lambda_k = 0$. For target modes $\mathcal{S}$, we report the support rank $r_\mathcal{S} = |\{k \in \mathcal{S} : \lambda_k > 0\}|$ separately from the supported-subspace condition $\kappa_\mathcal{S}^+ = \max_{k\in\mathcal{S}} \lambda_k / \min_{k\in\mathcal{S}:\lambda_k>0} \lambda_k$. The overall target condition is defined as $\kappa_\mathcal{S} = \infty$ whenever $r_\mathcal{S} < |\mathcal{S}|$, and equals $\kappa_\mathcal{S}^+$ otherwise. This prevents a missing mode from disappearing inside a finite condition number. We also report cross-mode coherence $|q_i^\top K q_j|/\sqrt{\lambda_i\lambda_j}$ when relevant.

Theorem 1 (Preconditioning cannot create tangent support)**.** *Every first-order functional update satisfies $\delta f = J\delta\theta \in \mathrm{Range}(J)$. For any parameter-space matrix M, $\mathrm{Range}(JMJ^\top) \subseteq \mathrm{Range}(J)$. If $P \succ 0$, then*

$$\mathrm{Null}(JPJ^\top) = \mathrm{Null}(JJ^\top) = \mathrm{Null}(J^\top),$$

and consequently $JPJ^\top$ and $JJ^\top$ have the same range. A differentiable coordinate map satisfies $\mathrm{Range}(J_\phi) \subseteq \mathrm{Range}(J_\theta)$; local invertibility makes the ranges equal.

Indeed, $v^\top JPJ^\top v = \|P^{1/2}J^\top v\|_2^2$, which vanishes exactly when $J^\top v = 0$. The theorem separates two interventions: non-invertible architectural augmentation can add support, whereas an SPD metric can only reshape supported directions. Five random rank-deficient controls preserve rank and range projectors to 5.4×10^{-13}.

Theorem 2 (Parameterization-level nullspace)**.** *Under the uniform periodic inner product, or a paired empirical analogue, if $f_\theta(x+\pi) = -f_\theta(x)$ on an open parameter set, then every parameter derivative is half-wave antisymmetric and $J_\theta^\top q_{2m} = \lambda_{2m} = 0$ for every even harmonic.*

Differentiating the symmetry identity proves the claim and yields $\inf_\theta \|f_\theta - y\|_N^2 \geq \|y_e\|_N^2$; for $y = \sin x + 0.5\cos 2x$, the floor is 0.125. The open-set condition matters: an odd output at one initialization does not imply an odd parameterization.

3.1 BLIND SCHUR-GATED SIDEBAND GROWTH

The construction below accepts any differentiable candidate dictionary H. For nonsmooth multiscale growth, H can concatenate localized edge, compact hat/wavelet, and windowed oscillatory atoms at several spatial scales. For the image experiments, we instead use learned-feature sidebands as follows. For a learned feature $h_j(x)$ and an unlabeled continuous carrier $\omega \in \mathbb{R}^d$, our candidate bank exposes

$$h_j(x)\cos(\omega^\top x), \qquad h_j(x)\sin(\omega^\top x). \tag{3}$$

Multiplication shifts the spectrum of h_j by $\pm\omega$, so these additions can create frequencies not present in the original feature. A constant channel adds ordinary Fourier features to the same bank as distractors and as a matched baseline. Selection uses neither frequency labels nor a target FFT.

Let $H \in \mathbb{R}^{N\times q}$ contain all candidates and write $C = J^\top J$, $D = J^\top H$, and $E = H^\top H$. For $\lambda > 0$, define

$$B_\lambda = (C + \lambda I)^{-1} D, \qquad R_\lambda = H - JB_\lambda, \tag{4}$$

$$K_\lambda = R_\lambda^\top R_\lambda + \lambda B_\lambda^\top B_\lambda = E - D^\top B_\lambda. \tag{5}$$

Thus B_λ is the ridge map predicting the candidate block from the old tangent, while K_λ measures both unexplained functional motion and the compensating parameter displacement needed to produce it.

The same metric supplies a blind block score. Let g be the output-space loss gradient, $c = R_\lambda^\top g$, and A a candidate subset. The squared coordinate-gradient norm after exact Schur whitening is, up to a common factor,

$$Q(A) = c_A^\top (K_\lambda)_{AA}^{-1} c_A. \tag{6}$$

Starting from $A = \varnothing$, we add the candidate maximizing $Q(A\cup\{i\}) - Q(A)$. This is an exact conditional Schur-complement gain: a large candidate that duplicates existing or already-selected

motion receives little marginal value. It replaces frequency-oracle selection and the earlier hand-chosen capacity penalty.

For the selected block, set $G_\lambda = (K_\lambda/N)^{-1/2}$ and introduce the nonlinear coordinates

$$F(\psi, \phi) = f_{\psi - B_\lambda G_\lambda \phi} + H_{\psi - B_\lambda G_\lambda \phi} G_\lambda \phi, \qquad \phi_0 = 0. \tag{7}$$

The sideband branch co-adapts with the effective base parameters. Derivatives of H are multiplied by $\phi_0 = 0$, so the new birth tangent is exactly $R_\lambda G_\lambda$.

Theorem 3 (Exact damped Schur gate)**.** *Let*

$$M_\lambda = \frac{1}{N}\begin{bmatrix} C + \lambda I & D \\ D^\top & E \end{bmatrix}, \quad T_\lambda = \begin{bmatrix} I & -B_\lambda \\ 0 & I \end{bmatrix}, \quad W_\lambda = \mathrm{diag}(I, G_\lambda).$$

If $K_\lambda \succ 0$, then Eq. (7) *preserves the insertion function and*

$$W_\lambda^\top T_\lambda^\top M_\lambda T_\lambda W_\lambda = \mathrm{diag}((C + \lambda I)/N, I)\,. \tag{8}$$

Moreover, $\|B_\lambda G_\lambda\|_2 \le \sqrt{N/\lambda}$.

The off-diagonal block vanishes because $(C + \lambda I)B_\lambda = D$; the lower block is precisely K_λ/N. The bound follows from $(\lambda/N)B_\lambda^\top B_\lambda \preceq K_\lambda/N$. This factorization also characterizes the coupling. For arbitrary transport $\widetilde{B}$ and invertible gate $\widetilde{G}$, the transformed old–new block is $(D - (C + \lambda I)\widetilde{B})\widetilde{G}/N$; it vanishes exactly when $\widetilde{B} = B_\lambda$. The unique symmetric SPD gate that maps the remaining block K_λ/N to identity is G_λ. Moreover, $Q(A)$ is the squared gradient norm in these deployed coordinates. Changing these components independently generally loses at least one of these properties: conditional selection, damped decoupling, or identity scaling. A relative eigenvalue floor is used in float32.

Notice that this does not exactly diagonalize the undamped functional tangent Gram used by joint SGD. The same normal equations give

$$J^\top R_\lambda = D - CB_\lambda = \lambda B_\lambda, \qquad \frac{1}{N} J^\top R_\lambda G_\lambda = \frac{\lambda}{N} B_\lambda G_\lambda, \tag{9}$$

which is generally nonzero. We therefore report the damped cross residual and the actual normalized undamped cross block separately, along with the full undamped joint-tangent condition. Accordingly, the gate conditions the selected Schur block locally; it does not imply global decoupling once the full nonlinear model continues to adapt.

Holding the old coordinate fixed at birth makes the optimizer role explicit:

$$\nabla_\phi L = G_\lambda^\top R_\lambda^\top g, \qquad \Delta f_{\mathrm{new}} = -\eta R_\lambda G_\lambda G_\lambda^\top R_\lambda^\top g. \tag{10}$$

Sidebands create candidate support, transport discounts motion already available to the old model, and the architectural gate preconditions the selected new block.

3.2 One-sketch selection, transport, and gating

Exact calibration stores $[J, H] \in \mathbb{R}^{N \times (P+q)}$. Sketch-Schur-OST instead draws a single scaled row-sampling operator $S \in \mathbb{R}^{s \times N}$ and forms $\widehat{J} = SJ$, $\widehat{H} = SH$, and $\widehat{g} = Sg$. It runs the same conditional score, ridge transport, and Schur gate on these sketched quantities. After selecting $k \ll q$ columns, only those k sidebands are evaluated on all samples inside the nonlinear branch. Thus calibration storage falls from $O(N(P + q))$ to $O(s(P + q))$; unlike independently sketching the blocks, the shared S preserves their cross-Gram structure.

For $P \gg s$, we also avoid constructing the full sketched primal transport $B \in \mathbb{R}^{P \times q}$. Let $A = \widehat{J}\widehat{J}^\top + \lambda I_s$. Woodbury identities give the exact same selection statistics

$$\widehat{K}_\lambda = \lambda \widehat{H}^\top A^{-1} \widehat{H}, \qquad \widehat{c} = \lambda \widehat{H}^\top A^{-1} \widehat{g}. \tag{11}$$

We greedily select using only these $q \times q$ and q statistics, then form $B_A = \widehat{J}^\top A^{-1} \widehat{H}_A$ and $R_A = \lambda A^{-1} \widehat{H}_A$ only for the final k columns. The resulting indices, transport, and gate are algebraically identical to the primal method; storage becomes $O(sP + sq + q^2 + Pk)$ rather than additionally materializing Pq transport entries.

The following deterministic statement isolates what a valid sketch preserves. It conditions on the positive ridge λ realized at calibration; the implementation chooses it relative to the sketched top singular value.

Theorem 4 (Sketched Schur gate)**.** *Let H denote the full candidate bank and suppose S is an ϵ-subspace embedding for $[J, H]$, $0 \le \epsilon < 1$:*

$$(1-\epsilon)\|[J,H]z\|_2^2 \le \|S[J,H]z\|_2^2 \le (1+\epsilon)\|[J,H]z\|_2^2 \quad \textit{for all } z.$$

For any subset A selected from the sketch, let $M_{\lambda,A}$ be the full damped metric of Theorem 3, and construct $\widehat{T}_A, \widehat{W}_A$ from (SJ, SH_A) without eigenvalue clipping. Writing

$$\widehat{\mathcal{D}}_A = \widehat{W}_A^\top \widehat{T}_A^\top \widehat{M}_{\lambda,A} \widehat{T}_A \widehat{W}_A = \mathrm{diag}((\widehat{C} + \lambda I)/N, I),$$

the deployed full-data coordinates obey

$$\frac{1}{1+\epsilon}\widehat{\mathcal{D}}_A \preceq \widehat{W}_A^\top \widehat{T}_A^\top M_{\lambda,A} \widehat{T}_A \widehat{W}_A \preceq \frac{1}{1-\epsilon}\widehat{\mathcal{D}}_A. \tag{12}$$

In particular, the deployed selected-coordinate block has condition at most $(1+\epsilon)/(1-\epsilon)$. Its zero-readout insertion preserves the full nonlinear function exactly, independently of ϵ.

Because the embedding holds for the whole bank, it holds simultaneously for the data-dependent selected subset. The theorem guarantees geometry, not that the sketch chooses the same columns as the exact selector. The following result covers that distinct, gradient-dependent question. For a subset A, define the ridge-residual objective

$$F_A = \min_{u,v} \ \|g - Ju - H_A v\|_2^2 + \lambda\|u\|_2^2, \qquad F_\varnothing = \min_u \ \|g - Ju\|_2^2 + \lambda\|u\|_2^2,$$

and define $\widehat{F}_A$ by replacing the data residual with $\|S(g - Ju - H_A v)\|_2^2$. Block elimination gives $Q(A) = F_\varnothing - F_A$, exactly the score in Eq. (6).

Theorem 5 (Gradient-aware selection fidelity)**.** *Suppose S is an ϵ-subspace embedding for $[J, H, g]$, with $0 \le \epsilon < 1$. Then, simultaneously for every subset A,*

$$|\widehat{Q}(A) - Q(A)| \le \epsilon(F_\varnothing + F_A) \le 2\epsilon F_\varnothing. \tag{13}$$

For a greedy marginal $\Delta_i(A) = Q(A \cup \{i\}) - Q(A) = F_A - F_{A\cup\{i\}}$,

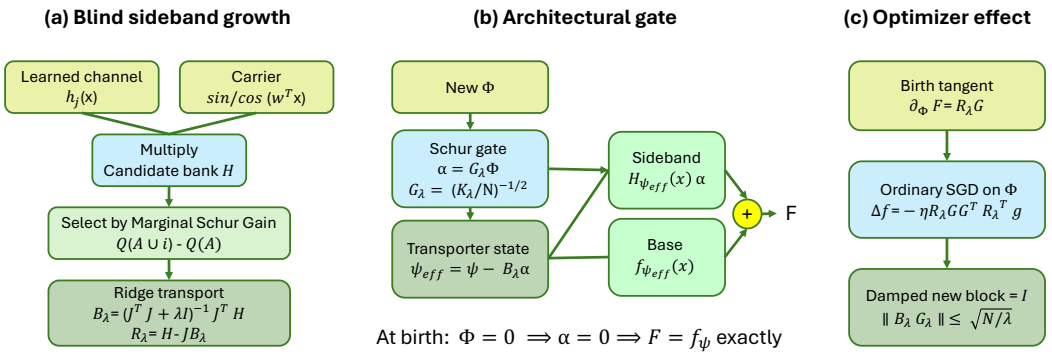
$$|\widehat{\Delta}_i(A) - \Delta_i(A)| \le 2\epsilon F_A. \tag{14}$$

Consequently, the sketched greedy choice is within $4\epsilon F_A$ of the best exact marginal, and it is identical whenever the best–second-best exact margin exceeds $4\epsilon F_A$.

The additive form is necessary: no sketch can guarantee small *relative* error for a selection gain that is arbitrarily close to zero. If $Q(A) \ge \rho F_\varnothing$, Eq. (13) additionally gives relative error at most $2\epsilon/\rho$. This theorem strengthens the required embedding by one gradient column; it complements rather than replaces the geometry guarantee of Theorem 4. We still use a second, independent row sketch to audit the fixed transport and gate. A relative eigenvalue floor handles float32 numerical rank; the geometry bound applies to the retained block.

Scope of the sketch theorems. Both results require an embedding of the *full-bank* range, not only the selected block; necessarily $s \ge \mathrm{rank}([J, H])$ (plus g for Theorem 5). The deployed $s = 256$ setting falls outside this sufficient condition: Appendix D reports a median full-bank rank of 604. We therefore treat that regime empirically and separately study how the selected set approaches exact selection as the sketch size increases.

Auxiliary coordinate controls. For a fixed linear head $\theta = G\phi$, gradient descent in ϕ is exactly $GG^\top$-preconditioned descent in θ, even when the upstream representation is trained jointly. Appendix B.1 collects the corresponding equivalence checks, including moving-coordinate transport, Adam non-equivariance, and refresh-time scale controls. It also verifies the complete transported implementation in a conventional nonlinear classifier, where architectural gated SGD and the corresponding explicit Schur optimizer follow the same trajectory to numerical precision. A separate five-seed classifier experiment examines finite-budget loss descent.



Sidebands propose support; Schur gain selects; ridge transport and the gate factor one damped metric.

Figure 1: Schur-gated growth. (a) Multiplicative carriers create candidate sidebands and ridge transport forms R_λ. (b) The Schur SPD gate $\alpha = G\phi$ drives both the compensating base shift and the co-adapting sideband branch; $\phi_0 = 0$ preserves the function. (c) Ordinary SGD through the gate induces $GG^\top$ preconditioning on the new functional directions. SKETCH-SCHUR-OST computes all three factors on one shared row sketch.

4 EXPERIMENTS

All main comparisons are paired and run in float32. Development and evaluation cases are kept separate, and the experimental settings are fixed before the evaluation runs. We foreground effect sizes and image-cluster intervals; the complete protocols are given in the appendix.

The experiments examine three main questions: whether the unified construction improves over sequential and ungated alternatives, whether the benefit of the full SPD gate persists as the optimization setting changes, and whether the same mechanism extends beyond the primary image-growth setting. We also study how sketch size affects selection fidelity and calibration cost, and report additional boundary controls in the appendix.

4.1 PRIMARY CAUSAL TEST ON LOCKED NATURAL-IMAGE GROWTH

We evaluate on six images from the BSDS500 test split (Arbelaez et al., 2011). A width-24 Tanh coordinate network first fits a 64×64 downsampled target for 400 Adam steps, then grows toward the 256×256 image for 300 full-batch SGD steps; an interleaved pixel quarter is held out. The old model has 697 parameters. Every growth arm has 745: 48 new scalar coordinates, or 12 GradMax-style Tanh neurons with four trainable scalars each. The shared sideband bank has the same 640 candidates as the exact study below.

SKETCH-SCHUR-OST uses one 3,072-row shared sketch for full-bank conditional selection, transport, and gate construction, plus a disjoint 1,536-row audit. A separate 1,024-row screen supplies the residual and random controls. Two excluded development images and seeds 20–22 select one shared rate 0.03 from $\{0.03, 0.1, 0.3\}$; it has the lowest mean AUC for every retained method. The untouched cohort comprises six other images and seeds 600–602, or 18 paired image–seed cases. The primary control uses the *identical* Schur-selected sidebands and ridge transport but replaces the 48×48 SPD gate by the identity. Additional controls use residual-selected gated sidebands, residual-selected pure Fourier features, random gated sidebands, parameter-matched Tanh growth, and continued training without growth.

In the primary ablation, the two models differ only in the gate. The full SPD parameterization lowers both training-loss AUC and held-out MSE, while the selected candidates, added capacity, ridge transport, learning rate, and insertion function remain unchanged. Table 1 also reports the scale-matched and architectural controls.

We separately froze the Introduction's direct sequential comparator: greedy raw-gradient selection, the same ridge transport, then whitening of $R_\lambda^\top R_\lambda / N$ without the $\lambda B_\lambda^\top B_\lambda / N$ term. The two excluded

Table 1: Locked BSDS500 evaluation at 256^2. The first row is the primary gate-only causal test; the scalar and diagonal rows isolate off-diagonal conditioning from global scale. The daggered sequential row uses a separately locked fresh six-image cohort. Positive values favor SKETCH-SCHUR-OST; intervals cluster-bootstrap images while retaining all three seeds per image. The Wins column counts paired cases positive on both reported endpoints.

Comparator	Wins	ΔAUC	95% CI	Δtest	95% CI
Same selected block, no gate	18/18	6.68%	[3.65, 10.13]%	5.08%	[2.66, 7.70]%
Sequential composition†	18/18	3.45%	[2.45, 4.50]%	3.54%	[2.59, 4.46]%
Scalar gate, max-curvature matched	18/18	7.52%	[4.01, 11.52]%	5.83%	[2.90, 9.05]%
Diagonal gate, max-curvature matched	18/18	7.15%	[3.83, 10.78]%	5.38%	[2.71, 8.14]%
Residual sideband + gate	18/18	8.68%	[6.45, 11.20]%	8.96%	[6.67, 11.52]%
GradMax-style Tanh growth	18/18	14.35%	[9.63, 19.23]%	14.68%	[9.67, 19.74]%

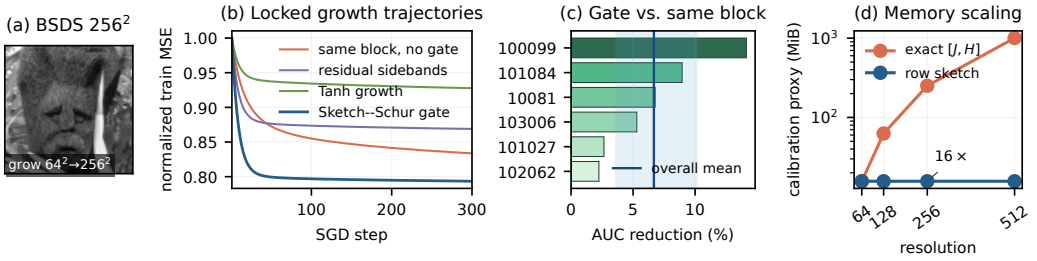


Figure 2: Locked scalable evaluation. (a) Representative BSDS image and resolution-growth task. (b) Solid lines are mean normalized trajectories over 18 paired runs; uncertainty bands are omitted for visual clarity. (c) Green bars are per-image mean gate-only AUC reductions; the blue line and band are the overall mean and image-cluster 95% interval. (d) Exact and sketched calibration-storage proxies; at 256^2 the reduction is $16\times$.

development images select rate 0.03 for this arm; the proposed rate was already 0.03. Evaluation uses six previously unused images and seeds 700–702, yielding 18 new paired cases.

Against the fresh sequential composition, SKETCH-SCHUR-OST reduces AUC by 3.45% (image-cluster 95% interval $[2.45, 4.50]\%$) and held-out MSE by 3.54% ($[2.59, 4.46]\%$). This comparison shows that using a common metric for selection, transport, and conditioning performs better than applying these steps sequentially. The identical-block ablation separately isolates the contribution of the gate. Relative to residual selection, AUC falls 8.68%; random, pure-Fourier, and no-growth controls are weaker by 17.57%, 16.08%, and 19.70%, respectively.

The same block and transport also receive scalar and diagonal gates normalized to the full gate's maximum transformed curvature at the shared rate 0.03. Full SPD improves AUC by $7.52\%/7.15\%$ and held-out MSE by $5.83\%/5.38\%$ versus scalar/diagonal; median condition is $276.7/236.4/1.00003$ for scalar/diagonal/full. Trace-matched stability controls are in the appendix.

The exact $[J, H]$ proxy is 250.69 MiB versus 15.67 MiB sketched ($16\times$); every insertion jump is zero. A disjoint audit gives median selected-block condition 2.23. Median damped/actual cross ratios are $1.32 \times 10^{-6}/0.0418$ on calibration and $0.1049/0.1846$ on the disjoint sketch. Thus the gate is calibrated locally, not globally decoupled. Figure 2 shows that the gain is not driven by one image. The unoptimized sideband implementation takes 61.7 seconds per growth run versus 12.2 for 12-neuron Tanh growth. The current implementation is therefore not faster in wall-clock time than the simpler Tanh-growth baseline; its advantage here is in optimization behavior and calibration memory.

4.2 COORDINATE-BUDGET SENSITIVITY

The primary comparison used the development-selected value $k = 48$. We therefore froze a separate sensitivity protocol before observing its full matrix and evaluated $k \in \{16, 32, 48, 64, 96\}$ on three previously unused BSDS500 images with seeds 730–732. The 640-candidate bank, sketches, ridge, eigenvalue floor, inherited model, data split, learning rate, and 300-step budget are unchanged. At

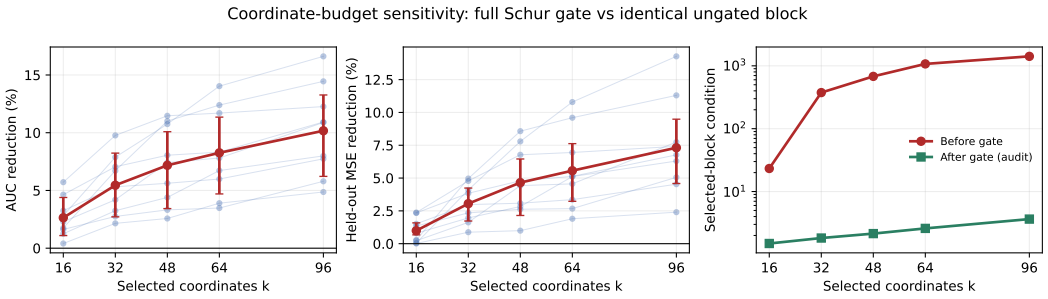


Figure 3: Fresh coordinate-budget sensitivity on three images and three seeds per image. Thin lines show paired cases; thick points and bands show means and descriptive image-cluster 95% intervals. Full SPD improves (a) normalized mean training-loss AUC and (b) mean held-out MSE at every tested budget. (c) The selected block becomes increasingly ill-conditioned with k, while the independent-audit condition remains small after gating.

every k, the only difference is again the full SPD gate versus the identity on the *identical* selected and ridge-transported block.

The gate improves both mean AUC and mean held-out MSE at every tested budget. Mean AUC reductions are 2.64%, 5.45%, 7.18%, 8.26%, and 10.17% as k increases; held-out MSE reductions are 0.99%, 3.05%, 4.65%, 5.56%, and 7.31%. The fresh $k = 48$ AUC result is close to the 6.68% locked headline effect despite using disjoint images and seeds. All 45 gate/no-gate pairs are positive for both endpoints (9/9 at each budget). Meanwhile, median pre-gate condition increases from 23.3 to 1,415, whereas the independent-audit post-gate condition remains between 1.49 and 3.64. Every insertion jump is zero, every run is stable, and mean calibration time stays within 5.54–5.66 seconds.

Figure 3 shows the paired image–seed effects and descriptive image-cluster intervals. The same pattern appears across the full range of tested coordinate budgets, rather than only at the original choice $k = 48$. The association between $\log_{10}$ pre-gate condition and AUC benefit is moderate ($r = 0.647$, Spearman $\rho = 0.600$). Because this correlation was examined after the first two extreme-budget runs, we treat it as exploratory rather than confirmatory. Full protocol and numerical summaries are in Appendix D.2.

4.3 Optimizer and horizon robustness

The primary comparison uses 300 post-growth SGD steps. To test whether its advantage is only an early transient, we froze a rerun on the same six images and three seeds before execution, retained the identical selected block, transport, insertion function, and rate 0.03, and extended both arms to 3,000 steps. Metrics were declared at 300, 1,000, and 3,000 steps. The 300-step checkpoint reproduces the headline result within 0.25 percentage points. At 1,000 steps, full SPD reduces AUC by 4.63% (image-cluster 95% interval $[2.45, 7.01]\%$) and held-out MSE by 2.88% ($[1.48, 4.23]\%$). At 3,000 steps the reductions remain 2.93% ($[1.51, 4.33]\%$) and 1.48% ($[0.62, 2.42]\%$), respectively. The advantage becomes smaller with longer training, but it is still present at 3,000 steps. We do not study asymptotic convergence here.

We separately froze an Adam transfer study. Two excluded development images select one common rate, 0.003, by the mean AUC of the gate and identity arms; the locked six-image cohort is then evaluated once for 300 steps. Full SPD reduces AUC by 1.74% ($[0.92, 2.56]\%$) and held-out MSE by 0.99% ($[0.31, 1.74]\%$). Every insertion jump is zero. The smaller Adam effect is consistent with partial redundancy from adaptive diagonal scaling, while its positive residual supports empirical optimizer transfer. Standard Adam is not rotation-equivariant, so we interpret this experiment as an empirical transfer test rather than an optimizer-equivalence result. Protocols and complete paired summaries are in Appendix D.1.

As a conventional-architecture diagnostic, Appendix B.2 reuses a locked five-seed MNIST MLP growth cohort in which only the gate changes. Full SPD lowers normalized training-loss AUC by 12.91% relative to the identity gate in all five pairs. It reaches normalized loss 0.40 in all five runs

Table 2: Frozen robustness against the identical selected and transported block without the gate. Brackets are image-cluster 95% intervals.

Optimizer	Steps	ΔAUC [CI]	Δtest [CI]
SGD	300	6.59 [3.65,10.04]%	5.01 [2.61,7.65]%
SGD	1,000	4.63 [2.45,7.01]%	2.88 [1.48,4.23]%
SGD	3,000	2.93 [1.51,4.33]%	1.48 [0.62,2.42]%
Adam	300	1.74 [0.92,2.56]%	0.99 [0.31,1.74]%

(median epoch 6), whereas the identity-gated block reaches it in none by epoch 10; matched scalar and diagonal gates reach it at median epochs 7 and 8. The gain is confined to optimization: validation and test cross-entropy do not improve in this cohort.

Heterogeneous coordinates and scale. A frozen five-signal stress test replaces image sidebands by multiscale edges, wavelets, and localized oscillations; full SPD reduces mean AUC/MSE by $64.61/39.56\%$ versus the identical ungated block. It wins training-loss AUC in 25/25 cases and held-out MSE in 24/25; the one negative Bumps endpoint is retained. Selected-block condition reaches 1.01×10^5 before gating, while every post-gate condition is at most 1.0048 (Appendix C.2). The row-space dual reduces mean AUC/MSE by $8.90/7.65\%$ on six 51,297-parameter networks and reaches 200,251 parameters in a calibration-only curve. A seven-size audit converges to exact selection, but the deployed 256-row sketch does not certify the full-bank embedding premise (Appendix D).

5 Discussion and Limitations

The main idea behind Sketch-Schur-OST is to use the same conditional geometry for choosing new directions and for parameterizing them. This distinguishes the method from sequential growth–projection–whitening, where the stages are designed independently. The sequential baseline, identical-block ablation, and scalar/diagonal controls separate the effects of the unified construction, the gate, and its full-matrix structure.

The results show a finite-horizon optimization benefit. The advantage becomes smaller with longer SGD training and is also reduced under Adam, suggesting that the gate mainly improves local optimization geometry rather than changing the final function class. Transport remains part of the same Schur factorization, although it is not uniformly better than functional-only whitening in every setting.

Several limitations remain. The main experiments use relatively small image cohorts and coordinate/function MLPs, so broader validation on conventional deep architectures is still needed. In the classifier study, the gate improves training-loss descent but does not improve held-out cross-entropy. The most aggressive deployed sketch also falls outside the sufficient full-bank embedding condition, and the current sideband implementation is slower than simple Tanh growth. In addition, stronger spectral INR models remain preferable when training from scratch is allowed. Extending the method to repeated growth, larger-scale matrix-free calibration, and safe metric refresh is an important direction for future work.

6 Conclusion

Sketch-Schur-OST makes conditional selection, ridge transport, and full-SPD gating factors of one damped metric while preserving the inherited function. Locked tests isolate the gate and unified construction, and show positive effects across coordinate budgets, a tenfold longer horizon, Adam, and heterogeneous coordinates. A conventional-classifier control further matches the architectural gate to an explicit Schur optimizer at numerical precision, and a separate locked cohort shows faster finite-budget training-loss descent while retaining its adverse held-out boundary. The result is a scoped growth–preconditioning method with explicit sketch limits, not a universal architecture or optimizer claim.

Reproducibility Statement

Source code, unit tests, and per-seed results will be made available to reviewers upon request and released publicly upon publication. Experiments use paired seeds and preserve inherited artifacts separately from locally reproduced results. Appendix D gives data, optimizer, precision, and evaluation details; Appendix E maps the main claims to their evidence.

AI Use Statement

Generative AI tools were used to assist with figure refinement and presentation, code cleanup, and language and grammar editing. All scientific content, analyses, results, and conclusions were developed and verified by the authors.

References

Nazanin Ahmadi Daryakenari, Khemraj Shukla, and George Em Karniadakis. Representation meets optimization: Training PINNs and PIKANs for gray-box discovery in systems pharmacology. *Computers in Biology and Medicine*, 201:111393, 2026. doi: 10.1016/j.compbiomed.2025.111393.

Ehsan Amid and Manfred K. Warmuth. Reparameterizing mirror descent as gradient descent. In *NeurIPS*, 2020.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011. doi: 10.1109/TPAMI.2010.161.

Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In *ICML*, 2020.

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2net: Accelerating learning via knowledge transfer. In *ICLR*, 2016.

Shin-Fang Chng, Hemanth Saratchandran, and Simon Lucey. Preconditioners for the stochastic training of neural fields. In *CVPR*, 2025.

Felix Dangel, Stefan Harmeling, and Philipp Hennig. Modular block-diagonal curvature approximations for feedforward architectures. In *AISTATS*, 2020.

Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, and Koray Kavukcuoglu. Natural neural networks. *arXiv preprint arXiv:1507.00210*, 2015.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *ICML*, 2017.

Wenyu Du, Tongxu Luo, Zihan Qiu, Zeyu Huang, Yikang Shen, Reynold Cheng, Yike Guo, and Jie Fu. Stacking your transformers: A closer look at model growth for efficient LLM pre-training. In *NeurIPS*, 2024.

Utku Evci, Bart van Merriënboer, Thomas Unterthiner, Max Vladymyrov, and Fabian Pedregosa. Gradmax: Growing neural networks using gradient information. In *ICLR*, 2022.

Amnon Geifman, Daniel Barzilai, Ronen Basri, and Meirav Galun. Controlling the inductive bias of wide neural networks by modifying the kernel's spectrum. *Transactions on Machine Learning Research*, 2024.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *ICML*, 2018.

Bendeguz Mate Györök, Jan H. Hoekstra, Johan Kon, Tamas Peni, Maarten Schoukens, and Roland Toth. Orthogonal projection-based regularization for efficient model augmentation. In *L4DC*, 2025.

Xue Han, Yitong Wang, Junlan Feng, Wenchun Gao, Qian Hu, and Chao Deng. LOIRE: Lifelong learning on incremental data via pre-trained language model growth efficiently. In *ICLR*, 2025.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

Anas Jnini, Elham Kiyani, Khemraj Shukla, Jorge F. Urbán, Nazanin Ahmadi Daryakenari, Johannes Müller, Marius Zeinhofer, and George Em Karniadakis. Curvature-aware optimization for high-accuracy physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 462:119289, 2026. doi: 10.1016/j.cma.2026.119289.

Siavash Khodakarami, Vivek Oommen, Nazanin Ahmadi Daryakenari, Maxim Beekenkamp, and George Em Karniadakis. Spectral bias in physics-informed and operator learning: Analysis and mitigation guidelines. *Computer Methods in Applied Mechanics and Engineering*, 461:119156, 2026. doi: 10.1016/j.cma.2026.119156.

Agustinus Kristiadi, Felix Dangel, and Philipp Hennig. The geometry of neural nets' parameter spaces under reparametrization. In *NeurIPS*, 2023.

Jonathan Lacotte and Mert Pilanci. Optimal randomized first-order methods for least-squares problems. In *ICML*, 2020.

Daniel LeJeune and Sina Alemohammad. An adaptive tangent feature perspective of neural networks. In *CoLLAs*, 2024.

Zeyu Liu, Jinhao Zhang, Yunquan Zhang, Guangming Tan, Xiang Gao, Fangming Liu, and Daning Cheng. Quantifying depth sufficiency in residual neural networks: A first-order criterion. *arXiv preprint arXiv:2608.14664*, 2026.

Dante Lok. Gate-zero growth: A geometric framework for function-preserving continual learning. *arXiv preprint arXiv:2607.14571*, 2026.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *ICML*, 2015.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *ICML*, 2019.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *NeurIPS*, 2016.

James Simon, Sajant Anand, and Michael DeWeese. Reverse engineering the neural tangent kernel. In *ICML*, 2022.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020.

Tao Wei, Changhu Wang, Yong Rui, and Chang Wen Chen. Network morphism. In *ICML*, 2016.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014. doi: 10.1561/0400000060.

Jiachen Yao, Chang Su, Zhongkai Hao, Songming Liu, Hang Su, and Jun Zhu. Multiadam: Parameter-wise scale-invariant optimizer for multiscale training of physics-informed neural networks. In *ICML*, 2023.

Yiqun Yao, Zheng Zhang, Jing Li, and Yequan Wang. Masked structural growth for 2x faster language model pre-training. In *ICLR*, 2024.

A Proofs and Algorithm Details

A.1 Proof of Theorem 1

The statements $J\delta\theta \in \mathrm{Range}(J)$ and $\mathrm{Range}(JMJ^\top) \subseteq \mathrm{Range}(J)$ follow directly from matrix multiplication and require no assumption on the optimizer. For any $v \in \mathbb{R}^N$,

$$v^\top JPJ^\top v = (J^\top v)^\top P(J^\top v) = \|P^{1/2}J^\top v\|_2^2.$$

Because $P \succ 0$, this quantity is zero if and only if $J^\top v = 0$. The same argument with $P = I$ gives $\mathrm{Null}(JPJ^\top) = \mathrm{Null}(JJ^\top) = \mathrm{Null}(J^\top)$. Both kernels are symmetric, so equality of their nullspaces implies equality of their ranges by orthogonal complementation. For any differentiable map $\theta = q(\phi)$, the chain rule gives $J_\phi = J_\theta Dq$, hence its range is a subset of the range of J_θ. Local invertibility makes $DqDq^\top \succ 0$, and the SPD statement yields range equality. The result is pointwise and does not require linearity of the network or loss.

A.2 Proof of Theorem 2

Differentiate $f_\theta(x+\pi) + f_\theta(x) = 0$ with respect to every component θ_j. This gives

$$\partial_{\theta_j} f_\theta(x+\pi) = -\partial_{\theta_j} f_\theta(x).$$

Every Jacobian column is therefore in the negative eigenspace of the half-period shift. Under the uniform periodic measure, or a paired empirical measure with equal orbit weights, this shift is unitary. Its positive and negative eigenspaces are therefore orthogonal. Even Fourier modes are in the positive eigenspace, so $J_\theta^\top q_{2m} = 0$ and $q_{2m}^\top J_\theta J_\theta^\top q_{2m} = 0$. Without a shift-invariant measure or paired grid, pointwise antisymmetry alone does not imply empirical orthogonality; the theorem makes no such claim.

For the residual-selected adapter, differentiation gives $\partial r_\phi/\partial\alpha_k = \sin(kx)$ and $\partial r_\phi/\partial\beta_k = \cos(kx)$. At least one derivative is the selected normalized mode up to a nonzero scale, so its projected empirical-NTK curvature is strictly positive on the periodic grid.

A.3 Proof and selection rule for Theorem 3

The ridge normal equations are $(C + \lambda I)B_\lambda = D$. Direct block multiplication therefore gives

$$T_\lambda^\top \begin{bmatrix} C+\lambda I & D \\ D^\top & E \end{bmatrix} T_\lambda = \begin{bmatrix} C+\lambda I & 0 \\ 0 & E - D^\top B_\lambda \end{bmatrix}.$$

To identify the lower block, substitute $R_\lambda = H - JB_\lambda$:

$$\begin{aligned} R_\lambda^\top R_\lambda + \lambda B_\lambda^\top B_\lambda &= E - D^\top B_\lambda - B_\lambda^\top D + B_\lambda^\top (C+\lambda I)B_\lambda \\ &= E - D^\top B_\lambda = K_\lambda. \end{aligned}$$

The final equality again uses the normal equations, which also show that the Schur block is symmetric. Multiplication by $W_\lambda = \mathrm{diag}(I, (K_\lambda/N)^{-1/2})$ proves Eq. (8). At $\phi = 0$, Eq. (7) equals f_ψ. Differentiating with respect to ϕ gives $-JB_\lambda G_\lambda + HG_\lambda = R_\lambda G_\lambda$ because every derivative of H_θ is multiplied by the zero coefficient. Finally, $(\lambda/N)B_\lambda^\top B_\lambda \preceq K_\lambda/N$ implies

$$(B_\lambda G_\lambda)^\top (B_\lambda G_\lambda) \preceq (N/\lambda) I,$$

which proves the spectral-norm bound.

For completeness, the greedy score in Eq. (6) can be updated without inverting every candidate subset. Let A be selected, k_{iA} the candidate-to-block row of K_λ, and $c = R_\lambda^\top g$. The exact marginal gain is

$$\frac{(c_i - k_{iA}K_{AA}^{-1}c_A)^2}{k_{ii} - k_{iA}K_{AA}^{-1}k_{Ai}}.$$

Our implementation solves the selected systems rather than forming explicit inverses, masks nonpositive numerical conditional variances, and adds one candidate at a time. This formula is also the test used for the distractor controls.

A.4 Row-space dual identities

Let $A = JJ^\top + \lambda I$. The push-through identity $(J^\top J + \lambda I)^{-1} J^\top = J^\top A^{-1}$ gives

$$B_\lambda = J^\top A^{-1} H, \qquad R_\lambda = H - JJ^\top A^{-1} H = \lambda A^{-1} H.$$

Substitution into Eq. (5) yields

$$K_\lambda = E - D^\top B_\lambda = H^\top \left(I - JJ^\top A^{-1}\right) H = \lambda H^\top A^{-1} H,$$

and $c = R_\lambda^\top g = \lambda H^\top A^{-1} g$. Thus Eq. (11) produces exactly the primal K_λ and c. Because greedy selection depends only on these quantities, its indices are identical; forming $J^\top A^{-1} H_A$ after selection recovers exactly the selected primal transport. Unit tests compare indices, transport, birth tangent, and gate to float64 tolerance.

A.5 Proof of Theorem 4

Let $A_A = [J, H_A]$ for the subset returned by the sketched selector. Because A_A is obtained by deleting columns of $[J, H]$, the assumed embedding for the full bank also holds for A_A. In Loewner form,

$$(1-\epsilon) A_A^\top A_A \preceq A_A^\top S^\top S A_A \preceq (1+\epsilon) A_A^\top A_A.$$

Let $\Lambda = \mathrm{diag}(\lambda I_P, 0) \succeq 0$. Adding the same deterministic ridge to the two data Grams preserves the multiplicative bound:

$$(1-\epsilon)(A_A^\top A_A + \Lambda) \preceq A_A^\top S^\top S A_A + \Lambda \preceq (1+\epsilon)(A_A^\top A_A + \Lambda).$$

For example, the left inequality follows because $\Lambda \succeq (1-\epsilon)\Lambda$; the right follows analogously. After division by N, these are $(1-\epsilon) M_{\lambda,A} \preceq \widehat{M}_{\lambda,A} \preceq (1+\epsilon) M_{\lambda,A}$.

Congruence by $L_A = \widehat{T}_A \widehat{W}_A$ preserves Loewner order. The exact Schur factorization applied to the sketched blocks gives $L_A^\top \widehat{M}_{\lambda,A} L_A = \widehat{\mathcal{D}}_A$; rearranging the two congruence inequalities yields Eq. (12). Restricting the quadratic form to vectors $(0, z)$ makes the corresponding block of $\widehat{\mathcal{D}}_A$ equal to I. Its deployed Rayleigh quotient is therefore between $1/(1+\epsilon)$ and $1/(1-\epsilon)$, proving the condition bound. Finally, the nonlinear insertion has $\phi = 0$, so both the sideband readout and compensating shift vanish exactly. This argument does not use the embedding assumption.

The theorem is deterministic conditional on a valid embedding. Standard randomized transforms can supply high-probability embeddings (Woodruff, 2014); the present implementation uses uniform rows without replacement, scaled by $\sqrt{N/s}$, because pixels are sampled on a dense regular grid. We do not assert a coherence-free probability guarantee for this sampler. Instead, every locked run evaluates the already fixed gate on a disjoint row sketch, which cannot influence selection or training.

A.6 Proof of Theorem 5

For any subset A, write $Z_A = [J, H_A]$, $w = (u, v)$, and $\Lambda_A = \mathrm{diag}(\lambda I_P, 0)$. Expanding the regularized least-squares objective and eliminating u first gives the old-only optimum minus

$$\left(H_A^\top g - D_A^\top (C + \lambda I)^{-1} J^\top g\right)^\top (K_\lambda)_{AA}^{-1} \left(H_A^\top g - D_A^\top (C + \lambda I)^{-1} J^\top g\right).$$

The vector in parentheses is $H_A^\top g - B_{\lambda,A}^\top J^\top g = R_{\lambda,A}^\top g = c_A$. Therefore $F_\varnothing - F_A = Q(A)$. The identical calculation on (SJ, SH, Sg) gives $\widehat{F}_\varnothing - \widehat{F}_A = \widehat{Q}(A)$.

Because every residual $g - Z_A w$ lies in the column span of $[J, H, g]$, the embedding assumption gives

$$(1-\epsilon)\|g - Z_A w\|_2^2 \le \|S(g - Z_A w)\|_2^2 \le (1+\epsilon)\|g - Z_A w\|_2^2.$$

Adding $w^\top \Lambda_A w$ preserves the same multiplicative inequalities, exactly as in the proof of Theorem 4. Taking the minimum over w yields $(1-\epsilon) F_A \le \widehat{F}_A \le (1+\epsilon) F_A$ simultaneously for every subset. Hence

$$|\widehat{Q}(A) - Q(A)| \le |\widehat{F}_\varnothing - F_\varnothing| + |\widehat{F}_A - F_A| \le \epsilon (F_\varnothing + F_A).$$

Candidate addition cannot increase the optimum, so $F_A \le F_\varnothing$, which proves Eq. (13). Applying the same argument to $F_A - F_{A \cup \{i\}}$ proves Eq. (14). Each sketched marginal is within $2\epsilon F_A$ of its exact counterpart; comparing the sketched maximizer to the exact maximizer gives the stated $4\epsilon F_A$ approximate-greedy and exact-margin conclusions.

A.7 ARCHITECTURAL-PRECONDITIONING EQUIVALENCE

Proposition 1 (Architectural preconditioning equivalence)**.** *Let $L(\psi, \theta)$ be differentiable and parameterize a linear head as $\theta = G\phi$ for fixed invertible G. Simultaneous gradient-descent steps on (ψ, ϕ) map exactly to the same update on ψ and*

$$\theta^+ = \theta - \eta G G^\top \nabla_\theta L(\psi, \theta).$$

Thus the coordinate map induces the SPD optimizer preconditioner $GG^\top$ on the head, even when upstream parameters learn. For full-column-rank least squares, $G = (X^\top X)^{-1/2}$ makes the reparameterized Hessian the identity.

By the chain rule, $\nabla_\phi L(\psi, G\phi) = G^\top \nabla_\theta L(\psi, \theta)$, while $\nabla_\psi L(\psi, G\phi) = \nabla_\psi L(\psi, \theta)$. Therefore

$$G\phi^+ = G\phi - \eta G G^\top \nabla_\theta L(\psi, \theta),$$

and the upstream update is identical, proving the claimed joint trajectory equivalence by induction. In the least-squares special case, the Hessian with respect to ϕ is $G^\top X^\top X G$. Full column rank makes $X^\top X$ SPD, so substituting its symmetric inverse square root gives the identity.

The same identity extends to the transported growth coordinates used by SCHUR-OST. Let the raw grown variables be $(\theta_{\text{eff}}, \alpha)$ and write

$$\theta_{\text{eff}} = u - BG\phi, \qquad \alpha = G\phi, \qquad P = GG^\top. \tag{15}$$

If $g_\theta = \nabla_{\theta_{\text{eff}}} L$ and $g_\alpha = \nabla_\alpha L$, the chain rule gives

$$u^+ = u - \eta g_\theta, \qquad \alpha^+ = \alpha - \eta P(g_\alpha - B^\top g_\theta). \tag{16}$$

Thus ordinary SGD through the architectural gate is exactly identity-gated transported growth with an explicit Schur-preconditioned coordinate update. Equivalently, in fully raw variables the induced SPD block preconditioner is

$$\begin{bmatrix} I + BPB^\top & -BP \\ -PB^\top & P \end{bmatrix} = \begin{bmatrix} I & -BG \\ 0 & G \end{bmatrix} \begin{bmatrix} I & 0 \\ -G^\top B^\top & G^\top \end{bmatrix}.$$

The factorization proves positive definiteness. Momentum remains equivalent when its coordinate state is transformed; for weight decay, the raw-coordinate update uses $Pg_\alpha + \gamma\alpha$, rather than applying P to the decay term. This distinction is enforced in the optimizer-side control below.

A.8 EXACT-UNMIXING AUXILIARY CONTROL

For $Y = [\sin x, \cos 8x]$, define $Z = YA_\rho$ with $A_\rho = \left[\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right]$.

Proposition 2 (Exact SPD unmixing)**.** *The unique map satisfying $ZG = Y$ is $G = A_\rho^{-1}$, which is SPD for $0 \le \rho < 1$. If $\rho \neq 0$, no scalar or diagonal G attains zero error.*

The two columns of Y are orthogonal and nonzero over a period, so Y has full column rank. Consequently, $YA_\rho G = Y$ implies $A_\rho G = I$ and $G = A_\rho^{-1}$. The matrix A_ρ is SPD for $0 \le \rho < 1$, hence its inverse is SPD. If G is diagonal and $\rho \neq 0$, $A_\rho G$ retains a nonzero off-diagonal entry and cannot equal I. A scalar gate is a special case.

B MECHANISM AND EQUIVALENCE CONTROLS

This section retains only controls that directly interpret the architectural gate: its equivalence to optimizer preconditioning, the need to transport coordinates when the map changes, and the complete transported implementation in a conventional nonlinear classifier. These are mechanism checks, not separate performance claims.

B.1 ARCHITECTURAL-PRECONDITIONER EQUIVALENCE CONTROL

We constructed five exact condition-10^4 least-squares problems in 12 dimensions. Inverse-square-root coordinates reduce the Hessian condition to 1; the architectural and explicit-preconditioner paths

differ by at most 4.4×10^{-16} and reach relative loss below 10^{-10} in eight steps, whereas plain gradient descent does not reach it within 200. We then co-trained a tanh feature map and its linear head. Although the relative feature-covariance drift reaches median 1.26, the two predictions remain within 4.9×10^{-15}, confirming that upstream feature motion alone does not break a fixed-head equivalence.

Finally, we rotate a condition-100 quadratic metric and compare explicit preconditioning with two architectural updates. Transporting $\phi \leftarrow G_{t+1}^{-1} G_t \phi$ preserves the represented coefficient vector and matches the explicit path within 2.4×10^{-15}. Carrying the old ϕ instead causes median coordinate shocks of 0.85 under every-step refresh and median post-burn-in loss 3.04, versus 3.84×10^{-25} with transport. This control isolates coordinate replacement from stale-metric conditioning without asserting that repeatedly replacing the coordinate map is universally beneficial.

We also ran a five-seed falsification suite beyond squared loss. Stochastic binary logistic regression matches the explicit preconditioner within 1.0×10^{-15} in parameters and 2.3×10^{-16} in loss. A stochastic three-class softmax head with a co-trained tanh trunk matches predictions within 3.8×10^{-15} despite median feature drift 2.11. Momentum remains equivalent when its velocity is transported, with maximum parameter discrepancy 3.3×10^{-14}. Coordinatewise Adam is a boundary: naively applying $GG^\top$ to a standard Adam step differs by as much as 3.4×10^{-2} on the first update because elementwise second moments are not rotation-equivariant. Thus the proposition covers arbitrary differentiable minibatch losses under gradient descent, and transformed-state momentum, but not standard Adam by mere insertion of the same SPD matrix.

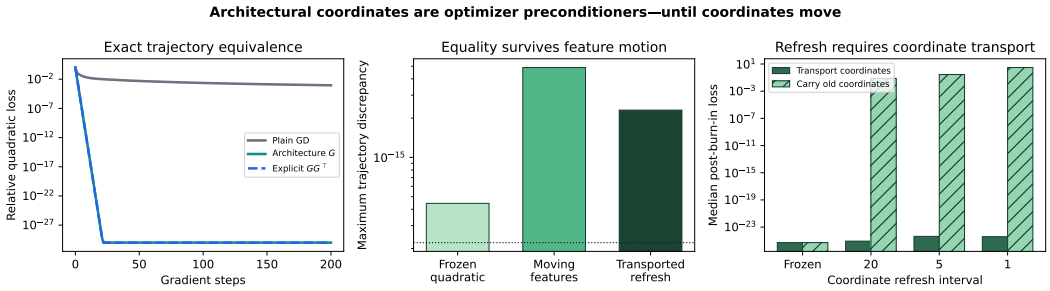


Figure 4: Direct tests of architectural preconditioning. A fixed coordinate map matches explicit preconditioning on a quadratic and while nonlinear upstream features learn. When the map changes, coordinate transport preserves the exact path; carrying old coordinates introduces refresh-dependent shocks.

Transported optimizer equivalence in a conventional classifier. To verify the complete transported implementation outside a coordinate MLP, we use a conventional $784 \to 128 \to 64 \to 10$ ReLU classifier and insert 32 zero-readout nonlinear residual neurons selected from a 256-candidate bank. On a fixed development checkpoint, we compare architectural gated SGD with the explicit optimizer in Eq. (16), including momentum, Nesterov acceleration, and correctly transformed weight decay. The insertion jump is zero. Across ten epochs, the maximum differences are 1.51×10^{-9} in training loss, 6.75×10^{-9} in validation loss, 1.14×10^{-5} in logits, and 1.61×10^{-6} in represented raw coordinates; accuracy is identical at every recorded epoch. This upgrades the earlier synthetic equivalence check to the complete transported growth parameterization in a conventional nonlinear classifier. It also clarifies the practical interpretation: the forward gate is an optimizer reparameterization that can be folded into the trained branch, rather than a different inference-time function class. This control verifies equivalence; it is not used as evidence of improved conventional-network generalization.

B.2 Finite-budget training-loss descent in conventional growth

We next ask whether the gate changes loss descent in the same conventional classifier, rather than merely reproducing an explicit optimizer. We use the previously frozen seeds 61–65 and the locked $784 \to 128 \to 64 \to 10$ ReLU MLP protocol. After five inherited epochs, each model receives the same 32 selected and transported residual coordinates from the same 256-candidate bank. Full

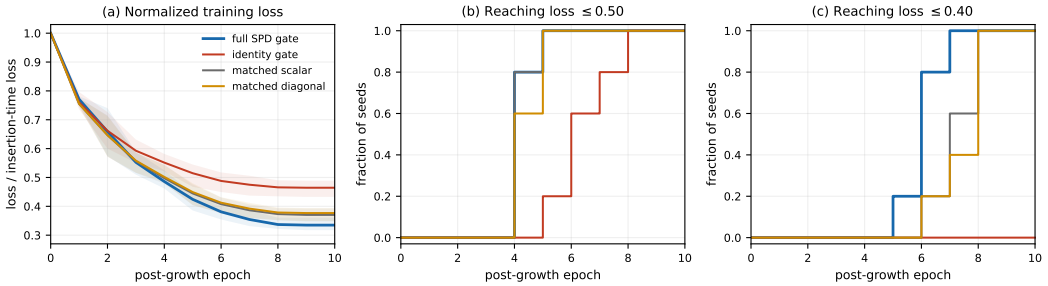


Figure 5: Finite-budget loss descent in locked conventional MNIST growth ($n = 5$ seeds). (a) Mean normalized fixed-probe training loss; bands span the five paired seeds. (b–c) Fraction of seeds reaching fixed fractions of the insertion-time loss. All arms use the identical selected and transported block; only the gate changes. These curves diagnose optimization, not generalization.

SPD, identity, matched scalar, and matched diagonal gates share minibatch order, momentum SGD, learning-rate schedule, and ten-epoch budget. Every insertion jump is zero.

For seed i, let $\widetilde{L}_i(t) = L_i(t)/L_i(0)$ denote fixed-probe training cross-entropy normalized at insertion, and define the hitting time

$$T_i(\tau) = \min\{t : \widetilde{L}_i(t) \le \tau\}.$$

The training-loss AUC endpoint was pre-specified in the locked protocol; the hitting-time view is a descriptive analysis of those frozen trajectories. Full SPD reduces AUC relative to the identical identity-gated block by 12.91% on average with five of five paired wins (exact two-sided sign-randomization $p = 0.0625$). The corresponding reductions are 3.04% versus the matched scalar gate and 3.72% versus the matched diagonal gate, again with five of five wins. Mean terminal normalized loss is 0.335 for full SPD, compared with 0.464, 0.370, and 0.377 for identity, scalar, and diagonal gates.

The time-to-target profiles make the finite-budget interpretation concrete. For $\tau = 0.50$, every method succeeds, but the median hitting epoch is 4 for full SPD and 6 for identity. For $\tau = 0.40$, full SPD succeeds in all five runs at median epoch 6; scalar and diagonal succeed in all five at median epochs 7 and 8, while identity succeeds in zero of five by epoch 10. At the harder $\tau = 0.35$ target, success counts are three of five for full SPD, one of five for each matched control, and zero of five for identity.

This distinction matters: full SPD worsens validation-loss AUC by 5.44% and final test cross-entropy by 9.54% relative to identity in this cohort, while mean test accuracy changes by only -0.05 percentage points. We therefore claim faster finite-budget training-loss descent, not improved held-out classification, a different limiting solution, or asymptotic convergence.

C Boundary and Generalization Studies

C.1 Higher resolution and strong INR baselines

We lock a bundled topobathymetry field resampled to 96×96. The growth architecture and 48-coordinate budget are unchanged; the coarse target is 24×24. On fresh seeds 600–609, the exact gate reduces post-growth AUC by 6.51% relative to the identical selected and transported block with no gate (paired-bootstrap 95% CI $[4.94, 8.49]\%$). Held-out MSE falls by 3.95%. It also reduces AUC by 4.98% versus residual-selected raw-whitened growth, 3.34% versus an earlier sequential transport baseline, and 18.44% versus pure-Fourier growth; every interval excludes zero. The raw selected-block condition has median 1342.6; the damped gated condition is 1.149 and the gated functional condition is 3.04. Every insertion jump is zero.

The stronger boundary is the endpoint comparison in Figure 6. We independently tune a standard two-layer SIREN (Sitzmann et al., 2020) and a multiscale Fourier-feature Tanh network (Tancik et al., 2020). Each has 751 trainable parameters versus 745 for the grown model, and each receives 700 full-target Adam steps; Schur-OST uses 400 coarse-target Adam steps followed by 300 full-target

SGD steps. Their locked mean held-out MSEs are 0.002555 and 0.002478, respectively, versus 0.004167 for SCHUR-OST. Thus the method does not beat a fresh spectral INR when restart is allowed. Its supported claim is function-preserving expansion.

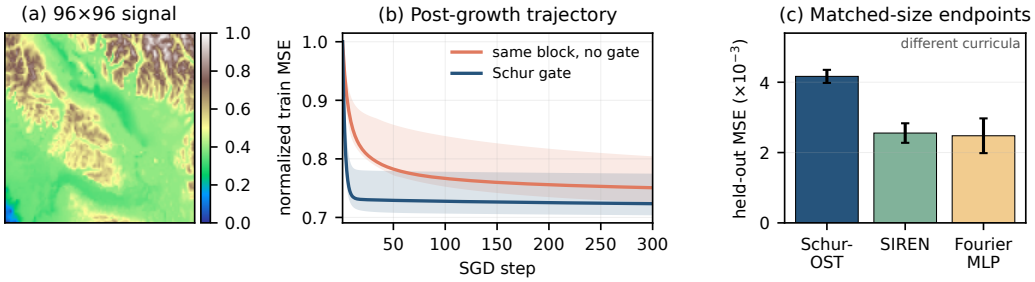


Figure 6: Locked 96×96 boundary study. (a) Normalized topobathymetry signal. (b) Median post-growth trajectories: the only intervention is the Schur gate. (c) Mean held-out MSE for matched-size endpoints; SIREN/Fourier use a different, from-scratch full-target curriculum.

C.2 FROZEN MULTISCALE/DISCONTINUOUS PROTOCOL AND FULL RESULTS

The complete configuration was frozen before evaluation. Development uses only Blocks with seed 40. A single common SGD rate is chosen from $\{0.003, 0.01, 0.03, 0.1\}$ by minimizing mean normalized training-loss AUC over the full, identity, scalar, and diagonal arms; unstable rates are ineligible and the smaller rate breaks ties. This selects 0.03. The evaluation seeds 50–54 are unused during development.

All signals use 512 points on $[0, 1]$ and an exact seeded random 75/25 train/test split. The inherited width-24 Tanh MLP receives 800 Adam steps at rate 0.005 on a 48-point linear coarse interpolation. Growth uses 400 full-batch SGD steps, relative ridge 0.003, eigenvalue-floor ratio 10^{-6}, and $k = 32$ selected coordinates. The fixed 528-column bank contains 96 smooth edge atoms, dyadic compact hats and Mexican hats, and Gaussian-windowed sine/cosine atoms with frequencies 4, 8, 16, and 32. Selection is conditional Schur gain computed from the exact calibration rows.

Writing the candidate design as H, the inherited functional Jacobian as J, and $S = JJ^\top + \lambda I$, the implemented dual quantities are

$$K_\lambda = \lambda H^\top S^{-1} H/N, \qquad B_\lambda = J^\top S^{-1} H_A, \qquad G_\lambda = (K_{\lambda,AA})^{-1/2}.$$

The expanded model evaluates $f_{\theta - B_\lambda G_\lambda \phi}(x) + H_A(x) G_\lambda \phi$ with $\phi_0 = 0$, making insertion exactly function preserving. Scalar and diagonal controls share A and B_λ and are scaled to the full gate's maximum transformed curvature.

Table 3: Frozen multiscale/discontinuous evaluation over five signals and five unused seeds. Positive reductions favor full SPD. Intervals cluster-bootstrap the five signal families while retaining seeds within signal.

Comparator	ΔAUC	95% CI	Δtest	95% CI
Identical block, no gate	64.61%	[53.90, 74.75]%	39.56%	[21.81, 51.21]%
Scalar, max-curvature matched	46.14%	[35.82, 60.28]%	20.29%	[7.85, 30.36]%
Diagonal, max-curvature matched	37.29%	[23.40, 54.40]%	10.68%	[0.68, 20.68]%

Mean AUC/MSE reductions versus identity by signal are Blocks $61.43/39.36\%$, HeaviSine $67.70/49.66\%$, Bumps $46.61/5.58\%$, Doppler $82.49/50.22\%$, and piecewise chirp $64.83/52.98\%$. The sole identity MSE loss is Bumps seed 51 (-59.78%); it is neither removed nor winsorized. The maximum post-gate condition is 1.00477 and the maximum recorded insertion jump is zero. Primary outcomes are AUC and held-out MSE; relative L_1/L_2, total variation, near-jump maximum error, Haar-band errors, and selected atom metadata are retained as mechanism diagnostics rather than additional claims.

This experiment tests function approximation, not conservation laws. A shock extension would require a weak or finite-volume-compatible residual and an entropy/admissibility diagnostic; pointwise strong-form PINN residual alone would not establish physically correct shock learning.

C.3 Support creation and negative controls

The half-wave construction supplies an exact support control. A bias-free odd network has zero even-harmonic tangent curvature on a paired periodic grid and cannot fit $\cos 2x + 0.5 \cos 4x$; its MSE remains at the exact target-energy floor 0.625. A four-parameter zero-readout Fourier adapter opens those derivatives and reaches median MSE 6.6×10^{-9}. Separately, the blind Schur selector recovers two useful harmonics from a bank containing supported and irrelevant frequencies, and rejects a candidate dominated by a $1000\times$ supported component.

This is the minimal support-creation control used by the paper: conditioning cannot create a missing tangent direction, whereas a zero-readout expansion can open it without changing the inherited function. It is not presented as a parallel method family or as evidence for semantic frequency recovery.

D Experimental Details

Table 4: Retained experimental configurations.

Experiment	Points	Seeds	Epochs	Precision	Optimizer
Multiscale stress test	512	1+25	800+400	float32	Adam/SGD
Minimal support control	128	5	exact/trajectory	float64	GD
Classifier equivalence	60,000	1	5+10	float32	SGD
Locked Schur-OST	64^2	10+10	400+300	float32	SGD
Topobathy extension	96^2	3+10	400+300/700	float32	SGD/Adam
Sketch-Schur-OST BSDS	256^2	2×3+6×3	400+300	float32	SGD
Sequential baseline	256^2	2×3+6×3	400+300	float32	SGD
Gate ablation	256^2	6×3	400+300	float32	SGD
Coordinate-budget sensitivity	256^2	3×3×5 budgets	400+300	float32	SGD
Long-horizon gate test	256^2	6×3	400+3000	float32	Adam/SGD
Adam gate transfer	256^2	2×3+6×3	400+300	float32	Adam
Dual scale BSDS	64^2	1+6 images	200+100	float32	Adam/SGD
Selection fidelity	64^2	1 image, 5 sketches/s	200+100	float32	Adam/SGD
Dual parameter scaling	64^2	1 image, 4 widths	20+calibration	float32	Adam/none

For the scalable study, six images from the BSDS500 test split are used only as coordinate-regression signals; segmentation annotations are not used. Each image is converted to grayscale and center-cropped to 256×256. The development images are 100007 and 100039 with seeds 20–22. The untouched evaluation images are 100099, 10081, 101027, 101084, 102062, and 103006 with seeds 600–602, yielding 18 paired cases. A deterministic seed-dependent interleaved pixel class holds out 16,384 of 65,536 pixels in every case.

The base is a width-24, two-hidden-layer Tanh coordinate MLP trained for 400 Adam steps at rate 0.01 on a 64×64 downsampled target. Growth uses 300 full-batch SGD steps at rate 0.03. The rate is selected before evaluation from $\{0.03, 0.1, 0.3\}$: 0.03 has the lowest six-case mean development AUC for every retained method, while 0.3 is unstable. All evaluation arms receive the same rate; there is no per-method rate advantage.

The bank contains 64 continuous two-dimensional carriers of maximum frequency 30, each modulating a constant and the four most variable standardized hidden channels, with sine and cosine phase, for 640 columns. The full bank is Schur-selected on 3,072 uniformly sampled training pixels with scaling $\sqrt{49152/3072} = 4$. The selector retains 48 columns, uses relative ridge 0.003, and floors gate eigenvalues at 10^{-3} of the largest. A disjoint 1,536-row sketch audits the fixed gate; a separate 1,024-row sketch supplies the residual/random controls. The exact full-data oracle is prohibited in evaluation. A single development-only 128^2 audit gives AUC 0.73223 for the sketch and 0.70967 for exact selection (3.18% relative difference), while reducing the materialized calibration proxy fourfold.

All growth controls add 48 trainable scalars. The parameter-matched Tanh arm adds 12 neurons whose two input weights, bias, and output coordinate are trainable; candidate neurons are selected by residual gradient magnitude and their selected raw block is whitened. The primary identity arm shares the exact selected indices and ridge transport with SKETCH-SCHUR-OST. Every arm begins at zero readout and the same represented function. The primary statistic is the paired relative reduction in normalized training-MSE AUC. The fixed-seed 20,000-draw primary interval resamples the six images and retains all three seeds of every sampled image; an ordinary 18-pair bootstrap gives the sensitivity interval $[4.59, 8.97]\%$.

The direct sequential comparison was frozen separately before full-resolution results were observed. Its baseline greedily selects 48 normalized raw-gradient-correlated columns with residual refitting, applies the same relative-ridge transport, and whitens only $R_\lambda^\top R_\lambda/N$, omitting $\lambda B_\lambda^\top B_\lambda/N$. The same two development images and seeds 20–22 select rate 0.03 by six-run mean AUC from $\{0.03, 0.1, 0.3\}$; rate 0.1 is worse and 0.3 is unstable in all six runs. The fresh cohort contains images 164046, 220003, 235098, 45000, 71076, and 81095 with seeds 700–702. These were sampled with fixed seed 20260917 after excluding development, headline, and prior A2 images, and did not occur in the existing result tree at lock time. Both arms therefore use rate 0.03, the same inherited model, bank, $k = 48$, ridge, zero readout, and 300-step budget.

Across the 18 fresh pairs, SKETCH-SCHUR-OST reduces AUC by 3.45% (image-cluster 95% interval $[2.45, 4.50]\%$), held-out MSE by 3.54% ($[2.59, 4.46]\%$), and final training MSE by 3.62% ($[2.60, 4.65]\%$). All image-level mean effects are positive. Raw and Schur selection overlap in 30.8 of 48 columns on average, so the test does not collapse to the identical-block ablation. Mean calibration time is 0.53 seconds for the sequential arm and 5.83 seconds for SKETCH-SCHUR-OST; mean 300-step training times are 33.74 and 34.14 seconds. The claim is therefore better optimization trajectory, not lower end-to-end wall time.

The full-versus-scalar/diagonal extension is frozen in advance. Every arm shares the inherited function, selected 48-column block, ridge transport, zero readout, rate 0.03, and 300 steps. For full gate G, the trace-matched scalar is cI, $c^2 = \mathrm{tr}(GG^\top)/48$; the inverse-standard-deviation diagonal gate is rescaled to the same trace. Trace equality holds to maximum relative error 2.39×10^{-7}. This aggressive common-rate comparison is stable throughout for the full gate, whereas the scalar and diagonal controls show occasional instability.

Before a second run, we froze the outcome-independent stability normalization: scalar and diagonal gates are rescaled so their maximum transformed Schur curvature equals the full gate's, and all retain rate 0.03. All 54 arms are stable. Full SPD reduces AUC by 7.52% versus scalar (image-cluster 95% interval $[4.01, 11.52]\%$) and 7.15% versus diagonal ($[3.83, 10.78]\%$). Held-out MSE reductions are 5.83% ($[2.90, 9.05]\%$) and 5.38% ($[2.71, 8.14]\%$). Median transformed conditions are 276.7, 236.4, and 1.00003 for scalar, diagonal, and full. Every insertion jump is zero, and rerun full/identity endpoints match the original locked files bit-for-bit.

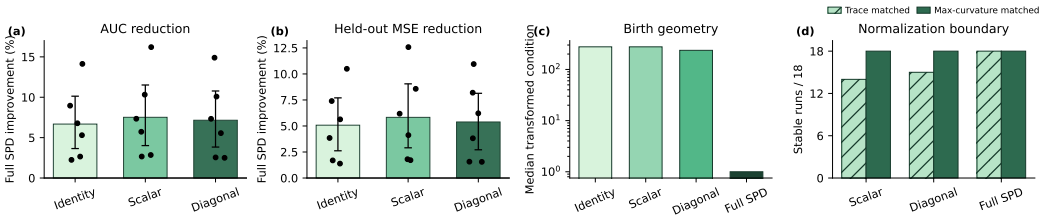


Figure 7: Headline gate ablation. (a,b) Full-SPD improvement over identity and maximum-curvature-matched scalar/diagonal controls; dots are six image-cluster means and bars show the overall mean with cluster-bootstrap 95% intervals. (c) Only the full gate removes within-block conditioning. (d) Trace matching exposes a stability boundary; maximum-curvature matching makes every arm stable without outcome-based tuning.

D.1 Optimizer and horizon robustness protocols

The long-horizon study was frozen before execution. It reuses the primary six-image cohort and seeds 600–602, but reruns rather than resumes both identical-block arms for 3,000 post-growth SGD steps at the already locked rate 0.03. Checkpoints at 300, 1,000, and 3,000 steps are specified in advance. The selected indices, ridge transport, zero readout, data split, and optimizer trajectory are paired within every image–seed case. The 300-step rerun differs from the published primary AUC and held-out effects by less than 0.25 percentage points, providing a direct implementation check before the longer endpoints are interpreted.

The Adam transfer protocol is frozen separately. On excluded development images 100007 and 100039 with seeds 20–22, rates $\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$ are screened for 300 Adam steps. The declared rule minimizes the mean AUC across both gate and identity arms, excludes unstable rates, and breaks exact ties toward the smaller rate; it selects 0.003. That single rate is frozen before one evaluation on the primary six-image cohort. Adam states begin at zero in both arms with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$; no per-arm or evaluation-case retuning is allowed.

Table 5: Frozen horizon and optimizer robustness. Positive values favor the full SPD gate over the identical selected and transported identity-gate block. Intervals cluster-bootstrap the six images while retaining all three seeds of each sampled image.

Optimizer	Steps	ΔAUC	95% CI	Δtest	95% CI
SGD	300	6.59%	[3.65, 10.04]%	5.01%	[2.61, 7.65]%
SGD	1,000	4.63%	[2.45, 7.01]%	2.88%	[1.48, 4.23]%
SGD	3,000	2.93%	[1.51, 4.33]%	1.48%	[0.62, 2.42]%
Adam	300	1.74%	[0.92, 2.56]%	0.99%	[0.31, 1.74]%

All paired AUC effects are positive. One 3,000-step SGD held-out effect and two Adam held-out effects are negative and are retained in the aggregates. Every insertion jump is exactly zero in the recorded float32 evaluations. The decreasing SGD effect supports persistence over the tested horizon, not a different asymptotic solution; the smaller Adam effect supports transfer to one adaptive optimizer, not coordinatewise-Adam equivalence.

D.2 Coordinate-budget sensitivity protocol and full results

The coordinate-budget study was frozen before its full result matrix was observed. Images 120003, 120093, and 123057 did not occur in the existing result tree at lock time and are disjoint from the headline, sequential, and embedding-scope cohorts. Seeds 730–732 are reused across $k \in \{16, 32, 48, 64, 96\}$ to form nine paired cases per budget. The inherited width-24 Tanh MLP, $64^2 \to 256^2$ task, 640-column bank, 1,024/3,072/1,536 screen/selection/audit rows, relative ridge 0.003, eigenvalue floor 0.001, rate 0.03, and 300 post-growth SGD steps are fixed. No per-budget, image, seed, or arm tuning is permitted. The gate and identity arms share selected columns, transport, initialization, split, and insertion function.

Table 6: Fresh coordinate-budget sensitivity. Each row contains nine paired gate/no-gate cases. Positive effects favor full SPD; intervals are descriptive image-cluster bootstrap intervals over three independent images. Pre and audit columns report median selected-block condition before gating and on the independent sketch after gating.

k	ΔAUC	95% CI	Δtest	95% CI	Pre cond.	Audit cond.
16	2.64%	[1.09, 4.39]%	0.99%	[0.67, 1.59]%	23.3	1.49
32	5.45%	[2.71, 8.23]%	3.05%	[1.73, 4.24]%	375.3	1.82
48	7.18%	[3.43, 10.09]%	4.65%	[2.15, 6.45]%	679.4	2.14
64	8.26%	[4.70, 11.35]%	5.56%	[3.23, 7.62]%	1,072.8	2.59
96	10.17%	[6.22, 13.27]%	7.31%	[4.58, 9.49]%	1,415.0	3.64

All training runs are numerically stable, and the mean AUC and held-out MSE effects are positive at every budget. Maximum insertion jump is zero. Median sketched post-gate conditions are 1.00001, 1.00003, 1.00011, 1.073, and 1.415, respectively; mean calibration time ranges only from 5.54 to

5.66 seconds. The locked interpretation criterion—a positive mean AUC effect with broad paired consistency for $k = 32$–96—is therefore met at every specified budget, including the small-block boundary $k = 16$.

The pre-gate condition was added to the summary after the first two extreme-budget validation jobs had completed, without changing any run or method setting. Across the 45 pairs, Pearson correlation between $\log_{10}$ condition and AUC benefit is 0.647 and Spearman correlation is 0.600. We report this only as exploratory support for the mechanism

$$\begin{aligned} \text{larger selected block} &\longrightarrow \text{greater anisotropy} \\ &\longrightarrow \text{greater value from full-matrix conditioning.} \end{aligned}$$

The data establish a positive mean trend on this cohort, not pairwise monotonicity or a universal scaling law.

At 256^2, materializing $[J, H]$ would require 250.69 MiB in float32; the 3,072-row design uses 15.67 MiB. Mean gate calibration is 0.97 seconds. Mean 300-step training time is 61.7 seconds for SKETCH-SCHUR-OST, 61.7 for its identity ablation, 62.8 for residual sidebands, and 12.2 for Tanh growth on the single-thread CPU implementation. These timings are descriptive and motivate caching/vectorizing the sideband branch rather than a wall-clock claim.

The damped-versus-undamped audit reruns the frozen 18 calibration states with one post-growth step only; it does not refit hyperparameters or create a new training claim. We define the actual normalized cross ratio as

$$\frac{\|J^\top R_\lambda G_\lambda / N\|_2}{\sqrt{\|J^\top J/N\|_2 \|G_\lambda^\top R_\lambda^\top R_\lambda G_\lambda / N\|_2}},$$

and the damped residual replaces the numerator by $\|J^\top R_\lambda G_\lambda / N - (\lambda/N) B_\lambda G_\lambda\|$. The joint functional condition uses the positive spectrum of $[J, R_\lambda G_\lambda][J, R_\lambda G_\lambda]^\top / N$ with the locked float32 tolerance. On the calibration sketches the median damped and actual ratios are 1.32×10^{-6} and 0.0418, respectively; the actual cross identity has maximum relative error 2.01×10^{-4}. On independent sketches the medians are 0.1049 and 0.1846. Median raw, transported, and gated joint conditions on the calibration sketch are $8.91, 8.98, 8.98 \times 10^4$, confirming that global condition reduction is not the gate claim.

The parameter-scale dual protocol is frozen in advance. Development uses training image 157036 and seed 810. Evaluation uses official BSDS500 training images 100075, 100080, 100098, 103041, 104022, and 105019 with seeds 820–825. Images are center-cropped to 64^2 and one interleaved pixel quarter is held out. The inherited width-224 Tanh MLP has 51,297 parameters and is trained for 200 Adam steps at rate 0.005 on a 24^2 coarse target. The shared sketch has 256 rows; the bank, $k = 48$, relative ridge, and gate floor match the headline study. Both identical-block arms independently sweep SGD rates $\{0.003, 0.01, 0.03, 0.1\}$ on the development image only. Their gated/ungated AUCs are respectively $0.8786/0.9576$, $0.7714/0.9002$, $0.7218/0.8406$, and unstable/unstable, so 0.03 is frozen for both arms. Every evaluation image is then rerun once at 0.03; no evaluation image or held-out pixel selects a rate. We report descriptive effects without a significance claim.

The theorem-scope audit is frozen separately. It reconstructs all six locked states exactly (including all selected indices) and accumulates the complete training-row spectra in float64 using exact row-NTK identities, while the model and calibration remain float32. Table 7 deliberately separates full-bank scope diagnostics from the geometry of the fixed selected block. Numerical rank uses a relative singular-value threshold; the 10^{-3} rank already exceeds 256 for every image. Consequently the standard $\epsilon < 1$ full-bank premise is not certified at the deployed sketch size. The combined effective dimension applies an isotropic ridge to $[J, H]$ only as a diagnostic; it is not substituted into Theorems 4–5.

For the selected-block rows, the full damped metric is recomputed on all 3,072 training pixels while retaining the sketch-selected indices, transport, and gate. The extrema are generalized Rayleigh quotients relative to the 256-row calibration metric, and $\epsilon = \max\{1 - \rho_{\min}, \rho_{\max} - 1\}$. This is labeled *selected-block empirical embedding distortion*: it audits one deployed block after selection and is not an empirical verification of Theorem 5.

The selection-fidelity protocol is frozen in advance. On official training image 187029 with seed 811, a width-24 model is trained for 200 steps before forming the complete 3,072-row float32 design.

Table 7: Locked 51k embedding-scope audit over six images. Entries are median [minimum, maximum]. "Selected" holds the deployed 48-column block fixed and does not test simultaneous selection fidelity.

Scope	Diagnostic	Median [range]
Full bank	$\mathrm{rank}_{10^{-3}}([J, H])$	603.5 [570, 636]
Full bank	$\mathrm{rank}_{10^{-4}}([J, H])$	639 [615, 658]
Full bank	stable rank of $[J, H]$	14.03 [12.99, 18.06]
Full bank	isotropic-ridge $d_{\mathrm{eff}}([J, H])$	292.68 [232.80, 322.90]
Full bank	ridge $d_{\mathrm{eff}}(J)$	4.61 [4.40, 4.91]
Selected	minimum Rayleigh quotient	0.524 [0.514, 0.529]
Selected	maximum Rayleigh quotient	3.33 [2.97, 3.82]
Selected	empirical ϵ	2.33 [1.97, 2.82]
Selected	full deployed condition	18.24 [6.54, 151.37]

The bank, $k = 48$, ridge, gate floor, and 100-step downstream rate 0.03 are fixed. Five independent uniform sketches use $s \in \{64, 128, 256, 512, 1024, 1536, 3072\}$; the last contains every training row. Median normalized terminal score error is 0.717, 0.870, 0.572, 0.267, 0.0646, 0.0177, and 1.55×10^{-6}. Median worst normalized marginal error decreases from 0.121 to 0.0105 at $s = 1536$ and 1.42×10^{-6} at full rows. Median exact-set overlap is 8.3%, 8.3%, 22.9%, 35.4%, 52.1%, 62.5%, and 100%. The 64-row gate is consistently unstable; among stable sizes, median downstream AUC excess over exact selection falls from 42.1% at $s = 128$ to 0.84% at $s = 1536$ and numerical zero at full rows. This audit uses one development image and is not an additional held-out performance claim.

The parameter-scaling protocol is frozen in advance. On training image 71046, widths $\{30, 98, 224, 445\}$ give 1,051, 10,095, 51,297, and 200,251 inherited parameters. All widths use a 256-row sketch, 640 candidates, 48 selected coordinates, and three timed dual solves after 20 base steps. At 200,251 parameters the sketched Jacobian is 195.6 MiB, the hypothetical full $P \times q$ transport is 488.9 MiB, and the actual selected $P \times k$ transport is 36.7 MiB. The explicit materialized-array proxy is at least 688.0 MiB for the primal path versus 235.8 MiB for the dual; the former omits any $P \times P$ solve. Median total calibration takes 0.78 s. Every insertion jump is exactly zero. This is a calibration scaling audit, not a predictive endpoint.

For the primary SCHUR-OST study, each seed trains a width-24, two-hidden-layer `tanh` coordinate MLP for 400 steps on a 12×12 coarse target; one interleaved pixel quarter is held out. The bank contains 64 continuous two-dimensional carriers modulating the four highest-variance hidden channels and a constant channel, giving 640 candidates without frequency labels. The blind marginal score in Eq. (6) selects 48 additions. Relative ridge 0.003 and eigenvalue floor 0.001 are inherited from development. Per-arm rates $\{0.03, 0.1, 0.3\}$ are screened on seeds 10–14; rate 0.1 is then locked for every retained arm. Evaluation uses untouched seeds 500–509 on both signals, 300 full-batch SGD steps, and float32 throughout. The four prespecified comparisons, paired bootstrap, exact one-sided signed-rank tests, and Holm correction are fixed in advance.

For the separately locked topobathymetry extension, the signal is bicubically resampled and min–max normalized to 96×96; the coarse target is 24×24. The base width, carrier bank, 48-coordinate budget, ridge, eigenvalue floor, and post-growth optimizer match the primary study. Development seeds 20–22 select rate 0.1 from $\{0.03, 0.1, 0.3\}$; evaluation uses untouched seeds 600–609. The SIREN has two width-25 sine layers and 751 trainable parameters. The multiscale baseline uses 16 fixed, geometrically spaced two-dimensional Fourier frequencies and two width-15 tanh layers, also totaling 751 parameters; the grown model has 745. Each baseline is trained from scratch on normalized coordinates for 700 full-target Adam steps. Independent rate grids are extended until both have an interior optimum; development selects 0.03 for each. The protocol is fixed in advance.

E CLAIM-TO-EVIDENCE LEDGER

Support versus conditioning. Theorems 1– 2 show that SPD coordinate maps preserve the tangent range; the minimal control reaches its exact error floor until a zero-readout adapter opens the excluded directions.

Exact Schur construction. Theorem 3 proves that ridge transport and the SPD gate factor one damped augmented tangent metric; float32/float64 tests verify the identities and exact insertion.

Sketch guarantees and scope. Theorems 4– 5 give guarantees under a shared embedding event. The fidelity audit converges to exact selection, while deployed $s = 256$ is explicitly outside the certified full-bank-rank regime.

Optimizer interpretation. A fixed architectural gate is exactly an SPD reparameterization for gradient descent and transported momentum, and the classifier implementation matches the explicit optimizer numerically. In a separate locked cohort, full SPD reduces training-loss AUC by 12.91% and reaches loss 0.40 in 5/5 rather than 0/5 runs, but worsens held-out cross-entropy. Standard coordinatewise Adam is not claimed equivalent.

Primary gate effect. On the locked BSDS500 cohort, full SPD improves training-loss AUC by 6.68% and held-out MSE by 5.08% over the identical ungated block, with 18/18 wins; matched scalar and diagonal gates also lose in 18/18 cases.

Unified versus sequential. Against a separately locked direct sequence of raw-gradient selection, the same ridge transport, and functional-only whitening, SKETCH-SCHUR-OST improves AUC by 3.45% and held-out MSE by 3.54% with 18/18 wins. No wall-clock superiority is claimed.

Budget and optimizer robustness. All 45 gate/no-gate comparisons are positive over $k = 16$–96. The effect attenuates but remains positive at 3,000 SGD steps and under Adam; no optimizer invariance or different limit is claimed.

Heterogeneous coordinates. In the frozen five-signal multiscale study, full SPD reduces mean AUC/MSE by $64.61\%/39.56\%$ with 25/25 and 24/25 wins. The retained Bumps loss and lack of a shock-PDE diagnostic bound the claim.

Dual and scaling evidence. The row-space dual preserves the gate effect on six 51,297-parameter networks; calibration reaches 200,251 inherited parameters with exact insertion. No end-to-end wall-clock speed is claimed.

Strong-model boundary. On the locked 96^2 field, parameter-matched SIREN and multiscale Fourier networks trained from scratch obtain lower held-out MSE. The supported use case is function-preserving expansion, not universal endpoint superiority.