FROM PREIMAGE SEARCH TO SOURCE-GROUNDED FEATURE INVERSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature inversion seeks to reveal what an internal representation extracts from a given input. When multiple inputs match a target representation, canonical preimage search alone does not specify the generating sample's inverse. We recast inversion as source-grounded upstream-state estimation along the target-generating directed acyclic graph, conditioned on source-local network geometry. At each boundary, a closed-form matrix Wiener map converts a mean-seed vector–Jacobian product from an adjoint signal into an upstream-state estimate. A second Wiener map estimates remaining state error from a pulled-back Jacobian–vector-product forward-consistency residual. Repaired states compose in one finite reverse pass. Fixed calibrated zero-intercept maps serve new inputs, depths, and tested channel–coordinate queries without query-specific optimisation. The formulation spans tensor components and visual distributions across convolutional networks and vision Transformers. Target–operator controls show that selected features and source-local geometry jointly determine architecture-dependent inverses. For GPT-2 small terminal-token queries, the same construction yields continuous input-embedding inverses without a text decoder. At an exploratory layer-6 contextual peak, they show semantic-cue selectivity and number-agreement routing. Across modalities, prediction-conditioned atlases display selected input-domain structure through inversion; independent interventions on corresponding original representations provide causal evidence of their decision effects. Results support source-grounded feature inversion as a reusable interface for interpreting internal representations.

1 INTRODUCTION

Modern neural networks predict through hierarchical internal representations, but an activation records a feature-space state without showing which structures it extracts from a particular input. Output explanations and sensitivity maps can locate influential input regions but do not show a selected internal state in the input domain at an arbitrary depth or channel. We therefore ask: what does this feature, at this layer and channel, extract from this input? Input-domain coordinates would put heterogeneous layers and architectures in a common space, revealing progression through network.

Feature inversion provides this interface (Figure 1). A query specifies a frozen model f, source input x, target node T, channel set S, and coordinate set Q. We seek a sample-specific input-domain state revealing what the masked target feature $P_{S,Q}H_T(x)$ extracts through the forward computation generated by x. The source specifies the forward operating point rather than a unique inverse label: the task is not reconstruction from a compressed code, and similarity to x does not define correctness.

Canonical inversion searches for target-matching preimages or learns representation-to-input mappings from paired data. Iterative search commonly uses hand-designed natural-image regularisers (Mahendran & Vedaldi, 2015; Olah et al., 2017): $z^\star = \arg\min_z d\big(H_T(z), H_T(x)\big) + \mathcal{R}(z)$. For non-injective representations, matching defines a class of compatible preimages; initialisation, regularisation, the input prior, and optimisation trajectory select among them (Figure 1a). The frozen network's differential operators are evaluated along the candidate trajectory, rather than used to condition the inverse on the forward point that generated $H_T(x)$. Independently trained reverse models typically target particular representation families or output domains in vision (Dosovitskiy & Brox, 2016; Nash et al., 2019) and language (Song & Raghunathan, 2020; Morris et al., 2023). Their recoveries demonstrate substantial retained input information but depend jointly on the target repre-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

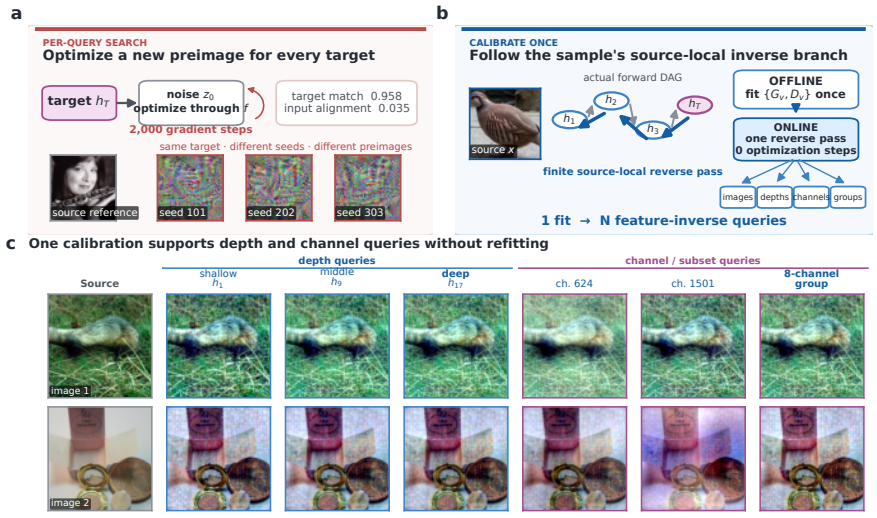


Figure 1: **From target-preimage search to source-grounded feature inversion.** **a**, Iterative inversion repeats target-matching optimisation from noise per query; three seeds attain similar feature match but weak input-instance alignment. **b**, Reusable local repairs compose in one finite reverse pass through source-local boundary DAG. Nodes denote forward states; G_v, D_v are two local repairs. **c**, ResNet50 map family serves image, depth, channel, and channel-group queries without refitting.

sentation, decoder architecture, and data prior. For sample-specific inverse, neither target matching nor learned reverse mapping alone supplies the generating local computation as explicit condition.

An internal feature depends on its computation: x determines forward activations, nonlinear operating regions, branch states, attention routes and local Jacobians on the computational directed acyclic graph (DAG) producing $H_T(x)$. We therefore recast inversion as source-grounded upstream-state estimation on that DAG, querying (f, x, T, S, Q) rather than a target tensor alone. This *source-grounded feature inversion* uses source-local geometry at the selected feature's generating forward point (Figure 1b). The input supplies only frozen forward activations and local differential operators; the selected terminal state seeds reverse recursion, with no direct additive input-to-inverse path. Local estimation on that DAG yields successive upstream-state estimates along a sample-specific local inverse branch. The formulation applies to frozen finite differentiable DAGs exposed by autograd.

The conditional mean defines the MSE-optimal local upstream-state estimation target. A vector–Jacobian product (VJP), however, follows reverse dependencies, transporting cotangent q through the local Jacobian's adjoint as J^*q rather than estimating an upstream forward state. Parent and child metrics M_v and M_u give the natural adjoint $J^\sharp = M_v^{-1} J^* M_u$; with covariance-induced metrics, this adjoint equals local Wiener inversion under the centred local linear-Gaussian reference model. In nonlinear networks, boundary- and source-point-dependent metrics and Jacobians make explicit online metric estimation impractical. We instead calibrate a zero-intercept full-channel matrix Wiener map G_v from a mean-seed VJP observable to the upstream activation. This projection amortises missing metric correction over calibration distribution, retaining source-local Jacobian in VJP input.

A shared G_v can capture the dominant correction for fixed or weakly sample-dependent local operators; stronger sample dependence from normalisation, attention or data-dependent gating or routing can cause operating-point-specific mismatch under this map. A Jacobian–vector product (JVP) applies the source-point differential to the first-stage estimate; comparing its response with child states gives a consistency residual. A second Wiener map D_v predicts the remaining state error from the pulled-back residual. Repaired states compose through the selected boundary DAG in one finite reverse pass, without end-to-end representation-to-input decoder. Fixed calibrated local maps serve new inputs, depths, and tested channel–coordinate queries without query-specific optimisation.

These upstream-state estimates differ from sensitivity, visualisation, routing, and attribution signals. Input gradients measure local sensitivity (Simonyan et al., 2013); deconvolutional visualisation and Guided Backpropagation display activation patterns (Zeiler & Fergus, 2014; Springenberg et al.,

2014); Grad-CAM provides class-gradient spatial localisation (Selvaraju et al., 2017). In language models, attention weights provide routing references (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), while gradient-times-input and Integrated Gradients yield token-level sensitivity or attribution scores (Ancona et al., 2018; Sundararajan et al., 2017). Some saliency outputs can remain visually similar after model or data randomisation (Adebayo et al., 2018), motivating method-specific model-dependence tests beyond visual sharpness. These references neither estimate a representation's continuous upstream forward state nor recursively compose estimates through the source-local DAG.

Representation content and model use require different evidence. Network Dissection measures unit–concept alignment (Bau et al., 2017); TCAV tests prediction sensitivity along user-defined concept directions (Kim et al., 2018); ACE automatically discovers concepts and measures their importance (Ghorbani et al., 2019). Probing tests whether linguistic structure is decodable from hidden states (Hewitt & Manning, 2019); intervention and mediation test internal components' behavioural effects (Vig et al., 2020). Decodability alone does not establish model use. Our prediction-conditioned atlases therefore pair inversion with independent intervention on the same original feature group: inversion displays its input-domain structure, while intervention measures the group's effect on the current decision. Further method descriptions are in section E.

Across CNN and vision Transformer families, tensor components, and visual distributions, inverses reveal architecture-specific spatial and frequency structure. Query studies reuse fixed calibrated maps across inputs, depths, and tested channel–coordinate selections (section 3.1). Zero targets yield zero inverses, channel selections produce distinct structures, and operating-point swaps redirect the inverse to the new source-local branch. These controls show that selected features and their generating geometry jointly determine each visualisation. Independent interventions on original representations yield larger prediction-margin drops for top-positive channel groups than matched random controls across all tested visual settings. Beyond vision, we apply the same construction to GPT-2 small terminal-token queries (Radford et al., 2019). Selected attention-output states map to token-resolved continuous input-embedding inverses without a text decoder. In controlled semantic and number-agreement queries, inverses expose context-specific evidence carried by selected representation. Independent interventions connect the corresponding feature groups to GPT-2's current decision.

Our contributions are threefold: (1) We recast feature inversion from target-matching preimage search to source-grounded upstream-state estimation along a target-generating finite differentiable computational DAG, making its source-local geometry part of the inverse object. (2) We derive two-stage local closed-form matrix Wiener repairs that convert VJP adjoints into upstream-state estimates, predict corrections from operating-point-specific consistency residuals, and compose repaired states through the DAG in one finite reverse pass, with a finite-DAG error-propagation analysis. (3) We validate the same construction across CNNs, vision Transformers, and GPT-2 small, with modality-native inverses and calibrated query reuse. Target–operator controls verify source grounding, while independent interventions test the corresponding groups' decision effects.

2 Source-Grounded Feature Inversion

We estimate upstream states along the source-local computational DAG using two calibrated local matrix Wiener repairs in one reverse pass (Figure 1). The construction proceeds from the statistical inverse to its geometric correction, then to two observable-based state estimators.

Let f be a frozen differentiable model with finite computational DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ exposed by autograd. A query (f, x, T, S, Q) selects target node T, channels S and coordinates Q at source input x. Let $H_v(x)$ be the forward state at boundary v, and $\mathcal{C}_v$ its non-overlapping child frontier on paths to T, avoiding double-counted downstream computation. Holding external subgraph inputs not descended from v at their source values defines, for each $u \in \mathcal{C}_v$,

$$H_u(x) = f_{u \leftarrow v, x}(H_v(x)), \qquad J_{u \leftarrow v}(x) = \left. \frac{\partial f_{u \leftarrow v, x}}{\partial H_v} \right|_{H_v = H_v(x)}. \tag{1}$$

The Jacobian includes every v-to-u path in this subgraph. Locality means boundary-DAG adjacency, which can encompass global attention or pooling (section A.1). The input instantiates this forward point and its derivatives; the selected terminal state drives the reverse computation. True states supply calibration targets; deployment uses already inverted children. The output estimates an upstream state in the input domain; source similarity is a descriptive diagnostic.

2.1 Statistical Inversion and the Local Geometric Gap

For square-integrable states on data distribution, MSE-optimal estimator from downstream states is

$$\mathcal{T}^\star \in \arg\min_{\mathcal{T}} \mathbb{E}\|\mathcal{T}(H_u) - H_v\|_2^2, \qquad \mathcal{T}^\star(h_u) = \mathbb{E}[H_v \mid H_u = h_u]. \tag{2}$$

This generally unavailable conditional law motivates local estimation along the source computation. Linearising a local map gives $\delta H_u = J\delta H_v$, with $J = J_{u\leftarrow v}(x)$. Ordinary backpropagation uses the Euclidean adjoint to transport a child cotangent q through the VJP J^*q (Rumelhart et al., 1986; Baydin et al., 2018), yielding a differential signal rather than an upstream-state estimate. Here $*$ denotes conjugate transpose, equal to $\top$ for real-domain Jacobians. To relate adjoint transport to state inversion, equip the upstream parent and downstream child with Hermitian positive-definite metrics M_v, M_u and inner product $\langle a, b\rangle_M = a^*Mb$. The natural adjoint satisfies, for all a_v, a_u,

$$\langle Ja_v, a_u\rangle_{M_u} = \langle a_v, J^\sharp a_u\rangle_{M_v} \quad \Longleftrightarrow \quad J^\sharp = M_v^{-1}J^*M_u. \tag{3}$$

The child metric converts a downstream vector to a cotangent before the J^* pullback; the parent inverse metric converts the result to an upstream vector. The statistical meaning of these metrics follows from a centred local reference model with fixed J. Vectorise states, stack children and Jacobians when needed, and denote centred deviations by tildes. Assume $\widetilde{H}_u = J\widetilde{H}_v + \eta$ with $\mathbb{E}[\eta\widetilde{H}_v^*] = 0$, and let Σ_v, Σ_u be the parent and child covariances. For nonsingular Σ_u, minimising $\mathbb{E}\|A\widetilde{H}_u - \widetilde{H}_v\|_2^2$ yields the linear minimum-MSE (LMMSE) map A_{LMMSE}. When both covariances are positive definite, their precision metrics identify this local Wiener map with the natural adjoint:

$$M_v = \Sigma_v^{-1}, \quad M_u = \Sigma_u^{-1} \quad \Longrightarrow \quad J^\sharp = \Sigma_v J^*\Sigma_u^{-1} = A_{\mathrm{LMMSE}}. \tag{4}$$

For jointly Gaussian centred deviations, its output also equals the conditional mean: $A_{\mathrm{LMMSE}}\widetilde{H}_u = \mathbb{E}[\widetilde{H}_v \mid \widetilde{H}_u]$. In nonlinear networks, Jacobians and metrics vary with the source point and boundary, making explicit online metric estimation impractical. We instead compute a source-local Euclidean VJP observable and calibrate its relation to upstream states. The centred reference explains this transition; deployed regressions use uncentred states and zero intercepts. Conditional-mean and adjoint/Wiener derivations, including singular cases, are in sections B.1 and B.2.

2.2 VJP Observable and First Wiener Repair

For a child with C_u physical channels, the channel-averaged energy defines the mean seed and its source-local VJP observable:

$$s_u = \nabla_{H_u}\frac{\|H_u\|_F^2}{2C_u} = \frac{H_u}{C_u}, \qquad R_v(x) = \sum\nolimits_{u\in\mathcal{C}_v} J_{u\leftarrow v}(x)^* s_u(x). \tag{5}$$

The divisor applies once per child, without averaging coordinates or children. R_v retains the source Jacobian but uses a Euclidean seed without explicit covariance metrics. Calibration on (R_v, H_v) thus targets $\mathbb{E}[H_v \mid R_v]$, which we approximate in a tractable linear class. Query transfer and architecture/component studies test its structural predictability (sections 3.1, 3.2 and C.4). Write tensors as $B \times C \times d_1 \times \cdots \times d_k$: batch supplies calibration samples, channels are mixed, and the k coordinate axes are Fourier transformed by $\mathcal{F}_k$ ($\mathcal{F}_0 = I$ when no coordinate axes exist). Let Ω_v^{st} be the stored non-redundant rFFT bins and Π_v a partition into nonempty groups, with $g(\omega)$ containing ω. Both repairs use the zero-intercept operator class

$$W_v(X) = \mathcal{F}_k^{-1}\left[W_{v,g(\omega)}(\mathcal{F}_k X)(\omega)\right], \qquad W_{v,g} \in \mathbb{C}^{C\times C}. \tag{6}$$

Full-channel matrices mix channels, not frequencies. Singletons fit one matrix per frequency; all-shared groups tie one across stored bins, retaining observable frequency content. This tractable calibration requires no network translation equivariance. Vision uses coarse/stage boundaries and singleton groups; GPT-2 uses custom boundaries and all-shared complex gains for both repairs.

Expectations average over the calibration set $\mathcal{D}_{\mathrm{cal}}$ using evaluation-mode models without cross-example coupling. For any calibration pair (X_v, Y_v), define $X_v' = \mathcal{F}_k X_v$, $Y_v' = \mathcal{F}_k Y_v$ and uncentred moments $S_v^{YX}(\omega) = \mathbb{E}[Y_v'(\omega)X_v'(\omega)^*]$, $S_v^{XX}(\omega) = \mathbb{E}[X_v'(\omega)X_v'(\omega)^*]$. Write $\overline{S}_{v,g}^{AB} = |g|^{-1}\sum_{\omega\in g} S_v^{AB}(\omega)$. With $\lambda_W > 0$, the common matrix Wiener fit is

$$W_{v,g}^\star = \arg\min_W \left\{\frac{1}{|g|}\sum_{\omega\in g}\mathbb{E}_{\mathcal{D}_{\mathrm{cal}}}\|WX_v'(\omega) - Y_v'(\omega)\|_2^2 + \lambda_W\|W\|_F^2\right\} = \overline{S}_{v,g}^{YX}(\overline{S}_{v,g}^{XX} + \lambda_W I)^{-1}. \tag{7}$$

Stored rFFT bins have equal weight; Parseval multiplicities change the objective. Positive ridge yields a unique second-moment solution without Gaussianity or gradient-based training (Wiener, 1949). The first stage sets $(W, X, Y, \lambda_W) = (G, R, H, \lambda_G)$ in eq. (7), giving $G_{v,g} = \overline{S}^{HR}_{v,g}(\overline{S}^{RR}_{v,g} + \lambda_G I)^{-1}$. G_v fits the full Euclidean-observable-to-state relation from calibration statistics to compensate systematic R_v–H_v differences and amortise missing geometric correction in the selected linear class. Each R_v retains its input's source Jacobian. For a fixed boundary DAG, fit each node independently in the two-stage order below. Fourier, bin-weighting and relative-ridge details: sections A.3 to A.5.

2.3 Source-Point Residual and Second Wiener Repair

A shared G_v can capture the dominant correction for fixed or weakly sample-dependent local operators; normalisation, attention and data-dependent gating or routing can leave stronger operating-point-specific mismatch. Fix G_v and form $y^0_v = G_v(R_v)$, $\Delta_v = H_v - y^0_v$. JVP forward-consistency correction (JVP-FC) observes this mismatch by applying original source differential to tangent y^0_v:

$$e_{u|v} = H_u - J_{u\leftarrow v} y^0_v = b_{u|v} + J_{u\leftarrow v}\Delta_v, \qquad b_{u|v} := H_u - J_{u\leftarrow v} H_v. \tag{8}$$

$J_{u\leftarrow v}\Delta_v$ is the child response to first-stage state error at the source point. The base term $b_{u|v}$ captures offsets and non-homogeneity, allowing residuals even at $y^0_v = H_v$. The JVP is a linear tangent response, not a nonlinear re-forward with y^0_v. Pulling $e_{u|v}$ back to parent coordinates gives

$$C_v = \sum\nolimits_{u\in\mathcal{C}_v} J^*_{u\leftarrow v} e_{u|v} = -\nabla\Phi_v(y^0_v), \qquad \Phi_v(y) = \frac{1}{2}\sum\nolimits_{u\in\mathcal{C}_v} \|H_u - J_{u\leftarrow v} y\|^2_F. \tag{9}$$

Thus C_v is a local child-consistency descent cotangent without further channel averaging in calibration or deployment. Regressing parent-state error Δ_v on C_v via eq. (7) with $(W, X, Y, \lambda_W) = (D, C, \Delta, \lambda_D)$, the same frequency partition, and independent positive ridge gives

$$D_{v,g} = \overline{S}^{\Delta C}_{v,g}(\overline{S}^{CC}_{v,g} + \lambda_D I)^{-1}, \qquad y_v = y^0_v + D_v(C_v). \tag{10}$$

Here $C'_v = \mathcal{F}_k C_v$ and $\Delta'_v = \mathcal{F}_k \Delta_v$ supply the second stage's moments and relative ridge scale. D_v estimates state error predictable from C_v within the chosen class; both maps remain fixed online. Calibration algorithms, moments and residual analysis are in sections A.6, A.7 and B.5.

2.4 Finite-DAG Composition and Scope

Deployment replaces calibration's true children with inverted states. For nonempty channel and coordinate sets (S, Q), shape-preserving projector $P_{S,Q}$ zeros unselected entries. With $C_S = |S|$, set

$$y_T = P_{S,Q} H_T, \qquad s^{\mathrm{on}}_T = \frac{y_T}{C_S} = \nabla_{H_T} \frac{\|P_{S,Q} H_T\|^2_F}{2C_S}. \tag{11}$$

No coordinate-count divisor is used. Full-target inversion selects all channels and coordinates, so $P_{S,Q} = I$ and $C_S = C_T$. Visit each remaining boundary once in reverse topological order, after its children are ready. Nonterminal children use their physical channel counts, $s^{\mathrm{on}}_u = y_u / C_u$ for $u \neq T$; terminal children retain the selected-channel normalisation above. At the original source point, apply

$$R^{\mathrm{on}}_v = \sum\nolimits_{u\in\mathcal{C}_v} J^*_{u\leftarrow v} s^{\mathrm{on}}_u, \qquad y^0_v = G_v(R^{\mathrm{on}}_v), \tag{12}$$

$$C^{\mathrm{on}}_v = \sum\nolimits_{u\in\mathcal{C}_v} J^*_{u\leftarrow v}(y_u - J_{u\leftarrow v} y^0_v), \qquad y_v = y^0_v + D_v(C^{\mathrm{on}}_v). \tag{13}$$

Residuals use repaired child states y_u, not normalised seeds. Native autograd handles branches, concatenations and shared tensors; input-boundary y_v is the reported inverse. Changing (S, Q) changes terminal initialisation and subsequent states while reusing G_v, D_v. Projector details, the algorithm and topology rules are in sections A.2, A.8 and A.9. At fixed source point, query and calibrated maps, VJP, JVP, both repairs and residual additions are linear in child states. Hence the complete update is $y_v = \mathcal{K}_v y_{\mathcal{C}_v}$. For full-target inversion, define $\delta_v = y_v - H_v$ and the true-child local defect $\epsilon_v = \mathcal{K}_v H_{\mathcal{C}_v} - H_v$. Substituting $y_{\mathcal{C}_v} = H_{\mathcal{C}_v} + \delta_{\mathcal{C}_v}$ and subtracting H_v gives

$$\delta_v = \epsilon_v + \mathcal{K}_v \delta_{\mathcal{C}_v}, \qquad \delta_T = 0. \tag{14}$$

This separates local approximation from propagated child error. Zero local defects yield exact full-target inversion; nonzero defects propagate along finite DAG paths. Section B.6 gives the path

bound, sufficient uniform-gain stability condition and reference-family formulation for general subset queries, all for one finite reverse pass. Zero intercepts also imply zero output for a zero terminal state, excluding a fixed target-independent additive template (section B.7). Frequency-domain repair can spread coordinate energy, so raw VJP dependency support need not survive repair. The construction accepts full, coarse/stage or custom boundaries with declared tensor axes and matrix-free VJP/JVP support. With fixed weights, maps serve new inputs, compatible depths and tested channel–coordinate queries without refitting; each input defines its forward point and derivatives. Weight changes require recalibration (section C.13). For boundary choices, architecture instances, numerical settings, calibration sizes and costs, see sections A.10 to A.12, A.14 and A.15.

3 EXPERIMENTS

We tested query reuse with fixed calibrated maps, local composition and repair, how target content and source geometry shape inverses, and visualised feature groups' decision effects via independent interventions, followed by a GPT-2 instance beyond vision. Unless marked as a control, results use Final JVP-FC (section 2): zero-intercept full-channel G_v, D_v and repaired online child states. Default visual experiments calibrated models separately on 4,096 ImageNet training images and evaluated them on 1,024 disjoint validation images (Deng et al., 2009; Russakovsky et al., 2015); cN/eM denotes calibration/evaluation sizes, with study-specific exceptions stated. Pixel cosine measures alignment with the source input. Defaults, metric definitions and reproducibility records: section C.1.

3.1 ONE CALIBRATION, MANY SOURCE-GROUNDED QUERIES

One ResNet50 c4096 calibration served 1,024 images, four full-state target depths, four individual terminal channels, and an eight-channel terminal subset with fixed maps, boundary DAG, and ridge. Across depths, full-state pixel cosine ranged from 0.928 to 0.945; channel queries produced distinct inverses (fig. 1c). These queries varied T or S with all coordinates retained. We then fixed all 768 ViT-B/32 channels and varied Q: 49 non-CLS patches at layers 2, 8, and 14 on 128 fixed, correctly classified ImageNet images, yielding 18,816 inverses from one c4096 map family (fig. 2b,c). Input-domain support sums the inverse's per-pixel channel ℓ_2 magnitudes by input patch, then normalises across patches. Self-patch support is mass at the query coordinate's input patch; profile entropy measures dispersion and support radius spatial extent. Across layers 2/8/14, self support fell (0.417/0.139/0.065), entropy rose (0.649/0.915/0.964), and radius increased. Deeper patch queries thus expressed more distributed input-domain structure, not query-coordinate crops. Independent patch interventions: section 3.4; query atlases, support definitions and intervals: sections C.2 and C.3.

3.2 LOCAL DAG COMPOSITION AND TWO-STAGE REPAIR

ResNet18 controls held target and calibration/evaluation data fixed, replacing one terminal map with local DAG composition. Global repair remained far below local repair; a second global stage slightly reduced alignment (table 1a). A second fitted stage thus did not replace local composition.

With local topology fixed, we measured child NMSE, the normalised squared difference between a local JVP prediction and its child state; intermediate-state relative ℓ_2, comparing the estimated upstream state with its forward reference; and pixel cosine. Final improved all three aggregate readouts on ResNet18, ConvNeXt-B, and ViT-B/32 (table 1b), connecting residual repair to state estimation and the final output. Boundary-level changes were non-monotonic (section C.7). ResNet50 already had G_v-only/Final pixel cosines of 0.889/0.931. Absolute second-stage pixel-cosine gains were larger for ConvNeXt-B, Swin-T and ViT-B/32: 0.111, 0.325 and 0.391, respectively (table 7; fig. 2a). These gains show complementary repair roles: G_v captures dominant distributional correction, while local operators such as LayerNorm and attention vary with the source point. Stage two estimates state error remaining after G_v from consistency residual under actual source differential.

Separate capacity controls tested full-channel coupling and frequency specificity with the visual two-stage workflow and c4096/e1024 budget fixed. On ViT-B/32, per-frequency diagonal maps gave pixel cosine 0.062, shared-frequency full-channel maps 0.337, and per-frequency full-channel maps 0.702. Both factors aid input-instance alignment at this budget. Diagonal-plus-rank-8 maps recovered much of ResNet18's alignment (0.834) but little of ViT-B/32's (0.050), so compression effects depend on architecture. Full controls and diagnostics: sections C.7 and C.8.

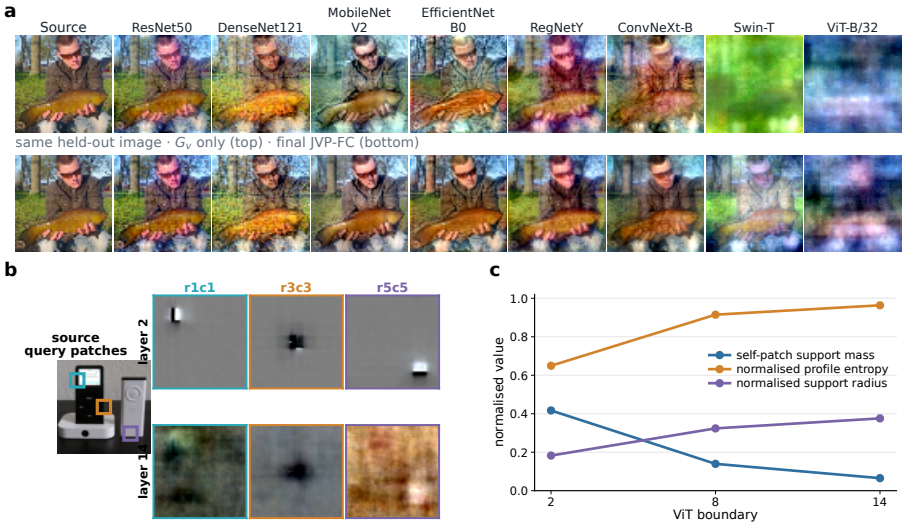


Figure 2: **Two-stage repair across architectures and coordinate-query reuse. a**, G_v-only (top) and Final (bottom) inverses of one held-out image across eight backbone families; models calibrated separately with the same recipe. **b**, Three fixed ViT patch queries (layers 2 and 14, first evaluation image); boxes mark coordinates. **c**, Mean self-patch support, normalised entropy and support radius (128 images; 95% image-bootstrap intervals). Panels b,c reuse one ViT map family across coordinates and depths. Architecture results, coordinate panels and selection protocols: sections C.3 and C.4.

Table 1: **Local composition and second-stage repair. a**, Pixel cosine for matched ResNet18 topology controls. **b**, With local topology fixed, c4096/e1024 stage controls measure output, child, and intermediate-state changes; decreases compare Final with G_v only.

a Topology (ResNet18)

Method	Pixel cosine
Raw VJP	0.0002
Global G	0.0722
Global $G + D$	0.0671
Local G_v	0.9009
Local Final	0.9380

b Stage controls (local DAG)

Model	Pixel cosine $G_v \rightarrow$ Final	Child NMSE decrease	State-error decrease
ResNet18	$0.9009 \rightarrow 0.9375$	23.0%	1.52%
ConvNeXt-B	$0.7343 \rightarrow 0.8450$	15.4%	0.45%
ViT-B/32	$0.3116 \rightarrow 0.7021$	16.2%	5.92%

Final improved pixel alignment over G_v alone in all nine separately calibrated ImageNet models spanning eight families. The same public path passed all 12 formal component runs across 11 tensor-component families; full architecture results and boundary-capture audits are in section C.4. Output structure differed: Swin-T had lower pixel cosine than ViT-B/32 (0.504 vs. 0.702), yet higher high-frequency luminance cosine (0.614 vs. 0.248) and SSIM (0.519 vs. 0.314); ViT retained stronger low-frequency alignment (0.789 vs. 0.484). The shared recipe thus exposes spatial and frequency profiles beyond a pixel-alignment ranking, consistent with the image biases of global patch-token mixing and hierarchical windowed computation (Dosovitskiy et al., 2021; Liu et al., 2021).

3.3 What Determines the Inverse?

We crossed terminal states H_A, H_B with source-local operator families J_A, J_B to separate content and geometry effects. On ResNet18, H_A inverted with J_B had pixel cosine 0.939 to source B and 0.034 to A. In the same c4096/e512 information-control study, terminal-state shuffling at fixed source geometry changed ResNet18 alignment from 0.9380 to 0.9379, but ViT-B/32's from 0.6922 to 0.0995. Source geometry thus strongly selected the branch, with architecture-dependent terminal modulation. Every zero target gave exactly zero output, excluding a fixed target-independent additive template; channel changes gave distinct inverses. Swap, interpolation, noise and query controls: section C.9.

Under dataset-specific checkpoints and calibration, the same recipe gave ResNet50 pixel cosines of 0.931/0.957/0.940 on ImageNet/Pets/CUB, while ConvNeXt-B gave 0.845/0.288/0.651 (section C.5).

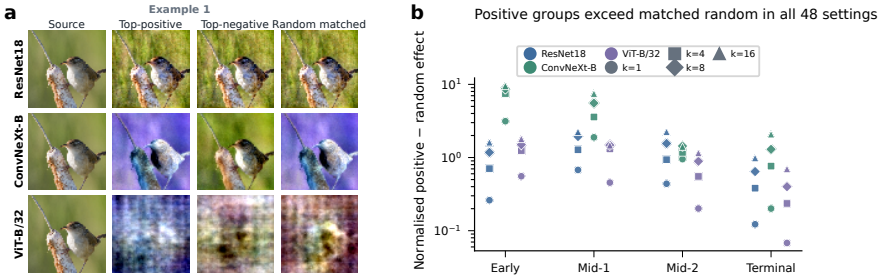


Figure 3: **Paired structural and decision readouts of the same feature groups.** **a**, The first predeclared CUB example per model, with positive, negative, and random-group inverses. **b**, Normalised positive-minus-random intervention effects for all 48 model–depth–group-size settings under calibration-mean replacement. Inverses display groups; independent interventions on the original representations measure their effects. The atlas, depth-wise deletion AUC, and controls are in fig. 18.

Interpreting an inverse therefore requires its model and distribution context. To understand low alignment, we tested reconstructed states' source-local path consistency. In the lowest pixel-cosine quartile of a CUB-trained and CUB-calibrated ConvNeXt-B (128 of 512 held-out CUB images, mean cosine -0.135), we fixed reconstructed parent and child states and compared JVP consistency under matched versus source-shuffled Jacobians. Path NMSE was 0.433 versus 0.719. Low pixel alignment thus coexisted with stronger matched-path consistency; local reversals and distribution controls are in sections C.5 and C.6. Estimation error also shapes the output: using the same maps with true rather than online child states increased ViT-B/32 pixel cosine from 0.7005 to 0.9410 in the nondeployable c4096/e512 diagnostic (section C.11). These results support a target-modulated source-local estimate whose visible structure reflects both network geometry and the deployed estimator's approximation.

3.4 Independent Interventions Connect Feature Groups to Decisions

CUB-trained ResNet18, ConvNeXt-B and ViT-B/32 each used 16 fixed, correctly classified test images. We ranked channels by activation's inner product with the original top-1–top-2 logit-margin gradient (Ancona et al., 2018). At four depths and group sizes $k \in \{1, 4, 8, 16\}$, each model's fixed c4096 maps inverted top-positive, top-negative and size-matched random groups; independent interventions set them to calibration means in the original representation. A separate forward pass recomputed that margin; margin drop is its pre- minus post-intervention value. Positive-group drops exceeded the within-image random mean in all 48 model–depth–k settings (all paired image-bootstrap intervals above zero). All 12 terminal negative-group controls gave oppositely signed (negative) drops (fig. 3). In ResNet18 and ConvNeXt-B depth atlases, both groups share inverse structure despite opposite intervention signs. Visible content thus complements signed decision role. Expanded atlases, deletion AUC, zero-replacement robustness and all control groups: section C.12.

The ViT study adds a complementary patch-level test on its 128 ImageNet images. Gradient-times-activation selected score-top and score-counter patches against a paired random patch; intervention zeroed the entire selected patch-token vector and recomputed the original class-pair margin. Layer-8 top-minus-random was 0.373 (95% image-bootstrap interval 0.217–0.544) and counter-minus-random -0.268 (-0.388 to -0.162); layer 11 retained both signs. Layer-14 patch effects were numerically zero under CLS-only readout (section C.3). Input-domain inverses display groups and are not re-input.

3.5 A GPT-2 Instance Beyond Vision

We applied this construction to GPT-2 small (Radford et al., 2019), querying all 768 channels of the layer-6 attention output before residual addition at the terminal position of 16-token contexts. The full reverse path yields a continuous input-embedding inverse without a text decoder; inverse energy is each embedding's squared norm normalised over all positions. Two independent all-shared spectral map families were calibrated on 4,096 complete WikiText-2 windows each (Merity et al., 2016); terminal masks were applied only at query time (section A.12). Layer 6 maximised cue-minus-random in an exploratory six-layer sweep. We averaged both calibration outputs before word/template-cluster bootstrapping; intervals condition on this selection (section C.14).

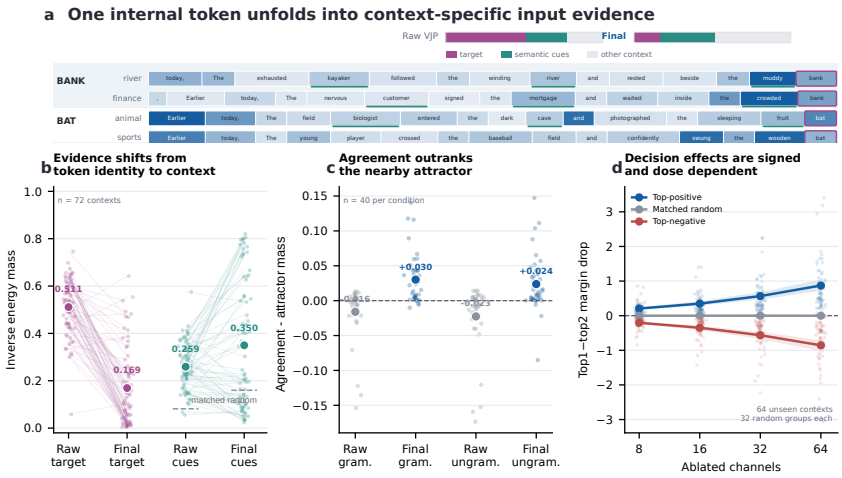


Figure 4: **Controlled continuous inversion in GPT-2. a**, BANK/BAT examples: shading shows Final inverse energy; teal underlines mark cues, magenta outlines terminal queries. Header: Raw VJP/Final means (72 contexts). **b**, Terminal/cue mass (72 contexts). **c**, Agreement-minus-attractor energy (40 grammatical, 40 ungrammatical contexts). **d**, Independent channel-group interventions (64 unseen contexts; 95% context-bootstrap intervals). Selection rules and full results: section C.14.

Semantic and syntactic profiles. Across 72 polysemy contexts, Final reduced terminal mass and increased predefined-cue mass relative to Raw VJP (fig. 4a,b). Against random subsets matched for cue count and position, Final cue-minus-random was 0.190 (95% word-cluster interval 0.096–0.297); its paired gain over Raw VJP remained unresolved. The 80 syntax contexts crossed subject-head number and auxiliary grammaticality. Agreement-minus-attractor subtracts energy at the nearer noun of opposite number to the head from energy at the number-marking auxiliary *was*/*were*; inversion queries the terminal predicate. Raw VJP's difference was negative and Final's positive in both grammaticality conditions (fig. 4c). Paired comparisons placed Final above Raw VJP but below G_v (section C.14), showing that the stages changed different profile aspects. A target-only global linear embedding map and attention showed little cue selectivity; gradient-times-input and integrated gradients showed positive cue contrasts but favoured the attractor. These references did not reproduce both patterns; attention and gradient profiles serve as interpretability references (table 17). **Source dependence and decision use.** Terminal state A inverted on geometry B was closer to the matched B inverse than the matched A inverse. On 64 unseen WikiText contexts, we ranked terminal channels by activation times the original top-1–top-2 margin gradient. Fixed maps inverted the same top-positive, top-negative, and size-matched random groups that were independently zeroed in the original terminal attention output; a new forward pass recomputed the original class-pair margin. At $k = 32$, margin drops were 0.567 for positive groups, -0.564 for negative groups, and approximately zero for random controls (fig. 4d). A separate joint full-coordinate control reused the maps, changing inverse profiles and signed intervention effects; this extension and calibration reproducibility are detailed in section C.14. Isolated nonterminal queries remain outside the validated scope: controls showed substantial future-position inverse energy under the deployed spectral surrogate.

4 Conclusion

We recast preimage search as source-grounded upstream-state estimation on the target-generating DAG. Two local closed-form matrix Wiener repairs estimate states, then remaining state error. One finite reverse pass reuses calibrated maps across inputs, depths and tested channel–coordinate queries without per-query optimisation. CNN and vision Transformer inverses reveal architecture-dependent spatial and frequency structure. Controlled GPT-2 small terminal-token queries yield token-resolved continuous input-embedding inverses without a text decoder. Target–operator controls confirm joint dependence on selected features and source-local geometry. Inversion displays input-domain feature structure; independent interventions on corresponding original representations test decision effects, together providing a reusable, modality-native interpretation interface (section D).

AI USE STATEMENT

We developed the core research ideas and methodological framework and provided the initial proof outlines and manuscript draft. OpenAI GPT and Google Gemini supported specific tasks under our direction. Following our specifications, they helped write code implementing our methods, prepare validation and experiment scripts, and refine experimental plans. Within the proof outlines we supplied, they assisted with completing derivation steps and checking for potential gaps. We manually checked the equations and proofs, reviewed and ran the code, and verified the correspondence between experimental outputs and reported figures and tables.

For literature discovery, these tools suggested relevant papers and provided summaries, which we checked against the original publications. We verified the support for cited claims and obtained BibTeX entries through Google Scholar. For writing, they helped review and refine our existing draft, including its language, organisation, and allocation of content between the main text and appendices. We made the final scientific and editorial decisions and take responsibility for the manuscript, code, and reported results.

ETHICS STATEMENT

We used existing datasets and pretrained models in accordance with their applicable licences and terms of use, alongside controlled language test contexts constructed for this study. No human participants were recruited, and no new private or sensitive personal data were collected. The method is intended to support the interpretation of neural-network representations by estimating upstream states conditioned on a supplied source input.

REPRODUCIBILITY STATEMENT

Appendix A specifies the calibration and online inversion algorithms, implementation conventions, and numerical settings. Appendix B provides the theoretical assumptions and proofs. Appendix section C.1 reports the vision experimental protocol, data splits, random seeds, metrics, and software and hardware settings; Appendix section C.14 details the language benchmarks and statistical procedures. We will provide an anonymised version of the code with the submission to support reproduction of the reported results.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3319–3327, 2017.

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4829–4837, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2668–2677, 2018.

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*, pp. 10012–10022, 2021.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5188–5196, 2015.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. URL https://arxiv.org/abs/1609.07843.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12448–12460. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.765. URL https://aclanthology.org/2023.emnlp-main.765/.

Charlie Nash, Nate Kushman, and Christopher K. I. Williams. Inverting supervised representations with autoregressive neural density models. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1620–1629, 2019.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 377–390. ACM, 2020. doi: 10.1145/3372297.3417270. URL https://doi.org/10.1145/3372297.3417270.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Norbert Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. MIT Press, 1949.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014.

A SUPPLEMENTARY METHOD DETAILS

This appendix expands the estimator defined in section 2 and fixes the axis, seed, stored-frequency tying, topology, and closed-form solve conventions needed for reproducibility. It proceeds from notation and mean-seed construction to the two calibration solves, the online recursion, topology and boundary instantiation, the reported GPT-2 instance, restricted operator families, complexity, and the final reproducibility checklist.

A.1 NOTATION AND AXIS CONVENTION

For each selected tensor node v in the computational DAG, we use the following objects:

- $H_v(x)$: the forward activation at node v.
- $\mathcal{C}_v$: the non-overlapping selected child frontier of v on paths to the target.
- $R_v(x)$: the local mean-seed VJP source in eq. (5).
- G_v: the first-stage local matrix Wiener repair.
- y_v^0: the first-stage state estimate $G_v R_v$.
- $C_v(x)$: the pulled-back JVP residual source.
- D_v: the JVP forward-consistency correction map.
- $S_v^{HR}(\omega)$ and $S_v^{RR}(\omega)$: first-stage cross- and auto-spectral matrices.
- $S_v^{\Delta C}(\omega)$ and $S_v^{CC}(\omega)$: correction cross- and auto-spectral matrices.
- Ω_v^{st}: the stored non-redundant rFFT-bin set at node v.
- Π_v: a declared partition of Ω_v^{st} used for spectral tying.
- $P_{S,Q}$: a shape-preserving terminal channel-coordinate mask projector.

The implementation uses a deepest non-overlapping fitting frontier for $\mathcal{C}_v$. If two candidate selected children u_1 and u_2 satisfy $u_1 \rightsquigarrow u_2$ in the selected boundary DAG, then u_1 is removed from the fitting frontier. This prevents the same downstream computation from contributing through both a branch-internal child and its later merge or output child. The unpruned structural edges are retained for ancestor coverage and boundary placement; only the local source frontier is pruned.

For an edge $v \to u$ in this fitting frontier, the local Jacobian is the partial derivative of the intervening subgraph with respect to H_v at the actual forward point. External inputs to that subgraph that are not descendants of v are held fixed at their forward values. All directed paths from v to u remain active, so residual branches and multi-use tensors contribute through native autograd semantics.

The scalar C_u denotes the channel count at node u. The calligraphic set $\mathcal{C}_v$ denotes child boundaries, whereas C_v denotes the JVP-FC residual source.

The batch dimension is always dimension zero. All reported models are frozen in evaluation mode and do not couple examples across this dimension, so each calibration example contributes one channel vector per stored frequency to the empirical statistics. The channel axis is the feature axis over which the full matrix Wiener map is applied. For CNN features this is usually dimension one, while for token streams such as (B, N, D) it is the hidden dimension D. The implementation may infer this axis from tensor shape and module type, and the user may override it explicitly. All remaining non-batch, non-channel axes are coordinate axes. If there are no coordinate axes, then $k = 0$ and $\mathcal{F}_0$ is the identity transform. For global-pooled nodes, singleton coordinate axes that only record the pooled spatial extent are trivial and may be collapsed before applying this rule. Thus $(B, C, 1, 1)$ is algebraically equivalent to the coordinate-free vector (B, C). For notation, tensors may be written as $(B, C, d_1, \ldots, d_k)$ after this axis identification; this does not require the original model tensor to store the channel axis in dimension one.

This convention is intentionally axis-based rather than module-name-based. The same estimator applies to convolutional features, token features, vector nodes, and attention tensors once architecture-aware boundary and axis metadata are specified. After the coordinate transform, each node also declares a stored-frequency partition Π_v. Frequencies in the same group share one complex full-channel gain matrix. This parameter tying remains frequency diagonal and does not imply that the underlying network is translation equivariant or that a causal support constraint has been imposed.

Table 2: One tensor rule for different node types.

Node type	Tensor shape	Channel axis	Coordinate transform
CNN feature	(B, C, H, W)	C	2D Fourier over (H, W)
Token stream	(B, N, D)	D	1D Fourier over N
Vector / GAP / MLP	(B, D)	D	None; one matrix
Attention tensor	(B, h, N, N)	h	2D Fourier over query-key axes

A.2 Mean-Seed Local Source

For a child boundary u with channel count C_u, the seed is

$$s_u = \frac{H_u}{C_u}.$$

This is the differential of the channel-averaged child energy

$$\Psi_u(H_u) = \frac{1}{2C_u}\|H_u\|_F^2.$$

For a single child, the local source is $J^*_{u\leftarrow v} s_u$. For multiple selected children, the source is the native multi-output VJP

$$R_v = \sum_{u\in\mathcal{C}_v} J^*_{u\leftarrow v}\frac{H_u}{C_u}.$$

The division by C_u is applied once, at the child seed. Coordinate positions are not divided out, and there is no second division by the number of children. This convention preserves the standard autograd gauge: multi-output backpropagation sums cotangents contributed by each output.

At online inversion time, let s_u^{on} denote the seed attached to child state y_u. For a nonterminal child, the same channel convention is used. For the terminal child, choose a nonempty channel set $S \subseteq \{1, \dots, C_T\}$ and a nonempty coordinate set Q. For a tensor with multiple coordinate axes, elements of Q are coordinate tuples; for a coordinate-free vector, Q contains the single trivial coordinate. Define channel and coordinate masks that preserve the original tensor shape and their orthogonal product projector

$$P_{S,Q} = P_S^{\mathrm{ch}} P_Q^{\mathrm{coord}}, \qquad P^*_{S,Q} = P_{S,Q} = P^2_{S,Q}. \tag{15}$$

Unselected entries are set to zero rather than removed from the tensor. With $C_S = |S|$, the terminal state and seed are

$$y_T = P_{S,Q} H_T, \qquad s_T^{\mathrm{on}} = \frac{P_{S,Q} H_T}{C_S},$$

which is the differential of

$$\Psi_T^{(S,Q)}(H_T) = \frac{1}{2|S|}\,\|P_{S,Q} H_T\|_F^2\,.$$

The coordinate-set cardinality is therefore not part of the seed divisor. For the complete online recursion,

$$s_u^{\mathrm{on}} = \begin{cases} y_T/C_S, & u = T, \\ y_u/C_u, & u \neq T. \end{cases} \qquad R_v^{\mathrm{on}} = \sum_{u\in\mathcal{C}_v} J^*_{u\leftarrow v} s_u^{\mathrm{on}}.$$

The full target has $P_{S,Q} = I$ and $C_S = C_T$. A single-channel, all-coordinate query has $C_S = 1$ and $Q = Q_{\mathrm{all}}$, where Q_{all} is the complete coordinate set. If Q is restricted separately, only the selected coordinates are retained. The reported GPT-2 terminal-token query retains all hidden channels and uses $Q = \{L-1\}$, so its divisor is the hidden width rather than the number of selected token positions. Thus channel normalisation is always part of the child seed, never part of the fitted Wiener map.

A.3 Full Matrix Wiener Derivation

For a tensor with C channels and k coordinate axes, the repair family applies a full channel matrix at each coordinate frequency:

$$G_v(X) = \mathcal{F}_k^{-1}[G_v(\omega)(\mathcal{F}_k X)(\omega)], \qquad G_v(\omega) \in \mathbb{C}^{C\times C}. \tag{16}$$

The transform acts only on coordinate axes, leaving batch and channel axes unchanged; with no coordinate axes, $\mathcal{F}_0 = I$ and the repair is one channel matrix. This zero-intercept family is frequency diagonal: it mixes channels within each frequency without mixing different frequencies. The correction map D_v uses the same application rule with its own matrices. For a fixed node v, let $H'_v = \mathcal{F}_k H_v$ and $R'_v = \mathcal{F}_k R_v$. We take $\mathcal{F}_k$ to be the full unitary transform in the derivation. Parseval's identity gives

$$\mathbb{E}\|G_v(R_v) - H_v\|^2 = \sum_{\omega} \mathbb{E}\left\|G_v(\omega)R'_v(\omega) - H'_v(\omega)\right\|^2. \tag{17}$$

Under the singleton partition, the unknown matrix at frequency ω appears only in the ω term, so the objective decomposes across coordinate frequencies.

For one frequency, write $R = R'_v(\omega) \in \mathbb{C}^C$, $H = H'_v(\omega) \in \mathbb{C}^C$, and $G = G_v(\omega)$, with all vectors treated as columns. The frequency-domain repair uses ordinary complex matrix multiplication GR; it is linear, not conjugate-linear, in R. The ridge objective is

$$J(G) = \mathbb{E}\|GR - H\|_2^2 + \lambda\|G\|_F^2. \tag{18}$$

The problem decomposes over output channels. Let w_i be the i-th row of G and H_i the i-th component of H. For that row,

$$J_i(w_i) = \mathbb{E}|w_i R - H_i|^2 + \lambda\|w_i\|_2^2.$$

The complex least-squares stationarity condition is

$$w_i\left(\mathbb{E}[RR^*] + \lambda I\right) = \mathbb{E}[H_i R^*].$$

It follows by treating w_i and its complex conjugate as independent variables and setting the derivative with respect to the conjugate row to zero. Stacking the C row equations yields

$$G\left(\mathbb{E}[RR^*] + \lambda I\right) = \mathbb{E}[HR^*].$$

With

$$S_v^{RR}(\omega) = \mathbb{E}[RR^*], \qquad S_v^{HR}(\omega) = \mathbb{E}[HR^*],$$

we obtain the singleton-group case of the closed form in eq. (21). Since $S_v^{RR}(\omega)$ is Hermitian positive semidefinite, $S_v^{RR}(\omega) + \lambda I$ is strictly positive definite for $\lambda > 0$, so the solution is unique. If $\lambda = 0$, uniqueness requires $S_v^{RR}(\omega)$ to be nonsingular. In finite calibration, expectations are replaced by empirical averages. This Parseval derivation directly establishes the singleton, per-frequency case. When parameters are tied across stored rFFT bins, the precise objective is the declared stored-spectrum objective below.

For real tensors, the Fourier spectrum is conjugate symmetric. Implementations may therefore use a real FFT and store only the non-redundant spectrum. This does not change the represented real operator family or the per-frequency normal equations, provided the forward and inverse transforms use a consistent convention and the inverse restores the corresponding real signal. The relative-ridge average below is defined over the stored spectrum, so results must report and retain that storage convention. The evaluated implementation uses orthonormal real transforms (`norm=ortho`). It converts channel-first tensors to float32 and stores spectral statistics and fitted gains in complex64. A nonunitary FFT normalisation rescales both stored second moments; the effective relative ridge then rescales by the same factor.

A.4 Stored-Frequency Partitions and Grouped Closed Forms

Let Ω_v^{st} be the non-redundant frequency-bin set stored by the real-transform implementation, and let Π_v partition this set into nonempty groups. Frequencies in group $g \in \Pi_v$ share one matrix $G_{v,g}$, so $G_v(\omega) = G_{v,g}$ for every $\omega \in g$. To state the constrained problem independently of a particular

storage weighting, assign a positive weight a_ω to every stored bin and define $a_g = \sum_{\omega \in g} a_\omega$. The group objective is

$$\mathcal{L}_{v,g}(A) = \frac{1}{a_g} \sum_{\omega \in g} a_\omega \mathbb{E} \left\| AR'_v(\omega) - H'_v(\omega) \right\|_2^2 + \lambda \|A\|_F^2. \tag{19}$$

Define the weighted pooled moments

$$\overline{S^{HR}}_{v,g} = \frac{1}{a_g} \sum_{\omega \in g} a_\omega S_v^{HR}(\omega), \qquad \overline{S^{RR}}_{v,g} = \frac{1}{a_g} \sum_{\omega \in g} a_\omega S_v^{RR}(\omega). \tag{20}$$

The row-wise stationarity argument above then gives the exact equality-constrained closed form

$$G^\star_{v,g} = \overline{S^{HR}}_{v,g} \left(\overline{S^{RR}}_{v,g} + \lambda I \right)^{-1}, \qquad G^\star_v(\omega) = G^\star_{v,g} \quad (\omega \in g). \tag{21}$$

The reported implementation uses $a_\omega = 1$ for every stored non-redundant rFFT bin. Its grouped solve is therefore exact for the equally weighted stored-spectrum objective in eq. (19) with $a_\omega = 1$. It does not claim equivalence to a coordinate-domain MSE in which full-spectrum Parseval multiplicities weight ordinary positive frequencies twice and DC and Nyquist frequencies once. Changing to those multiplicity weights would define a different estimator and is not used for any reported map.

The singleton partition $\Pi_v = \{\{\omega\} : \omega \in \Omega_v^{\mathrm{st}}\}$ recovers the per-frequency solve. The all-shared partition $\Pi_v = \{\Omega_v^{\mathrm{st}}\}$ fits one complex full-channel gain and broadcasts it to every stored bin under the declared rFFT/irFFT convention. Spectral tying removes bin-specific gain variation; it does not remove frequency content already present in the source-local VJP or JVP observables and does not impose causal support.

The same partition and weights are used for JVP-FC. With correction moments $S_v^{\Delta C}(\omega)$ and $S_v^{CC}(\omega)$, their weighted group averages replace $\overline{S^{HR}}_{v,g}$ and $\overline{S^{RR}}_{v,g}$ in eq. (21), yielding $D^\star_{v,g}$ in eq. (23).

A.5 Ridge and Zero-Intercept Solve

Both G_v and D_v are solved as zero-intercept maps. The zero-frequency coefficient is treated like every other coordinate frequency: it has a full channel matrix and no additive term. The ridge is implemented as a relative ridge with respect to the source auto-spectral second moment of the current solve:

$$\lambda_{\mathrm{eff}} = \rho \, \max\left\{ \mathrm{mean}_{\omega,c} S^{XX}_{v,cc}(\omega), \epsilon_{\mathrm{ridge}} \right\}, \qquad \rho = 0.01, \quad \epsilon_{\mathrm{ridge}} = 10^{-30}, \tag{22}$$

where $X = R$ for the first-stage map G_v and $X = C$ for the correction map D_v. This keeps the same nominal relative coefficient ρ meaningful across nodes with different source scales. In the closed-form solve, λ_{eff} is the value added to the diagonal of the current source second moment. Because $S_v^{XX}(\omega)$ is Hermitian positive semidefinite, its diagonal entries are real and nonnegative; implementations use the real diagonal when computing the relative ridge. The average is taken over the same stored frequency convention used by the solve, for example over the non-redundant spectrum when a real FFT implementation is used. The numerical floor guarantees a positive ridge for a degenerate zero source when $\rho > 0$; it has no practical effect at ordinary source scales. With N calibration examples, an empirical source second moment has rank at most $\min(C, N)$ at each singleton frequency. For a group g, the pooled moment contains $N|g|$ spectral coefficient vectors and has algebraic rank at most $\min(C, N|g|)$, although these coefficients arise from only N independent calibration inputs. A positive ridge guarantees an invertible solve even when the empirical moment is rank deficient.

A.6 Local Calibration Algorithm for G_v

The first-stage map G_v is calibrated from local source-target pairs (R_v, H_v). It does not use repaired downstream maps as part of its source construction. Given frozen model f, target T, selected boundary nodes $\mathcal{V}_b$, and calibration set $\mathcal{D}_{\mathrm{cal}}$:

1. For each selected node $v \neq T$ on a path to T:
   (a) find the deepest non-overlapping child frontier $\mathcal{C}_v$ on paths to T;

(b) initialise spectral accumulators $S_v^{HR}(\omega)$ and $S_v^{RR}(\omega)$;
(c) for each calibration batch:
  i. run the forward pass and cache H_v and $\{H_u : u \in \mathcal{C}_v\}$;
  ii. form child seeds H_u/C_u;
  iii. compute $R_v = \sum_{u\in\mathcal{C}_v} J^*_{u\leftarrow v}(H_u/C_u)$ by autograd;
  iv. transform H_v and R_v over coordinate axes;
  v. accumulate $H'_v(\omega)R'_v(\omega)^*$ and $R'_v(\omega)R'_v(\omega)^*$.
(d) average the per-bin accumulators over calibration inputs;
(e) pool the moments within each declared group $g \in \Pi_v$ using the stored-bin weights a_ω and solve eq. (21);
(f) broadcast $G_{v,g}$ to the stored bins in g for application under the declared inverse-transform convention.

Since each G_v uses only local child activations and local VJPs, the G_v fits are independent once the boundary DAG is fixed. The implementation may still traverse nodes in a particular order for memory management, but the estimator itself is local.

A.7 Calibration Algorithm for JVP-FC Correction

After all first-stage maps G_v are available, the correction map D_v is calibrated from residual pairs. Each first-stage source already contains the local Jacobian of its calibration sample, but the fitted G_v is shared across the calibration distribution. For fixed or weakly sample-dependent local operators, this first projection can already account for most of the local correction. JVP-FC is principally motivated by operators whose differential changes more strongly with the source operating point, including normalisation, attention, and data-dependent gating or routing. It remains a general residual regression that estimates first-stage error from its correction source within the chosen linear class. For each selected node $v \neq T$:

1. For each calibration batch, compute the local source R_v exactly as in the G_v calibration.
2. Apply the first-stage map $y_v^0 = G_v R_v$. In the following JVP, y_v^0 is supplied as a tangent vector at the cached forward point rather than as a replacement activation.
3. For every child $u \in \mathcal{C}_v$, compute the local JVP
$$\widehat{H}_{u|v} = J_{u\leftarrow v} y_v^0.$$
4. Form child residuals $e_{u|v} = H_u - \widehat{H}_{u|v}$. This residual contains both the zero-intercept base term $H_u - J_{u\leftarrow v}H_v$ and the pushed-forward first-stage error $J_{u\leftarrow v}(H_v - y_v^0)$; it is not assumed to vanish at $y_v^0 = H_v$. The base term vanishes for a positively one-homogeneous local map but is generally nonzero in the presence of offsets or non-homogeneous nonlinearities.
5. Pull the residuals back:
$$C_v = \sum_{u\in\mathcal{C}_v} J^*_{u\leftarrow v} e_{u|v}.$$
This is the negative gradient at y_v^0 of $\frac{1}{2}\sum_{u\in\mathcal{C}_v} \|H_u - J_{u\leftarrow v}y\|_F^2$. The source therefore uses the Euclidean child-state residual gauge and is not divided by C_u.
6. Accumulate $S_v^{\Delta C}(\omega)$ and $S_v^{CC}(\omega)$, pool them under the same partition Π_v and stored-bin weights as G_v, and solve $D_{v,g}$ by eq. (23).

Thus D_v uses the same axis rule and Fourier decomposition as G_v. For $\Delta_v = H_v - y_v^0$, let $C'_v = \mathcal{F}_k C_v$ and $\Delta'_v = \mathcal{F}_k \Delta_v$. The regression target is the parent-state error Δ_v, with the pulled-back residual C_v as its source; eq. (35) states the corresponding grouped ridge objective. The per-frequency correction moments are

$$S_v^{\Delta C}(\omega) = \mathbb{E}[\Delta'_v(\omega)C'_v(\omega)^*], \qquad S_v^{CC}(\omega) = \mathbb{E}[C'_v(\omega)C'_v(\omega)^*].$$

For the reported equal stored-bin weights, the pooled moments are

$$\overline{S}_{v,g}^{\Delta C} = \frac{1}{|g|}\sum_{\omega\in g} S_v^{\Delta C}(\omega), \qquad \overline{S}_{v,g}^{CC} = \frac{1}{|g|}\sum_{\omega\in g} S_v^{CC}(\omega).$$

Its stored group matrices are

$$D^{\star}_{v,g} = \overline{S}^{\Delta C}_{v,g} \left(\overline{S}^{CC}_{v,g} + \lambda_D I \right)^{-1}, \qquad D^{\star}_{v}(\omega) = D^{\star}_{v,g} \quad (\omega \in g). \tag{23}$$

The correction ridge λ_D is obtained from its own source moments S^{CC}_v using eq. (22). This is a second closed-form regression, not query-specific iterative preimage optimisation. The JVP and VJP are matrix-free applications of the frozen local subgraph differential at the current forward point.

A.8 Online Feature-Inversion Algorithm

At inversion time, the computation proceeds from the target toward the input. Choose nonempty terminal channel and coordinate sets (S, Q) and let $C_S = |S|$. Initialise the terminal state and seed as

$$y_T = P_{S,Q} H_T, \qquad s^{\mathrm{on}}_T = \frac{y_T}{C_S} = \frac{P_{S,Q} H_T}{C_S},$$

where $P_{S,Q}$ is the shape-preserving projector defined in eq. (15). The full target has $P_{S,Q} = I$ and $C_S = C_T$. A single-channel, all-coordinate query has $C_S = 1$ and $Q = Q_{\mathrm{all}}$, where Q_{all} is the complete coordinate set. If Q is restricted separately, only the selected coordinates are retained. Visit each remaining selected node v once in reverse topological order, so all of its child states have already been inverted to their boundaries. At each node:

1. attach the seed
$$s^{\mathrm{on}}_u = \begin{cases} y_T / C_S, & u = T, \\ y_u / C_u, & u \neq T, \end{cases}$$
to each child, and compute the online local source
$$R^{\mathrm{on}}_v = \sum_{u \in \mathcal{C}_v} J^{*}_{u \leftarrow v} s^{\mathrm{on}}_u;$$
2. apply the first-stage map $y^0_v = G_v R^{\mathrm{on}}_v$;
3. compute the online child residual source
$$C^{\mathrm{on}}_v = \sum_{u \in \mathcal{C}_v} J^{*}_{u \leftarrow v} \left(y_u - J_{u \leftarrow v} y^0_v \right);$$
4. return $y_v = y^0_v + D_v C^{\mathrm{on}}_v$.

The input-node state is the reported input-domain feature inverse. Only the terminal child uses C_S in its seed. Every inverted nonterminal child uses its physical channel count C_u. The residual source uses the state y_u itself, without seed normalisation. The calibrated G_v and D_v maps are unchanged when (S, Q) changes.

A.9 DAG Topology: Branches and Multi-Use Nodes

The method relies on autograd's VJP semantics for graph topology. No special mathematical rule is introduced for residual connections, concatenations, or branching.

Residual addition. For $z = a + b$, the VJP distributes the cotangent of z to both a and b. Each branch receives its own cotangent and is repaired at its own tensor nodes.

Concatenation. For $z = \mathrm{concat}(a, b)$, the VJP slices the cotangent of z back into the coordinates or channels corresponding to a and b. Repairs are then applied on the branch tensors.

One-to-many use. If a tensor feeds multiple downstream operations, autograd sums the cotangents contributed by each downstream user. The node repair acts on this accumulated incoming cotangent. If the capture granularity exposes separate edge tensors, the same local repair rule can be applied at those edge nodes instead.

A.10 BOUNDARY MODES

The theoretical maximal form is Full-DAG: every captured differentiable floating tensor node with a specified channel axis has a well-defined local matrix Wiener repair. Scalar nodes, non-floating tensors, and bookkeeping operations without a meaningful channel axis are traversed by autograd but are not instantiated as repair nodes. The implementation exposes three modes:

- **Full-DAG**: instantiate every captured differentiable floating tensor node with a specified channel axis.
- **Coarse/stage block**: instantiate the input, stem or patch embedding, and architecture block or stage edges; this is the default efficient cross-architecture mode.
- **Custom list**: instantiate user-specified nodes after full DAG capture.

Skipping nodes is an implementation choice, not a change in the local estimator. The selected nodes are repair boundaries, not a replacement for the full forward graph: every intervening computation remains inside the corresponding frozen autograd subgraph. The formulation requires matrix-free VJP and JVP support for these subgraphs. The current JVP implementation uses a VJP-of-probes identity and therefore requires double-backward support; fused attention kernels are replaced by a mathematically equivalent kernel with the required derivative support when necessary.

A.11 ARCHITECTURE-SPECIFIC INSTANCES

CNNs. Feature maps have shape (B, C, H, W). The channel axis is C and the coordinate axes are (H, W), giving a 2D Fourier transform and one $C \times C$ matrix per 2D frequency.

DenseNet and ConvNeXt. The same tensor rule applies. Coarse/stage block boundaries provide a compact cross-model instantiation, while Full-DAG can be used when memory permits.

ViT token streams. For tokens with shape (B, N, D), the hidden dimension D is the channel axis and token position N is the coordinate axis. The repair uses a 1D Fourier transform over tokens and a full $D \times D$ matrix per token frequency. When a class token is present, it remains at its stored sequence position and is included in this transform unless an experiment explicitly specifies a different token convention.

Causal language token streams. For a language-token tensor shaped (B, L, D), the hidden width D is the channel axis and token position L is the coordinate axis. Learned positional embeddings and causal attention remain part of the frozen forward subgraphs and their source-local VJP/JVP operators. A stored-frequency partition controls only how the repair gains are tied; it does not replace or certify those causal dependencies.

Attention tensors. For attention logits or weights shaped (B, h, N, N), the head dimension h is treated as channel and query-key axes as coordinates. The repair acts on cotangents or reverse-state estimates rather than a probability matrix that is reinserted into the forward model, so row-stochastic softmax constraints are not explicitly enforced. The estimator remains algebraically well defined, but both this relaxation and the query-key frequency-diagonal approximation remain empirical questions.

Vector, GAP, and MLP nodes. For (B, D) tensors, $k = 0$, so there is no Fourier transform and the repair is a single $D \times D$ matrix. Global-pooled tensors stored as $(B, D, 1, 1)$ are treated equivalently because the singleton-coordinate Fourier transform is the identity.

A.12 REPORTED GPT-2 SMALL INSTANTIATION

The reported language experiment instantiates the same estimator on pretrained GPT-2 small. A length-$L = 16$ sequence enters the inversion API as continuous input embeddings with shape $(B, 768, 16)$; inside GPT-2, token-stream boundaries have shape $(B, 16, 768)$. The target is the complete layer-6 attention-output sequence before residual addition, rather than the layer-6 residual stream. Both G_v and D_v are calibrated on these complete sequence tensors before any terminal-position mask is applied.

The selected repair boundaries are the input embeddings, the outputs of the second and fourth GPT-2 blocks, and the layer-6 attention output. These four boundaries define three fitted nonterminal nodes, each with one G_v map and one D_v map. They do not omit the intervening model computation: blocks and operations between consecutive repair boundaries remain inside the actual frozen autograd subgraphs used to evaluate their local VJPs and JVPs.

At every fitted boundary, the channel matrices are full and zero-intercept. The length-16 real transform exposes nine stored rFFT bins, and the all-shared partition ties one complex full-channel gain matrix across all nine bins for each map. The reported implementation gives every stored bin equal weight in the pooled objective. The relative ridge coefficient is $\rho = 0.01$ for both G_v and D_v; the diagonal shift applied at each solve is its source-moment-dependent λ_{eff}, not an absolute ridge of 0.01.

Each formal map family is fitted from 4,096 fixed-seed WikiText-2 windows, and the reproducibility analysis uses independently sampled calibration sets with seeds 123 and 321. At deployment, the reported token query uses $S = \{1, \dots, 768\}$ and $Q = \{L - 1\}$, yielding

$$y_T = P_{S,\{L-1\}} H_T, \qquad s_T^{\text{on}} = \frac{P_{S,\{L-1\}} H_T}{768}.$$

The same full-sequence maps are reused without per-query optimisation or refitting, and the reported inverse remains in the continuous input-embedding domain without a learned text decoder or autoregressive decoding. The general projector permits other coordinate sets algebraically. The formal semantic and syntactic claims use terminal-token queries; the supplementary joint full-coordinate channel-subset control does not validate isolated nonterminal-position queries.

A.13 Restricted Operator Families

The full-channel singleton per-frequency partition is the final vision configuration. The visual ablations impose additional restrictions in both G_v and D_v:

- **Diagonal-channel**: retain a separate map at each frequency but force every channel matrix to be diagonal.
- **Shared-frequency**: retain full channel matrices but tie one gain across stored coordinate-frequency bins. This removes frequency-specific gain variation from the repair maps; it does not remove frequency content already present in the source-local VJP and JVP observables.
- **Gamma-only**: force diagonal channel gains and share those gains across frequencies.
- **Diagonal plus low-rank**: use a diagonal channel map plus a low-rank channel correction at each frequency.

Diagonal-channel removes cross-channel second-order coupling. Shared-frequency is a capacity ablation relative to the vision singleton partition, while the same equality-constrained closed-form family is the declared GPT-2 instantiation on a different coordinate domain; it is not treated as a GPT-2 ablation. The matched quantitative and visual comparisons are reported in tables 11 and 12 and fig. 13. Those experiments evaluate not only global input-instance alignment but also the spatial and spectral structures retained by each restricted family.

A.14 Complexity and Memory

Let F be the number of stored coordinate frequencies. Let $K_v = |\Pi_v|$ be the number of distinct stored-frequency groups at node v. For a node with C channels, storing the unpooled full spectral statistics requires $O(FC^2)$ memory. Accumulating either source-target statistic costs $O(NFC^2)$ arithmetic for N calibration examples, excluding VJP/JVP and FFT costs. Pooling the stored moments costs $O(FC^2)$, and the dense grouped solves cost $O(K_v C^3)$ in the worst case.

Each fitted map contains K_v distinct full matrices, or $K_v C^2$ distinct complex gain coefficients. The two stages together contain $2K_v C^2$ distinct complex coefficients per fitted boundary. These are not real-parameter counts. The current implementation may broadcast each group gain to its member bins for application, in which case physical map storage remains FC^2 complex coefficients per map. Complex64 storage uses eight bytes per stored complex coefficient.

Applying one fitted spectral map costs $O(BFC^2)$ arithmetic, in addition to coordinate-transform cost $O(BCP \log P)$ for P coordinate positions. At each frequency, each calibration example contributes one C-dimensional Fourier coefficient. A singleton group therefore has N coefficient vectors. A group g pools $N|g|$ coefficient vectors but still contains only N independent input sequences; its empirical auto-spectrum has algebraic rank at most $\min(C, N|g|)$. Vector nodes have $F = 1$ and are therefore comparatively cheap. ViT token nodes have $C = D$ and F proportional to the token sequence frequency count, so calibration size and statistic accumulation can be substantial. In the reported GPT-2 instance, $F = 9$ and $K_v = 1$ at every fitted boundary.

Engineering controls include frequency chunking, solve chunking, CPU/GPU placement of maps, cached calibration tensors, and cotangent caching. These optimisations do not change the estimator as long as they preserve the local source-target pairs defined above. JVP-FC additionally requires an autodiff path with second-order derivative support because the matrix-free JVP is evaluated through a VJP-of-probes identity. For fused attention implementations, a mathematically equivalent kernel with double-backward support must be enabled when the default kernel lacks it.

A.15 Reproducibility Details

Reported defaults. Unless stated otherwise, all reported default configurations use mean-full calibration seeds: all child channels are retained and divided by their channel count. Both stages use full-channel zero-intercept maps, no activation or cotangent pre-normalisation, relative ridge coefficient $\rho = 0.01$, and unit JVP-FC residual and correction weights. The headline vision configuration uses coarse/stage boundaries and the singleton per-frequency partition; the reported GPT-2 configuration uses selected custom boundaries and all-shared frequency tying. Full-channel second-moment estimation makes calibration size a practical consideration at high-dimensional token nodes. The ViT token-stream runs use 4,096 calibration images, and the reported GPT-2 map families use 4,096 calibration windows. These are the reported experimental settings, rather than a universal optimal calibration size.

Each experiment should report:

- model architecture, frozen weights, and evaluation mode,
- data split for calibration and evaluation,
- sampling scheme and random seeds for both splits,
- input preprocessing, including tokenisation and sequence construction when applicable,
- calibration size and batch size,
- target node, nonempty channel and coordinate sets (S, Q), and the shape-preserving terminal mask definition,
- boundary mode and the exact selected repair boundaries,
- non-overlapping child-frontier construction,
- calibration seed $s_u = H_u/C_u$, nonterminal online seed $s_u^{\mathrm{on}} = y_u/C_u$, and terminal subset seed $s_T^{\mathrm{on}} = P_{S,Q}H_T/|S|$,
- channel-axis convention or override,
- source-normalisation setting and zero-intercept linear convention,
- channel solve mode, stored-frequency partition Π_v, and group weights a_ω,
- FFT normalisation, stored-spectrum convention, numerical precision, and whether group gains are broadcast for storage,
- first-stage and correction ridge values and whether relative ridge is used,
- nominal relative ridge coefficient ρ and effective ridge definition,
- JVP-FC child-target mode, residual scale, and correction weight,
- map device and chunking settings,
- autodiff or attention-kernel settings needed for JVP double backward,
- post-calibration online runtime, hardware, software version, and memory constraints.

These conventions define the deployed estimator used by the supplementary theory and by every headline experiment.

B Supplementary Theory: From the Statistical Ideal to DAG Stability

This appendix expands the derivation and finite-DAG results in section 2. As in the main text, the estimated objects are upstream tensor states along the source-local inverse branch, not generic reconstructions of the complete source input from a compressed code. It first proves the conditional-mean optimum, then connects local Wiener inversion to a covariance-induced natural adjoint. It next formalises the mean-seed observable and the two distributional Wiener projections, before deriving channel–coordinate-conditioned finite-DAG stability and the remaining operator boundaries.

B.1 MSE-Optimal Conditional-Mean Inversion

Let H_v and H_u be square-integrable random states on the data distribution, with H_v upstream of H_u. Define

$$m(H_u) = \mathbb{E}[H_v \mid H_u].$$

For any measurable estimator $\mathcal{T}(H_u)$, write

$$H_v - \mathcal{T}(H_u) = \big(H_v - m(H_u)\big) + \big(m(H_u) - \mathcal{T}(H_u)\big).$$

Conditional-expectation orthogonality gives

$$\mathbb{E}\left[\langle H_v - m(H_u), m(H_u) - \mathcal{T}(H_u)\rangle\right] = 0.$$

Therefore

$$\mathbb{E}\|H_v - \mathcal{T}(H_u)\|_2^2 = \mathbb{E}\|H_v - m(H_u)\|_2^2 + \mathbb{E}\|m(H_u) - \mathcal{T}(H_u)\|_2^2. \tag{24}$$

The second term is nonnegative, so the conditional mean is the MSE-optimal inverse, uniquely up to almost-sure equality.

This statistical ideal uses only the downstream observation. The deployed method is input-conditioned in a different sense: the input being explained is available to instantiate the frozen forward point and its local Jacobians. It is not included as an extra variable in the conditional mean, because for a deterministic frozen model the full input already determines every forward state. The method does not estimate the full nonlinear conditional law in eq. (24). It uses the forward point to construct a sequence of local observable inverse problems on the boundary DAG.

B.2 Local Wiener Inversion and the Natural Adjoint

For a local map $y = f(h)$ linearised at the current activation h_0, the differential is $\delta y = J(h_0)\delta h$ and backpropagation of a seed q gives $r = J(h_0)^\top q$. We use $\top$ for real-domain Jacobians and $*$ for conjugate transpose in complex or Fourier coordinates. The Euclidean adjoint is generally different from the pseudoinverse, $J^\top \neq J^\dagger$, except in partial-isometry cases where every nonzero singular value of J equals one. Raw VJPs therefore follow the computational dependencies but carry adjoint transport geometry rather than that of an upstream-state inverse.

Vectorise the local tensors and hold the Jacobian J fixed at one forward point for this unrestricted reference calculation. Write $\widetilde{H}_v$ and $\widetilde{H}_u$ for centred parent and child deviations, and assume

$$\widetilde{H}_u = J\widetilde{H}_v + \eta, \qquad \mathbb{E}[\eta\widetilde{H}_v^*] = 0. \tag{25}$$

Multiple child states can be stacked, with J formed by vertically stacking their local Jacobians. Define the covariances

$$\Sigma_v = \mathbb{E}[\widetilde{H}_v\widetilde{H}_v^*], \qquad \Sigma_\eta = \mathbb{E}[\eta\eta^*], \qquad \Sigma_u = \mathbb{E}[\widetilde{H}_u\widetilde{H}_u^*] = J\Sigma_v J^* + \Sigma_\eta.$$

For a linear estimator $A\widetilde{H}_u$, the normal equation for $\min_A \mathbb{E}\|A\widetilde{H}_u - \widetilde{H}_v\|_2^2$ is

$$A\Sigma_u = \mathbb{E}[\widetilde{H}_v\widetilde{H}_u^*] = \Sigma_v J^*.$$

If Σ_u is nonsingular, the unique best linear estimator is

$$A^\star = \Sigma_v J^* \Sigma_u^{-1}. \tag{26}$$

Under joint Gaussianity, the estimate satisfies $A^\star \widetilde{H}_u = \mathbb{E}[\widetilde{H}_v \mid \widetilde{H}_u]$, recovering the zero-mean conditional-mean inverse. Without Gaussianity, it remains the best zero-intercept linear estimator.

The same operator arises from weighted geometry when both covariances are positive definite. Use the precision metrics

$$M_v = \Sigma_v^{-1}, \qquad M_u = \Sigma_u^{-1},$$

and the inner products $\langle a, b \rangle_M = a^* M b$. The Riesz adjoint $J^\sharp$ is determined by

$$\langle J\delta_v, s_u \rangle_{M_u} = \langle \delta_v, J^\sharp s_u \rangle_{M_v}.$$

Expanding both inner products gives

$$\delta_v^* J^* M_u s_u = \delta_v^* M_v J^\sharp s_u.$$

Equality for all δ_v and s_u therefore implies $M_v J^\sharp = J^* M_u$, and hence

$$J^\sharp = M_v^{-1} J^* M_u = \Sigma_v J^* \Sigma_u^{-1} = A^\star. \tag{27}$$

The operators M_u and M_v are the child and parent metrics themselves. The factor M_v^{-1} appears when the adjoint is represented in parent coordinates: the child metric acts before the pullback and the parent inverse metric acts after it. Thus, after the metric-induced identification of vectors and covectors, local Wiener inversion and the covariance-induced natural adjoint are the same ideal operator under the centred linear model. If a covariance is singular, the LMMSE map can be written with a pseudoinverse on its support. A positive-definite metric on the full space instead requires regularisation, for example replacing Σ by $\Sigma + \tau I$ with $\tau > 0$. Centring is used only for this ideal identity. The deployed estimator retains the uncentred, zero-intercept source-target objectives defined in the Method and does not fit an additive bias.

This identity motivates the geometry that an inverse correction should supply. It is not an algebraic characterisation of the deployed G_v or D_v. In a nonlinear network, the local Jacobian and appropriate covariance geometry vary with the input and boundary. Explicit deployment would require estimating and applying both metrics online. The projected local-oracle experiment in section C.8 tests the Wiener–natural-adjoint identity under its controlled assumptions.

B.3 Variational Interpretation of the Mean-Seed Observable

For a child activation $H_u \in \mathbb{R}^{B \times C_u \times \cdots}$, define

$$\Psi_u(H_u) = \frac{1}{2C_u} \|H_u\|_F^2. \tag{28}$$

Its differential with respect to H_u is H_u / C_u. Therefore

$$J_{u \leftarrow v}^* \frac{H_u}{C_u}$$

is the exact adjoint pullback of a scalar child-state observable.

It is also the arithmetic mean of the channel-energy pullbacks. Writing $H_{u,c}$ for channel c and $\Psi_{u,c} = \|H_{u,c}\|_F^2 / 2$ gives

$$\frac{1}{C_u} \sum_{c=1}^{C_u} J_{u \leftarrow v}^* \frac{\partial \Psi_{u,c}}{\partial H_u} = J_{u \leftarrow v}^* \frac{H_u}{C_u}.$$

This equality uses only VJP linearity.

For a node with multiple selected children, define

$$\Psi_{\mathcal{C}_v} = \sum_{u \in \mathcal{C}_v} \frac{1}{2C_u} \|H_u\|_F^2.$$

Its differential with respect to H_v is

$$R_v = \sum_{u \in \mathcal{C}_v} J^*_{u \leftarrow v} \frac{H_u}{C_u}. \tag{29}$$

Thus summing child VJPs is the differential of a well-defined local observable. Dividing again by $|\mathcal{C}_v|$ would define a different observable and change the child gauge.

The channel–coordinate terminal seed has the same variational interpretation. For a nonempty channel set S and nonempty coordinate set Q, let $P_{S,Q}$ be the shape-preserving orthogonal projector in eq. (15), and let $C_S = |S|$. Then

$$\Psi_T^{(S,Q)}(H_T) = \frac{1}{2C_S}\|P_{S,Q}H_T\|_F^2, \qquad \nabla_{H_T}\Psi_T^{(S,Q)} = \frac{P_{S,Q}H_T}{C_S}. \tag{30}$$

The coordinate-set cardinality does not enter the divisor. The full target selects every terminal channel and coordinate, so $P_{S,Q} = I$ and $C_S = C_T$.

B.4 Distributional Wiener Projection and the Proxy Bridge

The ideal operator in eq. (27) applies a child precision before the adjoint and a parent covariance after it. The deployed source in eq. (29) instead uses an observable Euclidean VJP. Calibration therefore estimates a ridge-regularised distributional map from this source to the parent state. Once (R_v, H_v) is chosen as the observable source-target pair, the MSE-optimal nonlinear predictor is $\mathbb{E}[H_v \mid R_v]$. The deployed projection restricts this predictor to the frequency-diagonal, full-channel linear class with the declared stored-frequency partition.

Let $g \in \Pi_v$ be a stored-frequency group, assign each stored bin a positive weight a_ω, and define $a_g = \sum_{\omega \in g} a_\omega$. The grouped ridge objective is

$$\mathcal{L}_{v,g}(G) = \frac{1}{a_g}\sum_{\omega \in g} a_\omega \mathbb{E}\,\|GR'_v(\omega) - H'_v(\omega)\|_2^2 + \lambda\|G\|_F^2.$$

Using the weighted pooled moments in eq. (20), its normal equation is

$$G\left(\overline{S^{RR}}_{v,g} + \lambda I\right) = \overline{S^{HR}}_{v,g}.$$

For $\lambda > 0$, the unique equality-constrained solution is

$$G^\star_{v,g} = \overline{S^{HR}}_{v,g}\left(\overline{S^{RR}}_{v,g} + \lambda I\right)^{-1}, \qquad G^\star_v(\omega) = G^\star_{v,g} \quad (\omega \in g). \tag{31}$$

The full complex least-squares derivation and Fourier decomposition are given in section A.3. No Gaussian assumption is required for optimality of this ridge objective within the selected linear class. The reported estimator sets $a_\omega = 1$ for every stored non-redundant rFFT bin. The singleton and all-shared partitions are therefore two equality-constrained instances of the same equally weighted stored-spectrum objective; neither is claimed to equal a coordinate-domain Parseval objective with different bin weights.

The statistics in eq. (31) encode the relation induced jointly by the input-dependent Jacobian, the data distribution, and the mean-seed observable. They do not separately identify the Jacobian or either covariance metric. The fitted G_v is therefore a closed-form distributional projection surrogate for the unavailable metric-aware inverse. During calibration, $R_v(x)$ combines the forward child states with the source-local Jacobians through its VJP. During deployment, $R_v^{\mathrm{on}}(x)$ instead uses the supplied repaired child states, while the fitted projection G_v remains shared across samples. When local operators and covariance geometry are fixed or vary weakly across source inputs, one shared distributional projection can capture the dominant geometric correction. Stronger source dependence can leave an operating-point-specific mismatch in the relation between $R_v(x)$ and $H_v(x)$. Error propagated by the supplied child states is accounted for separately by the finite-DAG recursion in section B.6.

Dependency support. By the chain rule, a child seed affects node v only through computational paths connecting v to that child. Entries outside the corresponding Jacobian sparsity pattern receive zero cotangent. This is a theorem-level dependency statement.

State-structure proxy. Dependency support does not imply that nonzero cotangent support equals forward activation support. The working assumption is weaker: the coordinate or semantic organisation of R_v remains predictive of the organisation carried by H_v. This assumption is evaluated through state-inversion diagnostics, modality-native input-domain structure, and transfer.

Local transfer. Different terminal targets change the seed and therefore the observable source. Using one calibrated map family across compatible target depths and channel–coordinate masks tests whether the dominant correction is node-local rather than a stored query-specific reconstruction.

B.5 JVP-FC AS A SECOND RESIDUAL REGRESSION

Let $y_v^0 = G_v R_v$ be the first-stage estimate. JVP-FC evaluates this estimate under the local differential of the actual source operating point and thereby constructs a second, operating-point-conditioned observable. This mechanism is especially relevant for normalisation, attention, and data-dependent gating or routing, although the residual regression itself is not restricted to those components. In the JVP below, y_v^0 is interpreted as a tangent vector at the fixed forward point, not as a replacement forward activation. For each selected child, define the JVP residual

$$e_{u|v} = H_u - J_{u\leftarrow v} y_v^0. \tag{32}$$

The exact decomposition is

$$e_{u|v} = \left(H_u - J_{u\leftarrow v} H_v\right) + J_{u\leftarrow v}(H_v - y_v^0). \tag{33}$$

The first term is the offset or non-homogeneous base component missing from the zero-intercept differential. The residual therefore need not vanish at the exact parent state. For a positively one-homogeneous local map, including the bias-free linear case, Euler's identity makes this base term zero at differentiable points. The second term transports the first-stage state error through the current sample's local differential and exposes its source-specific child response.

Pulling the residual back gives

$$C_v = \sum_{u\in\mathcal{C}_v} J_{u\leftarrow v}^* e_{u|v}. \tag{34}$$

Define

$$\Phi_v(y) = \frac{1}{2}\sum_{u\in\mathcal{C}_v} \|H_u - J_{u\leftarrow v} y\|_F^2.$$

Direct differentiation gives $C_v = -\nabla\Phi_v(y_v^0)$. The unnormalised child residual therefore corresponds to a specific Euclidean child-consistency objective. Let $\Delta_v = H_v - y_v^0$, $C_v' = \mathcal{F}_k C_v$, and $\Delta_v' = \mathcal{F}_k \Delta_v$. Using the same partition and stored-bin weights as G_v, the second map solves

$$D_{v,g}^\star = \arg\min_D \frac{1}{a_g}\sum_{\omega\in g} a_\omega \mathbb{E}\,\|DC_v'(\omega) - \Delta_v'(\omega)\|_2^2 + \lambda_D \|D\|_F^2. \tag{35}$$

Let $\overline{S}_{v,g}^{CC}$ and $\overline{S}_{v,g}^{\Delta C}$ denote the correction moments pooled with the same weights as in eq. (20). Its normal equation and unique solution for $\lambda_D > 0$ are

$$D\left(\overline{S}_{v,g}^{CC} + \lambda_D I\right) = \overline{S}_{v,g}^{\Delta C}, \qquad D_{v,g}^\star = \overline{S}_{v,g}^{\Delta C}\left(\overline{S}_{v,g}^{CC} + \lambda_D I\right)^{-1}.$$

Thus D_v has the same grouped ridge-regularised Wiener form as G_v, with C_v as source and Δ_v as target. Its guarantee is optimality for the stated distributional ridge objective, not improvement at every boundary or on every metric. This does not make every residual component observable. Any child-residual component in the null space of the summed pullback is absent from C_v, and the fitted map can estimate only parent-error components that are predictable from the resulting source within the chosen linear class.

The online version replaces each true child state by the already inverted state:

$$C_v^{\mathrm{on}} = \sum_{u\in\mathcal{C}_v} J_{u\leftarrow v}^*\left(y_u - J_{u\leftarrow v} y_v^0\right). \tag{36}$$

This is deployable because y_u is available during target-to-input recursion. When the downstream state estimate is accurate, the online residual approaches the calibration residual. Otherwise it exposes the child state that is actually propagated.

The term $J_{u\leftarrow v}y_v^0$ is a JVP at the original forward point. It is not a nonlinear evaluation of $f_{u\leftarrow v}$ on y_v^0. The correction is therefore a residual regression in local differential coordinates, not a claim that the nonlinear subgraph is homogeneous. The matched ablations in section C.7 examine aggregate residual reduction and the non-monotonic boundary-level behaviour of this distributional projection.

B.6 Channel–Coordinate-Conditioned Finite-DAG Error Propagation and Stability

The local closed-form objectives do not by themselves control how errors accumulate during the target-to-input recursion. We first give an exact recursion relative to an arbitrary channel–coordinate-conditioned reference family, then obtain the forward-state theorem as the full-target special case.

Fix an input x, a nonempty terminal channel set S, and a nonempty terminal coordinate set Q. Let $C_S = |S|$. Equip each selected node space with a fixed norm and each block between node spaces with the corresponding induced operator norm. For a child node u, define the online seed weight

$$\alpha_u^{(S)} = \begin{cases} 1/C_S, & u = T, \\ 1/C_u, & u \neq T. \end{cases} \tag{37}$$

For a nonterminal boundary node v, stack arbitrary child states as $z_{\mathcal{C}_v} = (z_u)_{u\in\mathcal{C}_v}$ and define

$$\mathcal{P}_v^{(S)} z_{\mathcal{C}_v} = \sum_{u\in\mathcal{C}_v} J_{u\leftarrow v}^* \alpha_u^{(S)} z_u, \qquad \mathcal{Q}_v z_{\mathcal{C}_v} = \sum_{u\in\mathcal{C}_v} J_{u\leftarrow v}^* z_u, \tag{38}$$

$$\mathcal{L}_v = \sum_{u\in\mathcal{C}_v} J_{u\leftarrow v}^* J_{u\leftarrow v}. \tag{39}$$

At the fixed forward point, the complete update driven by these child states is

$$y_v^0 = G_v \mathcal{P}_v^{(S)} z_{\mathcal{C}_v}, \tag{40}$$

$$C_v = \mathcal{Q}_v z_{\mathcal{C}_v} - \mathcal{L}_v G_v \mathcal{P}_v^{(S)} z_{\mathcal{C}_v}, \tag{41}$$

$$y_v = \mathcal{K}_v^{(S)} z_{\mathcal{C}_v}, \qquad \mathcal{K}_v^{(S)} = G_v \mathcal{P}_v^{(S)} + D_v \left(\mathcal{Q}_v - \mathcal{L}_v G_v \mathcal{P}_v^{(S)} \right). \tag{42}$$

Thus the local update is linear in the supplied child states, although its Jacobian-dependent coefficients remain input-conditioned. The displayed operator uses the unit residual and correction weights evaluated in the paper; any fixed scalar weights can be absorbed into the effective correction operator in $\mathcal{K}_v^{(S)}$. The superscript S records dependence through the terminal normalisation $C_S = |S|$; at a fixed operating point, the operator does not depend on the identity of the selected channels or on Q. The channel and coordinate identities enter through the terminal state $P_{S,Q}H_T$.

Let $\{\bar{H}_v^{(S,Q)}\}$ be any channel–coordinate-conditioned reference family satisfying

$$\bar{H}_T^{(S,Q)} = P_{S,Q} H_T. \tag{43}$$

Define the local defect

$$\epsilon_v^{(S,Q)} = \mathcal{K}_v^{(S)} \bar{H}_{\mathcal{C}_v}^{(S,Q)} - \bar{H}_v^{(S,Q)}. \tag{44}$$

For online errors $\delta_v^{(S,Q)} = y_v - \bar{H}_v^{(S,Q)}$, linearity gives

$$\boxed{\delta_v^{(S,Q)} = \epsilon_v^{(S,Q)} + \mathcal{K}_v^{(S)} \delta_{\mathcal{C}_v}^{(S,Q)}}. \tag{45}$$

The terminal error is zero because the online and reference terminal states are both $P_{S,Q}H_T$.

For full-target inversion, $S = \{1, \ldots, C_T\}$ and $Q = \Omega_T^{\mathrm{coord}}$, the complete terminal coordinate set, so $P_{S,Q} = I$ and $C_S = C_T$. Choosing $\bar{H}_v^{(S,Q)} = H_v$ then recovers the forward-state errors and local defects used in the main text. For a strict channel or coordinate subset, eq. (45) remains valid relative to any chosen channel–coordinate-conditioned reference family. The theorem controls

propagation around that reference but does not define a unique semantic ground truth for channel- or coordinate-specific inversion. The reported GPT-2 query selects every hidden channel but only the terminal position, so it is a strict coordinate query rather than this full-target special case.

Write $\mathcal{K}_{vu}^{(S)}$ for the block acting on child u and $\kappa_{vu}^{(S)} = \|\mathcal{K}_{vu}^{(S)}\|_{\mathrm{op}}$. Then

$$\|\delta_v^{(S,Q)}\| \le \|\epsilon_v^{(S,Q)}\| + \sum_{u \in \mathcal{C}_v} \kappa_{vu}^{(S)} \|\delta_u^{(S,Q)}\|. \tag{46}$$

Because the graph is acyclic and the terminal error is zero, this inequality can be unrolled over all paths from v toward the target. If $w \succeq v$ denotes that w is a selected descendant of v and $\mathrm{Paths}(v, w)$ is the set of selected-DAG paths from v to w, then

$$\|\delta_v^{(S,Q)}\| \le \sum_{w \succeq v,\, w \neq T} \left(\sum_{p \in \mathrm{Paths}(v,w)} \prod_{(a,b) \in p} \kappa_{ab}^{(S)} \right) \|\epsilon_w^{(S,Q)}\|, \tag{47}$$

where the empty path for $w = v$ has product one.

If every local defect is zero, reverse topological induction gives zero error at every selected node. If the local reverse gain satisfies

$$\sup_{v \neq T} \sum_{u \in \mathcal{C}_v} \kappa_{vu}^{(S)} \le q < 1, \tag{48}$$

then

$$\max_v \|\delta_v^{(S,Q)}\| \le \frac{\max_{v \neq T} \|\epsilon_v^{(S,Q)}\|}{1 - q}. \tag{49}$$

Indeed, if $M = \max_v \|\delta_v^{(S,Q)}\|$ and $E = \max_{v \neq T} \|\epsilon_v^{(S,Q)}\|$, then eq. (46) gives $M \le E + qM$. This condition is sufficient rather than necessary. When it fails, eq. (47) still identifies the boundaries and paths that can amplify local defects. These are finite-pass stability statements, not iterative convergence claims. The empirical path-gain study in section C.8 does not reliably rank global error across all nodes or architectures; it is reported as a conservative stability diagnostic rather than evidence of contraction or global convergence.

B.7 Zero Target-Independent Injection

The two stages have the zero-intercept form

$$y_v^0 = G_v R_v, \qquad y_v = y_v^0 + D_v C_v.$$

If the selected target state satisfies $P_{S,Q} H_T = 0$, its terminal seed is zero. Induction through the DAG then gives zero local VJP sources, zero first-stage states, zero JVP residual sources, and zero inverted states. The calibrated matrices can encode a calibration-distribution-dependent linear prior, but they cannot add a fixed target-independent component.

This is narrower than support non-expansion. A frequency-diagonal multiplier is convolutional in coordinate space and can spread energy. The guarantee is only that every inverted component is induced by the selected target, the frozen model derivatives, and fixed zero-intercept maps.

B.8 The Frequency-Diagonal Structural Boundary

The main structural restriction is coordinate-frequency diagonality. At each frequency, channels may mix through a full matrix, but different coordinate frequencies do not mix inside a repair map. The stored-frequency partition Π_v may additionally impose equality constraints that tie the matrices assigned to selected bins. For a linear operator on the selected discrete Fourier grid, this is equivalent to a matrix-valued circular convolution under that grid's boundary convention. The equivalence describes the repair operator, not the original network.

Without this restriction, a node with C channels and P coordinate positions would require an unrestricted $(CP) \times (CP)$ linear map. Frequency diagonality instead gives F channel-matrix blocks, while a partition with $K_v = |\Pi_v|$ groups reduces the number of distinct fitted matrices to K_v. This provides the tractability needed for full-channel calibration.

Frequency diagonality and fixed distributional projections are modelling choices. The conditional-mean, natural-adjoint/Wiener, ridge, and finite-DAG results hold under their stated conditions. A frozen finite differentiable DAG defines computational scope; it does not by itself guarantee observable sufficiency, well-conditioned calibration, or uniformly accurate inversion. Whether the optimal repair is adequately approximated in this class is an empirical question. The evaluated CNN and vision-Transformer boundaries, including the strict ViT-B/32 patch-coordinate queries, together with the controlled GPT-2 small terminal-token setting and joint full-coordinate channel-subset control, support this operator class in the studied configurations. Internal attention-logit or attention-weight matrices, arbitrary nonterminal language-token queries, and transfer to other language-model architectures remain outside the present empirical validation. The supplementary experiments next test each empirical bridge separately: observable transfer, residual repair, finite-DAG diagnostics, information flow, and the practical calibration and deployment boundaries.

C Supplementary Experiments

This supplement expands the five main experimental questions into a detailed protocol–query–mechanism–validation sequence. It records the detailed evaluation protocol, expanded quantitative results, negative controls, deployment audits, and the optional breadth studies. Visual experiments use image-space metrics to describe alignment with the known target-generating input, whereas the language experiment uses token-position energy profiles and modality-specific routing controls. Neither family is treated as a universal ground-truth correctness measure for a feature inverse. The progression is protocol and reuse, structural and distributional scope, source-path faithfulness, mechanism and theory bridge, inverse meaning, deployment audits, decision-grounded interpretation, optional visual extensions, and finally language-Transformer validation.

C.1 Reproducibility, Metric Definitions, and Final-Result Criteria

Fixed protocol. Unless a row is explicitly marked as an ablation or diagnostic, all results use the final public method: local mean seeds, zero-intercept full-channel maps for both G_v and D_v, repaired online child states, declared architecture-aware boundaries, no activation or cotangent pre-normalisation, and relative ridge $\rho = 0.01$. The default vision configuration uses the singleton per-frequency partition; the reported GPT-2 configuration uses all-shared spectral tying over stored token-frequency bins. The calibration and evaluation sets are disjoint. For the headline ImageNet experiments, 4,096 training images are sampled with seed 123 and 1,024 validation images are sampled with seed 456. Dataset-specific experiments use ImageNet, Oxford-IIIT Pets, and CUB-200-2011 (Deng et al., 2009; Russakovsky et al., 2015; Parkhi et al., 2012; Wah et al., 2011). Pets experiments use all 3,680 available training images for calibration; other dataset-specific calibrations use 4,096 images. We use cN/eM as shorthand for N calibration images and M evaluation images.

The final vision runs used Python 3.10.4, PyTorch 2.5.1 with CUDA 12.1, TorchVision 0.20.1, and timm 1.0.25. Runtime measurements used one NVIDIA GeForce RTX 3090 GPU with 24 GiB of memory on a host with two Intel Xeon Gold 6426Y processors. Each run was bound to one GPU. The fitted-map placement followed the reported deployment configuration: ResNet18 maps remained on the GPU, while the larger headline map families were applied from CPU memory.

Each vision run records the model and weight identifier, dataset root, preprocessing transform, calibration and evaluation indices, target node, fitted boundary DAG, tensor shape, channel axis, coordinate axes, frequency shape, ridge, source-code hashes, map metadata, and per-image metrics. The formal component validator additionally requires the same public code path, no component-specific inverse, finite outputs, and a complete target-axis record. Nondeployable oracle conditions are labelled diagnostic-only and are never used as headline method rows.

Tensor and Fourier convention. A floating tensor is represented as batch, channel, and zero or more coordinate axes as defined in section A.1. Real-valued coordinate axes use an orthonormal real Fourier transform. The calibrated map is full over channels and diagonal over coordinate frequency, while the declared partition Π_v may tie selected stored-frequency matrices. Vector states therefore have one frequency bin, while token streams use the token coordinate as their Fourier axis. The partition changes operator capacity but not the source-local VJP/JVP observables or the boundary

DAG. The forward model fixes the target node and boundary topology before calibration; no inverse component is manually inserted at evaluation time.

Input-instance alignment. For flattened prediction $\widehat{x}$ and known target-generating input x, pixel cosine is

$$\cos_{\mathrm{pix}}(\widehat{x}, x) = \frac{\langle \widehat{x}, x \rangle}{\|\widehat{x}\|_2 \|x\|_2}. \tag{50}$$

The same cosine is computed after separating luminance low-pass, luminance high-pass, and chroma components. SSIM and LPIPS provide complementary structural and perceptual descriptions (Wang et al., 2004; Zhang et al., 2018). Relative ℓ_2 is $\|\widehat{x} - x\|_2 / \|x\|_2$. These metrics quantify input-instance alignment for a known source reference. A low value for a channel or channel-subset target does not by itself invalidate the inverse, because the selected target is $P_{S,Q} H_T$ rather than the complete input.

Representation and mechanism diagnostics. Re-encoding cosine compares the representation of the inverse with the target representation and is reported only as representation consistency. Child NMSE measures the local JVP prediction error before and after D_v. Intermediate-state relative ℓ_2 measures the corresponding state-space error. Runtime is reported for post-calibration online application. Calibration complexity is described separately by calibration-set size, fitted-node count, map storage, and peak memory rather than by cross-architecture fit wall time. Map storage includes both G_v and D_v; peak memory is measured separately during calibration and evaluation.

Uncertainty and visual selection. Unless a study specifies a subset, ImageNet means use all 1,024 fixed evaluation images (Deng et al., 2009; Russakovsky et al., 2015). Calibration-estimation variation is assessed with independently sampled calibration subsets and is kept separate from evaluation-image uncertainty. Operating-point curves use deterministic 10,000-resample bootstrap intervals. The decision-conditioned study aggregates random controls within each image before deterministic 20,000-resample image-cluster bootstrap intervals, which prevents random-control repeats from being treated as independent images. The language benchmarks average the two calibration-map outputs before 10,000 percentile cluster-bootstrap resamples over polysemous words or lexical templates; decision-intervention intervals resample held-out contexts. Visual sample indices, ordering, targets, and display transformations are fixed before figure assembly. Inverse tensors are scaled only for display after all quantitative measurements.

C.2 Large-Scale Atlas, No-Refit Reuse, Target Depth, and Optimisation Details

One fit, many queries. The canonical ResNet50 atlas uses one c4096 fitted map family. It serves 1,024 evaluation images, four full-state depths, four terminal channels, and one eight-channel terminal subset without refitting any map or changing the ridge. The four full-state target keys are `full_n1`, `full_n5`, `full_n9`, and `full_n17`. Their pixel cosines are 0.9448, 0.9282, 0.9339, and 0.9333, respectively. The four individual terminal channels have pixel cosines 0.8254, 0.5979, 0.9243, and 0.9199; the fixed eight-channel subset reaches 0.8932. The channel results demonstrate public-API reuse and output diversity. They are not evaluated as attempts to reproduce the complete input from one channel.

Figure 5 shows all 24 predeclared source examples and the complete fixed depth strip. Sample indices, ordering, targets, and display transformations were fixed before rendering.

Canonical iterative preimage optimisation. The external numerical baseline starts from noise and minimises the terminal feature discrepancy for 2,000 Adam steps per image (Kingma & Ba, 2015). The learning rate is 0.05, the total-variation and ℓ_2 weights are 0.1 and 0.001, and random spatial jitter is bounded by four pixels (Rudin et al., 1992). A fixed pilot sweep from 1,000 to 3,000 steps selected this setting because additional steps produced only a small alignment change while increasing cost. The target layer, preprocessing, evaluation images, and feature objective are matched to the closed-form run. Three predeclared initialisation seeds are used. No learned decoder or saliency method is included because the intended numerical comparison is between canonical feature-preimage search and the proposed source-grounded closed-form regime.

Across the matched 16-image set, iterative optimisation obtains mean target re-encoding cosine 0.9575 and feature NMSE 0.0794, but pixel cosine is 0.0351. Mean pairwise output cosine across the

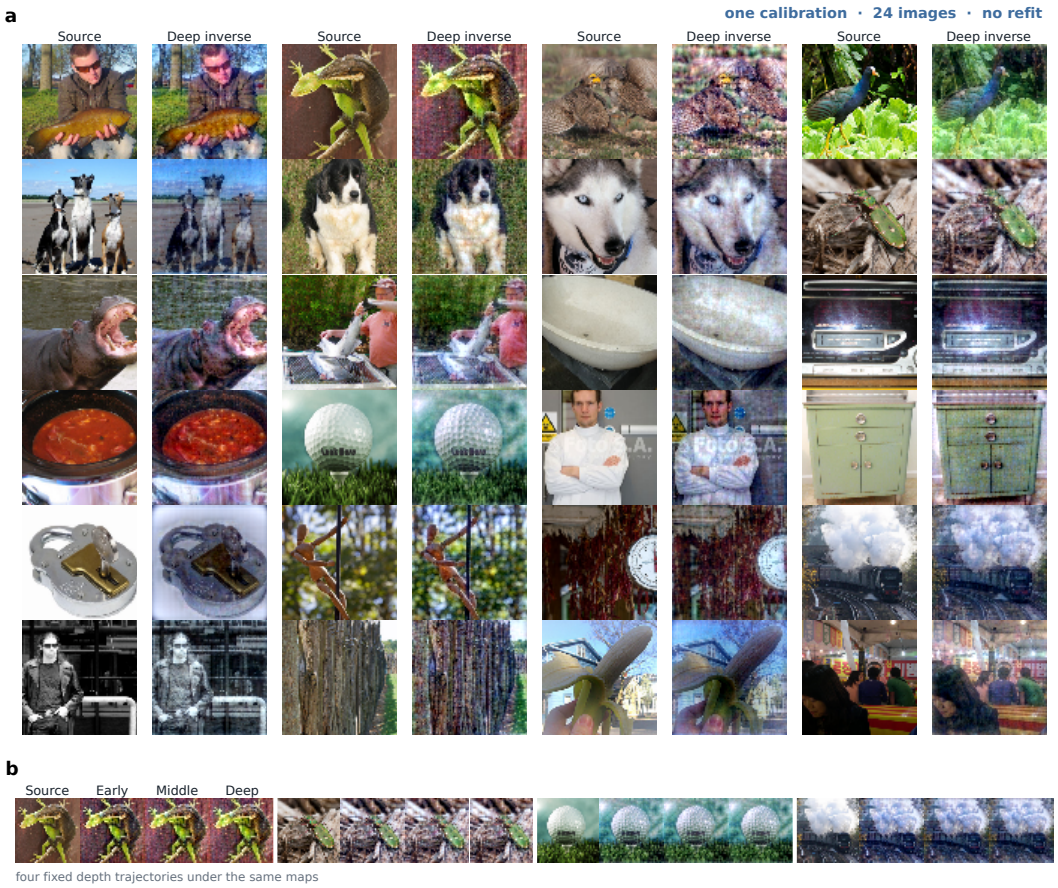


Figure 5: **Feature inversion at scale under one calibration.** **a**, The dominant plate contains 24 predeclared ImageNet source references paired with deep ResNet50 feature inverses. **b**, The lower strip shows four fixed examples across early, middle, and deep targets. All outputs use one calibrated map family, one public inversion path, and no query-specific optimisation or refitting. The source images are references for input-instance alignment, not unique inverse labels.

three initialisations is 0.1010, demonstrating that the converged preimages remain seed dependent. The closed-form estimator uses the target-generating sample's local geometry to select a source-conditioned inverse branch. After calibration, the final closed-form method obtains pixel cosine 0.9388 and requires 0.9339 seconds for all 16 online queries, or 0.0584 seconds per query in the measured batched application run. Raw VJP has pixel cosine 2.5×10^{-5} on the same set. The optimisation result is not called an incorrect preimage. It demonstrates that target-only preimage search and source-local feature inversion use different information and can select different solutions of an underdetermined feature constraint.

The matched regime comparison in table 3 uses a separate sequential batch-one timing: the optimisation value is an eight-query mean, and the closed-form value is the median of three isolated repeats. The resulting 24.379 versus 0.139 seconds is a 176× reduction in post-calibration query time. It does not include calibration.

Initialisation sensitivity and post-calibration latency. The three optimisation seeds reach similar target objectives, while their outputs remain seed dependent. We therefore report the predeclared seed mean as the baseline result and retain the median only as a diagnostic. In an isolated three-repeat benchmark, closed-form batch-one online latency is 0.1385 seconds when images vary, 0.1104 seconds when depths vary, 0.1408 seconds when channels vary, and 0.1274 seconds for a mixed workload. The corresponding 2,000-step iterative-optimisation references require 23.5421, 21.7281,

Table 3: **Canonical feature-inversion regimes on a matched 16-image ResNet18 target.** Pixel cosine is the mean descriptive input-instance alignment on the same fixed 16 images. Online time uses sequential batch-one image queries: the optimisation value is the mean over eight queries, while ours is the median over three isolated repeats. The optimisation reference uses 2,000 steps per image.

Method	Source-local geometry	Per-query search	Online/query (s)	Pixel cosine
Iterative preimage optimisation	No	2,000 steps	24.379	0.035
Ours, final JVP-FC	Yes	0 steps	**0.139**	**0.939**

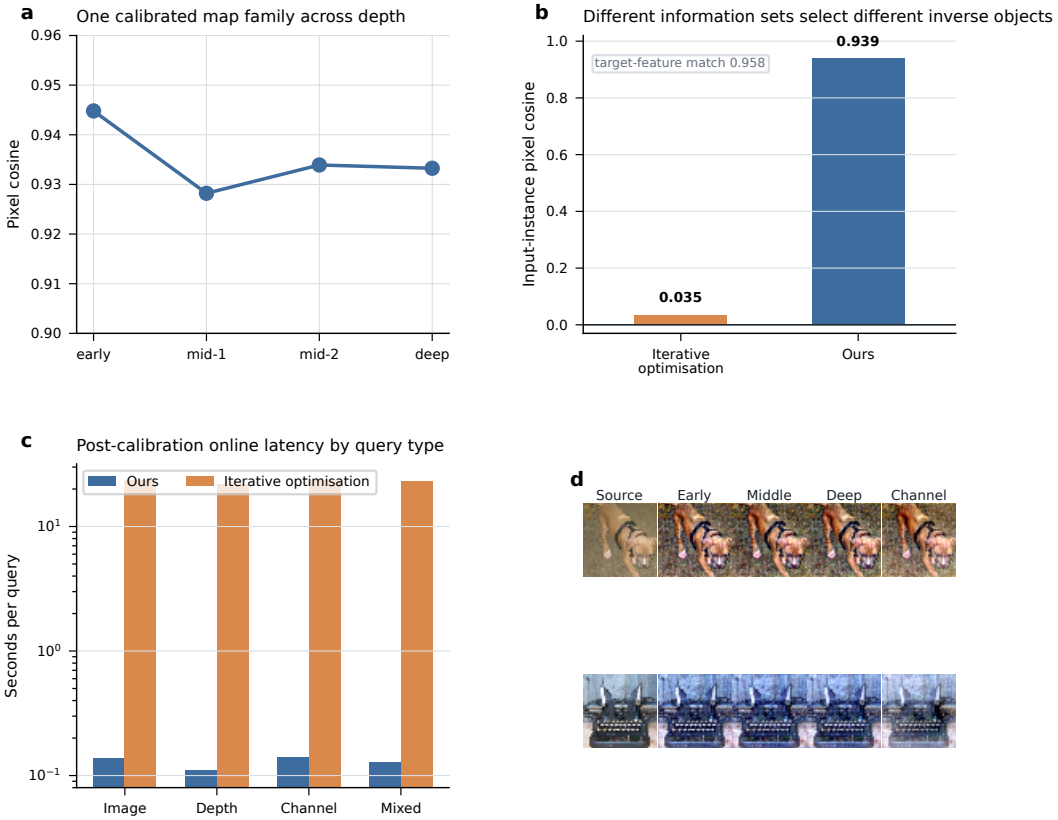


Figure 6: **Expanded no-refit reuse and optimisation-regime evidence. a**, One ResNet50 calibration across four full-state depths. **b**, Descriptive input-instance alignment for the matched iterative and closed-form rows; the iterative target-feature match is annotated separately because it is a representation-consistency diagnostic. **c**, Post-calibration online latency across image, depth, channel, and mixed queries. **d**, Fixed depth and channel queries under one fitted map family.

23.5781, and 23.0087 seconds per query. All values in this comparison begin after calibration so that they measure the deployment-time scaling of each query type.

Figure 6 combines the no-refit atlas and fixed depth/channel queries with the matched regime and online-latency comparisons.

Figure 7 reports depth-dependent structure under fixed maps, the separation between feature matching and input alignment during optimisation, and cumulative online cost for mixed queries.

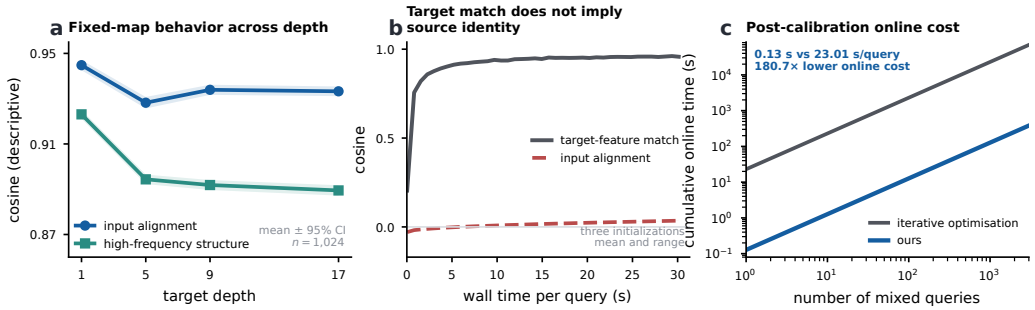


Figure 7: **Depth behaviour, optimisation trajectories, and query cost. a**, Descriptive input-instance alignment and high-frequency structure across ResNet50 target depths under the same fitted maps (mean and 95% confidence interval over $n = 1{,}024$ images). **b**, Optimisation target-feature match and input-instance alignment versus wall time (mean and range over three initialisation seeds). **c**, Cumulative online time for repeated mixed queries, using the measured per-query rates of 0.1274 s for our method and 23.0087 s for iterative optimisation. Offline calibration is excluded.

C.3 Channel–Coordinate Query Generalisation in ViT

Protocol and support definition. The formal patch-coordinate experiment used the ImageNet-pretrained torchvision ViT-B/32 checkpoint and a target token stream of shape $B \times 50 \times 768$, comprising one CLS token and a 7×7 patch grid. The hidden width remained the channel axis and the token index remained the coordinate and Fourier axis. Each query retained all hidden channels and one non-CLS patch coordinate,

$$S = \{1, \dots, 768\}, \qquad Q = \{q\}, \qquad q \in \{1, \dots, 49\}.$$

One isolated c4096 map family was reused for all queries, images, and target depths without patch-specific refitting. Evaluation used 128 fixed, correctly classified ImageNet validation images and layers 2, 8, and 14, giving $128 \times 49 \times 3 = 18{,}816$ inversions.

For inverse $\widehat{x}_{l,q}$ from queried internal patch q, its input-domain support at input patch r is

$$A_l(q, r) = \frac{M_r(\widehat{x}_{l,q})}{\sum_{r'} M_{r'}(\widehat{x}_{l,q})}, \qquad M_r(\widehat{x}) = \sum_{(h,w)\in r} \|\widehat{x}_{:,h,w}\|_2. \tag{51}$$

Thus the reported normalised inverse-magnitude support mass first computes a channel L_2 magnitude at every pixel, sums it inside each 32×32 input patch, and normalises over the 49 input patches. It is not squared inverse energy. The matrix reports where a patch-selected inverse is expressed in input coordinates; it is not attention, attribution, or causal importance.

All 49 patch queries. The average layer-2 support matrix was strongly diagonal and locally banded. With depth, self-patch and local-neighbour mass decreased while profile entropy and expected support distance increased (table 4 and fig. 8). The direct observation is a transition from local, coordinate-aligned support to distributed input-domain support. Contextual mixing is a compatible interpretation of the deeper pattern, but semantic content is not inferred from these scalar summaries alone.

Table 4: **ViT patch-query input-domain support from one c4096 calibration.** Entries are means with 95% image-bootstrap intervals over 128 fixed validation images. Support radius is the expected patch-grid distance normalised by the maximum grid distance.

Layer	Self-patch mass	Profile entropy	Support radius	Local 3×3 mass
2	0.417 (0.413–0.421)	0.649 (0.645–0.654)	0.183 (0.181–0.184)	0.720 (0.716–0.725)
8	0.139 (0.136–0.143)	0.915 (0.912–0.918)	0.324 (0.321–0.327)	0.396 (0.389–0.402)
14	0.065 (0.063–0.067)	0.964 (0.962–0.965)	0.376 (0.374–0.378)	0.273 (0.268–0.278)

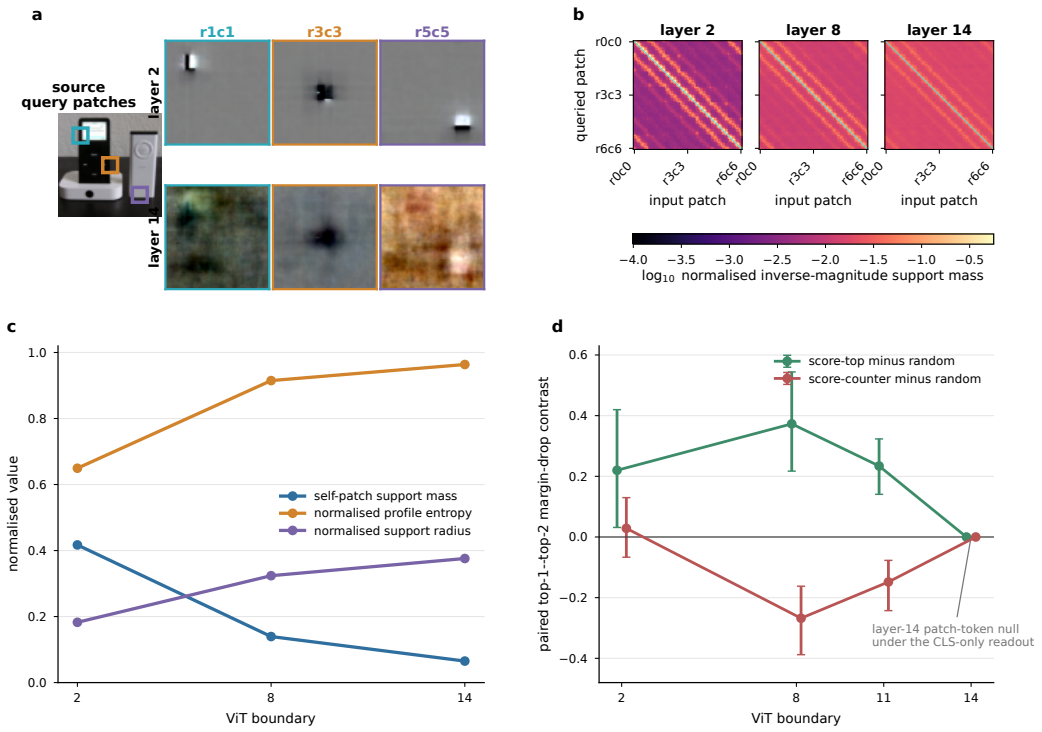


Figure 8: **One calibration supports all ViT patch-coordinate queries. a**, Representative inverses for three fixed patch queries at layers 2 and 14 on the first fixed evaluation image. Coloured boxes mark the query coordinates; the same signed visualisation rule is applied to each inverse. **b**, Patch-query input-domain support matrices averaged over 128 correctly classified ImageNet validation images. Rows index queried internal patches and columns index input patches; the cyan line marks equal coordinates. **c**, Mean self-patch support mass, normalised profile entropy, and normalised support radius; shading gives 95% image-bootstrap intervals. **d**, Paired exact top-1–top-2 margin-drop contrasts after independent internal zero intervention on each of the 49 patch-token states. Points are means and bars are 95% image-bootstrap intervals ($n = 128$). Patch ranking uses gradient times activation on the original representation; the exact intervention is a separate forward evaluation. Layer 14 is a patch-token null control under the CLS-only readout; no CLS intervention is shown.

Independent internal intervention. For the same 128 images, every one of the 49 patch-token vectors was independently zeroed at layers 2, 8, 11, and 14 and the fixed top-1–top-2 class margin was recomputed, giving 25,088 interventions. Gradient times activation was computed on the original representation and used only to identify the score-top and score-counter patch in each image; exact margin drop was measured by a separate forward evaluation. A deterministic random patch provided the paired reference.

Table 5: **Exhaustive ViT patch-token intervention.** Entries are means with 95% image-bootstrap intervals over 128 images. Spearman correlation compares the gradient-times-activation ranking with all 49 exact intervention effects within each image. It is undefined at layer 14 because every patch-token intervention is numerically zero under the CLS-only readout.

Layer	Score-top − random	Score-counter − random	Score/drop Spearman
2	0.220 (0.031–0.420)	0.029 (−0.067–0.130)	0.134 (0.104–0.164)
8	0.373 (0.217–0.544)	−0.268 (−0.388–−0.162)	0.354 (0.319–0.388)
11	0.234 (0.141–0.323)	−0.148 (−0.243–−0.077)	0.372 (0.324–0.419)
14	0.000 (0.000–0.000)	0.000 (0.000–0.000)	—

Layer 8 showed the clearest signed separation: score-top patch removal reduced the class margin relative to random removal, whereas score-counter removal shifted it in the opposite direction. Layer

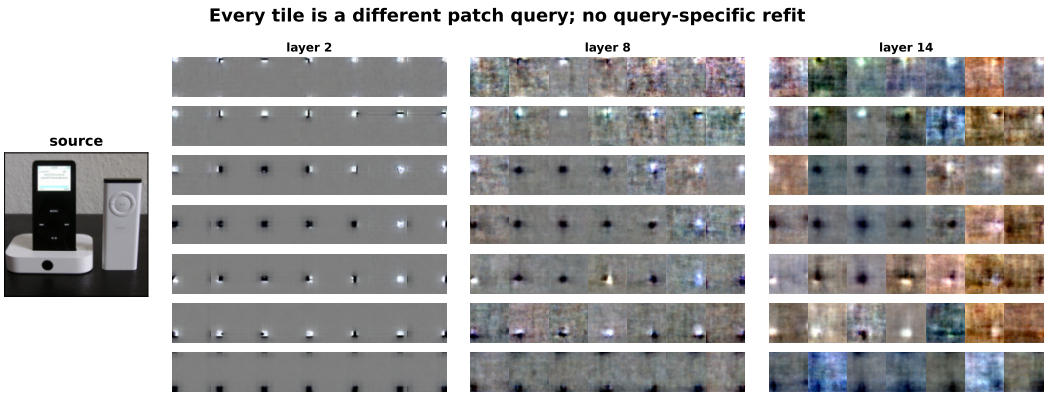


Figure 9: **Complete 49-query ViT inverse atlas for one fixed image.** The source is followed by every non-CLS patch query at layers 2, 8, and 14, ordered on the 7×7 patch grid. All 147 inverses use the same c4096 map family with no query-specific refit. Each inverse is displayed with the same signed visualisation rule and its own fixed 99th-percentile magnitude scale; quantitative support is measured before display scaling.

11 retained both effects. Layer-14 patch-token interventions were null to numerical precision (mean scale approximately 3×10^{-8}), as expected because the final torchvision classifier reads only the CLS token after the last normalisation. This is a patch-token null control under a CLS-only readout; no CLS ablation is included in the formal c4096/e128 experiment.

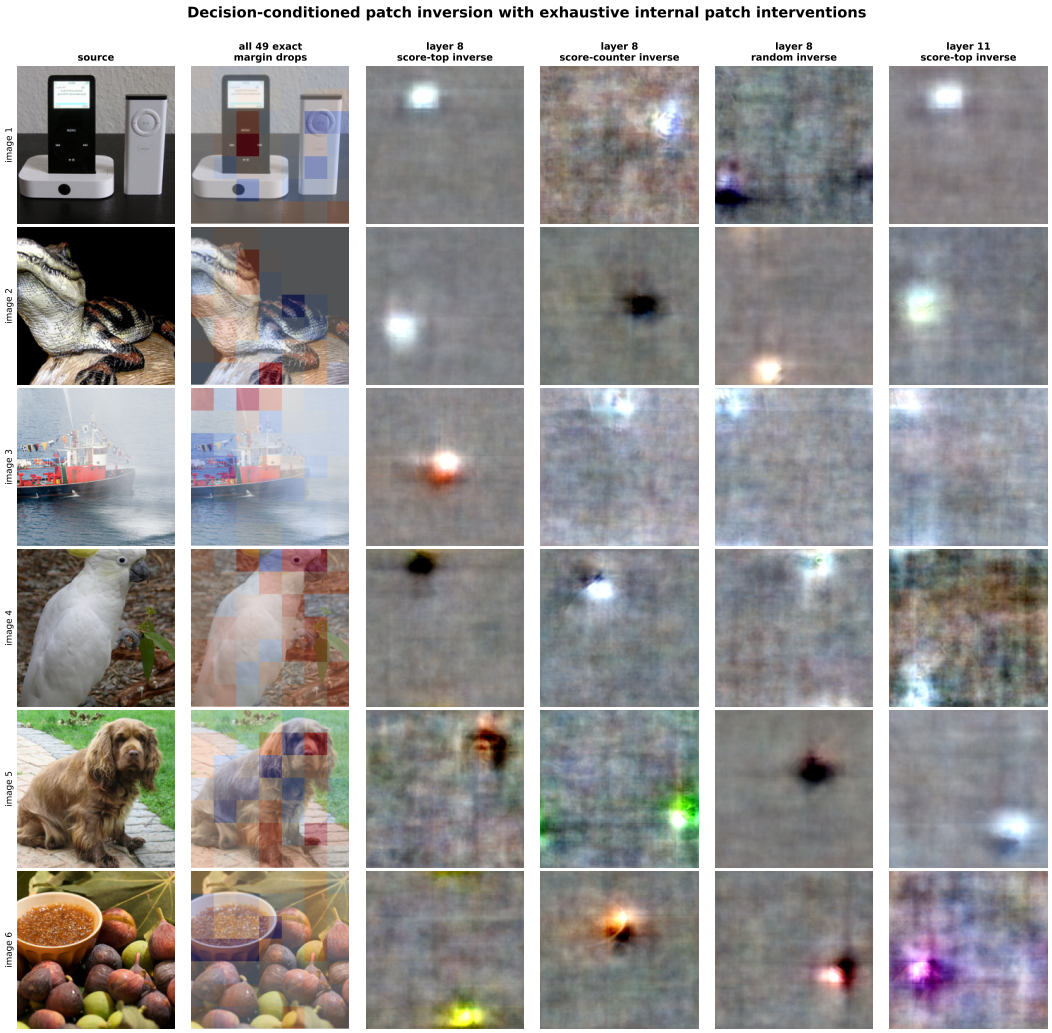


Figure 10: **Patch-query inverses paired with exhaustive internal interventions.** Rows are the first six fixed evaluation images. The second column overlays all 49 exact layer-8 patch-token margin drops on the source. Subsequent columns show the layer-8 score-top, score-counter, and random-query inverses, followed by the layer-11 score-top inverse. The inverse panels are descriptive input-domain views; causal interpretation is restricted to the separately evaluated internal patch-token interventions.

C.4 Expanded Architecture and Tensor-Component Results

Table 6 records the exact terminal tensor and axis convention for every headline ImageNet backbone (He et al., 2016; Huang et al., 2017; Sandler et al., 2018; Tan & Le, 2019; Radosavovic et al., 2020; Liu et al., 2022; 2021; Dosovitskiy et al., 2021).

The main text summarises nine ImageNet models spanning eight backbone families; table 7 gives their complete compact comparison, including the final-minus-G_v gain, and table 8 resolves their frequency components. All rows pass the final-code validator and use c4096/e1024 with the same public recipe. The full results are reproduced in table 8 to expose architecture-dependent structure rather than reducing generality to one scalar ranking.

A matched boundary audit showed that block-output-only capture was still too coarse for stage-entry MobileNetV2 and EfficientNet-B0 blocks. Stem-only changes and indiscriminate full-DAG refinement did not improve the matched c512 diagnostic. The decisive change was to expose the expansion/depthwise tensors of each non-residual stage-entry block while leaving residual blocks at

Table 6: **Headline ImageNet terminal and boundary setups.** All rows use full-state terminal targets, c4096/e1024, and registered model-native boundary DAGs. MobileNetV2 and EfficientNet-B0 factorise each stage-entry non-residual block at expansion/depthwise tensors, retain residual blocks at their add endpoints, and terminate at the final 320-channel feature block. Tensor shapes include the batch dimension; channel and coordinate axes use zero-based indexing.

Model	Captured terminal tensor	Shape	Channel axis	Coordinate axes	Fitted nodes
ResNet18	`op.add_7.add:out`	$1 \times 512 \times 7 \times 7$	1	2,3	9
ResNet50	`op.add_15.add:out`	$1 \times 2048 \times 7 \times 7$	1	2,3	17
DenseNet121	`op.batch_norm_120.batch_norm:out`	$1 \times 1024 \times 7 \times 7$	1	2,3	9
MobileNetV2	`features.17.conv.3:out`	$1 \times 320 \times 7 \times 7$	1	2,3	31
EfficientNet-B0	`features.7.0.block.3.1:out`	$1 \times 320 \times 7 \times 7$	1	2,3	30
RegNetY-1.6GF	`op.add_26.add:out`	$1 \times 888 \times 7 \times 7$	1	2,3	28
ConvNeXt-B	`op.add_35.add:out`	$1 \times 1024 \times 7 \times 7$	1	2,3	37
Swin-T	`norm:out`	$1 \times 7 \times 7 \times 768$	3	1,2	26
ViT-B/32	`encoder.ln:out`	$1 \times 50 \times 768$	2	1	14

Table 7: **One recipe across ImageNet backbone families.** All values are means over 1,024 validation images. Δ is the final-minus-G_v pixel-cosine gain. LPIPS is lower for stronger perceptual alignment.

Backbone	G_v pixel	Final pixel	Δ	SSIM	LPIPS
ResNet18	0.901	**0.938**	0.037	0.684	0.314
ResNet50	0.889	**0.931**	0.042	0.726	0.277
DenseNet121	0.751	**0.824**	0.074	0.588	0.419
MobileNetV2	0.781	**0.833**	0.052	0.710	0.260
EfficientNet-B0	0.659	**0.796**	0.137	0.722	0.272
RegNetY-1.6GF	0.872	**0.914**	0.043	0.597	0.423
ConvNeXt-B	0.734	**0.845**	0.111	0.617	0.399
Swin-T	0.178	**0.504**	0.325	0.519	0.520
ViT-B/32	0.312	**0.702**	0.391	0.314	0.757

their add endpoints. For EfficientNet-B0, making the squeeze-and-excitation multiplication itself a boundary caused sample-wise sign instability, so the gate and multiplication remain inside one local edge. MobileNetV2 uses pre-activation convolution tensors at these boundaries, whereas EfficientNet-B0 uses post-activation spatial tensors. The boundary change leaves the estimator, ridge, calibration size, and online API unchanged. The final c4096/e1024 rows reach pixel cosines 0.8333 and 0.7960 at the same last 320-channel targets; no shallower target was substituted.

Figure 11 shows the same fixed source across backbone families, the Swin–ViT spatial comparison, and representative distributions. Swin-T and ViT-B/32 expose different Transformer feature geometry. Swin has stronger SSIM and high-frequency structure, whereas ViT has stronger low-frequency and global pixel alignment. The method thus operates on Transformer computation graphs with distinct spatial and frequency profiles. The comparison contrasts complete architectures that differ in hierarchy, windowing, and patch size.

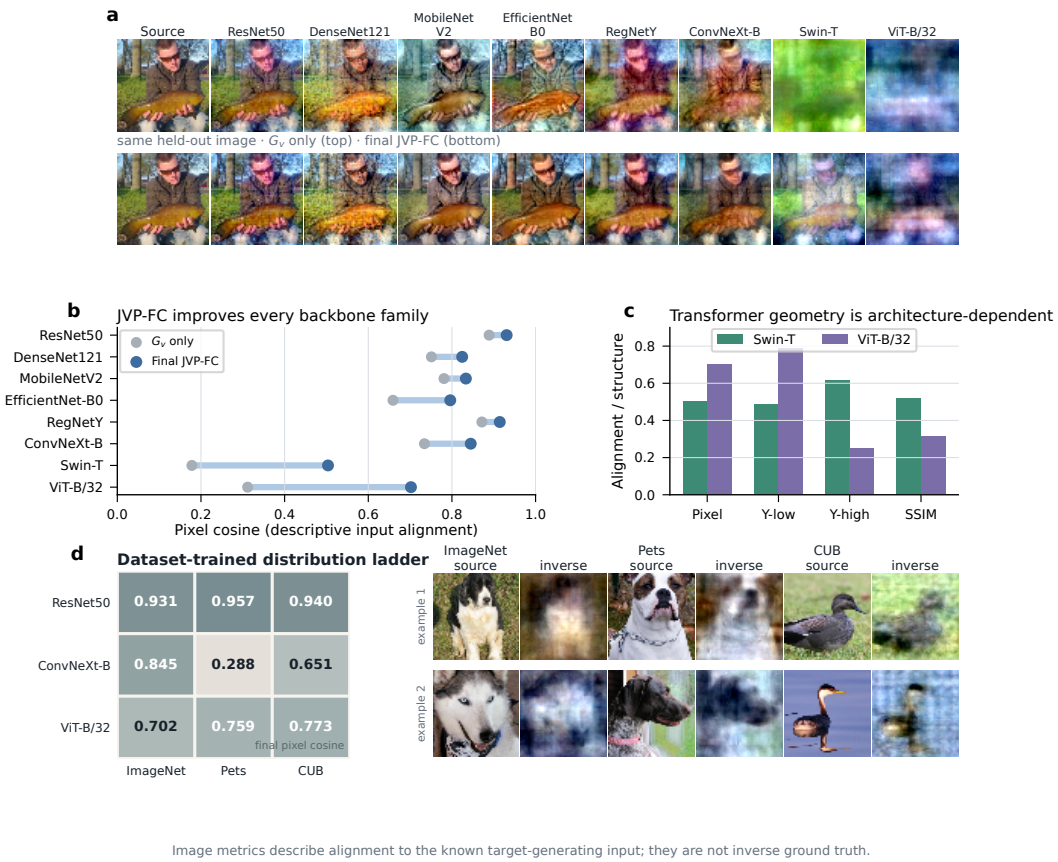


Figure 11: **Structural and distributional generality.** **a**, One fixed held-out image across eight backbone families under one recipe, with G_v-only outputs above the final JVP-FC outputs. **b**, Compact quantitative summary of input-instance alignment and second-stage gain. **c**, Swin-T and ViT-B/32 on identical images, showing architecture-dependent spatial and frequency structure. **d**, Representative ImageNet, Pets, and CUB results. Full per-image distributions and component-level results are reported in the supplement.

Table 8: **Expanded ImageNet architecture results.** Y-low and Y-high are luminance low- and high-frequency cosines. All metrics are descriptive alignment over 1,024 images.

Model	G_v	Final	Y-low	Y-high	Chroma	SSIM	LPIPS
ResNet18	0.9009	0.9375	0.9486	0.8624	0.9041	0.6837	0.3137
ResNet50	0.8889	0.9308	0.9178	0.8900	0.9260	0.7261	0.2772
DenseNet121	0.7510	0.8245	0.8335	0.7243	0.8593	0.5878	0.4188
MobileNetV2	0.7810	0.8333	0.8247	0.8484	0.8244	0.7103	0.2597
EfficientNet-B0	0.6591	0.7960	0.7398	0.8581	0.7729	0.7223	0.2720
RegNetY-1.6GF	0.8716	0.9143	0.9500	0.6553	0.8947	0.5966	0.4231
ConvNeXt-B	0.7343	0.8450	0.8626	0.7088	0.7724	0.6166	0.3985
Swin-T	0.1783	0.5037	0.4844	0.6144	0.4514	0.5187	0.5198
ViT-B/32	0.3116	0.7021	0.7891	0.2480	0.6044	0.3144	0.7568

Tensor-component zoo. The formal zoo contains 12 validated runs and 11 component families. It includes DenseNet concatenations, MobileNet depthwise and inverted-residual states, EfficientNet MBConv and squeeze-and-excitation states, ConvNeXt block and normalisation-adjacent tensors, Swin window-stage and patch-merging states, ViT token streams, and pooled or vector states. Representative results are shown in table 9. Every row uses the same tensor-axis interface and records `manual_component_inverse=False`. Across the five ConvNeXt normalisation-adjacent targets, input-instance pixel cosine ranges from 0.943 to 0.957 and target re-encoding cosine from 0.935 to 0.981.

Table 9: **Representative component-zoo targets.** Shapes include the batch dimension. Re-encoding is a representation diagnostic and is not required to track pixel alignment monotonically.

Model	Component	Target shape	Target example	Pixel	Re-enc.
DenseNet121	concatenation	$1\times256\times56\times56$	dense cat	0.8338	0.8353
MobileNetV2	inverted residual	$1\times160\times7\times7$	residual add	0.4674	0.1961
MobileNetV2	depthwise tensor	$1\times32\times112\times112$	depthwise conv	0.9861	0.9745
EfficientNet-B0	MBConv/SE	$1\times192\times7\times7$	MBConv add	0.3080	0.0248
ConvNeXt-B	normalisation-adjacent	$1\times56\times56\times128$	layer norm	0.9487	0.9445
ConvNeXt-B	MLP state	$1\times56\times56\times512$	MLP fc1	0.9475	0.9814
Swin-T	window stage	$1\times14\times14\times384$	residual add	0.7516	0.6485
Swin-T	patch merging	$1\times28\times28\times384$	merge pad	0.8924	0.9049
EfficientNet-B0	pooled vector	$1\times1280\times1\times1$	average pool	0.2622	-0.1710
EfficientNet-B0	logit vector	1×1000	classifier	0.2640	-0.0050

The component table establishes coverage, not uniform visual alignment. In particular, a pooled or logit vector contains a different selected feature object from a spatial tensor. Its inverse should therefore not be judged by whether it reproduces the complete input.

C.5 Distribution Behaviour and Calibration Transfer

Dataset-trained representative ladder. The representative study recalibrates a dataset-trained model on each dataset while keeping the method recipe fixed. Final pixel cosines for ImageNet/Pets/CUB are 0.9308/0.9566/0.9398 for ResNet50, 0.8450/0.2876/0.6507 for ConvNeXt-B, and 0.7021/0.7590/0.7728 for ViT-B/32. Representative inverses are shown in fig. 11d. The non-monotonic ConvNeXt result prevents a scalar "dataset complexity" interpretation. The supported conclusion is that feature-inverse behaviour depends jointly on the data distribution, checkpoint, and calibrated local statistics.

Eight-family breadth with ImageNet weights. The separate breadth experiment applies ImageNet-weight models to Pets and CUB and recalibrates their maps on the corresponding images. All 16 rows pass the strict artifact and protocol audit. Final pixel cosine ranges from 0.3864 to 0.9451 on Pets and from 0.3684 to 0.9182 on CUB. Swin reaches 0.6757 on Pets and 0.6098 on CUB, while ViT reaches 0.7808 and 0.7193. These rows broaden the visual-distribution evidence but remain distinct from dataset-trained model comparisons.

Calibration-source transfer. We fitted maps on ImageNet, Pets, or CUB and evaluated each map family on all three evaluation distributions. The model checkpoint and method remained fixed within each 3×3 matrix. The complete transfer matrices are shown directly in fig. 12 rather than repeated in a separate table.

The matched-minus-cross average gap is small for ResNet18 (approximately 0.937 versus 0.932) and larger for ViT-B/32 (approximately 0.721 versus 0.662). Thus fixed-map transfer is more sensitive for the token model. Recipe generality across datasets does not imply that one map family is distribution invariant.

Figure 12 collects the component-zoo breadth, the ImageNet-weight distribution study, and both calibration-transfer matrices.

C.6 Source-Path Faithfulness Under Low Pixel Alignment

Question and protocol. Pixel cosine describes alignment between an inverse and the known target-generating image, but does not directly test whether the reconstructed states follow the local computation instantiated by that image. We therefore performed a dedicated source-geometry intervention on independently fine-tuned ConvNeXt-B checkpoints for CUB-200-2011 and Oxford-IIIT Pets. The checkpoints reached 85.73% and 95.42% top-1 accuracy, respectively, excluding a degenerate classifier as the explanation for low pixel alignment. CUB used 4,096 calibration images and Pets used all 3,680 training images; both studies evaluated 512 held-out images with the final full-channel, per-frequency $G_v + D_v$ estimator, relative ridge 0.01, the 38-node ConvNeXt DAG, and terminal node 37. Seven preselected states spanning the path were probed at nodes $\{1, 4, 7, 20, 34, 35, 36\}$.

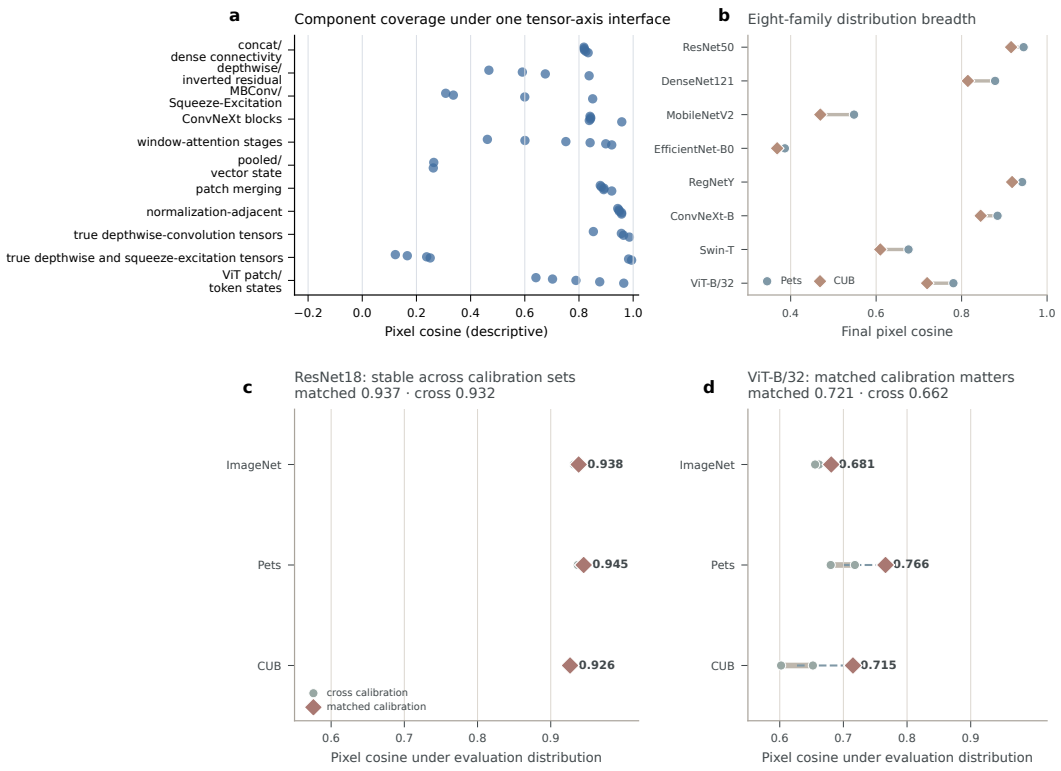


Figure 12: **Expanded component and distribution evidence.** **a**, All validated component-zoo targets grouped by tensor class. **b**, Paired Pets and CUB results for eight ImageNet-weight backbone families under dataset-specific recalibration. **c,d**, Calibration-transfer profiles for ResNet18 and ViT-B/32. Each row retains the two cross-calibration values and the matched-calibration value for one evaluation distribution.

Table 10: **Source-path faithfulness in the lowest pixel-cosine quartile of ConvNeXt-B inverses.** Each row contains 128 of 512 held-out images. Matched and shuffled path NMSE evaluate the same matched reconstructed states under J_A and J_B, respectively. Gap is shuffled minus matched NMSE, with a deterministic 10,000-resample 95% bootstrap interval. Node win is the fraction of the seven probed nodes at which the shuffled-path NMSE exceeds the matched-path NMSE. State margin is the cosine of a matched reconstructed state to the corresponding state of A minus its cosine to the paired state of B. Branch cosine compares the matched and geometry-swapped input-domain outputs.

Dataset	Pixel cosine	Matched NMSE ↓	Shuffled NMSE	Gap [95% CI]	Node win	State margin	Branch cosine
CUB-200-2011	−0.135	0.433	0.719	0.286 [0.239, 0.336]	0.800	0.343	0.334
Oxford-IIIT Pets	−0.037	0.443	0.636	0.194 [0.155, 0.235]	0.706	0.292	0.216

Source-geometry intervention. For an evaluation pair (A, B), the matched inverse used terminal target H_A and source-local geometry J_A, whereas the geometry-swapped inverse retained H_A but replaced the operating-point geometry by J_B. The primary path test held the reconstructed states from (H_A, J_A) fixed and evaluated their JVP forward consistency on either the matched graph J_A or the shuffled graph J_B. At every probed node, path NMSE is the normalised squared mismatch between the JVP-predicted child state and the reconstructed child state, averaged over the relevant children; lower values indicate greater compatibility with that graph's local differential path. This intervention does not require the geometry-swapped output to become the image or feature identity of B: J_B specifies the complete local input-to-feature geometry, while the fixed terminal state H_A continues to modulate the selected inverse branch.

Results. The lowest pixel-cosine quartile had mean pixel cosine -0.135 on CUB and -0.037 on Pets, yet matched-path NMSE remained substantially below shuffled-path NMSE on both datasets

(table 10). The 95% bootstrap intervals for the paired mean gaps excluded zero, and the matched source-state margins remained positive. The effect was strongest at deeper states: matched versus shuffled path NMSE was 0.370 versus 0.986 on CUB and 0.411 versus 0.874 on Pets. Changing only the source-local geometry also selected substantially different input-domain branches: in the lowest quartile, matched-versus-swapped output cosine was 0.334 on CUB and 0.216 on Pets, with relative ℓ_2 distances of 0.980 and 1.120.

Pixel cosine was nearly uncorrelated with the primary path diagnostics over all 512 images. Its Spearman correlation with matched state margin was 0.012 on CUB and 0.051 on Pets; with the log shuffled-to-matched path-NMSE ratio it was -0.049 and -0.035; and with node-win fraction it was -0.042 and 0.034. Thus poor input-instance alignment did not identify samples with weak source-path consistency.

Boundary of the claim. The comparison supports path-level faithfulness, not a claim that every local boundary must favour the matched graph. Six of the seven probed nodes had lower mean NMSE under J_A on both datasets, whereas the early node 4 showed a systematic reversal. Pets therefore had a small shallow-state reversal (matched/shuffled NMSE 0.466/0.458) despite a large deep-state advantage. In addition, geometry-swapped deep states remained modulated by the fixed terminal target rather than globally changing into the states of B. These observations are consistent with the intended object: the source-local geometry selects an inverse branch, while the terminal feature modulates the state transported through that branch. The experiment therefore shows that low pixel cosine alone is not a sufficient failure diagnostic. The low-alignment ConvNeXt outputs retain measurable fidelity to the target-generating source path.

C.7 Topology, Two-Stage Mechanism, and Operator Ablations

Figure 13 presents the complete fixed-example mechanism plate; table 11 gives the matched topology, stage, and operator results. The sections below connect their controls to the local residual, state, and output measurements and then expand the diagnostics.

Local topology versus a global repair. On the matched ResNet18 terminal target, raw VJP has pixel cosine 0.00019. A single global G reaches 0.07221, and a global $G + D$ reaches 0.06714. Local G_v reaches 0.90090, while local $G_v + D_v$ reaches 0.93795. The comparison holds operator calibration and target data fixed while changing whether repair is factorised along the fitted DAG. It therefore isolates local topological composition from the mere existence of a closed-form map.

Second-stage residual repair. On ResNet18, G_v alone reaches a pixel cosine of 0.9009 and the second stage adds 0.0367. The corresponding gain increases to 0.1107 on ConvNeXt-B and 0.3905 on ViT-B/32. Mean child NMSE decreases by 23.0%, 15.4%, and 16.2%, while intermediate-state relative ℓ_2 decreases by 1.52%, 0.45%, and 5.92%. These matched G_v-only/Final comparisons measure the second-stage contribution within each architecture. The larger output gains on ConvNeXt-B and ViT-B/32 highlight the role of source-specific repair. For sample-dependent operators such as normalisation and attention, JVP-FC uses the actual source differential to estimate the state error left by G_v. The measured gains span local consistency, upstream-state estimation, and end-to-end alignment. Boundary-level correlations are not uniformly positive. Across 86 pooled boundaries, the Spearman correlation between the pre-correction residual and state-error reduction is 0.276 with bootstrap interval $[0.068, 0.462]$; its correlation with residual reduction is -0.235 with interval $[-0.462, 0.004]$. Residual magnitude was positively associated with state-error reduction; its association with residual reduction was unresolved from zero.

Operator family. The complete matched operator comparison is reported in table 12. Full channel mixing and frequency specificity are separately required for strong input-instance alignment at the representative calibration budget.

Rank-8 recovers much of the ResNet result but not the ViT result. This architecture dependence becomes stronger at small calibration budgets. Repeated ResNet18 seeds reproduce the Gamma-only result near 0.006, excluding an accidental calibration subset as the explanation. The visual outputs are retained alongside these scalar values because restricted operators can retain coarse spatial structure while differing in scale, colour, or frequency.

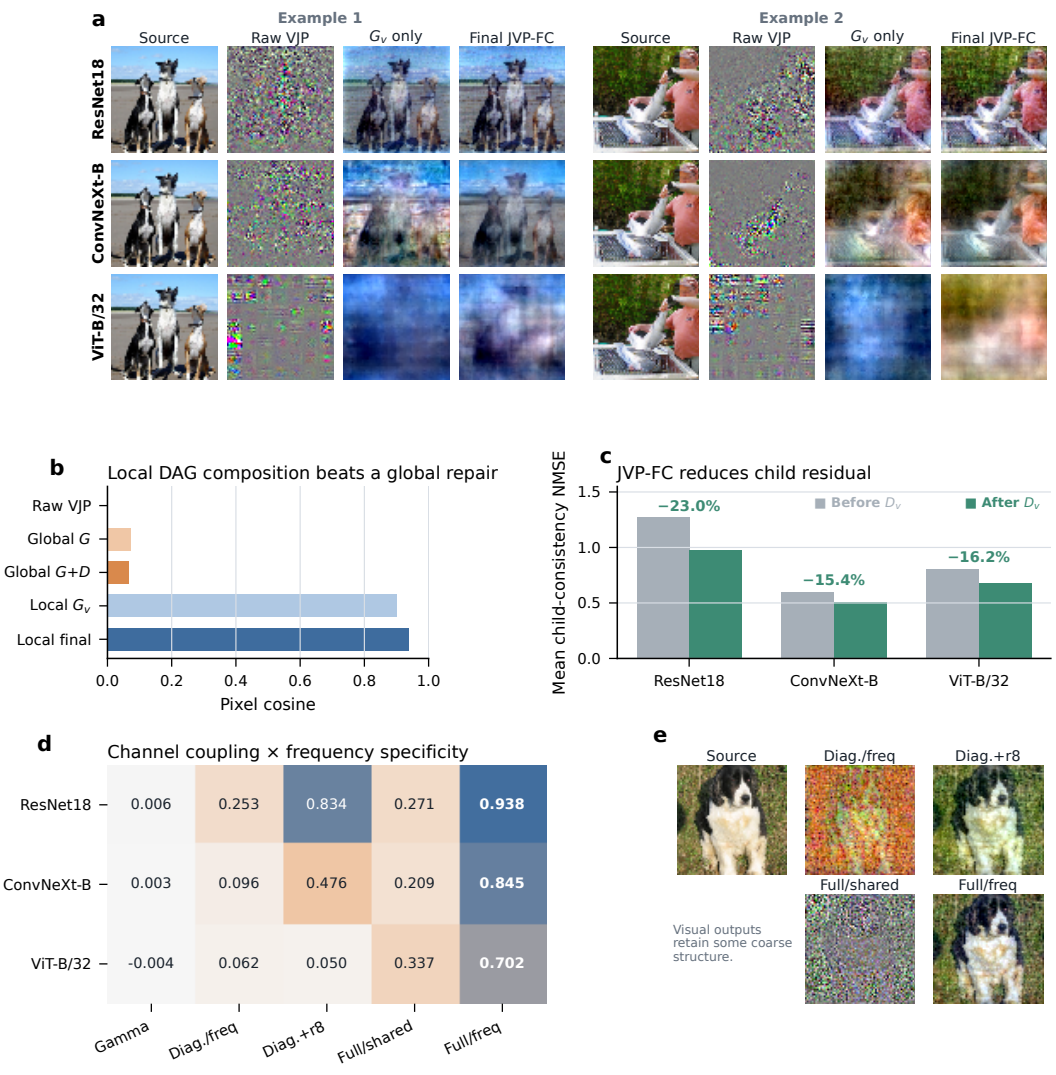


Figure 13: **Why topology-respecting two-stage repair works. a**, Raw VJP, distributional G_v-only repair, and final JVP-FC outputs on two predeclared examples per architecture. **b**, Global single-boundary repair versus local DAG composition. **c**, Child-consistency residuals before and after JVP-FC. **d**, Restricted operator families across three architectures. **e**, Matched visual examples showing why scalar image metrics must be interpreted together with spatial and spectral structure.

Frequency alignment control. Replacing the per-frequency maps with one coordinate-collapsed map reduces pixel cosine from 0.9375 to 0.2709 on ResNet18 and from 0.7021 to 0.3366 on ViT-B/32. Permuting otherwise fitted frequency maps reduces the values further to 0.00025 and 0.0288. The result shows that map capacity alone is insufficient; the learned second-order operator must remain aligned with the coordinate frequency on which it was calibrated.

Mean-seed gauge. The seed-source ablation includes heterogeneous-channel, multi-child nodes. On ResNet18, the full-state mean seed and an unweighted constant seed give numerically identical final pixel cosine, 0.9740235 in both cases. This is expected because a constant full-state rescaling is absorbed when separate zero-intercept relative-ridge maps are refitted. The mean gauge is retained because it defines the reusable terminal-subset API $s_T = P_{S,Q_{\mathrm{all}}} H_T / |S|$ across channel masks, where Q_{all} denotes the complete coordinate set. It is not claimed to improve full-state quality over an arbitrary constant gauge. Unit-energy seeding is nonfinite on the tested ResNet configuration, while a fixed uniform seed produces zero output, confirming that the source must remain target dependent.

Table 11: **Matched topology, stage, and operator ablations. a**, The topology control is a matched ResNet18 experiment. **b,c**, Stage and operator comparisons use matched c4096/e1024 runs across the three representative architectures. Values are pixel cosine unless marked otherwise. Restricted operators measure which input-aligned structures their operator class retains; low values do not define feature-inversion invalidity.

a Topology control (ResNet18)

Variant	Pixel	Computation
Raw VJP	0.0002	adjoint transport
Global G	0.0722	one terminal map
Global $G+D$	0.0671	one terminal map
Local G_v	0.9009	boundary DAG
Local final	**0.9380**	boundary DAG

b Two-stage correction

Model	G_v	Final	Δ pixel	Child NMSE red.
ResNet18	0.9009	**0.9375**	0.0367	23.0%
ConvNeXt-B	0.7343	**0.8450**	0.1107	15.4%
ViT-B/32	0.3116	**0.7021**	0.3905	16.2%

c Operator family

Operator	ResNet18	ConvNeXt-B	ViT-B/32
Gamma-only	0.0062	0.0034	−0.0044
Diagonal/per-frequency	0.2526	0.0961	0.0620
Diagonal+rank-8	0.8337	0.4763	0.0501
Full/shared-frequency	0.2709	0.2090	0.3366
Full/per-frequency	**0.9375**	**0.8450**	**0.7021**

Table 12: **Restricted operator families.** Entries are final pixel cosine at c4096/e1024. Restricted rows remain valid feature-inverse computations within their operator class; the values measure the input-aligned structures retained by that class.

Operator	ResNet18	ConvNeXt-B	ViT-B/32
Gamma-only	0.0062	0.0034	−0.0044
Diagonal, per-frequency	0.2526	0.0961	0.0620
Diagonal + rank-8, per-frequency	0.8337	0.4763	0.0501
Full-channel, shared-frequency	0.2709	0.2090	0.3366
Full-channel, per-frequency	**0.9375**	**0.8450**	**0.7021**

Nested-target composition. Restarting the same calibrated inverse at an intermediate repaired state and continuing upstream to the input-domain output is numerically identical to the original deep-target run: output cosine is at least 0.999999996 on ResNet18 and ViT-B/32. Directly beginning at the true forward state of the shallower target gives 0.9955 output cosine to the original deep-target run on ResNet18 and 0.8689 on ViT-B/32. Thus the implementation is compositionally consistent when the same repaired state is reused, while changing the declared target changes the feature-inverse object.

C.8 Metric-Aware Theory Bridge and Finite-DAG Stability Diagnostics

Metric-aware local oracle. We evaluated the natural-adjoint construction on held-out local perturbations at 16 operating points. In a tractable projected problem, the local LMMSE operator and the metric-aware natural adjoint agree to numerical precision. Both reach mean NMSE 0.0291, cosine 0.9858, and relative ℓ_2 0.1590. The raw Euclidean adjoint reaches NMSE 0.9999 and relative ℓ_2 1.0000. A calibrated scalar rescaling improves these values to 0.5226 and 0.7191 but does not close the metric gap. The deployed observable-plus-G surrogate reaches NMSE 0.4416, cosine 0.7682, and relative ℓ_2 0.6572.

This experiment has a deliberately limited role. It verifies the algebraic bridge between a covariance-induced natural adjoint and a local LMMSE inverse in a tractable projected setting. It does not claim that the deployed distributional surrogate estimates M_u and M_v online or equals the local oracle at every operating point.

Finite-DAG diagnostics. For each fitted boundary we recorded path depth, a conservative product-of-gain proxy, local condition statistics, child NMSE, intermediate-state error, and final pixel error. The proxy spans several orders of magnitude and describes the scale of the fitted reverse gains.

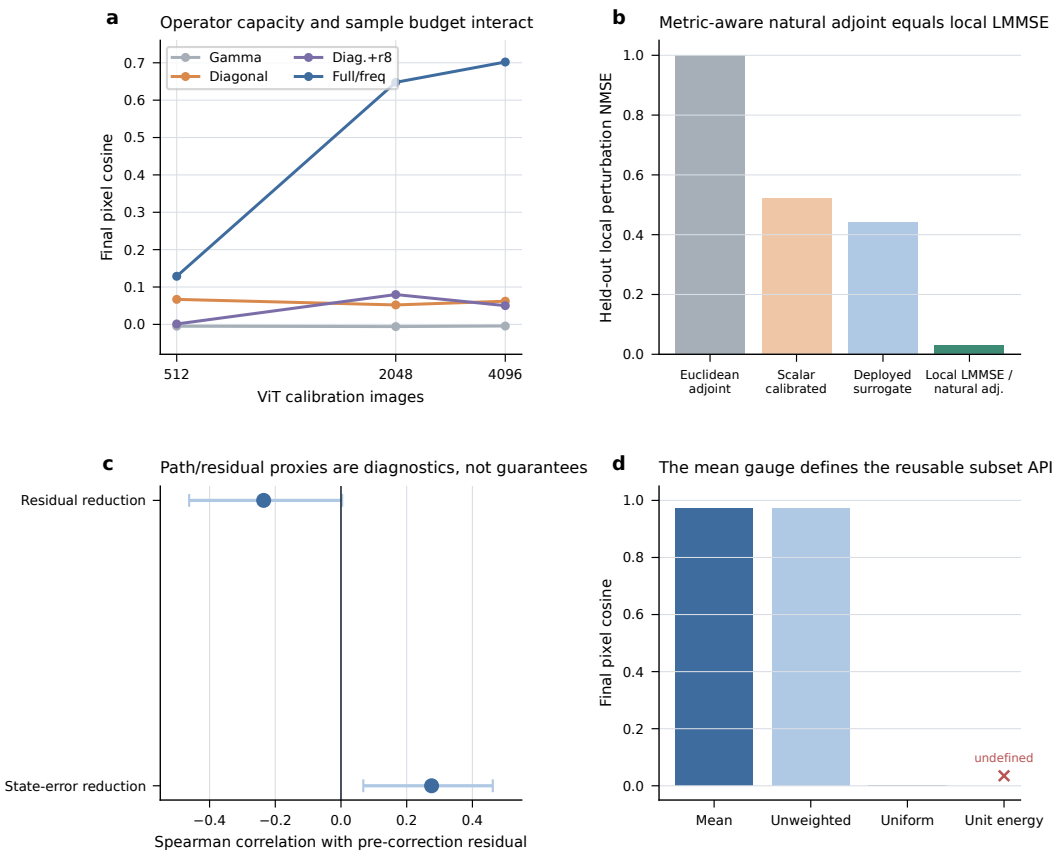


Figure 14: **Operator capacity, theory bridge, and stability diagnostics. a**, ViT operator-capacity by calibration-budget frontier. **b**, Held-out local perturbation NMSE for Euclidean, deployed, and metric-aware oracle operators. **c**, Pooled boundary correlations with 95% bootstrap intervals. **d**, Multi-child seed-source ablation.

However, it does not reliably rank final error across all nodes or architectures. The pooled Spearman correlation between pre-correction residual and state-error reduction is 0.276, with a 95% bootstrap interval of $[0.068, 0.462]$. Consequently, the empirical path-gain study is reported as a stability diagnostic, not as a contraction certificate or proof of global convergence. The formal finite-DAG result in section B.6 remains a conditional error-propagation bound.

Figure 14 places the operator-capacity frontier, metric-aware oracle, boundary correlations, and seed-source controls side by side.

C.9 Input Alignment, Representation Consistency, and Information-Flow Controls

Figure 15 shows the fixed-pair swap, continuous operating-point, shuffle, channel-diversity, and zero-target panels. Their quantitative protocols are separated below.

Zero and shuffle controls. The complete controls are summarised in table 13. All zero targets produce exactly zero output, so the zero-intercept maps cannot inject a target-independent additive image template. Source shuffling moves the output toward the new operating point. Terminal-state shuffling produces architecture-dependent modulation, which separates the role of the target state from that of the source-local operators.

Target–operator swap. For paired images A and B, we cross H_A, H_B with operator families J_A, J_B. ResNet18 conditions H_A, J_B and H_B, J_A align with the operator-point images at 0.9386

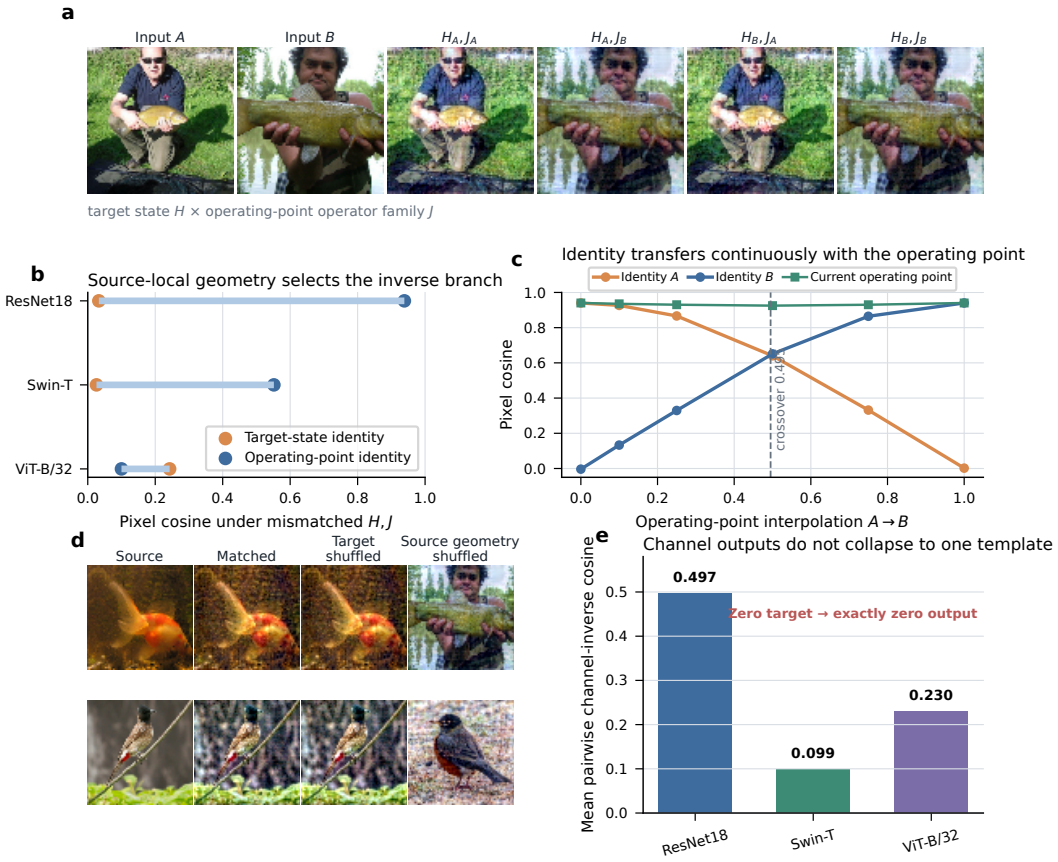


Figure 15: **What determines a feature inverse.** **a**, Target–operator swap for one fixed pair. **b**, Similarity to target-state and operating-point identities under mismatched conditions. **c**, Continuous operating-point interpolation with a fixed target. **d**, Matched, target-shuffled, and source-geometry-shuffled controls. **e**, Channel diversity and the exact zero-target result. The controls show that source-local geometry selects the inverse branch, while the terminal state modulates that branch in an architecture-dependent manner.

Table 13: **Information-flow controls.** All rows use the matched c4096/e512 target/operator-control protocol. "Same" is the matched target and operating point. Terminal shuffle reports alignment to the fixed operating-point input and to the shuffled target image. Source shuffle reports alignment to the old input and new source.

Model	Same	Zero norm ratio	Term. → input	Term. → target	Source → old	Source → new
ResNet18	0.9380	0	0.9379	0.0225	0.0186	0.9380
Swin-T	0.3890	0	0.3492	0.0312	0.0096	0.5056
ViT-B/32	0.6922	0	0.0995	0.2494	0.0158	0.6922

and 0.9386, while alignment with the target-state images is 0.0340 and 0.0316. Swin shows the same direction at 0.5509 and 0.5526. ViT has stronger target modulation: the mismatched conditions have paired identity values 0.2495/0.0959 and 0.1040/0.2358. These results show that source-local geometry selects the inverse branch while the selected terminal state modulates that branch.

Continuous operating-point locality. On 128 fixed ResNet18 image pairs, the operating point is interpolated from A to B while the terminal target remains H_A. Alignment to A decreases monotonically from 0.9402 to 0.0021, while alignment to B increases from -0.0036 to 0.9402. The two curves cross at interpolation coefficient 0.495. At every point, alignment to the current operating-point image remains between 0.9251 and 0.9402. By contrast, same-image noise with

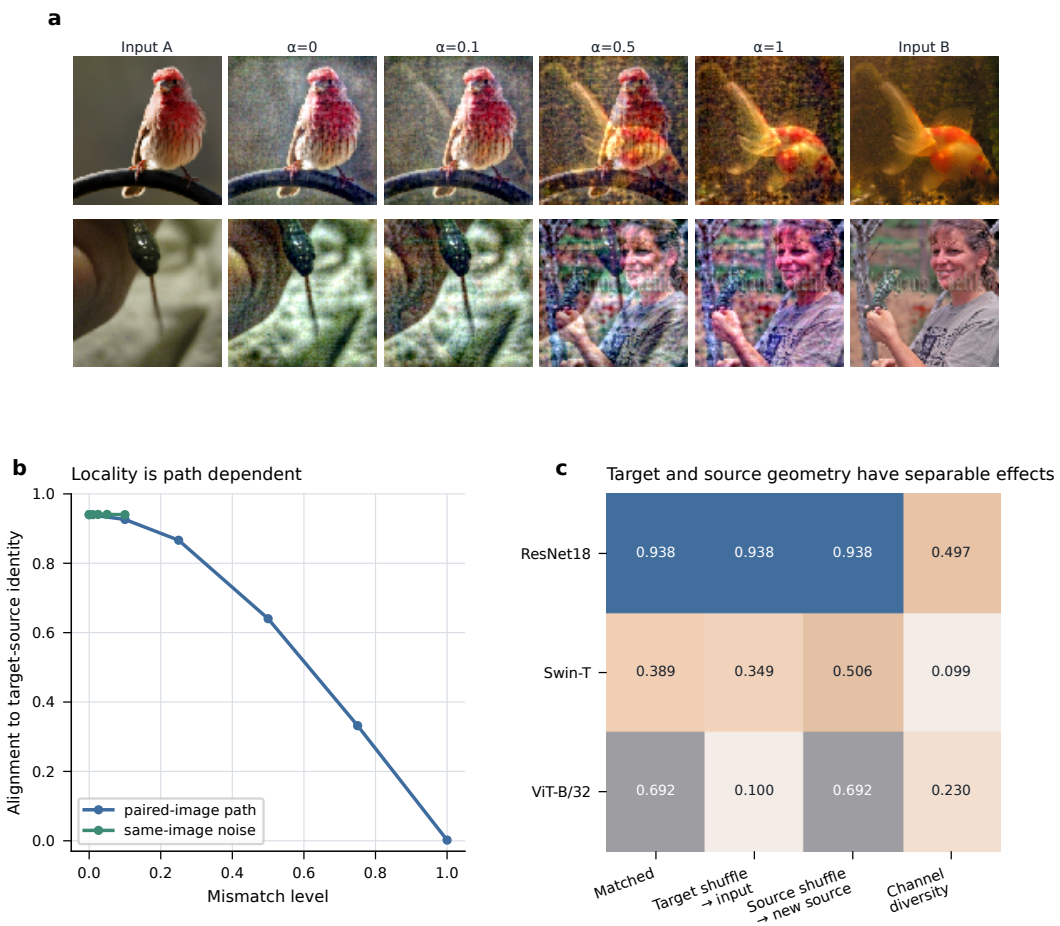


Figure 16: **Expanded operating-point locality and information-flow controls. a**, Fixed-target inverses along two paired-image operating-point paths. **b**, Paired-image interpolation versus same-image noise. **c**, Matched, shuffle, and channel-diversity summaries across three architectures.

relative magnitude up to 0.1 changes alignment to the clean source by at most 0.00066. The curve therefore characterises the local branch selected by the operating point. Its child residual does not increase monotonically with mismatch and must not be used as a source-match detector.

Channel diversity and representation checks. Mean pairwise pixel cosine among three fixed channel inverses is 0.4969 for ResNet18, 0.2304 for ViT-B/32, and 0.0993 for Swin-T. The outputs therefore do not collapse to one map-dependent template. Re-encoding is reported against architecture-specific random-pair nulls and remains separate from direct input alignment. The ROC AUC distinguishing matched from random-pair re-encoding is 1.000, 0.996, and 0.947 for ResNet18, Swin-T, and ViT-B/32, respectively. A high re-encoding value establishes representation consistency, while a low value does not prove that the selected feature inverse is false.

Failure taxonomy. The fixed 1,024-image evaluation sets are stratified by low-frequency structure, weak high-frequency alignment, and chroma drift. ConvNeXt and Swin low-alignment cases are often concentrated in colour or fine texture, whereas deep ViT outputs are dominated more strongly by low-frequency and patch-grid structure. These categories describe architecture-dependent feature-inverse geometry. Mismatch with the complete input is not, by itself, treated as feature-inversion failure.

Figure 16 expands the paired-image paths, same-image noise control, shuffle conditions, and channel-diversity summaries.

C.10 Calibration, Boundary, Ridge, Online Runtime, and Memory

Calibration size. The paired vision-Transformer sweep uses calibration sizes $\{512, 1024, 2048, 4096, 8192\}$. ViT-B/32 final pixel cosine is 0.1287, 0.3626, 0.6477, 0.7021, and 0.7243. Swin-T reaches 0.0467, 0.2901, 0.4575, 0.5037, and 0.5259. The largest increase occurs from 1,024 to 2,048 images for ViT-B/32 and from 512 to 1,024 for Swin-T. The 4,096-to-8,192 gain is approximately 0.022 for both models. This supports c4096 as a diminishing-return operating point, not as a universal sample optimum.

Independent c4096 calibration subsets yield pixel-cosine standard deviations of 0.00030 for ResNet18, 0.00362 for ConvNeXt-B, and 0.00464 for ViT-B/32. The corresponding SSIM standard deviations are 0.00178, 0.00060, and 0.00122. Thus the reported architecture differences are larger than the observed three-subset calibration variation.

Boundary density. At c512/e1024, the ResNet18 coarse boundary uses 9 fitted nodes and reaches pixel cosine 0.9293. The full-DAG mode uses 49 fitted nodes and reaches 0.9631. SSIM increases from 0.6747 to 0.7913 and LPIPS decreases from 0.3255 to 0.1664, while post-calibration online application increases from 53.9 to 74.7 ms/image. This diagnostic shows that finer boundaries can improve the inverse while raising online application, map, and topology-onboarding demands. The coarse policy provides a practical balance between alignment and deployment cost.

Ridge sensitivity. At c512, ResNet18 is stable over $\rho \in \{0.003, 0.01, 0.03\}$, with final pixel cosines 0.9247, 0.9278, and 0.9293. At c1024, ConvNeXt-B reaches 0.7233, 0.7584, and 0.7674. ViT-B/32 is more sensitive at the same small calibration budget, increasing from 0.0726 to 0.3124 and 0.5258. This interaction explains why low-capacity or low-calibration token-model results should not be used to claim universal operator compression.

Representative deployment comparison. Table 14 reports the four representative models, including the reciprocal throughput readout. Post-calibration application ranges from 58.4 ms/image on ResNet18 to 369.7 ms/image on ConvNeXt-B. The latter uses 37 fitted nodes and 15.51 GiB of maps, compared with 9 nodes and 0.83 GiB for ResNet18. These values describe each architecture's deployed reverse pass; hardware-specific calibration engineering is outside this timing comparison.

Table 14: **Representative post-calibration online deployment costs.** All online values are measured after calibration on the c4096/e1024 quality runs. Online time is the measured 1,024-image application total divided by 1,024 and is therefore not isolated batch-one latency. Throughput is its reciprocal. Map storage includes both G_v and D_v; evaluation peak is the measured peak allocated memory. These values describe the current public implementation rather than an optimised-kernel lower bound.

Model	Fitted nodes	Online (ms/image)	Throughput (images/s)	Map storage (GiB)	Eval peak (GiB)
ResNet18	9	58.4	17.1	0.83	1.71
ConvNeXt-B	37	369.7	2.70	15.51	16.78
Swin-T	26	129.0	7.75	5.98	6.56
ViT-B/32	14	105.1	9.52	2.99	3.61

Online runtime, storage, and memory. The expanded resource audit is shown in table 15. Online times correspond to post-calibration application in c4096/e1024 quality runs. Storage includes both stages.

Full-channel maps make calibration demand, storage, and online application architecture dependent. The practical benefit is therefore a change from repeated per-target search to direct queries after calibration, not zero total computation or memory.

C.11 Deployment Validity, Oracle Gaps, and Implementation Quality Control

Public-path equivalence. Across all nine architecture implementations, the public inversion call and the directly instrumented reference call agree to pixel cosine 0.99999997 or higher. Maximum absolute differences are zero or below 6.2×10^{-7}. This audit verifies that the reported output is produced by the public API rather than a private visualisation branch.

Table 15: **Expanded resource audit.** Online time is the measured 1,024-image application total divided by 1,024 after calibration; it is not isolated batch-one latency. Storage includes both G_v and D_v. Peak memory is GiB.

Model	Nodes	Online (ms/image)	Storage (GiB)	Calibration peak (GiB)	Eval peak (GiB)
ResNet18	9	58.4	0.83	1.24	1.71
ResNet50	17	223.7	25.42	19.44	17.09
DenseNet121	9	104.1	5.84	8.79	7.42
MobileNetV2	31	155.0	4.35	6.46	5.39
EfficientNet-B0	30	164.0	5.43	7.63	6.79
RegNetY-1.6GF	28	114.2	4.29	4.79	5.21
ConvNeXt-B	37	369.7	15.51	16.69	16.78
Swin-T	26	129.0	5.98	6.57	6.56
ViT-B/32	14	105.1	2.99	3.79	3.61

Legal online child state. At deployment, each D_v source uses the repaired online child state, not a cached true forward activation. The same-map true-child condition is retained only as a nondeployable c4096/e512 diagnostic upper reference. On ResNet18, replacing the online child with the true child changes pixel cosine from 0.9382 to 0.9391 and target re-encoding from 0.6522 to 0.7570. On ViT-B/32, the same diagnostic changes pixel cosine from 0.7005 to 0.9410 and re-encoding from 0.2481 to 0.7289. The larger ViT gap identifies online child estimation as an important future improvement direction.

Zero intercept, scaling, and superposition. Zero targets produce zero outputs in every information-flow audit. For fixed operating-point operators, scaling the target by $\alpha \in \{0, 0.25, 0.5, 1, 2\}$ scales the output with cosine at numerical unity and relative errors at or below 1.7×10^{-7} for ResNet18; ViT scaling is exact to recorded precision. Cross-target superposition has output cosine 0.999999996 for both models. Weighted recomposition of a complete target from channel partitions gives cosine 0.99999 for ResNet18 and 0.99950 for ViT. After weighting for the public subset mean gauge, the remaining recomposition residual reflects finite-precision application, not an additive image template.

Normalisation and residual gauge checks. The final path applies no source, activation, or cotangent pre-normalisation. The D_v source is the pulled-back local child residual in the same gauge used by the online JVP, while its regression target is $H_v - y_v^0$. Static checks validate the JVP direction, residual sign, stored map layout, and zero-intercept application. Runs that use true children, alternative correction weights, or nondefault internal boundaries are labelled diagnostic-only in their artifacts and are excluded from headline tables.

With the legal online path and its practical headroom fixed, the next section expands the decision-grounded interpretation study without changing the fitted maps or the inversion API.

Figure 17 summarises the token calibration curves, ridge sensitivity, storage demand, and online-child oracle gap.

C.12 Expanded Prediction-Conditioned Feature Atlas

Protocol. The decision-conditioned experiment uses CUB-trained ResNet18, ConvNeXt-B, and ViT-B/32 checkpoints with top-1 accuracies 70.59%, 85.81%, and 70.16%. For each model, 16 correctly classified test images are fixed before feature selection. One c4096 map family is reused across four depths, all images, and all selected channel groups. The depths are 3/5/7/9 for ResNet18, 4/7/34/37 for ConvNeXt-B, and 2/6/10/14 for ViT-B/32.

For the original predicted class $\hat{y}$ and runner-up class y_2, channels are ranked by their activation–gradient contribution to the decision margin (Ancona et al., 2018),

$$s_c = \left\langle H_{l,c}, \frac{\partial(z_{\hat{y}} - z_{y_2})}{\partial H_{l,c}} \right\rangle . \tag{52}$$

Nested groups of $k \in \{1, 4, 8, 16\}$ top-positive, top-negative, activation-matched, and 20-repeat random-matched channels are selected. Each group is inverted through the legal subset gauge $P_{S,Q_{\mathrm{all}}} H_l / |S|$. In a parallel branch, the same channels are replaced in the original internal represen-

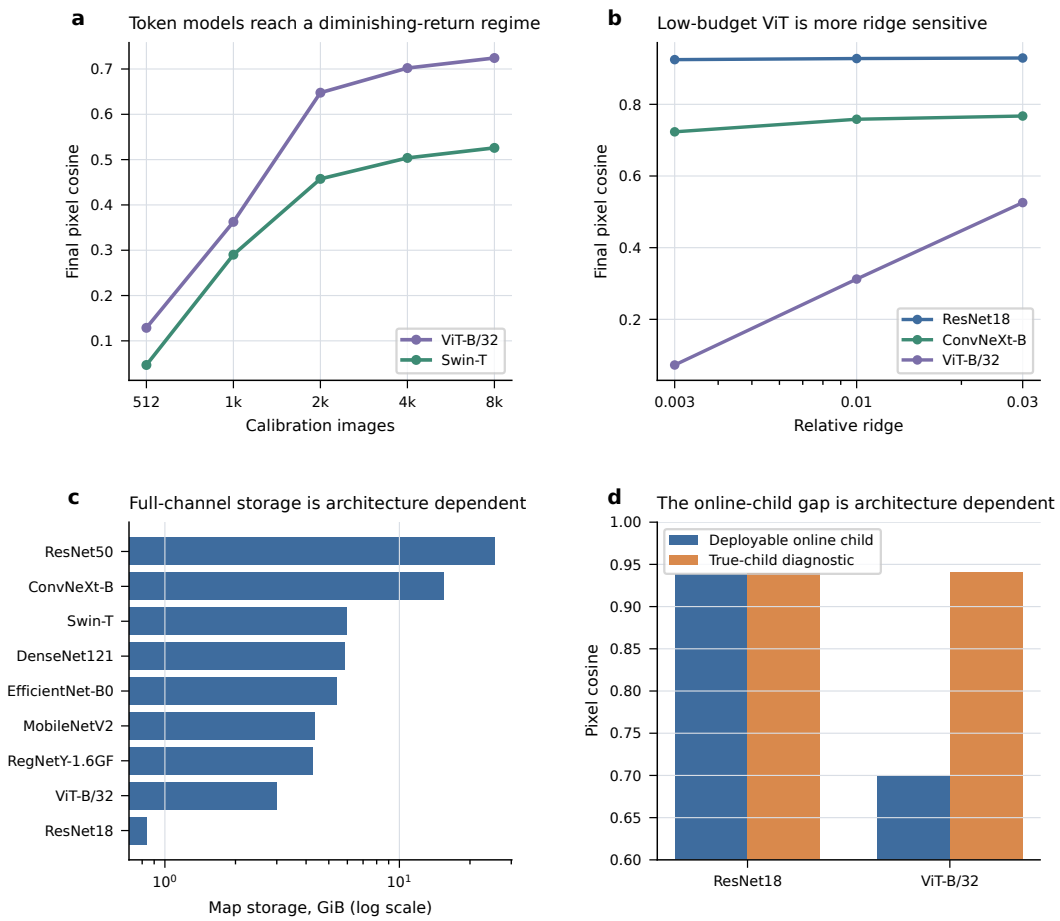


Figure 17: **Calibration and deployment diagnostics. a**, Paired Swin-T and ViT-B/32 calibration curves. **b**, Ridge sensitivity at the stated small-calibration budgets. **c**, Full $G_v + D_v$ storage across architecture families. **d**, Deployable online-child versus nondeployable true-child diagnostic.

tation. Calibration-mean replacement is primary; zero replacement is a robustness protocol. The feature inverse is never re-input as a causal stimulus.

The intervention is followed by an independent forward pass, retaining the original class pair. Margin drop is the margin before intervention minus the margin after intervention.

Figure 18 shows both predeclared examples, all model–depth–group-size contrasts, depth-wise decision-margin deletion AUC, the signed validation gate, and the paired evidence branches. The deletion AUC panel summarises the decision-margin deletion response across depths.

Cross-architecture gate. The strict audit covers 36,864 intervention rows, 3,072 group inversions, and 1,536 deletion curves. Under calibration-mean replacement, the top-positive margin drop exceeds the within-image random-matched mean in all 48 model–layer–k settings. Every image-cluster bootstrap interval is above zero. Zero replacement gives the same 48/48 result. All 12 terminal top-negative controls have intervals below zero, demonstrating the expected opposite signed effect. Table 16 summarises the signed gate for each architecture.

Normalised positive-minus-random effects are positive at every tested depth and group size. They range from 0.122 to 2.253 on ResNet18, 0.201 to 9.559 on ConvNeXt-B, and 0.068 to 1.805 on ViT-B/32. The effect scale differs across architectures, so the claim is based on within-image matched contrasts rather than direct comparison of raw logit units.

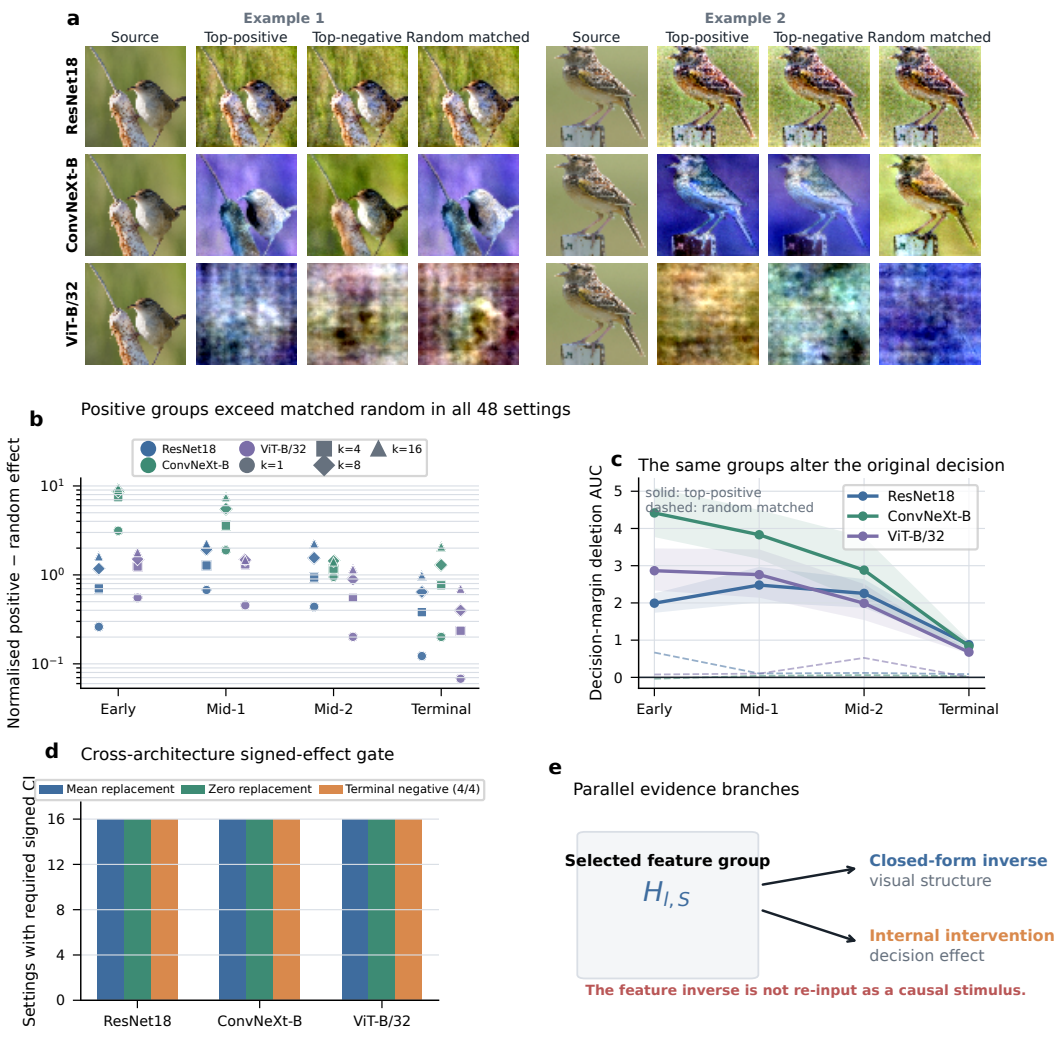


Figure 18: **Decision-grounded feature atlases across architectures. a**, Two predeclared source examples with top-positive, top-negative, and random-matched feature inverses for ResNet18, ConvNeXt-B, and ViT-B/32. **b**, Normalised positive-minus-random effects over all model–depth–group-size settings. **c**, Decision-margin deletion AUC across depth. **d**, Signed cross-architecture validation gate. **e**, Parallel inversion and internal-intervention evidence branches. The feature inverse is not re-input as a causal stimulus; causal evidence comes from intervention on the original internal representation.

Table 16: **Decision-conditioned cross-architecture gate.** Counts give the number of settings whose bootstrap interval has the required sign. The final column is the minimum–maximum prediction-flip fraction across the four tested depths at $k = 16$ under calibration-mean replacement.

Model	Primary > 0	Zero > 0	Terminal negative < 0	$k = 16$ flip range
ResNet18	16/16	16/16	4/4	0.313–0.813
ConvNeXt-B	16/16	16/16	4/4	0.125–0.750
ViT-B/32	16/16	16/16	4/4	0.188–1.000

Architecture-dependent inverse geometry. ResNet18 and ConvNeXt-B develop stronger deep common-mode inverse structure, while positive and negative groups have opposing decision effects. ViT-B/32 retains a patch-grid, low-frequency, globally mixed geometry. The shared result across architectures is the selection–inversion–intervention chain, not identical visible structure. CUB

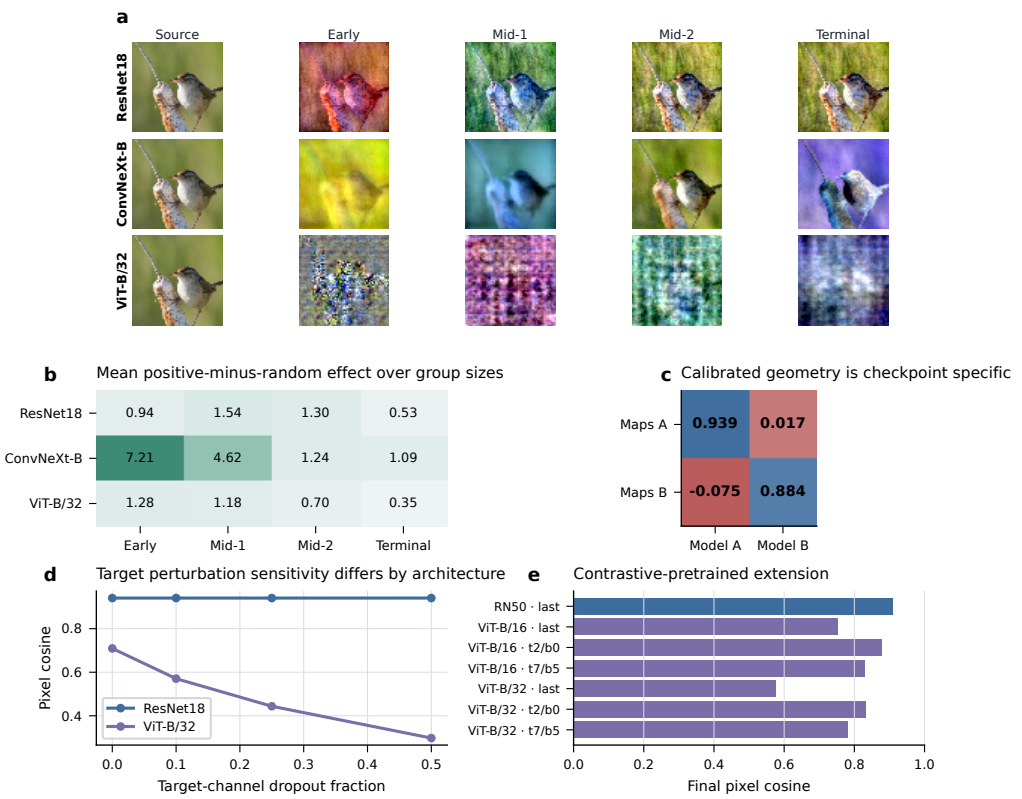


Figure 19: **Expanded decision-conditioned atlas and optional extensions.** **a**, One fixed CUB image across four depths and three architectures. **b**, Mean positive-minus-random intervention effect over group sizes. **c**, Same-architecture checkpoint/map mismatch. **d**, Channel-dropout sensitivity. **e**, Contrastive-pretrained CLIP extensions across architecture and depth.

segmentation and visible-part overlap remain secondary diagnostics and are not used to support the cross-architecture gate. They do not establish a unique semantic name for every channel group. Likewise, the CNN common-mode observation is consistent with prior holographic-representation evidence (Shu, 2026), but does not establish destructive interference without a dedicated signed cancellation experiment.

C.13 Optional Generalisation and Application Extensions

These experiments test additional boundaries after the core claim is already established. They are not required for the main paradigm comparison; Figure 19 summarises the decision-conditioned, checkpoint, perturbation, and CLIP evidence.

Target perturbations. ResNet18 is nearly invariant to relative Gaussian target noise up to 0.1 and to 8- or 16-bit quantisation. Its inverse remains at pixel cosine 0.9406, and output cosine to the clean inverse remains above 0.99999998 for Gaussian noise. ViT-B/32 remains close under Gaussian noise and 8-bit quantisation, but is more sensitive to channel dropout. Dropping 50% of target channels reduces pixel cosine from 0.7094 to 0.2977 and output cosine to the clean inverse to 0.4083. This contrast is consistent with the stronger channel-coupling demand observed in the ViT operator ablation.

Checkpoint specificity. Maps fitted on one ResNet18 checkpoint transfer poorly to a different same-architecture checkpoint. Matched map/model conditions reach pixel cosines 0.9388 and 0.8842, whereas mismatched conditions reach 0.0165 and -0.0754. The result confirms that calibrated maps encode checkpoint-specific local geometry rather than an architecture-level image decoder.

Blind architecture onboarding. Automatic MLP-Mixer onboarding (Tolstikhin et al., 2021) captures only the stem projection rather than the intended final Mixer norm. The captured stem target itself is inverted successfully, with final pixel cosine 0.9470, SSIM 0.8299, and LPIPS 0.0991. The experiment is therefore a verified boundary-discovery failure, not a closed-form inversion failure. It defines a concrete engineering limitation: new architectures require the boundary capture policy to expose the intended target and reverse DAG.

Contrastive-pretrained models. The same estimator and tensor-axis rule extend to CLIP-RN50 and CLIP ViTs (Radford et al., 2021). CLIP-RN50 reaches final pixel cosine 0.9102 at its last target. CLIP-ViT-B/16 reaches 0.7519 at the last target and 0.8799 at an early block. For CLIP-ViT-B/32, the best last-target result is 0.5774 with a finer attention/MLP boundary, while early and middle targets reach 0.8338 and 0.7829. Overly fine token-internal splits collapse, which reinforces the practical importance of boundary selection. These results broaden the architecture and pretraining evidence but remain supplement-only because they do not change the main claim.

C.14 Language-Transformer Protocol and Full Evidence

Controlled benchmark. The formal language experiment uses pretrained GPT-2 small and the exact instantiation in section A.12.

Inputs have length $L = 16$, hidden width 768 as the channel axis, and token positions as coordinates. The formal target is the layer-6 attention-module output before residual addition; it is distinct from the residual stream and attention-logit/weight matrices. The all-channel terminal query and online mean seed are

$$S = \{1, \dots, 768\}, \qquad Q = \{L-1\}, \qquad y_T = P_{S,Q} H_T, \qquad s_T^{\mathrm{on}} = P_{S,Q} H_T / 768. \tag{53}$$

The reverse boundaries are the input embeddings, block-2 and block-4 outputs, and the layer-6 attention output. Both stages use zero-intercept full-channel maps with all-shared gains over nine stored rFFT bins. Calibration uses the complete sequence tensor; coordinate masks are applied only when querying the fixed maps.

For inverse embedding $\widehat{e}_p(q)$ at position q from terminal query p, the token-position profile is

$$\mathcal{E}_{p \to q} = \frac{\|\widehat{e}_p(q)\|_2^2}{\sum_{r=0}^{L-1} \|\widehat{e}_p(r)\|_2^2}. \tag{54}$$

This *inverse energy* is measured after the complete reverse path; it is neither an attention weight nor a decoded-token probability. Each map family was calibrated on 4,096 fixed-seed WikiText-2 windows (Merity et al., 2016), and the formal evaluation averaged independently fitted seed-123 and seed-321 outputs before statistical resampling. The formal reverse path uses no per-example optimisation or refitting, learned text decoder, or autoregressive decoding. The semantic benchmark contained 72 contexts: 12 polysemous words, two senses per word, and three contexts per sense. Each context was deterministically padded and cropped to 16 GPT-2 tokens while keeping the target word at the terminal position. Each context contained three to five semantic cues specified before evaluation. For every context, 64 random subsets matched both cue count and the empirical cue-position distribution while excluding cue and target positions.

The syntactic benchmark contained 80 contexts from 20 lexical templates. Within each template, singular/plural heads and grammatical/ungrammatical auxiliaries were crossed while the nearer attractor noun carried the opposite number. All four variants shared the same token-position roles: head, attractor, auxiliary, and terminal predicate. The agreement site is the auxiliary token *was* or *were* that explicitly carries the subject–verb number agreement; it is distinct from the head noun, the nearer attractor, and the terminal predicate queried for inversion. The four variants therefore cross singular head/plural attractor with *was* (grammatical) or *were* (ungrammatical), and plural head/singular attractor with *were* (grammatical) or *was* (ungrammatical). The analysis asks whether the inverse routes through this agreement site rather than the nearer attractor; it does not interpret the energy difference as a probability of grammaticality.

Statistical resampling. The benchmark intervals used 10,000 percentile cluster-bootstrap resamples after averaging the two calibration-map outputs. Polysemy analyses resampled the 12 target-word

Table 17: **Controlled token-routing summary averaged over two c4096 calibrations.** Cue-minus-random uses all 72 semantic contexts. Agreement-to-attractor uses 40 grammatical and 40 ungrammatical contexts. Ratios are descriptive; inference on syntax uses the cluster-bootstrap agreement-minus-attractor differences reported in the text.

Method	Cue − random	Target mass	Gram. agr./attr.	Ungram. agr./attr.
Raw VJP	0.178	0.511	0.488	0.410
G_v only	0.161	0.147	**13.42**	**9.55**
Final JVP-FC	**0.190**	0.169	7.57	4.82
Direct global linear	0.004	0.053	0.870	0.869
Attention	0.006	0.039	1.79	1.66
Gradient × input	0.173	0.267	0.559	0.517
Integrated gradients	0.087	0.216	0.742	0.729

clusters and retained all six contexts for each sampled word. Syntax analyses resampled the 20 lexical-template clusters and retained all four number and grammaticality variants for each sampled template. These intervals therefore quantify variation across words or templates rather than treating contexts derived from one benchmark item as independent. Layer 6 was selected as the cue-minus-random peak in the exploratory sweep below, so the reported layer-6 intervals are conditional on that selected layer and do not include uncertainty from the six-layer search. Decision-intervention intervals separately resampled the 64 held-out WikiText contexts.

Semantic and syntactic readouts. Averaged over the two calibrations, Raw VJP's terminal-token and semantic-cue masses were 0.511 and 0.259; Final's were 0.169 and 0.350, against matched-random mass 0.160. Final cue-minus-random was 0.190 (95% word-cluster interval 0.096–0.297). This supports redistribution toward a cue-selective contextual profile; the paired comparison below reports the unresolved difference from Raw VJP in cue-minus-random.

For syntax, agreement-minus-attractor is energy at the *was*/*were* agreement site minus energy at the nearer attractor. Raw VJP's grammatical/ungrammatical differences were $-0.016/-0.023$; Final's were $+0.030$ (95% template-cluster interval 0.017–0.045) and $+0.024$ (0.009–0.039). The main-text summary (fig. 4) displays BANK and BAT, the first two benchmark words, with one fixed benchmark-order context per sense. Its stacked header summarises mean Raw VJP and Final energy allocations over all 72 semantic contexts. Figure 20 shows one fixed benchmark-order context for each sense of all 12 words, all semantic and syntactic summary panels, and the independent held-out decision study. Enlarged views of the same 24 contexts appear in figs. 26 to 28.

Depth and calibration reproducibility. An exploratory sweep evaluated layer-2, 4, 6, 8, 10, and 12 attention outputs under both independently fitted c4096 map families. Both calibrations selected layer 6 as the cue-minus-random peak. Across layers, mean continuous-inverse cosine was 0.915, energy-profile cosine was 0.959, top-five evidence-position overlap was 87.3%, and the cue-minus-random depth-curve cosine was 0.999. At layer 6, continuous-inverse cosine was 0.912, energy-profile cosine was 0.913, and top-five evidence-position overlap was 82.4% over all 152 controlled examples (72 semantic and 80 syntactic). The sweep did not support a monotonic lexical-to-semantic hierarchy: layer 2 remained target-token dominated, later layers reintegrated terminal evidence, and layer-12 target mass reached 0.772 under the seed-123 sweep. Figure 21 reports the complete depth and cross-calibration profiles.

Reference methods. Raw VJP, G_v only, and Final JVP-FC share the same target, operating point, boundaries, and terminal seed. The direct global linear reference was fitted on 4,096 WikiText windows and maps the complete layer-6 terminal feature directly to all 16 input embeddings; it does not use the source-local forward graph. Attention weights were averaged over heads at layer 6. Gradient-times-input and integrated gradients used the squared norm of the selected layer-6 attention feature as their scalar objective; integrated gradients used 16 steps from a zero-embedding baseline (Ancona et al., 2018; Sundararajan et al., 2017). These three quantities are attribution references, not feature-inversion baselines.

The references separate the semantic and syntactic patterns. The direct global map and layer-6 attention give cue-minus-random effects of 0.004 and 0.006. Gradient-times-input and integrated gradients give 0.173 and 0.087, but favour the attractor (grammatical agreement-to-attractor ratios

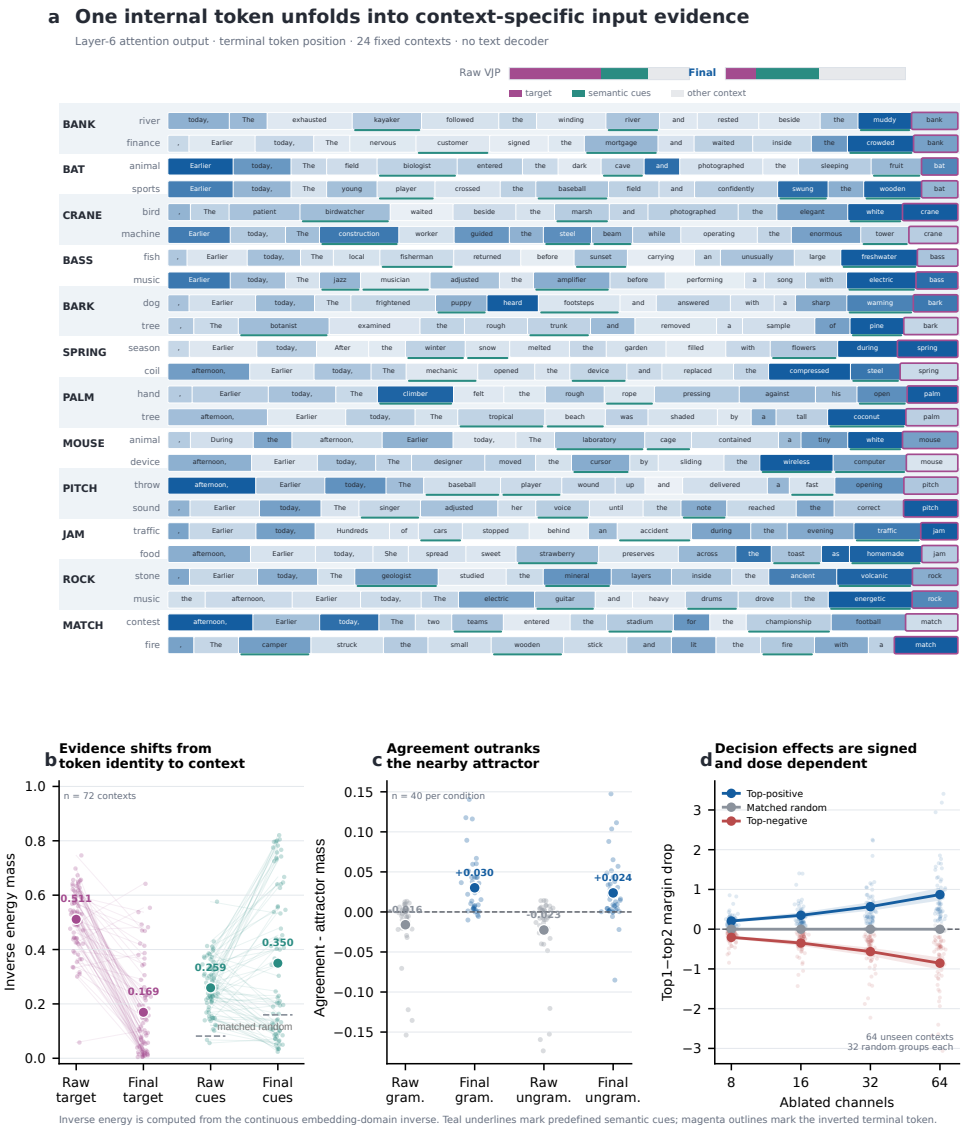


Figure 20: **Token-resolved source-grounded inversion in GPT-2.** The target is the complete layer-6 attention-output vector at the terminal token position. Layer 6 was identified as the cue-minus-random peak in an exploratory six-layer sweep, so the reported benchmark intervals are conditional on this selected layer. Panels a–c use continuous input-embedding inverses and no text decoder; panel d reports an independent intervention on the original representation. **a**, Rows show Final inverse energy for one fixed benchmark-order context from each sense of all 12 polysemous words; the stacked header compares mean Raw VJP and Final allocations. Fill intensity is inverse energy averaged over two independently fitted c4096 map families; teal underlines mark predefined semantic cues and magenta outlines mark the inverted terminal token. **b**, Per-context target-token and semantic-cue mass for Raw VJP and Final JVP-FC ($n = 72$); dashed grey marks position-matched random controls. **c**, Agreement-minus-attractor mass for 40 grammatical and 40 ungrammatical number-agreement controls. **d**, Internal intervention on decision-conditioned channel groups over 64 unseen WikiText contexts. Lines show means and 95% context-bootstrap intervals; faint points are individual contexts. The intervention acts on the original internal representation, not on a rendered inverse.

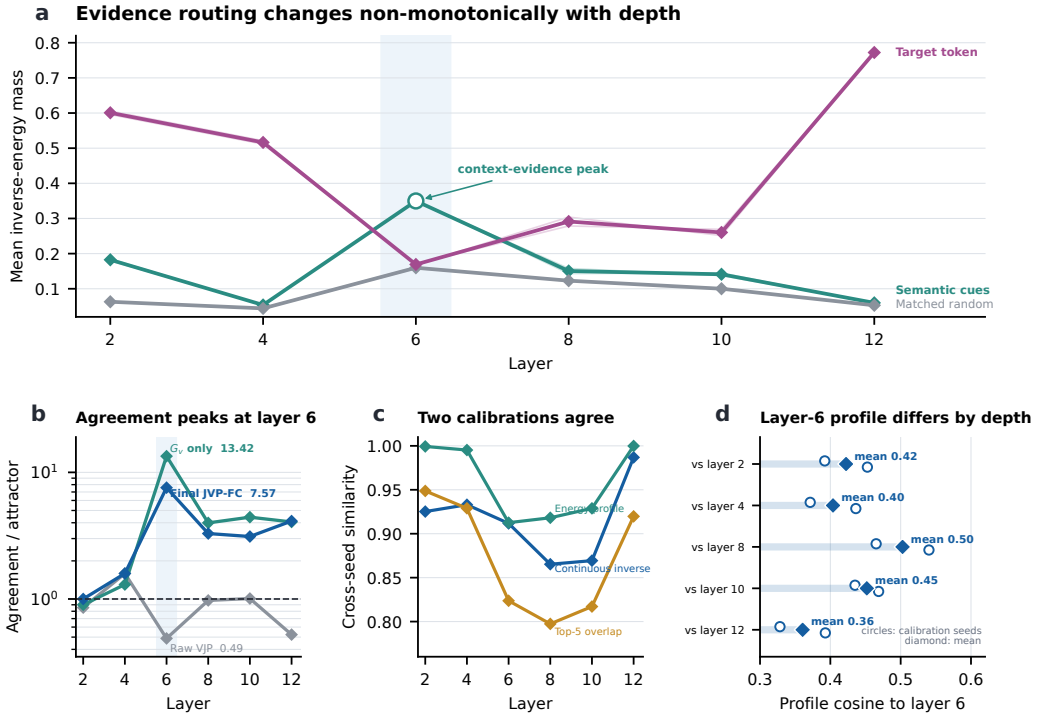


Figure 21: **Depth and calibration reproducibility of token-position inversion. a**, Target-token, semantic-cue, and position-matched-random mass for Final JVP-FC across layers 2, 4, 6, 8, 10, and 12. Thin lines are independent c4096 calibrations and thick lines are their means. **b**, Grammatical agreement-to-attractor ratio for Raw VJP, G_v only, and Final JVP-FC. **c**, Per-layer cross-seed continuous-inverse cosine, energy-profile cosine, and top-five evidence-position overlap. **d**, Layer-6 profile cosine relative to the other target depths. Circles are independent c4096 calibrations and diamonds are their means.

0.559 and 0.742). Final combines a 0.190 cue-minus-random effect with a descriptive agreement ratio of 7.57; G_v alone gives 13.42. None of these tested external references reproduces both patterns. The attention and gradient profiles are interpretability references, whereas the direct global map estimates from the target without the source-local DAG.

Paired stage comparisons. The scalar summaries in table 17 should not be read as a monotonic ranking of the two fitted stages. Relative to Raw VJP, Final changed cue-minus-random by $+0.012$ (95% word-cluster bootstrap interval -0.064–0.097), increased cue mass by 0.091 (0.024–0.158), and reduced terminal-token mass by 0.342 (0.294–0.389). Final also improved the pooled agreement-minus-attractor effect over Raw VJP by 0.046 (95% template-cluster bootstrap interval 0.027–0.066). However, the same effect was 0.031 lower than G_v alone (Final-minus-G_v interval -0.043 to -0.019). These comparisons support an interpretation in which G_v and JVP-FC alter different aspects of the inverse.

Source-local and query controls. For 24 fixed cross-context pairs, terminal state A was evaluated on operating-point geometry B. Averaged over the two calibrations, the swapped continuous inverse was closer to the geometry-B inverse than to the target-A inverse (0.496 versus 0.189); the corresponding profile cosines were 0.623 and 0.373. Every zero terminal state produced an exactly zero output. Replacing the terminal query with a predefined cue-position query also changed the inverse substantially: target-versus-cue cosine was 0.079 for the continuous inverse and 0.182 for the energy profile. These controls separate source-local geometry, selected terminal content, and selected token position. Figure 22 places the reference-method and source-local controls side by side, while fig. 23 reports every semantic and syntactic example.

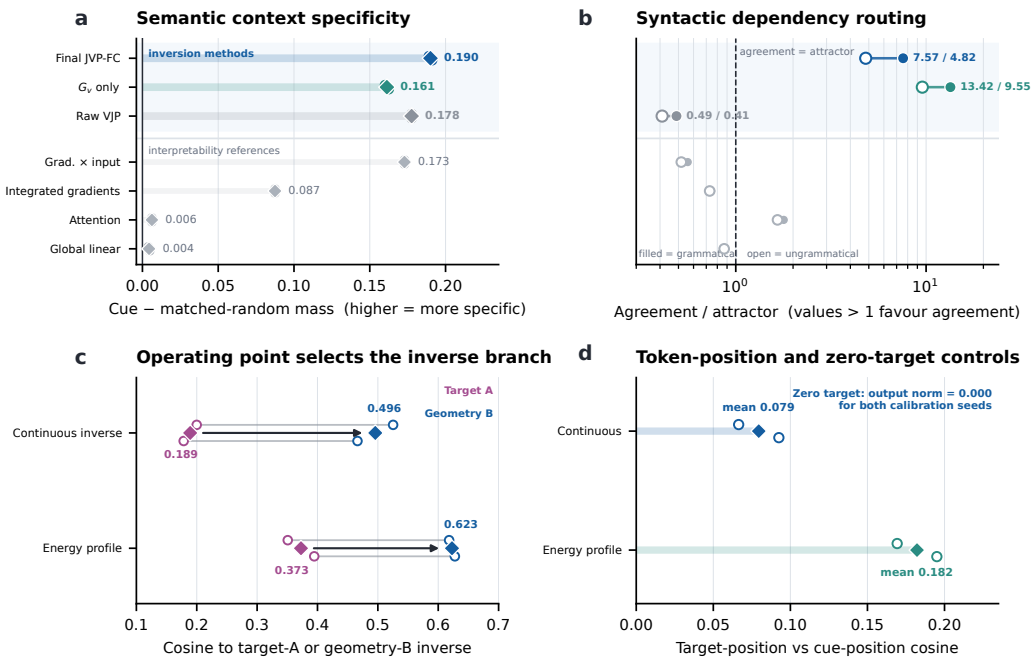


Figure 22: **Reference methods and source-local controls. a**, Semantic-cue mass above position-matched random controls for each inversion method and interpretability reference. Small circles are the two c4096 calibrations and diamonds are their means. **b**, Grammatical and ungrammatical agreement-to-attractor ratios in the same method order as panel a. Filled and open markers denote grammatical and ungrammatical contexts; the dashed line marks equal agreement and attractor mass. **c**, Terminal-state-A inverses evaluated on source-local geometry B. Paired seed points and mean arrows compare similarity to target-A and geometry-B inverses. **d**, Similarity between inverses of the terminal target position and a predefined cue position. An exactly zero terminal state gives exactly zero output under both calibrated map families.

Decision-conditioned protocol. Each of the 64 decision contexts was a held-out 16-token WikiText prefix. We computed the top1–top2 logit margin and ranked the 768 channels of the layer-6 terminal attention output by activation times margin gradient. Top-positive and top-negative groups contained the k most positive and most negative channels; each of 32 random controls sampled k channels uniformly without replacement. Ablation set the selected attention-output channels to zero at the terminal position.

A separate forward pass then recomputed the margin for the original top-1–top-2 token pair, and margin drop was measured before minus after intervention. The margin experiment therefore evaluates the feature group inside the original model. The corresponding inverse uses the fitted closed-form maps but is not fed back into GPT-2. At $k = 32$, the top-positive mean margin drop was 0.567 (0.463–0.680), the mean of the per-context random medians was 0.0001 (-0.0033–0.0038), and the top-negative mean was -0.564. Across group sizes $k \in \{8, 16, 32, 64\}$, the top-positive mean drops were 0.206, 0.349, 0.567, and 0.871, while random effects remained centred near zero. Figure 24 reports the complete group-size and context-level analysis together with fixed-order inverse profiles.

Joint full-coordinate channel-subset query. We next held a selected channel set fixed while changing its coordinate set. The target remained the GPT-2 small layer-6 attention output with shape $B \times 16 \times 768$, with token coordinates $p = 0, \ldots, 15$. For each of 128 fixed WikiText-2 contexts, channels were ranked by

$$\sum_{p=0}^{15} H_T[p, c] \, \frac{\partial(\ell_{\text{top1}} - \ell_{\text{top2}})}{\partial H_T[p, c]}.$$

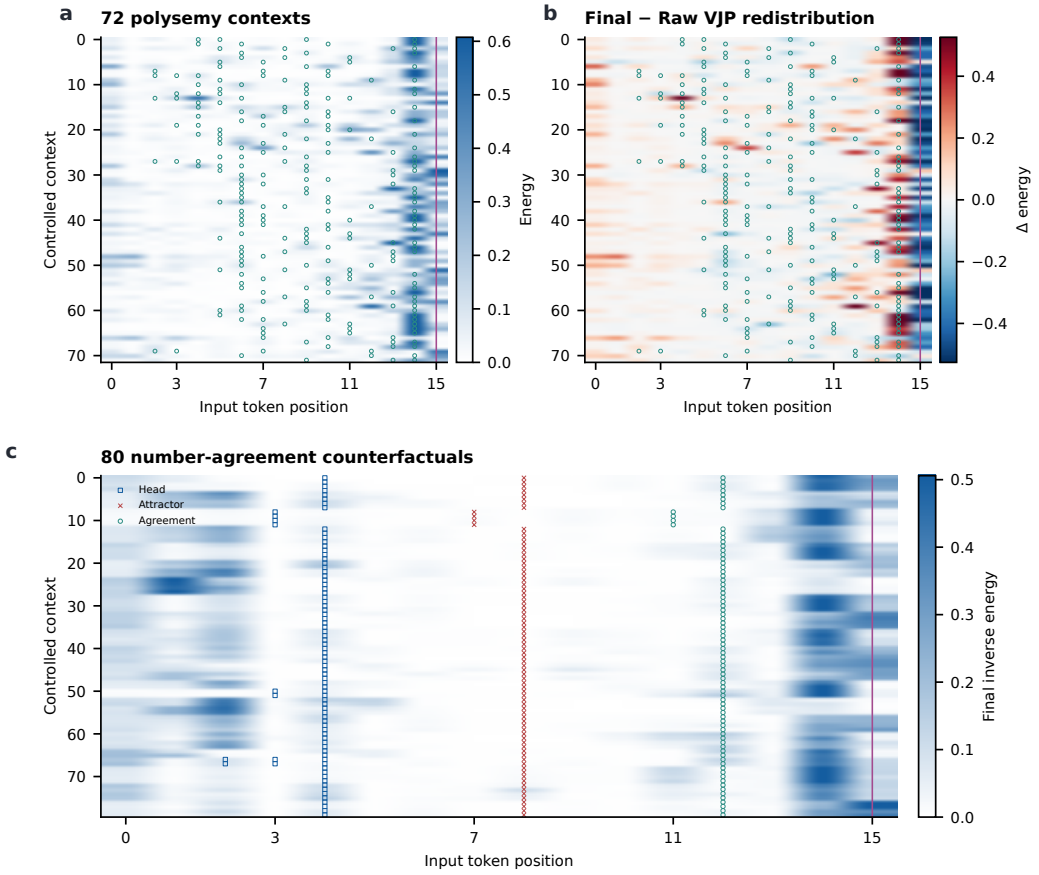


Figure 23: **Complete semantic and syntactic control matrices.** **a**, Final inverse energy for all 72 polysemy contexts. Teal circles mark predefined cues and the magenta line marks the terminal target. **b**, Final-minus-Raw-VJP energy redistribution for the same contexts. **c**, Final inverse energy for all 80 number-agreement counterfactuals; blue squares, red crosses, and teal circles mark heads, attractors, and agreement positions, respectively.

The 32 most positive and most negative channels formed the two decision-conditioned groups. Sixteen deterministic random groups excluded these channels; the group with the median all-position margin drop was retained as the representative random reference for that context. The same selected channels were queried either only at the terminal token, $Q = \{L-1\}$, or jointly over the complete token stream, $Q = Q_{\text{all}}$. The existing c4096 all-shared maps were read without refitting.

The evaluation windows were sampled independently of calibration. None of their complete 17-token context-plus-next-token intervals overlapped a c4096 calibration interval, and the 128 evaluation intervals were pairwise non-overlapping. The corresponding independent intervention zeroed the same channels either at the terminal position or at all 16 positions and recomputed the terminal top-1–top-2 logit margin.

Table 18: **GPT-2 joint full-coordinate channel-subset control.** Margin drop is measured by an independent intervention on the original attention-output representation. Paired differences report $Q_{\text{all}} - \{L-1\}$ as means with 95% context-bootstrap intervals over 128 strictly disjoint evaluation windows.

Channel group	Terminal-only margin drop	Full-coordinate margin drop	Paired difference
Top positive	0.447	0.677	0.230 (0.201–0.263)
Top negative	−0.442	−0.675	−0.233 (−0.270–−0.200)
Random median	0.0019	0.0024	0.0006 (−0.0030–0.0042)

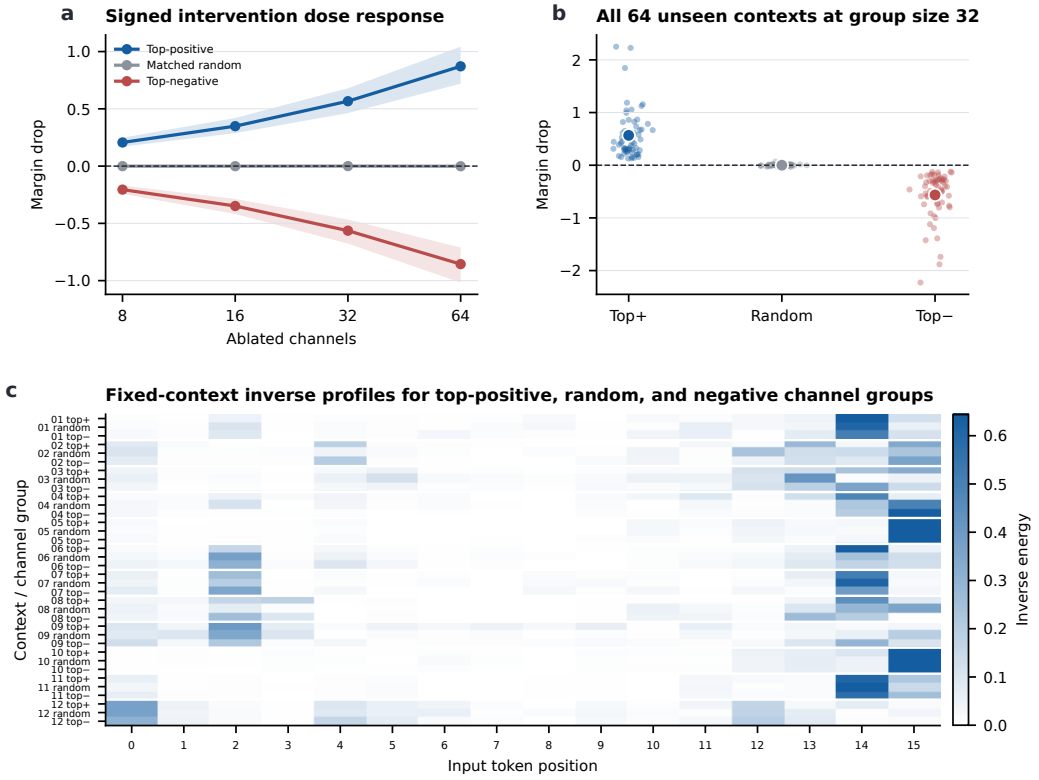


Figure 24: **Expanded decision-conditioned language analysis. a**, Mean signed intervention effect with 95% context-bootstrap intervals over group sizes 8, 16, 32, and 64. **b**, All 64 context-level effects at group size 32. **c**, Fixed-context inverse-energy profiles for top-positive, median random, and top-negative channel groups. Contexts are selected by fixed corpus order rather than effect size.

Expanding the top-positive query to Q_{all} increased the signed margin-drop effect in all 128 contexts, whereas the top-negative group changed in the opposite direction and the random contrast remained centred near zero (table 18). Holding S fixed while changing Q also changed the inverse: the mean token-energy-profile cosine was 0.489 (95% interval 0.444–0.535), the mean continuous-inverse cosine was 0.309, and normalised profile entropy increased by 0.144 (0.116–0.170). For the top-positive group, terminal-position energy mass decreased descriptively from 0.202 to 0.112. These measurements show that the joint full-coordinate target distributes the recovered input-embedding support over the context rather than reproducing the terminal-only query. They do not rank the correctness of two different targets.

Terminal-query scope. We additionally evaluated the complete 16×16 target-position matrix on 24 controlled contexts using the seed-123 map family without refitting. Raw VJP and attention assigned zero mean inverse energy to positions later than the queried token, consistent with the causal forward graph. The deployed closed-form surrogate assigned mean future-position energy of 0.426 after G_v and 0.427 after Final. Because the effect was already present after G_v, it is not specific to JVP-FC. We therefore do not use the deployed surrogate for arbitrary isolated nonterminal-position queries, and the formal semantic and syntactic routing claims remain restricted to terminal-token targets, for which no future positions exist. The joint Q_{all} channel-subset control above selects the complete token stream as one target and does not validate any individual nonterminal position. This diagnostic bounds the current deployment approximation; it does not alter the terminal-token results or the general source-local DAG formulation.

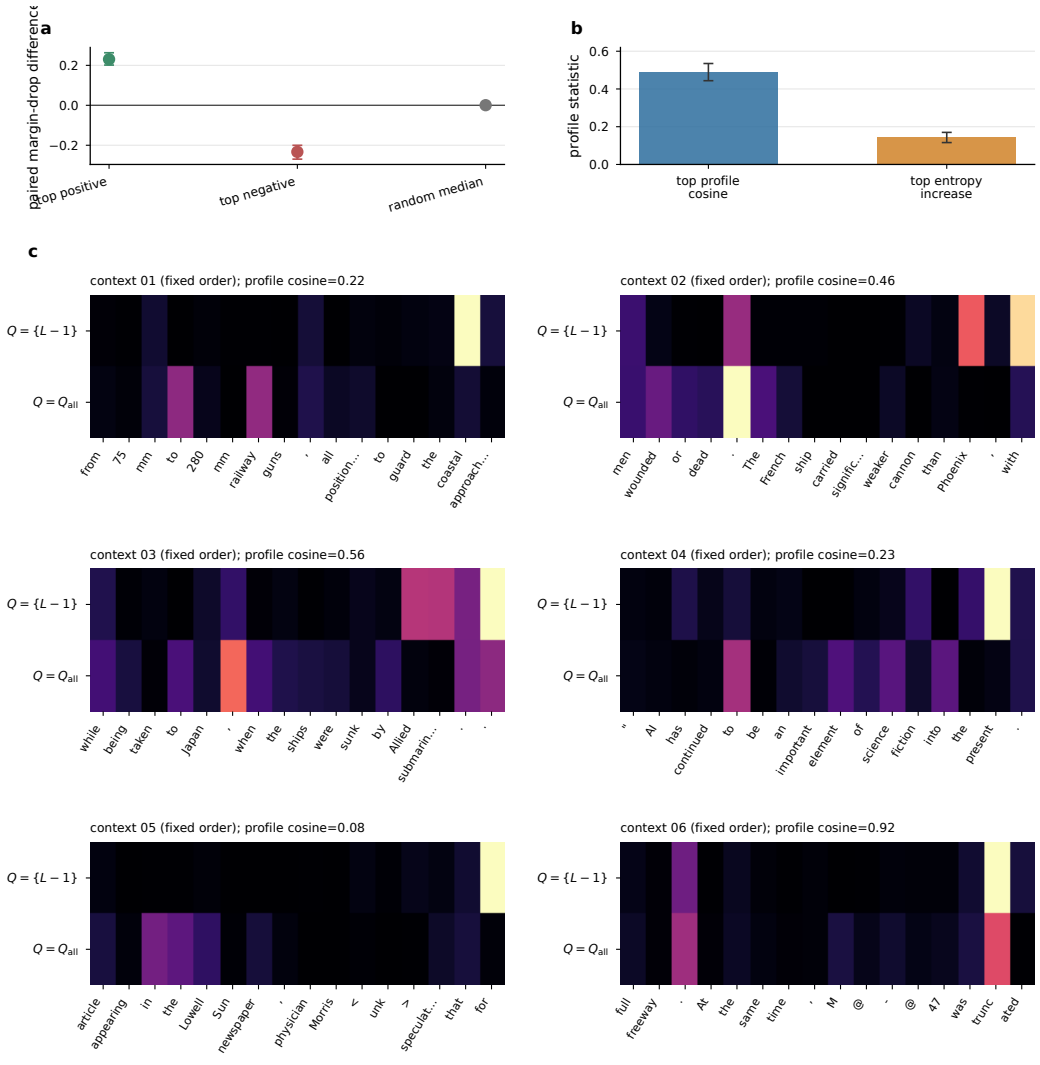


Figure 25: **GPT-2 terminal and joint full-coordinate channel-subset queries. a**, Paired intervention difference between Q_{all} and $Q = \{L-1\}$ for top-positive, top-negative, and representative random 32-channel groups. Points are means and bars are 95% context-bootstrap intervals over 128 strictly disjoint WikiText-2 windows. **b**, Token-energy-profile cosine and entropy increase for the top-positive group, with the same interval definition. **c**, Terminal-only and joint full-coordinate inverse-energy profiles for the first six evaluation contexts in fixed corpus order. Panels b–c use continuous input-embedding inverses and no text decoder; panel a reports an independent intervention on the original representation.

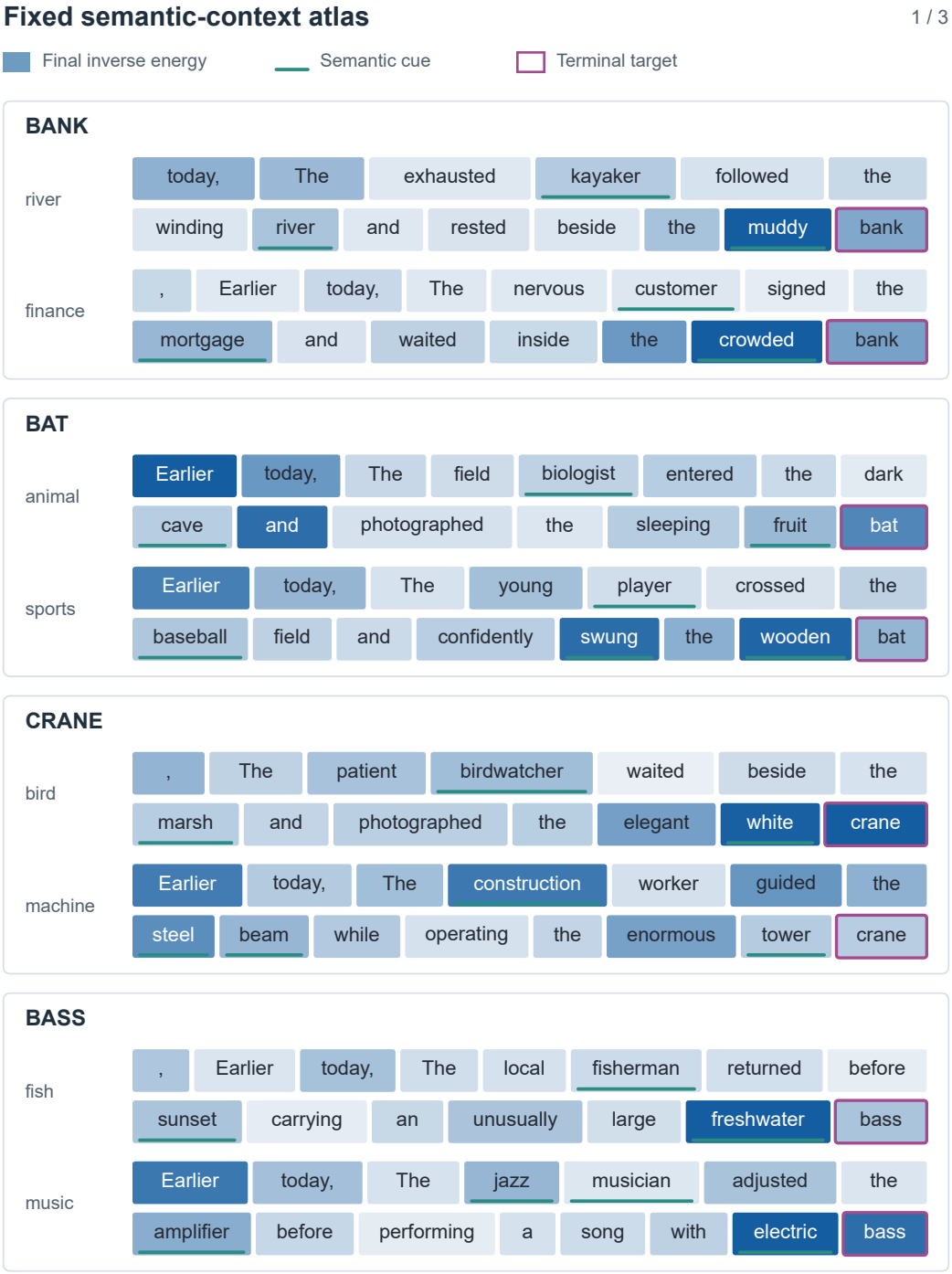


Figure 26: **Fixed semantic-context atlas, 1 of 3.** BANK, BAT, CRANE, and BASS, redrawn from fig. 20a with one fixed benchmark-order context per sense. Each context reads across its first line and then its second line. Cell colours show Final inverse energy averaged over two independently fitted c4096 map families; teal underlines mark predefined semantic cues and magenta outlines mark the queried terminal token.

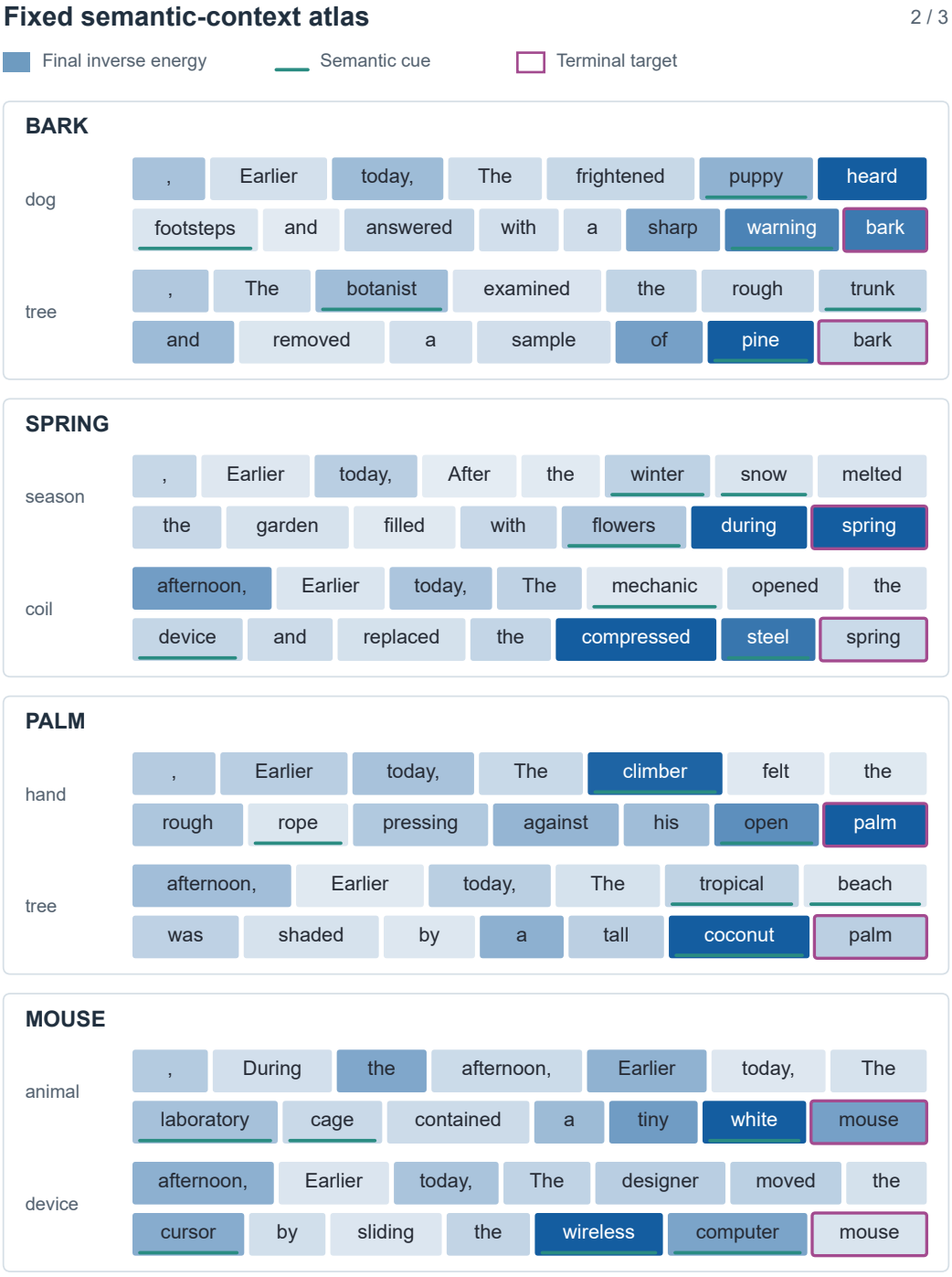


Figure 27: **Fixed semantic-context atlas, 2 of 3.** BARK, SPRING, PALM, and MOUSE contexts, redrawn from fig. 20a. The two-line reading order and all energy, cue, and target conventions are as in fig. 26.

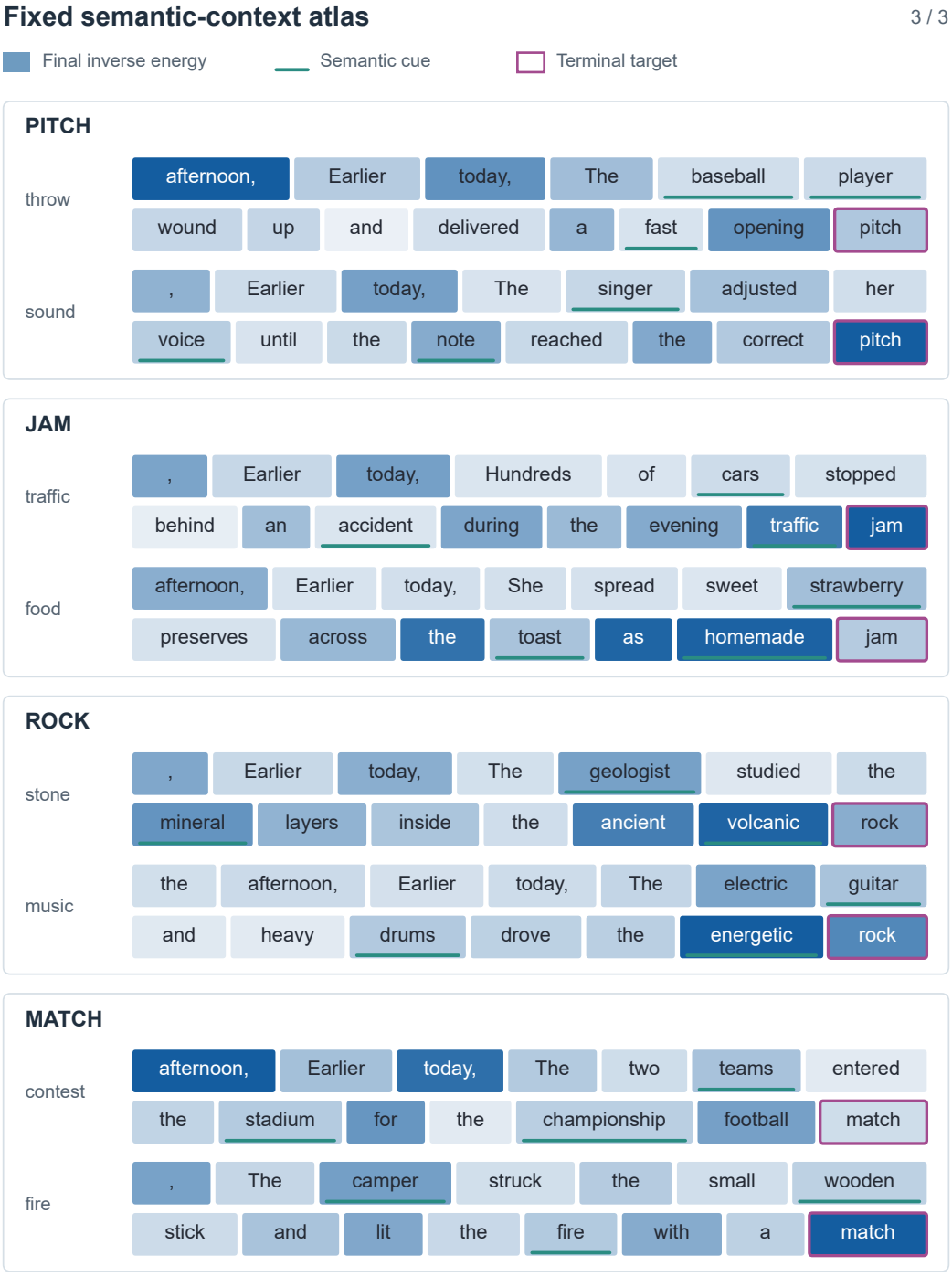


Figure 28: **Fixed semantic-context atlas, 3 of 3.** PITCH, JAM, ROCK, and MATCH contexts, redrawn from fig. 20a. The two-line reading order and all energy, cue, and target conventions are as in fig. 26.

C.15 Continuous Non-Residual Inversion and the Display Interface

The token-routing experiment uses no text renderer. We separately tested whether the complete sequence of a genuinely transformed non-residual feature could be inverted in the continuous embedding domain. The target was the full layer-6 MLP-output sequence. On 512 calibration-disjoint WikiText windows, mean input-embedding cosine was 0.030 for Raw VJP, 0.663 after G_v, and 0.683 after Final JVP-FC. Independent Euclidean nearest-vocabulary lookup recovered the exact source token at 2.3%, 39.7%, and 44.5% of positions, respectively. A calibration-centred matched-versus-cross re-encoding diagnostic gave AUC 0.526, 0.803, and 0.908. Re-encoding is reported as a source-specific representation-consistency check, not as the definition of a unique true inverse.

A zero-intercept rank-64 residual display interface contained 98,304 trainable parameters while GPT-2 and the inverse maps remained frozen. Under matched-capacity interfaces, Raw VJP reached 38.5% top-1 and 67.3% top-64 token recovery, whereas Final reached 59.0% and 88.1%. The same $L = 16$ inverse maps and interface transferred without refitting to lengths $L \in \{8, 12, 16, 24, 32\}$ by deterministic interpolation on the normalised rFFT grid. Final continuous embedding cosine remained between 0.670 and 0.716. These experiments demonstrate reusable continuous inversion and provide a readable display, but neither decoded fluency nor exact-token recovery is used as evidence for the token-resolved interpretability claim. Figure 29 separates the quantitative continuous inverse from the optional display interface.

Supplementary boundary. The formal language claim is that a source-grounded closed-form inverse operates on a canonical language Transformer and exposes controlled token evidence from a selected internal representation. The complete-state and display experiments broaden feasibility but do not convert feature inversion into sentence generation. Evidence outside GPT-2 small, substantially longer contexts, and other language-model architectures is outside the present claim. Isolated nonterminal-position queries are also excluded because the deployed surrogate can allocate inverse energy to later positions, as disclosed above; the joint full-coordinate channel-subset query does not remove this boundary.

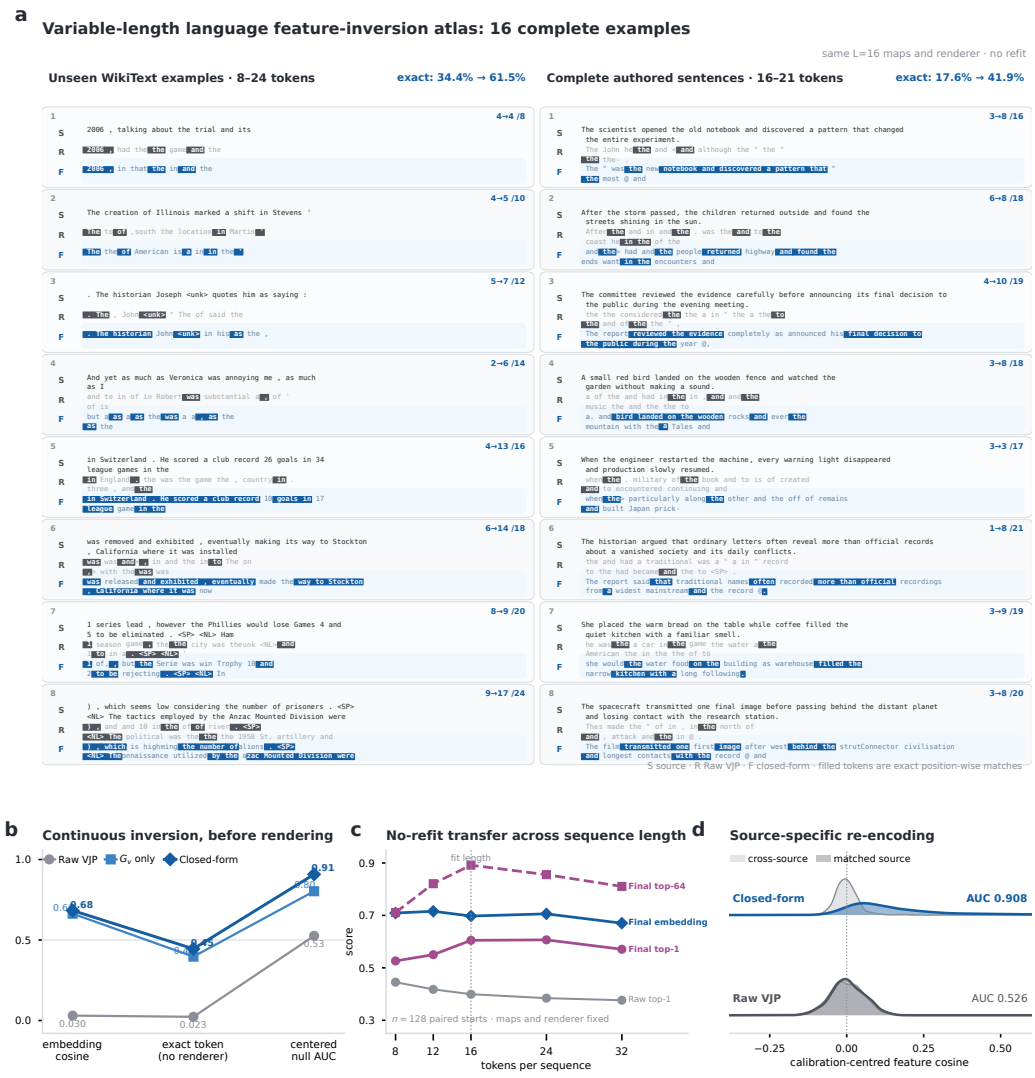


Figure 29: **Closed-form inversion of a non-residual GPT-2 feature.** **a**, Qualitative atlas over unseen WikiText and authored contexts using matched-capacity display interfaces. **b**, Continuous embedding cosine, pre-renderer exact-token recovery, and calibration-centred null AUC. **c**, No-refit transfer across sequence lengths 8–32. **d**, Calibration-centred matched- and cross-source re-encoding distributions. All quantitative inversion and re-encoding metrics are computed before the display interface.

D Discussion and Limitations

Source conditioning is part of the feature-inversion object: a target tensor from a non-injective nonlinear network does not identify its generating local branch. The forward operating point supplies this condition through its activations, nonlinear regimes, routing states, and local Jacobians. In visual target–operator and language cross-context swaps, changing that point redirected the inverse toward a new source-local branch; changing the terminal state, channel set, or coordinate query modulated the content transported through it. Zero targets give exactly zero outputs, excluding a fixed target-independent additive template. Distinct channel and coordinate inverses at the same operating point, together with poor map transfer across mismatched checkpoints, establish query and model dependence (sections C.9 and C.13). This dependence coexists with strong source-geometry dominance, as in the ResNet18 full-target swaps. The controls therefore support a feature inverse as a target-modulated upstream-state estimate traced through the local computation that generated the feature, with that computation supplying the branch condition absent from the target tensor alone.

Architecture-dependent inverses provide a representation-level view whose differences can reflect both source-local computation and estimation error. The ViT true-child diagnostic shows that online state estimation can substantially affect the output (section C.11). Convolutional models typically expose coherent spatial and high-frequency structure, while ViT-B/32 retains stronger patch-grid, low-frequency, and globally mixed geometry. Swin-T's stronger spatial and high-frequency organisation than ViT-B/32 despite lower pixel cosine shows that the closed-form estimator can retain spatial structure on Transformers. The contrast is consistent with the image biases of global patch-token mixing and hierarchical windowed computation (Dosovitskiy et al., 2021; Liu et al., 2021). Similarity to the target-generating image remains an input-instance alignment diagnostic. On Pets and CUB, independently fine-tuned ConvNeXt-B checkpoints remain accurate, and even their lowest-alignment quartiles are more consistent with matched than source-shuffled Jacobian paths (section C.6). Weak pixel alignment can thus coexist with fidelity to the source-local computation. Likewise, weak or sparse single-channel inverses may reflect little independently visible structure or evidence distributed across channel combinations, so appearance alone does not establish failure.

The ViT patch-query experiment makes the coordinate component of the terminal query explicit. Its normalised inverse-magnitude support is strongly coordinate aligned at shallow boundaries and increasingly distributed at deeper ones, consistent with contextual mixing. This readout describes where the recovered input-domain structure is expressed. Decision evidence comes from independent interventions on the original patch-token states: layers 8 and 11 show signed margin effects, while layer-14 patch-token interventions are numerically null under the CLS-only readout (section C.3).

The GPT-2 results extend the same two-stage construction to causal token sequences with different spectral configurations: singleton per-frequency maps in vision and all-shared tying in GPT-2. The formal semantic and agreement benchmarks use terminal queries in 16-token, author-constructed contexts. Layer 6 was selected as the cue-minus-random peak in an exploratory sweep, and the layer-6 benchmark intervals are conditional on that choice. The semantic result shows redistribution from terminal-token identity toward a cue-selective contextual profile, while the Final-minus-Raw change in cue-minus-random remains unresolved from zero. Final improves pooled agreement routing over Raw VJP but remains below G_v alone. The stages therefore alter complementary aspects of the inverse, with non-monotonic effects across diagnostics. The joint semantic and syntactic pattern, not reproduced by the direct global map or attribution references, provides constructive evidence beyond visual coordinates (section C.14).

Prediction-conditioned atlases pair two views of the same feature group: inversion reveals the input-domain structure selected by a layer, channel, or channel group, and independent internal intervention tests its effect on the current prediction. Visual top-positive groups exceed matched random controls across all tested model–layer–group-size settings. On held-out WikiText contexts, GPT-2 interventions separate top-positive, random, and top-negative groups with the expected signed effects (sections C.12 and C.14). Inversion displays the selected group's input-domain structure, while intervention on its original representation supplies the causal evidence. These paired readouts describe structure and decision role; assigning a unique human-semantic label remains a separate interpretive task.

The main limitations concern the deployed approximation to the ideal local inverse. The metric-aware operator $M_v^{-1} J^* M_u$ varies with the sample and boundary, whereas deployment uses fixed zero-

intercept Wiener projections calibrated over a distribution. JVP-FC uses pulled-back source-point consistency residuals to estimate the part of first-stage state error predictable within the calibrated linear class. This addresses sample-dependent mismatch without guaranteeing exact online metrics or complete local correction. The larger ViT-B/32 online-child oracle gap motivates more efficient sample-specific metric or state estimation. Frequency-diagonal, full-channel maps make calibration tractable but exclude cross-frequency mixing within each repair map; their calibration, storage, and online costs vary by architecture (sections B.8 and C.10). Cross-checkpoint failures confirm that the maps encode checkpoint-specific geometry and require recalibration when weights change (section C.13). More efficient metric-aware updates and validated structured operator families could reduce these costs while retaining source-local geometry.

The formulation covers frozen finite differentiable DAGs exposed by autograd, with declared tensor boundaries and channel–coordinate axes. Empirical validation covers the studied CNNs and vision Transformers, including all 49 non-CLS ViT-B/32 patch queries, and GPT-2 small. In language, the joint Q_{all} channel-subset control extends query reuse but does not validate isolated nonterminal positions. For such isolated queries, the deployed surrogate can allocate substantial inverse energy to future positions. This diagnostic identifies a limitation of the spectral deployment approximation, not of the source-local DAG formulation itself. Arbitrary nonterminal queries, internal attention-logit or attention-weight matrices, substantially longer contexts, and transfer to other language-model architectures remain outside validation. Supplementary full-sequence MLP-output inversion broadens continuous-inversion feasibility; its optional renderer provides a display interface rather than sentence reconstruction (sections C.14 and C.15). Causality- and position-structured operators and validation on additional differentiable architectures are further directions.

E Additional Context for Inversion and Interpretability Methods

Learned reverse mappings. Independently trained reverse models typically target a representation family or output domain. Convolutional decoders can invert fixed visual representations, while autoregressive density models represent image distributions conditioned on supervised features (Dosovitskiy & Brox, 2016; Nash et al., 2019). In language, embedding-inversion attacks partially recover words from sentence embeddings (Song & Raghunathan, 2020); encoder–decoder correctors can generate, re-embed, and iteratively revise text from dense embeddings (Morris et al., 2023). These recoveries depend jointly on the target representation, decoder architecture, and data prior. For the sample-specific inverse defined here, the generating local computation provides an additional explicit condition.

Interpretability signals. Input gradients measure local sensitivity (Simonyan et al., 2013); deconvolutional visualisation and Guided Backpropagation visualise activation patterns through modified reverse rules (Zeiler & Fergus, 2014; Springenberg et al., 2014). Grad-CAM aggregates class gradients for spatial localisation (Selvaraju et al., 2017). In language models, attention weights provide internal routing references whose interpretation depends on the claim and evaluation criterion (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Gradient-times-input and Integrated Gradients yield token-level sensitivity or attribution scores (Ancona et al., 2018; Sundararajan et al., 2017). These signals serve as interpretability references alongside the continuous upstream-state estimates produced by source-grounded feature inversion.

Concepts, probing, and intervention. Network Dissection measures unit alignment with a labelled concept vocabulary (Bau et al., 2017); TCAV tests prediction sensitivity along user-defined concept directions (Kim et al., 2018); ACE automatically discovers concepts and measures their importance (Ghorbani et al., 2019). Probing tests whether linguistic structure is decodable from hidden states; structural probes fit supervised maps from word representations to syntactic geometry (Hewitt & Manning, 2019). Intervention and mediation test internal components' behavioural effects (Vig et al., 2020). These distinctions motivate pairing an inverse with an independent intervention on the same original feature group: the inverse displays input-domain structure, and intervention measures the group's effect on the current decision.

SUPPLEMENTARY REFERENCES

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3319–3327, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4829–4837, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2668–2677, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. URL https://arxiv.org/abs/1609.07843.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12448–12460. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.765. URL https://aclanthology.org/2023.emnlp-main.765/.

Charlie Nash, Nate Kushman, and Christopher K. I. Williams. Inverting supervised representations with autoregressive neural density models. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1620–1629, 2019.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10428–10436, 2020.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, 1992.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Kaixiang Shu. Adjoint inversion reveals holographic superposition and destructive interference in CNN classifiers. *arXiv preprint arXiv:2604.27529*, 2026.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 377–390. ACM, 2020. doi: 10.1145/3372297.3417270. URL https://doi.org/10.1145/3372297.3417270.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, 2019.

Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP architecture for vision. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24261–24272, 2021.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.