THE GEOMETRY OF TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

A new target task may have little labeled data but access to a library of fine-tuned models whose source datasets are unavailable. Existing transferability scores rank source representations, while model-merging methods combine task vectors, but neither directly quantifies *which part of target adaptation is already contained in the source library*. We formulate transfer as projection in a target-conditioned functional tangent space. Each source checkpoint induces a tangent field on target inputs, while the target loss induces a curvature-normalized descent field. Their weighted L_2 geometry defines a principal transfer angle for each source subspace and a transfer completeness index measuring the fraction of target adaptation captured by a sparse source span. We introduce GEOTRANSFER, which estimates this geometry from source checkpoints and a small target sample, performs group-sparse source projection, learns an orthogonality-regularized innovation adapter for residual target structure, and uses a held-out risk gate for target-only fallback. We prove a local projection identity, first-order invariance to smooth reparameterization, a sparse-span oracle inequality, and a finite-sample guarantee that gated deployment does not exceed the target-only baseline under a prespecified bounded target risk. Simulations and real-data experiments show that the geometry predicts realized source utility, is stable under parameter rescaling, and improves few-shot transfer when useful and harmful source directions coexist.

1 INTRODUCTION

Transfer learning increasingly begins with a model library, not with a source dataset. A laboratory may retain fitted models from related cohorts after individual-level data have become inaccessible; a company may store compact task adapters or LoRA modules (Houlsby et al., 2019; Hu et al., 2022); and public model hubs contain large families of checkpoints descended from the same foundation model. When a new target task has only a small labeled sample, the operational problem is therefore multi-source and source-data-free: which fitted sources contain knowledge useful for the target, which source directions conflict, and how much genuinely new target adaptation remains to be learned?

The statistical literature gives a useful clue. In high-dimensional regression, Gu et al. (2025) show that source-target directional concordance can be more robust than Euclidean distance and can be exploited even when only source parameter estimates are available. A direct replacement of regression coefficients by neural-network parameters, however, is not satisfactory. Neural parameterizations contain rescaling and permutation symmetries (Dinh et al., 2017; Ainsworth et al., 2023), and different optima can be connected through low-loss paths (Garipov et al., 2018). Parameter space is therefore a poor proxy for predictive change; equal Euclidean perturbations can induce very different changes in target outputs at identical input points. Consequently, a small raw angle between two task vectors need not imply that they produce similar functional changes on the target distribution.

Transferability estimation and model merging address neighboring questions with fundamentally different estimands. H-score, LEEP, LogME, PACTran, and task-relatedness scores predict whether a pretrained representation will transfer (Bao et al., 2019; Nguyen et al., 2020; You et al., 2021; Ding et al., 2022; Mehra et al., 2024). Model soups and task-vector methods instead combine related checkpoints in weight space (Wortsman et al., 2022; Ilharco et al., 2023), while Fisher weighting, conflict resolution, gradient corrections, and low-rank subspaces seek to reduce merge interference (Matena and Raffel, 2022; Yadav et al., 2023; Daheim et al., 2024; Gargiulo et al., 2025; Sun et al., 2025). Closest to our source-data-free setting, multi-source transfer by model merging decomposes source models and adapts selected components on the target (Osial et al., 2025); recent work also considers curvature-

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.

aware experts, functional anchors, adaptive subspaces, and Fisher-Rao geometry (Nguyen et al., 2025; Shi et al., 2025; Crisostomi et al., 2026; Wang et al., 2026). These methods primarily score individual sources or construct a merged model. Our object is different: the target-conditioned projection of a target descent field onto the functional span induced by a source library, together with a completeness diagnostic measuring how much target adaptation that library can represent.

We make this object explicit. Let $\boldsymbol{\theta}_0$ be a common pretrained checkpoint and $\boldsymbol{\Delta}_s = \boldsymbol{\theta}_s - \boldsymbol{\theta}_0$ the update supplied by source s. On a target input x, source s induces the first-order functional perturbation $\phi_s(x) = P_C J_0(x) \boldsymbol{\Delta}_s$, where $J_0(x)$ is the Jacobian of the target logits at $\boldsymbol{\theta}_0$ and P_C removes the common-logit direction. This first-order function-space viewpoint is related to tangent-kernel descriptions of neural-network training (Jacot et al., 2018), but here the tangent dictionary is supplied by fitted source updates, not by individual parameter coordinates. The target loss and target distribution determine a local curvature-normalized descent field $r_0(x)$ in the same output space. After weighting output perturbations by the target loss curvature, ϕ_s and r_0 live in a Hilbert space whose inner product is determined by the target distribution. Transferability then has a geometric meaning; a source is useful when its tangent subspace has a small principal angle with the target descent field, and a library is locally sufficient when a sparse source span captures a large fraction of the target descent energy.

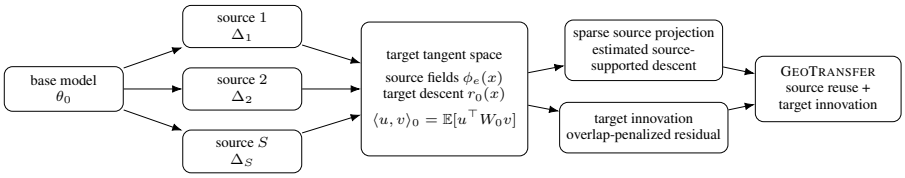


Figure 1: **The geometry of transfer.** Source updates map through the target Jacobian to functional tangent directions. GeoTransfer estimates a sparse approximation to the target descent field and learns residual innovation while penalizing overlap with the source span.

This formulation therefore yields a complete algorithm, not merely a transferability score. GeoTransfer constructs source atoms at the checkpoint, layer, adapter, or low-rank-component level; estimates a target-conditioned Gram matrix and source-target signal vector without materializing a parameter-space Hessian; solves a group-sparse projection that accounts jointly for source redundancy and orientation; and optionally learns a small target-specific residual penalized for functional overlap with the selected source span. A transferred candidate is deployed only when a held-out upper confidence bound on its prespecified bounded-risk difference from the target-only baseline is nonpositive; otherwise the procedure falls back to target-only. The decomposition therefore has a direct interpretation; source-supported descent is reused, redundant or unsupported source structure is suppressed, and target novelty is learned as innovation without being forced into the source span.

Our contributions are fourfold. First, we introduce the target functional transfer geometry, including principal transfer angles for source subspaces and a transfer-completeness index for a source library. Second, we derive GeoTransfer, a source-data-free multi-source procedure that performs sparse geometric projection and orthogonality-regularized innovation learning with a target-only fallback. Third, we show that the population source-span solution is the orthogonal projection of the target natural descent field, establish first-order invariance to smooth reparameterizations, and derive a sparse-group oracle inequality whose complexity depends on the active source span, not on the ambient neural-network dimension. Fourth, experiments evaluate source-utility ranking, robustness to parameter symmetries, source-span recovery, low-resource prediction, and negative-transfer protection, with full assumptions, proofs, optimization details, and extended results in the appendix.

2 Target-Conditioned Transfer Geometry

We observe S source checkpoints descended from a common base $\boldsymbol{\theta}_0$. Transfer uses prespecified shared coordinates and a target-compatible output map already defined at the base; source-specific heads are never transferred, preserving a common parameterization across all source atoms. The target law is P_0 over (X, Y), with $X \in \mathcal{X}$, and labeled data are split in advance as $\mathcal{D}_0 = \mathcal{D}_0^{\mathrm{fit}} \dot{\cup} \mathcal{D}_0^{\mathrm{val}} \dot{\cup} \mathcal{D}_0^{\mathrm{test}}$. Fit data estimate geometry and adaptation, validation data tune the finite candidate library and gate transfer, and test data are untouched until evaluation. We present multiclass classification with C target classes and a twice differentiable loss ℓ whose conditional logit curvature is positive semidefinite on the identifiable centered-logit space. Appendix D covers regression and token losses.

A whole checkpoint is generally too coarse a transfer unit. Let p denote the number of shared transferable parameter coordinates. We partition source s into prespecified atoms $\mathbf{\Delta}_{sm} \in \mathbb{R}^p,\ m \in [M_s] = \{1, \dots, M_s\}$, and $\mathbf{\Delta}_s = \sum_{m=1}^{M_s} \mathbf{\Delta}_{sm}$. The partition is fixed before target labels are used. Let $\mathcal{E} = \{(s,m) : s \in [S], m \in [M_s]\}$, $M = |\mathcal{E}|$, index atoms by $e \in [M]$, and let $\mathcal{E}_s \subset [M]$ denote the indices belonging to source s. For the target logit map $z_{\boldsymbol{\theta}}(x) \in \mathbb{R}^C$, define

$$J_0(x) = \left.\frac{\partial z_{\boldsymbol{\theta}}(x)}{\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \quad \mathbb{R}_0^C = \{v \in \mathbb{R}^C : \mathbf{1}^\top v = 0\}, \quad P_C = I_C - \frac{\mathbf{1}\mathbf{1}^\top}{C}, \tag{2.1}$$

where $\mathbf{1} \in \mathbb{R}^C$ is the all-ones vector and I_C is the $C \times C$ identity matrix. We work with centered logits, so atom e induces $\phi_e(x) = P_C J_0(x) \mathbf{\Delta}_e \in \mathbb{R}_0^C$. The field ϕ_e is the first-order identifiable change in target logits produced by the source atom on target input x relative to the common base checkpoint. Once the source update is available, it can be evaluated on target inputs without access to the source training data. A Jacobian-vector product or symmetric directional finite difference computes $\phi_e(x)$ without storing $J_0(x)$; the result is then centered by P_C.

We write $z_0(x) = z_{\boldsymbol{\theta}_0}(x)$ and $p_0(x) = \mathrm{softmax}\{z_0(x)\}$. On $\mathbb{R}_0^C$, define the conditional target score and curvature by

$$g_0(x) = P_C \mathbb{E}\left[\nabla_z \ell\{z_0(X), Y\} \mid X = x\right] \text{ and } W_0(x) = P_C \mathbb{E}\left[\nabla_z^2 \ell\{z_0(X), Y\} \mid X = x\right] P_C. \tag{2.2}$$

For cross-entropy, the Hessian is label free and $W_0(x) = \mathrm{diag}\{p_0(x)\} - p_0(x)p_0(x)^\top$ as an operator on $\mathbb{R}_0^C$. Here $\mathrm{diag}\{p_0(x)\}$ denotes the diagonal matrix with $p_0(x)$ on its diagonal. The theoretical results use this true loss curvature; a small ridge may be added only in numerical inverses and is treated as a geometry perturbation in Appendix B. For measurable perturbations $u, v : \mathcal{X} \to \mathbb{R}_0^C$ with finite weighted second moment, define

$$\langle u, v \rangle_0 = \mathbb{E}_{P_0}\left[u(X)^\top W_0(X) v(X)\right] \text{ and } \|u\|_0^2 = \langle u, u \rangle_0 . \tag{2.3}$$

We identify fields that differ only by a zero-norm element and let $\mathcal{H}_0$ denote the completion. Under the curvature condition in Section 4, this is a Hilbert space on the centered output coordinates, with no null directions. For any closed subspace $\mathcal{V} \subset \mathcal{H}_0$, we write $\mathrm{Proj}_{\mathcal{V}} h$ for the orthogonal projection of $h \in \mathcal{H}_0$ onto $\mathcal{V}$ under the inner product in (2.3). Let P_0^X denote the X-marginal of P_0. We assume $g_0(x)$ lies in the range of $W_0(x)$ for P_0^X-almost every x, as is automatic under Section 4's positive-definite centered-space condition. The target descent field is $r_0(x) = -W_0(x)^\dagger g_0(x)$, where $\dagger$ denotes the Moore-Penrose inverse on $\mathbb{R}_0^C$. Thus $W_0(x) r_0(x) = -g_0(x)$. This is a curvature-normalized descent direction; for softmax cross-entropy, W_0 is the output Fisher metric and the construction coincides locally with natural-gradient geometry (Amari, 1998; Martens, 2020).

Assume $\|r_0\|_0 > 0$. If $\|r_0\|_0 = 0$, then $g_0 = 0$ P_0^X-almost surely under the stated range condition, so the base is locally stationary. For source s, let $\mathcal{V}_s = \mathrm{span}\{\phi_e : e \in \mathcal{E}_s\} \subset \mathcal{H}_0$. Its principal transfer angle is

$$\cos^2 \Theta_s = \frac{\|\mathrm{Proj}_{\mathcal{V}_s} r_0\|_0^2}{\|r_0\|_0^2}, \quad \Theta_s \in [0, \pi/2]. \tag{2.4}$$

For a one-dimensional source this is the squared target-functional cosine; for a multi-atom source it measures the fraction of target-descent energy captured by that source span. Because a subspace contains both orientations of each atom, Θ_s measures span availability, not whether the stored forward update itself points in the favorable direction. The corresponding quadratic local span utility is

$$U_s = \frac{1}{2} \|\mathrm{Proj}_{\mathcal{V}_s} r_0\|_0^2. \tag{2.5}$$

Section 4 relates (2.5) to actual local target-risk improvement and gives the associated remainder.

The full library contributes $\mathcal{V} = \mathrm{span}\{\phi_e : e \in [M]\}$. For an atom budget $q \in [M]$, let $\mathcal{V}_A = \mathrm{span}\{\phi_e : e \in A\}$ and define the q-sparse transfer completeness

$$\chi_q = \max_{A \subset [M]; |A| \le q} \frac{\|\mathrm{Proj}_{\mathcal{V}_A} r_0\|_0^2}{\|r_0\|_0^2}. \tag{2.6}$$

Thus $\chi_q = 0$ means that no span generated by at most q atoms captures any component of first-order target descent, whereas $\chi_q = 1$ means that some set of at most q atoms spans the entire local descent field. Unlike source ranking, χ_q diagnoses the library itself; it quantifies the fraction of local target

adaptation representable by sparse source reuse. For $q < M$, completeness is defined relative to the prespecified atomization; changing the granularity can change χ_q, so the atomization is fixed before target labels are used and reported with every experiment.

The geometry has a finite representation on the source span. Define $G_{ee'} = \langle \phi_e, \phi_{e'} \rangle_0$, $b_e = \langle \phi_e, r_0 \rangle_0 = -\mathbb{E}[\phi_e(X)^\top g_0(X)]$, and collect these quantities as $\mathbf{G} = (G_{ee'})_{e,e' \in [M]} \in \mathbb{R}^{M \times M}$ and $\mathbf{b} = (b_1, \ldots, b_M)^\top \in \mathbb{R}^M$. For $\mathbf{a} = (a_1, \ldots, a_M)^\top \in \mathbb{R}^M$, we write $\Phi \mathbf{a} = \sum_{e=1}^M a_e \phi_e$. Then $\mathbf{G} \succeq 0$. For $A \subset [M]$, let $\mathbf{G}_{AA}$ denote the principal submatrix indexed by A, and let $\mathbf{b}_A$ and $\mathbf{a}_A$ denote the corresponding subvectors. The projection coefficients may be chosen as $\mathbf{a}_A^\star = \mathbf{G}_{AA}^\dagger \mathbf{b}_A$, and therefore $\| \operatorname{Proj}_{\mathcal{V}_A} r_0 \|_0^2 = \mathbf{b}_A^\top \mathbf{G}_{AA}^\dagger \mathbf{b}_A$. Hence the source dictionary together with $(\mathbf{G}, \mathbf{b})$ determines every source-span projection, while $(\mathbf{G}, \mathbf{b})$ alone determines its coefficients up to Gram-null directions and its projected descent energy. The diagonal G_{ee} is the target-functional size of atom e, and off-diagonal entries measure target-specific redundancy. If $G_{ee} = 0$, atom e has zero target-functional norm and A_e^{dir} is defined as zero by convention. For $G_{ee} > 0$, define the directed atom alignment

$$A_e^{\mathrm{dir}} = \frac{b_e}{\sqrt{G_{ee}}\, \|r_0\|_0} \in [-1, 1] \tag{2.7}$$

records orientation: $A_e^{\mathrm{dir}} > 0$ means the stored forward atom agrees with target descent, whereas $A_e^{\mathrm{dir}} < 0$ means its reverse direction agrees. Principal angles and χ_q are orientation free by construction. Applications requiring literal forward reuse may instead constrain source coefficients to be nonnegative, in which case A_e^{dir} distinguishes favorable from unavailable orientations.

Remark 2.1. *The construction reduces to familiar angle geometry in a special case. Suppose $z_\theta(x) = x^\top \theta$, $\ell(z, y) = \frac{1}{2}(y - z)^2$, and the target descent field is itself linear, $r_0(x) = x^\top \boldsymbol{\delta}_0$ for some $\boldsymbol{\delta}_0 \in \mathbb{R}^p$. Then $W_0(x) = 1$, $\phi_s(x) = x^\top \boldsymbol{\Delta}_s$, and, writing $\boldsymbol{\Sigma}_X = \mathbb{E}(XX^\top)$,*

$$\cos^2 \Theta_s = \frac{(\boldsymbol{\Delta}_s^\top \boldsymbol{\Sigma}_X \boldsymbol{\delta}_0)^2}{(\boldsymbol{\Delta}_s^\top \boldsymbol{\Sigma}_X \boldsymbol{\Delta}_s)(\boldsymbol{\delta}_0^\top \boldsymbol{\Sigma}_X \boldsymbol{\delta}_0)}.$$

If $\boldsymbol{\Sigma}_X = I$, this is the squared Euclidean angle between the source update and target descent coefficient; anisotropic covariates yield the covariance-weighted angle. Appendix D.1 gives the reduction and explains why deep models require target-functional geometry, not raw parameter geometry.

3 Geometric Transfer

The quantities needed for source selection are estimated from $\mathcal{D}_0^{\mathrm{fit}}$. We write $n_{\mathrm{fit}} = |\mathcal{D}_0^{\mathrm{fit}}|$ and index its observations by $i \in [n_{\mathrm{fit}}]$. For the multiclass implementation we use cross-entropy. Let $p_i = p_0(X_i)$, $W_i = \operatorname{diag}(p_i) - p_i p_i^\top$, $g_i = P_C \nabla_z \ell\{z_0(X_i), Y_i\}$, and $\phi_{ie} = P_C J_0(X_i) \boldsymbol{\Delta}_e$. Here W_i is regarded as an operator on $\mathbb{R}_0^C$. The empirical Gram matrix and source-target signal are

$$\hat{G}_{ee'} = \frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} \phi_{ie}^\top W_i \phi_{ie'} \text{ and } \hat{b}_e = -\frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} \phi_{ie}^\top g_i. \tag{3.1}$$

Conditional on the fixed source library, $\mathbb{E}(\hat{b}_e) = b_e$. We remove atoms satisfying $\hat{G}_{ee} \le \tau_G$, where the numerical tolerance $\tau_G > 0$ is fixed in advance, and rescale every retained pair $(\boldsymbol{\Delta}_e, \phi_e)$ by $\hat{G}_{ee}^{-1/2}$. After this step we relabel the retained atoms and reuse $\boldsymbol{\Delta}_e, \phi_e, \hat{\mathbf{G}}, \hat{\mathbf{b}}$, and $\mathcal{E}_s$ for the normalized dictionary. For stabilization, define

$$\hat{\mathbf{G}}_{\mathrm{st}}(\rho_G, \lambda_G) = (1 - \rho_G)\hat{\mathbf{G}} + \rho_G \operatorname{diag}(\hat{\mathbf{G}}) + \lambda_G I_M, \quad \rho_G \in [0, 1], \quad \lambda_G \ge 0, \tag{3.2}$$

where I_M is the $M \times M$ identity matrix after relabeling the retained atoms. Normalizing before stabilization prevents the ridge term from reintroducing arbitrary parameter scaling.

The normalization denominator $R_0 = \|r_0\|_0^2 = \mathbb{E}[g_0(X)^\top W_0(X)^\dagger g_0(X)]$ depends on the conditional score $g_0(X)$. Replacing $g_0(X_i)$ by the observed score g_i generally adds conditional label-noise energy and is therefore not a consistent estimator of R_0. We instead use K-fold cross-fitting, with $K \ge 2$ fixed in advance. Let $k(i)$ denote the fold containing observation i. A nuisance model is trained using only $\mathcal{D}_0^{\mathrm{fit}}$ outside fold $k(i)$ to estimate $g_0(X_i)$, producing $\hat{g}_{0,-k(i)}(X_i)$. For cross-entropy, let $\eta_0(x) = P_0(Y = \cdot \mid X = x)$. If $\hat{\eta}_{-k(i)}(X_i)$ is the corresponding out-of-fold class-probability estimate, then $\hat{g}_{0,-k(i)}(X_i) = p_i - \hat{\eta}_{-k(i)}(X_i)$. All nuisance-model classes and tuning rules are fixed in advance or selected using only the nuisance-training folds. Define

$$\hat{R}_{\mathrm{cf}} = \frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} \hat{g}_{0,-k(i)}(X_i)^\top W_i^\dagger \hat{g}_{0,-k(i)}(X_i). \tag{3.3}$$

Let $(\rho_G^{\mathrm{diag}}, \lambda_G^{\mathrm{diag}})$ be a stabilization pair fixed before validation and write $\hat{\mathbf{G}}_{\mathrm{diag}} = \hat{\mathbf{G}}_{\mathrm{st}}(\rho_G^{\mathrm{diag}}, \lambda_G^{\mathrm{diag}})$. If $\hat{R}_{\mathrm{cf}}$ is below a prespecified numerical tolerance, the normalized diagnostics are set to zero and the corresponding angles are not interpreted. Otherwise, with $[t]_0^1 = \min\{1, \max(0, t)\}$, define

$$\hat{c}_s^2 = \left[\frac{\hat{\mathbf{b}}_{\mathcal{E}_s}^\top \hat{\mathbf{G}}_{\mathrm{diag},\mathcal{E}_s\mathcal{E}_s}^\dagger \hat{\mathbf{b}}_{\mathcal{E}_s}}{\hat{R}_{\mathrm{cf}}} \right]_0^1, \quad \hat{\Theta}_s = \arccos(\sqrt{\hat{c}_s^2}), \quad \hat{\chi}_q = \left[\max_{\substack{A \subset [M] \\ |A| \le q}} \frac{\hat{\mathbf{b}}_A^\top \hat{\mathbf{G}}_{\mathrm{diag},AA}^\dagger \hat{\mathbf{b}}_A}{\hat{R}_{\mathrm{cf}}} \right]_0^1. \tag{3.4}$$

We also report $\hat{A}_e^{\mathrm{dir}} = \hat{b}_e / \sqrt{\hat{G}_{ee} \hat{R}_{\mathrm{cf}}}$ for $\hat{G}_{ee} > 0$, clipping it to $[-1, 1]$ only for display. When $\rho_G^{\mathrm{diag}} = \lambda_G^{\mathrm{diag}} = 0$, the numerators in (3.4) are the exact empirical source-span projection energies associated with the unshrunk Gram matrix. The denominator remains the cross-fitted estimator of the population energy R_0, so these ratios are normalized estimators, not literal finite-sample cosine measures. With stabilization, the numerators are regularized projection scores. Cross-fitting is used only for normalized diagnostics; estimation of the transferable coefficients uses the unbiased signal $\hat{\mathbf{b}}$. We compute $\hat{\chi}_q$ exactly for small source banks and otherwise use forward selection followed by local swaps. Appendix B gives the nuisance-estimation, Jacobian-vector, sketching, and complexity details.

For $\mathbf{a} = (a_1, \ldots, a_M)^\top \in \mathbb{R}^M$ and source-index set $\mathcal{E}_s \subset [M]$, let $\mathbf{a}_{\mathcal{E}_s} = (a_e : e \in \mathcal{E}_s)$ denote the subvector of coefficients associated with source s. For source s, let $w_s = \sqrt{|\mathcal{E}_s|}$ and define the empirical functional norm $\|h\|_{\hat{0}}^2 = n_{\mathrm{fit}}^{-1} \sum_{i=1}^{n_{\mathrm{fit}}} h(X_i)^\top W_i h(X_i)$. For trust radius $r > 0$, let

$$\mathcal{A}_r = \left\{ \mathbf{a} \in \mathbb{R}^M : \left\| \sum_{e=1}^M a_e \phi_e \right\|_{\hat{0}} \le r \right\} = \{ \mathbf{a} \in \mathbb{R}^M : \mathbf{a}^\top \hat{\mathbf{G}} \mathbf{a} \le r^2 \}. \tag{3.5}$$

For a candidate stabilization pair (ρ_G, λ_G), reusable source coefficients solve

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a} \in \mathcal{A}_r} \left\{ \frac{1}{2} \mathbf{a}^\top \hat{\mathbf{G}}_{\mathrm{st}}(\rho_G, \lambda_G) \mathbf{a} - \hat{\mathbf{b}}^\top \mathbf{a} + \lambda_{\mathrm{src}} \sum_{s=1}^S w_s \|\mathbf{a}_{\mathcal{E}_s}\|_2 + \lambda_{\mathrm{atom}} \|\mathbf{a}\|_1 \right\}. \tag{3.6}$$

The sparse-group penalty (Simon et al., 2013) selects source checkpoints while permitting atom-level sparsity within an active source, and the trust radius limits extrapolation beyond the local tangent approximation. Signed coefficients are allowed by default because the geometric object associated with an atom is the line that it spans. When literal forward reuse is required, $\mathcal{A}_r$ is intersected with $\{\mathbf{a} : a_e \ge 0 \; \forall e\}$; the sign policy is fixed before validation. The geometric interpretation is exact at the population, unregularized level. For every $\mathbf{a} \in \mathbb{R}^M$,

$$\frac{1}{2} \mathbf{a}^\top \mathbf{G} \mathbf{a} - \mathbf{b}^\top \mathbf{a} = \frac{1}{2} \|\Phi \mathbf{a} - r_0\|_0^2 - \frac{1}{2} \|r_0\|_0^2. \tag{3.7}$$

Thus (3.6) is a stabilized sparse empirical analogue of projecting a single target descent field onto the source-induced functional span. It is therefore distinct from multi-task model fusion. The off-diagonal Gram entries discount redundant source directions, whereas $\hat{\mathbf{b}}$ rewards alignment with target descent.

The source span need not contain all target structure. Let $\hat{A} = \{e : \hat{a}_e \ne 0\}$ up to a numerical solver tolerance fixed in advance. Let $\omega \in \mathbb{R}^{d_\omega}$ parameterize a prespecified parameter-efficient target update $u_\omega \in \mathbb{R}^p$, initialized so that $u_{\omega_0} = 0$, and define its tangent field on the fit observations by $\psi_{\omega i} = P_C J_0(X_i) u_\omega$. Also define the mean fit loss $\hat{L}_{\mathrm{fit}}(\boldsymbol{\theta}) = n_{\mathrm{fit}}^{-1} \sum_{i=1}^{n_{\mathrm{fit}}} \ell\{z_{\boldsymbol{\theta}}(X_i), Y_i\}$. Conditional on the selected source coefficients, the innovation parameters are estimated by

$$\hat{\omega} = \arg\min_{\omega} \left[\hat{L}_{\mathrm{fit}} \left(\boldsymbol{\theta}_0 + \sum_{e=1}^M \hat{a}_e \boldsymbol{\Delta}_e + u_\omega \right) + \lambda_\perp \sum_{e \in \hat{A}} \left\{ \frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} \phi_{ie}^\top W_i \psi_{\omega i} \right\}^2 + \lambda_\omega \|\omega\|_2^2 \right]. \tag{3.8}$$

The overlap penalty is evaluated in the same base-point tangent geometry as the source projection and discourages the target adapter from relearning selected source directions. It is an efficiency device, not a condition required by the projection or validation guarantees themselves.

Finally, every transferred candidate is compared with a target-only adapter trained solely on $\mathcal{D}_0^{\mathrm{fit}}$ using the same target-adapter architecture and optimization budget but no source atoms. Before inspecting $\mathcal{D}_0^{\mathrm{val}}$, we fix a finite candidate set Λ, a failure level $\delta \in (0,1)$, and a bounded gate loss $\ell_{\mathrm{gate}} \in [0, B]$, such as 0-1 error for classification or a prespecified clipped loss. Every candidate $\hat{\boldsymbol{\theta}}_\lambda$, $\lambda \in \Lambda$, and the target-only comparator $\hat{\boldsymbol{\theta}}_{\mathrm{TO}}$ are completely fitted using $\mathcal{D}_0^{\mathrm{fit}}$ before the validation split is inspected. Thus Λ may contain different choices of Gram stabilization, sparsity penalties, trust radii, innovation penalties, and other prespecified tuning parameters without using validation data to fit any candidate. For validation observation i and candidate λ, let $D_i(\lambda) = \ell_{\mathrm{gate}}\{f_{\hat{\boldsymbol{\theta}}_\lambda}(X_i), Y_i\} - \ell_{\mathrm{gate}}\{f_{\hat{\boldsymbol{\theta}}_{\mathrm{TO}}}(X_i), Y_i\}$. Let $n_{\mathrm{val}} = |\mathcal{D}_0^{\mathrm{val}}| \geq 2$, $\bar{D}_\lambda = n_{\mathrm{val}}^{-1} \sum_{i=1}^{n_{\mathrm{val}}} D_i(\lambda)$, and $\hat{v}_\lambda = (n_{\mathrm{val}} - 1)^{-1} \sum_{i=1}^{n_{\mathrm{val}}} \{D_i(\lambda) - \bar{D}_\lambda\}^2$. Because $D_i(\lambda) \in [-B, B]$, the one-sided empirical-Bernstein upper bound (Maurer and Pontil, 2009) is

$$U_\lambda = \bar{D}_\lambda + \sqrt{\frac{2\hat{v}_\lambda \log(2|\Lambda|/\delta)}{n_{\mathrm{val}}}} + \frac{14B \log(2|\Lambda|/\delta)}{3(n_{\mathrm{val}} - 1)}. \tag{3.9}$$

Among candidates satisfying $U_\lambda \leq 0$, we deploy the one with smallest $\bar{D}_\lambda$; if none qualifies, we deploy the target-only comparator. Proposition 4.4 shows that the resulting procedure has confidence level at least $1 - \delta$ for the prespecified bounded average target-risk comparison. Algorithm 1 gives the complete estimation procedure from geometry construction through gated deployment and fallback.

Algorithm 1 GEOTRANSFER: target-conditioned geometric transfer

Require: Base $\boldsymbol{\theta}_0$; source checkpoints/adapters $\{\boldsymbol{\theta}_s\}_{s=1}^S$; target $\mathcal{D}_0^{\mathrm{fit}}$, $\mathcal{D}_0^{\mathrm{val}}$; fixed atomization, diagnostic budget q, sign policy, candidate set Λ, and failure level δ.

1: Form source atoms $\{\boldsymbol{\Delta}_e\}_{e=1}^M$ on the shared coordinates.
2: On $\mathcal{D}_0^{\mathrm{fit}}$ compute p_i, W_i, g_i and $\phi_{ie} = P_C J_0(X_i)\boldsymbol{\Delta}_e$; estimate $(\hat{\mathbf{G}}, \hat{\mathbf{b}})$ by (3.1).
3: Remove numerically null atoms, normalize the retained atom dictionary, and form the prespecified diagnostic Gram stabilization.
4: Cross-fit $\hat{g}_{0,-k(i)}$ and compute $\hat{R}_{\mathrm{cf}}$, $\hat{\Theta}_s$, directed alignments, and $\hat{\chi}_q$ using (3.3)-(3.4).
5: Using only $\mathcal{D}_0^{\mathrm{fit}}$, fit every $\lambda \in \Lambda$ by solving (3.6) and, when specified, (3.8); also fit the capacity-matched target-only comparator.
6: On untouched $\mathcal{D}_0^{\mathrm{val}}$ compute U_λ by (3.9).
7: Deploy the accepted candidate with smallest $\bar{D}_\lambda$; if no candidate satisfies $U_\lambda \leq 0$, deploy target-only.
8: **return** Predictor, selected atoms, geometry diagnostics, innovation norm, and fallback indicator.

4 THEORETICAL PROPERTIES

We state the main results so that the geometric, estimation, and validation claims are distinct. Appendix A gives the assumptions and proofs. For $x \in \mathcal{X}$, let $\bar{\ell}_x(z) = \mathbb{E}\{\ell(z, Y) \mid X = x\}$ denote the conditional target risk, and define $\mathcal{T}(h) = \mathbb{E}_{P_0}\left[\ell\{z_0(X) + h(X), Y\} - \ell\{z_0(X), Y\}\right]$ and $\|h\|_\infty = \operatorname{ess\,sup}_{x \sim P_0^X} \|h(x)\|_2$. Assume $\bar{\ell}_x$ is three-times continuously differentiable throughout the Euclidean ball of radius $r_{\mathrm{loc}} > 0$ around $z_0(x)$ for P_0^X-almost every x. On $\mathbb{R}_0^C$, assume $0 < \underline{\kappa} \leq \lambda_{\min}\{W_0(x)\} \leq \lambda_{\max}\{W_0(x)\} \leq \bar{\kappa} < \infty$, and assume the operator norm of the third derivative of $\bar{\ell}_x$ is uniformly bounded by $L_3 < \infty$ throughout this neighborhood. Under these local regularity conditions, the target-risk increment admits a quadratic expansion whose optimizer over each source span is the projection of the target descent field onto that span.

Theorem 4.1. *For any $A \subset [M]$, let $\mathcal{V}_A = \operatorname{span}\{\phi_e : e \in A\}$ and $h_A^\star = \operatorname{Proj}_{\mathcal{V}_A} r_0$. Every $h \in \mathcal{V}_A$ satisfying $\|h\|_\infty \leq r_{\mathrm{loc}}$ obeys*

$$\mathcal{T}(h) = -\langle h, r_0 \rangle_0 + \frac{1}{2}\|h\|_0^2 + \mathrm{Rem}(h), \quad |\mathrm{Rem}(h)| \leq \frac{L_3}{6}\mathbb{E}\|h(X)\|_2^3.$$

The unique functional perturbation minimizing the quadratic part over $\mathcal{V}_A$ is $h_A^\star$, although its coefficient representation need not be unique. Define $\cos^2\Theta(A) = \|\operatorname{Proj}_{\mathcal{V}_A} r_0\|_0^2 / \|r_0\|_0^2$. The corresponding quadratic improvement over no transfer is

$$\mathcal{G}(A) = \frac{1}{2}\|h_A^\star\|_0^2 = \frac{1}{2}\|r_0\|_0^2 \cos^2\Theta(A).$$

If $\|h_A^\star\|_\infty \leq r_{\mathrm{loc}}$, then $|-\mathcal{T}(h_A^\star) - \mathcal{G}(A)| \leq \frac{L_3}{6}\mathbb{E}\|h_A^\star(X)\|_2^3$. Thus, χ_q is exactly the largest fraction of local natural-descent energy representable by at most q source atoms in the quadratic target model.

Theorem 4.1 separates target signal strength from geometric compatibility. A source may have a large parameter update but negligible local gain if its target-functional span is nearly orthogonal to r_0; adding a source whose functional directions are represented by the active span gives little additional quadratic gain. This characterization also raises a basic invariance question: whether the resulting transfer geometry depends on the particular parameterization used to represent the same predictor.

Proposition 4.2. *Let $\vartheta = \psi(\boldsymbol{\theta})$ be a C^2 local diffeomorphism on a neighborhood containing $\boldsymbol{\theta}_0$ and the source endpoints, with bounded first and second derivatives there. Let $\tilde{z}_\vartheta = z_{\psi^{-1}(\vartheta)}$, $\tilde{J}_0 = \frac{\partial \tilde{z}_\vartheta}{\partial \vartheta^\top}\big|_{\vartheta=\psi(\boldsymbol{\theta}_0)}$, and define $\tilde{\boldsymbol{\Delta}}_e = \psi(\boldsymbol{\theta}_0 + \boldsymbol{\Delta}_e) - \psi(\boldsymbol{\theta}_0)$, $\tilde{\phi}_e(X) = P_C \tilde{J}_0(X)\tilde{\boldsymbol{\Delta}}_e$. Then, uniformly on compact target-input sets, $\tilde{\phi}_e(X) = \phi_e(X) + O(\|\boldsymbol{\Delta}_e\|_2^2)$. Suppose the tangent dictionaries for diagnostics have locally constant rank, positive Gram eigenvalues bounded below by $\kappa_G > 0$, and nonzero atom norms for directed alignments bounded away from zero. Then the Gram matrices, source-span projections, principal angles, directed alignments, and χ_q are continuous under these perturbations. Thus the geometry is first-order invariant under smooth reparameterization. By contrast, an invertible linear reparameterization can move the Euclidean cosine between two noncollinear parameter directions arbitrarily close to any value in $[-1, 1]$ while preserving predictors.*

Proposition 4.2 transports a fixed atomization through ψ; it does not claim invariance when a coordinate-dependent atomization rule is redefined after reparameterization. This distinction parallels the atomization dependence of sparse completeness in (2.6).

For estimation, condition on the fitted source library. Let $h^\star = \mathrm{Proj}_{\mathcal{V}} r_0 = \Phi \mathbf{a}^\star$ have a coefficient representation supported on q atoms, and define its active source set by $\mathcal{S}^\star = \{s : \mathcal{E}_s \cap \mathrm{supp}(\mathbf{a}^\star) \neq \varnothing\}$ and $D_\star = \sum_{s \in \mathcal{S}^\star} |\mathcal{E}_s|$. The normal equations give $\mathbf{G}\mathbf{a}^\star = \mathbf{b}$. With $g(X, Y) = P_C \nabla_z \ell\{z_0(X), Y\}$, define the normal-equation score $\xi_e = -\phi_e(X)^\top \{g(X, Y) + W_0(X) h^\star(X)\}$. Then $\mathbb{E}(\xi_e) = 0$. Assume the normalized tangent fields and scores satisfy the sub-Gaussian envelope conditions in Appendix A with stochastic scale σ, and assume the sparse-group restricted-eigenvalue condition there holds with constant $\kappa_{\mathrm{RE}} > 0$. These conditions control empirical fluctuations in the source span and yield the following finite-sample sparse-estimation guarantee.

Theorem 4.3. *Consider the exact-tangent, unshrunk sparse-group estimator obtained from* (3.6) *with $\rho_G = \lambda_G = 0$ and $\mathcal{A}_r = \mathbb{R}^M$. For any $\delta \in (0, 1)$, choose $\lambda_{\mathrm{src}} \geq c_1 \sigma \sqrt{[1 + \log(S/\delta)]/n_{\mathrm{fit}}}$ and $\lambda_{\mathrm{atom}} \geq c_2 \sigma \sqrt{\log(M/\delta)/n_{\mathrm{fit}}}$, for sufficiently large universal constants $c_1, c_2 > 0$. If the empirical restricted-eigenvalue event specified in Appendix A holds, then an exact optimizer $\hat{\mathbf{a}}$ satisfies, with probability at least $1 - \delta$,*

$$\|\Phi(\hat{\mathbf{a}} - \mathbf{a}^\star)\|_0^2 \leq \frac{C_{\mathrm{or}}}{\kappa_{\mathrm{RE}}^2} \left\{ D_\star \lambda_{\mathrm{src}}^2 + q \lambda_{\mathrm{atom}}^2 \right\},$$

where $C_{\mathrm{or}} > 0$ is a universal constant. Hence the leading stochastic term is

$$O\left(\frac{\sigma^2 [D_\star \{1 + \log(S/\delta)\} + q \log(M/\delta)]}{\kappa_{\mathrm{RE}}^2 n_{\mathrm{fit}}} \right).$$

Appendix A gives sufficient sample-size conditions for the empirical restricted-eigenvalue event, with approximate-sparsity, finite-optimization-tolerance, and shrinkage/sketching perturbation extensions.

The rate depends on reusable-span complexity, not on the ambient neural parameter dimension. If both $\|h^\star\|_\infty \leq r_{\mathrm{loc}}$ and $\|\Phi\hat{\mathbf{a}}\|_\infty \leq r_{\mathrm{loc}}$, then Theorem 4.1 also gives

$$\mathcal{T}(\Phi\hat{\mathbf{a}}) - \mathcal{T}(h^\star) = \frac{1}{2}\|\Phi(\hat{\mathbf{a}} - \mathbf{a}^\star)\|_0^2 + \mathrm{Rem}(\Phi\hat{\mathbf{a}}) - \mathrm{Rem}(h^\star). \tag{4.1}$$

Thus the oracle bound controls the quadratic component of target-risk difference relative to the population source-span projection, with cubic remainders accounting for local nonlinearity. The theorem conditions on the fitted source library and analyzes estimation within its functional span; Proposition 4.4 separately covers deployment of the practical candidates fitted without validation observations.

Proposition 4.4. *For the risk gate, define $R_{\mathrm{gate}}(\boldsymbol{\theta}) = \mathbb{E}_{P_0} \ell_{\mathrm{gate}}\{f_{\boldsymbol{\theta}}(X), Y\}$. Condition on the fit data and hence on all candidate predictors $\{\hat{\boldsymbol{\theta}}_\lambda : \lambda \in \Lambda\}$ and the target-only comparator $\hat{\boldsymbol{\theta}}_{\mathrm{TO}}$, all of which are fitted without $\mathcal{D}_0^{\mathrm{val}}$. If the validation observations are independent and identically distributed from P_0 and $D_i(\lambda) \in [-B, B]$, then, with probability at least $1 - \delta$, $R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_\lambda) - R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{TO}}) \leq U_\lambda$ simultaneously for every $\lambda \in \Lambda$. Thus, Algorithm 1 satisfies $\mathbb{P}\{R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{deploy}}) > R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{TO}})\} \leq \delta$.*

5 NUMERICAL STUDIES

The numerical study isolates the theoretical claims. Simulations test source ranking, reparameterization stability, completeness, and target innovation; vision and language libraries test source-data-free few-shot transfer under matched target-data and adaptation budgets, with source-bank contamination studied separately in Appendix F. Candidate grids exclude test data, and summaries average independent seeds. In empirical ranking experiments, each source checkpoint is treated as one atom. For each source, the scalar coefficient $\alpha \in [-2, 2]$ is selected using the target validation split, and realized single-source gain is then evaluated on the held-out target test split. In the controlled simulations, we instead use the population single-source utility defined in Appendix C, where $\alpha \in [-2, 2]$ is optimized against the independent oracle population risk. Multi-source GEOTRANSFER uses the prespecified block or adapter atomization. Appendix Table F.1 reports nonnegative-coefficient sensitivity.

5.1 MONTE CARLO SIMULATIONS

We construct a nonlinear teacher-student classification problem with a two-layer ReLU network and 128-dimensional hidden representation. A base model is pretrained on a broad mixture. Twenty-four source checkpoints are fine-tuned along eight latent functional factors; useful, redundant, orthogonal, and forward-antagonistic directions are present. The target uses a sparse three-factor combination plus innovation absent from every source. This separates source coverage from redundancy. To test coordinate dependence, we apply function-preserving rescalings $W_1 \mapsto DW_1,\ W_2 \mapsto W_2 D^{-1}$ for positive diagonal D with condition number 1 to 10^4. This changes parameter geometry while preserving the network function. Target sizes are $n_{\text{fit}} \in \{32, 64, 128, 256\}$, with equal validation size and $10{,}000$ test observations. Appendix C gives the complete data-generating process.

Table 1: **Source ranking and transfer.** Rank ρ is Spearman correlation between source score and single-source gain ("–" when ranking is undefined); Top-3 regret is the loss gap to the oracle three-source subset under a three-source budget; NT is the fraction of replications worse than target-only; Reparam. drift is mean absolute score change after rescaling. Standard errors are parenthesized.

Method	Rank $\rho \uparrow$	Top-3 regret$\downarrow$	Test acc.$\uparrow$	NT$\downarrow$	Reparam. drift$\downarrow$
Euclidean task-vector cosine	.31(.03)	.084(.006)	71.8(.4)	.247	.41
Parameter distance	.18(.03)	.103(.007)	70.9(.4)	.284	.36
H-score	.49(.02)	.061(.005)	73.0(.3)	.176	.02
LEEP	.55(.02)	.052(.004)	73.6(.3)	.151	.01
LogME	.58(.02)	.048(.004)	73.8(.3)	.144	.01
Diagonal-Fisher cosine	.68(.02)	.037(.003)	74.7(.3)	.103	.08
Target gradient cosine	.72(.02)	.032(.003)	75.1(.3)	.091	.19
GEOTRANSFER angle only	.86(.01)	.017(.002)	76.4(.2)	.052	.01
GEOTRANSFER sparse projection	–	.010(.001)	77.3(.2)	.031	–
GEOTRANSFER + innovation + gate	–	.011(.001)	78.1(.2)	.008	–
Oracle source span	1.00	0	78.6(.2)	0	–

Table 1 separates ranking from prediction. Raw parameter cosines degrade under hidden-unit rescaling although functions are unchanged. Representation scores are more stable but ignore source-source redundancy. Accurate source ranking does not ensure that sources are complementary or cover the target descent field. Source-span geometry improves selection because off-diagonal Gram entries identify redundancy while the signal vector records target orientation. Innovation recovers structure omitted from the source library, and the gate sharply reduces negative transfer.

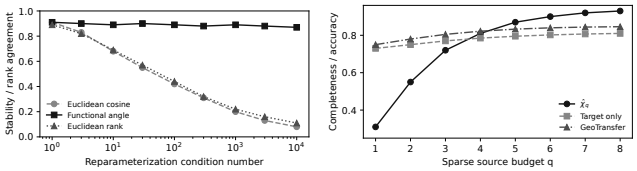


Figure 2: **Controlled geometry diagnostics.** Left: ranking stability under function-preserving reparameterization. Right: estimated completeness $\hat{\chi}_q$ tracks realized target gain.

Completeness provides information beyond source ranking; see Figure 2. It measures how much of the target descent field is captured by the best sparse source span under a prespecified source-cardinality budget q. Small $\hat{\chi}_q$ indicates limited coverage by any q-atom source span, whereas large $\hat{\chi}_q$ indicates substantial sparse source coverage. Appendix Table C.4 examines this relationship across changes in library coverage, redundancy, and innovation strength.

5.2 Empirical Applications

We use three classes of source-library setting: CLIP task libraries, cross-domain vision libraries, and language adapter libraries. CLIP-8 denotes an eight-task suite with a common ViT-B/32 base, whereas CLIP-14 denotes a fourteen-task suite with a common ViT-L/14 base; each target checkpoint is withheld and $k \in \{4, 16, 32, 64\}$ labeled target fit examples per class are used, with $k = 32$ serving as the reference condition for Table 2. PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019) use common-base domain libraries with DomainBed-style splits (Gulrajani and Lopez-Paz, 2021); we use $k \in \{16, 64\}$ target fit examples per class, with $k = 64$ as the reference condition. The language experiments use LoRA/adapter experts (Houlsby et al., 2019; Hu et al., 2022) from a common backbone and $n_{\text{fit}} \in \{32, 64, 128, 256\}$ labeled target examples, with $n_{\text{fit}} = 128$ as the reference condition. No method combines architectures or uses source training examples. Methods share identical target-label budgets within each experimental condition. Appendix E details the benchmark construction and Appendix F provides extended comparisons.

Table 2: **Source-data-free transfer on foundation-model libraries.** Panel A reports source ranking and Panel B few-shot performance. Top-1 regret is loss gap to the oracle source; NT is paired negative-transfer frequency relative to capacity-matched target-only. Both are macro-averaged across the six empirical library regimes after computing each statistic within each matched target replicate.

Panel A: predicting source utility before transfer

Method	CLIP-8 ρ	CLIP-14 ρ	PACS ρ	OfficeHome ρ	DomainNet ρ	Language ρ	Top-1 regret↓
H-score	.43	.39	.51	.46	.48	.37	.062
LogME	.59	.54	.63	.57	.57	.50	.045
PACTran	.61	.56	.64	.59	.59	.53	.042
Euclidean task-vector cosine	.36	.31	.42	.39	.38	.28	.073
Diagonal Fisher angle	.66	.61	.69	.64	.65	.58	.036
GeoTransfer principal angle	.79	.76	.82	.78	.80	.71	.023

Panel B: target performance under matched adaptation budget

Method	CLIP-8 Acc.	CLIP-14 Acc.	PACS Acc.	OfficeHome Acc.	DomainNet Acc.	Lang. score	NT↓
Target-only adapter	76.8	73.4	86.0	69.7	52.4	61.5	–
TIES	77.8	74.3	87.0	70.6	53.2	62.2	.171
TATR	78.2	74.7	87.3	70.9	53.7	62.5	.130
MASS-style adaptive subspace	78.3	74.8	87.4	71.0	54.0	62.6	.118
Sparse target-loss weights	78.6	75.0	87.6	71.2	54.1	62.8	.106
GeoTransfer projection	79.2	75.8	88.2	72.0	54.7	63.5	.060
GeoTransfer + innovation	79.8	76.4	88.8	72.8	55.3	64.2	.053
GeoTransfer + innovation + gate	79.6	76.2	88.6	72.6	55.1	64.0	.019

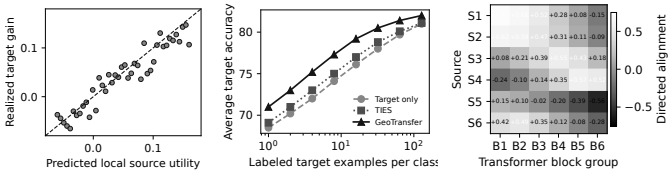


Figure 3: **Empirical transfer geometry on CLIP-8.** Left: predicted versus realized source utility. Middle: few-shot accuracy; GeoTransfer denotes projection + innovation + gate. Right: directed alignment from (2.7) changes sign across six prespecified diagnostic groups $B_1, \ldots, B_6$; this partition is visualization-only, while predictive transfer uses the three-group atomization of Appendix E.

Panel A shows that the principal angle better preserves realized source ordering. Panel B reports a prespecified reference target-data condition for each library: 32 labeled fit examples per class for CLIP-8 and CLIP-14, 64 per class for PACS, OfficeHome, and DomainNet, and 128 labeled target fit examples for the language suite. All methods within a library use identical target-label budgets and the same disjoint fit, validation, and test protocol. The results also show why ranking alone is insufficient; projection combines complementary sources, innovation captures residual structure, and the gate sharply reduces negative transfer. Figure 3 visualizes utility prediction, few-shot scaling, and block-level sign changes; Appendix Table F.2 gives the full comparison.

6 Concluding Remarks

GeoTransfer measures target adaptation supported by source functions, reuses it through sparse projection, and learns residual innovation with a target-only fallback. The framework turns source libraries into target-conditioned adaptation resources. Because the geometry is local, the trust radius and held-out gate cannot eliminate failures for distant checkpoints. Extending the framework beyond first-order tangents to nonlocal source interactions is a direction for future work.

AI USE STATEMENT

Generative AI tools were employed only to improve phrasing, readability, and editorial consistency. They played no role in formulating the methodology or deriving the theoretical results. AI-assisted revisions were independently reviewed and revised by the authors to ensure accuracy, clarity, and appropriate attribution. Responsibility for the manuscript and all of its claims rests entirely with the authors.

REPRODUCIBILITY STATEMENT

The appendix presents the statistical assumptions and complete proofs, details the Jacobian-vector and randomized-sketch estimators for constructing the transfer Gram matrix, specifies optimization and tuning procedures, and documents the controlled data-generating process and benchmark protocols. To preserve anonymity during double-blind review, code for source-atom construction, transfer-geometry estimation, and reproduction of all figures and tables will be released after peer review.

REFERENCES

Ainsworth, S. K., Hayase, J., and Srinivasa, S. (2023). Git Re-Basin: Merging models modulo permutation symmetries. *International Conference on Learning Representations*.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251-276.

Bao, Y., Li, Y., Huang, S.-L., Zhang, L., Zheng, L., Zamir, A., and Guibas, L. (2019). An information-theoretic approach to transferability in task transfer learning. *IEEE International Conference on Image Processing*, 2309-2313.

Crisostomi, D., Zirilli, A., Gargiulo, A. A., Bucarelli, M. S., Scardapane, S., Silvestri, F., Masi, I., and Rodolà, E. (2026). MASS: MoErging through adaptive subspace selection. *International Conference on Learning Representations*.

Daheim, N., Möllenhoff, T., Ponti, E. M., Gurevych, I., and Khan, M. E. (2024). Model merging by uncertainty-based gradient matching. *International Conference on Learning Representations*.

Ding, N., Chen, X., Levinboim, T., Changpinyo, B., and Soricut, R. (2022). PACTran: PAC-Bayesian metrics for estimating the transferability of pretrained models to classification tasks. *European Conference on Computer Vision*, 252-268.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. (2017). Sharp minima can generalize for deep nets. *Proceedings of the 34th International Conference on Machine Learning*, 70, 1019-1028.

Gargiulo, A. A., Crisostomi, D., Bucarelli, M. S., Scardapane, S., Silvestri, F., and Rodolà, E. (2025). Task singular vectors: Reducing task interference in model merging. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18695-18705.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of DNNs. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 8803-8812.

Gu, T., Han, Y., and Duan, R. (2025). Robust angle-based transfer learning in high dimensions. *Journal of the Royal Statistical Society Series B*, 87(3), 723-745.

Gulrajani, I. and Lopez-Paz, D. (2021). In search of lost domain generalization. *International Conference on Learning Representations*.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. *Proceedings of the 36th International Conference on Machine Learning*, 97, 2790-2799.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Schmidt, L., Hajishirzi, H., and Farhadi, A. (2023). Editing models with task arithmetic. *International Conference on Learning Representations*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 8580-8589.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. (2017). Deeper, broader and artier domain generalization. *IEEE International Conference on Computer Vision*, 5543-5551.

Martens, J. (2020). New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146), 1-76.

Matena, M. and Raffel, C. (2022). Merging models with Fisher-weighted averaging. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 17703-17716.

Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. *Proceedings of the 22nd Annual Conference on Learning Theory*.

Mehra, A., Zhang, Y., and Hamm, J. (2024). Understanding the transferability of representations via task-relatedness. *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 116513-116546.

Nguyen, C., Hassner, T., Seeger, M., and Archambeau, C. (2020). LEEP: A new measure to evaluate transferability of learned representations. *Proceedings of the 37th International Conference on Machine Learning*, 119, 7294-7305.

Nguyen, V. D., Nguyen, M. H., Nguyen, L. Q., Teo, R. S. Y., Nguyen, T. M., and Tran, L. D. (2025). CAMEx: Curvature-aware merging of experts. *International Conference on Learning Representations*.

Osial, M., Wójcik, B., Zieliński, B., and Cygert, S. (2025). Efficient multi-source knowledge transfer by model merging. *arXiv:2508.19353*.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. (2019). Moment matching for multi-source domain adaptation. *IEEE/CVF International Conference on Computer Vision*, 1406-1415.

Shi, K., Wen, Y., and Liu, W. (2025). Model merging with functional dual anchors. *arXiv:2510.21223*.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2), 231-245.

Sun, W., Li, Q., Wang, W., Geng, Y., and Li, B. (2025). Task arithmetic in trust region: A training-free model merging approach to navigate knowledge conflicts. *Proceedings of the 33rd ACM International Conference on Multimedia* , 5178-5187.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. *IEEE Conference on Computer Vision and Pattern Recognition*, 5385-5394.

Wang, J., Ye, Z., and Yin, W. (2026). Functionality-oriented LLM merging on the Fisher-Rao manifold. *arXiv:2603.04972*.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., and Schmidt, L. (2022). Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *Proceedings of the 39th International Conference on Machine Learning*, 162, 23965-23998.

Yadav, P., Tam, D., Choshen, L., Raffel, C., and Bansal, M. (2023). TIES-Merging: Resolving interference when merging models. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 7093-7115.

You, K., Liu, Y., Wang, J., and Long, M. (2021). LogME: Practical assessment of pre-trained models for transfer learning. *Proceedings of the 38th International Conference on Machine Learning*, 139, 12133-12143.

APPENDIX

A THEORETICAL DETAILS AND PROOFS

This part states the regularity conditions and gives complete proofs for the results in Section 4. Note that P_0 denotes the target law, P_0^X its X-marginal, $\mathbb{R}_0^C = \{v \in \mathbb{R}^C : \mathbf{1}^\top v = 0\}$ is the centered-logit space, P_C is the orthogonal projector onto $\mathbb{R}_0^C$, and $\{\phi_e : e \in [M]\}$ is the source-induced tangent dictionary. For $A \subset [M]$, v_A denotes the subvector of v indexed by A, $\mathbf{G}_{AA}$ the corresponding principal submatrix, and $\mathrm{supp}(v) = \{e : v_e \neq 0\}$. All inverses, square roots, and Moore-Penrose inverses of curvature operators below are understood on $\mathbb{R}_0^C$. Probability statements condition on the fitted source library and its prespecified atomization. Thus source checkpoints may be random through their own training process, but are treated as fixed when analyzing the low-target-sample problem.

A.1 FUNCTIONAL GEOMETRY, LOCAL REGULARITY, AND PROJECTION IDENTITIES

For $x \in \mathcal{X}$, define the conditional target risk $\bar{\ell}_x(z) = \mathbb{E}\{\ell(z, Y) \mid X = x\}$. At the base logits $z_0(x) = z_{\boldsymbol{\theta}_0}(x)$, define the centered conditional score and curvature by $g_0(x) = P_C \nabla_z \bar{\ell}_x\{z_0(x)\}$ and $W_0(x) = P_C \nabla_z^2 \bar{\ell}_x\{z_0(x)\} P_C$. These definitions coincide with those in Section 2 and make explicit that both quantities are restricted to the identifiable centered-logit coordinates. For measurable $u, v : \mathcal{X} \to \mathbb{R}_0^C$ with finite weighted second moments, write $\langle u, v \rangle_0 = \mathbb{E}\{u(X)^\top W_0(X) v(X)\}$, $\|u\|_0^2 = \langle u, u \rangle_0$, and define $\|u\|_{L_3(P_0^X)}^3 = \mathbb{E}\|u(X)\|_2^3$, $\|u\|_\infty = \operatorname{ess\,sup}_{x \sim P_0^X} \|u(x)\|_2$. The space $\mathcal{H}_0$ is the completion, modulo zero-norm elements, under $\|\cdot\|_0$. Under the curvature condition stated below, this norm is nondegenerate on the centered coordinates, so $\mathcal{H}_0$ is an ordinary Hilbert space. We next state the sampling and local regularity conditions used throughout the theoretical analysis.

Assumption 1. *The fit observations (X_i, Y_i), $i \in [n_{\mathrm{fit}}]$, are independent and identically distributed (i.i.d.) from P_0. The source library and atomization are fixed independently of the target labels in $\mathcal{D}_0^{\mathrm{fit}}$. The validation observations by the risk gate are independent of the fit sample and are i.i.d. from P_0.*

Assumption 1 separates randomness in the target sample from the previously fitted source library. Conditioning on the source library therefore does not require the source checkpoints themselves to be nonrandom. It permits source checkpoints trained on external datasets.

Assumption 2. *There exist constants $0 < \underline{\kappa} \leq \bar{\kappa} < \infty$, $r_{\mathrm{loc}} > 0$, and $L_3 < \infty$ such that, for P_0^X-almost every x, $\underline{\kappa} I \preceq W_0(x) \preceq \bar{\kappa} I$ on $\mathbb{R}_0^C$. Moreover, $\bar{\ell}_x$ is three-times continuously differentiable throughout the Euclidean ball $\{z_0(x) + h : h \in \mathbb{R}_0^C,\ \|h\|_2 \leq r_{\mathrm{loc}}\}$, and throughout this ball $\|\nabla_z^3 \bar{\ell}_x(z)\|_{\mathrm{op}} \leq L_3$. For a trilinear tensor T, its operator norm is*

$$\|T\|_{\mathrm{op}} = \sup_{\|u\|_2 = \|v\|_2 = \|w\|_2 = 1} |T[u, v, w]|.$$

The lower curvature bound is needed only on the identifiable centered-logit space. For softmax cross-entropy, $W_0(x) = \mathrm{diag}\{p_0(x)\} - p_0(x) p_0(x)^\top$ is positive definite on $\mathbb{R}_0^C$ whenever all class probabilities are positive; Assumption 2 strengthens this to uniform positivity over the target distribution. The upper bound controls equivalence between Euclidean and curvature-weighted output norms. The third-derivative bound controls local nonlinearity and is not a global convexity condition.

Under Assumption 2, $W_0(x)$ is invertible on $\mathbb{R}_0^C$ for P_0^X-almost every x, and its Moore-Penrose inverse there equals the ordinary inverse. Hence $r_0(x) = -W_0(x)^\dagger g_0(x) = -W_0(x)^{-1} g_0(x)$ is well defined and satisfies $W_0(x) r_0(x) = -g_0(x)$. Consequently, for every $h \in \mathcal{H}_0$,

$$-\mathbb{E}\{g_0(X)^\top h(X)\} = \mathbb{E}\{h(X)^\top W_0(X) r_0(X)\} = \langle h, r_0 \rangle_0 . \tag{A.1}$$

For each source atom define $\phi_e(x) = P_C J_0(x) \boldsymbol{\Delta}_e$. For $A \subset [M]$, let $\Phi_A : \mathbb{R}^{|A|} \to \mathcal{H}_0$, $\Phi_A a_A = \sum_{e \in A} a_e \phi_e$, and $\mathcal{V}_A = (\Phi_A) = \mathrm{span}\{\phi_e : e \in A\}$. Because A is finite, $\mathcal{V}_A$ is finite dimensional and therefore closed. Its Gram matrix and signal vector are $\mathbf{G}_{AA} = \big(\langle \phi_e, \phi_{e'} \rangle_0\big)_{e,e' \in A}$ and $\mathbf{b}_A = \big(\langle \phi_e, r_0 \rangle_0\big)_{e \in A}$. Equivalently, $\mathbf{G}_{AA} = \Phi_A^* \Phi_A$ and $\mathbf{b}_A = \Phi_A^* r_0$, where Φ_A^* denotes the Hilbert-space adjoint. In particular, $\mathbf{G}_{AA} \succeq 0$. These identities provide the finite-dimensional representation of the source-span geometry used in the projection argument below.

Lemma A.1. *Under Assumption 2, every measurable $h : \mathcal{X} \to \mathbb{R}_0^C$ satisfying $\|h\|_\infty \le r_{\mathrm{loc}}$ obeys*

$$\mathcal{T}(h) = -\langle h, r_0\rangle_0 + \frac{1}{2}\|h\|_0^2 + R_3(h), \text{ where } |R_3(h)| \le \frac{L_3}{6}\|h\|_{L_3(P_0^X)}^3.$$

Proof. We fix an x for which Assumption 2 holds and define $f_x(t) = \bar{\ell}_x\{z_0(x) + th(x)\},\ t \in [0,1]$. Because $\|h(x)\|_2 \le r_{\mathrm{loc}}$, the entire segment $z_0(x) + th(x)$ lies in the neighborhood on which $\bar{\ell}_x$ is three-times continuously differentiable. By the chain rule,

$$f_x'(0) = \nabla_z \bar{\ell}_x\{z_0(x)\}^\top h(x).$$

Since $h(x) \in \mathbb{R}_0^C = P_C\mathbb{R}^C$,

$$f_x'(0) = g_0(x)^\top h(x).$$

Similarly,

$$f_x''(0) = h(x)^\top W_0(x) h(x).$$

For $t \in [0,1]$,

$$f_x'''(t) = \nabla_z^3 \bar{\ell}_x\{z_0(x) + th(x)\}[h(x), h(x), h(x)].$$

Taylor's theorem with integral remainder gives

$$f_x(1) - f_x(0) = f_x'(0) + \frac{1}{2}f_x''(0) + \frac{1}{2}\int_0^1 (1-t)^2 f_x'''(t)\, dt.$$

Hence

$$\bar{\ell}_x\{z_0(x) + h(x)\} - \bar{\ell}_x\{z_0(x)\} = g_0(x)^\top h(x) + \frac{1}{2}h(x)^\top W_0(x) h(x) + R_{3,x}(h),$$

where

$$R_{3,x}(h) = \frac{1}{2}\int_0^1 (1-t)^2 \nabla_z^3 \bar{\ell}_x\{z_0(x) + th(x)\}[h(x), h(x), h(x)]\, dt.$$

By the operator-norm bound,

$$|R_{3,x}(h)| \le \frac{L_3}{2}\|h(x)\|_2^3 \int_0^1 (1-t)^2 dt = \frac{L_3}{6}\|h(x)\|_2^3.$$

Taking expectation over X gives

$$\mathcal{T}(h) = \mathbb{E}\{g_0(X)^\top h(X)\} + \frac{1}{2}\mathbb{E}\{h(X)^\top W_0(X) h(X)\} + R_3(h) \text{ with } |R_3(h)| \le \frac{L_3}{6}\mathbb{E}\|h(X)\|_2^3.$$

Finally, (A.1) converts the first term to $-\langle h, r_0\rangle_0$, and the second term is $\|h\|_0^2/2$.

□

Lemma A.1 converts target risk into the local quadratic geometry used throughout the paper. The first two terms are exact at the base point, while all departure from that quadratic model is isolated in the cubic remainder. This expansion is the key ingredient in the proof of Theorem 4.1, which identifies the optimal source-supported perturbation with the Hilbert-space projection of the target descent field.

Proof of Theorem 4.1. Define $Q(h) = -\langle h, r_0\rangle_0 + \frac{1}{2}\|h\|_0^2$. Completing the square gives

$$Q(h) = \frac{1}{2}\|h - r_0\|_0^2 - \frac{1}{2}\|r_0\|_0^2. \tag{A.2}$$

Because $\mathcal{V}_A$ is a closed subspace of $\mathcal{H}_0$, the projection theorem gives a unique element $h_A^\star = \mathrm{Proj}_{\mathcal{V}_A} r_0$ minimizing $\|h - r_0\|_0$ over $h \in \mathcal{V}_A$. Thus $h_A^\star$ is the unique functional perturbation minimizing $Q(h)$ over $\mathcal{V}_A$. If the atom dictionary is redundant, multiple coefficient vectors may represent the same function, but $h_A^\star$ remains unique.

Projection orthogonality gives $r_0 - h_A^\star \perp \mathcal{V}_A$ and $\langle r_0 - h_A^\star, h_A^\star\rangle_0 = 0$. Therefore

$$\|r_0\|_0^2 = \|h_A^\star\|_0^2 + \|r_0 - h_A^\star\|_0^2.$$

Evaluating (A.2) at 0 and $h_A^\star$ yields $Q(0) - Q(h_A^\star) = \frac{1}{2}\|h_A^\star\|_0^2$. Hence

$$\mathcal{G}(A) = \frac{1}{2}\|h_A^\star\|_0^2 = \frac{1}{2}\|r_0\|_0^2 \cos^2\Theta(A).$$

If $\|h_A^\star\|_\infty \le r_{\mathrm{loc}}$, Lemma A.1 gives $\mathcal{T}(h_A^\star) = Q(h_A^\star) + R_3(h_A^\star)$. Since $\mathcal{T}(0) = Q(0) = 0$,

$$-\mathcal{T}(h_A^\star) = \mathcal{G}(A) - R_3(h_A^\star),$$

so

$$|-\mathcal{T}(h_A^\star) - \mathcal{G}(A)| \le \frac{L_3}{6}\mathbb{E}\|h_A^\star(X)\|_2^3.$$

Finally,

$$\chi_q = \max_{A\subset[M];|A|\le q} \frac{\|\operatorname{Proj}_{\mathcal{V}_A} r_0\|_0^2}{\|r_0\|_0^2}$$

is exactly the maximum fraction of local natural-descent energy captured by a source span containing at most q atoms in the quadratic model.

□

Theorem 4.1 characterizes the optimal source-supported perturbation at the functional level. To connect this projection directly to the finite-dimensional source coefficients used by the estimator, the next lemma gives the corresponding Gram-system and Moore-Penrose representation.

Lemma A.2. *For every $A \subset [M]$, $\mathbf{b}_A \in (\mathbf{G}_{AA})$. The coefficient vectors representing $\operatorname{Proj}_{\mathcal{V}_A} r_0$ are exactly the solutions of $\mathbf{G}_{AA}a_A = \mathbf{b}_A$. The minimum-Euclidean-norm solution is*

$$a_A^\star = \mathbf{G}_{AA}^\dagger \mathbf{b}_A \text{ and } \|\operatorname{Proj}_{\mathcal{V}_A} r_0\|_0^2 = \mathbf{b}_A^\top \mathbf{G}_{AA}^\dagger \mathbf{b}_A.$$

Proof. Since $\mathbf{G}_{AA} = \Phi_A^*\Phi_A$, we first prove $\ker(\mathbf{G}_{AA}) = \ker(\Phi_A)$. If $\Phi_A a_A = 0$, then $\mathbf{G}_{AA}a_A = 0$. Conversely, if $\mathbf{G}_{AA}a_A = 0$, then

$$0 = a_A^\top \mathbf{G}_{AA} a_A = \|\Phi_A a_A\|_0^2,$$

so $\Phi_A a_A = 0$ in $\mathcal{H}_0$. Thus the kernels coincide. Because $\mathbf{G}_{AA}$ is symmetric and finite dimensional,

$$(\mathbf{G}_{AA}) = \ker(\mathbf{G}_{AA})^\perp = \ker(\Phi_A)^\perp = (\Phi_A^*).$$

Since $\mathbf{b}_A = \Phi_A^* r_0$, it follows that $\mathbf{b}_A \in (\mathbf{G}_{AA})$.

Let $h_A^\star = \operatorname{Proj}_{\mathcal{V}_A} r_0$. If a_A represents $h_A^\star$, then projection orthogonality gives

$$0 = \Phi_A^*(r_0 - \Phi_A a_A) = \mathbf{b}_A - \mathbf{G}_{AA}a_A,$$

so a_A solves the normal equation. Conversely, if the normal equation holds, then $r_0 - \Phi_A a_A$ is orthogonal to $(\Phi_A) = \mathcal{V}_A$, hence $\Phi_A a_A = h_A^\star$. Since $\mathbf{b}_A \in (\mathbf{G}_{AA})$, $\mathbf{G}_{AA}^\dagger \mathbf{b}_A$ is a solution and is the minimum-Euclidean-norm solution. Finally,

$$\|h_A^\star\|_0^2 = (a_A^\star)^\top \mathbf{G}_{AA} a_A^\star = (a_A^\star)^\top \mathbf{b}_A = \mathbf{b}_A^\top \mathbf{G}_{AA}^\dagger \mathbf{b}_A.$$

□

Lemma A.2 expresses the functional projection entirely through the Gram system and also clarifies how redundancy affects only the coefficient representation, not the projected function itself. Building on this representation, Corollary A.3 isolates the additional local gain contributed by a new atom after removing the component already explained by the active source span.

Corollary A.3. *Let $A \subset [M]$ and $e \notin A$, and define $\phi_{e|A} = \phi_e - \operatorname{Proj}_{\mathcal{V}_A}\phi_e$. If $\|\phi_{e|A}\|_0 > 0$, then*

$$\mathcal{G}(A\cup\{e\}) - \mathcal{G}(A) = \frac{1}{2}\frac{\langle \phi_{e|A}, r_0\rangle_0^2}{\|\phi_{e|A}\|_0^2}.$$

Proof. By construction, $\phi_{e|A} \perp \mathcal{V}_A$ and $\mathcal{V}_{A\cup\{e\}} = \mathcal{V}_A \oplus \operatorname{span}\{\phi_{e|A}\}$. Therefore

$$\operatorname{Proj}_{\mathcal{V}_{A\cup\{e\}}} r_0 = \operatorname{Proj}_{\mathcal{V}_A} r_0 + \operatorname{Proj}_{\operatorname{span}(\phi_{e|A})} r_0,$$

and the two summands are orthogonal. The one-dimensional projection formula gives

$$\mathrm{Proj}_{\mathrm{span}(\phi_{e|A})}\, r_0 = \frac{\langle \phi_{e|A}, r_0 \rangle_0}{\|\phi_{e|A}\|_0^2}\phi_{e|A},$$

whose squared norm is $\langle \phi_{e|A}, r_0 \rangle_0^2 / \|\phi_{e|A}\|_0^2$. Since $\mathcal{G}$ is one half of projection energy, the result follows.

□

A.2 Reparameterization and Estimation of the Transfer Geometry

We next formalize the reparameterization argument underlying Proposition 4.2. Let $\vartheta = \psi(\boldsymbol{\theta})$, $_0 = \psi(\boldsymbol{\theta}_0)$, and $A_0 = D\psi(\boldsymbol{\theta}_0)$; the derivative of the inverse coordinate map at $_0$ is then A_0^{-1}. The uniform tangent expansion used in the proof requires only the following local regularity condition.

Assumption 3. *For every compact target-input set $\mathcal{K}$ under consideration, $x \mapsto J_0(x)$ is continuous on $\mathcal{K}$ and $\sup_{x\in\mathcal{K}} \|J_0(x)\|_{\mathrm{op}} < \infty$. The coordinate map ψ is a C^2 local diffeomorphism on a neighborhood containing $\boldsymbol{\theta}_0$ and the source endpoints, with first and second derivatives bounded on that neighborhood.*

The bounded-Jacobian condition is used only to make the $O(\|\boldsymbol{\Delta}_e\|_2^2)$ tangent approximation uniform over compact target-input sets. It imposes no global bound on the network Jacobian.

Proof of Proposition 4.2. Under Assumption 3, Taylor expansion of ψ about $\boldsymbol{\theta}_0$ gives

$$\tilde{\boldsymbol{\Delta}}_e = \psi(\boldsymbol{\theta}_0 + \boldsymbol{\Delta}_e) - \psi(\boldsymbol{\theta}_0) = A_0\boldsymbol{\Delta}_e + q_e,$$

where $\|q_e\|_2 \le C_\psi \|\boldsymbol{\Delta}_e\|_2^2$ for a finite constant C_ψ determined by the local second-derivative bound for ψ. The predictor in the new coordinates is $\tilde{z}_\vartheta = z_{\psi^{-1}(\vartheta)}$. By the chain rule, $\tilde{J}_0(x) = J_0(x)A_0^{-1}$. Therefore

$$\begin{aligned}\tilde{\phi}_e(x) &= P_C \tilde{J}_0(x)\tilde{\boldsymbol{\Delta}}_e = P_C J_0(x) A_0^{-1}\{A_0\boldsymbol{\Delta}_e + q_e\} \\ &= P_C J_0(x)\boldsymbol{\Delta}_e + P_C J_0(x)A_0^{-1}q_e = \phi_e(x) + P_C J_0(x)A_0^{-1}q_e.\end{aligned}$$

By Assumption 3, $\sup_{x\in\mathcal{K}} \|P_C J_0(x)A_0^{-1}\|_{\mathrm{op}} < \infty$. Consequently,

$$\sup_{x\in\mathcal{K}} \|\tilde{\phi}_e(x) - \phi_e(x)\|_2 \le C\|\boldsymbol{\Delta}_e\|_2^2.$$

This proves the stated second-order tangent error and hence first-order invariance of the induced tangent field.

We next justify stability of the diagnostics. Let $d_e = \tilde{\phi}_e - \phi_e$ and $\varepsilon = \max_{e\in[M]} \|d_e\|_0$. For a fixed finite dictionary, all original tangent norms are finite, and

$$\begin{aligned}|\tilde{G}_{ee'} - G_{ee'}| &= |\,\langle \phi_e + d_e, \phi_{e'} + d_{e'}\rangle_0 - \langle \phi_e, \phi_{e'}\rangle_0\,| \\ &\le \|d_e\|_0\|\phi_{e'}\|_0 + \|\phi_e\|_0\|d_{e'}\|_0 + \|d_e\|_0\|d_{e'}\|_0 = O(\varepsilon).\end{aligned}$$

Likewise,

$$|\tilde{b}_e - b_e| = |\,\langle d_e, r_0\rangle_0\,| \le \varepsilon\|r_0\|_0.$$

For every source span entering the diagnostics, Proposition 4.2 assumes locally constant rank and a positive lower bound κ_G on the smallest nonzero Gram eigenvalue. On a fixed-rank matrix manifold, the Moore-Penrose inverse is continuous. Hence $\mathbf{G}_{AA}^\dagger \longmapsto \tilde{\mathbf{G}}_{AA}^\dagger$ continuously as the tangent dictionary is perturbed. By Lemma A.2, the projection energy $\mathbf{b}_A^\top \mathbf{G}_{AA}^\dagger \mathbf{b}_A$ is continuous for every fixed A. Since the library contains finitely many atoms, the maximum over the finitely many sets $|A| \le q$ is continuous; therefore χ_q is continuous under the tangent perturbation.

The same argument applies to the squared source-span cosine

$$\frac{\mathbf{b}_{\mathcal{E}_s}^\top \mathbf{G}_{\mathcal{E}_s\mathcal{E}_s}^\dagger \mathbf{b}_{\mathcal{E}_s}}{\|r_0\|_0^2}.$$

The principal angle is obtained through the continuous map $c \mapsto \arccos(\sqrt{c})$ on $[0, 1]$, hence it is continuous. Away from 0 and 1, this map is locally Lipschitz and the angle inherits the local perturbation order. At the endpoints, only continuity is claimed. For directed alignment,

$$A_e^{\mathrm{dir}} = \frac{b_e}{\sqrt{G_{ee}}\|r_0\|_0}.$$

The proposition assumes every nonzero atom norm entering this diagnostic is bounded away from zero. Its denominator is therefore locally bounded away from zero, so directed alignment is continuous under the same perturbation.

Finally, let $u, v \in \mathbb{R}^p$ be linearly independent. For any linearly independent $u', v' \in \mathbb{R}^p$, define a linear map on $\mathrm{span}\{u, v\}$ by $Au = u',\ Av = v'$, extend $\{u, v\}$ and $\{u', v'\}$ to bases of $\mathbb{R}^p$, and define A by the corresponding basis map. Then A is invertible. For any $c \in (-1, 1)$, choose u', v' with Euclidean cosine c. Thus an invertible linear coordinate change can assign any cosine in $(-1, 1)$ to the two parameter directions. Taking c arbitrarily close to 1 or -1 makes the transformed cosine arbitrarily close to either endpoint. The predictor is unchanged because this is only a coordinate change. Hence raw Euclidean task-vector cosine has no analogous coordinate-invariance property.

□

We next state the stochastic envelope used for finite-sample estimation. For a scalar random variable U, let $\|U\|_{\psi_2}$ and $\|U\|_{\psi_1}$ denote its sub-Gaussian and sub-exponential Orlicz norms. For a random vector V, write $\|V\|_{\psi_2,2} = \big\|\|V\|_2\big\|_{\psi_2}$. For the theoretical estimation results, atom scales are fixed in advance; equivalently, population unit-norm scaling may be absorbed into the fields so that $\|\phi_e\|_0 = 1$ for every retained atom e. This differs from empirical normalization in the practical algorithm; the latter is treated as a Gram perturbation after diagonal concentration.

Define

$$g(X, Y) = P_C \nabla_z \ell\{z_0(X), Y\},$$

$$Z_e(X) = W_0(X)^{1/2}\phi_e(X), \quad R(X, Y) = -W_0(X)^{-1/2} g(X, Y),$$

and, for the oracle analysis,

$$R^\star(X, Y) = -W_0(X)^{-1/2}\{g(X, Y) + W_0(X)h^\star(X)\}.$$

These quantities express the Gram, signal, and oracle-score fluctuations in the curvature geometry. Their finite-sample concentration is controlled by the following stochastic-envelope condition.

Assumption 4. *There exist finite constants $K_\phi, K_g, K_\star, K_{\mathrm{grp}}$ such that $\max_{e\in[M]} \|Z_e(X)\|_{\psi_2,2} \le K_\phi,\ \|R(X, Y)\|_{\psi_2,2} \le K_g$, and $\|R^\star(X, Y)\|_{\psi_2,2} \le K_\star$. For each source s, define*

$$\mathcal{Z}_s(X) = \left\{ \frac{1}{|\mathcal{E}_s|} \sum_{e\in\mathcal{E}_s} \|Z_e(X)\|_2^2 \right\}^{1/2}.$$

We also assume $\max_{s\in[S]} \|\mathcal{Z}_s(X)\|_{\psi_2} \le K_{\mathrm{grp}}$.

The atomwise bounds control individual Gram and signal entries, whereas the group-envelope condition controls the source-level dual norm in the sparse-group penalty. Normalization by $|\mathcal{E}_s|$ makes the source-level fluctuation rate depend on $\log S$, not on the total number of atoms in a source. Bounded tangent fields imply these conditions; more general sub-exponential product assumptions could be substituted with corresponding changes to constants.

For fit observation i, we write $Z_{ie} = Z_e(X_i)$ and $R_i = R(X_i, Y_i)$. For the cross-entropy estimator analyzed in Section 3, $W_i = W_0(X_i)$, so

$$\hat{G}_{ee'} = \frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} Z_{ie}^\top Z_{ie'} \text{ and } \hat{b}_e = \frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} Z_{ie}^\top R_i.$$

Let $\|A\|_{\max} = \max_{j,k} |A_{jk}|$. The next lemma quantifies the uniform sampling error of the empirical Gram matrix and signal vector that drive the estimated transfer geometry.

Lemma A.4. *Under Assumptions 1 and 4, there is a universal numerical constant C_0 such that, with probability at least $1-\delta$,*

$$\|\hat{\mathbf{G}} - \mathbf{G}\|_{\max} \le C_0 K_\phi^2 \left[\sqrt{\frac{\log(4M^2/\delta)}{n_{\text{fit}}}} + \frac{\log(4M^2/\delta)}{n_{\text{fit}}} \right]$$

and

$$\|\hat{\mathbf{b}} - \mathbf{b}\|_\infty \le C_0 K_\phi K_g \left[\sqrt{\frac{\log(4M/\delta)}{n_{\text{fit}}}} + \frac{\log(4M/\delta)}{n_{\text{fit}}} \right].$$

When $n_{\text{fit}} \gtrsim \log(M/\delta)$, both bounds have order $\sqrt{\log(M/\delta)/n_{\text{fit}}}$.

Proof. We fix e, e'. By Cauchy-Schwarz, $|Z_{ie}^\top Z_{ie'}| \le \|Z_{ie}\|_2 \|Z_{ie'}\|_2$. A product of sub-Gaussian scalar random variables is sub-exponential, so $\|Z_{ie}^\top Z_{ie'}\|_{\psi_1} \le C K_\phi^2$. Centering changes a ψ_1 norm by at most a universal factor. Therefore $Z_{ie}^\top Z_{ie'} - \mathbb{E}(Z_{ie}^\top Z_{ie'})$ is centered sub-exponential with norm at most $C K_\phi^2$. Bernstein's inequality gives, for every $t > 0$,

$$\mathbb{P}\left(|\hat{G}_{ee'} - G_{ee'}| > t \right) \le 2 \exp\left[-c n_{\text{fit}} \min\left\{ \frac{t^2}{K_\phi^4}, \frac{t}{K_\phi^2} \right\} \right].$$

A union bound over at most M^2 ordered pairs and inversion of this tail bound at failure probability $\delta/2$ yield the first display.

For the signal, $|Z_{ie}^\top R_i| \le \|Z_{ie}\|_2 \|R_i\|_2$, so $\|Z_{ie}^\top R_i\|_{\psi_1} \le C K_\phi K_g$. Moreover,

$$\mathbb{E}(Z_{ie}^\top R_i) = -\mathbb{E}\{\phi_e(X)^\top g(X,Y)\} = b_e.$$

Thus the centered signal summands are mean-zero sub-exponential variables with envelope $C K_\phi K_g$. Bernstein's inequality followed by a union bound over M atoms gives the second display. Combining the two failure probabilities proves the simultaneous result.

□

Randomized output sketching is not required for the oracle theorem. If $\hat{\mathbf{G}}_{\text{sk}}$ is a sketched Gram matrix satisfying $\|\hat{\mathbf{G}}_{\text{sk}} - \hat{\mathbf{G}}\|_{\max} \le \varepsilon_{\text{sk}}$, then

$$\|\hat{\mathbf{G}}_{\text{sk}} - \mathbf{G}\|_{\max} \le \|\hat{\mathbf{G}} - \mathbf{G}\|_{\max} + \varepsilon_{\text{sk}}.$$

Thus sketching contributes an additional Gram perturbation, whose construction and distortion control are detailed in Appendix B.

A.3 Sparse Estimation and Validation Guarantees

Define

$$\Omega_{\text{grp}}(a) = \sum_{s=1}^{S} w_s \|a_{\mathcal{E}_s}\|_2, \quad w_s = \sqrt{|\mathcal{E}_s|}, \quad \mathcal{P}(a) = \lambda_{\text{src}} \Omega_{\text{grp}}(a) + \lambda_{\text{atom}} \|a\|_1.$$

Let $h^\star = \text{Proj}_{\mathcal{V}} r_0 = \Phi a^\star$ for a coefficient representation $a^\star$ whose support has size q. If the representation is not unique, the theorem assumes the existence of one such sparse representation and fixes it for the analysis. Define

$$T = \text{supp}(a^\star), \quad \mathcal{S}^\star = \{s : a^\star_{\mathcal{E}_s} \neq 0\}, \quad G_\star = \bigcup_{s \in \mathcal{S}^\star} \mathcal{E}_s, \quad D_\star = |G_\star| = \sum_{s \in \mathcal{S}^\star} |\mathcal{E}_s| = \sum_{s \in \mathcal{S}^\star} w_s^2.$$

Thus q counts active atoms, whereas $D_\star$ counts all atom coordinates belonging to active source groups. Because $h^\star$ is the projection of r_0 onto the full source span, $r_0 - h^\star \perp \mathcal{V}$, and therefore

$$\mathbf{G} a^\star = \mathbf{b}. \tag{A.3}$$

For $u \in \mathbb{R}^M$, let $u_{\mathcal{S}^\star}$ denote the vector retaining all coordinates belonging to source groups in $\mathcal{S}^\star$ and zeroing the remaining groups. The notation u_T instead retains only the coordinates in T. Define

$$\mathcal{C}_\star = \left\{u : \lambda_{\mathrm{src}}\Omega_{\mathrm{grp}}(u_{(\mathcal{S}^\star)^c}) + \lambda_{\mathrm{atom}}\|u_{T^c}\|_1 \le 3\left[\lambda_{\mathrm{src}}\Omega_{\mathrm{grp}}(u_{\mathcal{S}^\star}) + \lambda_{\mathrm{atom}}\|u_T\|_1\right]\right\}. \tag{A.4}$$

This cone collects the estimation-error directions permitted by the sparse-group basic inequality. The next assumption requires the population Gram geometry to remain sufficiently nondegenerate only along these directions, not globally over the entire source dictionary.

Assumption 5. *There exists $\kappa_{\mathrm{RE}} > 0$ such that $u^\top \mathbf{G} u \ge \kappa_{\mathrm{RE}}^2 \|u_{G_\star}\|_2^2$ for every $u \in \mathcal{C}_\star$.*

The condition does not require $\mathbf{G}$ to be nonsingular on all of $\mathbb{R}^M$. Globally redundant source atoms are allowed. It only rules out degeneracy along error directions compatible with the active sparse-group structure. Define the single-observation normal-equation score

$$\xi_e(X, Y) = -\phi_e(X)^\top \{g(X, Y) + W_0(X)h^\star(X)\}.$$

Equation (A.3) implies $\mathbb{E}\{\xi_e(X, Y)\} = 0$. For the fit sample, we write

$$\xi_{ie} = \xi_e(X_i, Y_i) \text{ and } \bar{\xi}_e = \frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} \xi_{ie}.$$

Then

$$\bar{\xi} = \hat{\mathbf{b}} - \hat{\mathbf{G}} a^\star. \tag{A.5}$$

Thus $\bar{\xi}$ is the empirical normal-equation fluctuation that enters the basic inequality. The next lemma controls its source-group and coordinatewise dual norms at the penalty scales in the oracle analysis.

Lemma A.5. *Under Assumptions 1 and 4, there is a constant $\sigma = C_1 \max\{K_\phi K_\star, K_{\mathrm{grp}} K_\star\}$ such that, with probability at least $1 - \delta$,*

$$\max_{s \in [S]} \frac{\|\bar{\xi}_{\mathcal{E}_s}\|_2}{w_s} \le C_2 \sigma \left[\sqrt{\frac{1 + \log(2S/\delta)}{n_{\mathrm{fit}}}} + \frac{1 + \log(2S/\delta)}{n_{\mathrm{fit}}}\right]$$

and

$$\|\bar{\xi}\|_\infty \le C_2 \sigma \left[\sqrt{\frac{\log(2M/\delta)}{n_{\mathrm{fit}}}} + \frac{\log(2M/\delta)}{n_{\mathrm{fit}}}\right].$$

Consequently, when n_{fit} dominates the displayed logarithms, the two bounds have the rates used in Theorem 4.3.

Proof. We write $R_i^\star = R^\star(X_i, Y_i)$. Since $\xi_{ie} = Z_{ie}^\top R_i^\star$, we have $|\xi_{ie}| \le \|Z_{ie}\|_2 \|R_i^\star\|_2$. The product of the two sub-Gaussian norm envelopes is sub-exponential, so $\|\xi_{ie}\|_{\psi_1} \le C K_\phi K_\star$. Since $\mathbb{E}\xi_{ie} = 0$, scalar Bernstein followed by a union bound over $e \in [M]$ gives the coordinatewise display.

For source s, define $V_{is} = \frac{1}{w_s}(\xi_{ie} : e \in \mathcal{E}_s)$ and $w_s = \sqrt{|\mathcal{E}_s|}$. Then

$$\|V_{is}\|_2 = \frac{1}{\sqrt{|\mathcal{E}_s|}} \left[\sum_{e \in \mathcal{E}_s} (Z_{ie}^\top R_i^\star)^2\right]^{1/2} \le \|R_i^\star\|_2 \left\{\frac{1}{|\mathcal{E}_s|} \sum_{e \in \mathcal{E}_s} \|Z_{ie}\|_2^2\right\}^{1/2} = \|R_i^\star\|_2 \, \mathcal{Z}_s(X_i).$$

Assumption 4 makes $\|R_i^\star\|_2$ and $\mathcal{Z}_s(X_i)$ uniformly sub-Gaussian, so their product is sub-exponential and $\|\|V_{is}\|_2\|_{\psi_1} \le C K_\star K_{\mathrm{grp}}$. Also $\mathbb{E}V_{is} = 0$. The Bernstein inequality for sums of independent mean-zero random vectors in a Hilbert space gives, for every $t > 0$,

$$\mathbb{P}\left[\left\|\frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} V_{is}\right\|_2 > C K_\star K_{\mathrm{grp}} \left\{\sqrt{\frac{1+t}{n_{\mathrm{fit}}}} + \frac{1+t}{n_{\mathrm{fit}}}\right\}\right] \le 2e^{-t}.$$

Taking $t = \log(2S/\delta)$ and applying a union bound over $s \in [S]$ gives the source-group display. Intersecting this event with the coordinatewise event proves the lemma.

□

Lemma A.5 controls the stochastic component of the sparse-group basic inequality. To convert that control into an estimation bound, we must also ensure that the empirical Gram matrix preserves the population curvature along the same cone of admissible error directions. Lemma A.6 establishes this transfer from the population restricted-eigenvalue condition to its empirical counterpart.

Lemma A.6. *Define* $L_\star = 4\sqrt{D_\star} + 3\frac{\lambda_{\text{atom}}}{\lambda_{\text{src}}}\sqrt{q}$. *Every* $u \in \mathcal{C}_\star$ *satisfies* $\|u\|_1 \le L_\star \|u_{G_\star}\|_2$. *Consequently, if*

$$\|\hat{\mathbf{G}} - \mathbf{G}\|_{\max} L_\star^2 \le \frac{1}{2}\kappa_{\text{RE}}^2, \tag{A.6}$$

then

$$u^\top \hat{\mathbf{G}} u \ge \frac{1}{2}\kappa_{\text{RE}}^2 \|u_{G_\star}\|_2^2 \text{ for every } u \in \mathcal{C}_\star.$$

Proof. For the active source groups, $\|u_{G_\star}\|_1 \le \sqrt{D_\star}\|u_{G_\star}\|_2$. For the inactive groups,

$$\|u_{G_\star^c}\|_1 = \sum_{s \notin \mathcal{S}^\star} \|u_{\mathcal{E}_s}\|_1 \le \sum_{s \notin \mathcal{S}^\star} \sqrt{|\mathcal{E}_s|}\|u_{\mathcal{E}_s}\|_2 = \Omega_{\text{grp}}(u_{(\mathcal{S}^\star)^c}).$$

The cone condition therefore implies

$$\|u\|_1 \le \sqrt{D_\star}\|u_{G_\star}\|_2 + 3\Omega_{\text{grp}}(u_{\mathcal{S}^\star}) + 3\frac{\lambda_{\text{atom}}}{\lambda_{\text{src}}}\|u_T\|_1.$$

By Cauchy-Schwarz,

$$\Omega_{\text{grp}}(u_{\mathcal{S}^\star}) \le \left(\sum_{s \in \mathcal{S}^\star} w_s^2\right)^{1/2} \|u_{G_\star}\|_2 = \sqrt{D_\star}\|u_{G_\star}\|_2 \text{ and } \|u_T\|_1 \le \sqrt{q}\,\|u_{G_\star}\|_2.$$

Thus $\|u\|_1 \le L_\star \|u_{G_\star}\|_2$. Now

$$|u^\top(\hat{\mathbf{G}} - \mathbf{G})u| \le \|\hat{\mathbf{G}} - \mathbf{G}\|_{\max}\|u\|_1^2 \le \|\hat{\mathbf{G}} - \mathbf{G}\|_{\max} L_\star^2 \|u_{G_\star}\|_2^2.$$

Under (A.6),

$$u^\top \hat{\mathbf{G}} u \ge u^\top \mathbf{G} u - \frac{1}{2}\kappa_{\text{RE}}^2 \|u_{G_\star}\|_2^2.$$

Assumption 5 gives $u^\top \mathbf{G} u \ge \kappa_{\text{RE}}^2 \|u_{G_\star}\|_2^2$, which proves the empirical restricted-eigenvalue bound. □

Combining Lemma A.4 with (A.6) gives a sufficient sample-size condition. Let

$$\varepsilon_G(\delta) = C_0 K_\phi^2 \left[\sqrt{\frac{\log(8M^2/\delta)}{n_{\text{fit}}}} + \frac{\log(8M^2/\delta)}{n_{\text{fit}}}\right].$$

It suffices that $\varepsilon_G(\delta) L_\star^2 \le \frac{1}{2}\kappa_{\text{RE}}^2$. A sufficient order condition is

$$n_{\text{fit}} \gtrsim K_\phi^4 L_\star^4 \kappa_{\text{RE}}^{-4} \log(8M^2/\delta) + K_\phi^2 L_\star^2 \kappa_{\text{RE}}^{-2} \log(8M^2/\delta).$$

This makes explicit the empirical restricted-eigenvalue event invoked in the main theorem.

Proof of Theorem 4.3. The theorem concerns the exact-tangent, unshrunk estimator over $\mathbb{R}^M$, so $\rho_G = \lambda_G = 0$ and there is no trust-region constraint. Its objective is

$$\hat{Q}(a) = \frac{1}{2}a^\top \hat{\mathbf{G}} a - \hat{\mathbf{b}}^\top a + \mathcal{P}(a).$$

Let $\hat{u} = \hat{a} - a^\star$. Since $\hat{a}$ is an exact minimizer and $a^\star$ is an admissible comparator, $\hat{Q}(a^\star + \hat{u}) \le \hat{Q}(a^\star)$. Expanding and canceling common terms gives

$$\frac{1}{2}\hat{u}^\top \hat{\mathbf{G}} \hat{u} \le \bar{\xi}^\top \hat{u} + \mathcal{P}(a^\star) - \mathcal{P}(a^\star + \hat{u}). \tag{A.7}$$

Let

$$\mathcal{E}_{\text{score}} = \left\{ \max_s \frac{\|\bar{\xi}_{\mathcal{E}_s}\|_2}{w_s} \le \frac{\lambda_{\text{src}}}{2}, \quad \|\bar{\xi}\|_\infty \le \frac{\lambda_{\text{atom}}}{2} \right\}.$$

Under the sample-size regime in the theorem, the square-root terms dominate the linear Bernstein terms. Hence the theorem's choices of λ_{src} and λ_{atom}, with sufficiently large universal constants, ensure $\mathbb{P}(\mathcal{E}_{\text{score}}) \ge 1 - \delta/2$.

On $\mathcal{E}_{\text{score}}$, split the scalar product into two equal halves, $\bar{\xi}^\top \hat{u} = \frac{1}{2}\bar{\xi}^\top \hat{u} + \frac{1}{2}\bar{\xi}^\top \hat{u}$. For the first half, groupwise Cauchy-Schwarz gives

$$\frac{1}{2}\bar{\xi}^\top \hat{u} \le \frac{1}{2} \max_s \frac{\|\bar{\xi}_{\mathcal{E}_s}\|_2}{w_s} \Omega_{\text{grp}}(\hat{u}) \le \frac{1}{4}\lambda_{\text{src}}\Omega_{\text{grp}}(\hat{u}).$$

For the second half,

$$\frac{1}{2}\bar{\xi}^\top \hat{u} \le \frac{1}{2}\|\bar{\xi}\|_\infty \|\hat{u}\|_1 \le \frac{1}{4}\lambda_{\text{atom}}\|\hat{u}\|_1.$$

Therefore

$$\bar{\xi}^\top \hat{u} \le \frac{1}{4}\lambda_{\text{src}}\Omega_{\text{grp}}(\hat{u}) + \frac{1}{4}\lambda_{\text{atom}}\|\hat{u}\|_1. \tag{A.8}$$

Because $a^\star$ vanishes outside the active source groups,

$$\Omega_{\text{grp}}(a^\star) - \Omega_{\text{grp}}(a^\star + \hat{u}) \le \Omega_{\text{grp}}(\hat{u}_{\mathcal{S}^\star}) - \Omega_{\text{grp}}(\hat{u}_{(\mathcal{S}^\star)^c}).$$

Since $a^\star$ is supported on T,

$$\|a^\star\|_1 - \|a^\star + \hat{u}\|_1 \le \|\hat{u}_T\|_1 - \|\hat{u}_{T^c}\|_1.$$

Substituting these bounds and (A.8) into (A.7) gives

$$\frac{1}{2}\hat{u}^\top \hat{\mathbf{G}}\hat{u} + \frac{3}{4}\lambda_{\text{src}}\Omega_{\text{grp}}(\hat{u}_{(\mathcal{S}^\star)^c}) + \frac{3}{4}\lambda_{\text{atom}}\|\hat{u}_{T^c}\|_1 \le \frac{5}{4}\lambda_{\text{src}}\Omega_{\text{grp}}(\hat{u}_{\mathcal{S}^\star}) + \frac{5}{4}\lambda_{\text{atom}}\|\hat{u}_T\|_1. \tag{A.9}$$

Dropping the nonnegative quadratic term gives

$$\lambda_{\text{src}}\Omega_{\text{grp}}(\hat{u}_{(\mathcal{S}^\star)^c}) + \lambda_{\text{atom}}\|\hat{u}_{T^c}\|_1 \le \frac{5}{3}\left[\lambda_{\text{src}}\Omega_{\text{grp}}(\hat{u}_{\mathcal{S}^\star}) + \lambda_{\text{atom}}\|\hat{u}_T\|_1\right].$$

Since $5/3 < 3$, $\hat{u} \in \mathcal{C}_\star$.

Retaining the quadratic term in (A.9) and dropping only the inactive penalties,

$$\frac{1}{2}\hat{u}^\top \hat{\mathbf{G}}\hat{u} \le \frac{5}{4}\left[\lambda_{\text{src}}\Omega_{\text{grp}}(\hat{u}_{\mathcal{S}^\star}) + \lambda_{\text{atom}}\|\hat{u}_T\|_1\right].$$

By Cauchy-Schwarz,

$$\Omega_{\text{grp}}(\hat{u}_{\mathcal{S}^\star}) \le \sqrt{D_\star}\|\hat{u}_{G_\star}\|_2 \text{ and } \|\hat{u}_T\|_1 \le \sqrt{q}\,\|\hat{u}_{G_\star}\|_2.$$

Set $A_\star = \lambda_{\text{src}}\sqrt{D_\star} + \lambda_{\text{atom}}\sqrt{q}$. On the empirical restricted-eigenvalue event,

$$\hat{u}^\top \hat{\mathbf{G}}\hat{u} \ge \frac{1}{2}\kappa_{\text{RE}}^2\|\hat{u}_{G_\star}\|_2^2.$$

Therefore

$$\frac{1}{4}\kappa_{\text{RE}}^2\|\hat{u}_{G_\star}\|_2^2 \le \frac{5}{4}A_\star\|\hat{u}_{G_\star}\|_2.$$

If $\hat{u}_{G_\star} = 0$, the cone inequality forces $\hat{u} = 0$. Otherwise, dividing by $\|\hat{u}_{G_\star}\|_2$ gives

$$\|\hat{u}_{G_\star}\|_2 \le \frac{5A_\star}{\kappa_{\text{RE}}^2}. \tag{A.10}$$

Substitution into the quadratic inequality gives $\hat{u}^\top \hat{\mathbf{G}}\hat{u} \le \frac{25}{2}\frac{A_\star^2}{\kappa_{\text{RE}}^2}$.

To convert to the population source-span norm, use (A.6). Since $\hat{u} \in \mathcal{C}_\star$,

$$|\hat{u}^\top(\hat{\mathbf{G}} - \mathbf{G})\hat{u}| \le \frac{1}{2}\kappa_{\text{RE}}^2\|\hat{u}_{G_\star}\|_2^2.$$

Together with (A.10), this yields

$$\hat{u}^\top \mathbf{G}\hat{u} \le 25\frac{A_\star^2}{\kappa_{\mathrm{RE}}^2}.$$

Since $A_\star^2 \le 2D_\star\lambda_{\mathrm{src}}^2 + 2q\lambda_{\mathrm{atom}}^2$, we obtain

$$\|\Phi(\hat{a}-a^\star)\|_0^2 = \hat{u}^\top \mathbf{G}\hat{u} \le \frac{C_{\mathrm{or}}}{\kappa_{\mathrm{RE}}^2}\left\{D_\star\lambda_{\mathrm{src}}^2 + q\lambda_{\mathrm{atom}}^2\right\},$$

for a universal constant $C_{\mathrm{or}} > 0$.

The score event has probability at least $1-\delta/2$. Under the explicit sample-size condition following Lemma A.6, the Gram event and therefore the empirical restricted-eigenvalue event have probability at least $1-\delta/2$. Their intersection has probability at least $1-\delta$. Substituting the penalty levels gives

$$O\left(\frac{\sigma^2[D_\star\{1+\log(S/\delta)\} + q\log(M/\delta)]}{\kappa_{\mathrm{RE}}^2 n_{\mathrm{fit}}}\right).$$

Finally, assume $\|h^\star\|_\infty \le r_{\mathrm{loc}}$ and $\|\Phi\hat{a}\|_\infty \le r_{\mathrm{loc}}$. Both functions belong to the full source span $\mathcal{V}$. Since $h^\star = \mathrm{Proj}_{\mathcal{V}}\, r_0$, the projection identity gives

$$Q(\Phi\hat{a}) - Q(h^\star) = \frac{1}{2}\|\Phi(\hat{a}-a^\star)\|_0^2.$$

Applying Lemma A.1 separately to $\Phi\hat{a}$ and $h^\star$ yields

$$\mathcal{T}(\Phi\hat{a}) - \mathcal{T}(h^\star) = \frac{1}{2}\|\Phi(\hat{a}-a^\star)\|_0^2 + R_3(\Phi\hat{a}) - R_3(h^\star).$$

Thus the oracle inequality controls the quadratic component of target-risk difference relative to the population source-span projection, while the two cubic terms account for local nonlinear error.

□

Theorem 4.3 assumes that the population source-span projection admits an exactly sparse coefficient representation. The next corollary relaxes this requirement by allowing a sparse comparator that approximates the projection, with the resulting error separated into estimation and source-span approximation components.

Corollary A.7. *Let a_q be a q-sparse source coefficient vector. Define $T_q = \mathrm{supp}(a_q)$, $\mathcal{S}_q = \{s : (a_q)_{\mathcal{E}_s} \neq 0\}$, $G_q = \bigcup_{s\in\mathcal{S}_q}\mathcal{E}_s$, $D_q = |G_q|$, and let $\kappa_{\mathrm{RE},q} > 0$ denote the corresponding restricted-eigenvalue constant. Set $h_q = \Phi a_q$, $d_q = h^\star - h_q$, and define $c_q = \mathbf{b} - \mathbf{G}a_q = \Phi^* d_q$. Suppose the centered empirical fluctuation around a_q satisfies the same dual-score bounds as in Lemma A.5, and suppose*

$$\max_s \frac{\|(c_q)_{\mathcal{E}_s}\|_2}{w_s} \le \frac{\lambda_{\mathrm{src}}}{4}, \quad \|c_q\|_\infty \le \frac{\lambda_{\mathrm{atom}}}{4}.$$

If the corresponding restricted-eigenvalue condition holds for the support of a_q, then

$$\|\Phi\hat{a} - h^\star\|_0^2 \le 2\|d_q\|_0^2 + \frac{C}{\kappa_{\mathrm{RE},q}^2}\left\{D_q\lambda_{\mathrm{src}}^2 + q\lambda_{\mathrm{atom}}^2\right\}.$$

Proof. Define $\hat{\eta}_q = (\hat{\mathbf{b}}-\mathbf{b}) - (\hat{\mathbf{G}}-\mathbf{G})a_q$. Then $\hat{\mathbf{b}}-\hat{\mathbf{G}}a_q = \hat{\eta}_q + c_q$. By assumption, the centered fluctuation $\hat{\eta}_q$ and deterministic bias c_q together satisfy the same group and coordinate dual bounds used in (A.8), after enlarging universal constants if necessary. Repeating the full basic-inequality argument from Theorem 4.3, with a_q as comparator, gives

$$\|\Phi(\hat{a}-a_q)\|_0^2 \le \frac{C}{\kappa_{\mathrm{RE},q}^2}\left\{D_q\lambda_{\mathrm{src}}^2 + q\lambda_{\mathrm{atom}}^2\right\}.$$

Since $\Phi\hat{a} - h^\star = \Phi(\hat{a}-a_q) - d_q$, we use $\|u+v\|_0^2 \le 2\|u\|_0^2 + 2\|v\|_0^2$ to obtain

$$\|\Phi\hat{a} - h^\star\|_0^2 \le 2\|\Phi(\hat{a}-a_q)\|_0^2 + 2\|d_q\|_0^2.$$

Absorbing the factor two into the universal constant proves the claim.

□

For finite optimization error, define the penalty-valued enlarged cone

$$\mathcal{C}_\star^{\mathrm{opt}}(\varepsilon) = \left\{ u : \begin{array}{c} \lambda_{\mathrm{src}}\Omega_{\mathrm{grp}}(u_{(\mathcal{S}^\star)^c}) + \lambda_{\mathrm{atom}}\|u_{T^c}\|_1 \\ \leq 3\left[\lambda_{\mathrm{src}}\Omega_{\mathrm{grp}}(u_{\mathcal{S}^\star}) + \lambda_{\mathrm{atom}}\|u_T\|_1\right] + 4\varepsilon \end{array} \right\}. \tag{A.11}$$

The additive term has the same units as the objective gap. When $\varepsilon = 0$, this reduces to $\mathcal{C}_\star$. This enlarged cone captures the controlled deviation from the exact sparse-group geometry caused by approximate optimization. Corollary A.8 uses it to extend the oracle bound to practical objectives with numerical and geometry perturbations.

Corollary A.8. *Consider a practical penalized quadratic criterion with symmetric Gram matrix $\tilde{\mathbf{G}}$, signal $\tilde{\mathbf{b}}$, and returned coefficient vector $\tilde{a}$. Let $u = \tilde{a} - a^\star$. Assume (i) the practical normal-equation score $\tilde{\mathbf{b}} - \tilde{\mathbf{G}}a^\star$ obeys the same source-group and coordinate dual bounds used in* (A.8)*; (ii) the returned objective is within $\varepsilon_{\mathrm{opt}} \geq 0$ of the exact minimum; (iii) for the realized error u, the practical quadratic form satisfies $u^\top\tilde{\mathbf{G}}u \geq \frac{1}{2}\kappa_{\mathrm{RE}}^2\|u_{G_\star}\|_2^2$; and (iv) the geometry perturbation is small on the realized error, $|u^\top(\tilde{\mathbf{G}} - \mathbf{G})u| \leq \frac{1}{4}\kappa_{\mathrm{RE}}^2\|u_{G_\star}\|_2^2$. Then $u \in \mathcal{C}_\star^{\mathrm{opt}}(\varepsilon_{\mathrm{opt}})$ and*

$$\|\Phi(\tilde{a} - a^\star)\|_0^2 \leq \frac{C}{\kappa_{\mathrm{RE}}^2}\left\{D_\star\lambda_{\mathrm{src}}^2 + q\lambda_{\mathrm{atom}}^2\right\} + C\varepsilon_{\mathrm{opt}}.$$

Proof. Let $\tilde{Q}(a)$ denote the practical penalized quadratic objective. By $\varepsilon_{\mathrm{opt}}$-optimality,

$$\tilde{Q}(\tilde{a}) \leq \inf_a \tilde{Q}(a) + \varepsilon_{\mathrm{opt}} \leq \tilde{Q}(a^\star) + \varepsilon_{\mathrm{opt}}.$$

Writing $u = \tilde{a} - a^\star$ and expanding gives

$$\frac{1}{2}u^\top\tilde{\mathbf{G}}u \leq (\tilde{\mathbf{b}} - \tilde{\mathbf{G}}a^\star)^\top u + \mathcal{P}(a^\star) - \mathcal{P}(a^\star + u) + \varepsilon_{\mathrm{opt}}. \tag{A.12}$$

By condition 1,

$$(\tilde{\mathbf{b}} - \tilde{\mathbf{G}}a^\star)^\top u \leq \frac{1}{4}\lambda_{\mathrm{src}}\Omega_{\mathrm{grp}}(u) + \frac{1}{4}\lambda_{\mathrm{atom}}\|u\|_1.$$

The decomposability inequalities used in Theorem 4.3 remain valid

$$\Omega_{\mathrm{grp}}(a^\star) - \Omega_{\mathrm{grp}}(a^\star + u) \leq \Omega_{\mathrm{grp}}(u_{\mathcal{S}^\star}) - \Omega_{\mathrm{grp}}(u_{(\mathcal{S}^\star)^c}),$$

and

$$\|a^\star\|_1 - \|a^\star + u\|_1 \leq \|u_T\|_1 - \|u_{T^c}\|_1.$$

Substituting these relations into (A.12) gives

$$\begin{aligned} \frac{1}{2}u^\top\tilde{\mathbf{G}}u + \frac{3}{4}\lambda_{\mathrm{src}}\Omega_{\mathrm{grp}}(u_{(\mathcal{S}^\star)^c}) + \frac{3}{4}\lambda_{\mathrm{atom}}\|u_{T^c}\|_1 \\ \leq \frac{5}{4}\lambda_{\mathrm{src}}\Omega_{\mathrm{grp}}(u_{\mathcal{S}^\star}) + \frac{5}{4}\lambda_{\mathrm{atom}}\|u_T\|_1 + \varepsilon_{\mathrm{opt}}. \end{aligned} \tag{A.13}$$

Dropping the nonnegative quadratic term yields

$$\lambda_{\mathrm{src}}\Omega_{\mathrm{grp}}(u_{(\mathcal{S}^\star)^c}) + \lambda_{\mathrm{atom}}\|u_{T^c}\|_1 \leq \frac{5}{3}\left[\lambda_{\mathrm{src}}\Omega_{\mathrm{grp}}(u_{\mathcal{S}^\star}) + \lambda_{\mathrm{atom}}\|u_T\|_1\right] + \frac{4}{3}\varepsilon_{\mathrm{opt}}.$$

Since $5/3 < 3$ and $4/3 < 4$, this proves $u \in \mathcal{C}_\star^{\mathrm{opt}}(\varepsilon_{\mathrm{opt}})$.

Retaining the quadratic term in (A.13), dropping the inactive penalty terms, and applying $\Omega_{\mathrm{grp}}(u_{\mathcal{S}^\star}) \leq \sqrt{D_\star}\|u_{G_\star}\|_2$, $\|u_T\|_1 \leq \sqrt{q}\|u_{G_\star}\|_2$, gives

$$\frac{1}{2}u^\top\tilde{\mathbf{G}}u \leq \frac{5}{4}A_\star\|u_{G_\star}\|_2 + \varepsilon_{\mathrm{opt}},$$

where $A_\star = \lambda_{\mathrm{src}}\sqrt{D_\star} + \lambda_{\mathrm{atom}}\sqrt{q}$. Condition 3 therefore gives

$$\frac{1}{4}\kappa_{\mathrm{RE}}^2\|u_{G_\star}\|_2^2 \leq \frac{5}{4}A_\star\|u_{G_\star}\|_2 + \varepsilon_{\mathrm{opt}}.$$

For nonnegative a, b, c, t, the implication

$$at^2 \le bt + c \quad \Longrightarrow \quad t \le \frac{b}{a} + \sqrt{\frac{c}{a}}$$

holds. Hence

$$\|u_{G_\star}\|_2 \le C\left\{\frac{A_\star}{\kappa_{\mathrm{RE}}^2} + \frac{\sqrt{\varepsilon_{\mathrm{opt}}}}{\kappa_{\mathrm{RE}}}\right\}.$$

Returning to the preceding upper bound on $u^\top \tilde{\mathbf{G}} u$ and using $2ab \le a^2 + b^2$ gives

$$u^\top \tilde{\mathbf{G}} u \le C\left\{\frac{A_\star^2}{\kappa_{\mathrm{RE}}^2} + \varepsilon_{\mathrm{opt}}\right\}.$$

Finally, condition 4 gives

$$u^\top \mathbf{G} u \le u^\top \tilde{\mathbf{G}} u + \frac{1}{4}\kappa_{\mathrm{RE}}^2 \|u_{G_\star}\|_2^2 \le C\left\{\frac{A_\star^2}{\kappa_{\mathrm{RE}}^2} + \varepsilon_{\mathrm{opt}}\right\}.$$

Since $A_\star^2 \le 2D_\star\lambda_{\mathrm{src}}^2 + 2q\lambda_{\mathrm{atom}}^2$ and $u^\top \mathbf{G} u = \|\Phi u\|_0^2$, the stated bound follows. □

The preceding theory conditions on the fitted source checkpoints. Unconditional analysis would additionally propagate source-estimation error through perturbation of the tangent dictionary; that is distinct from the low-target-sample problem studied here. We now turn to the validation stage, where the remaining randomness comes from the held-out target sample used by the deployment gate. Lemma A.9 provides the simultaneous finite-sample risk bound required for that selection step.

Lemma A.9. *Condition on the fit data and hence on all candidate predictors. Suppose* $\{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{val}}}$ *are i.i.d. from* P_0 *and* $D_i(\lambda) \in [-B, B]$ *for every* $\lambda \in \Lambda$. *Let*

$$\bar{D}_\lambda = \frac{1}{n_{\mathrm{val}}}\sum_{i=1}^{n_{\mathrm{val}}} D_i(\lambda) \text{ and } \hat{v}_\lambda = \frac{1}{n_{\mathrm{val}} - 1}\sum_{i=1}^{n_{\mathrm{val}}}\{D_i(\lambda) - \bar{D}_\lambda\}^2.$$

Then, with conditional probability at least $1 - \delta$, *simultaneously for every* $\lambda \in \Lambda$,

$$\mathbb{E}\{D_i(\lambda) \mid \mathcal{D}_0^{\mathrm{fit}}\} \le \bar{D}_\lambda + \sqrt{\frac{2\hat{v}_\lambda \log(2|\Lambda|/\delta)}{n_{\mathrm{val}}}} + \frac{14B\log(2|\Lambda|/\delta)}{3(n_{\mathrm{val}} - 1)}.$$

Proof. We fix $\lambda \in \Lambda$ and define $Z_i(\lambda) = \frac{D_i(\lambda)+B}{2B}$. Then $0 \le Z_i(\lambda) \le 1$. Conditionally on the fit data, the candidate predictor is fixed and the validation observations are i.i.d., so the $Z_i(\lambda)$ are i.i.d. bounded random variables. Apply the one-sided empirical-Bernstein inequality of Maurer and Pontil (2009) at failure probability $\delta/|\Lambda|$

$$\mathbb{E}Z_i(\lambda) \le \bar{Z}_\lambda + \sqrt{\frac{2\hat{v}_{Z,\lambda}\log(2|\Lambda|/\delta)}{n_{\mathrm{val}}}} + \frac{7\log(2|\Lambda|/\delta)}{3(n_{\mathrm{val}} - 1)}.$$

Because $D_i(\lambda) = 2BZ_i(\lambda) - B$, the sample means and variances satisfy $\bar{D}_\lambda = 2B\bar{Z}_\lambda - B$ and $\hat{v}_\lambda = 4B^2\hat{v}_{Z,\lambda}$. Multiplying by $2B$ and subtracting B gives

$$\mathbb{E}D_i(\lambda) \le \bar{D}_\lambda + \sqrt{\frac{2\hat{v}_\lambda \log(2|\Lambda|/\delta)}{n_{\mathrm{val}}}} + \frac{14B\log(2|\Lambda|/\delta)}{3(n_{\mathrm{val}} - 1)}.$$

This fails for a fixed candidate with probability at most $\delta/|\Lambda|$. A union bound over the finite set Λ gives simultaneous validity with failure probability at most δ. No independence between different candidates is required. □

Proof of Proposition 4.4. Condition on $\mathcal{D}_0^{\mathrm{fit}}$. By construction, every transferred candidate $\hat{\boldsymbol{\theta}}_\lambda$ and the target-only comparator $\hat{\boldsymbol{\theta}}_{\mathrm{TO}}$ are fixed functions because none was fitted using validation observations. For each candidate,

$$D_i(\lambda) = \ell_{\mathrm{gate}}\{f_{\hat{\boldsymbol{\theta}}_\lambda}(X_i), Y_i\} - \ell_{\mathrm{gate}}\{f_{\hat{\boldsymbol{\theta}}_{\mathrm{TO}}}(X_i), Y_i\}.$$

Therefore

$$\mathbb{E}\{D_i(\lambda) \mid \mathcal{D}_0^{\mathrm{fit}}\} = R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_\lambda) - R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{TO}}).$$

Lemma A.9 implies that, on an event $\mathcal{E}_{\mathrm{gate}}$ with conditional probability at least $1-\delta$,

$$R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_\lambda) - R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{TO}}) \le U_\lambda$$

for every $\lambda \in \Lambda$ simultaneously.

If Algorithm 1 deploys a transferred candidate, that candidate satisfies $U_\lambda \le 0$. Hence, on $\mathcal{E}_{\mathrm{gate}}$, $R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_\lambda) \le R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{TO}})$. If no candidate qualifies, the algorithm deploys $\hat{\boldsymbol{\theta}}_{\mathrm{TO}}$ itself, so the excess risk relative to target-only is zero. Thus, on $\mathcal{E}_{\mathrm{gate}}$, $R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{deploy}}) \le R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{TO}})$. It follows conditionally that

$$\mathbb{P}\left\{R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{deploy}}) > R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{TO}}) \mid \mathcal{D}_0^{\mathrm{fit}}\right\} \le \delta.$$

Taking expectation with respect to the fit sample preserves the same bound

$$\mathbb{P}\left\{R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{deploy}}) > R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{TO}})\right\} \le \delta.$$

This proves Proposition 4.4.

□

B Estimation, Optimization, and Computation

This part gives the computational details needed to reproduce the estimators and optimization procedures in Section 3. We retain the notation of the main paper and Appendix A. In particular, $n_{\mathrm{fit}} = |\mathcal{D}_0^{\mathrm{fit}}|$, M is the number of retained source atoms, $\phi_{ie} = \phi_e(X_i)$ is the centered output-space tangent of source atom e at target observation i, g_i is the observed centered output score, and W_i is the base curvature on $\mathbb{R}_0^C$. All curvature inverses, square roots, and pseudoinverses below are taken on the centered output space. The source atomization, nuisance-model class, numerical tolerances, candidate grids, and sign policy are fixed without using validation or test labels. The default numerical constants used in the implementation described below are

$$c_{\mathrm{fd}} = 10^{-3}, \quad \varepsilon = 10^{-6}, \quad \tau_{\mathrm{inner}} = 10^{-6}, \quad \tau_{\mathrm{trust}} = 10^{-6}, \quad \tau_{\mathrm{supp}} = 10^{-8}, \quad \tau_\perp = 10^{-12}.$$

Atoms whose parameter norm is at most $10^{-12}\max\{1, \|\boldsymbol{\theta}_0\|_2\}$ are removed before tangent construction. The Gram-diagonal threshold is

$$\tau_G = 10^{-10}\,\mathrm{median}\{\hat{G}_{ee}^{\mathrm{raw}} : \hat{G}_{ee}^{\mathrm{raw}} > 0\}.$$

These constants are numerical safeguards, not statistically tuned hyperparameters, and are held fixed across all targets.

B.1 Geometry Construction, Cross-Fitting, and Stabilization

For source atom $\boldsymbol{\Delta}_e$, the exact first-order output perturbation is $\phi_{ie} = P_C J_0(X_i)\boldsymbol{\Delta}_e$. A forward-mode Jacobian-vector product (JVP) computes this quantity without forming the parameter Jacobian $J_0(X_i)$. The JVP is evaluated only on the prespecified shared coordinates to which source transfer is restricted; any source-specific output head has already been removed from the atom definition. In automatic-differentiation systems with efficient forward mode, one JVP has a cost comparable to a constant number of forward passes. When the output dimension is small, reverse-over-forward implementations provide an equivalent way to obtain the same product.

When an exact JVP is unavailable, we use a centered finite difference along the source direction. For every retained atom let

$$\epsilon_{\mathrm{fd},e} = c_{\mathrm{fd}}\frac{\max\{1, \|\boldsymbol{\theta}_0\|_2\}}{\|\boldsymbol{\Delta}_e\|_2},$$

and define

$$\tilde{\phi}_{ie} = P_C\frac{z_{\boldsymbol{\theta}_0+\epsilon_{\mathrm{fd},e}\boldsymbol{\Delta}_e}(X_i) - z_{\boldsymbol{\theta}_0-\epsilon_{\mathrm{fd},e}\boldsymbol{\Delta}_e}(X_i)}{2\epsilon_{\mathrm{fd},e}}. \tag{B.1}$$

The scaling keeps the norm of the parameter perturbation approximately constant across source atoms. The atom-norm threshold stated at the beginning of this appendix prevents division by a numerically negligible source direction.

To make the approximation error explicit, we fix (i, e) and set $f_{ie}(t) = P_C z_{\boldsymbol{\theta}_0 + t\boldsymbol{\Delta}_e}(X_i)$. If f_{ie} is three-times continuously differentiable on $[-\epsilon_{\mathrm{fd},e}, \epsilon_{\mathrm{fd},e}]$ and

$$\sup_{|t| \le \epsilon_{\mathrm{fd},e}} \|f_{ie}^{(3)}(t)\|_2 \le L_{z,3} \|\boldsymbol{\Delta}_e\|_2^3,$$

then Taylor expansion about zero gives

$$\left\| \tilde{\phi}_{ie} - \phi_{ie} \right\|_2 \le \frac{L_{z,3}}{6} \epsilon_{\mathrm{fd},e}^2 \|\boldsymbol{\Delta}_e\|_2^3.$$

Indeed, expanding $f_{ie}(\epsilon)$ and $f_{ie}(-\epsilon)$ to third order, subtracting the two expansions, and dividing by 2ϵ cancels the even-order terms and leaves a third-order remainder of the displayed order. This bound applies directly to smooth network maps. For piecewise-smooth networks such as ReLU models, a global third-derivative bound is not available across activation boundaries; in that case (B.1) is treated purely as a numerical fallback. Whenever finite differences are used, we recompute the relevant geometry at step sizes $\epsilon_{\mathrm{fd},e}/10,\ \epsilon_{\mathrm{fd},e},\ 10\epsilon_{\mathrm{fd},e}$ and verify that source rankings and reported normalized diagnostics are stable. The default implementation uses exact automatic-differentiation JVPs whenever they are available.

The empirical signal entry can be computed without storing ϕ_{ie}. Define the target parameter gradient

$$\hat{\mathbf{g}}_\theta = \frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} J_0(X_i)^\top g_i.$$

Because $P_C g_i = g_i$ and $\phi_{ie} = P_C J_0(X_i) \boldsymbol{\Delta}_e$,

$$-\hat{\mathbf{g}}_\theta^\top \boldsymbol{\Delta}_e = -\frac{1}{n_{\mathrm{fit}}} \sum_i g_i^\top J_0(X_i) \boldsymbol{\Delta}_e = -\frac{1}{n_{\mathrm{fit}}} \sum_i g_i^\top \phi_{ie} = \hat{b}_e.$$

Thus a single reverse-mode pass over the target fit sample produces $\hat{\mathbf{g}}_\theta$, after which all M signal entries are obtained by inner products with the source atoms. JVPs are still required for the Gram matrix, source-span diagnostics, and the innovation-overlap penalty.

The population normalization in (2.4) is

$$R_0 = \|r_0\|_0^2 = \mathbb{E}\{g_0(X)^\top W_0(X)^\dagger g_0(X)\},$$

which depends on the conditional mean score $g_0(X)$, not the corresponding single-label score $g(X, Y)$. Using the latter directly would include irreducible label-noise energy and would therefore generally overstate the normalizing descent energy.

For multiclass cross-entropy, $p_i = \mathrm{softmax}\{z_0(X_i)\}$, $W_i = \mathrm{diag}(p_i) - p_i p_i^\top$, and $g_0(X_i) = p_i - \eta_0(X_i)$, where $\eta_0(x) = P_0(Y = \cdot \mid X = x)$. We partition $\mathcal{D}_0^{\mathrm{fit}}$ into $K = 5$ folds $\{\mathcal{I}_k\}_{k=1}^K$. For each k, a ridge-regularized multinomial logistic nuisance model is fitted on frozen base features using only $\mathcal{D}_0^{\mathrm{fit}} \setminus \mathcal{I}_k$. If $i \in \mathcal{I}_k$, write $k(i) = k$ and define $\hat{g}_{0,-k(i)}(X_i) = p_i - \hat{\eta}_{-k(i)}(X_i)$. The same construction extends to a general smooth loss by cross-fitting a regression of the observed score g_i on frozen base features. Any nuisance-model hyperparameters are either prespecified or selected entirely inside the corresponding nuisance-training portion. Neither the validation split nor the test split is used for this nuisance step. The cross-fitted energy estimator in (3.3) is

$$\hat{R}_{\mathrm{cf}} = \frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} \hat{g}_{0,-k(i)}(X_i)^\top W_i^\dagger \hat{g}_{0,-k(i)}(X_i).$$

The following deterministic calculation isolates exactly what is required for consistency. Let

$$e_i = \hat{g}_{0,-k(i)}(X_i) - g_0(X_i), \quad R_n^0 = \frac{1}{n_{\mathrm{fit}}} \sum_i g_0(X_i)^\top W_i^\dagger g_0(X_i), \quad E_n = \frac{1}{n_{\mathrm{fit}}} \sum_i e_i^\top W_i^\dagger e_i.$$

Expanding the squared norm gives

$$\hat{R}_{\mathrm{cf}} - R_n^0 = \frac{2}{n_{\mathrm{fit}}} \sum_i g_0(X_i)^\top W_i^\dagger e_i + E_n.$$

By Cauchy-Schwarz in the product Hilbert space with inner product

$$\langle a, b\rangle_n = \frac{1}{n_{\mathrm{fit}}} \sum_i a_i^\top W_i^\dagger b_i, \quad \left| \frac{1}{n_{\mathrm{fit}}} \sum_i g_0(X_i)^\top W_i^\dagger e_i \right| \le \sqrt{R_n^0 E_n}.$$

Therefore

$$|\hat{R}_{\mathrm{cf}} - R_n^0| \le 2\sqrt{R_n^0 E_n} + E_n. \tag{B.2}$$

Consequently, if the nuisance learner satisfies $E_n = o_p(1)$ and a law of large numbers gives $R_n^0 \to_p R_0$, then $\hat{R}_{\mathrm{cf}} \to_p R_0$. Cross-fitting by itself does not assert that a particular nuisance model is consistent; its role is to prevent an observation's label from being used to fit its own conditional score. Equation (B.2) makes the remaining nuisance-estimation requirement explicit.

For numerical inversion, the implementation uses $W_{i,\varepsilon} = W_i + \varepsilon P_C$ and $\varepsilon = 10^{-6}$, with the value fixed before validation. All inverses are still taken on $\mathbb{R}_0^C$. Under Assumption 2,

$$\|W_{i,\varepsilon}^{-1} - W_i^{-1}\|_{\mathrm{op}} \le \frac{\varepsilon}{\underline{\kappa}(\underline{\kappa} + \varepsilon)} \le \frac{\varepsilon}{\underline{\kappa}^2}.$$

Likewise,

$$u^\top W_{i,\varepsilon} v - u^\top W_i v = \varepsilon u^\top P_C v.$$

Hence, for bounded tangent and score fields, the induced inverse-curvature and Gram perturbations are $O(\varepsilon)$ uniformly. The population theory in Appendix A corresponds to $\varepsilon = 0$. The nuisance estimator is used only to normalize principal angles, directed alignments, and completeness. The empirical signal $\hat{\mathbf{b}}$ itself remains the observed-score estimator in (3.1). We also compute

$$\hat{R}_{\mathrm{raw}} = \frac{1}{n_{\mathrm{fit}}} \sum_i g_i^\top W_i^\dagger g_i$$

only as a numerical scale for the trust-radius grid. Because it contains single-label score noise, $\hat{R}_{\mathrm{raw}}$ is not interpreted as an estimator of R_0.

For large-output models, particularly language models, storing every token-logit tangent vector can dominate memory. We partition the fit data into mini-batches $\{\mathcal{B}_b\}_{b=1}^{B_{\mathrm{geom}}}$. Let d_b denote the total number of centered output coordinates entering the loss in batch b, and define the stacked, curvature-whitened tangent $Z_{be} = \mathrm{col}_{i\in\mathcal{B}_b}\{W_i^{1/2}\phi_{ie}\} \in \mathbb{R}^{d_b}$. For each batch draw an independent Rademacher matrix $R_b \in \mathbb{R}^{d_b \times r_{\mathrm{sk}}}$, $(R_b)_{jk} \in \{-\frac{1}{\sqrt{r_{\mathrm{sk}}}}, \frac{1}{\sqrt{r_{\mathrm{sk}}}}\}$ with independent entries, and use the same R_b for every source atom in that batch. Define $\tilde{Z}_{be} = R_b^\top Z_{be}$. For the language experiments we use $r_{\mathrm{sk}} = 64$ by default; the unsketched geometry is used whenever the centered output dimension is small enough to store directly. The sketched Gram estimator is

$$\tilde{G}_{ee'} = \frac{1}{n_{\mathrm{fit}}} \sum_{b=1}^{B_{\mathrm{geom}}} \tilde{Z}_{be}^\top \tilde{Z}_{be'}. \tag{B.3}$$

Conditional on the observed tangents, $\mathbb{E}_R(R_b R_b^\top) = I_{d_b}$, and therefore $\mathbb{E}_R\{\tilde{Z}_{be}^\top \tilde{Z}_{be'} \mid Z_{be}, Z_{be'}\} = Z_{be}^\top Z_{be'}$. Thus (B.3) is conditionally unbiased for the unsketched empirical Gram. For completeness, we record the inner-product distortion used by the implementation. Conditional on fixed vectors $u, v \in \mathbb{R}^{d_b}$, a Rademacher sketch with the above normalization satisfies the bilinear concentration bound

$$\mathbb{P}_R\left(\left|u^\top R_b R_b^\top v - u^\top v\right| > \epsilon_{\mathrm{sk}} \|u\|_2 \|v\|_2\right) \le 2\exp(-c r_{\mathrm{sk}} \epsilon_{\mathrm{sk}}^2), \quad 0 < \epsilon_{\mathrm{sk}} < 1,$$

for a universal constant $c > 0$. This is the standard sub-Gaussian inner-product form of the Johnson-Lindenstrauss concentration inequality and follows, for example, by applying a bilinear Hanson-Wright bound to the independent Rademacher columns of R_b. Apply the bound conditionally to every pair $(Z_{be}, Z_{be'})$, every $e, e' \in [M]$, and every geometry batch b. A union bound shows that if

$$r_{\mathrm{sk}} \ge C\epsilon_{\mathrm{sk}}^{-2} \log\left(\frac{2B_{\mathrm{geom}} M^2}{\delta}\right),$$

then with probability at least $1 - \delta$, simultaneously for all batches and atom pairs,

$$\left|\tilde{Z}_{be}^\top \tilde{Z}_{be'} - Z_{be}^\top Z_{be'}\right| \le \epsilon_{\mathrm{sk}} \|Z_{be}\|_2 \|Z_{be'}\|_2.$$

Consequently,

$$\left|\tilde{G}_{ee'} - \hat{G}_{ee'}\right| \le \frac{\epsilon_{\mathrm{sk}}}{n_{\mathrm{fit}}} \sum_{b=1}^{B_{\mathrm{geom}}} \|Z_{be}\|_2 \|Z_{be'}\|_2$$

simultaneously over e, e'. This is the deterministic Gram perturbation in Appendix A. Independent sketches across batches avoid unnecessary coupling of approximation errors, while using one common sketch across all source atoms within a batch is essential for preserving their pairwise geometry.

Let $\hat{\mathbf{G}}^{\mathrm{raw}}$ and $\hat{\mathbf{b}}^{\mathrm{raw}}$ denote the empirical Gram and signal before atom normalization. An atom is removed if $\hat{G}_{ee}^{\mathrm{raw}} \le \tau_G$. The default relative value of τ_G was stated at the beginning of this appendix and is computed using the fit split only. If no positive diagonal entry exists, the source dictionary is declared empty and the procedure falls back to the target-only candidate. For every retained atom define $d_e = \sqrt{\hat{G}_{ee}^{\mathrm{raw}}}$, $\bar{\mathbf{\Delta}}_e = \frac{\mathbf{\Delta}_e}{d_e}$, and $\bar{\phi}_{ie} = \frac{\phi_{ie}}{d_e}$. If $D = \mathrm{diag}(d_e)$ over the retained atoms, then $\bar{\mathbf{G}} = D^{-1}\hat{\mathbf{G}}^{\mathrm{raw}}D^{-1}$, $\bar{\mathbf{b}} = D^{-1}\hat{\mathbf{b}}^{\mathrm{raw}}$, and $\bar{G}_{ee} = 1$. Normalization is performed before stabilization. After this step we relabel $(\bar{\mathbf{\Delta}}_e, \bar{\phi}_{ie}, \bar{\mathbf{G}}, \bar{\mathbf{b}})$ as $(\mathbf{\Delta}_e, \phi_{ie}, \hat{\mathbf{G}}, \hat{\mathbf{b}})$ to simplify notation. The stabilized Gram used in the projection criterion is

$$\hat{\mathbf{G}}_{\mathrm{st}}(\rho_G, \lambda_G) = (1-\rho_G)\hat{\mathbf{G}} + \rho_G \,\mathrm{diag}(\hat{\mathbf{G}}) + \lambda_G I_M, \quad 0 \le \rho_G \le 1, \quad \lambda_G \ge 0. \tag{B.4}$$

The off-diagonal shrinkage reduces sensitivity to uncertain source-source interactions, while the ridge term controls numerical singularity. Source redundancy itself is not removed: the unshrunk Gram $\hat{\mathbf{G}}$ continues to define the empirical functional norm and the trust constraint. If a denotes a coefficient vector in the normalized dictionary, the parameter update is

$$\sum_e a_e \bar{\mathbf{\Delta}}_e = \sum_e \frac{a_e}{d_e} \mathbf{\Delta}_e^{\mathrm{raw}}.$$

Thus the coefficient on the original source-update scale is $a_e^{\mathrm{raw}} = a_e/d_e$. This inverse rescaling is performed before the final source update is assembled. For the normalized geometry diagnostics, we use the fixed stabilization pair $(\rho_G^{\mathrm{diag}}, \lambda_G^{\mathrm{diag}}) = (0.25, 10^{-4})$, chosen before validation, and define $\hat{\mathbf{G}}_{\mathrm{diag}} = \hat{\mathbf{G}}_{\mathrm{st}}(\rho_G^{\mathrm{diag}}, \lambda_G^{\mathrm{diag}})$. This diagnostic stabilization is distinct from the candidate-specific pair (ρ_G, λ_G) used in the sparse projection objective. Consequently, $\hat{\Theta}_s$, $\hat{A}_e^{\mathrm{dir}}$, and $\hat{\chi}_q$ are computed once from a fixed target geometry and do not vary across transfer candidates. The regularized projection energies are computed with $\hat{\mathbf{G}}_{\mathrm{diag}}$ and $\hat{R}_{\mathrm{cf}}$ from (3.3). Squared normalized quantities are clipped to $[0,1]$ only because finite-sample and regularization error can otherwise produce values slightly outside the population range. If $\hat{R}_{\mathrm{cf}}$ falls below a prespecified numerical floor, we follow the zero-signal convention of the main text: normalized angle diagnostics are reported as zero and the unnormalized projection energies are retained.

The empirical directed alignment itself is

$$\hat{A}_e^{\mathrm{dir}} = \frac{\hat{b}_e}{\sqrt{\hat{G}_{ee}}\sqrt{\hat{R}_{\mathrm{cf}}}}.$$

For display only, we report $[\hat{A}_e^{\mathrm{dir}}]_{-1}^{1} = \max\{-1, \min(1, \hat{A}_e^{\mathrm{dir}})\}$. The unclipped value is retained internally, so clipping does not alter ranking or optimization. Exact computation of

$$\hat{\chi}_q = \max_{\substack{A \subset [M] \\ |A| \le q}} \frac{\hat{\mathbf{b}}_A^\top \hat{\mathbf{G}}_{\mathrm{diag},AA}^\dagger \hat{\mathbf{b}}_A}{\hat{R}_{\mathrm{cf}}}$$

is combinatorial. In the small controlled simulation we enumerate all eligible subsets. For larger source libraries we use a deterministic forward-selection and local-swap approximation. Starting from $A_0 = \varnothing$, at step t we add the atom maximizing the actual increment

$$\Delta_e(A_t) = \hat{\mathbf{b}}_{A_t\cup\{e\}}^\top \hat{\mathbf{G}}_{\mathrm{diag},A_t\cup\{e\},A_t\cup\{e\}}^\dagger \hat{\mathbf{b}}_{A_t\cup\{e\}} - \hat{\mathbf{b}}_{A_t}^\top \hat{\mathbf{G}}_{\mathrm{diag},A_tA_t}^\dagger \hat{\mathbf{b}}_{A_t}.$$

Without stabilization, this reduces to the incremental projection gain in Corollary A.3. After q forward additions, we repeatedly evaluate every one-for-one swap between the selected and unselected sets and accept the best strictly improving swap until no such swap remains. Because each accepted move increases the objective, this procedure terminates after finitely many distinct subsets. It is an algorithmic lower bound on the exact combinatorial maximum, not a claim of global optimality. In large-library experiments the reported completeness value is this greedy-swap approximation; the exact definition remains the quantity used in the theory.

In the nonlinear stress experiment of Appendix C only, we additionally compute on fit inputs the tangent-nonlinearity diagnostic

$$\hat{E}_{\mathrm{nl}}(a) = \frac{\left\| z_{\boldsymbol{\theta}_0 + \sum_e a_e \boldsymbol{\Delta}_e} - z_{\boldsymbol{\theta}_0} - \sum_e a_e J_0 \boldsymbol{\Delta}_e \right\|_{\mathrm{fit}}}{\left\| \sum_e a_e J_0 \boldsymbol{\Delta}_e \right\|_{\mathrm{fit}} + 10^{-8}}.$$

Candidate combinations exceeding a prespecified threshold are rejected in that stress-test analysis. This additional screening is used only to study breakdown of the local approximation and is not part of the default procedure in Algorithm 1.

B.2 Sparse Projection and Innovation Fitting

All coefficient optimization is performed on the unit-functional-norm atom dictionary constructed above. For a fixed candidate $(\rho_G, \lambda_G, \lambda_{\mathrm{src}}, \lambda_{\mathrm{atom}}, r)$, define

$$F(a) = \frac{1}{2} a^\top \hat{\mathbf{G}}_{\mathrm{st}} a - \hat{\mathbf{b}}^\top a + \lambda_{\mathrm{src}} \sum_{s=1}^{S} w_s \|a_{\mathcal{E}_s}\|_2 + \lambda_{\mathrm{atom}} \|a\|_1,$$

where $w_s = \sqrt{|\mathcal{E}_s|}$ and $\hat{\mathbf{G}}_{\mathrm{st}} = \hat{\mathbf{G}}_{\mathrm{st}}(\rho_G, \lambda_G)$. The trust-region problem in (3.6) is

$$\min_{a \in \mathbb{R}^M} F(a) \quad \text{subject to} \quad a^\top \hat{\mathbf{G}} a \le r^2. \tag{B.5}$$

The constraint uses the raw normalized Gram $\hat{\mathbf{G}}$ and excludes $\hat{\mathbf{G}}_{\mathrm{st}}$ because $a^\top \hat{\mathbf{G}} a = \| \sum_e a_e \phi_e \|_{\hat{0}}^2$ is the empirical functional size of the source update.

For $\nu \ge 0$, we use the Lagrangian

$$\mathcal{L}_\nu(a) = F(a) + \frac{\nu}{2} \{ a^\top \hat{\mathbf{G}} a - r^2 \}.$$

Its smooth Hessian is $H_\nu = \hat{\mathbf{G}}_{\mathrm{st}} + \nu \hat{\mathbf{G}}$, and the smooth gradient is $H_\nu a - \hat{\mathbf{b}}$. The nonsmooth penalty is the nonoverlapping sparse-group penalty. For a proximal step size $\eta > 0$, let $v = a - \eta(H_\nu a - \hat{\mathbf{b}})$. The exact proximal map is obtained in two stages. First apply elementwise soft thresholding

$$u_e = \mathrm{sign}(v_e) \left(|v_e| - \eta \lambda_{\mathrm{atom}} \right)_+ .$$

Then, for every source group,

$$a^+_{\mathcal{E}_s} = \left(1 - \frac{\eta \lambda_{\mathrm{src}} w_s}{\|u_{\mathcal{E}_s}\|_2} \right)_+ u_{\mathcal{E}_s},$$

with $a^+_{\mathcal{E}_s} = 0$ when $u_{\mathcal{E}_s} = 0$. The formula is the exact proximal operator because the source groups form a nonoverlapping partition and the ℓ_1 penalty is coordinatewise separable. Under the optional nonnegative sign policy, the first stage becomes $u_e = (v_e - \eta \lambda_{\mathrm{atom}})_+$, after which the same group shrinkage is applied. The signed policy is the default; the nonnegative policy is a prespecified sensitivity analysis and is not chosen from target test performance.

At fixed ν, we use accelerated proximal gradient with backtracking. Backtracking begins from the previous accepted step size and reduces it until the standard quadratic majorization inequality for the smooth part holds. Equivalently, any $0 < \eta \le \|H_\nu\|_{\mathrm{op}}^{-1}$ is valid. We monitor the proximal-gradient mapping

$$\mathcal{G}_\eta(a) = \frac{1}{\eta} \left[a - \mathrm{prox}_{\eta \mathcal{P}} \{ a - \eta(H_\nu a - \hat{\mathbf{b}}) \} \right]$$

and stop the inner solve when

$$\|\mathcal{G}_\eta(a)\|_\infty \le \tau_{\mathrm{inner}} \{ 1 + \|\hat{\mathbf{b}}\|_\infty \},$$

where τ_{inner} is fixed before validation. Warm starts are used between neighboring multiplier values. If the minimizer at $\nu = 0$ satisfies $a^\top \hat{\mathbf{G}} a \le r^2$, it is also the solution of (B.5). Otherwise the constrained optimum lies on the trust boundary. This follows from convexity and the KKT conditions. Because $r > 0$, $a = 0$ is strictly feasible, so Slater's condition holds and strong duality applies.

The trust norm of the Lagrangian solution is monotone in ν, which justifies bisection. To see this, let a_1 and a_2 minimize $\mathcal{L}_{\nu_1}$ and $\mathcal{L}_{\nu_2}$ with $0 \le \nu_1 < \nu_2$. Optimality gives

$$F(a_1) + \frac{\nu_1}{2} q(a_1) \le F(a_2) + \frac{\nu_1}{2} q(a_2) \text{ and } F(a_2) + \frac{\nu_2}{2} q(a_2) \le F(a_1) + \frac{\nu_2}{2} q(a_1),$$

where $q(a) = a^\top \hat{\mathbf{G}} a$. Adding the inequalities yields

$$(\nu_2 - \nu_1)\{q(a_2) - q(a_1)\} \le 0,$$

so $q(a_2) \le q(a_1)$. In the implementation $\lambda_G > 0$ for the stabilized computational candidates, making the smooth quadratic coercive and ensuring that the multiplier can be bracketed by repeated doubling. Bisection then locates a multiplier whose converged primal solution satisfies the trust-boundary KKT condition to the requested tolerance. Algorithm B.1 gives the sparse source-span projection procedure. The returned normalized-dictionary coefficients are mapped back to the original source-update scale using the inverse normalization described above before constructing the transferred parameter update.

Algorithm B.1 Sparse source-span projection

Require: normalized $\hat{\mathbf{G}}, \hat{\mathbf{b}}$; stabilized $\hat{\mathbf{G}}_{\text{st}}$; source groups; λ_{src}, λ_{atom}; trust radius $r > 0$; sign policy; $\tau_{\text{inner}}, \tau_{\text{trust}}$.

Set $\nu = 0$ and solve the sparse-group objective by accelerated proximal gradient with backtracking.
if $a^\top \hat{\mathbf{G}} a \le r^2 + \tau_{\text{trust}}(1 + r^2)$ **then**
    Return the unconstrained solution and $\hat{\nu} = 0$.
end if
Set $\nu_{\text{lo}} = 0$ and initialize $\nu_{\text{hi}} > 0$.
while the converged solution at ν_{hi} satisfies $a^\top \hat{\mathbf{G}} a > r^2$ **do**
    Set $\nu_{\text{lo}} \leftarrow \nu_{\text{hi}}$ and $\nu_{\text{hi}} \leftarrow 2\nu_{\text{hi}}$.
end while
while the trust-boundary residual exceeds τ_{trust} **do**
    Set $\nu \leftarrow (\nu_{\text{lo}} + \nu_{\text{hi}})/2$.
    Warm-start and solve the penalized problem with smooth Hessian $H_\nu = \hat{\mathbf{G}}_{\text{st}} + \nu \hat{\mathbf{G}}$.
    if $a^\top \hat{\mathbf{G}} a > r^2$ **then**
        Set $\nu_{\text{lo}} \leftarrow \nu$.
    else
        Set $\nu_{\text{hi}} \leftarrow \nu$.
    end if
end while
return $\hat{a} = a$, the final multiplier $\hat{\nu}$, and the optimization residuals.

Let $\hat{A} = \{e : |\hat{a}_e| > \tau_{\text{supp}}\}$, $\tau_{\text{supp}} = 10^{-8}$, and $\boldsymbol{\theta}_{\text{src}} = \boldsymbol{\theta}_0 + \sum_e \hat{a}_e \boldsymbol{\Delta}_e$ after mapping the coefficients back to the original source-update scale. The target innovation is represented by a LoRA or bottleneck adapter ω with parameter-space update u_ω and zero functional initialization $u_{\omega_0} = 0$. For factored adapters, "zero initialization" refers to zero initial function update; the internal factor initialization may be chosen in the standard way that preserves nonzero optimization gradients. The actual target loss is evaluated at the nonlinear predictor $f_{\boldsymbol{\theta}_{\text{src}} + u_\omega}$. The overlap penalty, however, is measured in the same base functional geometry used to construct the source span $\psi_{\omega i} = P_C J_0(X_i) u_\omega$. Define the empirical overlap coordinate

$$m_e(\omega) = \frac{1}{n_{\text{fit}}} \sum_{i=1}^{n_{\text{fit}}} \phi_{ie}^\top W_i \psi_{\omega i}, \quad e \in \hat{A}.$$

Then the innovation objective in (3.8) is

$$\hat{L}_{\text{innov}}(\omega) = \frac{1}{n_{\text{fit}}} \sum_{i=1}^{n_{\text{fit}}} \ell\{f_{\boldsymbol{\theta}_{\text{src}} + u_\omega}(X_i), Y_i\} + \lambda_\perp \sum_{e \in \hat{A}} m_e(\omega)^2 + \lambda_\omega \|\omega\|_2^2. \tag{B.6}$$

Thus $\lambda_\perp$ discourages the target adapter from relearning directions supplied by the selected source span, while the target loss itself remains the actual nonlinear loss, not merely its local quadratic approximation. The penalty is therefore orthogonality-regularizing, not an exact hard-orthogonality constraint.

For batchwise implementation, let $F_b = [\phi_{ie}]_{i\in\mathcal{B}_b,\ e\in\hat{A}}$, interpreted as the stacked selected-source tangent matrix, and let W_b be the corresponding block-diagonal curvature operator. The batch contribution to the overlap vector is

$$m_b(\omega) = \frac{1}{n_{\text{fit}}} F_b^\top W_b \psi_{\omega,b}.$$

We accumulate $m(\omega) = \sum_b m_b(\omega)$ over the complete fit split and only then square $L_\perp(\omega) = \|m(\omega)\|_2^2$. This accumulation order is important. Squaring each mini-batch overlap before summing would optimize a different objective containing additional batch-variance terms. Accumulating first reproduces exactly the empirical penalty in (B.6). Because the fit sample is small in the intended few-shot regime, this exact accumulation is practical. The source tangents F_b are cached during geometry construction; only the current adapter tangent $\psi_{\omega,b}$ must be recomputed. The functional overlap uses J_0 and W_i at the common base, excluding Euclidean parameter orthogonality entirely. This keeps the innovation penalty tied to the same target-conditioned output geometry used for source selection. Parameter-space orthogonality is included only as an ablation.

B.3 Tuning and Computational Complexity

Every candidate is fully constructed and fitted using $\mathcal{D}_0^{\text{fit}}$ before its validation loss is examined. The default finite grid is

$$\rho_G \in \{0, 0.25, 0.5\} \text{ and } \lambda_G \in \{10^{-5}, 10^{-4}, 10^{-3}\}.$$

After unit-functional-norm normalization, the theoretical rate scales from Theorem 4.3 are instantiated as

$$\lambda_{\text{src}}(c_{\text{src}}) = c_{\text{src}} \sqrt{\frac{1 + \log(2S)}{n_{\text{fit}}}}, \quad c_{\text{src}} \in \{0.5, 1, 2\},$$

and

$$\lambda_{\text{atom}}(c_{\text{atom}}) = c_{\text{atom}} \sqrt{\frac{\log(2M)}{n_{\text{fit}}}}, \quad c_{\text{atom}} \in \{0.5, 1, 2\}.$$

For cross-entropy with normalized source tangents, the bounded-score scale is absorbed into the finite multipliers. These multipliers are tuning constants, not estimates of the unknown constants in the oracle theorem. The trust-radius grid is

$$r = c_r \hat{R}_{\text{raw}}^{1/2}, \quad c_r \in \{0.25, 0.5, 1\}.$$

The quantity $\hat{R}_{\text{raw}}$ is used here only to put trust radii on a comparable numerical scale across targets; it is never substituted for $\hat{R}_{\text{cf}}$ in normalized transfer diagnostics.

The innovation-overlap multipliers are $\lambda_\perp \in \{0, 0.1, 1, 10\}$, and the adapter-weight penalty is fixed at $\lambda_\omega = 10^{-4}$. Fixing λ_ω avoids enlarging the validation candidate family solely to tune ordinary adapter weight decay. To avoid candidate-dependent units, the overlap penalty is divided by a fit-only reference scale. For each source-projection candidate, starting from the common zero-functional-update adapter initialization, we take the same prespecified first nonzero pilot update using the unpenalized target loss and compute $s_\perp = \max\{L_\perp(\omega^{(1)}), \tau_\perp\}$, where $\tau_\perp > 0$ is a fixed numerical floor. The normalized penalty $L_\perp / s_\perp$ is then used for all four $\lambda_\perp$ values attached to that source-projection candidate. The pilot step and scale use fit data only and are completed before validation evaluation. The target-adapter architecture, parameter budget, optimizer family, and stopping rule are identical across transferred candidates and the target-only comparator.

Let Λ denote the resulting finite set of completely fitted transferred candidates. For classification and multiple-choice tasks, the bounded gate loss is

$$\ell_{\text{gate}}(f(x), y) = \mathbf{1}\{\hat{y}(x) \neq y\}, \quad B = 1.$$

For short-form QA evaluated by per-example F1, we use

$$\ell_{\text{gate}}(f(x), y) = 1 - \text{F1}\{f(x), y\}, \quad B = 1.$$

When exact match is the task metric, its corresponding 0-1 error is used. These gate losses are fixed before validation. The signed coefficient policy is the default. The nonnegative policy is run as a

separately declared sensitivity analysis and is not chosen by validation unless it is explicitly included in Λ before any validation labels are examined. The source atom partition is never tuned on target labels. On the validation split, for every $\lambda \in \Lambda$ compute

$$D_i(\lambda) = \ell_{\mathrm{gate}}\{f_{\hat{\boldsymbol{\theta}}_\lambda}(X_i), Y_i\} - \ell_{\mathrm{gate}}\{f_{\hat{\boldsymbol{\theta}}_{\mathrm{TO}}}(X_i), Y_i\}$$

and its empirical-Bernstein upper bound U_λ from Section 3. Candidates satisfying $U_\lambda \le 0$ are admissible. Among admissible candidates, the algorithm chooses the one with the smallest validation mean $\bar{D}_\lambda$; deterministic ties are resolved by the fixed candidate ordering. If no candidate is admissible, the target-only comparator is deployed. Test observations are used only after this choice has been made.

The relevant dimensions are the transferred parameter count p_{tr}, the number of fit observations n_{fit}, the number of source atoms M, the number of selected atoms $\hat{q} = |\hat{A}|$, the output-sketch rank r_{sk}, and, when used, the low-rank geometry dimension r_{geom}. The empirical signal requires one reverse-mode pass through the fit data to obtain $\hat{\mathbf{g}}_\theta$, followed by M dot products with source updates. Its arithmetic cost is linear in Mp_{tr} once the gradient is available, and it requires no $n_{\mathrm{fit}} \times M$ tangent store. Exact Gram construction requires one JVP per source atom and fit batch. If a batch has stacked centered-output dimension d_b, forming all pairwise inner products directly costs $O(d_b M^2)$ after the M tangents are available. With the Rademacher sketch, the contraction cost becomes $O(r_{\mathrm{sk}} M^2)$ per batch and the temporary tangent-sketch storage is $O(Mr_{\mathrm{sk}})$. The dense Gram itself requires $O(M^2)$ memory. Cross-fitting the denominator requires $K = 5$ nuisance fits on frozen base features and one held-out prediction for every fit observation. Its cost is governed by the nuisance-feature dimension, not by the neural parameter dimension p_{tr}. At fixed trust multiplier ν, a dense proximal-gradient iteration requires one multiplication by H_ν, costing $O(M^2)$, and an $O(M)$ sparse-group proximal map. Bisection requires

$$O\left(\log \frac{\nu_{\mathrm{hi}} - \nu_{\mathrm{lo}}}{\tau_\nu}\right)$$

outer multiplier updates after bracketing, where $\tau_\nu > 0$ is the fixed multiplier-bracket tolerance; warm starts substantially reducing the number of inner iterations. This regime is practical for the intended source libraries, where M is tens to a few hundred even when p_{tr} is extremely large.

For larger libraries, we do not materialize the dense Gram. If the whitened tangent matrix admits a rank-r_{geom} factorization

$$\hat{\mathbf{G}} \approx UU^\top + D, \quad U \in \mathbb{R}^{M \times r_{\mathrm{geom}}},$$

then a Gram-vector product costs $O(Mr_{\mathrm{geom}})$ and storage is $O(Mr_{\mathrm{geom}})$ in addition to the diagonal. The same factorization is used inside proximal-gradient and trust-multiplier iterations. This changes only the numerical linear algebra; any approximation error enters the perturbed-geometry conditions of Corollary A.8. During innovation fitting, the selected source tangents are cached. Computing the exact overlap vector requires one current adapter tangent and inner products with $\hat{q}$ cached source directions. In unsketched form the contraction is linear in the relevant output dimension and $\hat{q}$; with the same output sketch it is $O(r_{\mathrm{sk}}\hat{q})$ per batch. The dominant remaining cost is therefore ordinary target-adapter training. All stochastic sketches, nuisance-fold assignments, optimization seeds, and tie-breaking orders are recorded. Candidate construction uses only the fit split, the finite risk gate uses only the validation split, and the test split remains untouched until final evaluation. These conventions ensure that the computational approximations in this appendix do not alter the sample-separation conditions used by the theoretical guarantees.

C Complete Monte Carlo Design and Additional Simulations

This part gives the complete synthetic design used to isolate the geometric, estimation, and deployment claims of the paper. The simulation is carefully constructed so that reusable source structure, source redundancy, target innovation, coordinate reparameterization, and departure from the local tangent approximation can be varied separately. Unless a factor is explicitly varied, all fitting, stabilization, tuning, and gating choices are those specified in Appendix B. Fit, validation, test, and oracle samples are mutually independent. All methods within a Monte Carlo replication receive the same generated source bank and the same target splits, and all random seeds and optimizer diagnostics are recorded.

C.1 Data-Generating Mechanism and Oracle Targets

We generate $X \sim \mathcal{N}(0, I_{40})$ and use a two-layer ReLU network with $C = 5$ output classes and hidden width $H = 128$,

$$z_{\boldsymbol{\theta}}(x) = W_2\sigma(W_1x), \quad W_1 \in \mathbb{R}^{128\times 40}, \quad W_2 \in \mathbb{R}^{5\times 128}.$$

The base parameter is $\boldsymbol{\theta}_0 = (B_1, B_2)$ and $z_0(x) = B_2\sigma(B_1x)$. Before normalization, entries of B_1 and B_2 are drawn independently from centered Gaussian distributions with variance equal to the reciprocal of the corresponding input dimension. The columns of B_1 and the rows of B_2 are then normalized to unit Euclidean norm. The same generated base is used by every method within a replication. The base class probabilities and curvature are

$$p_0(x) = \mathrm{softmax}\{z_0(x)\}, \quad W_0(x) = \mathrm{diag}\{p_0(x)\} - p_0(x)p_0(x)^\top,$$

with $W_0(x)$ understood on the centered-logit space $\mathbb{R}^5_0$ as in Section 2.

We generate $K_{\mathrm{fac}} = 8$ low-rank parameter directions $V_k = (V_{1k}, V_{2k}),\ k \in [8]$, each perturbing both layers. The raw first-layer perturbation is generated as

$$V_{1k}^{\mathrm{raw}} = A_{1k}B_{1k}^\top, \quad A_{1k} \in \mathbb{R}^{128\times r_V}, \quad B_{1k} \in \mathbb{R}^{40\times r_V},$$

and the raw second-layer perturbation as

$$V_{2k}^{\mathrm{raw}} = A_{2k}B_{2k}^\top, \quad A_{2k} \in \mathbb{R}^{5\times r_V}, \quad B_{2k} \in \mathbb{R}^{128\times r_V},$$

with all factor matrices having independent standard-normal entries. The reference rank is $r_V = 4$.

Let $\phi_k^{\mathrm{raw}}(x) = P_C J_0(x) V_k^{\mathrm{raw}}$ denote the raw tangent field. To separate source coverage from source redundancy, the eight latent factors are orthogonalized under the population target-functional inner product induced by the base curvature. Specifically, using an independent construction sample $\{X_j^{\mathrm{geom}}\}_{j=1}^{N_{\mathrm{geom}}}$, we apply Gram-Schmidt to the parameter directions using

$$\langle u, v\rangle_{\mathrm{geom}} = \frac{1}{N_{\mathrm{geom}}} \sum_{j=1}^{N_{\mathrm{geom}}} u(X_j^{\mathrm{geom}})^\top W_0(X_j^{\mathrm{geom}}) v(X_j^{\mathrm{geom}}).$$

Because the tangent map is linear in the parameter direction, subtracting a projection in tangent space is implemented by subtracting the corresponding linear combination of parameter directions. The resulting V_k are rescaled so that $\|\phi_k\|_{\mathrm{geom}} = 1,\ \phi_k(x) = P_C J_0(x) V_k$. Thus, before controlled source-level correlation is introduced, $\langle \phi_k, \phi_{k'}\rangle_{\mathrm{geom}} = 0\ (k \neq k')$. The reference construction uses $N_{\mathrm{geom}} = 200{,}000$. This large unlabeled construction sample is part of the simulation generator and is never supplied to a fitted method.

The reusable component of the target parameter is

$$\boldsymbol{\theta}^\star_{\mathrm{supp}} = \boldsymbol{\theta}_0 + \sum_{k=1}^{8} \alpha_k^\star V_k, \quad \boldsymbol{\alpha}^\star = (1.0, 0.7, 0.5, 0, \ldots, 0)^\top.$$

Hence exactly three of the eight latent transferable factors enter the target-supported component. To introduce target structure unavailable from every source, generate an independent two-layer innovation module

$$u_{\mathrm{raw}}(x) = D_2\sigma(D_1x), \text{ with } D_1 \in \mathbb{R}^{r_{\mathrm{innov}}\times 40}, \quad D_2 \in \mathbb{R}^{5\times r_{\mathrm{innov}}},$$

and independent Gaussian entries scaled by their input dimensions. The reference innovation width is $r_{\mathrm{innov}} = 8$. Center the output and remove its component in the latent source span

$$u_\perp = P_C u_{\mathrm{raw}} - \mathrm{Proj}_{\mathrm{span}\{\phi_1,\ldots,\phi_8\}}(P_C u_{\mathrm{raw}}),$$

where the projection uses the same population functional metric as above. Finally rescale $u_\perp$ to unit functional norm. The target teacher logits are

$$z^\star(x) = z_{\boldsymbol{\theta}^\star_{\mathrm{supp}}}(x) + \beta_{\mathrm{innov}} u_\perp(x), \tag{C.1}$$

with $\beta_{\mathrm{innov}} = 0.6$ in the reference condition. Labels are drawn from

$$Y \mid X = x \sim \mathrm{Categorical}\left[\eta^\star(x)\right], \quad \eta^\star(x) = \mathrm{softmax}\{z^\star(x)\}. \tag{C.2}$$

For cross-entropy, the exact conditional score at the base is therefore

$$g_0(x) = p_0(x) - \eta^\star(x).$$

This makes the population natural-descent field, the source-span completeness, and all oracle source utilities available to arbitrary Monte Carlo precision without fitting a nuisance model.

Each synthetic source checkpoint is one source atom throughout Appendix C; hence $M = S$ in the Monte Carlo study. For source s, let $c_s = (c_{s1}, \dots, c_{s8})^\top$ be its latent-factor coefficient vector. Before adding source-specific nuisance, define

$$\Delta_s^{\mathrm{fac}} = \sum_{k=1}^{8} c_{sk} V_k.$$

The source bank contains equal numbers of four source types.

For a relevant source, draw $\zeta_s \sim \mathcal{N}(0, 0.15^2 I_3)$ and set

$$(c_{s1}, c_{s2}, c_{s3}) \propto \left[(1.0, 0.7, 0.5)^\top + \zeta_s\right]_+, \quad c_{sk} = 0,\ k \geq 4,$$

followed by Euclidean normalization of c_s. For each redundant source, choose one relevant source as its parent. Let u_s be a unit vector orthogonal to the parent's coefficient vector within the active-factor subspace, and set

$$c_s = \rho_{\mathrm{red}} c_{\mathrm{par}(s)} + \sqrt{1 - \rho_{\mathrm{red}}^2}\, u_s. \tag{C.3}$$

Because the factor tangents are orthonormal before nuisance is added, ρ_{red} is the source-level tangent cosine induced by the latent coefficients, up to the source-specific nuisance term. The reference value is $\rho_{\mathrm{red}} = 0.6$. For an orthogonal source, draw a standard Gaussian vector supported only on factors $4, \dots, 8$ and normalize it to unit length. For an antagonistic source, let

$$a_{\mathrm{act}} = \frac{(1.0, 0.7, 0.5, 0, \dots, 0)^\top}{\|(1.0, 0.7, 0.5)\|_2}$$

and draw a unit vector v_s supported on factors $4, \dots, 8$. Set

$$c_s = -\rho_{\mathrm{ant}} a_{\mathrm{act}} + \sqrt{1 - \rho_{\mathrm{ant}}^2}\, v_s, \quad \rho_{\mathrm{ant}} = 0.7.$$

The first component gives negative orientation relative to reusable target descent, while the inactive-factor component prevents antagonistic sources from being exact sign reversals of one another.

Finally, every source receives a small nuisance parameter direction ξ_s. A raw rank-r_V Gaussian direction is projected, in tangent space, onto the orthogonal complement of $\{\phi_1, \dots, \phi_8\}$ and rescaled so that

$$\|P_C J_0 \xi_s\|_{\mathrm{geom}} = \tau_\xi, \quad \tau_\xi = 0.10.$$

The stored source update and checkpoint are

$$\boldsymbol{\Delta}_s = \Delta_s^{\mathrm{fac}} + \xi_s, \quad \boldsymbol{\theta}_s = \boldsymbol{\theta}_0 + \boldsymbol{\Delta}_s. \tag{C.4}$$

All source updates are generated relative to the same base checkpoint. No source labels or source training samples are provided to the transfer procedure. Table C.1 collects the complete reference Monte Carlo configuration.

Table C.1: **Monte Carlo configuration.** All quantities are fixed unless explicitly varied in Appendix C. Large construction and oracle samples belong to the simulation generator and are never available to a fitted method.

Quantity	Value	Quantity	Value	Quantity	Value
Input dimension	40	Classes C	5	Hidden width H	128
Latent factors	8	Active target factors	3	Factor rank r_V	4
Innovation width	8	β_{innov}	.6	τ_ξ	.10
Sources $S = M$	24	ρ_{red}	.6	ρ_{ant}	.7
n_{fit}	128	n_{val}	128	n_{test}	10,000
Cross-fit folds	5	Completeness budget q	3	Gate level δ	.05
$\kappa(D)$	100	N_{geom}	200,000	N_{oracle}	200,000
Ordinary replications	500	Gate replications	2,000	Reference source scale γ	1

Table C.2: **Reference simulation source bank.** The bank is heterogeneous so that source selection, redundancy control, sign handling, and innovation learning are separately identifiable. The same four-way composition is retained when S is varied.

Source type	Count	Active-factor overlap	Redundancy	Orientation
Relevant	6	High	Low-moderate	Positive
Redundant variants	6	High	ρ_{red}	Positive
Orthogonal	6	Zero by design	Low	Near zero
Antagonistic	6	Moderate	Low-moderate	Negative

The resulting 24-source reference bank is summarized in Table C.2. Its four source types separate positive target overlap, source redundancy, near-orthogonality, and antagonistic orientation so that these mechanisms can be evaluated independently. When $S \in \{8, 16, 24, 48, 96\}$ is varied, the four-way composition is preserved; the bank contains exactly $S/4$ sources of each source type.

To test coordinate dependence without changing any represented predictor, for a prescribed condition number $\kappa \geq 1$ define

$$d_j = \kappa^{-1/2+(j-1)/(127)}, \quad j = 1, \ldots, 128, \quad D_\kappa = \mathrm{diag}(d_1, \ldots, d_{128}).$$

Then $\kappa(D_\kappa) = \kappa$. Reparameterize every two-layer network by

$$W_1 \longmapsto D_\kappa W_1, \quad W_2 \longmapsto W_2 D_\kappa^{-1}. \tag{C.5}$$

ReLU homogeneity gives

$$W_2 D_\kappa^{-1} \sigma(D_\kappa W_1 x) = W_2 \sigma(W_1 x)$$

exactly. The transformation is applied to the common base and every source endpoint, after which

$$\tilde{\boldsymbol{\Delta}}_s = \psi_\kappa(\boldsymbol{\theta}_0 + \boldsymbol{\Delta}_s) - \psi_\kappa(\boldsymbol{\theta}_0).$$

Thus the experiment uses endpoint transport in Proposition 4.2 and does not redefine atomization after reparameterization. For the controlled invariance diagnostic only, define

$$\boldsymbol{\Delta}^\star_{\mathrm{supp}} = \sum_{k=1}^{8} \alpha_k^\star V_k.$$

The raw Euclidean score is

$$S_s^{\mathrm{Euc}} = \frac{(\boldsymbol{\Delta}_s^\top \boldsymbol{\Delta}^\star_{\mathrm{supp}})^2}{\|\boldsymbol{\Delta}_s\|_2^2 \|\boldsymbol{\Delta}^\star_{\mathrm{supp}}\|_2^2},$$

and the functional score is

$$S_s^{\mathrm{fun}} = \frac{b_s^2}{G_{ss} R_0}.$$

The oracle parameter direction is used only to diagnose coordinate dependence; it is never supplied to a fitted method. For either score,

$$\mathrm{Drift}(\kappa) = \frac{1}{S} \sum_{s=1}^{S} \left| S_s(\kappa) - S_s(1) \right|.$$

The rank column is Spearman correlation with realized single-source utility defined below. Within a replication, the same bank and target data are reused across all κ values.

Table C.3: **Function-preserving reparameterization stress.** Rank ρ is Spearman correlation with realized source utility and drift is the mean absolute score change relative to $\kappa(D) = 1$. The represented functions are exactly unchanged; small nonzero functional drift reflects finite endpoint linearization.

$\kappa(D)$	Euclidean rank ρ	Euclidean drift	Functional rank ρ	Functional drift
1	.89	.000	.88	.000
10	.69	.118	.88	.006
10^2	.44	.236	.87	.009
10^3	.22	.351	.87	.012
10^4	.11	.437	.86	.015

Table C.3 reports the resulting coordinate-stress experiment. Because (C.5) preserves every network function exactly, deterioration of the Euclidean score reflects coordinate dependence, not an actual

change in predictive behavior, whereas the functional score is expected to remain stable up to finite-endpoint linearization error. As Table C.3 shows, Euclidean ranking and score values deteriorate rapidly as the parameterization becomes ill-conditioned, whereas the target-functional ranking remains nearly unchanged. The small nonzero functional drift is consistent with Proposition 4.2, which gives first-order but not exact finite-endpoint invariance.

For any candidate parameter $\boldsymbol{\theta}$, define population cross-entropy risk

$$L_0(\boldsymbol{\theta}) = \mathbb{E}_X \left[-\sum_{c=1}^{5} \eta_c^\star(X) \log p_{\boldsymbol{\theta},c}(X) \right],$$

where $p_{\boldsymbol{\theta}}(x) = \mathrm{softmax}\{z_{\boldsymbol{\theta}}(x)\}$. All oracle risks and population geometric quantities are approximated on an independent oracle covariate sample of size $N_{\mathrm{oracle}} = 200{,}000$. Using teacher probabilities, not the sampled oracle labels, removes irreducible label Monte Carlo error. This sample is regenerated independently in every replication and is never available to fitted methods. For source s, realized single-source utility is

$$G_s^{\mathrm{real}} = L_0(\boldsymbol{\theta}_0) - \min_{\alpha \in [-2,2]} L_0(\boldsymbol{\theta}_0 + \alpha \boldsymbol{\Delta}_s). \tag{C.6}$$

The scalar minimization is initialized on a fixed dense grid and refined by bounded one-dimensional optimization to tolerance 10^{-8}.

For source subset A, define

$$L_0^\star(A) = \min_{\alpha_A \in [-2,2]^{|A|}} L_0 \left(\boldsymbol{\theta}_0 + \sum_{e \in A} \alpha_e \boldsymbol{\Delta}_e \right).$$

For the top-three comparison,

$$A_{3,\mathrm{risk}}^\star \in \arg \min_{A \subset [S],\, |A|=3} L_0^\star(A).$$

For the 24-source reference bank, all three-source subsets are enumerated. Within each subset, the bounded three-dimensional optimization is solved from a deterministic coarse grid followed by multistart L-BFGS-B; additional starts are added until the best risk changes by less than 10^{-8}. The top-three regret of an estimated three-source set $\hat{A}_3$ is

$$\mathrm{Regret}_3(\hat{A}_3) = L_0^\star(\hat{A}_3) - L_0^\star(A_{3,\mathrm{risk}}^\star). \tag{C.7}$$

Thus selected and oracle subsets are evaluated under the same source budget and coefficient box. The population geometric completeness target is computed separately. Let

$$\mathcal{A}_3^\star = \arg \max_{\substack{A \subset [S] \\ |A| \le 3}} \mathbf{b}_A^\top \mathbf{G}_{AA}^\dagger \mathbf{b}_A \text{ and } \chi_3 = \max_{\substack{A \subset [S] \\ |A| \le 3}} \frac{\mathbf{b}_A^\top \mathbf{G}_{AA}^\dagger \mathbf{b}_A}{R_0}.$$

If several supports attain projection energies within 10^{-8} of the maximum, all are retained as population-equivalent optimal supports.

C.2 Simulation Factors and Evaluation Protocol

The reference condition is summarized in Table C.1. We vary one factor at a time

$$n_{\mathrm{fit}} \in \{32, 64, 128, 256, 512\}, \quad S \in \{8, 16, 24, 48, 96\},$$

$$\beta_{\mathrm{innov}} \in \{0, 0.3, 0.6, 1.0\}, \quad \kappa(D) \in \{1, 10, 10^2, 10^3, 10^4\}, \quad \rho_{\mathrm{red}} \in \{0, 0.3, 0.6, 0.9\}.$$

Whenever n_{fit} varies, we set $n_{\mathrm{val}} = n_{\mathrm{fit}}$ and retain $n_{\mathrm{test}} = 10{,}000$. The $n_{\mathrm{fit}} = 512$ condition is an appendix-only extension beyond the target sizes displayed in the main text. Ordinary conditions use 500 independent Monte Carlo replications. Gate-failure probabilities use 2,000 replications. Except in the dedicated gate-calibration experiment, $\delta = 0.05$. Because this is a classification simulation, the bounded gate loss is 0-1 misclassification loss and $B = 1$. The default atomization has one atom per source checkpoint. GeoTransfer uses the five-fold nuisance cross-fit, the fixed diagnostic stabilization

$$(\rho_G^{\mathrm{diag}}, \lambda_G^{\mathrm{diag}}) = (0.25, 10^{-4}),$$

and the remaining candidate grids in Appendix B. Candidate construction uses only fit data, the risk gate uses only validation observations, and the test split is untouched until deployment.

The completeness error is

$$\mathrm{CompErr} = |\hat{\chi}_3 - \chi_3|.$$

For support recovery, let $\hat{A}_3$ be the three-atom set returned by the same forward-selection plus local-swap routine used to approximate $\hat{\chi}_3$. Define

$$\mathrm{F1}_{\mathrm{supp}} = \max_{A^\star \in \mathcal{A}_3^\star} \mathrm{F1}(\hat{A}_3, A^\star). \tag{C.8}$$

The maximum over population-equivalent optimal supports avoids penalizing a method for selecting a functionally equivalent redundant support. Test accuracy is the fraction of the independent test sample classified correctly. Negative transfer is the replication-level event

$$\mathrm{NT} = \mathbf{1}\left\{\mathrm{Acc}_{\mathrm{deploy}} < \mathrm{Acc}_{\mathrm{TO}}\right\},$$

and the reported NT rate is its Monte Carlo average. Gate failure is the population event

$$\mathbf{1}\left\{R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{deploy}}) > R_{\mathrm{gate}}(\hat{\boldsymbol{\theta}}_{\mathrm{TO}})\right\},$$

with both population gate risks evaluated on the oracle sample. Rank ρ is Spearman correlation with G_s^{real}, and top-three regret is (C.7). For a scalar replication statistic T_r, the Monte Carlo standard error is

$$\mathrm{MCSE}(\bar{T}) = \frac{\mathrm{sd}(T_1, \ldots, T_R)}{\sqrt{R}};$$

for an event frequency it is

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{R}}.$$

Table C.4 reports the one-factor-at-a-time sensitivity analysis under this protocol. The four blocks vary target sample size, source-bank size, target innovation, and source redundancy, respectively, while holding all remaining quantities at their reference values in Table C.1.

Table C.4: **Additional Simulation Sensitivity**

Condition	$\hat{\chi}_3$	Comp. error↓	Support F1↑	Acc.↑	NT↓	Gate failure
$n_{\mathrm{fit}} = 32$	.54	.112	.51	73.1	.031	.037
$n_{\mathrm{fit}} = 64$	.58	.076	.63	75.4	.019	.031
$n_{\mathrm{fit}} = 128$	.61	.048	.74	78.1	.008	.026
$n_{\mathrm{fit}} = 256$	.62	.031	.82	79.2	.006	.023
$n_{\mathrm{fit}} = 512$	.63	.020	.88	80.0	.004	.020
$S = 8$	.48	.070	.78	76.9	.011	.025
$S = 16$	.57	.055	.76	77.7	.009	.025
$S = 24$	.61	.048	.74	78.1	.008	.026
$S = 48$	.66	.051	.69	78.4	.010	.028
$S = 96$	.69	.060	.61	78.5	.013	.031
Innovation $= 0$	.91	.041	.77	81.3	.006	.022
Innovation $= .3$	.76	.044	.76	79.9	.007	.023
Innovation $= .6$	.61	.048	.74	78.1	.008	.026
Innovation $= 1.0$	.39	.052	.70	75.2	.009	.028
Redundancy $= .0$	.64	.043	.80	78.5	.006	.024
Redundancy $= .3$	.62	.046	.77	78.3	.007	.024
Redundancy $= .6$	.60	.050	.73	77.9	.009	.025
Redundancy $= .9$	.58	.061	.62	77.1	.014	.027

Table C.4 separates four distinct effects. The sample-size block isolates finite-target estimation; completeness error decreases and support recovery improves as n_{fit} grows. The source-bank block shows the tradeoff between greater potential source coverage and a harder finite-sample selection problem. Increasing β_{innov} lowers source-span completeness because more target structure lies outside the reusable span, whereas increasing ρ_{red} makes population-equivalent source supports harder to distinguish and consequently degrades support recovery.

C.3 Additional simulations and nonlinear stress tests

We compare the raw individual-score denominator, an in-sample conditional score fit, two-fold and five-fold cross-fitting, and the oracle conditional score. All other geometry and selection steps are held fixed. The five-fold version is the default used elsewhere in the paper. Table C.5 quantifies the resulting bias in $\hat{\chi}_3$ together with its consequences for source ranking and support recovery, which shows that using the observed individual-label score in the denominator produces persistent downward bias. Cross-fitting significantly reduces this bias, with the five-fold default approaching the oracle conditional-score benchmark while retaining a fully data-driven procedure.

Table C.5: **Normalization and nuisance-estimation ablation.** Bias is $\mathbb{E}(\hat{\chi}_3 - \chi_3)$; rank ρ is Spearman correlation with realized single-source utility. Rank and support F1 are evaluated at the reference $n_{\mathrm{fit}} = 128$. The raw individual-score denominator contains conditional label-noise energy and is shown only as a diagnostic ablation.

Normalization	$n_{\mathrm{fit}} = 64$ bias	$n_{\mathrm{fit}} = 128$ bias	$n_{\mathrm{fit}} = 256$ bias	Rank ρ	Support F1
Raw individual-score energy	-.146	-.129	-.118	.80	.66
In-sample conditional-score fit	-.061	-.047	-.033	.84	.70
2-fold cross-fit	-.028	-.019	-.013	.87	.73
5-fold cross-fit (default)	-.017	-.010	-.006	.88	.74
Oracle conditional score	-.004	-.002	-.001	.90	.77

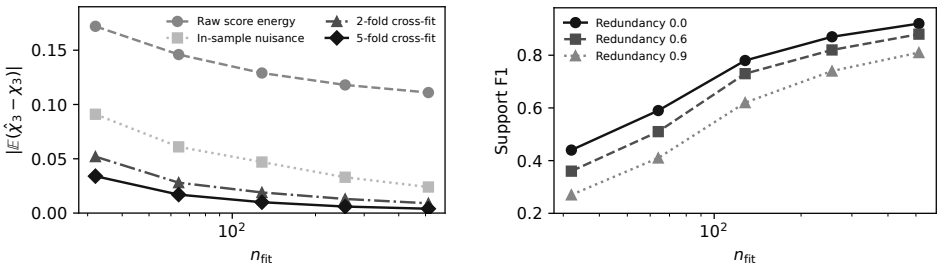


Figure C.1: **Finite-sample geometry diagnostics.** Left: cross-fitting removes the persistent label-noise bias of the naive descent-energy denominator. Right: three-atom support recovery improves with target sample size and degrades as source redundancy makes functionally similar supports harder to distinguish.

Figure C.1 visualizes the same two finite-sample mechanisms. Panel (a) traces normalization bias as n_{fit} increases, while Panel (b) shows how support recovery depends jointly on target sample size and source redundancy. The two panels complement Table C.5; cross-fitting removes the dominant normalization bias as information grows, whereas high source redundancy continues to make exact support identification more difficult even when the underlying transferable subspace is well estimated.

Multiply every stored source update by

$$\gamma \in \{0.25, 0.5, 1, 1.5, 2\}, \quad \boldsymbol{\Delta}_s^{(\gamma)} = \gamma \boldsymbol{\Delta}_s,$$

while holding the teacher distribution and target samples fixed. For an individual atom define

$$E_{\mathrm{nl}}(e;\gamma) = \frac{\left\| P_C\{z_{\boldsymbol{\theta}_0 + \boldsymbol{\Delta}_e^{(\gamma)}} - z_{\boldsymbol{\theta}_0}\} - P_C J_0 \boldsymbol{\Delta}_e^{(\gamma)} \right\|_{L_2(P_0^X)}}{\left\| P_C J_0 \boldsymbol{\Delta}_e^{(\gamma)} \right\|_{L_2(P_0^X)} + 10^{-8}}. \tag{C.9}$$

On fit inputs define

$$\|h\|_{\mathrm{fit}}^2 = \frac{1}{n_{\mathrm{fit}}} \sum_{i=1}^{n_{\mathrm{fit}}} \|P_C h(X_i)\|_2^2 \text{ and } \hat{E}_{\mathrm{nl}}(a) = \frac{\left\| z_{\boldsymbol{\theta}_0 + \sum_e a_e \boldsymbol{\Delta}_e} - z_{\boldsymbol{\theta}_0} - \sum_e a_e J_0 \boldsymbol{\Delta}_e \right\|_{\mathrm{fit}}}{\left\| \sum_e a_e J_0 \boldsymbol{\Delta}_e \right\|_{\mathrm{fit}} + 10^{-8}}. \tag{C.10}$$

The centering inside $\|\cdot\|_{\mathrm{fit}}$ makes this diagnostic insensitive to common logit shifts. The default GeoTransfer candidate still uses only the functional trust constraint in (3.6). In this stress experiment we additionally consider a diagnostic-screen variant rejecting a candidate when

$$\hat{E}_{\mathrm{nl}}(a) > \tau_{\mathrm{nl}}, \quad \tau_{\mathrm{nl}} = 0.25.$$

The threshold is fixed before evaluation. This screen is used only to study breakdown of the local approximation and is not part of Algorithm 1. Its numerator uses exact forward evaluations of the nonlinear predictor on fit inputs, not a finite-difference derivative.

Figure C.2 compares the deterioration of the tangent approximation with the resulting negative-transfer behavior as γ increases. It separately displays the unprotected procedure, the default functional trust-radius procedure, and the additional stress-only nonlinearity screen so that the role of the default safeguard is not conflated with the auxiliary diagnostic. As Figure C.2 shows, tangent error is small near the base and grows as the source endpoints are moved farther away. The default trust radius increasingly limits harmful extrapolation in this regime; the additional $\hat{E}_{\mathrm{nl}}$ screen provides a stronger stress-test diagnostic but is not required by the default GeoTransfer algorithm.

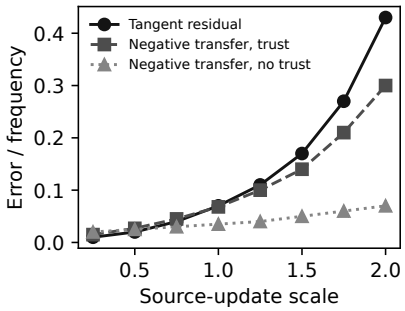


Figure C.2: **Breakdown of the tangent approximation under increasingly distant source checkpoints.** The tangent residual increases with source-update scale. The default functional trust radius limits negative transfer as the local approximation deteriorates, while the additional nonlinear screen used only in this stress experiment provides a diagnostic upper layer of protection. The nonlinear screen is not part of the default GeoTransfer procedure.

We rerun the reference simulation for $2{,}000$ independent replications at each requested failure level. All candidates are fitted before validation is inspected. The gate loss is 0-1 misclassification loss, so $B = 1$, and population gate risks used only for evaluation are approximated on the independent oracle sample. Table C.6 directly examines the finite-sample statement of Proposition 4.4 by comparing the requested failure level with the realized probability that the deployed predictor has greater population gate risk than target-only. It shows that empirical failure remains below the corresponding requested level across the reported settings, while candidate acceptance increases as the requested failure level is relaxed. The failure event is evaluated using population gate risk, not the finite test sample itself, matching Proposition 4.4. This experiment is therefore distinct from the ordinary negative-transfer frequency in Table C.4, which compares realized finite-test predictive performance.

Table C.6: **Finite-sample calibration of the risk gate.** Failure is the event that the deployed predictor has larger population misclassification risk than target-only. Candidate acceptance is the fraction of replications in which at least one transferred candidate passes the empirical-Bernstein gate; mean gain is reported only among accepted transferred deployments.

Requested δ	Empirical failure	Candidate acceptance	Mean gain when accepted
.01	.006	.41	.029
.025	.014	.53	.031
.05	.026	.64	.032
.10	.057	.76	.033

D Model Extensions

This part clarifies the scope of the target-functional geometry through extensions to regression, token-level losses, heterogeneous architectures, semi-supervised estimation, and nonlocal updates.

D.1 Regression and Reduction to Angle-Based Transfer

For squared-error regression, $f_{\boldsymbol{\theta}(x)\in\mathbb{R}}$ and

$$\ell(f, y) = \frac{1}{2}(y - f)^2.$$

Let $m_0(x) = \mathbb{E}(Y \mid X = x)$. At the base predictor $f_0 = f_{\boldsymbol{\theta}_0}$,

$$W_0(x) = 1, \quad g_0(x) = f_0(x) - m_0(x), \quad r_0(x) = m_0(x) - f_0(x).$$

Consequently the target-functional inner product reduces to the ordinary $L_2(P_0^X)$ inner product,

$$\langle u, v \rangle_0 = \mathbb{E}\{u(X)v(X)\} \text{ and } \langle \phi_e, \phi_{e'} \rangle_0 = \mathbb{E}\{\phi_e(X)\phi_{e'}(X)\}.$$

In this case the quadratic expansion of Theorem 4.1 is exact as a statement about output perturbations. Indeed, for every square-integrable perturbation h,

$$\begin{aligned} \mathcal{T}(h) &= \mathbb{E}\left[\frac{1}{2}\{Y - f_0(X) - h(X)\}^2 - \frac{1}{2}\{Y - f_0(X)\}^2\right] \\ &= -\mathbb{E}\{h(X)r_0(X)\} + \frac{1}{2}\mathbb{E}\{h(X)^2\} = -\langle h, r_0 \rangle_0 + \frac{1}{2}\|h\|_0^2. \end{aligned} \tag{D.1}$$

Hence there is no third-order risk remainder under squared loss, and the unique function minimizing (D.1) over any finite source span $\mathcal{V}_A$ is exactly $h_A^\star = \mathrm{Proj}_{\mathcal{V}_A} r_0$. The distinction between functional and parameter linearization remains; for a nonlinear predictor, replacing the exact finite parameter displacement $f_{\boldsymbol{\theta}_0+\boldsymbol{\Delta}} - f_{\boldsymbol{\theta}_0}$ by the tangent $J_0\boldsymbol{\Delta}$ may still incur model-linearization error. Squared loss removes the loss-expansion remainder, not the nonlinear-predictor remainder.

The connection with parameter-angle transfer is especially transparent in linear regression. Let $f_{\boldsymbol{\beta}(x)} = x^\top \boldsymbol{\beta}$, $\mathbb{E}(Y \mid X = x) = x^\top \boldsymbol{\beta}^\star$, and let $\boldsymbol{\beta}_0$ be the common base. We write $\delta^\star = \boldsymbol{\beta}^\star - \boldsymbol{\beta}_0$ for the target descent coefficient and let $w = \boldsymbol{\beta}_s - \boldsymbol{\beta}_0$ be a source displacement. Then

$$\phi_w(x) = x^\top w \text{ and } r_0(x) = x^\top \delta^\star.$$

Writing $\Sigma_X = \mathbb{E}(XX^\top)$, the one-source principal transfer angle satisfies

$$\cos^2 \Theta_w = \frac{(w^\top \Sigma_X \delta^\star)^2}{(w^\top \Sigma_X w)(\delta^{\star\top} \Sigma_X \delta^\star)}. \tag{D.2}$$

The corresponding directed alignment is

$$A_w^{\mathrm{dir}} = \frac{w^\top \Sigma_X \delta^\star}{\sqrt{w^\top \Sigma_X w}\sqrt{\delta^{\star\top} \Sigma_X \delta^\star}}. \tag{D.3}$$

Thus the principal angle is orientation free, whereas A_w^{dir} retains the sign of source-target concordance, exactly as in Section 2. If $\Sigma_X = I$, (D.2) is the squared Euclidean angle between the source and target coefficient directions, and (D.3) is their signed Euclidean cosine. For anisotropic target covariates, the same quantities are measured in the prediction geometry induced by Σ_X. In a nonlinear model, $J_0(X)$ replaces the linear design vector and $W_0(X)$ additionally weights functional changes by target loss curvature. This calculation explains the relation to the angle-based transfer procedure discussed in Section 1. That approach uses directional concordance between source and target regression coefficients to regularize a high-dimensional target estimator. GeoTransfer instead treats one or more fitted source updates as a functional dictionary and projects the target descent field onto their induced span. The relationship is therefore a reduction of the geometry in the linear-regression special case, not an algorithmic equivalence between the two procedures.

D.2 Token-Level and Structured-Output Losses

The same construction applies to autoregressive or other token-level losses. Consider a target observation consisting of an input X and an output sequence $Y_{1:T} = (Y_1, \ldots, Y_T)$. At position t, let $H_t = (X, Y_{<t})$ denote the teacher-forced history and let $z_{\boldsymbol{\theta},t}(H_t) \in \mathbb{R}^V$ be the vocabulary logits, where V is the target vocabulary size. A general weighted sequence loss can be written

$$\ell_{\mathrm{seq}}(\boldsymbol{\theta}; X, Y) = \sum_{t=1}^{T} \omega_t \left[-\log p_{\boldsymbol{\theta},t,Y_t}(H_t)\right], \tag{D.4}$$

where $\omega_t \geq 0$ is fixed by the task's masking and normalization convention. Taking $\omega_t = 1$ gives a summed sequence loss, whereas normalizing the weights gives a length-normalized loss.

Let $p_{0,t}(H_t) = \mathrm{softmax}\{z_{\boldsymbol{\theta}_0,t}(H_t)\}$, and let $\eta_{0,t}(H_t) = P_0(Y_t = \cdot \mid H_t)$ be the conditional target next-token law. On the centered vocabulary-logit space, $P_V = I_V - \frac{1}{V}\mathbf{1}\mathbf{1}^\top$, the conditional score and curvature are

$$g_{0,t}(H_t) = p_{0,t}(H_t) - \eta_{0,t}(H_t) \text{ and } W_{0,t}(H_t) = \mathrm{diag}\{p_{0,t}(H_t)\} - p_{0,t}(H_t)p_{0,t}(H_t)^\top.$$

For source atom e, define the token-level tangent field $\phi_{e,t}(H_t) = P_V J_{0,t}(H_t)\boldsymbol{\Delta}_e$. The sequence-level target-functional inner product is then

$$\langle u, v\rangle_0 = \mathbb{E}_{P_0}\left[\sum_{t=1}^{T} \omega_t u_t(H_t)^\top W_{0,t}(H_t) v_t(H_t)\right]. \tag{D.5}$$

Accordingly, $r_{0,t}(H_t) = -W_{0,t}(H_t)^\dagger g_{0,t}(H_t)$, and the finite source-span representation becomes

$$G_{ee'} = \mathbb{E}\left[\sum_{t=1}^{T} \omega_t \phi_{e,t}(H_t)^\top W_{0,t}(H_t)\phi_{e',t}(H_t)\right], \tag{D.6}$$

$$b_e = -\mathbb{E}\left[\sum_{t=1}^{T} \omega_t \phi_{e,t}(H_t)^\top g_{0,t}(H_t)\right]. \tag{D.7}$$

Thus the token positions form a weighted direct-sum functional space, and all projection identities in Section 2 continue to hold after replacing the single-output inner product by (D.5). The large vocabulary dimension affects computation, not directly the definition of the geometry; Appendix B describes the randomized output sketch used to approximate (D.6). This formulation covers the language settings used in the experiments. For multiple-choice tasks, the loss is accumulated over the candidate answer tokens; for label-token classification it is accumulated over the prescribed label-token representation; and for short-form generation it is the usual teacher-forced sequence negative log likelihood. All such choices specify the weights and masks in (D.4) before target validation is inspected. The loss used to construct the geometry need not equal the bounded loss used by the deployment gate. Token negative log likelihood is unbounded, whereas Proposition 4.4 requires a prespecified bounded gate loss. Accordingly, the language implementation uses the bounded task-level gate losses specified in Appendix B, such as classification or exact-match error and $1-\mathrm{F1}$ for per-example F1. The projection geometry and the deployment guarantee therefore address different but explicitly specified risks.

D.3 Heterogeneous Architectures and Semi-Supervised Geometry

The deployment procedure in the main paper assumes a common transferable parameterization; every stored source atom $\boldsymbol{\Delta}_e$ is expressed relative to the same base coordinates so that $\boldsymbol{\theta}_0 + \sum_e a_e \boldsymbol{\Delta}_e$ is a valid target model. The geometric diagnostic itself requires less. Suppose a source model with a different architecture can be mapped to a target-compatible functional field $\phi_e(x) \in \mathbb{R}_0^C$ on the same target output coordinates. Once such a field is fixed,

$$G_{ee'} = \langle \phi_e, \phi_{e'}\rangle_0, \quad b_e = \langle \phi_e, r_0\rangle_0$$

and the principal transfer angles, directed alignments, and completeness indices remain well defined.

Such a functional field could in principle be constructed by distillation, a fixed logit transport, or a representation map into the target output space. This observation does not by itself extend the deployment theorem. A favorable transported function must still be realized by a parameter update in the target architecture, requiring an inverse problem of the form

$$u_e^\star \in \arg\min_u \mathbb{E}\,\|P_C J_0(X)u - \phi_e(X)\|_2^2$$

or a nonlinear analogue. Approximation error from this realization would enter both the transferred predictor and its risk analysis. Moreover, if a transport is learned from target data, it must be estimated entirely inside the fit sample and the theory must condition on, or explicitly analyze, that additional estimation step. We therefore keep heterogeneous architectures outside the claims of the present algorithm and avoid implicitly assuming an unrestricted cross-architecture transport.

For multiclass cross-entropy,

$$W_0(x) = \mathrm{diag}\{p_0(x)\} - p_0(x)p_0(x)^\top$$

depends only on the base prediction and the target covariate. Likewise, $\phi_e(x) = P_C J_0(x)\mathbf{\Delta}_e$ requires no target label. The Gram matrix can therefore use additional unlabeled target inputs even though the source-target signal still requires labeled observations.

Let $\{(X_i^\ell, Y_i^\ell)\}_{i=1}^{n_\ell} \overset{\text{iid}}{\sim} P_0$ be labeled target observations and let $\{X_j^u\}_{j=1}^{n_u} \overset{\text{iid}}{\sim} P_0^X$ be an independent unlabeled sample from the same target covariate law. Define

$$\hat{G}_{ee'} = \frac{1}{n_\ell + n_u}\left[\sum_{i=1}^{n_\ell} \phi_e(X_i^\ell)^\top W_0(X_i^\ell)\phi_{e'}(X_i^\ell) + \sum_{j=1}^{n_u} \phi_e(X_j^u)^\top W_0(X_j^u)\phi_{e'}(X_j^u)\right], \tag{D.8}$$

while

$$\hat{b}_e = -\frac{1}{n_\ell}\sum_{i=1}^{n_\ell} \phi_e(X_i^\ell)^\top g(X_i^\ell, Y_i^\ell). \tag{D.9}$$

Under the same sub-Gaussian and sub-exponential envelope conditions used in Appendix A, union bounds give

$$\|\hat{\mathbf{G}} - \mathbf{G}\|_{\max} = O_p\left\{\sqrt{\frac{\log M}{n_\ell + n_u}} + \frac{\log M}{n_\ell + n_u}\right\}, \tag{D.10}$$

$$\|\hat{\mathbf{b}} - \mathbf{b}\|_\infty = O_p\left\{\sqrt{\frac{\log M}{n_\ell}} + \frac{\log M}{n_\ell}\right\}. \tag{D.11}$$

Thus unlabeled target covariates improve estimation of source-source functional geometry, whereas estimation of target orientation remains label limited. This distinction also applies to the normalized diagnostics and the deployment gate. The cross-fitted descent-energy denominator $\hat{R}_{\text{cf}}$ requires estimation of the conditional target score and therefore remains governed by the labeled sample unless additional structure is imposed. The held-out risk gate likewise requires labeled validation observations. Hence the semi-supervised extension improves curvature and Gram estimation; it does not remove the labeled-data requirement for $\mathbf{b}$, normalized target-descent diagnostics, or safe deployment. If the unlabeled covariate distribution differs from P_0^X, (D.8) estimates a different geometry unless an appropriate covariate-shift correction is introduced.

D.4 Beyond the Local Additive Model

The main method uses fixed stored parameter atoms. For such atoms the parameter updates themselves commute

$$\boldsymbol{\theta}_0 + \alpha_e\mathbf{\Delta}_e + \alpha_{e'}\mathbf{\Delta}_{e'} = \boldsymbol{\theta}_0 + \alpha_{e'}\mathbf{\Delta}_{e'} + \alpha_e\mathbf{\Delta}_e.$$

Thus there is no order effect in the additive parameter model used by GeoTransfer. Nonlocal behavior can nevertheless arise because the predictor is nonlinear in its parameters. For $h = \sum_e \alpha_e\mathbf{\Delta}_e$, a second-order expansion gives

$$z_{\boldsymbol{\theta}_0+h}(x) - z_{\boldsymbol{\theta}_0}(x) = J_0(x)h + \frac{1}{2}D^2 z_{\boldsymbol{\theta}_0}(x)[h,h] + O(\|h\|_2^3). \tag{D.12}$$

Expanding the quadratic term reveals both individual and pairwise interactions,

$$D^2 z_{\boldsymbol{\theta}_0}[h,h] = \sum_e \alpha_e^2 D^2 z_{\boldsymbol{\theta}_0}[\mathbf{\Delta}_e, \mathbf{\Delta}_e] + 2\sum_{e<e'} \alpha_e\alpha_{e'} D^2 z_{\boldsymbol{\theta}_0}[\mathbf{\Delta}_e, \mathbf{\Delta}_{e'}].$$

These terms are absent from the first-order dictionary $\{\phi_e\}$ and explain why a combination of individually reasonable source directions can cease to be well approximated when its total finite-step displacement becomes large. The nonlinear diagnostic $\hat{E}_{\text{nl}}$ in Appendix C.3 measures the aggregate effect of precisely this tangent remainder. A richer nonlocal geometry could augment the dictionary with second-order functional atoms such as

$$q_{ee'}(x) = P_C D^2 z_{\boldsymbol{\theta}_0}(x)[\mathbf{\Delta}_e, \mathbf{\Delta}_{e'}],$$

at the cost of a quadratic increase in dictionary size and substantially more expensive derivative computation. We do not estimate these terms in the present method.

A genuinely different extension arises when the update associated with an atom is itself state dependent. Let

$$F_e(\boldsymbol{\theta}) = \boldsymbol{\theta} + \alpha_e \boldsymbol{\Delta}_e(\boldsymbol{\theta}),$$

where $\boldsymbol{\Delta}_e(\boldsymbol{\theta})$ is a differentiable vector field, not simply a fixed stored vector. For two such updates,

$$\begin{aligned} F_{e'}\{F_e(\boldsymbol{\theta})\} = \boldsymbol{\theta} &+ \alpha_e \boldsymbol{\Delta}_e(\boldsymbol{\theta}) + \alpha_{e'} \boldsymbol{\Delta}_{e'}(\boldsymbol{\theta}) \\ &+ \alpha_e \alpha_{e'} D\boldsymbol{\Delta}_{e'}(\boldsymbol{\theta}) \boldsymbol{\Delta}_e(\boldsymbol{\theta}) + O\big(|\alpha_e|^2 |\alpha_{e'}| + |\alpha_e||\alpha_{e'}|^2\big), \end{aligned} \tag{D.13}$$

whereas reversing the order replaces the cross term by $\alpha_e \alpha_{e'} D\boldsymbol{\Delta}_e(\boldsymbol{\theta}) \boldsymbol{\Delta}_{e'}(\boldsymbol{\theta})$. Consequently

$$F_{e'}\{F_e(\boldsymbol{\theta})\} - F_e\{F_{e'}(\boldsymbol{\theta})\} = \alpha_e \alpha_{e'} [\boldsymbol{\Delta}_e, \boldsymbol{\Delta}_{e'}](\boldsymbol{\theta}) + O\left(|\alpha_e|^2 |\alpha_{e'}| + |\alpha_e||\alpha_{e'}|^2\right), \tag{D.14}$$

with $[\boldsymbol{\Delta}_e, \boldsymbol{\Delta}_{e'}] = D\boldsymbol{\Delta}_{e'}\, \boldsymbol{\Delta}_e - D\boldsymbol{\Delta}_e\, \boldsymbol{\Delta}_{e'}$. Such commutator terms describe order dependence of sequential, state-dependent adaptation operators. They are not present in the fixed additive task-vector model analyzed in the paper and are not estimated by Algorithm 1. Extending the transfer geometry to second-order functional interactions or noncommuting state-dependent operators is a natural direction for genuinely nonlocal transfer.

E Implementation and Reproducibility Details

This part specifies the empirical source libraries, target-data protocols, baseline information sets, and comparison rules used in Section 5. Every source library is defined relative to a common base parameterization, source training examples are unavailable at transfer time, and test observations are not used for source ranking, coefficient estimation, hyperparameter selection, or deployment gating.

Table E.1: **Empirical benchmark protocol.** Every row is a separate common-coordinate source library. Target-label counts refer to observations available for geometry construction and target adaptation; validation observations are disjoint and are used only for candidate selection and the risk gate. Boldface denotes the reference target-data condition reported in Table 2.

Suite	Base / architecture	Source unit	Held-out target	Target labels	Primary metric
CLIP-8	CLIP ViT-B/32	task fine-tunes	one of 8 tasks	4/16/32/64 per class	accuracy
CLIP-14	CLIP ViT-L/14	task fine-tunes	one of 14 tasks	4/16/32/64 per class	accuracy
PACS	shared ViT backbone	domain fine-tunes	one domain	16/64 per class	accuracy
OfficeHome	shared ViT backbone	domain fine-tunes	one domain	16/64 per class	accuracy
DomainNet	shared ViT backbone	domain fine-tunes	one domain	16/64 per class	accuracy
Language	shared instruction backbone	LoRA/adapters	one task	32/64/128/256	normalized score

E.1 Vision Model Libraries and Cross-Domain Protocols

The CLIP task libraries use a common pretrained checkpoint within each reported suite (Radford et al., 2021). CLIP-8 uses a ViT-B/32 base and contains eight independently fine-tuned source-task checkpoints; CLIP-14 uses a ViT-L/14 base and contains fourteen. For each target task, its own fine-tuned checkpoint is removed before the remaining checkpoints form the source bank. The experiment therefore measures cross-task transfer, not simple retrieval of a stored target expert. No parameter-space operation is performed across the ViT-B/32 and ViT-L/14 libraries. Target geometry uses stratified $k \in \{4, 16, 32, 64\}$ labeled fit examples per class. The 32-shot condition is the reference condition used for the primary empirical comparison in Table 2; the remaining shot levels are used for target-data sensitivity analyses. When enough labeled training observations remain, the validation split has the same nominal size and is disjoint from the fit split. Otherwise the available labeled target sample is partitioned once into 60% fit and 40% validation observations. The fit split constructs the geometry and all fitted candidates; validation data tune the finite candidate family and apply the risk gate. The held-out test set is accessed only after deployment has been fixed. Source training examples are never loaded. The default vision atomization used for all predictive transfer results partitions each source checkpoint into three prespecified transferable groups: patch embedding plus early transformer blocks, middle transformer blocks, and late transformer blocks. The projection or classification head remains target specific. This three-group partition is fixed before target labels are inspected and is used for the results in Table 2 and Table F.2. For diagnostic visualization only, we additionally refine the transformer backbone into six prespecified block groups, denoted $B_1, \ldots, B_6$. This six-group decomposition is used for the directed alignment atlas in Figure 3 and the block-stability analysis in Appendix F.2; it does not replace the three-group atomi-

zation used by the default predictive procedure and is not selected by target validation.

PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019) are evaluated as separate common-coordinate domain libraries using the DomainBed-style protocol in Section 5. Within each benchmark, one domain is held out as target and source checkpoints are obtained from the remaining domains using a common base and label space. The target-only comparator receives exactly the same labeled target fit observations as GeoTransfer and the same parameter-efficient adaptation capacity as the innovation component. Whole-source selection chooses one stored source checkpoint using permitted target tuning data. Sparse target-loss weighting uses the same stored source task vectors but estimates their coefficients directly from target fit loss, without reliance on the functional Gram projection, thereby controlling for the benefit of continuous source weighting. Table E.1 summarizes the empirical libraries and target-data budgets. Each row is a separate source library; no parameter-space operation crosses architectures or benchmark-specific coordinate systems.

E.2 Language Adapters and Competing Methods

The language experiments use a common instruction-tuned transformer and a library of frozen LoRA or adapter source experts. Each source expert is trained on one source task family, the target task is excluded from source training, and only the parameter update relative to the common backbone is retained. The innovation adapter and the target-only comparator use the same adapter family, rank, initialization rule, trainable-parameter budget, optimizer family, and stopping rule. The transfer geometry is estimated from token-level negative log likelihood on the target fit split using the sequence construction in Appendix D.2. The output-space Gram is sketched to $r_{\mathrm{sk}} = 64$ by default, as specified in Appendix B; the unsketched geometry is used when the centered output dimension is small enough to store directly. For multiple-choice tasks, the training loss is accumulated over answer strings; for classification tasks, we use normalized label-token likelihood; and for short-form QA, we optimize the standard teacher-forced sequence loss and evaluate exact match or F1. The reported language score is the unweighted mean of the prespecified task-normalized metrics.

The geometry loss and the bounded deployment loss are separated. For classification and multiple-choice tasks, the gate uses 0-1 prediction error. For short-form QA evaluated by per-example F1, it uses $1 - \mathrm{F1}$; when exact match is the task metric, the corresponding 0-1 exact-match error is used. In all cases the gate loss is bounded by one and is fixed before validation, matching the conditions of Proposition 4.4. H-score, LEEP, LogME, and PACTran are used as source-ranking baselines and are computed from permitted target examples according to their standard transferability-estimation definitions. When a score requires source classifier outputs that are not compatible with the target label space, we use its representation-based form with a target linear probe and avoid introducing an artificial source-to-target label map. These scores rank sources only; subsequent target adaptation is evaluated under the same target-label budget as GeoTransfer.

Table E.2: **Information used by empirical baselines.** "Fit" and "validation" refer to the target splits in Table E.1. No method receives source training examples in the source-data-free comparisons.

Method	Source data	Source checkpoints	Target labels	Target derivative information	Test-time routing
H-score / LEEP / LogME / PACTran	No	Yes	Fit	No	No
Task arithmetic / TIES	No	Yes	Validation	No	No
Fisher-weighted merging	No	Yes	Fit	Fisher	No
Gradient matching	No	Yes	Fit	Gradient	No
TATR	No	Yes	Fit	Gradient / mask	No
TSV-Merge	No	Yes	Fit or validation	No	No
MASS-style adaptive subspace	No	Yes	Target support	Method specific	Yes
Sparse target-loss weights	No	Yes	Fit / validation	Gradient	No
GeoTransfer	No	Yes	Fit / validation	Fisher + score	No

The model-merging and task-vector baselines are subject to the same source-data-free and target-data restrictions. Task arithmetic uses a single mixing coefficient chosen on target validation data. TIES resolves parameter-sign conflicts before target-selected scaling. Fisher-weighted merging uses target Fisher information when source Fisher matrices are unavailable, so the baseline does not receive source training data. Gradient matching uses target gradients to score or weight source updates. TATR applies its trust-region masking before target-selected aggregation, and TSV-Merge decomposes source task updates into singular directions before combination. MASS-style adaptive subspace procedures are included only in benchmarks for which their required target support construction is compatible with the common source-data-free protocol. Sparse target-loss weighting directly opti-

mizes continuous source coefficients on target fit loss and uses the same validation restrictions as the other fitted competitors. Table E.2 records the information supplied to the implementations used in this paper. This comparison makes explicit which methods use target labels, target derivative information, or test-time routing. It also ensures that differences in empirical performance are not attributable to unreported access to source data or additional target information.

E.3 Optimization Fairness, Metrics, and Reporting

All methods within a target replicate use the same base checkpoint, target fit and validation observations, preprocessing, and held-out test set. Whenever a method performs target adaptation, its maximum number of target-gradient updates is matched to the corresponding capacity-matched comparator unless the published procedure requires a different optimization primitive. The target-only and innovation adapters use identical architecture, rank, initialization rule, optimizer family, and stopping rule. Additional scalar source coefficients introduced by a competitor are reported separately from the adapter parameter budget and are never silently treated as free adaptation capacity.

GeoTransfer's geometry construction, stabilization, sparse projection, innovation fitting, and finite validation family are exactly those in Appendix B. In particular, normalized geometry diagnostics use the fixed pre-validation stabilization pair $(\rho_G^{\mathrm{diag}}, \lambda_G^{\mathrm{diag}}) = (0.25, 10^{-4})$, which is distinct from candidate-specific stabilization in the sparse projection. With the default signed coefficient policy and fixed adapter weight penalty $\lambda_\omega = 10^{-4}$, the Cartesian candidate family contains 972 transferred hyperparameter configurations before fit-time duplicates or degenerate solutions are removed. The family is fixed before validation. The nonnegative coefficient policy is a separately declared sensitivity analysis and is not chosen by validation in the default experiments. The comparison protocol is summarized in Table E.3. It separates fit-time estimation, validation-time selection, and test-time evaluation so that neither GeoTransfer nor any fitted baseline receives test information during construction. The target-only comparator is completely fitted on the fit split before validation is inspected; its validation predictions are used only in the paired gate comparison.

Table E.3: **Optimization and comparison protocol.** The same target split is used across methods within each replicate. GeoTransfer-specific candidate grids and tolerances are in Appendix B.

Component	Fit data	Validation data	Test data before deployment	Capacity / selection rule
Transferability scores	Yes	No	No	source ranking only
Task arithmetic / TIES	No	Yes	No	one validation-selected scale
Gradient / sparse source weighting	Yes	Yes	No	source coefficients only
GeoTransfer geometry and sparse projection	Yes	No	No	fixed candidate construction
GeoTransfer innovation adapter	Yes	No	No	matched to target-only adapter
GeoTransfer candidate selection and gate	No	Yes	No	finite predeclared family
Target-only comparator	Yes	Evaluation only	No	matched adapter capacity
Final evaluation	No	No	Yes	predictor fixed before access

For geometry construction, the empirical source-target signal uses one reverse-mode pass through the fit data followed by inner products with the stored source updates, while exact Gram construction uses one JVP per retained source atom and fit batch. The randomized output sketch reduces the stored output dimension in the language experiments. Appendix B gives the resulting arithmetic and memory complexity. These geometry-construction operations are performed offline before the final transferred parameter update is assembled; the deployed GeoTransfer model requires no additional source model and no test-time router. The primary source-ranking metric is Spearman correlation between the estimated source score and realized held-out single-source transfer gain. Top-k regret compares the selected source set with an oracle set under the same source budget, and a harmful source is one whose independently evaluated single-source gain is negative. Ranking experiments treat each source checkpoint as one atom; the multi-source transfer experiments use the prespecified block or adapter atomization above. Prediction performance is reported as benchmark accuracy, F1, exact match, or the prespecified normalized language score. Paired negative-transfer frequency is the fraction of matched target replicates in which a method underperforms the capacity-matched target-only comparator. Fallback frequency is the fraction for which the validation gate deploys target only. Additional geometric diagnostics reported where relevant include $\hat{\chi}_q$ and the selected source count.

F Additional Empirical Results

This part provides additional empirical analyses supporting the main results in Section 5. We examine

sensitivity to coefficient-sign constraints, ablations of the proposed functional geometry, robustness to contaminated source libraries, stability of selected source blocks, and computational scaling. Unless otherwise stated, scalar empirical diagnostics in this appendix use the CLIP-8 suite. Each of its eight tasks is held out in turn as the target, and results are macro-averaged over held-out target tasks and independent seeds, giving each target equal weight. The fit, validation, and test construction follows Appendix E. Table F.3 varies the target fit budget explicitly; except as otherwise specified, the remaining scalar CLIP-8 ablations use the 32-shot reference condition reported in Table 2.

F.1 Source Selection, Sign Policy, and Few-Shot Transfer

The default geometry treats each source atom as a line and therefore permits signed coefficients. Some applications instead require literal forward reuse of a stored source update. We evaluate that restriction by intersecting the feasible set in (3.6) with $a_e \geq 0$ while leaving every other tuning rule unchanged. For the positive bank, atoms with negative fit-sample directed alignment are removed before selection; the mixed bank retains the original source library. All rows in Table F.1 use the innovation adapter, and the final row additionally applies the held-out validation gate.

Table F.1: **Sensitivity to the coefficient-sign policy.** "Positive bank" removes atoms with negative directed alignment, whereas "mixed bank" retains them. The nonnegative variant cannot reverse a stored source atom.

Source bank	Method	CLIP-14 Acc.	OfficeHome Acc.	Language score
Positive bank	Signed GeoTransfer + innovation	76.3	72.7	64.1
	Nonnegative GeoTransfer + innovation	76.2	72.6	64.0
Mixed bank	Signed GeoTransfer + innovation	76.4	72.8	64.2
	Nonnegative GeoTransfer + innovation	75.8	72.1	63.6
	Nonnegative GeoTransfer + innovation + gate	75.9	72.2	63.7

Table F.1 shows that the two coefficient policies are nearly identical when the retained bank is forward aligned. The signed formulation gains in the mixed bank because useful target descent can be stored in the reverse orientation; the nonnegative formulation must instead discard such directions. This is exactly the distinction between the orientation-free principal-angle utility in (2.4) and the directed diagnostic in (2.7). Table F.2 expands Panel B of Table 2 by reporting additional merging, gradient-based, and adaptive-subspace comparators across the same six empirical library regimes. It therefore uses the same benchmark columns as the main table and does not introduce a different aggregation in the appendix. The expanded comparison preserves the ordering in the main table while showing that the advantage is not created by omission of intermediate merging or gradient-based baselines. The innovation component improves the capacity-matched transferred predictor, whereas the gate trades a small amount of average performance for a substantial reduction in harmful deployments.

Table F.2: **Expanded target-performance comparison for the empirical libraries reported in Table 2.** All methods use the matched target-data and adaptation protocol in Appendix E. NT is paired negative-transfer frequency relative to the capacity-matched target-only comparator.

Method	CLIP-8 Acc.	CLIP-14 Acc.	PACS Acc.	OfficeHome Acc.	DomainNet Acc.	Lang. score	NT↓
Target-only adapter	76.8	73.4	86.0	69.7	52.4	61.5	–
Best-average source	75.9	72.1	85.1	68.4	51.4	60.7	.309
Task arithmetic	77.1	73.8	86.5	70.0	52.7	61.8	.219
Fisher merging	77.4	74.0	86.8	70.2	52.9	62.0	.192
TIES	77.8	74.3	87.0	70.6	53.2	62.2	.171
Gradient matching	78.0	74.5	87.2	70.7	53.5	62.4	.146
TATR	78.2	74.7	87.3	70.9	53.7	62.5	.130
TSV-Merge	78.4	74.9	87.5	71.0	53.9	62.6	.121
MASS-style adaptive subspace	78.3	74.8	87.4	71.0	54.0	62.6	.118
Sparse target-loss weights	78.6	75.0	87.6	71.2	54.1	62.8	.106
GeoTransfer projection	79.2	75.8	88.2	72.0	54.7	63.5	.060
GeoTransfer + innovation	79.8	76.4	88.8	72.8	55.3	64.2	.053
GeoTransfer + innovation + gate	79.6	76.2	88.6	72.6	55.1	64.0	.019

Table F.3 reports target-data scaling on the CLIP-8 suite. The 32-shot reference result is reported in Table 2 and is omitted here to emphasize scaling away from the reference condition. At each shot level, accuracy is first averaged over independent seeds within each held-out target task and then macro-averaged over the eight targets, giving every target equal weight. NT is computed over the corresponding matched target-seed replicates. "Full target" uses all target training observations permitted by the benchmark split while preserving the disjoint validation and held-out test sets. The

gain from source reuse is largest in the lowest-data regime and narrows as the target-only adapter receives enough labels to estimate its own adaptation. The same pattern appears in NT; additional target information reduces the cost of imperfect source selection, while the validation gate maintains the lowest harmful-deployment frequency.

Table F.3: **Few-shot target scaling on CLIP-8.** Accuracy is macro-averaged over the eight held-out target tasks with equal target weight; NT is the paired negative-transfer frequency over the same matched target-seed replicates. "Full target" uses all permitted target training observations while preserving the disjoint validation and test sets.

	4-shot		16-shot		64-shot		Full target	
Method	Acc.	NT	Acc.	NT	Acc.	NT	Acc.	NT
Target only	68.2	–	74.6	–	79.8	–	84.1	–
Task arithmetic	69.7	.24	75.4	.20	80.2	.17	84.2	.14
LogME-selected source	70.4	.19	76.1	.16	80.5	.13	84.3	.12
Sparse target-loss weights	71.3	.15	76.9	.12	81.0	.10	84.5	.08
GeoTransfer projection	73.1	.08	78.4	.06	82.0	.05	84.9	.04
GeoTransfer + innovation	74.0	.07	79.2	.05	82.8	.04	85.4	.03
GeoTransfer + innovation + gate	73.8	.02	79.0	.02	82.6	.02	85.3	.01

To isolate the source-target geometry itself, all entries in Table F.4 use the reference CLIP-8 condition from Table 2. The downstream adaptation and innovation specification is held fixed and the deployment gate is disabled; only the geometry used for source ranking and projection is changed. Moving from raw parameter angles to target-weighted functional geometry improves source ranking and utility calibration; retaining off-diagonal Gram structure improves multi-source selection; and cross-fitted normalization further improves calibration of the normalized geometry. Reparameterization robustness is evaluated separately in the controlled experiment of Section 5.1 and Figure 2, where the function-preserving transformation is known exactly.

Table F.4: **Ablation of the target-functional geometry on CLIP-8.** Results are macro-averaged over the eight held-out CLIP-8 target tasks under the 32-shot reference condition. Calibration error is the absolute difference between predicted local utility and realized single-source gain after linear calibration.

Geometry	Rank $\rho \uparrow$	Calibration error↓	Top-3 regret↓	NT↓
Raw parameter cosine	.36	.061	.072	.143
Layer-normalized parameter cosine	.48	.051	.058	.121
Diagonal Fisher angle	.66	.039	.035	.087
Target gradient cosine	.70	.036	.031	.079
Functional Gram, no off-diagonals	.73	.031	.028	.071
Functional Gram, no cross-fit norm.	.76	.028	.025	.064
Full target-functional geometry	.79	.021	.020	.054

F.2 Robustness and Mechanism Diagnostics

We next perturb the source library under the reference CLIP-8 empirical condition while leaving the target observations and adaptation budget unchanged. Irrelevant-source contamination augments the bank with source updates intended to carry no useful target signal, whereas the sign-reversal stress replaces the indicated fraction of stored source updates by their negatives. The contamination level is fixed before target validation. Table F.5 reports the resulting selection, estimated completeness, gain over target-only, and fallback behavior. As contamination rises, Table F.5 shows that the selected source count and estimated completeness decline while fallback becomes more frequent. The behavior is conservative in the intended direction: less credible reusable signal produces less aggressive source deployment. Figure F.1 gives the same mechanism continuously, not only at the discrete contamination levels in Table F.5.

Table F.5: **Robustness to irrelevant and sign-reversed source injection on CLIP-8.** Results are macro-averaged over the eight held-out target tasks under the 32-shot reference data budget using the independent replicate schedule of this contamination experiment.

Contamination	Selected sources	$\hat{\chi}_4$	ΔAcc. vs TO	Fallback
0% irrelevant	3.1	.68	+3.1	.09
20% irrelevant	3.0	.66	+3.0	.12
50% irrelevant	2.8	.62	+2.7	.18
100% irrelevant	2.6	.57	+2.2	.26
25% sign-reversed	2.7	.59	+2.4	.23
50% sign-reversed	2.2	.48	+1.5	.39

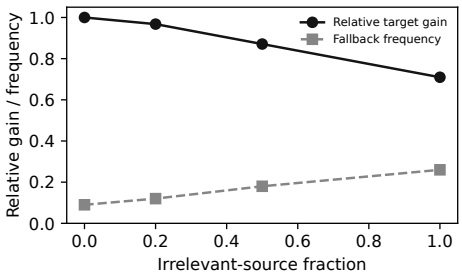


Figure F.1: **Robustness to source-bank contamination on CLIP-8.** As irrelevant or sign-reversed atoms are injected, source-supported gain decreases and target-only fallback becomes more frequent. Curves are macro-averaged over the eight held-out CLIP-8 target tasks.

The block-level diagnostics use the same six-group refinement $B_1, \ldots, B_6$ introduced for Figure 3, not merely the default three-group transfer atomization. The six-group partition is fixed in advance and is used only to obtain a higher-resolution view of where source-target alignment occurs across depth. The source-by-block heatmap in the main paper reports the directed alignment

$$\hat{A}^{\mathrm{dir}}_{s\ell} = \left[\frac{\hat{b}_{s\ell}}{\sqrt{(\hat{\mathbf{G}}_{\mathrm{diag}})_{s\ell,s\ell}}\sqrt{\hat{R}_{\mathrm{cf}}}} \right]^{1}_{-1},$$

where $\hat{\mathbf{G}}_{\mathrm{diag}} = \hat{\mathbf{G}}_{\mathrm{st}}(\rho_G^{\mathrm{diag}}, \lambda_G^{\mathrm{diag}})$ uses the fixed pre-validation diagnostic stabilization specified in Appendix B. We additionally compute the partial angle after conditioning on already selected blocks using residualized tangent fields from Corollary A.3. Early blocks often exhibit broad positive alignment across related vision domains, whereas later blocks are more target specific. This pattern is not assumed by the method; source and block selection arise from the target-conditioned geometry.

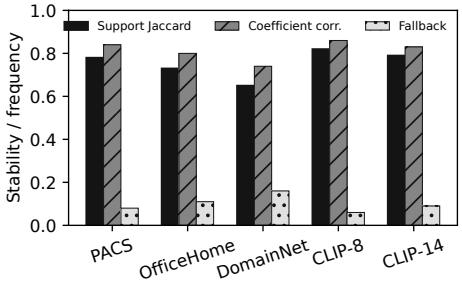


Figure F.2: **Stability of the six-group diagnostic block selection across independent target samples.** Results are shown separately for PACS, OfficeHome, DomainNet, CLIP-8, and CLIP-14. Within each library, support Jaccard agreement and coefficient correlation are computed using the fixed diagnostic partition $B_1, \ldots, B_6$; dashed lines give the corresponding target-only fallback frequency. The default predictive procedure continues to use the three-group atomization of Appendix E.

Table F.6: **Mechanism ablations on CLIP-8.** ΔAcc. is the macro-averaged accuracy change relative to full GEOTRANSFER over the eight held-out target tasks; NT and fallback are frequencies over the same matched target-seed replicates.

Variant	ΔAcc.↑	Rank ρ ↑	NT↓	Fallback
Full GEOTRANSFER	0	.79	.019	.11
No group sparsity	-0.4	.79	.058	.07
No innovation adapter	-0.8	.79	.061	.10
No orthogonality penalty	-0.3	.79	.067	.09
No risk gate	+0.2	.79	.054	0
Finite-difference tangent	-0.2	.76	.025	.12

Figure F.2 evaluates whether the routing pattern is stable to the particular few-shot target sample. DomainNet is included in this stability diagnostic with PACS, OfficeHome, CLIP-8, and CLIP-14; the figure concerns support reproducibility, not merely the aggregate performance columns of Table F.2. Table F.6 isolates algorithmic components not already covered by the geometry and contamination experiments. Each row removes one component while leaving the remaining fitting and target-data protocol

unchanged. The component ablation separates predictive gain from deployment protection. Removing the innovation adapter primarily reduces average accuracy; removing the risk gate slightly increases average accuracy but raises NT; and removing group sparsity or the orthogonality penalty weakens selective source reuse without changing the underlying one-source ranking score. The finite-difference row shows the cost of replacing exact tangent products by the approximation described in Appendix B.

F.3 Computation and Sensitivity

Table F.7 separates one-time source-ranking computation from deployment cost. "Source passes" means forward evaluation of target fit examples through stored source models when required by a transferability score; it does not mean access to source training examples. JVP counts are per target fit batch. The additional JVP cost is paid only during geometry construction. Once the selected update and innovation adapter have been assembled, GeoTransfer requires neither the source models nor a test-time router, so its deployed forward cost matches the corresponding target model.

Table F.7: **Offline and deployment computation.** JVP counts are per target fit batch. Deployment FLOPs exclude one-time geometry construction.

Method	Source passes	JVPs	Trainable params	Rel. test FLOPs
Target only	0	0	1.0×	1.00
Task arithmetic	0	0	scalars	1.00
Transferability score	S	0	1.0×	1.00
Gradient matching	0	0	scalars	1.00
MASS-style routing	offline	0	small	1.85
GeoTransfer angle only	0	M	scalars	1.00
GeoTransfer projection	0	M	scalars	1.00
GeoTransfer + innovation	0	M	1.0×	1.00

The sparse-completeness budget q is used only to evaluate how rapidly the source library covers the target descent field; it does not constrain the penalized transfer estimator. Table F.8 therefore varies q while keeping the fitted GeoTransfer predictor fixed. The reference empirical summary uses $q = 4$. Estimated completeness increases rapidly over the first few atoms and then levels off, indicating diminishing additional coverage of the target descent field. This diagnostic does not alter the support selected by the penalized projection. Figure F.3 complements Table F.7 by showing how the one-time geometry cost changes with library size.

Table F.8: **Sensitivity of empirical transfer completeness to the diagnostic budget on CLIP-8.** Results are macro-averaged over the eight held-out CLIP-8 target tasks under the 32-shot reference condition. The fitted transfer procedure is unchanged as q varies.

q	1	2	4	6	8
$\hat{\chi}_q$	.34	.52	.68	.75	.79

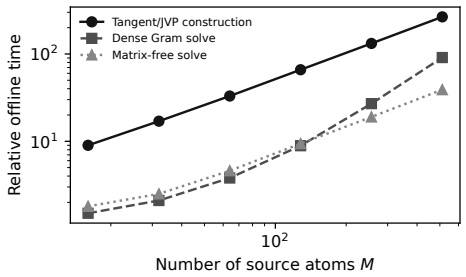


Figure F.3: **Offline geometry-construction scaling.** Batched JVP construction grows approximately linearly with the number of atoms; matrix-free Gram products reduce the additional solve cost for larger libraries. Deployment cost is unchanged after the transferred update is formed.

References

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. (2017). Deeper, broader and artier domain generalization. *IEEE International Conference on Computer Vision*, 5543-5551.

Maurer, A. and Pontil, M. (2009). Empirical Bernstein bounds and sample variance penalization. *Proceedings of the 22nd Annual Conference on Learning Theory*.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. (2019). Moment matching for multi-source domain adaptation. *IEEE/CVF International Conference on Computer Vision*, 1406-1415.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748-8763.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. *IEEE Conference on Computer Vision and Pattern Recognition*, 5385-5394.