CONTROLLING MODEL UPDATES AT THE GRANULARITY OF OUTPUT PERMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model updating adapts a deployed model to revised training annotations. Such updates can also change outputs beyond their intended scope. For example, when relation annotations are added to a document, only the scores of entity pairs that gain relations are permitted to change. The scores of the remaining pairs are protected outputs that should stay unchanged. Existing methods control an update either for the whole model through compatible training or for each input through edit-scope routing. Preservation constraints instead restrict non-target distributions within an edit prompt or enforce specified output properties. However, a decision made for a whole input applies equally to the pairs permitted to change and to the pairs that must be protected. In this paper, we ask what shared controls cost and when that cost disappears. We compare document-scope switching with full redeployment for annotation revisions from DocRED to Re-DocRED. Switching only documents gaining relations retains 99.98% of full-redeployment root mean square (RMS) change in protected relation margins. This cost arises because a shared control scales the permitted and protected changes within each group by the same factor. For fixed updates with locally linear gains on permitted outputs, we derive in closed form the minimum mean squared protected change of each grouping, which reaches zero exactly when some group with useful slope has no protected response. Building on this analysis, we propose authorization-granular endpoint routing (AGER), which estimates whether each class score or entity pair may change and scales its update accordingly. On ImageNet, AGER reduces the mean squared protected change by a median of 97.4% relative to the strongest uniformly scaled update that meets the target gain. On document relation extraction, AGER retains 0.376 and 0.436 of the protected RMS change of retrained DREEAM and ATLOP while changing micro-F1 by -1.4 and -2.2 points. These results indicate that model updates should be controlled at the granularity of output permissions.

1 INTRODUCTION

Model updating adapts a deployed model to revised annotations while seeking to preserve behavior outside the intended scope (Yan et al., 2021; Cai et al., 2022). Document-level relation extraction (DocRE) makes this distinction concrete because entity pairs share a document's context (Yao et al., 2019; Tan et al., 2022). Adding an employment relation permits that person–organization pair's scores to change. The remaining pairs are protected outputs whose annotations remain unchanged.

Positive-Congruent Training emphasizes previously correct predictions (Yan et al., 2021), while backward-compatible training keeps new embeddings compatible with the earlier classifier (Shen et al., 2020). Weight interpolation balances improvement against prediction changes (Schumann et al., 2024) and reranking reduces changes to complete structures (Cai et al., 2022). Finer adaptation selects teachers per token (Echterhoff et al., 2024) or learns class-specific feature adjustments (Kim & Do, 2025).

Memory-based editors learn where revised behavior applies (Mitchell et al., 2022b; Hartvigsen et al., 2023; Wang et al., 2024a). Complementary constraints preserve non-target distributions (Jeong et al., 2026) or specified output properties (Tao & Thakur, 2024), and protect retained knowledge through null-space projection (Fang et al., 2025).

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.

However, an input-level deployment decision still applies jointly to outputs with different permissions. Selecting the revised model for a document changes both the newly annotated employment pair and protected pairs. We therefore ask what shared controls cost when permissions differ within an input, and when that cost disappears.

We define an authorization unit as the smallest output unit with its own permission. For DocRE it is a directed entity pair whose relation scores may change when annotations gain relations. For image classification it is a class score permitted to change on images belonging to that class. We separate improvement on permitted outputs, called utility, from protected-output changes (Section 3).

From shared controls to output permissions. For fixed updates with locally linear utility, we derive each grouping's minimum mean squared protected change at a positive attainable target (Section 4). Splitting groups preserves or lowers this minimum. We propose authorization-granular endpoint routing (AGER) to estimate permissions from base-model scores and scale each unit's response. AGER selects the interface and scale with the smallest protected-response upper bound among candidates whose utility lower bounds meet the target.

Evidence and contributions. We evaluate DocRED to Re-DocRED revisions using a retrained relation extractor (Ma et al., 2023). Document-scope switching retains 99.98% of full-redeployment protected root mean square (RMS) change, as reported in Section 6.2. On ImageNet (Deng et al., 2009), AGER reduces mean squared protected change by a median of 97.4% relative to the strongest utility-qualified scalar (Table 1). Our contributions are:

- We introduce the authorization unit, the smallest output unit that carries its own permission, and formalize model updating as the joint control of permitted and protected outputs at that unit (Section 3).
- We prove that a grouping of outputs into shared controls has a closed-form minimum mean squared protected change at a required improvement, that refining a grouping never raises this minimum, and that it reaches zero exactly when some group with positive useful slope carries no protected mass (Theorem 1).
- We propose authorization-granular endpoint routing (AGER), which estimates each unit's permission from base-model scores and selects an interface and multiplier under a finite-sample utility lower bound; it removes 62% of the protected RMS response at a cost of 1.4 micro-F1 points on DREEAM (Section 6.2).

Matching controls to output permissions makes within-input protection part of the update decision.

2 Related Work

Editing mechanisms and their scope. Learned editors (Sinitsin et al., 2020; Mitchell et al., 2022a) and factual rewrites (Meng et al., 2022; 2023) modify model parameters, whereas retrieval-based editors route instead of overwriting: SERAC retrieves edits to select a predictor (Mitchell et al., 2022b), GRACE replaces local activations (Hartvigsen et al., 2023), WISE routes queries between memories (Wang et al., 2024a) and SeqMMR routes between models (Qiao et al., 2025).

Preserving behavior during updates. Parameter penalties (Kirkpatrick et al., 2017) and gradient constraints (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019; Farajtabar et al., 2020; Saha et al., 2021; Lin et al., 2022; Yu et al., 2020; Fang et al., 2025) restrict interfering updates, while output matching takes a complementary route (Li & Hoiem, 2018; Benjamin et al., 2019) through Mean Teacher (Tarvainen & Valpola, 2017), R-Drop (Liang et al., 2021) and self-distillation (Furlanello et al., 2018). Recent variants instead select targets or distributions: MUSCLE selects token-level distillation targets (Echterhoff et al., 2024), KLOD constrains non-target distributions (Jeong et al., 2026) and PREPARED enforces specified output properties (Tao & Thakur, 2024). Locality evaluations separate required factual propagation (Cohen et al., 2024) from protected-query changes (Liu et al., 2026; Wang et al., 2024b), and certified updates restrict the weights an update may reach (Elmecker-Plakolm et al., 2026). All of these act on the editing loss or the parameter domain; we instead hold the update fixed and vary only the grouping of outputs that share one control, measuring protected changes at matched useful improvement.

Conditional routing and selective adaptation. Expert routing allocates computation (Shazeer et al., 2017; Fedus et al., 2022; Hazimeh et al., 2021; Zhou et al., 2022) and selective adaptation modifies blocks or channels (Guo et al., 2019; Abati et al., 2020). Class-dependent adjustments (Valada et al., 2017; Kim & Do, 2025) motivate AGER's finer control over exposed output changes.

Selective prediction and finite-sample selection. Selective prediction learns when to reject (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019) or defer to an expert (Mozannar & Sontag, 2020). Candidate testing (Angelopoulos et al., 2025) and prediction-set calibration (Bates et al., 2021; Angelopoulos et al., 2024) provide finite-sample risk control. Our selection minimizes a protected-response upper bound subject to a utility lower bound using simultaneous concentration bounds (Hoeffding, 1963).

Document-level relation extraction. DocRE predicts entity-pair relations from shared context (Yao et al., 2019), with revised annotations correcting omissions (Tan et al., 2022). ATLOP combines pair-specific thresholds and context (Zhou et al., 2021), Eider and DREEAM add evidence guidance (Xie et al., 2022; Ma et al., 2023), graph and rule methods enrich relational reasoning (Zeng et al., 2020; Xu et al., 2021; Ru et al., 2021; Fan et al., 2022), and recent objectives refine relation margins and threshold learning (Duan et al., 2025; Xu et al., 2026).

3 Endpoint Updates and Authorization Units

We formulate model updating when outputs within an input have different permissions. The endpoint and permission definitions identify what may change independently of the editor or model architecture. We then require useful improvement before assessing protected response. Suppressing an entire update would preserve every output while leaving the requested revision unapplied. Section 4 then derives the smallest protected response a shared control reaches at a required improvement, and Section 5 builds routes at the permission unit.

3.1 Endpoints and permissions

An endpoint collects the model scores exposed to an update. For $F_\theta(x) \in \mathbb{R}^{K_x}$ with endpoint width K_x, the exact response is $\boldsymbol{d}_{\boldsymbol{\Delta}}(x) = F_{\theta+\boldsymbol{\Delta}}(x) - F_\theta(x)$ and its local approximation is $J_\theta(x)\boldsymbol{\Delta}$. Here θ denotes the model parameters and J_θ is the endpoint Jacobian. Exact responses compare checkpoints, while local responses connect update directions to useful improvement. An authorization unit has its own permission: an image–class coordinate or a directed DocRE pair sharing permission across its relation scores. The supplied mask $\boldsymbol{a}(x) \in \{0,1\}^{K_x}$ defines writable coordinates $W(x) = \{j : a_j(x) = 1\}$ and protected coordinates $P(x) = \{j : a_j(x) = 0\}$. An interface consists of a route $\boldsymbol{q}(x) \in [0,1]^{K_x}$ and multiplier $s \geq 0$, exposing $F_\theta(x) + s\boldsymbol{q}(x) \odot \boldsymbol{d}_{\boldsymbol{\Delta}}(x)$. A zero route value preserves the base score, whereas a value of one applies the full response at $s = 1$. Scalar routes share a strength across the input, whereas fine routes assign strengths to authorization units. We write $\boldsymbol{r}$ for fitted route values.

3.2 Utility and protected response

Protection is measured at a required useful improvement. Let $X_{q,s}(\xi)$ be the benefit over the base endpoint on an independent utility unit $\xi \sim \mathcal{D}_U$, with mean $U(q,s) = \mathbb{E}X_{q,s}(\xi)$ and $U(q,0) = 0$. ImageNet uses negative log-likelihood benefit and DocRE uses added-relation margin gain. For inputs $x \sim \mathcal{D}_P$ containing protected coordinates, we minimize

$$R(q,s) = \mathbb{E}\left[\frac{1}{|P(x)|}\sum_{j\in P(x)}(sq_j(x)d_{\boldsymbol{\Delta},j}(x))^2\right] \quad \text{subject to } U(q,s) \geq t > 0. \tag{1}$$

Root mean square (RMS) response is $\sqrt{R}$. The inner average is over protected coordinates and the outer average is over inputs. Reported DocRE values pool protected coordinates across documents, and we report the document-by-document comparison alongside them. Task accuracy and micro-F1 measure prediction outcomes separately.

4 STRUCTURAL RESULTS AND DESIGN REQUIREMENTS

Shared update controls couple permitted changes to protected outputs. We quantify this coupling for a fixed response, then for a required improvement, and finally for accumulating protected mass.

4.1 BROADCAST RESPONSE AND COARSE-INTERFACE EFFICIENCY

Broadcasting increases strengths on outputs that a fine route could suppress. For fixed $\boldsymbol{d}$ and $\boldsymbol{r}$, max-union broadcast $u = \max_j r_j$ adds squared response $\sum_j (u^2 - r_j^2) d_j^2 \geq 0$ (Proposition 2). The identity also holds on protected coordinates. To require benefit, partition outputs into shared-control cells $g \in G$ with useful slopes $B_g = \sum_{j \in W \cap g} b_j d_j$ and protected masses $P_g = \sum_{j \in P \cap g} d_j^2$. Here $\boldsymbol{b}$ contains benefit weights on writable outputs. With cell route values $c_g \in [0, 1]$, effective strengths $v_g = s c_g \geq 0$ give local utility $U = \sum_g v_g B_g$ and response $R = m^{-1} \sum_g v_g^2 P_g$ with $m = |P| \geq 1$.

Theorem 1 (Coarse-interface efficiency ceiling) *Let $H_G = \sum_{g:B_g>0} B_g^2/P_g$ and assume $0 < H_G < \infty$. Among shared-control routes at any target $t > 0$, the minimum response is*

$$\min_{U \geq t} R = \frac{t^2}{m H_G}, \qquad \sup \eta = m H_G, \qquad \eta = U^2/R \quad (U > 0). \tag{2}$$

The floor is attained by strengths proportional to useful slope over protected mass, and it remains positive whenever every cell with positive useful slope also carries protected mass. Refining the partition cannot raise it. All 3,794 documents that gain relations also contain protected pairs, and those pairs respond to the revised checkpoint, so document-level control pays a positive floor on this workload. A shared strength cannot match a mixed permission mask either, and its smallest squared error is the writable count times the protected count divided by the endpoint width (Proposition 1). Attaining strengths and zero-mass cases are in Appendix A.

4.2 SEQUENTIAL DEPLETION UNDER GROWING PROTECTED MASS

Reusing shared controls can expose an accumulating protected set. Let $P_g^{(\tau)} = P_g^{(0)} + (\tau - 1)\pi_g$ with $\pi_g > 0$ and fixed useful slopes B_g. Here π_g is the per-round increase and $H_+^{(\tau)} = \sum_{g:B_g>0} B_g^2/P_g^{(\tau)}$.

Theorem 2 (Sequential depletion) *Assume a positive useful slope exists and $1 \leq m_\tau \leq K < \infty$ independently of the horizon. The efficiency ceiling decays as $O(1/\tau)$ and the minimum response at a fixed positive target grows as $\Omega(\tau)$. With a fixed initial target and fixed per-round response budget, cumulative installed utility is $O(\sqrt{T})$.*

Each reuse of a shared control adds protected mass to its cells, so a fixed per-round response budget installs total utility of order $\sqrt{T}$ after T rounds (proof in Appendix A).

5 AUTHORIZATION-GRANULAR ENDPOINT ROUTING

At the permission unit a writable pair carries no protected mass, so its floor is zero (Appendix A). Reaching that unit requires knowing which outputs may change, and authorization-granular endpoint routing (AGER) estimates this from base-model scores. The route construction defines matched controls that isolate the effects of granularity and permission alignment. We then select among fitted candidates using separate bounds on useful improvement and protected response, because the smallest response alone could correspond to an update that achieves none of the intended revision.

5.1 ROUTES AND MATCHED CONTROLS

Permission estimates assign different strengths within each input. For ImageNet we calibrate the base margin $g_j = z_j - \max_{k \neq j} z_k$ into permission probability π_j and use $r_j = r_{\min} + (1 - r_{\min})\pi_j^p$. Here $\boldsymbol{z}$ contains base logits and $r_{\min}$ is the route floor. The power p controls concentration and $\boldsymbol{h}(x)$ is the feature vector. We estimate permissions with balanced logistic fits to standardized margins. ImageNet

uses class-specific parameters and DocRE uses a shared calibrator. Each ImageNet interface fits its own coefficients $\boldsymbol{v}_q$ in the feature perturbation $\boldsymbol{h}(x) \odot B\boldsymbol{v}_q$. Interfaces share the orthonormal Gaussian basis B and use identical samples and regularized cross-entropy objectives.

For DocRE a directed pair is writable when revised annotations add a relation. A relation margin is its score minus the no-relation score. We calibrate its largest base non-null margin into π_{ef} and set $r_{ef} = r_{\min} + (1 - r_{\min})\pi_{ef}$ across its relation margins. AGER scales the exact updated-minus-base margin response by sr_{ef}. Revised-negative pairs have no revised relations, while pure protected-positive pairs retain relations without additions. The calibrators and fitting objectives are specified in Appendix B. Max-union broadcast and an exposure-matched scalar share strengths across the input. The exposure-matched scalar uses the RMS of fine route values. For DocRE the mean-broadcast scalar instead assigns the mean pair-route value within a document to every pair in that document. Raw-confidence and within-input permutation controls isolate calibration and permission alignment. An authorization-mask Oracle uses known permissions as a capacity reference outside the deployable candidate family.

5.2 Utility-qualified finite-grid selection

Selection requires simultaneous control of estimation error. We apply fit-then-evaluate risk control (Angelopoulos et al., 2025) with Hoeffding bounds (Hoeffding, 1963) to M fitted interface–multiplier candidates fixed before support evaluation. Each candidate c has independent utility observations $X_{c,i} \in [-B_U, B_U]$ and response observations $Y_{c,i} \in [0, B_R]$ within disjoint supports of sizes n_U, n_P. The independent units are images or complete documents. Their means are U_c, R_c, with sample means $\widehat{U}_c, \widehat{R}_c$. The simultaneous Hoeffding radii ϵ_U, ϵ_R bound these candidate means at total error probability δ (Appendix B). AGER minimizes $\widehat{R}_c + \epsilon_R$ among candidates with $\widehat{U}_c - \epsilon_U \geq t$, using a fixed tie order and returning no update if none qualifies.

Theorem 3 (Finite-grid guarantee) *Under the stated independence and bounds, with probability at least $1 - \delta$ every returned candidate $\widehat{c}$ satisfies $U_{\widehat{c}} \geq t$. For every candidate with $U_c \geq t + 2\epsilon_U$, a candidate is returned and $R_{\widehat{c}} \leq R_c + 2\epsilon_R$.*

When observations are clipped, the guarantee concerns their clipped population means. A returned candidate therefore carries the required improvement and a response bound at the same confidence level (proof in Appendix B).

6 Experiments

We test whether permission-aligned routing reduces protected response at required useful improvement. ImageNet class scores and DocRE relation margins give complementary comparisons with separately fitted directions and a shared checkpoint response. We report the design and primary results, then mixed permissions, selection behavior, further output units and deployment costs.

6.1 Experimental design

We compare ImageNet updates with ResNet-50 and DenseNet-121 alongside MobileNetV3-Large (Deng et al., 2009; He et al., 2016; Huang et al., 2017; Howard et al., 2019), and aligned DocRED and Re-DocRED annotations that identify added relations under DREEAM BERT-base and ATLOP RoBERTa-large (Yao et al., 2019; Tan et al., 2022; Ma et al., 2023; Zhou et al., 2021). Alignment requires identical sentences and canonical entity order so that permissions refer to the same pairs in the base and revised checkpoints, whose difference is shared across all DocRE routing controls. Each ImageNet interface instead fits its own direction in the same feature basis and budget on 20 writable and 20 protected classes, using separate images for route fitting and direction training. Utility and response support estimate their objectives before the selected update is evaluated on held-out images, separating the control of an existing update from the interface's effect on learning a useful direction.

We compare calibrated and raw fine routes with max-union and exposure-matched scalars; permutation preserves the route strengths while changing their permission alignment. External methods include SERAC-style routing and AlphaEdit-style projection (Mitchell et al., 2022b; Fang et al., 2025), while OGD and EWC test gradient projection and parameter regularization (Farajtabar et al., 2020;

Table 1: **AGER reaches both utility targets with less protected change.** Paired entries correspond to $t = 0.01/0.02$. LCB is the simultaneous utility lower bound. Evaluation/target divides held-out utility by the target. Reduction compares mean squared response with the strongest eligible scalar.

Model	Utility LCB ↑	Eval./target ↑	Reduction ↑
ResNet-50	0.0109 / 0.0202	1.61 / 1.41	96.8% / 95.9%
DenseNet-121	0.0113 / 0.0207	1.67 / 1.52	97.5% / 97.4%
MobileNetV3-Large	0.0103 / 0.0200	1.76 / 2.24	98.7% / 99.0%

Kirkpatrick et al., 2017) and L2-SP and LoRA serve as further weight-update comparators (Li et al., 2018; Hu et al., 2022). Native full-head fitting supplies an ordinary ImageNet reference. External methods update the classifier head or a low-dimensional endpoint with 8,096 to 2,050,049 trainable parameters, and DocRE comparisons add document-scope deployment and protected-gradient head projection. The known-mask Oracle is handed the true permissions and thus holds more information than AGER, measuring what the interface attains when permissions need not be estimated.

Larger utility and smaller protected RMS are preferable. ImageNet targets $t \in \{0.01, 0.02\}$ are floors on clipped negative log-likelihood benefit, so held-out utility may exceed them; for DocRE a scalar qualifies when support utility reaches at least half the target within a relative gap of 0.25, and we choose the qualifying scalar with the smallest support response. Route fitting and support evaluation use samples disjoint from the held-out evaluation, whose intervals use 20,000 bootstrap replicates over ImageNet authorization classes or complete DocRE documents to preserve dependence among outputs sharing an input. The complete sample allocations and candidate grids appear in Appendix C.

6.2 Fine routing reduces protected change at required utility

Matched ImageNet updates. Against the strongest utility-qualified scalar, fine routing keeps the required improvement and removes most of the protected change (Table 1). Every selected route reaches its support target, and held-out utility ranges from 1.41 to 2.24 times the required value across the six model–target combinations. The median reduction in mean squared protected response is 97.4%, so the permitted improvement is met on every model while the protection is set by the granularity of the control rather than by the strength chosen at that granularity. With one fixed update direction across all routes, the same six combinations give a median reduction of 96.95% (Appendix E).

Matched DocRE updates. With the checkpoint response held fixed, control granularity alone changes overall protected RMS by more than a factor of two. Table 2 reports scalar-to-pair RMS ratios, with larger values favoring pair routing. Overall and revised-negative ratios exceed two at both endpoints and their lower interval endpoints remain above one. Even the best document-level strengths reach 1.94 times the pair route's protected RMS at matched utility on the design documents (Appendix D). Parameter-space methods apply one learned change to every pair, so we scale their responses to the common utility before comparing, and the matched LoRA and head projection updates then produce greater overall protected RMS. Both mean-broadcast intervals lie below one on pure protected-positive pairs, the only stratum in Table 2 where the scalar is preferable. These comparisons behave as a shared strength must when the strata it pools disagree in permission, not as an artefact of the strength chosen. When permissions are supplied exactly the known-mask Oracle leaves no protected response at matched support utility, which is the zero-mass boundary that Theorem 1 identifies between a reduced floor and a vanishing one.

External and native ImageNet updates. We compare external methods with AGER at the clipped-utility target $t = 0.01$. At $t = 0.01$ every method-to-route RMS interval has a lower endpoint of at least 3.2. AlphaEdit-style projection and SERAC-style routing give the closest external points. The smallest comparator-to-AGER RMS ratios are 9.04, 6.64 and 3.75 for ResNet-50, DenseNet-121 and MobileNetV3-Large. The discrete scale grid produces greater utility overshoot for native fitting and EWC on DenseNet, whereas the external ordering also holds on backbones with comparable overshoot. Across both targets the EWC penalty keeps at least 2.398 times the route's protected RMS, and at least 17.93 times where it removes 66.35% of the Fisher-weighted update energy (Appendix G). It moves protected RMS by at most 0.1502% against the unpenalized update, so the EWC and

Table 2: **Pair routing reduces overall and revised-negative response.** Comparator/pair RMS ratios use the 160 design documents and DREEAM BERT-base unless ATLOP RoBERTa-large is specified. The last columns cover revised-negative and pure protected-positive pairs. Brackets are endpoint-wise 95% document-bootstrap intervals. Dashes mark absent stratum intervals because the permuted control was bootstrapped only for overall response.

Comparator	Overall ↑	Revised-neg. ↑	Pure-prot. ↑
Mean-broadcast scalar	2.467 [2.416, 2.521]	2.600 [2.537, 2.665]	0.938 [0.886, 0.990]
Mean-broadcast (ATLOP)	2.132 [2.096, 2.169]	2.337 [2.293, 2.382]	0.871 [0.804, 0.955]
Permuted-pair control	3.39 [3.32, 3.47]	–	–
LoRA r=16	11.17 [10.72, 11.64]	11.83 [11.33, 12.36]	2.64 [2.54, 2.76]
Head projection k=8	9.35 [8.95, 9.75]	9.91 [9.48, 10.36]	1.82 [1.75, 1.91]

Table 3: **External ImageNet updates have greater protected RMS.** Ratios divide method RMS by fine-route RMS at $t = 0.01$. Brackets are endpoint-wise 95% authorization-class bootstrap intervals. Configurations are selected on support data.

Method	ResNet-50 ↑	DenseNet-121 ↑	MobileNetV3-L ↑
SERAC query routing	9.17 [7.57, 11.90]	9.26 [7.17, 12.78]	3.75 [3.21, 4.28]
AlphaEdit projection	9.04 [7.12, 12.36]	6.64 [5.18, 9.28]	3.85 [3.49, 4.25]
OGD	18.20 [14.50, 24.70]	9.13 [7.14, 12.65]	4.17 [3.84, 4.52]
EWC	17.90 [14.29, 24.34]	17.79 [13.95, 24.69]	4.13 [3.79, 4.47]
L2-SP	25.59 [20.49, 34.67]	14.40 [11.27, 19.86]	6.73 [6.18, 7.31]
LoRA	22.35 [17.91, 30.28]	12.61 [9.84, 17.43]	9.40 [8.59, 10.28]
Native full-head	17.90 [14.30, 24.21]	17.79 [13.99, 24.61]	4.13 [3.79, 4.47]

native full-head point estimates in Table 3 coincide. In the separate native-head comparison the route-to-native response ratios have upper interval endpoints of at most 0.445, and every model meets the utility noninferiority margin of -0.005. The $t = 0.02$ panel, the utility operating points and the projection-basis comparison appear in Appendix E. The ranking is stable across both targets while the discrete utility scale is not, so protected response answers a question the utility scale cannot resolve.

DocRE deployment and task outcomes. On the deployed checkpoints, document-level switching protects nothing measurable, while pair routing removes 62% of the protected RMS response at a cost of 1.4 micro-F1 points (Table 4). Document-scope switching retains 99.98% of full-redeployment protected RMS because 442 of the 443 test documents contain a writable pair, so selecting the revised checkpoint opens 169,034 of the 169,090 directed pairs while only 8,224 of them may change. Its full-panel pair/native RMS ratios are 0.376 for DREEAM and 0.436 for ATLOP, at -1.4 and -2.2 micro-F1 points. For DREEAM the micro-F1 change comes from added relations that the route installs less often, where the added-recall difference is -0.022. Retained recall changes by -0.003 $[-0.008, +0.001]$, so the cost concentrates on the added relations rather than on the relations already present. The 91 documents outside route development give 0.376 $[0.356, 0.398]$ for DREEAM and 0.438 $[0.416, 0.460]$ for ATLOP. The pooled scope-to-pair response ratio is 2.66 [2.60, 2.72]. On pure protected-positive pairs the pair route responds more than full redeployment (4.486 versus 4.204 RMS). The ratio there passes one at multiplier 1.217, below the 1.299 that matches utility (Appendix D). Pair routing therefore protects the pairs that a document decision cannot separate, so the micro-F1 cost is paid for protection that no document-level rule can reach.

6.3 Permission structure and selection behavior

Document switching protects little because protected response lives in the same documents as the additions, and no document-level decision separates them. On the test panel the documents that gain relations hold 160,810 of the 160,866 protected pairs. Across the 3,826 aligned documents, 98.1% of those with writable pairs also contain pure protected-positive pairs. The median document has 342 directed pairs and only 13 writable pairs. A document decision therefore couples a small intended revision to a much larger protected set. Figure 1 separates typical documents from the upper tail: panel (a) reports medians, nearest-rank ninetieth percentiles and maxima on a logarithmic count axis,

Table 4: **Pair-level deployment reduces exposure within writable documents.** Results use 443 aligned test documents at the reported deployment scales. Exposure counts protected pairs with route values at least one half. Overall and pure-prot. columns report RMS pooled over protected coordinates. Pure-prot. denotes pure protected-positive pairs.

Method	Exposed pairs ↓	Overall ↓	Pure-prot. ↓	Micro-F1 ↑
Full redeploy	160,866	5.186	4.204	0.7472
Document-scope route	160,810	5.185	4.204	0.7472
Union broadcast	160,866	5.180	4.199	0.7469
Mean-broadcast scalar	360	4.845	4.300	0.7170
Pair route (ours)	22,922	1.951	4.486	0.7333

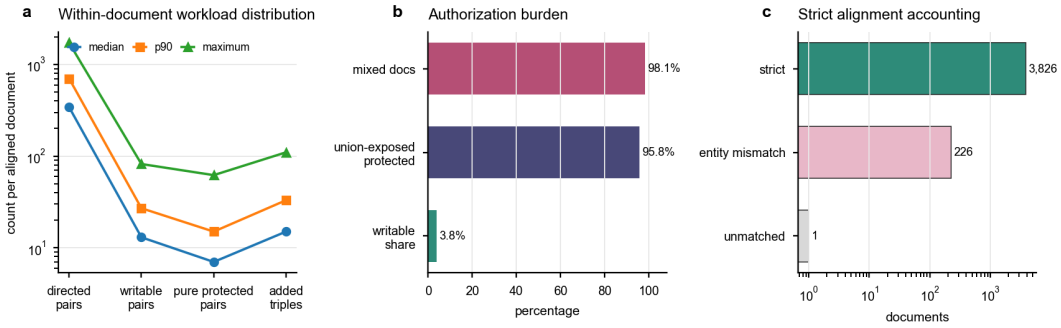


Figure 1: **Most writable documents also contain protected pairs.** Statistics describe the complete strictly aligned DocRE population. Panel (a) uses a logarithmic count axis for medians and nearest-rank ninetieth percentiles alongside maxima. Panel (b) reports mixed-document frequency and pair-exposure percentages. Panel (c) counts alignment outcomes. Markers and bars are population summaries without uncertainty intervals.

panel (b) reports mixed-document frequency and pair-exposure percentages, and panel (c) counts alignment outcomes.

Numerical response comparisons isolate the same mechanism. The mean union-to-fine squared-response ratio rises from 1.0 to 20.4 as endpoint width increases from 1 to 32 and falls back to 1.0 when every coordinate is authorized, tracing the coupling Section 4.1 derives: the penalty is widest while a shared control pools both permissions and disappears only once the grouping stops mixing them.

Permission alignment also matters. Permuting the DocRE pair strengths increases protected response while preserving their distribution, which rules out the strength distribution alone as an explanation for the aligned route's protection in Table 2. Appendix F reports the same comparison directly, where values above the dotted unit reference favor pair routing. Retained-positive pairs favor scalar control because their base scores assign high route values.

ImageNet alignment perturbations likewise distinguish a fine interface from a useful fine interface. With uniform permissions the fine and union routes coincide. After sufficient assignment permutation the scalar becomes preferable, with the observed transition fraction between 0.50 and 0.75. On unperturbed ImageNet routes the measured local efficiencies are $\eta_{\mathrm{loc}} = 12.47$, 6.48 and 2.46 for ResNet-50, DenseNet-121 and MobileNetV3-Large. Across 36 reachable alignment settings $R\eta_{\mathrm{loc}}/\widehat{U}^2$ exceeds one with mean 1.13, so the local prediction overstates the attainable ratio. Here η_{loc} uses the local utility slope rather than the achieved support utility $\widehat{U}$. The workload distributions and alignment comparisons appear in Appendix F.

The utility bound determines whether any candidate can be selected. Raising the ImageNet target from 0.01 to 0.02 shrinks the eligible set for every backbone, most sharply for MobileNetV3-Large from 632 to 283. Figure 2 shows the selection margins behind those counts: panel (a) reports the utility lower bound minus target for each model–target combination, panel (b) the held-out squared-response reduction relative to the eligible scalar, and panel (c) the eligible candidate count. The

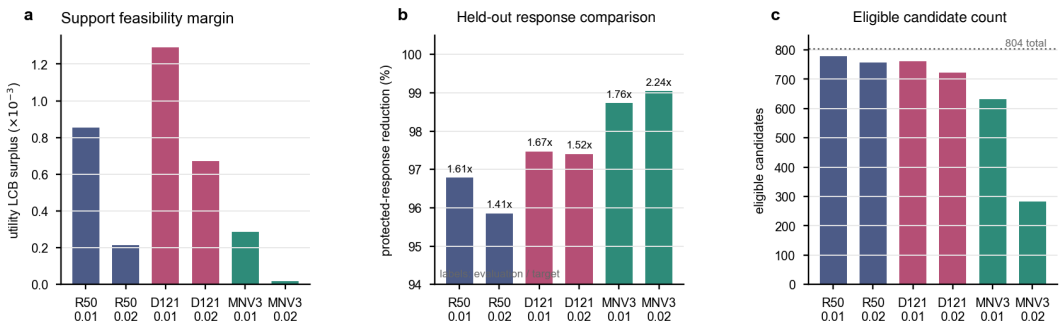


Figure 2: **Selected fine routes retain positive utility-bound margins.** Bars represent ImageNet model–target combinations. Panel (a) reports utility LCB minus target. Panel (b) reports held-out squared-response reduction relative to the eligible scalar, with labels giving held-out utility divided by target. Panel (c) counts eligible candidates. Bars are point summaries and larger values indicate more margin or reduction. R50 denotes ResNet-50, D121 denotes DenseNet-121 and MNV3 denotes MobileNetV3-Large.

selected multiplier increases within each architecture. Above the largest utility lower bound the selector returns no update.

Permuting fitting labels changes which route is useful, causing the full family to select the unchanged raw route while calibrated-only candidates incur larger protected responses, so a misdirected fit is rejected rather than returned at a larger protected cost. Thus the utility requirement rejects ineffective candidates and the protected-response objective chooses among those that still provide the requested benefit, which is the order Theorem 3 prescribes: no candidate is returned unless it first clears the target.

6.4 Additional endpoints and deployment costs

Constructed responses on CodRED, CoNLL-2003, AG News and CIFAR-100 test the structural results across cross-document, token-level and class-level authorization units (Yao et al., 2021; Tjong Kim Sang & De Meulder, 2003; Zhang et al., 2015; Krizhevsky, 2009). Fine routing has higher efficiency than shared controls on the first three endpoints and most CIFAR backbones, and partition refinement never raises the response floor. Ceiling decay follows protected-mass growth, as Theorem 2 predicts (Appendix G).

For actual GPT-2 XL edits from ROME and constrained fine-tuning (Meng et al., 2022), input-scope routing retains most protected response. Known prompt–token masks preserve identical writable change with zero protected response, realizing the zero-mass case of Theorem 1 (Appendix G).

ImageNet routing fits 32 coefficients. DocRE routing adds no encoder training but evaluates both checkpoints. Distillation reduces deployment to one checkpoint: the DREEAM student recovers 85% of the original-to-revised micro-F1 gain, with lower overall protected RMS and greater utility than matched LoRA and projection (Appendix H).

7 Conclusion

Output permissions determine the protection achievable by an update interface. For a fixed response, shared controls impose a positive response floor when every useful group also contains protected response. AGER addresses this coupling through output-level permission estimates and route selection under utility and protected-response bounds, with abstention when no candidate qualifies. ImageNet and DocRE experiments show lower overall protected response at required utility, while deployment comparisons quantify task-performance costs. These results make control granularity a measurable design choice: assess both the attainable protection floor and the utility of the resulting update.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure. Generative AI tools helped develop the theoretical models and conceptual frameworks, helped formulate the mathematical claims, provided critical ingredients for proving those claims, and assisted in the writing of the proofs. The mathematical claims, the proofs, and every experimental result were checked and verified by the authors. We additionally used generative AI tools for tasks with recommended disclosure, namely to aid and polish the writing and to edit the paper for readability. We have reviewed all AI-assisted work, including the mathematical derivations and the text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. Conditional channel gated networks for task-aware continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3931–3940, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Abati_Conditional_Channel_Gated_Networks_for_Task-Aware_Continual_Learning_CVPR_2020_paper.html.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2), 2025. doi: 10.1214/24-AOAS1998.

Daniele Angioni, Luca Demetrio, Maura Pintor, Luca Oneto, Davide Anguita, Battista Biggio, and Fabio Roli. Robustness-congruent adversarial training for secure machine learning model updates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7457–7469, 2025. doi: 10.1109/TPAMI.2025.3573237.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Ari S. Benjamin, David Rolnick, and Konrad Kording. Measuring and regularizing networks in function space. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkMwpiR9Y7.

Deng Cai, Elman Mansimov, Yi-An Lai, Yixuan Su, Lei Shu, and Yi Zhang. Measuring and reducing model update regression in structured prediction for NLP. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19384–19397, 2022. doi: 10.52202/068431-1409.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644.

Santiago Cortes-Gomez, Carlos Miguel Patiño, Yewon Byun, Zhiwei Steven Wu, Eric Horvitz, and Bryan Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iOMnn1hSBO.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Zhichao Duan, Tengyu Pan, Zhenyu Li, Xiuxing Li, and Jianyong Wang. COMM: Concentrated margin maximization for robust document-level relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23841–23849, 2025. doi: 10.1609/aaai.v39i22.34556. URL https://ojs.aaai.org/index.php/AAAI/article/view/34556.

Jessica Maria Echterhoff, Fartash Faghri, Raviteja Vemulapalli, Ting-Yao Hu, Chun-Liang Li, Oncel Tuzel, and Hadi Pouransari. MUSCLE: A model update strategy for compatible LLM evolution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7320–7332, 2024. doi: 10.18653/v1/2024.findings-emnlp.430.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Leo Elmecker-Plakolm, Pierre Fasterling, Philip Sosnin, Calvin Tsay, and Matthew Wicker. Provably safe model updates. In *2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2026. URL https://arxiv.org/abs/2512.01899.

Shengda Fan, Shasha Mo, and Jianwei Niu. Boosting document-level relation extraction by mining and injecting logical rules. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10311–10323. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.704. URL https://aclanthology.org/2022.emnlp-main.704/.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HvSytvg3Jh.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://www.jmlr.org/papers/v23/21-0998.html.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1607–1616, 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. SpotTune: Transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4805–4814, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Guo_SpotTune_Transfer_Learning_Through_Adaptive_Fine-Tuning_CVPR_2019_paper.html.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=Oc1SIKxwdV.

Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. DSelect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f5ac21cd0ef1b88e9848571aeb53551a-Abstract.html.

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

Bryce Hinkley and Peyman Najafirad. Residual paving: Diagnosing the routing bottleneck in selective refusal editing. *arXiv preprint arXiv:2605.20262*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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

Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru, and Kuldeep Singh. Revisiting document-level relation extraction with context-guided link prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18327–18335, 2024. doi: 10.1609/aaai.v38i16.29792. URL https://ojs.aaai.org/index.php/AAAI/article/view/29792.

Hojun Jeong, Gyunyeop Kim, and Sangwoo Kang. KLOD: Locality-preserving knowledge editing via non-target distribution preservation. arXiv:2608.27839. To appear in Findings of the Association for Computational Linguistics: EMNLP 2026, 2026.

Jaeik Kim and Jaeyoung Do. Exploring and leveraging class vectors for classifier editing. In *Advances in Neural Information Processing Systems*, volume 38, pp. 54894–54927. Curran Associates, Inc., 2025. doi: 10.52202/085713-1837. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4f0484db17ae7d2edb3c2031b885ba76-Abstract-Conference.html.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Siyuan Li, Youyuan Zhang, Ruitong Liu, Junxi Wang, and Jing Li. Multimodal knowledge edit-scoped generalization for online recursive MLLM editing. *arXiv preprint arXiv:2607.01978*, 2026.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. URL https://arxiv.org/abs/1802.01483.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.14448.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. TRGP: Trust region gradient projection for continual learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=iEvAf8i6JjO.

Wei Liu, Haomei Xu, Hongkai Liu, Zhiying Deng, Ruixuan Li, Heng Huang, Yee Whye Teh, and Wee Sun Lee. Are we evaluating the edit locality of LLM model editing properly? *arXiv preprint arXiv:2601.17343*, 2026. doi: 10.48550/arXiv.2601.17343. URL https://arxiv.org/abs/2601.17343.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6467–6476, 2017.

Youmi Ma, An Wang, and Naoaki Okazaki. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1971–1983, 2023. doi: 10.18653/v1/2023.eacl-main.145.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=0DcZxeWfOPt.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15817–15831, 2022b.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Shanbao Qiao, Xuebing Liu, Akshat Gupta, and Seung-Hoon Na. SeqMMR: Sequential model merging and LLM routing for enhanced batched sequential knowledge editing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16932–16947, 2025. doi: 10.18653/v1/2025.findings-acl.870.

Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. Learning logic rules for document-level relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1239–1250. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.95. URL https://aclanthology.org/2021.emnlp-main.95/.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ZK4Cvv5Eci.

Raphael Schumann, Elman Mansimov, Yi-An Lai, Nikolaos Pappas, Xibin Gao, and Yi Zhang. Backward compatibility during data updates by weight interpolation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2846–2861, 2024. doi: 10.18653/v1/2024.eacl-long.174.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/CVPR42600.2020.00640. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Shen_Towards_Backward-Compatible_Representation_Learning_CVPR_2020_paper.pdf.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitry Pyrkin, Sergei Popov, and Artem Babenko. Editable neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJedXaEtvS.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting DocRED: Addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8472–8487, 2022. doi: 10.18653/v1/2022.emnlp-main.580.

Zhe Tao and Aditya V. Thakur. Provable editing of deep neural networks using parametric linear relaxation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3615.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1703.01780.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0076. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/047c84ec50bd8ea29349b996fc64af4b-Abstract-Conference.html.

Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. AdapNet: Adaptive semantic segmentation in adverse environmental conditions. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 4644–4651, 2017. doi: 10.1109/ICRA.2017.7989540. URL http://ais.informatik.uni-freiburg.de/publications/papers/valada17icra.pdf.

Ke Wang, Yiming Qin, Nikolaos Dimitriadis, Alessandro Favero, and Pascal Frossard. MEMOIR: Lifelong model editing with minimal overwrite and informed retention for LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/809a560bf60984be05180c071c352eaa-Abstract-Conference.html.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://openreview.net/forum?id=VJMYOfJVC2.

Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. EasyEdit: An easy-to-use knowledge editing framework for large language models. In Yixin Cao, Yang Feng, and Deyi Xiong (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 82–93, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.9. URL https://aclanthology.org/2024.acl-demos.9/.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of ACL*, pp. 257–268, 2022. doi: 10.18653/v1/2022.findings-acl.23. URL https://aclanthology.org/2022.findings-acl.23/.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14149–14157, 2021. doi: 10.1609/aaai.v35i16.17665. URL https://ojs.aaai.org/index.php/AAAI/article/view/17665.

Huangming Xu, Fu Zhang, Zhixuan Yang, Lu Zhang, and Jingwei Cheng. ATGL: An adaptive-threshold global loss for document-level relation extraction. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34702–34716, 2026. doi: 10.18653/v1/2026.acl-long.1603. URL https://doi.org/10.18653/v1/2026.acl-long.1603.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14294–14303, 2021. doi: 10.1109/CVPR46437.2021.01407.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 764–777, 2019. doi: 10.18653/v1/P19-1074.

Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. CodRED: A cross-document relation extraction dataset for acquiring knowledge in the wild. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4452–4472, 2021. doi: 10.18653/v1/2021.emnlp-main.366.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1630–1640. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.127. URL https://aclanthology.org/2020.emnlp-main.127/.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28*, pp. 649–657, 2015.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14612–14620, 2021. doi: 10.1609/aaai.v35i16.17717.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0515. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2f00ecd787b432c1d36f3de9800728eb-Abstract-Conference.html.

A PROOFS OF THE STRUCTURAL RESULTS

This section supports the representation barrier in Section 3 and the structural guarantees in Section 4. It gives the complete derivations and the secondary efficiency results.

A.1 ENDPOINT UPDATES AND AUTHORIZATION UNITS

The definitions below support the endpoint model in Section 3.

We formalize an update through the outputs it may change and the outputs it should preserve. We begin with output permissions and the response produced by an update. We then define useful improvement and protected change before asking whether one shared control can represent mixed permissions.

A.1.1 ENDPOINT, POLICY, AND RESPONSE

The endpoint and policy definitions below support the routing model in Section 3.

An update can affect several outputs of the same input even when their permissions differ. We therefore specify the output vector before defining how to control its change. For an input x, let $F_\theta(x) \in \mathbb{R}^{K_x}$ denote the endpoint exposed by the base model with parameters θ. An endpoint is the collection of output scores whose changes we study. Its width K_x is fixed for selected image-class scores and varies with the number of entity pairs in a document. For a model update $\boldsymbol{\Delta}$, we define its exact endpoint response as

$$\boldsymbol{d}_{\boldsymbol{\Delta}}(x) = F_{\theta+\boldsymbol{\Delta}}(x) - F_\theta(x), \tag{3}$$

where $\boldsymbol{d}_{\boldsymbol{\Delta}}(x)$ records the change in each selected output. A zero coordinate means that the update leaves that output unchanged. For local analysis, we use the first-order response $\boldsymbol{d}^{\mathrm{loc}}_{\boldsymbol{\Delta}}(x) = J_\theta(x)\boldsymbol{\Delta}$. Here $J_\theta(x)$ is the Jacobian mapping the declared update space to endpoint coordinates. The exact response supports comparisons between fixed checkpoints, while the local response connects update directions to useful improvement.

Permissions specify which of these changes are intended. An *authorization unit* is the smallest output unit with its own permission. For example, a directed entity pair forms one unit when all its relation scores share a permission. We encode an externally supplied policy by an authorization mask $\boldsymbol{a}(x) \in \{0,1\}^{K_x}$. A value $a_j(x) = 1$ permits coordinate j to change, whereas $a_j(x) = 0$ protects it. The corresponding coordinate sets are

$$W(x) = \{j : a_j(x) = 1\}, \qquad P(x) = \{j : a_j(x) = 0\}, \tag{4}$$

where $W(x)$ and $P(x)$ contain writable and protected outputs. Coordinates in the same authorization unit share the same mask value. The supplied policy defines the learning target and the protection objective.

To control the response in Eq. (3), we assign a route value $q_j(x) \in [0,1]$ to each coordinate. Let $\boldsymbol{q}(x)$ collect these values and let $s \geq 0$ be an overall multiplier. The routed endpoint is

$$F_\theta(x) + s\boldsymbol{q}(x) \odot \boldsymbol{d}_{\boldsymbol{\Delta}}(x), \tag{5}$$

where $\odot$ denotes coordinate-wise multiplication. The route sets the relative strength of each output change, and s controls their common scale. At $s = 0$ the endpoint equals the base endpoint. At $s = 1$, a route value of zero preserves its coordinate and a value of one applies its full response. We write q for a route as an object and $\boldsymbol{q}(x)$ for its vector at input x. A *scalar route* has the form $\boldsymbol{q}(x) = \gamma(x)\mathbf{1}$ with $\gamma(x) \in [0,1]$ and an all-one vector $\mathbf{1}$. A *fine route* assigns values separately to authorization units. Thus the policy specifies which outputs may change, while the route determines how much of the update they receive.

A.1.2 UTILITY AND PROTECTED RESPONSE

The measurement definitions below support the comparison objective in Section 3.

Equation (5) specifies the controlled output. We still need to measure useful improvement and protected change. A route that suppresses the entire update also suppresses the desired correction. We therefore compare protection only after requiring a specified improvement on writable outputs.

Table 5: Domain-specific endpoints under the shared authorization policy. The table defines the semantic objects used by the method. Dataset sizes and empirical outcomes are reported in the experimental sections.

Domain	Endpoint coordinate	Minimum authorization unit	Useful objective	Protected response and independent unit
ImageNet	selected class logit within a fixed 1000-class endpoint	image × update-class coordinate	held-out negative log-likelihood benefit on writable-class members	writable-logit RMS on protected-class members, with independent support images and authorization-class blocks for confirmation
DocRE	non-null relation margin for a directed entity pair	document × directed entity pair	margin gain on newly added relation coordinates	relation-margin RMS on nonwritable pairs, with complete documents as independent units

Let ξ be an independent utility unit drawn from the utility population $\mathcal{D}_U$. We measure its improvement over the base endpoint by $X_{q,s}(\xi)$ and define mean utility as

$$U(q,s) = \mathbb{E}_{\xi\sim\mathcal{D}_U}[X_{q,s}(\xi)], \tag{6}$$

where $U(q,s)$ averages the benefit on the writable objective. Positive utility means that the routed update improves that objective, and $U(q,0) = 0$. We use negative log-likelihood benefit for ImageNet and added-relation margin gain for document-level relation extraction (DocRE). Task-level accuracy and micro-F1 are reported separately.

Protection concerns outputs outside this objective. Let $\mathcal{D}_P$ be the protected population of inputs with at least one protected coordinate. We define mean squared protected response as

$$R(q,s) = \mathbb{E}_{x\sim\mathcal{D}_P}\left[\frac{1}{|P(x)|}\sum_{j\in P(x)}\left(sq_j(x)d_{\boldsymbol{\Delta},j}(x)\right)^2\right], \tag{7}$$

where $d_{\boldsymbol{\Delta},j}(x)$ is coordinate j of the endpoint response and $|P(x)|$ counts protected coordinates. The inner average measures the change per protected output, and the outer expectation averages over inputs. For a target $t > 0$, a route–multiplier pair is feasible when $U(q,s) \geq t$. Among feasible candidates, we seek small $R(q,s)$. We report $\sqrt{R(q,s)}$ as root mean square (RMS) response when that scale is easier to interpret. Table 5 maps these quantities to the two domains.

A.1.3 Scalar representation barrier

The representation result below supports the scalar-control limitation in Section 3.

The objective in Section A.1.2 requires useful change and protection within the same endpoint. Before analyzing the update itself, we ask whether a shared route can express those different permissions. We quantify the closest scalar approximation to the authorization mask. For the minimal mixed mask $\boldsymbol{a} = (1,0)$, every scalar route has the form (γ,γ). For a general fixed input, let $w = |W(x)|$ and $p = |P(x)|$ count writable and protected coordinates. We abbreviate the endpoint width as $K = K_x = w + p$.

Proposition 1 (Scalar representation gap) *If $w > 0$ and $p > 0$, no scalar route $\gamma\mathbf{1}$ equals the binary authorization mask $\boldsymbol{a}$. Its minimum squared representation error is*

$$\min_{\gamma\in[0,1]} \|\gamma\mathbf{1} - \boldsymbol{a}\|_2^2 = \frac{wp}{K}, \qquad \gamma_\star = \frac{w}{K}, \tag{8}$$

where $\gamma_\star$ is the best shared route value.

The objective is $w(\gamma-1)^2 + p\gamma^2$. Differentiation gives $\gamma_\star = w/K$, and substitution gives Eq. (8). The best scalar therefore compromises between opening writable outputs and closing protected outputs. For the mask $(1,0)$, this compromise is $(1/2, 1/2)$ with squared error $1/2$. The positive gap depends on the permissions alone. It motivates separate controls for authorization units before considering the magnitude of an update.

This representation problem connects to the selection rule developed in Section 5. Its final line previews the decision: require sufficient utility evidence and then choose the smallest protected-response bound.

The setup separates permissions from responses, and the next section uses these objects to derive the protection attainable under shared controls.

A.2 Structural Results and Design Requirements

The derivations below support the structural guarantees in Section 4.

We connect the representation gap in Eq. (8) to protected response. We begin with the exact cost of broadcasting one route value across a fixed update. We then compare fitted directions at matched utility and derive the best protection achievable by each grouping of outputs. Finally, we study repeated updates and translate the resulting bounds into design requirements.

A.2.1 Exact max-union broadcast excess

The broadcast derivation below supports the protected-response identity in Section 4.

Equation (8) establishes a mismatch in permissions without measuring its effect on model outputs. To measure that effect, we compare fine and scalar routing of the same response. Let $\boldsymbol{r}(x)$ be a fine route and define its *max-union* scalar as $u(x) = \max_j r_j(x)$. Broadcasting this maximum keeps the strongest route value and applies it to every endpoint coordinate. For one fixed input, we suppress x and write $\boldsymbol{d} = \boldsymbol{d}_{\boldsymbol{\Delta}}(x)$ and $K = K_x$.

Proposition 2 (Exact broadcast excess) *For a fine route $\boldsymbol{r} \in [0,1]^K$ and a fixed response $\boldsymbol{d} \in \mathbb{R}^K$, the excess squared response from max-union routing is*

$$\|u\boldsymbol{d}\|_2^2 - \|\boldsymbol{r} \odot \boldsymbol{d}\|_2^2 = \sum_{j=1}^{K} (u^2 - r_j^2) d_j^2 \geq 0, \tag{9}$$

where $u = \max_j r_j$. The identity also holds after restricting the sum to protected coordinates.

Expanding the squared norms proves the identity. Each summand is nonnegative because $u \geq r_j$. Broadcasting therefore adds response at coordinates whose fine route value is below the maximum. Fine routing controls this excess by reducing the values assigned to protected outputs.

To connect this identity to update directions, we express the local protected response as a quadratic form. Let $J_P(x)$ be the Jacobian restricted to protected coordinates and let $\boldsymbol{q}_P(x)$ be the corresponding route subvector. The protected response matrix is

$$C_q = \mathbb{E}_{x \sim \mathcal{D}_P}\left[\frac{1}{|P(x)|} J_P(x)^\top \operatorname{Diag}\big(\boldsymbol{q}_P(x)^2\big) J_P(x)\right], \tag{10}$$

where Diag places its vector argument on the diagonal and the vector square is coordinate-wise. For a local update $\boldsymbol{\Delta}$, the unit-multiplier response is $\boldsymbol{\Delta}^\top C_q \boldsymbol{\Delta}$. The matrix combines the endpoint's sensitivity with the protection assigned by the route.

Applying Eq. (9) to this quadratic form gives the matrix comparison

$$\begin{aligned} C_u - C_r = \mathbb{E}_{x \sim \mathcal{D}_P}\Bigg[\frac{1}{|P(x)|} J_P(x)^\top \\ \operatorname{Diag}\big(u(x)^2 \mathbf{1} - \boldsymbol{r}_P(x)^2\big) J_P(x)\Bigg] \succeq 0, \end{aligned} \tag{11}$$

where C_u and C_r correspond to max-union and fine routing. The notation $\succeq 0$ means positive semidefinite. Thus fine routing produces no larger local protected response for any shared update direction. This conclusion allows arbitrary correlations among update coordinates.

The excess in Eq. (9) is zero precisely when $r_j = u$ at every coordinate with $d_j \neq 0$. For the protected version, only responding protected coordinates enter this condition. The fine and scalar routes therefore coincide in response when every responding coordinate receives the maximum route value. The difference grows when a small open subset coexists with responding coordinates assigned smaller values.

A binary route gives a simple width-dependent example. Let $k_{\text{open}} > 0$ coordinates be open and let every coordinate have squared response $d_0^2 > 0$. The fine and union squared responses are $k_{\text{open}} d_0^2$ and $K d_0^2$. Their ratio is K/k_{open}. This special case predicts larger amplification for wider endpoints or sparser authorization, which we examine in Section 6.2.

A.2.2 MATCHED UTILITY AND AUTHORIZATION EFFICIENCY

The efficiency derivation below supports the matched-utility comparison in Section 4.

Equation (9) compares routes applied to the same response. When each interface fits its own update direction, useful improvement also changes. We therefore compare protected response at a common utility target. The following local model identifies the quantity that governs this comparison.

Let $\boldsymbol{\Delta}_q$ be the fitted direction associated with route q. We define its useful-benefit vector $\boldsymbol{b}_q$ through the first-order expansion

$$U(q, s \mid \boldsymbol{\Delta}_q) - U(q, 0 \mid \boldsymbol{\Delta}_q) = s\boldsymbol{b}_q^\top \boldsymbol{\Delta}_q + o(s\|\boldsymbol{\Delta}_q\|_2)\,, \tag{12}$$

where $\boldsymbol{b}_q$ gives the sensitivity of utility to the declared update coordinates. The remainder is of smaller order as the scaled update approaches zero. We call $A_q(\boldsymbol{\Delta}_q) = \boldsymbol{b}_q^\top \boldsymbol{\Delta}_q$ the useful alignment. A positive alignment means that moving along $\boldsymbol{\Delta}_q$ increases utility to first order.

Within this linear utility model, a direction with $A_q(\boldsymbol{\Delta}_q) > 0$ reaches target t at multiplier $s_q = t/A_q(\boldsymbol{\Delta}_q)$. Substituting this scale into the local quadratic response gives

$$\begin{aligned} R_q(t, \boldsymbol{\Delta}_q) &= t^2 \frac{\boldsymbol{\Delta}_q^\top C_q \boldsymbol{\Delta}_q}{(\boldsymbol{b}_q^\top \boldsymbol{\Delta}_q)^2} = \frac{t^2}{\eta_q(\boldsymbol{\Delta}_q)}, \\ \eta_q(\boldsymbol{\Delta}_q) &= \frac{(\boldsymbol{b}_q^\top \boldsymbol{\Delta}_q)^2}{\boldsymbol{\Delta}_q^\top C_q \boldsymbol{\Delta}_q}, \end{aligned} \tag{13}$$

where $R_q(t, \boldsymbol{\Delta}_q)$ is protected response at the target and η_q is *authorization efficiency*. The numerator measures useful alignment squared, while the denominator measures unit-scale protected response. For positive alignment with zero response, we use $\eta_q = +\infty$ and $R_q = 0$. Larger efficiency therefore means less protected change at the same useful improvement. The broadcast identity controls the denominator for a shared update. Comparing independently fitted directions also requires their useful alignments.

The same ratio identifies the best direction in an unrestricted local update space. For $C_q \succ 0$ and $\boldsymbol{b}_q \neq 0$, Cauchy–Schwarz gives

$$\max_{\boldsymbol{\Delta} \neq 0} \eta_q(\boldsymbol{\Delta}) = \boldsymbol{b}_q^\top C_q^{-1} \boldsymbol{b}_q, \qquad \boldsymbol{\Delta}_\star \propto C_q^{-1} \boldsymbol{b}_q, \tag{14}$$

where $\boldsymbol{\Delta}_\star$ denotes an efficiency-maximizing direction. We choose its sign to give positive useful alignment. The inverse response matrix favors useful directions with small protected effects. When C_q is positive semidefinite and $\boldsymbol{b}_q$ lies in its range, the Moore–Penrose inverse gives the corresponding value. If $\boldsymbol{b}_q = 0$, every direction has zero first-order utility.

For routes q and r with positive useful alignments that reach the same target, Eq. (13) gives the comparison

$$R_r(t, \boldsymbol{\Delta}_r) < R_q(t, \boldsymbol{\Delta}_q) \quad \Longleftrightarrow \quad \eta_r(\boldsymbol{\Delta}_r) > \eta_q(\boldsymbol{\Delta}_q), \tag{15}$$

where each interface uses its own fitted direction. The better interface is the one with higher useful alignment relative to its protected response. This criterion motivates comparisons at matched utility and a permutation control that changes coordinate alignment while preserving route values. The selector in Section 5.2 evaluates finite-target utility directly.

A.2.3 AN EFFICIENCY CEILING FOR COARSE INTERFACES

The ceiling proof below supports the protection benchmark in Section 4.

Equation (13) compares particular fitted interfaces. To determine what granularity alone permits, we now optimize over every route with the same grouping of outputs. We keep the response fixed and derive the best possible allocation of shared strengths.

Fix one input with response $\boldsymbol{d}$ and a local benefit vector $\boldsymbol{b}$ supported on W. Here b_j weights the utility contribution of output change d_j. Partition the endpoint coordinates into cells $g \in G$. We call a route G-measurable when its values are constant within each cell. A scalar route uses one cell for

the whole input. A DocRE pair route uses one cell per directed pair, and an ImageNet class route uses one cell per selected coordinate. Within a cell, the useful slope and summed protected mass are

$$B_g = \sum_{j \in W \cap g} b_j d_j, \qquad P_g = \sum_{j \in P \cap g} d_j^2, \tag{16}$$

where B_g measures useful change at unit strength and P_g measures the accompanying protected change. A cell with $B_g > 0$ contributes useful improvement. If it also has $P_g = 0$, opening it changes no protected output.

Let $m = |P| \geq 1$ be the number of protected coordinates. Write $c_g \in [0, 1]$ for a cell's route value and $v_g = sc_g$ for its effective strength. The local utility is $\sum_g v_g B_g$, and the per-input mean squared response is $R = m^{-1} \sum_g v_g^2 P_g$. The next theorem solves this allocation problem for every positive attainable target.

Restatement of Theorem 1. Let $H_G = \sum_{g:\, B_g > 0} B_g^2 / P_g$. A term with $P_g = 0$ reads as $+\infty$, and an empty sum is zero. If $H_G = 0$, no G-measurable route reaches a positive utility target. If $H_G > 0$, consider all G-measurable routes and multipliers $s \geq 0$ with positive useful alignment $\sum_g c_g B_g > 0$. Their authorization efficiency obeys

$$\sup \eta = m H_G = m \sum_{g:\, B_g > 0} \frac{B_g^2}{P_g},$$

where positive alignment with zero protected response counts as $\eta = +\infty$. If $0 < H_G < \infty$, every such route reaching target $t > 0$ has $R \geq t^2/(mH_G)$. Equality holds for $v_g^\star = tB_g/(H_G P_g)$ on cells with $B_g > 0$ and $v_g^\star = 0$ on other cells. These strengths are realized by $s = \max_g v_g^\star$ and $c_g = v_g^\star / s$. If $H_G = +\infty$, opening one cell with $B_g > 0$ and $P_g = 0$ at $s = t/B_g$ reaches the target with zero response. The substitution $v_g = sc_g$ maps feasible route strengths onto the nonnegative orthant. For positive utility and finite H_G, cells with $B_g \leq 0$ add nonnegative response and no positive utility. Cauchy–Schwarz therefore gives

$$\Big(\sum_g v_g B_g \Big)^2 \leq \Big(\sum_{g:\, B_g > 0} v_g B_g \Big)^2 \leq H_G \sum_g v_g^2 P_g = m H_G R,$$

where the first inequality removes nonpositive useful contributions and the second weights the positive slopes by protected mass. Thus $\eta \leq mH_G$, with equality at the stated strengths. The construction using a cell with zero protected mass proves the zero-response case.

Theorem 1 states that optimal shared controls allocate useful change in proportion to each cell's useful slope relative to its protected mass. It specializes Cauchy–Schwarz to the permission-constrained allocation problem and supplies an attaining route. For fixed responses, the grouping determines which useful and protected contributions must receive the same strength. Refining that grouping can separate these contributions.

The positive-alignment condition makes the ceiling correspond to the positive utility target. For comparisons that allow either sign, the squared numerator in Eq. (13) also counts negative alignment. Define $H_G^+ = H_G$ and $H_G^- = \sum_{g:\, B_g < 0} B_g^2 / P_g$ with the same zero-denominator convention. If at least one $B_g \neq 0$, the two-sided efficiency supremum is $m \max(H_G^+, H_G^-)$. Nonzero alignment with zero protected response again gives infinite efficiency. The positive-target floor remains $t^2/(mH_G)$ because reaching $t > 0$ requires positive alignment. Appendix G compares these two ceilings on units with negative alignment.

Merging cells can only decrease or preserve the sum in Eq. (2). Conversely, partition refinement preserves or increases H_G and preserves or lowers the minimum response at an attainable positive target. A coordinate-wise route has an infinite ceiling whenever a writable coordinate has $b_j d_j > 0$. For the DocRE pair route, a writable pair with positive net slope B_g has zero protected mass because all its margin coordinates are writable. This is the pair-level counterpart of the free-coordinate case.

The minimum response at an attainable positive target is strictly positive exactly when every cell with $B_g > 0$ also has $P_g > 0$. For a document with positive net utility slope, the document-scope minimum is zero exactly when its protected margins have zero response. The mixed-permission

frequency in Section 6.2 describes where shared permissions arise. The quantities B_g and P_g determine their response cost for a given update.

To examine attainment of the bound, we compare its closed-form minimum with direct numerical minimization over the route box. Across 2,000 random cell instances, their relative difference is below 10^{-12}. On 10,000 additional instances, sampled routes respect the floor to numerical precision. Across 2,000 nested partition pairs, finer partitions preserve or lower the response floor. These comparisons support the predicted optimum and its ordering under refinement.

A.2.4 Sequential updates deplete coarse interfaces

The depletion proof below supports the sequential guarantee in Section 4.

Theorem 1 describes the best protection for one update. Successive updates can change the protected mass within each shared-control cell. We study this effect by holding useful slopes fixed while allowing protected mass to grow.

In a write-once-then-protect workload, units installed in earlier rounds become protected in later rounds. Consider rounds $\tau = 1, \dots, T$ over a fixed cell structure with slopes B_g. We model the responding protected mass as $P_g^{(\tau)} = P_g^{(0)} + (\tau - 1)\pi_g$ with $\pi_g > 0$. Here $P_g^{(0)}$ is the initial protected mass and π_g is its increase per round. Fixed slopes describe continued installation through cells that already contain protected outputs. This model isolates the effect of growing protection requirements under cell reuse.

Restatement of Theorem 2. Let $1 \le m_\tau \le K < \infty$ be the protected coordinate count in round τ. Assume K is independent of T and at least one cell has $B_g > 0$. Under the stated growth model, the positive-target ceiling $m_\tau H_+^{(\tau)}$ is governed by

$$\begin{aligned} H_+^{(\tau)} &= \sum_{g:\, B_g > 0} \frac{B_g^2}{P_g^{(0)} + (\tau - 1)\pi_g} \\ &\le \frac{C_+}{\tau - 1} \quad (\tau \ge 2), \\ C_+ &= \sum_{g:\, B_g > 0} \frac{B_g^2}{\pi_g}, \end{aligned} \tag{17}$$

where $H_+^{(\tau)}$ is the positive-slope cell sum and C_+ is its growth constant. For $\tau \ge 2$, the minimum response at a fixed target $t > 0$ obeys $R^{(\tau)} \ge t^2/(m_\tau H_+^{(\tau)}) \ge t^2(\tau - 1)/(KC_+)$. Under a per-round response budget ρ, every achievable positive target satisfies $t_\tau \le \sqrt{\rho m_\tau H_+^{(\tau)}} \le \sqrt{\rho K C_+/(\tau - 1)}$. Fix K, ρ and C_+ independently of the horizon. For a fixed initial target t_1, cumulative installed utility obeys $\sum_{\tau=1}^{T} t_\tau \le t_1 + 2\sqrt{\rho K C_+ (T - 1)} = O(\sqrt{T})$. Allowing either alignment sign gives $m_\tau \max(H_+^{(\tau)}, H_-^{(\tau)}) \le K \max(C_+, C_-)/(\tau - 1)$ for $\tau \ge 2$. Here $H_-^{(\tau)}$ and C_- replace B_g by $-B_g$ in Eq. (17). A singleton-cell route sustains a positive target with zero floor while a positively aligned writable coordinate remains free of responding protected mass. The positive-target bounds follow by substituting the growth model into Theorem 1 and using $m_\tau \le K$. Summing $(\tau - 1)^{-1/2}$ gives the cumulative bound. Applying the same argument to B_g and $-B_g$ proves the two-sided statement. Thus reusing cells with growing protected mass increases the minimum response required for a fixed useful target. With a fixed response budget, the attainable per-round target decreases at rate $O(\tau^{-1/2})$.

The fixed-slope assumption isolates growth in the denominator of the efficiency ratio. The bound on K keeps endpoint expansion from offsetting that growth. The scale of π_g against $P_g^{(0)}$ determines how quickly the growth term dominates within a finite horizon. A fine interface retains a zero floor in rounds with a positive-slope writable cell and zero protected mass. This condition explains the design value of separating newly writable outputs from previously protected outputs.

To compare the predicted decay with direct optimization, we use 3 simulated schedules of 512 rounds. At the evaluated rounds, minimizing $\sum_g c_g^2 P_g^{(\tau)}/(\sum_g c_g B_g)^2$ over positive-alignment routes

agrees with $1/H_+^{(\tau)}$. The fitted exponents are -1.00 for $H_+^{(\tau)}$ and -0.50 for $\sqrt{H_+^{(\tau)}}$. They match the respective theoretical rates -1 and $-1/2$ at the reported precision. For a fixed protected count m and budget ρ, the attainable target $\sqrt{\rho m H_+^{(\tau)}}$ has the latter exponent. Table 22 reports replays on additional task endpoints.

Theorem 2 attributes the decay to shared controls under the stated growth model. This complements studies of degradation from sequential weight-space editing (Fang et al., 2025) by isolating the contribution of the output interface. Theorem 3 applies to selection within each round.

A.2.5 Requirements induced by the analysis

The design implications below support the structural interpretation in Section 4.

Equation (17) shows how accumulated protected response changes the efficiency available to a shared interface. The remaining design question is which quantities routing can control. For a fixed update and policy, the coordinate responses and permissions are given. We can choose how outputs share controls and how strongly each control applies.

Equation (8) requires an interface that can represent the authorization unit. Equation (9) then identifies the avoidable response caused by broadcasting a larger route value to protected coordinates. AGER addresses these effects by assigning route values to class scores or entity pairs.

Equation (2) makes the role of granularity explicit. The partition determines which useful slopes B_g share controls with protected masses P_g. Refinement can increase H_G, while the route values allocate strength within the chosen partition. For independently fitted directions, Eq. (13) additionally makes useful alignment part of the comparison. We therefore evaluate the resulting interfaces at a common utility target.

Finally, the population quantities U and R must be estimated from finite samples. We fit candidates before evaluating these samples and estimate utility separately from protected response. The selection rule uses simultaneous bounds over the candidate family and returns no update when no utility lower bound reaches the target. The next section combines authorization-unit routing with this evidence-based choice of interface and scale.

B Construction Details for Authorization-Granular Routing

This section supports the routing and selection mechanisms in Section 5. It specifies the domain constructions and proves Theorem 3.

B.1 Authorization-Granular Endpoint Routing

The construction details below support the routing method in Section 5.

We now construct authorization-granular endpoint routing (AGER) from the requirements in Section A.2.5. We describe the complete procedure before instantiating its routes for class scores and entity pairs. We then construct matched controls and select an interface and scale using utility and protected-response bounds. The final subsection accounts for fitting, selection and endpoint computation.

B.1.1 Framework overview

The construction overview below supports the routing method in Section 5.

Section A.2.5 identifies separate output controls and matched utility as the central design requirements. Implementing them requires a route that estimates each unit's permission from information available at prediction time. The key idea is to scale each authorization unit's endpoint change using a permission score estimated from the base model.

We begin with the endpoint and authorization labels supplied by the task. We fit a route from base-model scores, then construct the update response that it will control. For ImageNet, we fit a direction for each interface in a common update space. For DocRE, we use the exact difference between the base and updated checkpoints. We combine the fitted interfaces with candidate multipliers and evaluate utility and protected response on separate support samples. We select the candidate with the smallest response upper bound among those whose utility lower bounds reach the target. If none qualifies, we return the base endpoint. We then evaluate the selected decision on held-out confirmation data.

The data roles follow these dependencies. Route fitting uses base scores and authorization labels. Direction training fits the ImageNet update after constructing the route. Utility support and protected support estimate the two selection objectives on disjoint samples. Confirmation data evaluate the resulting decision independently of fitting and selection. These roles keep the candidate construction independent of the observations used to bound its performance.

A fully fitted interface q includes its route and endpoint response. For ImageNet, it also includes its fitted direction $\boldsymbol{\Delta}_q$. A candidate $c = (q, s)$ pairs this interface with the multiplier in Eq. (5). For fixed-response comparisons, fine and max-union routes share $\boldsymbol{d}_{\boldsymbol{\Delta}}$ as required by Eq. (9). For fitted comparisons, we hold the update space and optimization budget constant and compare interfaces at the same useful target. This distinction connects the structural analysis to the domain constructions below.

B.1.2 ImageNet class-coordinate routing

The image-route construction below supports the class-coordinate method in Section 5.

The framework in Section 5 leaves the authorization unit and response construction task dependent. For image classification, we must identify which class score may change and construct a route for that score. We use the image's class label to define authorization and the base logits to estimate it.

Let $\mathcal{K}_{\mathrm{upd}}$ be the selected update classes and let $K_I = |\mathcal{K}_{\mathrm{upd}}|$. The endpoint consists of these selected logits within the fixed 1000-class classifier, so $K_x = K_I$. An image labeled $j \in \mathcal{K}_{\mathrm{upd}}$ authorizes coordinate j and protects the other selected coordinates. An image labeled outside $\mathcal{K}_{\mathrm{upd}}$ has an all-zero authorization mask. Logits outside the selected endpoint remain unchanged. Thus one image-class coordinate is an authorization unit, and s scales the response across the selected coordinates.

Constructing a shared update space. Comparing routing interfaces requires a common set of directions in which their logits can change. We keep the backbone and classifier fixed and fit coefficients in a shared feature-update basis. Let $\boldsymbol{h}(x) \in \mathbb{R}^D$ be the backbone feature and let $\boldsymbol{w}_j^\top$ be classifier row j. We construct an orthonormal basis $B \in \mathbb{R}^{D\times d}$ from a Gaussian matrix using a fixed model-specific random draw. The subspace dimension $d = 32$ controls the number of trainable coefficients, as specified in Section 6. For coefficients $\boldsymbol{v} \in \mathbb{R}^d$, the feature change and its selected-logit response are

$$\delta\boldsymbol{h}_{\boldsymbol{v}}(x) = \boldsymbol{h}(x) \odot B\boldsymbol{v}, \qquad d_j(x, \boldsymbol{v}) = \boldsymbol{w}_j^\top \delta\boldsymbol{h}_{\boldsymbol{v}}(x), \tag{18}$$

where $\delta\boldsymbol{h}_{\boldsymbol{v}}$ is a multiplicative feature change and d_j is its effect on class logit j. We embed only selected-logit responses into the complete classifier output. This construction lets every interface optimize within the same update space. The fitted coefficients determine useful alignment for each interface within that shared update space.

Estimating permissions from base scores. The route must assign comparable strengths to different class coordinates. Raw margins can have different scales, so we fit a balanced Platt calibrator that maps each margin through a logistic sigmoid. Let $\boldsymbol{z}(x) \in \mathbb{R}^{1000}$ be the complete base-logit vector. For selected class j, define its top margin as $g_j(x) = z_j(x) - \max_{k\neq j} z_k(x)$. A positive margin means that the base model ranks j above every other class. We turn this score into a permission estimate and route value through

$$\begin{aligned} \pi_j(x) &= \sigma\left(\alpha_j \frac{g_j(x) - \mu_j}{\kappa_j} + \beta_j\right), \\ r_j(x) &= r_{\min} + (1 - r_{\min})\pi_j(x)^{p_{\text{route}}}, \end{aligned} \tag{19}$$

where σ is the logistic sigmoid and $\pi_j(x)$ is the fitted permission score. The parameters μ_j and $\kappa_j > 0$ center and scale the margin. We fit a nonnegative slope α_j and an intercept β_j. The route floor $r_{\min} \in [0, 1]$ retains a minimum fraction of the response, and the power $p_{\text{route}} > 0$ controls concentration. A score approaching zero gives a route approaching $r_{\min}$, while a score approaching one gives a route approaching one. With $r_{\min} = 1$, all route values equal one and the full response is applied at unit multiplier.

The calibrator predicts authorization labels from base scores. We give writable and nonwritable fitting members equal total weight so class prevalence does not set the route scale. We use the fitted scores as update strengths rather than as replacement class predictions. Section 6 gives the fitting data and the chosen floor and powers.

To examine the role of this fitted mapping, we also construct a route directly from the base logits:

$$r_j^{\text{raw}}(x) = r_{\min} + (1 - r_{\min}) \left[\exp\left(z_j(x) - \max_k z_k(x)\right)\right]^{p_{\text{raw}}}, \tag{20}$$

where $p_{\text{raw}} > 0$ sets the concentration of the uncalibrated route. The exponential factor equals one for a top-scoring class and decreases for lower-scoring classes. This control shares the base logits and route floor with Eq. (19). When the powers differ, the comparison measures the combined effect of calibration and concentration.

Fitting the useful direction. Equation (13) shows that route values affect both protection and useful alignment. We therefore fit a coefficient vector separately for each interface while keeping the update space fixed. For direction-training example i, let $J_i \in \mathbb{R}^{K_I\times d}$ map coefficients to selected-logit changes. Write $\boldsymbol{z}_i$ for its complete base logits and $\boldsymbol{q}_i = \boldsymbol{q}(x_i)$ for its route. We minimize

$$\min_{\|\boldsymbol{v}_q\|_2 \le R_v} \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{CE}}\big(\boldsymbol{z}_i + \iota_{\mathcal{K}_{\text{upd}}}(\boldsymbol{q}_i \odot J_i\boldsymbol{v}_q), y_i\big) + \frac{\lambda}{2}\|\boldsymbol{v}_q\|_2^2, \tag{21}$$

where n is the direction-training sample size and y_i is the class label. The operator $\iota_{\mathcal{K}_{\text{upd}}}$ embeds selected responses into the complete logit vector. The loss ℓ_{CE} is complete 1000-class cross-entropy. Its data term fits the desired class predictions, while the L2 term with weight λ controls coefficient magnitude. The radius R_v bounds the coefficient norm. We use the same basis and training members for every interface. The regularization and projected optimization settings are also shared, with values given in Section 6. Candidate multipliers are applied after fitting $\boldsymbol{v}_q$. This yields a class-coordinate route and its fitted response for the selector.

B.1.3 DocRE directed-pair routing

The pair-route construction below supports the document endpoint method in Section 5.

Equation (19) assigns one strength to each selected class score. For DocRE, several relation scores share the permission of their entity pair. We adapt the same construction by estimating one route value per directed pair and applying it to every relation margin in that pair.

For a document x, let $\mathcal{P}_x$ be its directed entity pairs and let K_R count non-null relation labels. The endpoint contains all non-null relation margins, so $K_x = |\mathcal{P}_x|K_R$. A pair is writable when the revised annotation adds at least one relation for it. All its margin coordinates then belong to $W(x)$, while coordinates of nonwritable pairs belong to $P(x)$. We measure utility only on relation coordinates that were actually added. Among nonwritable pairs, a retained-relation pair is a pure protected positive and a pair without a revised positive relation is a revised negative. Both remain protected under the supplied policy.

To estimate pair permissions, we use the base model's strongest non-null margin. Let $L_{ef,\ell}(x)$ denote the base logit of relation ℓ for pair (e, f). The index $\ell = 0$ denotes the no-relation label (NA). We define the endpoint margin and the pair score as

$$\begin{aligned} m_{ef,\ell}(x) &= L_{ef,\ell}(x) - L_{ef,0}(x), \\ g_{ef}(x) &= \max_{\ell \neq 0} m_{ef,\ell}(x), \end{aligned} \tag{22}$$

where $m_{ef,\ell}$ compares a relation with NA and g_{ef} selects the largest such comparison. A positive pair score means that at least one relation logit exceeds the NA logit. This score uses the same base endpoint that will later receive the routed update.

We use the balanced logistic mapping from Section 5.1 to estimate writability from this pair score. The pair route is

$$\begin{aligned} \pi_{ef}(x) &= \sigma\left(\alpha \frac{g_{ef}(x) - \mu}{\kappa} + \beta\right), \\ r_{ef}(x) &= r_{\min} + (1 - r_{\min})\pi_{ef}(x), \end{aligned} \tag{23}$$

where μ and $\kappa > 0$ center and scale the score. The nonnegative slope α and intercept β are shared across pairs. The score π_{ef} estimates pair writability, and $r_{\min}$ has the same route-floor role as in Eq. (19). Writable and nonwritable pairs receive equal total fitting weight. Repeating r_{ef} across the pair's non-NA coordinates makes the route index agree with the authorization unit. This is the same logistic route construction with a shared pair calibrator and unit power.

For this endpoint, the update is supplied by two trained checkpoints. We compute the exact difference of their relation margins:

$$d_{ef,\ell}(x) = \left[L^{\mathrm{upd}}_{ef,\ell}(x) - L^{\mathrm{upd}}_{ef,0}(x)\right] - \left[L^{\mathrm{base}}_{ef,\ell}(x) - L^{\mathrm{base}}_{ef,0}(x)\right], \tag{24}$$

where the superscripts identify updated and base checkpoint logits. The two bracketed terms are their respective margins against NA. AGER adds $sr_{ef}(x)d_{ef,\ell}(x)$ to each base margin. It therefore assigns separate strengths to pairs within the same document while sharing the global multiplier s. When all pair route values coincide, this operation reduces to document-level scalar routing. The construction changes the authorization unit and response source while retaining the selection objective of Section A.1.2.

B.1.4 Matched route family and controls

The candidate definitions below support the interface comparison in Section 5.

The ImageNet class routes and DocRE pair routes in Section 5.1 assign strengths to authorization units. Their protected response can differ because of granularity or simply because they apply less total change. To distinguish these effects, we construct controls from the same fine route and evaluate their useful alignment at matched utility.

Write $\boldsymbol{q}_{\mathrm{fine}} = \boldsymbol{r}$ for the fine route. The identity route $\boldsymbol{q}_{\mathrm{id}} = \mathbf{1}$ applies the complete response at unit multiplier. The max-union route $\boldsymbol{q}_{\mathrm{union}} = (\max_j r_j)\mathbf{1}$ broadcasts the strongest route value as in

Eq. (9). To separate route granularity from overall squared exposure, we also use

$$\boldsymbol{q}_{\mathrm{L2}}(x) = \sqrt{\frac{1}{K_x}\sum_{j=1}^{K_x} r_j(x)^2}\,\mathbf{1}, \tag{25}$$

where the scalar is the RMS of the fine route values within the input. Its repeated values have the same summed square as the fine route. Thus this control preserves total squared route exposure while sharing one value across outputs.

A deterministic within-input permutation preserves all fine route values and changes their coordinate alignment. It tests whether matching route strengths to the intended units matters beyond their distribution. For DocRE, we construct max-union and exposure-matched routes within each document and permute values between pairs in that document. This preserves the input boundary and the shared permission of each pair's coordinates.

The raw route in Eq. (20) compares fitted calibration with a direct score mapping. An authorization-mask Oracle route uses $\boldsymbol{a}(x)$ itself to measure the capacity available with known permissions. This Oracle is a comparison reference rather than a deployable candidate. Native task updates and external editing methods form separate baselines, as specified in Section 6. The scalar controls recover shared endpoint scaling by setting every route value within an input equal.

Let $\mathcal{C}$ be the deployable family of candidates $c = (q, s)$ supplied to selection. Each ImageNet interface includes its fitted direction, while the DocRE interfaces share the checkpoint response in Eq. (24). The family combines eligible interfaces with a fixed multiplier grid. Mechanism-only identity and permutation controls stay outside this family. We next choose among its candidates while accounting for sampling uncertainty.

B.1.5 Utility-qualified finite-grid selection

The selection algorithm and proof below support Theorem 3.

Section 5.1 gives a finite candidate family with different utility and protected response. Choosing the most favorable sample estimate also selects its estimation error. We therefore bound all candidates simultaneously before requiring utility and minimizing protected change. We use Hoeffding's inequality (Hoeffding, 1963) to control bounded sample means and a union bound to cover the family. This follows the fit-then-evaluate separation used in risk-controlled selection (Angelopoulos et al., 2025). Here the decision combines a lower bound on useful improvement with an upper bound on protected change.

Let $M = |\mathcal{C}|$ count the fully fitted interface–multiplier pairs constructed before support evaluation. For candidate c, utility-support unit i gives $X_{c,i} \in [-B_U, B_U]$ with mean U_c. Protected-support unit i gives $Y_{c,i} \in [0, B_R]$ with mean R_c. Here B_U and B_R are fixed observation bounds. With clipped observations, these means refer to the corresponding clipped utility and response statistics specified in Section 6. We use disjoint utility and protected supports with n_U and n_P independent units. Their means are

$$\widehat{U}_c = \frac{1}{n_U}\sum_{i=1}^{n_U} X_{c,i}, \qquad \widehat{R}_c = \frac{1}{n_P}\sum_{i=1}^{n_P} Y_{c,i}, \tag{26}$$

where the hats denote estimates of the candidate's population utility and response. We aggregate dependent coordinates within each observation before forming these averages. For example, we use complete DocRE documents as independent units.

Let $\delta \in (0, 1)$ be the total error probability. We set the observation bounds and any clipping rules before reading support outcomes. Hoeffding's inequality gives the simultaneous radii

$$\epsilon_U = B_U\sqrt{\frac{2\log(4M/\delta)}{n_U}}, \qquad \epsilon_R = B_R\sqrt{\frac{\log(4M/\delta)}{2n_P}}, \tag{27}$$

where ϵ_U and ϵ_R account for utility and response estimation error across the complete grid. Larger supports reduce these radii, while more candidates increase them through $\log M$. The different leading factors reflect the range lengths $2B_U$ and B_R.

Utility-qualified route and multiplier selection
Input: fixed candidates $\mathcal{C}$ and target t. Fixed bounds B_U, B_R and error probability δ. Disjoint supports and fixed tie order.
1. Evaluate utility on utility support.
2. Evaluate squared response on protected support.
3. Compute the simultaneous radii in Eq. (27).
4. Retain candidates whose utility lower bound reaches t.
5. Return no update if the feasible set is empty.
6. Otherwise minimize the response upper bound. Evaluate the selected decision on held-out confirmation data.

Figure 3: The selector separates fitted candidate construction, support-based choice, and held-out confirmation. Grid size enters the confidence penalty logarithmically.

We admit a candidate only when its utility lower bound reaches the fixed target:

$$\widehat{\mathcal{F}}(t) = \{c \in \mathcal{C} : \widehat{U}_c - \epsilon_U \geq t\}, \tag{28}$$

where $\widehat{\mathcal{F}}(t)$ is the utility-qualified feasible set. Subtracting the radius requires evidence that utility exceeds the target despite estimation error. When this set is nonempty, we choose

$$\widehat{c} = \arg\min_{c \in \widehat{\mathcal{F}}(t)} \left(\widehat{R}_c + \epsilon_R \right), \tag{29}$$

where $\widehat{c}$ is the selected route–multiplier candidate. A fixed order resolves ties. If the feasible set is empty, AGER returns no update. The response upper bound ranks candidates only after their utility lower bounds qualify them. Figure 3 gives the complete selection algorithm.

The algorithm separates the useful-update requirement from the protection objective. The following theorem states what this separation guarantees after choosing a candidate.

Restatement of Theorem 3. Assume the candidate family is fixed before support evaluation. Let observations obey the stated bounds and be independent at the declared domain unit within each support. With probability at least $1 - \delta$, all candidate utility and response means lie within their respective radii. On this event, any returned candidate $\widehat{c}$ satisfies $U_{\widehat{c}} \geq t$. For every $c \in \mathcal{C}$ with $U_c \geq t + 2\epsilon_U$, the feasible set is nonempty and

$$R_{\widehat{c}} \leq R_c + 2\epsilon_R, \tag{30}$$

where $R_{\widehat{c}}$ is the population response of the selected candidate.

Hoeffding bounds for the two bounded variables and a union bound over $2M$ deviation events establish the simultaneous statement. The utility range has length $2B_U$, and the response range has length B_R. Allocating probability $\delta/(2M)$ to each event gives Eq. (27). On the simultaneous event, a candidate with $U_c \geq t + 2\epsilon_U$ enters the feasible set. The selected utility lower bound also implies $U_{\widehat{c}} \geq t$. Empirical optimality then gives the response comparison

$$R_{\widehat{c}} \leq \widehat{R}_{\widehat{c}} + \epsilon_R \leq \widehat{R}_c + \epsilon_R \leq R_c + 2\epsilon_R.$$

The outer inequalities apply the simultaneous response bounds, and the middle inequality follows from selection. This proves Eq. (30).

The theorem guarantees the requested population utility whenever the selector returns a candidate. Its response is within $2\epsilon_R$ of every candidate whose true utility exceeds the target by $2\epsilon_U$. The utility margin covers estimation uncertainty in admission, while the response term covers uncertainty in comparing candidates. Candidate independence from support outcomes lets the same simultaneous event remain valid after selection. Candidates may share support examples because the simultaneous bound allows dependence between their estimates. The required independence is between the fitted candidate family and support outcomes, and between observations within each support. The bound compares deployable candidates in $\mathcal{C}$. The authorization-mask Oracle route remains a separate capacity reference. Held-out confirmation then measures the performance of the chosen decision.

B.1.6 IMPLEMENTATION AND COMPLEXITY

The implementation details below support the computational analysis in Section 5.

Equation (29) specifies the decision once candidate observations have been computed. Its practical cost also includes constructing those candidates and applying the selected route. We separate these stages to identify which computations are reused.

Offline construction caches ImageNet backbone features and logits or base-checkpoint DocRE relation logits. Route fitting estimates class-wise Platt parameters or a shared pair calibrator. For ImageNet, we construct the $D \times d$ basis from Eq. (18) and fit one d-dimensional coefficient vector per eligible interface. Cached features are reused across route and multiplier candidates. For DocRE, the endpoint construction requires one base-checkpoint evaluation and one updated-checkpoint evaluation per document. The route itself trains no new encoder.

After candidate observations are available, selection takes $O(M(n_U + n_P))$ scalar accumulations and can be streamed. The computational dependence on M is linear, whereas the confidence penalty depends on $\log M$. At endpoint use, route computation is linear in the selected class width for ImageNet and in the number of directed pairs for DocRE. We report feature extraction and fitting separately from support evaluation and endpoint inference. Measured comparisons use matched devices and batch sizes. They also match numerical precision and cache state. AGER now provides a fitted route and a utility-qualified selection rule, which the next section evaluates through protected response and task-level performance.

Table 6: **The comparisons span different authorization units.** Each row identifies the endpoint and comparison purpose. Independent support images and authorization-class blocks for confirmation distinguish the ImageNet sampling units. Complete documents are the DocRE units. This is a design summary with no estimated statistics.

Environment	Models or endpoints	Authorization unit	Comparison purpose	Support or confirmation unit
Utility-qualified ImageNet	ResNet-50, DenseNet-121, MobileNetV3-Large	image × update-class coordinate	uncalibrated row route versus strongest eligible scalar at targets 0.01 and 0.02	independent support images and authorization-class blocks for confirmation
Native ImageNet	same three pretrained backbones	image × class coordinate	support-matched uncalibrated row route versus ordinary 1000-way linear-head update	authorization class
Portable DocRE	DREEAM BERT-base and ATLOP RoBERTa-large	document × directed pair	pair route versus the support-selected document scalar and permutation	complete document
Native DocRE	DREEAM BERT-base checkpoint pair	document × directed pair	pair-routed response versus revised-checkpoint inference	complete document
Numerical endpoint	fixed finite-response generator	endpoint coordinate	fine, max-union, and route-value-preserving permutation	numerical setting
Additional endpoint CodRED (Yao et al., 2021)	NEPD base checkpoint	bag × relation coordinate	pair route with an unclipped Platt slope versus scalar, union, and permutation	cross-document unit
Additional endpoint CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003)	BERT-base NER	token × label margin	entity-token fine route versus sentence scalar	document
Additional endpoint AG News (Zhang et al., 2015)	RoBERTa-base classifier	example × class margin	row route versus example scalar under the all-class writable map	composite example block
Additional endpoint CIFAR-100 (Krizhevsky, 2009)	ResNet-56, VGG-16-BN, MobileNetV2	image × class margin	per-interface fitted directions at matched utility	class block
CounterFact editor panel	GPT-2 XL with ROME (Meng et al., 2022) and constrained fine-tuning	prompt × vocabulary coordinate	global, input-scope, output-scope, and prompt–token pair routing of each actual edited-minus-base response	CounterFact case

C Experimental Design and Sample Allocation

This section supports the comparisons in Section 6. It records the datasets, sample allocations and comparator settings used to measure utility and protected response.

C.1 Experiments

The allocations and candidate settings below complement the experimental design in Section 6.1.

C.1.1 Evaluation Setup

The sample allocations and settings below support the experimental design in Section 6.1.

Datasets and models. We use ImageNet class scores to compare learned update directions within a common feature space (Deng et al., 2009). ResNet-50 and DenseNet-121 cover different convolutional architectures, while MobileNetV3-Large adds a compact backbone (He et al., 2016; Huang et al., 2017; Howard et al., 2019). We use aligned DocRED and Re-DocRED annotations to identify naturally added relations within otherwise protected documents (Yao et al., 2019; Tan et al., 2022). DREEAM BERT-base and ATLOP RoBERTa-large supply base and revised checkpoints for the pair-response construction in Eq. (24) (Ma et al., 2023; Zhou et al., 2021). Alignment requires identical sentences and canonical entity order, so permissions refer to the same entity pairs. Table 6 specifies the remaining endpoints and independent units used to examine transfer beyond these primary comparisons.

Table 7: **Disjoint samples separate fitting from evaluation.** Values specify the final experimental design. Sample counts refer to the units indicated in each row. Optimization settings are shared within the stated comparison to isolate routing effects. These are configuration values rather than estimated outcomes.

Setting	Values and purpose
Utility-qualified ImageNet classes	20 writable and 20 protected classes, with 48,000 unique images
Writable-class samples	100 route-fitting, 300 direction-training, 500 utility-support and 300 evaluation images per class
Protected-class samples	100 route-fitting, 600 response-support and 500 evaluation images per class
ImageNet input processing	RGB, resize to 256 and center crop to 224×224, with ImageNet normalization
ImageNet candidate family	Four interfaces and 201 scales from 0 to 1 at increments of 0.005, giving 804 candidates per model
ImageNet targets and bounds	$t \in \{0.01, 0.02\}$, utility in $[-0.1, 0.1]$ and normalized squared response in $[0, 1]$
Simultaneous uncertainty	Global error probability 0.05 across three backbones, with $\epsilon_U = 0.0049$ and $\epsilon_R = 0.0225$
ImageNet direction fitting	Float64, $d = 32$, 320 projected steps, $\lambda = 0.02$ and $R_v = 1$
ImageNet route settings	$r_{\min} = 0.02$, $p_{\text{route}} = 1$ and $p_{\text{raw}} = 2$
Native ImageNet samples	17 writable and 17 protected classes, 19,380 development images and 3,400 evaluation images
Native ImageNet selection	121 route scales from 0 to 0.6, with support utility margin 0.004 above native utility
Native ImageNet fitting	Three seeds and three L2 penalties, five epochs and 201 scales, giving 1,809 candidates per architecture
Aligned DocRE splits	2,931 train, 452 development and 443 test documents
DocRE evaluation partition	160 design, 96 route-fitting, 96 support and 91 documents outside development
DocRE model inputs	Maximum sequence length 1024 and 97 relation outputs including NA
DocRE scalar eligibility	At least half the target utility and relative utility gap at most 0.25
DocRE external updates	LoRA (Hu et al., 2022) ranks 4 and 16, projection ranks 8 and 32, with selected ranks 16 and 8
Distilled student (Hinton et al., 2015)	Candidate weights 0.2 and 0.4, selected $\lambda = 0.4$ subject to retaining at least three quarters of pair-route utility
Numerical responses	256 vectors per setting and sixteen seeds, giving 12,000 settings and 3,072,000 vectors
Numerical factor ranges	Width 1 to 32, authorization fraction $1/32$ to 1, mixing 0 to 1 and multiplier 0.25 to 4.0
CounterFact comparisons	8 development cases and 100 evaluation cases per editor, with GPT-2 XL in float32

Additional endpoints. We use CodRED to test controls over cross-document relation evidence and CoNLL to examine token-level permissions (Yao et al., 2021; Tjong Kim Sang & De Meulder, 2003). AG News distinguishes uniform permissions within an example from mixed permissions across grouped examples (Zhang et al., 2015). CIFAR comparisons test separately fitted directions across additional convolutional backbones (Krizhevsky, 2009). These comparisons use constructed update responses so the structural quantities can be evaluated on real endpoint geometry. CounterFact (Meng et al., 2022) complements them with actual weight edits and known prompt–token permissions, separating the output interface from the choice of editor.

Sampling and fitting. Separate samples fit routes and ImageNet directions, estimate the selection objectives and evaluate the selected updates. This separation makes support observations independent of candidate fitting as required by Section 5.2. Table 7 gives sample counts and optimization settings. The shared ImageNet basis and optimization budget isolate the effect of routing within the same trainable space. The route floor retains a minimum response, while the calibrated and raw powers compare permission calibration with direct score concentration. For numerical experiments we use sixteen seeds to vary finite responses, and native ImageNet fitting uses three seeds to vary optimization. Complete documents remain the sampling units for DocRE because their pair coordinates share context.

Comparators. Max-union routing measures the effect of broadcasting the strongest permission score, and the L2 scalar matches total squared route exposure. Within-input permutation preserves route values while testing their alignment with authorization units. The raw route tests the combined effect of calibration and concentration described in Section 5.1. We include SERAC-style routing and AlphaEdit-style projection to compare input selection with parameter-space preservation (Mitchell et al., 2022b; Fang et al., 2025). OGD and EWC test gradient projection and parameter regularization (Farajtabar et al., 2020; Kirkpatrick et al., 2017). L2-SP and LoRA compare preservation around the original weights with low-rank adaptation, while full-head fitting supplies a native update reference (Li et al., 2018; Hu et al., 2022). For DocRE we additionally compare document-scope deployment and protected-gradient head projection. The authorization-mask Oracle uses known permissions as a capacity reference, including the nested masks applied after CounterFact weight edits.

Utility and protection. We measure utility as ImageNet negative log-likelihood benefit or DocRE added-relation margin gain, following Eq. (6). Larger utility is better, while smaller mean squared protected response or its RMS is better. For ImageNet selection the reported means are means of clipped utility and clipped normalized squared response, with ranges in Table 7. Accuracy measures correct class predictions, and prediction-change rates measure disagreement with the base classifier. Added-relation recall measures recovered additions, retained-relation recall measures preserved relations and micro-F1 balances relation precision with recall. Higher accuracy and recall are preferable, as is higher micro-F1. Tables state the numerator of every response ratio so its direction is explicit.

Selection and uncertainty. ImageNet compares route–scale candidates using the simultaneous bounds of Section 5.2. Its native-head comparison uses a separate uncalibrated route matched to native support utility with the stated margin. For DocRE a scalar qualifies when its support utility reaches at least half the target within a relative gap of 0.25, which admits the mean-broadcast scalar at both endpoints. We select the qualifying scalar with the smallest support response. Native DocRE outcomes are threshold-based comparisons under the reporting thresholds listed with their results. We use 20,000 bootstrap replicates to quantify confirmation uncertainty over authorization classes or complete documents. These endpoint-wise intervals preserve within-input dependence, while the simultaneous ImageNet bounds govern selection.

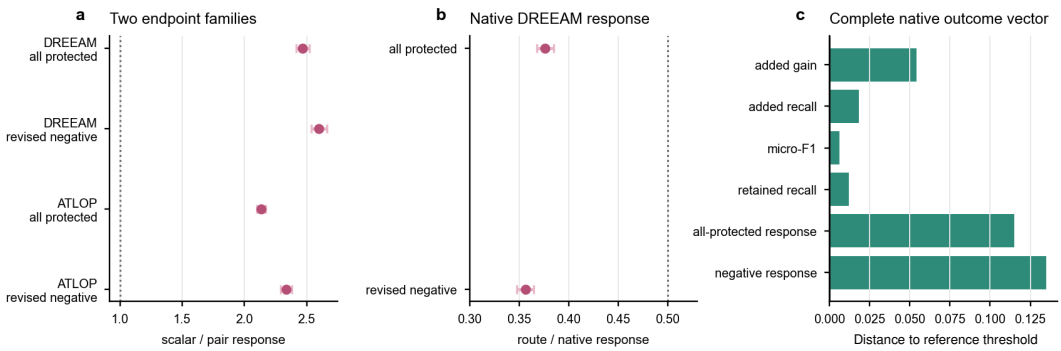


Figure 4: **Lower protected response accompanies a task-level trade-off.** Panel (a) divides document-scalar RMS by pair-route RMS, where larger values favor the pair route. Panel (b) divides pair-route RMS by native DREEAM RMS, where smaller values favor the route. The dotted line is the 0.50 reporting threshold. Error bars are endpoint-wise 95% intervals from 20,000 complete-document bootstrap replicates. Panel (c) plots signed distances of interval endpoints from the task and response thresholds.

D Complete Results for the Document-Level Relation Extraction Endpoint

This section supports the document-level relation extraction results in Section 6.2. It retains the endpoint comparisons and the task-level results for each evaluation population.

DocRE deployment and task outcomes. We evaluate document-level deployment to connect the protected-response objective to the workload described in the introduction. Table 4 shows that switching entire writable documents retains almost all the protected RMS of full redeployment. Pair routing reduces that response while giving up 1.39 micro-F1 points at this deployment scale. The trade-off follows from reducing the strengths applied to relation margins, which can change thresholded predictions. Figure 4 and Table 8 report the corresponding gain and recall comparisons. Table 9 retains the endpoint and subset differences in those task criteria. These results quantify the protection gained by controlling pairs within documents alongside its task-level consequences.

The attainable document-level floor. Theorem 1 gives the smallest mean squared protected response that any assignment of one strength per document can reach at a required utility. We evaluate that floor on the 160 design documents, where the per-document useful slopes and protected masses of the DREEAM endpoint are available. At the utility the pair route attains there, the document-level floor is 0.290 $[0.256, 0.332]$ in mean squared protected response, against 0.077 for the pair route. In the ratio convention of Table 2 the floor is 1.94 times the pair-route RMS, below the 2.467 of the mean-broadcast scalar and above the pair route itself. Refining a document partition into pairs therefore reaches responses that no document-level strength assignment attains, rather than only improving on one particular document-level rule. The interval resamples complete design documents at the observed utility and recomputes the optimal strengths. The remaining documents of the 443-document panel do not carry per-document slopes and masses in the released aggregates, so we report this floor for the design documents alone.

The retained-relation stratum. Pure protected-positive pairs are the one stratum where the pair route responds more than full redeployment. Because the route is fixed, the response of this stratum grows in proportion to the deployment multiplier, and the pair-to-native RMS ratio equals the multiplier times the root mean square of the route values weighted by squared native change. That weighted value is 0.822, so the ratio reaches one at multiplier 1.217 $[1.188, 1.247]$. The multiplier that matches support utility is 1.299, which gives a ratio of 1.067 $[1.041, 1.093]$ on this stratum. Lowering the multiplier to the crossing point would retain 93.7% of the matched support utility. The arithmetic mean route value of 0.903 describes the same exposure but does not determine the crossing point, since the reversal depends on how route values line up with the largest native changes.

Table 8: **The pair route reduces protected RMS within the full-panel task criteria.** This threshold-based DREEAM comparison covers 443 documents. Gain is route/native and task differences are route minus native. Higher gain and task differences are preferable, while lower response ratios are preferable. Brackets are endpoint-wise 95% document-bootstrap intervals. LCB and UCB denote lower and upper confidence bounds.

Outcome	Estimate [interval]	Requirement (task ↑, response ↓)
Added-gain ratio	+0.975 [+0.954, +0.997]	LCB ≥ 0.900
Added-recall difference	-0.022 [-0.032, -0.012]	LCB ≥ -0.050
Micro-F1 difference	-0.014 [-0.019, -0.009]	LCB ≥ -0.025
Retained-recall difference	-0.003 [-0.008, +0.001]	LCB ≥ -0.020
All-protected response ratio	+0.376 [+0.368, +0.385]	UCB ≤ 0.500
Revised-negative response ratio	+0.356 [+0.348, +0.365]	UCB ≤ 0.500

Table 9: **Task effects vary across DocRE endpoints and evaluation populations.** Entries are route-minus-native micro-F1 differences and route/native RMS ratios. Higher differences and lower response ratios are preferable. Brackets are 95% complete-document bootstrap intervals. The reference micro-F1 lower bound is -0.025. The final column lists task criteria whose lower bounds fall below their reference thresholds.

Endpoint	Documents	Micro-F1 difference ↑	RMS ratio ↓	Criteria below reference
DREEAM BERT-base	443	-0.014 [-0.019, -0.009]	0.376 [0.368, 0.385]	None
DREEAM BERT-base	91	-0.015 [-0.026, -0.004]	0.376 [0.356, 0.398]	Micro-F1, retained recall
ATLOP RoBERTa-large	443	-0.022 [-0.027, -0.017]	0.436 [0.426, 0.447]	Added recall, micro-F1
ATLOP RoBERTa-large	91	-0.021 [-0.031, -0.011]	0.438 [0.416, 0.460]	Added recall, micro-F1

Table 10: **AGER reaches the utility target with less protected change.** LCB denotes the simultaneous utility lower confidence bound. Evaluation/target divides held-out utility by the target. Response reduction compares mean squared response with the strongest eligible scalar. Entries summarize each selected update.

Model	Target	Scale	Utility LCB ↑	Evaluation/target ↑	Response reduction ↑
ResNet-50	0.010	0.030	0.0109	1.61	96.8%
ResNet-50	0.020	0.055	0.0202	1.41	95.9%
DenseNet-121	0.010	0.050	0.0113	1.67	97.5%
DenseNet-121	0.020	0.095	0.0207	1.52	97.4%
MobileNetV3-Large	0.010	0.135	0.0103	1.76	98.7%
MobileNetV3-Large	0.020	0.465	0.0200	2.24	99.0%

Table 11: **Holding the update direction fixed leaves the response reduction in place.** Each row compares the fine route with the strongest utility-qualified scalar under the selection rule of Table 1, with one update direction shared by every route. Reduction is the decrease in mean squared protected response and RMS ratio divides fine-route protected RMS by the scalar's, so reduction equals one minus the squared ratio. Larger reductions and smaller ratios favor the fine route. Brackets are endpoint-wise 95% authorization-class bootstrap intervals.

Model	Target	Scalar route	Reduction ↑	RMS ratio ↓ [interval]
ResNet-50	0.01	Exposure-matched	96.07%	0.1983 [0.1562, 0.2383]
ResNet-50	0.02	Max-union	95.95%	0.2012 [0.1612, 0.2422]
DenseNet-121	0.01	Max-union	97.09%	0.1707 [0.1249, 0.2229]
DenseNet-121	0.02	Max-union	96.81%	0.1785 [0.1306, 0.2331]
MobileNetV3-Large	0.01	Max-union	98.77%	0.1110 [0.09673, 0.1306]
MobileNetV3-Large	0.02	Max-union	98.88%	0.1056 [0.09206, 0.1243]

E Complete Results for the Image Endpoint

This section supports the ImageNet results in Section 6.2. It gives the external-method comparisons and the native-risk results at their reported operating points.

Selected ImageNet operating points. Table 10 retains the selected scales alongside the results in Table 1.

ImageNet comparison under one fixed update direction. Table 1 lets each interface fit its own direction, so Table 11 repeats that comparison with one update direction shared by every route. The comparator, the utility target and the selection rule are those of Table 1, and only the direction changes. Across the same six model–target combinations the median reduction in mean squared protected response is 96.95%, and the largest difference from the corresponding entry of Table 1 is 0.731 percentage points. Clipped utility differs from the scalar's by at most 8.66% and is the larger of the two in three of the six combinations, so the reduction does not come from additional utility.

External ImageNet methods. We compare the same protection objective with external update methods to test whether existing preservation mechanisms recover the routing benefit. Table 3 reports method-to-route RMS ratios at the shared utility target. At $t = 0.01$ every external method produces greater protected RMS, with interval lower bounds of at least 3.2 times the route. AlphaEdit-style projection and SERAC-style routing give the closest external points, consistent with their restriction of update directions or input scope. The discrete scale grid causes larger utility overshoot for native fitting and EWC on DenseNet, as quantified in Table 13. The ordering also holds on the other backbones with comparable overshoot, supporting authorization-unit control beyond these existing mechanisms.

At $t = 0.02$ the smallest method-to-route interval lower endpoint is 2.20 (Table 12).

Native ImageNet updating. To compare against ordinary task updating, we evaluate the separately support-matched uncalibrated route against full-head fitting. Table 14 reports absolute utility along-

Table 12: **External ImageNet comparisons at the higher utility target.** Ratios divide method RMS by fine-route RMS at $t = 0.02$. Brackets are endpoint-wise 95% authorization-class bootstrap intervals. Configurations are selected on support data.

Method	ResNet-50 ↑	DenseNet-121 ↑	MobileNetV3-L ↑
SERAC query routing	8.34 [6.87, 10.87]	6.50 [5.06, 9.01]	2.79 [2.38, 3.18]
AlphaEdit projection	10.05 [7.91, 13.77]	6.67 [5.19, 9.35]	2.42 [2.20, 2.67]
OGD	19.86 [15.90, 26.90]	9.61 [7.53, 13.34]	2.42 [2.23, 2.62]
EWC	19.53 [15.60, 26.47]	9.36 [7.33, 13.01]	2.40 [2.20, 2.59]
L2-SP	26.07 [20.86, 35.29]	12.99 [10.18, 17.95]	4.41 [4.05, 4.79]
LoRA	24.38 [19.46, 33.06]	10.77 [8.43, 14.89]	4.99 [4.60, 5.42]
Native full-head	19.53 [15.59, 26.32]	9.36 [7.33, 13.02]	2.40 [2.20, 2.59]

Table 13: **The external-method comparison separates utility overshoot from response reduction.** ImageNet values use the selected configurations. Higher utility/target means more overshoot. Lower AlphaEdit/native RMS means greater protection relative to native fitting. Bounds are endpoint-wise class-bootstrap bounds and other entries are point summaries. Sensitivity changes the projection-basis construction.

Quantity	ResNet-50	DenseNet-121	MobileNetV3-L
Route utility/target	1.58	1.62	1.53
Native utility/target	1.73	2.80	1.69
AlphaEdit/native RMS ↓	0.51	0.37	0.93
AlphaEdit/native upper bound	0.53	0.39	0.99
EWC/native maximum relative RMS difference across targets: 0.04%			
Alternative projection basis: largest ratio shift 0.02, MobileNet upper bound 1.002			
External fitting: 96 configurations, 8,096 to 2,050,049 trainable parameters			

side RMS, so the protection difference can be read at comparable improvement. All models satisfy the utility noninferiority criterion, and the largest response-ratio upper bound is 0.445. The routed update changes selected output coordinates, whereas native fitting changes the complete classifier head. This comparison confirms that endpoint control can preserve the native update's useful benefit with smaller protected changes.

Table 14: **The support-matched route preserves utility with smaller protected RMS.** The native ImageNet comparison uses the uncalibrated route. Absolute values are route/native and utility difference is route minus native. Higher utility and lower RMS are preferable. Brackets are endpoint-wise 95% authorization-class bootstrap intervals. The utility noninferiority margin is -0.005.

Model	Utility route/native ↑	Utility diff. ↑ [interval]	Response RMS route/native ↓	Response ratio ↓ [interval]	Params native/route
ResNet-50	0.0281/0.0230	+0.0051 [+0.0018, +0.0086]	0.00553/0.06869	0.080 [0.049, 0.109]	2.05M/32
DenseNet-121	0.0287/0.0239	+0.0048 [-0.0005, +0.0102]	0.00522/0.05683	0.092 [0.051, 0.142]	1.02M/32
MobileNetV3-Large	0.0217/0.0185	+0.0031 [+0.0004, +0.0061]	0.01424/0.03395	0.419 [0.396, 0.445]	1.28M/32

Table 15: **Writable pairs form a small part of the typical document.** Counts cover strictly aligned DocRE documents. Columns report the median and nearest-rank ninetieth percentile (P90), followed by the maximum. These are population distributions rather than repeated-run estimates.

Quantity	Median	P90	Maximum
Directed pairs	342	702	1722
Writable pairs	13	27	82
Pure protected pairs	7	15	62
Revised-negative pairs	322	677	1673
Added triples	15	33	110

Table 16: **Document-wide exposure exceeds the authorized pair width.** Counts describe aligned DocRE documents. Width ratios divide document-wide pair counts by writable-pair counts, so larger ratios indicate greater exposure. Percentages use writable documents as the denominator. Values are exact counts or population summaries at the original reporting precision.

Quantity	Value
Original / revised / aligned documents	4,051 / 4,053 / 3,826
Added relation triples / writable pairs	65,751 / 56,412
Writable documents / mixed with protected positives	3,794 / 3,722
Mixed rate including revised negatives	100.0%
Pairs within writable documents	1,481,282
Aggregate union / writable width	26.26
Per-document union ratio, median / mean / P90	26.25 / 37.34 / 70
Protected-only union / writable width	25.26
Test-panel union / writable width	20.55
Test-panel protected-only / writable width	19.55
Document-scope / full-redeployment protected RMS	99.98%
Document-scope / pair RMS [interval]	2.66 [2.60, 2.72]

F Permission Structure and Route Selection

This section supports the mechanism and selection conclusions in Section 6. It retains the workload distributions and the comparisons of route alignment and selection.

F.0.1 Mixed Permissions Explain the Granularity Advantage

The workload measurements below support the permission-structure interpretation in Section 6.3.

Mixed permissions within documents. Figure 1 summarizes the aligned population. Table 15 separates typical documents from the upper tail, while Table 16 gives aggregate counts and the union-to-writable width ratios beyond the median quantities reported in Section 6.3.

Figure 1 reports this distribution in the main text; Table 15 separates typical documents from the upper tail and Table 16 gives the aggregate counts.

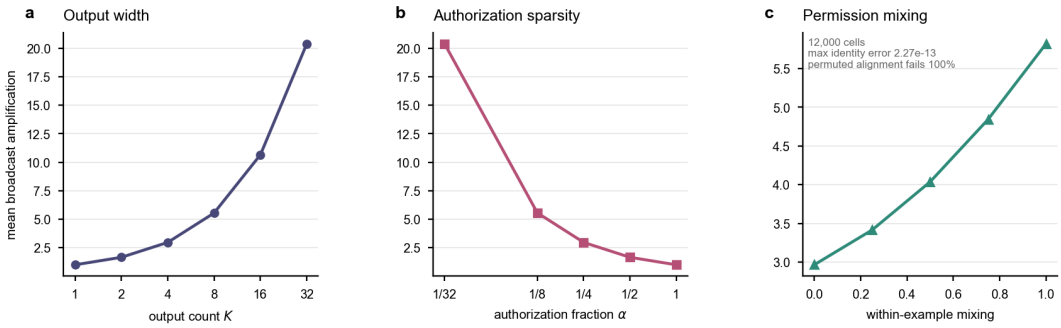


Figure 5: **Broadcast amplification grows with width and permission mixing.** Panels vary output width, authorization fraction and within-input mixing. Amplification divides union squared response by fine squared response. Larger values indicate more excess response from broadcasting. Points are means over the numerical factor grid and connecting lines describe the evaluated levels. Authorization sparsity in panel (b) is evaluated at width 32.

Table 17: **The numerical trends track width and permission mixing.** Amplification is mean union-to-fine squared response over the numerical grid. Larger amplification means greater broadcast excess. Width, authorization fraction and mixing are varied separately. The lower rows summarize multiplier scaling and alignment comparisons on the same generated responses.

Variable levels	Mean amplification ↑
Width: 1 / 2 / 4 / 8 / 16 / 32	1.000 / 1.658 / 2.965 / 5.543 / 10.653 / 20.383
Authorization fraction at width 32: $1/32$ / 1	20.383 / 1.000
Mixing: 0 / 1	2.965 / 5.818
Quadratic scaling in s	Relative deviation below 10^{-10}
Permutation of route values	Lower efficiency in all 8,000 applicable settings

Width and authorization determine broadcast amplification. Figure 5 separates width and authorization effects from within-input permission mixing. Table 17 gives the complete numerical levels and the multiplier comparison underlying the broadcast trends in Section 6.3.

Alignment distinguishes useful routing from small route values. Figure 6 reports route values by authorization stratum alongside utility and response ratios. Table 18 retains the absolute utilities of aligned and permuted pair routes together with the scalar utilities, complementing the protected-response comparison in Table 2.

Figure 6 reports the route values, utilities and response ratios in the main text; Table 18 retains the absolute utilities behind those ratios.

F.0.2 Utility Bounds and Alignment Determine the Selected Interface

The candidate and perturbation measurements below support the selection interpretation in Section 6.3.

Utility bounds determine feasibility. Figure 2 reports utility-bound surpluses for the selected updates. Table 19 gives the eligible counts for every architecture alongside the largest utility lower bounds, including the fitting-label perturbations that determine whether a target is attainable by the fitted candidate family.

Figure 2 reports the utility-bound margins and eligible counts in the main text; Table 19 gives the eligible counts for every architecture.

Authorization alignment determines the preferred granularity. We vary permissions and route alignment to identify the conditions separating fine routing from scalar control. Table 20 shows

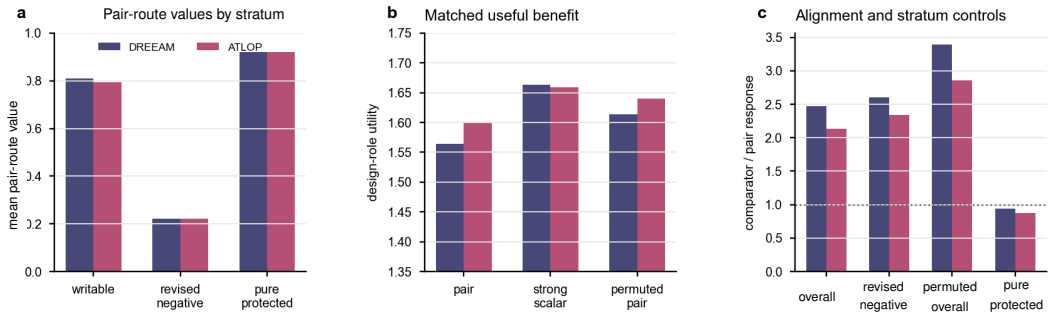


Figure 6: **Permission alignment separates pair routing from its scalar and permuted alternatives.** Results use 160 complete DocRE documents. Panel (a) reports mean route values by authorization stratum. Panel (b) reports mean utility, where larger values are preferable. Panel (c) divides comparator RMS by aligned-pair RMS, so values above the dotted unit reference favor pair routing. Bars are point summaries for each endpoint.

Table 18: **Permuting pair strengths increases protected response.** Utilities are mean added-relation margin gains on the 160-document DocRE comparison. Higher utility is better. Response ratios divide permuted RMS by aligned-pair RMS, so larger ratios favor alignment. Entries are point summaries.

Quantity	DREEAM BERT-base	ATLOP RoBERTa-large
Pair utility ↑	1.564	1.600
Scalar utility ↑	1.663	1.660
Permuted utility ↑	1.613	1.640
Permuted / pair RMS	3.394	2.855

equality under uniform permissions and a scalar advantage after sufficient alignment perturbation. With permuted fitting labels the full candidate family selects the unchanged raw route, while calibrated-only candidates incur larger protected responses. Table 21 records those changes and the associated loss of authorization efficiency. The utility requirement responds to attainable benefit, whereas efficiency also reflects where that benefit changes protected outputs. Thus the feasible set and authorization efficiency answer complementary questions when choosing an interface.

Table 19: **Higher targets reduce the eligible candidate set.** Counts refer to the utility-qualified ImageNet family. Maximum utility LCBs refer to the separate alignment-perturbation comparison. Targets above those maxima return no update. Counts and bounds are support summaries, with LCB denoting a simultaneous lower confidence bound.

Model	Eligible at 0.01	Eligible at 0.02	Maximum LCB ↑	Permuted-label maximum LCB ↑
ResNet-50	779	758	0.0473	0.0463
DenseNet-121	762	722	0.0464	0.0446
MobileNetV3-Large	632	283	0.0239	0.0236
Minimum selected utility-bound surplus in the utility-qualified comparison: 1.74×10^{-5}				

Table 20: **Permissions and alignment distinguish the preferred interface.** Conditions combine the theory with perturbations on the 48,000-image ImageNet panel. Equality counts are exact comparisons. The perturbation interval brackets the observed transition across architectures, while utility thresholds are maximum support lower bounds. Lower protected response is preferred among utility-qualified candidates.

Observable condition	Decision
Uniform permissions within each input	Scalar and fine routes coincide
Mixed permissions with qualified fine utility and higher efficiency	Select the fine candidate minimizing response UCB
Scalar efficiency at least as high after alignment perturbation	Retain the scalar. Transition fraction lies in $(0.50, 0.75)$
Target exceeds the maximum utility LCB	Return no update. Bounds are in Table 19
Permuted route-fitting labels with qualified raw-route utility	Select the unchanged raw route from the complete family

Table 21: **Alignment perturbation reduces authorization efficiency.** ImageNet comparisons use the 48,000-image panel. Response multiples compare label-permuted fitting with true-label fitting at matched utility. Larger multiples indicate more protected change. Efficiency decreases compare complete assignment permutation with the unperturbed route. Entries are point summaries or ranges across architectures.

Quantity	ResNet-50	DenseNet-121	MobileNetV3-L
Row-route response multiple	6.38	6.25	14.19
Unperturbed efficiency	12.47	6.48	2.46
Efficiency decrease factor	166	199	689
Assignment-permutation fractions: $0,\ 0.25,\ 0.5,\ 0.75,\ 1$			
Uniform-permission comparisons: fine equals union in 9/9 settings			
All 18 uniform-permission reduction intervals include zero, mixed-permission reductions: 54.5% to 72.6%			
Permuted-label selections at $t = 0.01$: 18 updates, all selecting the unchanged raw route			
Calibrated-family response multiple: 6.2 to 13.9			
Protected coordinates with route value at least 0.5: 81.6% to 93.2% after permutation			
The corresponding true-label proportion is at most 1.1%			
Fine/scalar response at permutation fraction 0.5: 0.53 to 0.75			
Fine/scalar response at permutation fraction 0.75: 1.11 to 1.16			
MobileNet maximum support utility under complete assignment permutation: 0.0028			
Across 36 reachable settings: $R\eta_{\mathrm{loc}}/\widehat{U}^2 > 1$, mean 1.13			

Table 22: **The fine-route advantage transfers across multiple endpoint geometries.** Efficiency ratios divide fine-route efficiency by scalar efficiency, so larger values favor fine routing. Brackets are 95% bootstrap intervals over the units in Table 6. CoNLL reports the median document ratio and other families report panel ratios. Win rate is the fraction of paired units favoring the fine route. The remaining columns report installed protected mass and descriptive log–log slopes of $H^{(\tau)}$. Dashes denote unreported entries.

Family	Model	Efficiency ratio ↑ [95% CI]	Win rate ↑	Installed mass	$H^{(\tau)}$ slope
CodRED	NEPD	1.35 [1.14, 1.66]	75.6%	$13.06 \times P^{(0)}$ per round	-0.95
CoNLL-2003	BERT-base NER	826 [657, 1,022]	100%	0.83% over the run	-0.012
AG News	RoBERTa-base, calibrated row	1.13 [1.04, 1.22]	61.9%	3.2% over the run	+0.074
AG News	RoBERTa-base, uncalibrated row	1.56 [1.37, 1.78]	75.0%	–	–
CIFAR-100	ResNet-56	1.85 [1.53, 2.18]	82.0%	$1.26 \times P^{(0)}$ over T=20	-0.64
CIFAR-100	VGG-16-BN	0.13 [0.00, 7.26]	91.0%	$1.11 \times P^{(0)}$ over T=20	-0.91
CIFAR-100	MobileNetV2	2.37 [1.98, 2.79]	92.0%	$1.44 \times P^{(0)}$ over T=20	-0.61

Table 23: **The ceiling comparison depends on useful alignment.** Structural comparisons use constructed update responses on the additional endpoints. Refinement counts enumerate nested partitions with nonincreasing response floors. One-sided exceedances occur at negative alignment and none exceed the two-sided bound. Other entries describe the authorization geometry or point efficiency ratios, where larger ratios favor fine routing.

Family	Structural observations
CodRED geometry	602 bags, 160 units, 276 relation coordinates per bag and 166,152 coordinates overall
CodRED authorization	198 writable bags, all selected units mixed, population mixed rate 37.0%
CodRED calibration	Clipping the fitted negative Platt slope to zero makes all route values equal, giving zero broadcast excess on 160 units
CodRED alignment	Permuted fine/scalar efficiency 0.42, authorization-mask Oracle 2,499
CodRED ceiling	8,904 positive-alignment comparisons respect the ceiling, with refinement monotonicity on 6,400 partition pairs
CoNLL geometry	160 documents, 295,936 coordinates and 3,145 writable tokens, with a direction fitted on 47 training error tokens
CoNLL authorization	100% mixed documents, median 14 writable and 152.5 protected tokens
CoNLL comparison	Fine efficiency is larger on 150 eligible documents, with refinement monotonicity on 2,093 partition pairs
AG News geometry	3,680 examples, four margins each, target-only mixed rate 4.3%
AG News authorization	All-class permissions are uniform within each example and mixed within composite blocks
AG News ceiling	658 negative-alignment one-sided exceedances, 0 at positive alignment and 0 two-sided exceedances
AG News refinement	5,635 nested partition pairs preserve the ordering
CIFAR authorization	Permissions are uniform within images and mixed when images share class-column controls
CIFAR ceiling	Negative-alignment one-sided exceedances: 21 / 24 / 12 for ResNet / VGG / MobileNet
CIFAR signed comparison	0 positive-alignment and 0 two-sided exceedances, with monotonicity on 16,800 partition pairs

G Structural Comparisons and Weight Updates

This section supports the transfer conclusions in Section 6. It presents the additional endpoints, sequential settings and comparisons with weight updates.

G.0.1 Transfer Across Endpoints and Weight Updates

The complete endpoint comparisons below support the transfer results in Section 6.4.

Structural comparisons on additional tasks. We apply the same quantities to constructed responses on released task endpoints to test their dependence on endpoint geometry. Table 22 reports larger fine-route efficiency on CodRED and CoNLL, as well as AG News and most CIFAR backbones. For VGG-16-BN the panel ratio is 0.13 [0.00, 7.26] and the unit-level win rate is 91.0%. High-response units receive more weight in the panel ratio. Reporting both quantities exposes this weighting effect within Eq. (13). Table 23 also records ceiling attainment and partition refinement, with the two-sided ceiling covering negative-alignment comparisons. The broadcast excess is nonnegative across these endpoints, as predicted for a fixed response. The numerical minimizers in Section 4.1 support attainment of the predicted minimum. These results connect the protection benchmark to task-specific authorization structures beyond the primary endpoints.

Table 24: **Ceiling trajectories reflect different protected-mass growth regimes.** Values summarize write-once-then-protect schedules. Growth is relative to the stated initial protected pool. More negative log–log slopes mean faster ceiling decrease. Realized trajectories and fixed-slope model predictions are listed separately. The reported CIFAR slopes use summed protected masses.

Quantity	Value
CodRED per-round growth, nonwritable / all non-added coordinates	13.06 / 0.94
CodRED slope against round / shifted round	-0.95 / -0.88
CoNLL realized / model slope	-0.012 / -0.013
CoNLL consumed-writable-mass slope	-1.24
AG News realized / model slope	+0.074 / -0.010
AG News example-scope schedules with infinite ceiling	45/45
DocRE per-round growth, pure protected / nonwritable	0.447 / 0.0063
DocRE schedule length	$T = 8$

Table 25: **A distilled checkpoint retains useful update behavior with lower overall RMS than the parameter-space baselines.** DREEAM comparisons use $\lambda = 0.4$. Higher retained utility and micro-F1 are preferable. Response ratios state their numerator explicitly and smaller values mean less response from it. Brackets are 95% complete-document bootstrap intervals. Deployment and utility-matched quantities use their respective scales.

Quantity	Value (utility/F1 ↑, RMS ratios ↓)
Deployed student / pair RMS	7.37 [6.84, 8.17]
Student utility / pair utility at the pair scale	5.2
Deployed student / matched LoRA RMS	0.66 [0.61, 0.74]
Deployed student / matched projection RMS	0.78 [0.73, 0.88]
Pure protected-positive student / matched LoRA RMS	1.14 [1.04, 1.26]
Pure protected-positive student / matched projection RMS	1.65 [1.52, 1.83]
Student micro-F1	0.683
Original / revised micro-F1	0.329 / 0.747
Utility-matched student / pair RMS	1.47
Unscaled LoRA / pair utility	7.9
Unscaled projection / pair utility	8.8
Unscaled LoRA / projection micro-F1	0.299 / 0.240
Writable training pairs with retained relations / writable pairs	8,789 / 39,656

Protected-mass growth determines sequential behavior. Table 24 gives growth relative to each initial protected pool and separates realized trajectories from fixed-slope model predictions, retaining the DocRE schedule and the alternative CodRED protected sets used to interpret the sequential behavior in Section 6.4.

Distillation retains part of the routing benefit. Table 25 separates the deployed student's comparisons from its utility-matched response. The latter remains close to the pair route overall. The student has greater pure protected-positive response than matched LoRA and projection because its teacher assigns larger strengths to retained relations. The table also reports the utilities and micro-F1 of unscaled LoRA and projection together with the number of writable training pairs that retain relations, giving the quantities behind the calibration effect described in Section 6.2.

Known permissions after actual weight edits. We apply nested authorization masks to CounterFact weight edits to test the interface with actual parameter changes. We use single-token targets so each intended output has an explicit vocabulary coordinate, and each edit begins from the same base model. Utility averages the new-target logit change over writable prompts, while protected response sums squared changes over the remaining prompt–token coordinates. Table 26 compares ROME with

Table 26: **The known pair mask preserves writable utility with zero protected response.** CounterFact comparisons apply each mask to the same GPT-2 XL edited-minus-base response at the same scale. Utility and margin entries are case medians with 95% case-bootstrap intervals. Input/global and output/global are median protected-response fractions with the same interval definition. Lower fractions are preferable and pair max is the largest protected response across cases. Positive utility is the fraction of cases with improvement.

Editor	Positive utility ↑	Target-logit change ↑	Margin gain ↑	Input/global ↓	Output/global ↓	Pair max ↓
ROME	100.0%	12.08 [11.27, 13.17]	13.47 [12.16, 14.69]	99.7% [99.6%, 99.8%]	7.9×10^{-7} [5.4×10^{-7}, 1.4×10^{-6}]	0
Constrained FT	100.0%	5.67 [5.45, 6.02]	5.93 [5.37, 6.26]	73.2% [65.1%, 78.5%]	4.6×10^{-5} [2.8×10^{-5}, 7.5×10^{-5}]	0

Table 27: **Both weight editors increase target preference.** CounterFact results use 100 independent cases per editor. Preference is the mean per-case fraction favoring the new target over the original target. Larger edited preference is better. Weight norms are medians of changed-weight Frobenius norms and establish the size of the parameter changes.

Editor	Base preference	Edited preference ↑	Median weight norm
ROME	24.3%	97.7%	8.10
Constrained FT	24.3%	67.7%	1.53

constrained fine-tuning through their resulting responses (Meng et al., 2022). Input scope retains most protected response, whereas output scope removes nearly all of it. The prompt–token pair mask preserves identical writable utility with zero protected response because it closes every protected coordinate. Table 27 confirms increased target preference for both editors. Known permissions thus realize the protection attainable by matching the endpoint mask to the authorization unit.

Parameter regularization at penalties that change the update. Table 28 compares AGER with the ImageNet parameter-regularization update (Kirkpatrick et al., 2017) over a penalty range wide enough for the penalty to change the fitted update. Eight penalty settings and six unpenalized references give fourteen working points. Every protected-RMS ratio against AGER lies between 2.397 and 19.56, and the lowest interval endpoint among them is 2.200. At the setting that removes 66.35% of the Fisher-weighted update energy, protected RMS changes by at most 0.1502% and clipped utility by at most 4.181×10^{-4} against the same update with no penalty. Penalizing movement along Fisher-weighted directions therefore leaves the protected outputs close to where the unpenalized update leaves them.

Table 28: **Parameter regularization keeps its protected response once the penalty changes the update.** Rows are the fourteen working points of the ImageNet penalty comparison, where λ is the penalty weight and $\lambda = 0$ marks the unpenalized reference at each target. Fisher energy divides the Fisher-weighted energy of the penalized update by that of the same update with no penalty, so values below one mean the penalty changes the update. Ratios divide each of these updates' protected RMS by AGER's at the same target, and smaller values favor the penalized update. Brackets are endpoint-wise 95% authorization-class bootstrap intervals.

Model	Target	λ	Fisher energy	RMS ratio ↓ [interval]
ResNet-50	0.01	0	1.000	17.90 [14.27, 24.22]
ResNet-50	0.01	10^{-4}	0.9006	17.90 [14.27, 24.22]
ResNet-50	0.01	10^{-2}	0.3365	17.93 [14.30, 24.25]
ResNet-50	0.02	0	1.000	19.53 [15.57, 26.42]
ResNet-50	0.02	10^{-4}	0.9006	19.53 [15.57, 26.42]
ResNet-50	0.02	10^{-2}	0.3365	19.56 [15.60, 26.46]
DenseNet-121	0.01	0	1.000	17.79 [13.92, 24.65]
DenseNet-121	0.01	10^{-4}	0.8862	17.79 [13.92, 24.66]
DenseNet-121	0.02	0	1.000	9.361 [7.326, 12.98]
DenseNet-121	0.02	10^{-4}	0.8862	9.363 [7.328, 12.98]
MobileNetV3-Large	0.01	0	1.000	4.128 [3.788, 4.468]
MobileNetV3-Large	0.01	10^{-4}	0.7841	4.130 [3.790, 4.470]
MobileNetV3-Large	0.02	0	1.000	2.397 [2.200, 2.594]
MobileNetV3-Large	0.02	10^{-4}	0.7841	2.398 [2.200, 2.596]

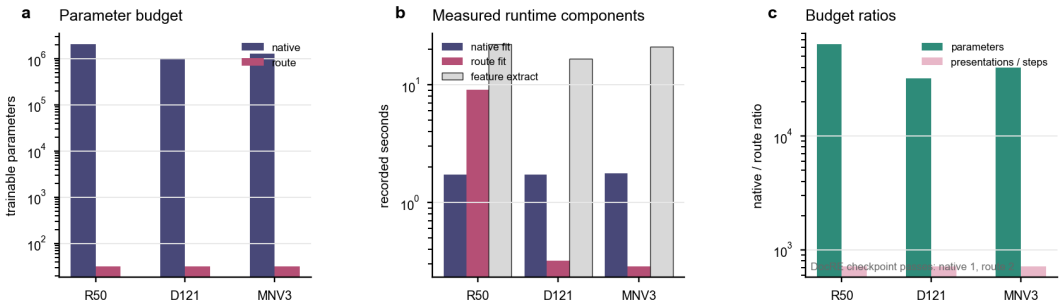


Figure 7: **Endpoint routing uses fewer fitted parameters with task-dependent computation.** Panel (a) compares trainable parameters on a logarithmic axis. Panel (b) reports ImageNet fitting and feature-extraction seconds on a logarithmic axis. Panel (c) gives native-to-route parameter and optimization-budget ratios. Bars are measurements from one CUDA execution rather than repeated-run estimates. The DocRE annotation counts checkpoint evaluations. R50 denotes ResNet-50, D121 denotes DenseNet-121 and MNV3 denotes MobileNetV3-Large.

Table 29: **Feature extraction and fitting contribute different costs.** ImageNet times are seconds from one CUDA execution with Torch 1.11.0+cu113 and feature-extraction batch size 64. Smaller counts or times mean lower cost within each column. The final column lists native sample presentations followed by route optimization steps, which are different work units. Values are recorded measurements without repeated-run uncertainty.

Model	Native params ↓	Route params ↓	Native fit ↓	Route fit ↓	Feature extract ↓	Presentations/steps ↓
ResNet-50	2,049,000	32	1.722	9.119	22.114	229,500/320
DenseNet-121	1,025,000	32	1.737	0.320	16.650	229,500/320
MobileNetV3-Large	1,281,000	32	1.760	0.288	21.125	229,500/320

H Fitting and Inference Costs

This section supports the computational comparison in Section 6. It separates parameter counts from measured fitting and inference costs.

H.0.1 Fitting and Inference Costs

The cost measurements below support the fitting and inference comparison in Section 6.1.

Figure 7 and Table 29 separate fitted parameter counts from measured fitting and feature-extraction times, complementing the inference construction described in Section 6.4.

I Extended Discussion and Limitations

This section supports the scope statements in Section 7. It develops the implications of authorization uncertainty and changes in the deployment setting.

I.1 Discussion and Limitations

The extended discussion below supports the interpretation of the results in Section 7.

The results support a precise statement. When permissions differ across outputs inside one example, repeating an open scalar route creates an exact nonnegative response excess for a fixed endpoint response. A fine-output route can remove that excess while retaining useful benefit when its authorization efficiency is high enough. Utility-qualified selection makes this second condition testable and permits abstention.

The analysis also gives an operational rule. A finer interface is warranted only when the workload demonstrates mixed authorization, the protected endpoint is specified, and a fine candidate survives the utility lower bound. A scalar remains adequate with one output or a common responding route value. It is also adequate under full authorization or when no fine candidate has sufficient utility evidence. The method is therefore a conditional interface change rather than a default replacement for sample-wide routing. This links Eq. (8) to the empirical selector.

Authorization-label uncertainty lies outside the current selector. Expanding the protected set changes the writable objective and response population, so the selector treats the policy as fixed. Sensitivity to permission errors remains future work.

The exact identity compares routes applied to the same response. The route-strength-preserving permutation control isolates the additional alignment requirement for matched-utility comparisons. Portable route-form and native procedures use different inference semantics. The native ImageNet intervals quantify sampling uncertainty over authorization classes and do not account for configuration selection. The native DocRE panel compares checkpoint responses against fixed reference thresholds. Theorem 3 applies to declared bounded or clipped statistics under independent examples within each support role. It requires the route family to be chosen before reading support outcomes. The protected-response gains concern the overall and revised-negative strata. The pure protected-positive stratum favors scalar control because its scalar-to-pair ratio is below one at both endpoints. At the deployment scales the pair route responds more than full redeployment on that stratum.

The endpoint and update spaces are also limited. ImageNet keeps the backbone fixed and studies linear-classifier or shared low-dimensional endpoint updates. DocRE binds specific original and revised checkpoints. Neither domain proves global nonlinear behavior for arbitrary full-model fine-tuning. ImageNet class-block and DocRE document-block intervals address their declared sampling units. The four additional endpoint families use constructed responses. These include fitted gradient directions and added-relation responses, as well as per-interface fitted directions. The responses are not induced by an external revision event. These comparisons test the structural quantities on real endpoint geometry without adding a second natural revision workload. The CounterFact panel adds actual GPT-2 XL weight updates from two editors and raises the combined coverage to seven datasets and twelve models. Its permissions are constructed from prompt roles and one target token, and its cases are independent single edits. It does not establish accumulated multi-edit behavior or universal cross-domain validity.

Authorization is supplied by the workload. The method does not infer who should be allowed to update an output, and it cannot repair an incorrect permission policy. Protected response is a measured endpoint quantity rather than a claim about fairness, privacy, or semantic harmlessness. Deployment therefore requires a separately justified permission source and task-specific checks for events that squared response may not capture.

The reported boundary is structural rather than universal. Equality holds when authorization is already uniform or only one output is present. Weak alignment can remove the matched-utility advantage even when the exact broadcast excess is positive. The selector also returns no update when no candidate satisfies the simultaneous utility and response bounds. These regimes specify when a fine route is unnecessary, unsupported, or outside the claim.

I.1.1 AUTHORIZATION AND ENDPOINT UNCERTAINTY

The uncertainty discussion below supports the treatment of supplied permissions in Section 7.

The experiments treat authorization as fixed because routing cannot determine which outputs are legitimately writable. In practice permissions may be incomplete or disputed. They may also be time dependent. A false writable label can authorize harmful change, while a false protected label can make the useful objective unreachable. These two errors are asymmetric and should be evaluated by the application owner before route selection.

Permission uncertainty also changes the estimand. Expanding the protected set adds coordinates to response and can remove observations from useful benefit. It is therefore not equivalent to ordinary label noise on a fixed outcome. One possible extension is a robust selector that requires utility and response requirements to hold over a declared family of plausible authorization masks. The current theorem does not supply that guarantee.

Endpoint choice introduces a second uncertainty. Squared logit or margin response captures smooth changes before a prediction flips, while prediction change captures a discrete downstream event. A route can improve one without uniformly improving the other. We report both where available and avoid replacing the declared endpoint after seeing which metric is more favorable.

I.1.2 DYNAMIC AND NONLINEAR UPDATE SETTINGS

The dynamic-setting discussion below supports the scope of the response analysis in Section 7.

The exact broadcast identity remains valid for any fixed finite endpoint response, including a response produced by a nonlinear full-model update. The matched-utility criterion is more limited because its local form uses a declared update direction and quadratic response. Large updates can alter features, route inputs, and endpoint geometry simultaneously. Recomputing the route after such an update can also introduce feedback that is absent from the fixed comparison.

A continual system adds temporal dependence. Previously protected outputs may later become writable, and successive routed updates can accumulate even when each single-step response is small. A deployment would need cumulative response accounting, a rule for revising permissions, and a new confirmation split after material model or workload drift. The present experiments concern one declared update transition at a time.

The finite selector also assumes a fixed candidate family. Expanding the family after observing support outcomes invalidates the original simultaneous event. A revised family therefore requires new data roles and a new simultaneous bound.

Finally, lower endpoint response is not automatically lower social or application harm. A small logit change can matter near a high-stakes decision threshold, while a larger change can be harmless in an unused coordinate. Authorization-granular routing supplies a technical control surface, not a replacement for domain-specific impact assessment or a justified permission policy.

I.1.3 METRIC AND CALIBRATION TRANSFER

The metric discussion below supports the interpretation of utility and response in Section 7.

The utility estimand also depends on the evaluation design. ImageNet uses clipped NLL benefit so a small number of extreme probabilities cannot dominate support selection. Clipping makes the finite bound explicit, but it also changes the target from unclipped expected NLL. A deployment that cares about tail calibration would need an additional endpoint and a bound suited to that endpoint.

Response ratios require a stable comparator denominator. A small native or pair-route response can make a ratio large even when both absolute responses are small. Applications that adopt the same ratio thresholds should therefore also inspect absolute protected responses.

The interval upper bound one and DocRE upper bound 0.50 encode different practical requirements. They are not exchangeable significance levels. Transporting the method to a new endpoint requires a new margin justified by that endpoint's tolerance, followed by a new comparison on held-out data.

Calibration drift can affect both route input and useful performance. The ImageNet route reads relative logits, and the DocRE route reads relation margins. Changes in temperature, label prevalence, or checkpoint confidence can move these quantities without changing authorization policy. Monitoring should therefore include route-strength distributions and useful-benefit slack, not only protected response. The permutation probe makes this concrete: damaged route fitting keeps utility feasible and satisfies the selection criterion in all 18 permuted configurations. Authorization efficiency falls by more than two orders of magnitude, so η belongs in the monitored set.

Finally, the reported block intervals describe uncertainty over the fixed class or document population. They do not include uncertainty from checkpoint training, authorization-policy construction, or alternative endpoint metrics. Those sources can be incorporated only by expanding the experimental unit and adopting a correspondingly broader evaluation design.

I.1.4 Additional connections to prior work

Repeated editing uses memory-based routing (Mitchell et al., 2022b; Wang et al., 2024a; 2025) and explicit scope mechanisms (Qiao et al., 2025; Li et al., 2026; Hinkley & Najafirad, 2026) to separate edited predictions from other queries as stored updates accumulate. Function-space penalties on predictive changes (Benjamin et al., 2019) complement behavioral preservation (Angioni et al., 2025; Elmecker-Plakolm et al., 2026). Preservation can also concern specified output properties (Tao & Thakur, 2024) or non-target distributions (Jeong et al., 2026). Evaluation studies (Traub et al., 2024) and utility-directed prediction (Cortes-Gomez et al., 2025) connect selective decisions to their consequences. Other approaches refine link prediction (Jain et al., 2024) or margins (Duan et al., 2025) to predict which relations hold.