UPDATE TOMOGRAPHY: REUSABLE LEARNING RECORDS VIA SUPERVISION SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Unlike homogeneous models, updates from heterogeneous architectures cannot be directly compared in a shared parameter space. We therefore turn to supervision space and show that updates from different architectures can be represented exactly as linear measurements of the same supervision contrast. Moreover, a target update remains uniquely determined as long as the supervision ambiguity left unresolved by these measurements cannot affect that target, even when the full supervision itself is not uniquely identified.We formalize these results as Update Tomography (UT), a framework for constructing updates across heterogeneous architectures without parameter alignment. Experiments on vision and language models agree with the theoretical predictions: even when full supervision cannot be recovered, UT reproduces direct-supervision learning whenever the information required by the target is retained. This learning information remains usable by heterogeneous models introduced after the original models are no longer available, with relative trajectory error remaining on the order of 10^{-5} across ten successive learning events.Thus, cross-architecture update reuse need not treat model-specific parameter changes as the object to be transferred, but can instead characterize and reuse the information that a learning event leaves in supervision space.

1 INTRODUCTION

Model merging and task arithmetic combine learned parameter changes, while federated learning aggregates local updates (Wortsman et al., 2022; Ilharco et al., 2023; McMahan et al., 2017). When models share parameter coordinates, these changes can be compared and combined directly. Across heterogeneous architectures, however, parameter dimensions, semantics, and correspondences may differ, so parameter space no longer provides a common reference. Yet models learning from the same event share the same supervision. We therefore turn from parameter correspondence to supervision space, asking how the shared learning information in heterogeneous updates can be characterized, when it is sufficient to determine another model's update, and how it can be reused.

We establish two results. First, for losses whose output gradients are affine in supervision, including cross-entropy and mean-squared error, updates from different architectures become exact linear measurements of the same supervision contrast after known model- and state-dependent terms are removed, without linearizing the neural networks themselves. Supervision space thus provides a common measurement domain for otherwise incomparable updates. Second, we prove that full supervision need not be recovered: the target's supervision-dependent gradient contribution is uniquely determined exactly when every feasible supervision difference invisible to the retained measurements is also invisible to that target. We call this necessary and sufficient condition *target-relative observability*. Given the target computation and specified learning rule, the determined contribution can then be converted into the target's parameter update.

These results lead to *Update Tomography* (UT). Analogous to medical tomography, where observations from different views provide complementary measurements of the same underlying object, UT interprets heterogeneous model updates as complementary measurements of shared supervision. It pools the information relevant to the target and lets the target form its own update from its current state through its backward pass and optimizer, without establishing a parameter correspondence between source and target models (Figure 1). We test whether recorded transitions match the predicted

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.

measurements, vary the retained information and target requirements to probe the determination boundary, and construct whole-network updates in heterogeneous deep models. The experiments also examine reuse by later models after the original sources and event labels are removed.

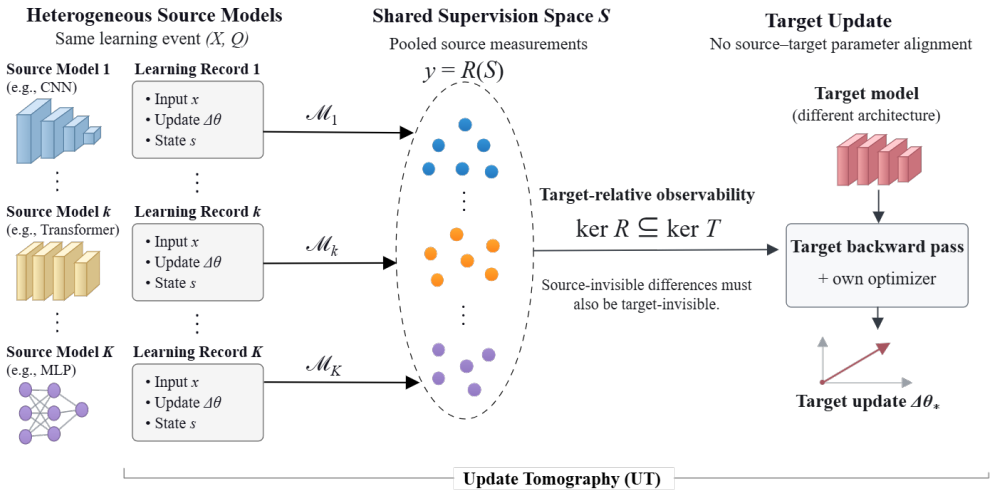


Figure 1: Update Tomography. Heterogeneous models process the same learning event and leave model-specific records. Correcting and pooling these records yields measurements $y = R(S)$ in a shared supervision space. When every feasible source-invisible difference is also target-invisible, equivalently $\ker R_{\mathcal{L}} \subseteq \ker T_{\mathcal{L}}$, the target can form its own update through its backward pass and optimizer without aligning source and target parameters.

2 Related Work

Objects of cross-model transfer. Model merging and task arithmetic reuse learned changes directly in parameter space, typically assuming comparable parameter coordinates (Ilharco et al., 2023; Ortiz-Jimenez et al., 2023; Yadav et al., 2023). Alignment methods establish correspondence through permutations between compatible parameterizations (Ainsworth et al., 2023), task-vector re-basing (Rinaldi et al., 2025), representation-level matching (Rinaldi et al., 2026b), or target-side gradient structure (Rinaldi et al., 2026a). Other approaches transfer predictions through distillation (Hinton et al., 2015; Liu et al., 2022a; Ahn et al., 2019), synthetic data through gradient or trajectory matching (Zhao et al., 2021; Cazenavette et al., 2022), or learning trajectories across models (Chijiwa, 2024). Recent theory also relates task vectors to the gradients that generate them (Zhou et al., 2026). UT instead uses the supervision shared by a learning event as the common reference for constructing an update in another architecture.

What updates reveal. Gradient inversion and label-recovery methods show that gradients and model updates can reveal the data or supervision that generated them (Zhu et al., 2019; Geiping et al., 2020; Zhao et al., 2020; Chen & Vikalo, 2024; Zhu et al., 2023). Functional observability and optimal recovery ask whether a quantity of interest can be determined without reconstructing the entire latent object (Darouach, 2000; Montanari et al., 2022; Donoho, 1994). We ask when information retained in heterogeneous updates determines another architecture's supervision-dependent gradient contribution without full supervision recovery. The target then constructs its update through its specified learning rule.

3 What Heterogeneous Updates Measure

Consider an ordered batch of n inputs with supervision u, processed by M heterogeneous source models. One optimizer step on this batch defines a learning event. Examples are aligned across models row by row, although model-specific preprocessing may differ. At construction time, the event inputs and interpretable source records are available, but the event supervision is not supplied.

For source m, let θ_m denote the specified trainable parameters, $z_m(\theta_m)$ its vectorized outputs on the event, $J_m = \partial z_m/\partial\theta_m$ the corresponding Jacobian, and $g_m(u) = \nabla_{\theta_m} L_m$ the event gradient. Here L_m is the averaged event data loss; optimizer terms such as decoupled weight decay are handled separately when the transition is corrected. All quantities below are evaluated at the recorded pre-update state, model mode, and realized randomness.

We assume that the output gradient is affine in a supervision representation $\psi(u)$:

$$\nabla_{z_m} L_m = a_m - D_m\psi(u), \tag{1}$$

where a_m and D_m are known at the recorded computation and independent of the unknown supervision. The shared object is $\psi(u)$, not the models' output coordinates: D_m may differ across architectures and may include known mappings of the supervision.

Exact supervision measurement. Choose a known reference u_0 and define the supervision contrast

$$s := \psi(u) - \psi(u_0).$$

The chain rule gives

$$g_m(u_0) - g_m(u) = J_m^\top D_m s.$$

Proposition 3.1 (Exact supervision-measurement identity)**.** *For any known normalization $\rho_m > 0$,*

$$b_m := \rho_m\big[g_m(u_0) - g_m(u)\big] = B_m s, \qquad B_m := \rho_m J_m^\top D_m. \tag{2}$$

Each architecture therefore induces its own measurement operator B_m, while the measured supervision contrast s is shared. The identity is exact at the recorded state and does not linearize the neural network.

The measurement can be obtained from an actual parameter transition. For vanilla SGD with learning rate η_m and post-update parameters θ_m^+,

$$b_m = \frac{\rho_m}{\eta_m}(\theta_m^+ - \theta_m) + \rho_m g_m(u_0). \tag{3}$$

For the stateful optimizers considered here, the event gradient is recovered from the recorded optimizer-state transition.

For average C-class cross entropy, let $F_m \in \mathbb{R}^{n\times C}$ denote the pre-update logits, $P_m = \mathrm{softmax}(F_m)$ the rowwise predictions, and $Q \in \Delta_C^n$ the supervision matrix, with one probability vector per row. Here $z_m = \mathrm{vec}(F_m)$. Define

$$Q_0 := \frac{1}{C}\mathbf{1}_n\mathbf{1}_C^\top, \qquad S := Q - Q_0.$$

Since $\nabla_{F_m} L_m = (P_m - Q)/n$, choosing $\rho_m = n$ gives

$$b_m = \frac{n}{\eta_m}(\theta_m^+ - \theta_m) + J_m^\top \mathrm{vec}(P_m - Q_0) = J_m^\top \mathrm{vec}(S). \tag{4}$$

Thus the corrected transition measures the centered supervision exactly. For the averaged mean-squared error $L_m = \|z_m - u\|_2^2/(2n)$, the same choice $\rho_m = n$ gives $b_m = J_m^\top(u - u_0)$.

From source records to a target response. A record need not retain the complete measurement b_m. Let A_m denote a known linear retention map and define

$$y_m := A_m b_m.$$

Pooling the retained source measurements gives

$$y = R(s) := \big(A_1B_1s, \ldots, A_MB_Ms\big). \tag{5}$$

Hence R describes the supervision information preserved by the source records; A_m may retain the complete measurement or only a selected part.

For a target model at its current state, let J_* and D_* denote its corresponding operators, choose a known normalization $\rho_* > 0$, and define

$$T(s) := \rho_* J_*^\top D_* s. \tag{6}$$

Its event gradient is

$$g_*(u) = g_*(u_0) - \frac{1}{\rho_*} T(s). \tag{7}$$

The source records provide $R(s)$, whereas the target needs $T(s)$. We ask whether $R(s)$ determines $T(s)$ even if the contrast s remains ambiguous.

Complete loss derivations and optimizer-state corrections are given in Appendix A.

4 Target-Relative Information Sufficiency

The source record may leave s ambiguous. *Target-relative observability* asks whether every feasible contrast consistent with that record nevertheless gives the same $T(s)$.

Theorem 4.1 (Target-relative observability)**.** *Let $\mathcal{K}$ be the feasible set of supervision contrasts, assumed convex with $0 \in \mathcal{K}$, and define the feasible difference space*

$$\mathcal{L} := \operatorname{span}(\mathcal{K} - \mathcal{K}).$$

Let $R_{\mathcal{L}} = R|_{\mathcal{L}}$ and $T_{\mathcal{L}} = T|_{\mathcal{L}}$. The source record uniquely determines $T(s)$ over $\mathcal{K}$, i.e.,

$$R(s_1) = R(s_2) \quad \Longrightarrow \quad T(s_1) = T(s_2), \qquad s_1, s_2 \in \mathcal{K},$$

if and only if

$$\ker R_{\mathcal{L}} \subseteq \ker T_{\mathcal{L}} \quad \Longleftrightarrow \quad T_{\mathcal{L}} = T_{\mathcal{L}} R_{\mathcal{L}}^{\dagger} R_{\mathcal{L}}, \tag{8}$$

where $R_{\mathcal{L}}^{\dagger}$ is the Moore–Penrose pseudoinverse. When this condition holds,

$$T(s) = T_{\mathcal{L}} R_{\mathcal{L}}^{\dagger} y, \qquad y = R(s). \tag{9}$$

Otherwise, no rule using only y can be correct for every $s \in \mathcal{K}$.

To see the criterion, let $d = s_1 - s_2$. Equal source records imply $d \in \ker R_{\mathcal{L}}$, while equal target responses require $d \in \ker T_{\mathcal{L}}$. Convexity ensures that every nonzero vector in $\mathcal{L}$ has a nonzero multiple in $\mathcal{K} - \mathcal{K}$. A source-null direction that changes T thus gives two feasible contrasts with the same record and different target responses. The complete proof is given in Appendix B.1.

We refer to $\ker R_{\mathcal{L}} \subseteq \ker T_{\mathcal{L}}$ as *target-relative coverage*. It is strictly weaker than identifying the supervision itself:

$$\underbrace{\ker R_{\mathcal{L}} = \{0\}}_{\text{identify supervision}} \quad \Longrightarrow \quad \underbrace{\ker R_{\mathcal{L}} \subseteq \ker T_{\mathcal{L}}}_{\text{determine this target}}, \tag{10}$$

but not conversely. A record may therefore leave the supervision non-unique while determining one target exactly and leaving another underdetermined. Likewise, supervision directions may be discarded without loss for targets that are insensitive to them.

Hereafter, when no ambiguity arises, R and T denote their restrictions to $\mathcal{L}$, and $R^{\dagger} := (R_{\mathcal{L}})^{\dagger}$.

A feature-space criterion. For linear output heads under cross entropy, let $X_m \in \mathbb{R}^{n \times d_m}$ denote the recorded feature design of source m, with a column of ones included when the classifier has a bias. Let $X_{\mathcal{S}} = [X_1, \ldots, X_M]$ denote the pooled source design, and let X_* denote the target design. The corresponding source and target maps are $S \mapsto X_{\mathcal{S}}^{\top} S$ and $S \mapsto X_*^{\top} S$.

Corollary 4.2 (Feature-span criterion)**.** *For $C \geq 2$ on the full probability domain,*

$$\ker X_{\mathcal{S}}^{\top} \subseteq \ker X_*^{\top} \quad \Longleftrightarrow \quad \operatorname{col}(X_*) \subseteq \operatorname{col}(X_{\mathcal{S}}) \quad \Longleftrightarrow \quad \operatorname{rank}([X_{\mathcal{S}}, X_*]) = \operatorname{rank}(X_{\mathcal{S}}). \tag{11}$$

Thus the sources need only span the sample-space directions used by the chosen target. For unprojected head measurements in a common label space, the stronger condition $\operatorname{rank}(X_{\mathcal{S}}) = n$ identifies the supervision contrast and therefore supports any fixed target response. For a general retained record, supervision identification instead requires $\ker R_{\mathcal{L}} = \{0\}$; full sample rank after a class projection identifies only the projected contrast. For rank-deficient records and deeper targets, the general condition in Theorem 4.1 is required. The proof is given in Appendix B.2.

Stability. For a perturbed record

$$y = R(s) + e,$$

where e denotes record error,

$$\|TR^{\dagger}y - T(s)\| \leq \underbrace{\|T(I - R^{\dagger}R)s\|}_{\text{missing target information}} + \underbrace{\|TR^{\dagger}\|_{\text{op}}\|e\|}_{\text{noise amplification}}. \tag{12}$$

The first term measures information missed by the record but required by the target; target-relative coverage makes it vanish. The second is controlled by the conditioning of the retained measurements. Information sufficiency and numerical stability are therefore distinct. A complete derivation and the corresponding worst-case bounds are given in Appendix B.3.

Global coverage is stronger than what is required for one particular event. For an observed record y, exact determination only requires T to be constant on

$$\mathcal{K}_y := \{v \in \mathcal{K} : R(v) = y\}.$$

Feasibility constraints can therefore remove ambiguity for a particular event even when global coverage fails. Section 5 uses this distinction for constrained construction; the local extension beyond affine supervision is given in Appendix B.4.

5 Update Tomography

Update Tomography (UT) uses source measurements to construct a target update through the target's specified learning rule. An event is first converted into a reusable record; a target then uses that record at its own current state.

Algorithm 1: Update Tomography

Input: event inputs; source transitions with the states required for correction, or corrected/compiled records with their measurement maps; target model and optimizer state; update scope. **Output:** target update.

1. **Correct.** Isolate the contribution of the current learning event from each source transition and remove known model- and state-dependent terms, yielding the supervision measurement $y_m = A_m B_m s$.
2. **Pool.** Combine the source measurements into $y = R(s)$ to form a joint record of the learning event, together with the event inputs and measurement information required for later use.
3. **Represent.** Construct a supervision representative $s_{\text{rep}} \in \mathcal{L}$ consistent with the joint record. Under target-relative coverage, one valid choice is $s_{\text{rep}} = R_{\mathcal{L}}^{\dagger} y$.
4. **Realize.** Form the target output-gradient signal from s_{rep}, backpropagate through the chosen target parameters, and apply the target's own optimizer to obtain the final parameter update.

Construction. The correction step removes source-specific prediction and training-state effects, leaving the supervision measurement derived in Proposition 3.1. Source optimizer or random state is needed only when a raw transition must be interpreted; it may instead be consumed at capture time to produce the corrected record. Pooling combines complementary measurements without aligning source parameters.

Under target-relative coverage, every record-consistent representative in $\mathcal{L}$ gives the same target response, so the true supervision need not be recovered. For classification when global coverage is not established, we can restrict the representative to valid probability contrasts,

$$S_{\Omega}^{\star} \in \arg\min_{V \in \Omega} \frac{1}{2}\|R(V) - y\|_2^2, \qquad \Omega = \{Q - Q_0 : Q \in \Delta_C^n\}. \tag{13}$$

This can reduce ambiguity for the recorded event but does not restore global exactness.

Target realization and reuse. For cross entropy in a shared output coordinate system, the chosen representative gives

$$r_* = \frac{P_* - Q_0 - S_{\mathrm{rep}}}{n}, \qquad \widehat{g}_* = J_*^\top \operatorname{vec}(r_*).$$

The target obtains $\widehat{g}_*$ by an ordinary backward pass and applies its own optimizer to produce the update in its own parameter space. Thus UT does not transport a source parameter displacement; the source record supplies the learning information from which the target forms its own update.

Storing the corrected measurement with a source-independent representation of its measurement map allows reuse without loading the source model or its training state. The record can be used at later target states whenever it remains sufficient for the corresponding target response. Optimizer correction, known label mappings, record compilation, coverage checks, constrained solvers, and temporal reuse are detailed in Appendix C.

6 Experiments

We test whether recorded training transitions yield the supervision measurements of Section 3, whether the information they retain determines a specified target response as predicted by Section 4, and whether UT can construct and later reuse updates across heterogeneous architectures. The sufficiency tests include cases in which the original supervision remains ambiguous.

For each learning event, the event inputs and source transitions are available to UT, but the event supervision is withheld. That supervision generates the recorded transitions and supplies the independent reference measurements and target updates. We report relative measurement error e_{meas}, target-update error e_θ, and trajectory error E_t; task metrics are reported where they clarify the corresponding learning effect. Metric definitions, experimental settings, and supporting results are given in Appendix D.

6.1 Recorded Transitions as Supervision Measurements

On Oxford-IIIT Pet (Parkhi et al., 2012), ConvNeXt-Tiny (Liu et al., 2022c), Swin-T (Liu et al., 2021), and ViT-B/32 (Dosovitskiy et al., 2021) each process three fixed learning events under Momentum or AdamW (Loshchilov & Hutter, 2019), giving 18 recorded transitions. For each transition, the recorded optimizer state is used to isolate the current-event gradient and form the corrected retained classifier-head measurement $\widehat{y}_m$. We compare it with the reference $A_m B_m s$ computed independently from the event supervision, while a weights-only control uses the same parameter transition without optimizer-state correction.

Table 1: **Measurement errors from recorded Pet transitions.** Results aggregate three architectures over three fixed 800-example Pet events per optimizer. Entries are relative measurement errors $e_{\mathrm{meas}} \downarrow$.

Optimizer	Corrected median	Corrected maximum	Weights-only median
Momentum	1.08×10^{-7}	1.65×10^{-7}	1.057
AdamW	1.07×10^{-7}	3.26×10^{-7}	46.353

Across all 18 corrected transitions, the error against the predicted measurement is at most 3.26×10^{-7}. Reading the same transitions from parameter changes alone instead produces order-one errors under Momentum and errors in the tens under AdamW. The small corrected residuals agree with the exact supervision-measurement relation derived in Section 3.

Target-side controls examine consecutive events and stochastic computations: omitting optimizer history or mismatching the realized computation increases target-update error. The source-measurement relation also holds in a separate 20-event mini-batch AdamW stream. Results are reported in Appendix D.1.

Table 2: **Target-relative information sufficiency.** Panel A compares matched and mismatched record–target pairs for a whole-network EfficientNet-B0 target on CIFAR-100. Panel B fixes one rank-deficient Sensorless source record and changes the target requirement.

A. Deep CIFAR-100: matched and mismatched record–target pairs

Official-20 has full sample rank $(800/800)$, while its class projection has a 64,000-dimensional kernel in the fine-supervision contrast space.

Record	Target	Fine-100 supervision	$e_\theta \downarrow$
Full-100	Fine-100	Identified	1.37×10^{-6}
Official-20	Official-20	Ambiguous	1.04×10^{-6}
Official-20	Fine-100	Ambiguous	2.768
Random-20	Official-20	Ambiguous	2.748

B. Sensorless: fixed record, changed target

Same rank-$433/512$ record; 79 sample-space directions remain unresolved.

Target requirement	Theory	$e_\theta \downarrow$
Inside pooled source span	Determined	2.83×10^{-14}
Adds source-invisible directions	Obstructed	5.36×10^{-2}

6.2 Target-Relative Information Sufficiency

To test whether ambiguity left by a record can change the specified target response, we vary the retained class information and the target requirement. On CIFAR-100 (Krizhevsky & Hinton, 2009), MobileNetV3-small (Howard et al., 2019) and ResNet-18 (He et al., 2016) process the same 800-example fine-100 learning event, and we construct matched and mismatched record–target pairs.

The official-20 record leaves the original fine-100 supervision non-unique, yet reconstructs the corresponding whole-network EfficientNet-B0 (Tan & Le, 2019) update with $e_\theta = 1.04 \times 10^{-6}$ (Table 2A). The same record gives an error of 2.768 when the target requires fine-100 distinctions; an equally sized random-20 record gives an error of 2.748 for the official-20 target. In these tests, record size alone does not determine whether a target's update can be reconstructed.

Sensorless Drive (Bator, 2013) isolates the same criterion with the source record held fixed (Table 2B). A target whose required directions lie in the pooled source span has $e_\theta = 2.83 \times 10^{-14}$; adding source-invisible directions raises the error to 5.36×10^{-2}.

These controls support the distinction in Theorem 4.1: a record can determine a target response without resolving supervision distinctions to which that target is insensitive. In the obstructed cases, explicit source-null directions change the target response.

6.3 Cross-Architecture Update Construction and Reuse

Target-native construction. On Oxford-IIIT Pet (Parkhi et al., 2012), ConvNeXt-Tiny (Liu et al., 2022c), Swin-T (Liu et al., 2021), and ViT-B/32 (Dosovitskiy et al., 2021) rotate as the target, while the other two architectures provide only classifier-head records from the same 800-example event. UT constructs all trainable target parameters, spanning 27.55M–87.48M coordinates with median relative errors from 4.17×10^{-5} to 1.20×10^{-4} (Table 3A). UT computes each update through the target's own backward pass, without transferring source parameter coordinates.

Persistent records. Once the source transitions are corrected, their pooled measurement can be compiled at capture time. Cold replay then runs without loading the source checkpoint, optimizer state, or random-number-generator (RNG) state. The compiled-to-raw update errors are 1.051×10^{-5} on Pet and 4.758×10^{-6} on DomainNet (Table 3B). The source training state is needed to interpret the raw transition, but not to reuse the resulting record. The compilation results are detailed in Appendix D.3.2.

Historical replay. Ten CIFAR-100 (Krizhevsky & Hinton, 2009) events are compiled into selective official-20 records and sealed before any future target is instantiated. Replay uses the archived event inputs and records, without accessing source models, training states, or event la-

Table 3: **Target updates and historical replay from retained records. A**, whole-network updates from source classifier-head records. **B**, source-independent FP32 cold replay; errors are e_θ. **C**, archived event inputs and sealed records reused by heterogeneous targets introduced after capture. In C2, task accuracy matches the direct-supervision reference at the displayed precision.

A. Whole-network construction on Pet

$n = 800$; two source-head records; 28,453 retained coordinates/source; pooled rank $800/800$.

Target	Updated params.	Median $e_\theta \downarrow$	Min–max e_θ
ConvNeXt-Tiny	27.85M	6.05×10^{-5}	$[5.15, 6.58] \times 10^{-5}$
Swin-T	27.55M	4.17×10^{-5}	$[2.27, 4.68] \times 10^{-5}$
ViT-B/32	87.48M	1.20×10^{-4}	$[1.05, 1.21] \times 10^{-4}$

B. Compiled cold replay without source state

Runtime audit: zero source opens; no source checkpoint, optimizer state, or RNG state loaded.

Dataset	n / pooled rank	Raw UT / reference	Compiled / reference	Compiled / raw UT
Pet	800 / 800	5.415×10^{-5}	5.449×10^{-5}	1.051×10^{-5}
DomainNet	2070 / 2070	1.906×10^{-4}	1.907×10^{-4}	4.758×10^{-6}

C1. Selectivity controls on the sealed CIFAR archive

Ten historical events; whole-network future targets; no source state or event labels at replay.

Archived information	Future target	Trajectory	Final $E_{10} \downarrow$
Official-20, matched	Official-20 EfficientNet-B0	$\sim 10^{-5}$	$(2.03 \pm 0.10) \times 10^{-5}$
Same official-20 archive	Fine-100 EfficientNet-B0	$O(1)$	1.81 ± 0.50
Random-20 archive	Official-20 EfficientNet-B0	$O(1)$	3.626 ± 0.058

C2. Future-target and temporal breadth

Archive	Future target	Scope	Updates	Final $E_t \downarrow$	Task result
Same sealed CIFAR	ViT-B/32	All params.	10	2.25×10^{-5}	Matches ref.
DomainNet	ViT-L/16	Block+head	20	2.04×10^{-5}	Acc. 6.783% (matches ref.)
DomainNet	ViT-L/16	Block+head	60	3.61×10^{-5}	Acc. 15.073% (matches ref.)
DomainNet	EfficientNet-B0	Block+head	20	1.85×10^{-5}	Acc. 8.232% (matches ref.)
DomainNet	EfficientNet-B0	Block+head	60	3.11×10^{-5}	Acc. 34.493% (matches ref.)

bels. A later whole-network EfficientNet-B0 target remains close to the direct-supervision trajectory, whereas changing either the target requirement or the retained grouping produces order-one error (Table 3C1).

The same sealed CIFAR archive later supports a whole-network ViT-B/32 target. A separate 20-event DomainNet (Peng et al., 2019) archive supports two future architectures; replaying each event three times gives 60 target updates (Table 3C2). The archive results are given in Appendix D.3.3.

The compiled and sealed records produce target updates without source-model access during replay, provided that each record retains the information needed by its target.

7 Discussion

A parameter update belongs to the model that produced it, but the learning information measured by that update can be reused by another architecture. UT separates the two: source transitions yield measurements of a learning event, while a target realizes the retained information through its own parameters. The full original supervision need not be recovered when the record already determines the response of the specified target.

Historical replay tests whether this separation persists without access to the source models. Source-side state is needed to interpret a transition at capture time, but the compiled record can later provide an update for a target instantiated after capture. Such records could, in principle, preserve infor-

mation from rare events, costly annotations, or expert interventions across model upgrades. Their usefulness remains target-dependent: a record sufficient for one target may be insufficient for another. The current construction also requires event inputs. Old checkpoints alone cannot generally be converted into valid records after the fact. Efficiently certifying target-relative coverage for large, rank-deficient deep targets remains open.

The records may contain sensitive supervision information. The present experiments test selective retention and replay without source state; they do not establish compression or privacy guarantees. Determining what a record must retain for future targets, and how to protect the information it carries, are open questions.

8 Conclusion

Updates from heterogeneous architectures cannot generally be transferred through shared parameter coordinates. Under the stated loss and record conditions, correcting source transitions instead yields exact linear measurements of a shared supervision contrast. On a convex feasible set containing zero, those measurements determine the specified target response for every feasible supervision if and only if every source-invisible direction in the feasible difference space is also target-invisible. This can hold even when the original supervision is not identifiable. Given the target computation and specified learning rule, UT turns the determined response into an update through the target's own parameters.

The experiments trace this sequence from optimizer-corrected source measurements through target-relative sufficiency tests to whole-network updates in different architectures. Compiled records also closely track direct-supervision trajectories in targets instantiated after capture, with source training state unavailable at replay. Thus, a learning event can leave information that a later model can use without receiving a source parameter update.

AI use statement

Generative AI tools were used to assist with language editing and manuscript polishing, reviewer-style critique of intermediate drafts, and suggestions for additional robustness and control experiments. They were also used for limited code drafting and refactoring, reuse and adaptation of existing analysis code, and the organization and formatting of appendix tables. All AI-assisted outputs were reviewed and verified by the authors; experimental decisions, reported results, and final scientific claims were determined by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9163–9171, 2019.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023.

Martyna Bator. Dataset for sensorless drive diagnosis. *UCI Machine Learning Repository*, 2013. doi: 10.24432/C5VP5F.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. doi: 10.1561/2200000016.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Huancheng Chen and Haris Vikalo. Recovering labels from local updates in federated learning. In *International Conference on Machine Learning (ICML)*, 2024.

Daiki Chijiwa. Transferring learning trajectories of neural networks. In *International Conference on Learning Representations (ICLR)*, 2024.

Mohamed Darouach. Existence and design of functional observers for linear systems. *IEEE Transactions on Automatic Control*, 45(5):940–943, 2000.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

David L. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1): 238–270, 1994.

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

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients – how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. Software available at `https://github.com/mlfoundations/open_clip`, 2021.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, Jingting Ding, and Liang Li. Cross-architecture knowledge distillation. In *Asian Conference on Computer Vision (ACCV)*, 2022a.

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

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12009–12019, 2022b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Arthur N. Montanari, Chao Duan, Luis A. Aguirre, and Adilson E. Motter. Functional observability and target state estimation in large-scale networks. *Proceedings of the National Academy of Sciences (PNAS)*, 119(1):e2113750119, 2022.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Filippo Rinaldi, Giacomo Capitani, Lorenzo Bonicelli, Donato Crisostomi, Federico Bolelli, Elisa Ficarra, Emanuele Rodolà, Simone Calderara, and Angelo Porrello. Update your Transformer to the latest release: Re-Basin of task vectors. In *International Conference on Machine Learning (ICML)*, 2025.

Filippo Rinaldi, Aniello Panariello, Giacomo Salici, Fengyuan Liu, Marco Ciccone, Angelo Porrello, and Simone Calderara. Gradient-sign masking for task vector transport across pre-trained models. In *International Conference on Learning Representations (ICLR)*, 2026a.

Filippo Rinaldi, Aniello Panariello, Giacomo Salici, Angelo Porrello, and Simone Calderara. Transporting task vectors across different architectures without training. *arXiv preprint arXiv:2602.12952*, 2026b.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. iDLG: Improved deep leakage from gradients. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations (ICLR)*, 2021.

Luca Zhou, Daniele Solombrino, Donato Crisostomi, Maria Sofia Bucarelli, Giuseppe Alessio D'Inverno, Fabrizio Silvestri, and Emanuele Rodolà. On task vectors and gradients. In *Proceedings of UniReps: the Third Edition of the Workshop on Unifying Representations in Neural Models*, volume 322 of *Proceedings of Machine Learning Research*, pp. 398–417, 2026.

Junyi Zhu, Ruicong Yao, and Matthew B. Blaschko. Surrogate model extension (SME): A fast and accurate weight update attack on federated learning. In *International Conference on Machine Learning (ICML)*, pp. 43228–43257, 2023.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

A Heterogeneous Update Measurements: Derivations

A.1 Exact supervision measurement

Recall

$$\nabla_{z_m} L_m = a_m - D_m \psi(u),$$

where L_m is the averaged event data loss, with all quantities evaluated at the same recorded pre-update computation. Optimizer regularization and decoupled weight decay are handled separately when a transition is corrected. The chain rule gives

$$g_m(u) = J_m^\top \nabla_{z_m} L_m = J_m^\top [a_m - D_m \psi(u)]\,. \tag{14}$$

For the reference supervision u_0,

$$\begin{aligned} g_m(u_0) - g_m(u) &= J_m^\top D_m \left[\psi(u) - \psi(u_0)\right] \\ &= J_m^\top D_m s, \end{aligned} \tag{15}$$

where $s = \psi(u) - \psi(u_0)$. Multiplying by the known normalization ρ_m gives

$$b_m = \rho_m [g_m(u_0) - g_m(u)] = \rho_m J_m^\top D_m s = B_m s, \tag{16}$$

with $B_m = \rho_m J_m^\top D_m$. This proves Proposition 3.1 without linearizing the network: J_m, a_m, and D_m come from the same recorded computation for both u and u_0.

A.2 From recorded transitions to measurements

For vanilla SGD,

$$\theta_m^+ - \theta_m = -\eta_m g_m(u),$$

where θ_m and θ_m^+ are the parameters before and after the recorded event. Substituting into (16) gives

$$b_m = \frac{\rho_m}{\eta_m}(\theta_m^+ - \theta_m) + \rho_m g_m(u_0), \tag{17}$$

which is (3).

For a stateful optimizer, the parameter difference alone need not identify the event gradient. The momentum and AdamW states used here isolate it as

$$\begin{aligned} v^+ &= \mu v^- + g + \lambda\theta, & g &= v^+ - \mu v^- - \lambda\theta, \\ m^+ &= \beta_1 m^- + (1-\beta_1) g, & g &= \frac{m^+ - \beta_1 m^-}{1-\beta_1}, \end{aligned} \tag{18}$$

for momentum with coupled weight decay and the AdamW first-moment transition (Loshchilov & Hutter, 2019), respectively. Here $v^\pm$ are momentum buffers and $m^\pm$ are first-moment states before and after the event.

The record also specifies any preprocessing, model mode, and realized randomness that affects the forward or backward map. The corrected measurement is then interpreted through B_m.

A.3 Loss instantiations

Cross entropy. For average soft-label cross entropy,

$$L_m = -\frac{1}{n} \sum_{i,c} Q_{ic} \log P_{m,ic},$$

where $Q \in \Delta_C^n$ is the supervision matrix and P_m contains the source probabilities. Since each row of Q sums to one,

$$\frac{\partial L_m}{\partial F_{m,ic}} = \frac{P_{m,ic} - Q_{ic}}{n}. \tag{19}$$

Thus

$$\psi(u) = \mathrm{vec}(Q), \qquad a_m = \frac{\mathrm{vec}(P_m)}{n}, \qquad D_m = \frac{I}{n}.$$

Let

$$Q_0 = \frac{1}{C}\mathbf{1}_n\mathbf{1}_C^\top, \qquad S = Q - Q_0.$$

With $\rho_m = n$,

$$b_m = n[g_m(Q_0) - g_m(Q)] = J_m^\top \operatorname{vec}(S). \tag{20}$$

For vanilla SGD this becomes

$$b_m = \frac{n}{\eta_m}(\theta_m^+ - \theta_m) + J_m^\top \operatorname{vec}(P_m - Q_0), \tag{21}$$

recovering (4).

A known fixed mapping of the supervision can be absorbed into D_m, allowing different architectures to measure known linear views of the same supervision.

Mean-squared error. For

$$L_m = \frac{1}{2n}\|z_m - u\|_2^2,$$

we have

$$\nabla_{z_m} L_m = \frac{z_m - u}{n}.$$

Taking $\psi(u) = u$, $D_m = I/n$, and $\rho_m = n$ yields

$$b_m = J_m^\top(u - u_0). \tag{22}$$

More generally, for the fixed weighted quadratic loss with $W_m = W_m^\top \succeq 0$,

$$L_m = \frac{1}{2n}(z_m - u)^\top W_m(z_m - u),$$

the measurement becomes

$$b_m = J_m^\top W_m(u - u_0).$$

Only the output gradient's dependence on supervision must be affine; the neural network need not be linear.

A.4 Retained source records and the target response

A source may retain only a known linear part of its complete measurement. For a retention map A_m,

$$y_m = A_m b_m = A_m B_m s. \tag{23}$$

Stacking M source records gives

$$y = R(s) = \big(A_1 B_1 s, \ldots, A_M B_M s\big), \tag{24}$$

which is (5).

For a target model at its current state and a known normalization $\rho_* > 0$, the same calculation gives

$$T(s) = \rho_* J_*^\top D_* s, \qquad g_*(u) = g_*(u_0) - \rho_*^{-1} T(s). \tag{25}$$

The specified target learning rule and optimizer state convert this gradient to a parameter update.

B Target-Relative Observability: Proofs and Extensions

B.1 Proof of the observability criterion

Proof of Theorem 4.1. Write

$$\mathcal{L} = \operatorname{span}(\mathcal{K} - \mathcal{K}), \qquad R = R_{\mathcal{L}}, \qquad T = T_{\mathcal{L}}.$$

Suppose first that $\ker R \subseteq \ker T$. For any $s_1, s_2 \in \mathcal{K}$,

$$R(s_1) = R(s_2) \quad \Longrightarrow \quad s_1 - s_2 \in \ker R \quad \Longrightarrow \quad s_1 - s_2 \in \ker T,$$

and hence $T(s_1) = T(s_2)$.

Conversely, suppose equal source records always imply equal target responses. Choose $s_0 \in \mathrm{relint}(\mathcal{K})$ and $d \in \ker R$. Since $d \in \mathcal{L}$, relative interiority gives $s_0 + \epsilon d \in \mathcal{K}$ for all sufficiently small $\epsilon > 0$. Because

$$R(s_0 + \epsilon d) = R(s_0),$$

observability implies $T(s_0 + \epsilon d) = T(s_0)$, and therefore $T(d) = 0$. Thus $\ker R \subseteq \ker T$.

Finally, $R^\dagger R$ is the orthogonal projector onto $(\ker R)^\perp$. Therefore

$$\ker R \subseteq \ker T \quad \Longleftrightarrow \quad T(I - R^\dagger R) = 0 \quad \Longleftrightarrow \quad T = TR^\dagger R. \tag{26}$$

For $y = R(s)$, this yields

$$T(s) = TR^\dagger y.$$

If kernel inclusion fails, some $d \in \ker R$ satisfies $T(d) \neq 0$. The same relative-interior construction then produces two feasible supervision contrasts with identical source records but different target responses. Hence no rule depending only on y can be correct over all of $\mathcal{K}$. □

When $0 \in \mathcal{K}$, no reference offset is needed. For any feasible reference s_0, applying the argument to $\mathcal{K} - s_0$ gives

$$T(s) = T(s_0) + T_{\mathcal{L}} R_{\mathcal{L}}^\dagger \big[R(s) - R(s_0)\big]. \tag{27}$$

Event-specific determination. For one observed record y, target determination requires only its feasible fiber:

$$\mathcal{K}_y := \{s \in \mathcal{K} : R(s) = y\}.$$

That event is target-determined exactly when

$$T(s_1) = T(s_2), \qquad \forall\, s_1, s_2 \in \mathcal{K}_y. \tag{28}$$

Section 5 uses this condition for constrained construction.

B.2 Retained records and the feature-space criterion

A retained record may contain only a fixed linear image of a complete measurement. If $b = Bs$ is replaced by

$$y = ABs,$$

then

$$\ker B \subseteq \ker(AB). \tag{29}$$

Retention cannot reduce ambiguity. It determines a target T precisely when

$$\ker(AB) \subseteq \ker T$$

on the feasible difference space. Invertible changes of retained coordinates leave this condition unchanged.

For linear-head cross-entropy measurements, identify the supervision contrast with the centered matrix

$$S = Q - Q_0, \qquad \mathcal{L} = \{S \in \mathbb{R}^{n \times C} : S\mathbf{1}_C = 0\}.$$

Let $X_{\mathcal{S}}$ denote the pooled source design and X_* the target design. The corresponding maps are

$$R(S) = X_{\mathcal{S}}^\top S, \qquad T(S) = X_*^\top S.$$

Proof of Corollary 4.2. The standard null-space and column-space identities give

$$\ker X_{\mathcal{S}}^\top \subseteq \ker X_*^\top \quad \Longleftrightarrow \quad \mathrm{col}(X_*) \subseteq \mathrm{col}(X_{\mathcal{S}}) \quad \Longleftrightarrow \quad \mathrm{rank}([X_{\mathcal{S}}, X_*]) = \mathrm{rank}(X_{\mathcal{S}}). \tag{30}$$

It remains to show that failure of the inclusion is realized within the probability domain. Suppose

$$v \in \ker X_{\mathcal{S}}^\top, \qquad X_*^\top v \neq 0.$$

Choose $a \neq 0$ with $a^\top \mathbf{1}_C = 0$ and set

$$Q_\pm = Q_0 \pm tva^\top, \qquad 0 < t \leq \left(C \max_{i,c} |v_i a_c| \right)^{-1}. \tag{31}$$

Then $Q_\pm \in \Delta_C^n$ and

$$X_{\mathcal{S}}^\top (Q_+ - Q_-) = 0,$$

whereas

$$X_*^\top (Q_+ - Q_-) = 2t(X_*^\top v)a^\top \neq 0.$$

Hence identical source-head records can produce different target-head responses whenever the span condition fails. □

For unprojected head measurements in a common label space, condition

$$\operatorname{rank}(X_{\mathcal{S}}) = n \quad \Longrightarrow \quad \ker X_{\mathcal{S}}^\top = \{0\} \tag{32}$$

identifies the supervision contrast and hence any fixed target response. For a retained or class-projected record, supervision identification requires $\ker R_{\mathcal{L}} = \{0\}$. Target-relative coverage does not require this.

B.3 Approximation and noisy records

In this subsection, R and T denote their restrictions to the feasible difference space. Let

$$y = R(s) + e, \qquad P_R := R^\dagger R,$$

where e is record error. The pseudoinverse construction satisfies

$$TR^\dagger y - T(s) = -T(I - P_R)s + TR^\dagger e. \tag{33}$$

Taking norms gives

$$\|TR^\dagger y - T(s)\| \leq \|T(I - R^\dagger R)s\| + \|TR^\dagger\|_{\mathrm{op}} \|e\|, \tag{34}$$

which yields (12) in the main text.

The corresponding worst-case error is minimal among linear maps:

$$\inf_{H \text{ linear}} \|T - HR\|_{\mathrm{op}} = \|T(I - R^\dagger R)\|_{\mathrm{op}}. \tag{35}$$

Indeed, every $v \in \ker R$ satisfies $(T - HR)v = Tv$, while $H = TR^\dagger$ attains the bound.

The same ambiguity limits arbitrary estimators, including nonlinear ones. Define

$$\omega_{R,T}(\delta) := \sup_{\substack{s_1, s_2 \in \mathcal{K} \\ \|R(s_1 - s_2)\| \leq \delta}} \|T(s_1 - s_2)\|. \tag{36}$$

Under bounded observation error $\|e\| \leq \epsilon$,

$$\inf_{\widehat{t}} \sup_{\substack{s \in \mathcal{K} \\ \|e\| \leq \epsilon}} \left\| \widehat{t}(Rs + e) - T(s) \right\| \geq \frac{1}{2} \omega_{R,T}(2\epsilon). \tag{37}$$

To see this, take any pair satisfying $\|R(s_1 - s_2)\| \leq 2\epsilon$. Their midpoint record

$$y_0 = \frac{R(s_1) + R(s_2)}{2}$$

is compatible with either supervision under admissible noise, so the triangle inequality forces error at least $\|T(s_1 - s_2)\|/2$ on one of the two cases.

At $\epsilon = 0$, any equal-record pair with different target responses rules out exact reconstruction by every estimator.

B.4 Local extension beyond affine supervision

For arbitrary maps

$$Y : \mathcal{K}_u \to \mathcal{Y}, \qquad G : \mathcal{K}_u \to \mathcal{G},$$

G is determined by Y if and only if it is constant on every fiber of Y. Equivalently, a map H on the reachable records satisfies

$$G = H \circ Y.$$

For a local analogue, let $\mathcal{L}_u := \operatorname{span}(\mathcal{K}_u - \mathcal{K}_u)$ and assume that Y and G are twice continuously differentiable near $u_0 \in \operatorname{relint}(\mathcal{K}_u)$. All derivatives and pseudoinverses below are restricted to this feasible difference space. Define

$$D_Y := DY(u_0)|_{\mathcal{L}_u}, \qquad D_G := DG(u_0)|_{\mathcal{L}_u}, \qquad H_0 := D_G D_Y^{\dagger}.$$

For a feasible perturbation δ and observation error e, consider

$$\widehat{G} = G(u_0) + H_0\big[Y(u_0 + \delta) + e - Y(u_0)\big]. \tag{38}$$

Writing

$$Y(u_0 + \delta) - Y(u_0) = D_Y\delta + r_Y, \qquad G(u_0 + \delta) - G(u_0) = D_G\delta + r_G,$$

gives

$$G(u_0 + \delta) - \widehat{G} = D_G(I - D_Y^{\dagger}D_Y)\delta + r_G - H_0 r_Y - H_0 e. \tag{39}$$

If

$$\|r_Y\| \le \frac{M_Y}{2}\|\delta\|^2, \qquad \|r_G\| \le \frac{M_G}{2}\|\delta\|^2,$$

then

$$\begin{aligned}\|G(u_0 + \delta) - \widehat{G}\| \le{}& \|D_G(I - D_Y^{\dagger}D_Y)\delta\| \\ &+ \frac{1}{2}\left(M_G + \|H_0\|_{\mathrm{op}}M_Y\right)\|\delta\|^2 + \|H_0 e\|.\end{aligned} \tag{40}$$

The first-order term vanishes for every feasible perturbation exactly when

$$\ker D_Y \subseteq \ker D_G \quad \Longleftrightarrow \quad D_G(I - D_Y^{\dagger}D_Y) = 0. \tag{41}$$

The noise-free local error is then $O(\|\delta\|^2)$. If instead $v \in \ker D_Y$ but $D_G v \neq 0$, the perturbation $\delta = rv$ produces a nonzero first-order $O(r)$ target change while the source record changes only at second order.

If DY has constant rank and

$$\ker DY(u) \subseteq \ker DG(u)$$

throughout a sufficiently small relative neighborhood, the constant-rank theorem further yields a local factorization

$$G = H \circ Y$$

on the reachable records.

C Update Tomography: Construction Details

For Algorithm 1, assume event inputs and either corrected source measurements or source transitions with the states needed for correction (Appendix A.2).

C.1 Correct: forming the source measurements

For source m, UT uses the retained measurement

$$y_m = \rho_m A_m\big[g_m(u_0) - g_m(u)\big] = A_m B_m s. \tag{42}$$

Here A_m is the retention map chosen for that record. It may keep the full measurement, select coordinates, or apply a known linear projection.

Once corrected, UT uses y_m and its measurement map without the raw parameter transition.

C.2 POOL AND COMPILE: FORMING THE EVENT RECORD

The corrected source measurements are pooled as

$$y = R(s) = (A_1 B_1 s, \dots, A_M B_M s). \tag{43}$$

The event record contains the inputs, pooled measurement, a representation of the measurement map, and metadata needed to interpret the supervision coordinates.

For an unprojected linear-head record in a common label space, full row rank of the pooled design identifies the supervision contrast, and thus any fixed target response. For a general record, supervision identification requires $\ker R_{\mathcal{L}} = \{0\}$. For a rank-deficient linear-head record, the feature-span criterion in Corollary 4.2 gives the target-specific test. More generally, if the columns of N span $\ker R_{\mathcal{L}}$, then

$$\ker R_{\mathcal{L}} \subseteq \ker T_{\mathcal{L}} \quad \Longleftrightarrow \quad T_{\mathcal{L}} N = 0. \tag{44}$$

A source-null direction with nonzero target response disproves coverage; finitely many sampled directions cannot certify it.

Source-independent compilation. For the pooled linear-head case, write the retained measurement as

$$Y = X^\top S,$$

where $X \in \mathbb{R}^{n \times d}$ is the pooled source design. Let

$$X = U_r \Sigma_r V_r^\top \tag{45}$$

be its rank-r thin SVD. Multiplying the record by the known inverse factors gives

$$Z := \Sigma_r^{-1} V_r^\top Y = U_r^\top S. \tag{46}$$

Thus the source equations

$$X^\top S = Y$$

are equivalent to

$$U_r^\top S = Z.$$

Storing (U_r, Z) with the inputs and supervision-coordinate metadata preserves the feasible supervision set and its target-relative information; the source model, optimizer, and RNG state are not retained.

A known class mapping is handled in the same record. Let $H \in \mathbb{R}^{C \times C'}$ map the source supervision coordinates to the target coordinates. From $Y = X^\top S$, the retained and compiled mapped records are

$$Y_H = YH = X^\top SH, \qquad Z_H = \Sigma_r^{-1} V_r^\top Y_H = U_r^\top SH. \tag{47}$$

A C'-class target uses $\bar{S}_{\mathrm{rep}} = U_r Z_H$ in

$$r_* = \frac{P_* - Q_0 H - \bar{S}_{\mathrm{rep}}}{n}. \tag{48}$$

The reference $Q_0 H$ is uniform over the target classes only for equally sized groups, as in the five-to-one official CIFAR-100 grouping used below.

C.3 REPRESENT: DIRECT AND CONSTRAINED CONSTRUCTION

Under target-relative coverage, any $s_{\mathrm{rep}} \in \mathcal{L}$ satisfying

$$R(s_{\mathrm{rep}}) = y$$

produces the correct target response. A convenient minimum-norm choice is

$$s_{\mathrm{rep}} = R_{\mathcal{L}}^{\dagger} y. \tag{49}$$

For the compiled linear-head record in (46), the corresponding representative is simply

$$S_{\mathrm{rep}} = U_r Z. \tag{50}$$

When global coverage is not established, classification restricts the search to valid centered probability matrices,

$$\Omega = \{Q - Q_0 : Q \in \Delta_C^n\},$$

and solves

$$S_\Omega^\star \in \arg\min_{V \in \Omega} \frac{1}{2}\|R(V) - y\|_2^2. \tag{51}$$

For a compiled record, the original least-squares objective satisfies

$$\|X^\top V - Y\|_F^2 = \|\Sigma_r(U_r^\top V - Z)\|_F^2 + \|(I - V_r V_r^\top)Y\|_F^2. \tag{52}$$

The second term is independent of V. The factor Σ_r preserves the original residual metric; omitting it whitens that metric. In the exact feasible case, both have the same zero-residual solutions. Under noise, quantization, or nonzero optimal residual, they can select different representatives. The constrained experiments scale each source by its spectral norm and minimize the original residuals, equivalently applying (52) to the pooled scaled record. Restricting the search to Ω can reduce ambiguity for the observed event without changing the global coverage condition.

For head measurements $R(V) = X^\top V$, we use the alternating direction method of multipliers (ADMM) (Boyd et al., 2011) with the split $V = Z_c$, $Z_c \in \Omega$:

$$\begin{aligned}(XX^\top + \gamma I)V^{k+1} &= XY + \gamma(Z_c^k - U^k),\\ Z_c^{k+1} &= \Pi_\Omega(V^{k+1} + U^k),\\ U^{k+1} &= U^k + V^{k+1} - Z_c^{k+1},\end{aligned} \tag{53}$$

where $\gamma > 0$ is the penalty parameter and Π_Ω denotes rowwise projection onto the centered probability domain. The implementation uses the equivalent probability coordinate $Q = V + Q_0$, initialized at Q_0 with zero dual variable. We set $\gamma = 0.1$, use relative primal and dual tolerances 10^{-7} in the standard ADMM residual tests, and cap the solver at 20,000 iterations. A projected minimum-norm candidate is accepted when its source-normalized objective is at most 10^{-20} times the measurement scale.

For $S \in \Omega$, let

$$y = R(S) + e, \qquad \|e\| \le \epsilon,$$

and define the event-specific target ambiguity

$$\omega_S(\delta) = \max_{\substack{V \in \Omega \\ \|R(V-S)\| \le \delta}} \|T(V - S)\|.$$

Theorem C.1 (Constrained representative stability)**.** *Let $f(V) = \|R(V) - y\|^2/2$. If $\widetilde{S} \in \Omega$ satisfies*

$$f(\widetilde{S}) \le \min_{V \in \Omega} f(V) + \frac{\tau^2}{2},$$

then

$$\|R(\widetilde{S} - S)\| \le \epsilon + \tau, \qquad \|T(\widetilde{S} - S)\| \le \omega_S(\epsilon + \tau). \tag{54}$$

Proof. Let $\mathcal{C} = R(\Omega - S)$ and $z^\star = P_{\mathcal{C}}(e)$. Projection optimality gives $\|z^\star\| \le \epsilon$, while the optimization gap gives

$$\|R(\widetilde{S} - S) - z^\star\| \le \tau.$$

Hence $\|R(\widetilde{S} - S)\| \le \epsilon + \tau$, and the target bound follows from the definition of ω_S. □

C.4 Realize: forming the target update

Given s_{rep}, the target evaluates the learning signal at its current state. In the general affine setting,

$$\widehat{g}_* = g_*(u_0) - \rho_*^{-1} T(s_{\mathrm{rep}}). \tag{55}$$

For cross entropy,

$$r_* = \frac{P_* - Q_0 - S_{\mathrm{rep}}}{n}, \qquad \widehat{g}_* = J_*^\top \operatorname{vec}(r_*). \tag{56}$$

The Jacobian–transpose product is evaluated by an ordinary backward pass; a full neural-network Jacobian is never formed. The target then applies its own optimizer state, step counter, hyperparameters, and decay convention to $\widehat{g}_*$. For a known label mapping, the target instead uses the mapped signal in (48).

C.5 TEMPORAL USE OF AN EVENT RECORD

The record stores an event measurement; the parameter displacement depends on the target's current state. When the record is reused after that state changes, UT recomputes the target response.

Let x_k collect the target parameters and continuous optimizer state, and let Φ_k and $\widehat{\Phi}_k$ denote the supervised and UT transitions at step k under the same realized target randomness. Assume

$$\|\Phi_k(\widehat{x}_k) - \Phi_k(x_k)\| \le L_k \|\widehat{x}_k - x_k\|.$$

Define

$$e_k := \|\widehat{x}_k - x_k\|, \qquad \delta_k := \|\widehat{\Phi}_k(\widehat{x}_k) - \Phi_k(\widehat{x}_k)\|.$$

Theorem C.2 (Fixed-event trajectory bound)**.** *For $K \ge 1$,*

$$e_K \le \left(\prod_{\ell=0}^{K-1} L_\ell \right) e_0 + \sum_{j=0}^{K-1} \delta_j \prod_{\ell=j+1}^{K-1} L_\ell. \tag{57}$$

In particular, identical initial states and $\delta_k = 0$ at every reached state yield identical trajectories.

Proof. Adding and subtracting $\Phi_k(\widehat{x}_k)$ gives $e_{k+1} \le L_k e_k + \delta_k$. Iterating this recursion yields (57). □

No contraction is required. Exact reuse requires the record to determine the target response at every reached state, not only at the initial state.

For successive learning events, each event supplies a record and representative. The target parameters and optimizer state continue from the preceding update. Reference target checkpoints are used only for evaluation and are never inputs to UT.

D EXPERIMENTAL DETAILS AND SUPPORTING RESULTS

D.1 RECORDED TRANSITIONS AS SUPERVISION MEASUREMENTS

Evaluation quantities. We use one source-side error and three target-side errors. For a source model, let $\widehat{y}_m = A_m \widehat{b}_m$ denote the retained classification-head measurement obtained from its recorded transition after the optimizer-specific correction. The matching oracle is $A_m B_m s$, computed directly from the pre-update head design and the formal supervision contrast $s = \mathrm{vec}(Q - Q_0)$. For a target parameter set, let $\Delta\theta_*$ be the supervised reference and $\widehat{\Delta\theta}_*$ its constructed counterpart. We use

$$e_{\mathrm{meas}} = \frac{\|\widehat{y}_m - A_m B_m s\|_2}{\|A_m B_m s\|_2}, \tag{58}$$

$$e_\theta = \frac{\|\widehat{\Delta\theta}_* - \Delta\theta_*\|_2}{\|\Delta\theta_*\|_2}, \qquad e_F = \frac{\|\Delta\widehat{F}_* - \Delta F_*\|_F}{\|\Delta F_*\|_F}. \tag{59}$$

Here ΔF_* and $\Delta\widehat{F}_*$ are the changes in the target logits induced by the supervised reference and constructed updates, respectively. They are evaluated either on the event inputs or on a disjoint held-out set, as stated with each experiment. For trajectories starting from the same target state θ_*^0, we use

$$E_t = \frac{\|\widehat{\theta}_*^t - \theta_{*,Q}^t\|_2}{\|\theta_{*,Q}^t - \theta_*^0\|_2}, \qquad t \ge 1. \tag{60}$$

A zero denominator is undefined. All four errors are reported as ratios. Accuracy and any other percentage-valued columns are marked explicitly. Where task metrics are reported, negative log-likelihood (NLL) is the mean true-class negative log probability, and the Brier score is the per-example sum of squared class-probability errors, averaged over examples.

Direct measurement realization. Both terms in e_{meas} subtract the same known uniform reference, but are computed independently. The record-derived side uses pre/post parameters and optimizer states; the oracle side uses the pre-update design and event supervision. Supervision generates the source event and scores the oracle, but is not an input to the record-derived calculation.

The experiment uses three fixed 800-example Pet events. Each of ConvNeXt-Tiny, Swin-T, and ViT-B/32 is evaluated under Momentum and AdamW after three warm-up steps on 128 separate examples. Momentum uses learning rate 10^{-2} and coefficient 0.9; AdamW uses learning rate 3×10^{-4}, $\beta = (0.9, 0.999)$, and $\epsilon = 10^{-8}$. Both use weight decay 10^{-4}. Model transitions are executed in FP32, while extracted states and the measurement calculations are evaluated in FP64.

Table 4: Complete source-measurement realization results. Entries are e_{meas} for the three fixed Pet partitions.

Reading	Source	Partition 1	Partition 2	Partition 3
Momentum				
Corrected	ConvNeXt	1.0860×10^{-7}	1.0824×10^{-7}	1.0619×10^{-7}
Corrected	Swin	1.0531×10^{-7}	9.9392×10^{-8}	9.9626×10^{-8}
Corrected	ViT	1.5790×10^{-7}	1.6522×10^{-7}	1.5132×10^{-7}
Weights only	ConvNeXt	1.0708	1.2687	1.0272
Weights only	Swin	1.1903	1.2294	1.0567
Weights only	ViT	0.8714	0.8653	0.8746
AdamW				
Corrected	ConvNeXt	9.9793×10^{-8}	9.8866×10^{-8}	1.0015×10^{-7}
Corrected	Swin	1.0448×10^{-7}	1.1010×10^{-7}	1.0661×10^{-7}
Corrected	ViT	2.2967×10^{-7}	3.2564×10^{-7}	2.0850×10^{-7}
Weights only	ConvNeXt	44.3912	43.9844	43.0311
Weights only	Swin	47.1064	46.2820	46.3528
Weights only	ViT	52.1217	51.8550	49.7618

Table 4 shows that state correction keeps all residuals near numerical precision; reading the same transitions from weights alone gives much larger errors, especially under AdamW.

Effects of state correction on target updates. The one-transition Pet study rotates the three architectures through the target role over three partitions and two optimizers, giving 18 target construction cases. These are distinct from the 18 source transitions in Table 4. Three warm-up steps on 128 examples precede one 800-example event; the target updates its final block and classifier, and e_F is computed from its logit change on 512 disjoint held-out images. The weights-only control keeps the checkpoints, event inputs, and parameter changes fixed but omits the optimizer-state correction.

The four-transition study uses ConvNeXt and Swin as sources and ViT as the target. Its six cases comprise three fixed partitions under Momentum and AdamW. Each case contains four disjoint 400-example Pet events; the target and optimizer states advance after every event, and the four records are read without resetting the target. The DomainNet control uses five $n = 2070$ repetitions after three balanced warm-up steps. The active ConvNeXt/Swin/ViT parameter counts are 19.49M/12.97M/12.95M, and the target advances with its own AdamW state.

The realized-randomness control uses nine Pet target–partition pairs at $n = 800$. The active ConvNeXt and Swin blocks have stochastic-depth probabilities 0.1 and 0.2, while the selected ViT block has no dropout. We compare deterministic computation, reuse of the recorded random state, and source-only, target-only, or combined random-state mismatches. Source mismatch changes the recorded prediction and measurement map; target mismatch changes the backward map used to evaluate the update. A target-only mismatch leaves the selected ViT case unchanged.

Table 5: Effects of optimizer and realized computation state on target construction. Panel A reports median error ratios with case counts shown; Panel B aggregates nine Pet target–partition pairs.

A. Optimizer state

Setting	Metric	Cases	**Corrected**	Weights only
Pet, one transition	e_θ	18	2.76×10^{-6}	1.12
Pet, four transitions	E_4	6	2.94×10^{-5}	1.46
DomainNet, AdamW	e_θ	5	6.08×10^{-6}	1.33

B. Realized randomness: Pet, $n = 800$

Computation	Median e_θ	Median held-out e_F
Deterministic	1.46×10^{-4}	4.52×10^{-6}
Recorded RNG reused	1.55×10^{-4}	5.18×10^{-6}
Source RNG mismatch	0.108	0.0203
Target RNG mismatch	0.119	0.0202
Both mismatched	0.174	0.0487

Using the recorded optimizer state reduces the one-transition error in all 18 Pet target cases, and the same separation persists across four transitions and the larger DomainNet control. Reusing the realized random state gives errors comparable to the deterministic control, whereas a mismatch produces nonzero error whenever it changes the active source or target computation. Figure 2 shows the error after each of the four transitions.

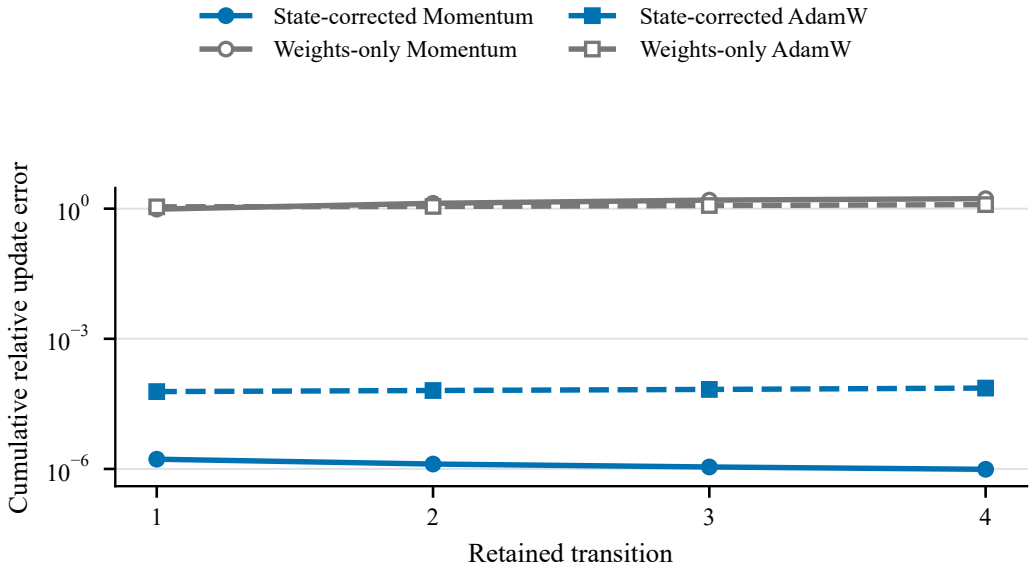


Figure 2: State correction across four retained Pet transitions. Corrected Momentum and AdamW records keep cumulative error low, whereas reading the same transitions from weights alone remains order one.

Mini-batch AdamW stream. On CIFAR-100, MobileNetV3-small and ResNet-18 classifier heads process 20 disjoint 128-example events. Their parameters and AdamW states advance continuously, and adjacent first-moment states isolate the gradient of each current event. A whole-network EfficientNet-B0 target and its AdamW state also advance continuously; UT and the direct-supervision reference start each event from their corresponding preceding target states. All three optimizers use learning rate 3×10^{-4}, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and weight decay 10^{-4}. The weights-only control reads the same source parameter changes without the AdamW history.

Table 6: CIFAR-100 mini-batch AdamW stream over 20 disjoint events. Maximum values retain the worst observed corrected error; the weights-only row reports the minimum error to show that the control fails throughout the stream.

Quantity	Summary over events	Value
Corrected source measurement error	Maximum	2.49×10^{-7}
Weights-only source measurement error	Minimum	31.205
Pooled-record reconstruction error	Maximum	1.99×10^{-7}
Whole-network trajectory error E_t	Maximum	1.41×10^{-4}
Whole-network trajectory error E_{20}	Event 20	3.13×10^{-5}

Direct supervision and UT both move accuracy from 68.0% to 64.7%. Because E_t is normalized by cumulative reference displacement, the smaller value at event 20 does not imply that the absolute trajectory error decreases.

D.2 Target-Relative Observability: Boundary and Extensions

Section 6.2 tests the criterion with a deep selectively grouped record and with a fixed rank-deficient record evaluated against two targets.

Deep selective supervision on CIFAR-100. A fixed 800-example CIFAR-100 event is processed by MobileNetV3-small (Howard et al., 2019) and ResNet-18 (He et al., 2016). Their active source coordinates are the standardized classifier weights and bias, and each recorded transition is one vanilla-SGD head step of size 0.3 in evaluation mode. The fine supervision uses $\varepsilon = 0.1$ smoothing,

$$Q_{ic} = \begin{cases} 1 - \varepsilon, & c = y_i, \\ \varepsilon/99, & c \neq y_i. \end{cases}$$

After correcting the two fine-100 source records, we retain either those full records, their product with the official five-to-one class map $H_{\mathrm{off}} \in \mathbb{R}^{100\times 20}$, or their product with a seeded balanced random map H_{rnd} (PCG64 seed 20260925). Thus the grouped records are computed from source transitions rather than from grouped event labels.

EfficientNet-B0 (Tan & Le, 2019) is initialized from the fixed ImageNet checkpoint and cached pre-event head, run in evaluation mode, and updated over all trainable parameters by a vanilla-SGD step of size 0.01. Event gradients are accumulated deterministically in 16-example batches in FP32; e_F is evaluated on 1,000 test images, ten per fine class. Numerical rank uses the cutoff $10^{-9}\sigma_{\max}$.

The official projection has an 80-dimensional kernel per example within the fine-supervision contrast space, hence a 64,000-dimensional kernel over the event. This is a linear nullity, not a claim about an unconstrained ambient parameter count. Because every entry of the smoothed Q is positive, sufficiently small row-centered redistributions within an official superclass are also feasible probability perturbations. At the same time, the retained official-20 record has full sample rank $800/800$. Full sample rank within this projected representation therefore identifies SH_{off}, not the original fine-100 contrast S.

The two source-measurement errors are 2.076×10^{-8} and 2.198×10^{-8} for MobileNetV3-small and ResNet-18, respectively. The pooled retained-space rank is $800/800$. The target contains 4.14M trainable parameters in the fine-100 setting and 4.03M in the official-20 setting. Complete target-update and functional errors are reported in Table 7.

Table 7: **Deep target-relative CIFAR-100 intervention.** The official-20 record has full sample rank in its retained representation, while its class projection has a 64,000-dimensional kernel in the fine-supervision contrast space. e_θ is the relative target-update error and e_F is the relative functional-displacement error.

Record	Target	Fine-100 supervision	e_θ	e_F
Full-100	Fine-100	Identified	1.371×10^{-6}	1.055×10^{-6}
Official-20	Official-20	Ambiguous	1.044×10^{-6}	1.846×10^{-6}
Official-20	Fine-100	Ambiguous	2.768	0.8440
Random-20	Official-20	Ambiguous	2.748	0.9398

The obstruction in the two mismatched rows is structural, rather than inferred from their errors. For the fine-100 target, move a small amount of probability between two fine classes inside one official superclass. This feasible direction satisfies $dH_{\text{off}} = 0$, while the fine-100 classifier-bias gradient changes, so $T(d) \neq 0$. For the official-20 target, the seeded random partition contains fine classes from different official superclasses in a common random group; shifting mass between such a pair gives $dH_{\text{rnd}} = 0$ but $dH_{\text{off}} \neq 0$, again changing the target-bias gradient. These witnesses establish the source-null, target-visible directions required by Theorem 4.1; the large errors in Table 7 show their effect on the tested construction. The matched official-20 record, by contrast, determines the official-20 whole-network response even though the original fine supervision remains non-identifiable.

Sensorless: one record, two targets. A label-blind permutation of Sensorless Drive (Bator, 2013) defines training, development, update, and evaluation subsets. Training fixes standardization, feature maps, and logistic heads; one development batch fixes numerical choices before five disjoint 512-example events. Source 1 uses raw coordinates 1–24 plus 128 SiLU features; source 2 uses coordinates 25–48 plus 256 GELU features. The in-span target uses all 48 raw coordinates, and the outside-span target adds 128 independent tanh features. These maps remain fixed. With one shared intercept, pooled rank is 433. The 0.1 label smoothing makes Q positive, so small nonzero source-null contrasts preserve probability feasibility. This establishes non-unique supervision for the realized events. Table 8 reports paired results, with

$$d_{\text{span}} = \frac{\|(I - P_{X_S})X_*\|_F}{\|X_*\|_F}. \tag{61}$$

Table 8: Complete paired Sensorless results. Every row shares one source record between the two targets. Multiply entries by the displayed column factors.

	d_{span}		e_θ	
Event	In ($\times 10^{-15}$)	Out	In ($\times 10^{-14}$)	Out
1	2.96	0.03852	1.33	0.04881
2	4.63	0.04166	4.17	0.05302
3	2.49	0.04106	2.83	0.05357
4	2.82	0.04045	3.25	0.05612
5	2.57	0.04212	2.40	0.06620
Median	**2.82**	**0.04106**	**2.83**	**0.05357**

Thus the same incomplete record determines the in-span target but not the target that depends on directions left unresolved by the record.

Coverage sweep under a fixed target. Five label-blind partitions each contain 1,200 update and 8,800 disjoint evaluation images, with nested events of size 400, 800, and 1200. MobileNetV3-small and ResNet-18 are the sources, and EfficientNet-B0 is the target (Howard et al., 2019; He et al., 2016; Tan & Le, 2019). Each model updates its final feature block and head. The source-1/source-2/pooled ranks are $400/400/400$, $577/513/800$, and $577/513/1089$ at the three sizes. At $n = 400$, each source identifies the supervision contrast; at $n = 800$, only their pooled record does.

These first two cases therefore support the deeper target because the retained source information has full row rank. At $n = 1200$, the pooled rank is only 1089, while augmenting it with target features raises the rank to 1200 and exposes a target-visible missing direction. These are feature ranks, not parameter-coordinate counts: without full row rank, a deeper target requires the general target-relative condition rather than the head-span test alone.

Continuous MSE targets on Pet. With continuous mean-squared-error targets, ConvNeXt and Swin are sources and ViT is the target. Four Pet trimap attributes (Parkhi et al., 2012)—foreground area, two center coordinates, and root-mean-square radial spread—are scaled to $[-1, 1]$. Seven empty trimaps are excluded by a rule fixed before model loading. Three label-blind partitions provide nested sizes $n = 24, 48, 96$. Each model uses a frozen random projection, GELU, and a trainable four-output layer. Corrected output-layer coordinates form the source record, while the target updates its final backbone block and output layer. At $n = 96$, a direction chosen from the source kernel alone changes the actual target update in every partition; reported response norms range from 8.57×10^{-4} to 9.06×10^{-4}.

Table 9: Complete source-information intervention under classification and continuous supervision. Errors are ratios; ranks are source-1/source-2/pooled feature ranks, not parameter-Jacobian ranks.

n	Partition	Ranks	Source 1	Source 2	**Pooled**
A. CIFAR-100 / cross-entropy					
400	1	400/400/400	1.54×10^{-13}	2.55×10^{-13}	3.27×10^{-14}
400	2	400/400/400	1.45×10^{-13}	1.98×10^{-13}	3.35×10^{-14}
400	3	400/400/400	2.08×10^{-13}	2.29×10^{-13}	3.70×10^{-14}
400	4	400/400/400	2.17×10^{-13}	2.26×10^{-13}	3.61×10^{-14}
400	5	400/400/400	1.58×10^{-13}	2.34×10^{-13}	3.23×10^{-14}
400	**Median**	400/400/400	1.58×10^{-13}	2.29×10^{-13}	3.35×10^{-14}
800	1	577/513/800	0.01640	0.01790	2.89×10^{-13}
800	2	577/513/800	0.01710	0.01874	2.89×10^{-13}
800	3	577/513/800	0.01744	0.01903	2.94×10^{-13}
800	4	577/513/800	0.01688	0.01869	3.21×10^{-13}
800	5	577/513/800	0.01707	0.01827	3.06×10^{-13}
800	**Median**	577/513/800	0.01707	0.01869	2.94×10^{-13}
1200	1	577/513/1089	0.02640	0.02908	0.01037
1200	2	577/513/1089	0.02806	0.03094	0.01092
1200	3	577/513/1089	0.02801	0.03103	0.01126
1200	4	577/513/1089	0.02728	0.03049	0.01092
1200	5	577/513/1089	0.02720	0.02993	0.01079
1200	**Median**	577/513/1089	0.02728	0.03049	**0.01092**
B. Oxford-IIIT Pet / continuous MSE					
24	1	24/24/24	1.13×10^{-13}	1.13×10^{-13}	1.13×10^{-13}
24	2	24/24/24	9.40×10^{-14}	9.39×10^{-14}	9.38×10^{-14}
24	3	24/24/24	1.13×10^{-13}	1.13×10^{-13}	1.13×10^{-13}
24	**Median**	24/24/24	1.13×10^{-13}	1.13×10^{-13}	1.13×10^{-13}
48	1	33/33/48	0.0982	0.1183	1.14×10^{-13}
48	2	33/33/48	0.0794	0.0901	1.15×10^{-13}
48	3	33/33/48	0.1287	0.1345	1.26×10^{-13}
48	**Median**	33/33/48	0.0982	0.1183	1.15×10^{-13}
96	1	33/33/65	0.1378	0.1242	0.0890
96	2	33/33/65	0.1090	0.1037	0.0581
96	3	33/33/65	0.1246	0.1237	0.0737
96	**Median**	33/33/65	0.1246	0.1237	**0.0737**

Both losses in Table 9 exhibit the same transition. Pooling closes the target-relevant gap at CIFAR $n = 800$ and Pet–MSE $n = 48$; once a target-visible source-null direction remains, the pooled error is nonzero at CIFAR $n = 1200$ and Pet–MSE $n = 96$.

Non-affine local boundary. Two one-step pseudo-Huber systems test the error orders in the nonlinear analysis. Their source records and target quantities are smooth but non-affine functions of a three-dimensional hidden supervision vector. We evaluate two independently chosen settings. Each contains three cases: an exact nonlinear factorization; a source-invisible direction that affects the target first at second order, for which derivative coverage holds; and a source-invisible direction that affects the target at first order, for which derivative coverage fails.

For reproducibility, let $A = \left[\begin{smallmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{smallmatrix}\right]$ and $h_\delta(r) = r/\sqrt{1+(r/\delta)^2}$, applied coordinatewise. At $u_0 = 0$ the source map is $Y(u) = h_\delta(p_s - Au)$. The factorable target is $G_0(u) = w^\top h_\delta(p_t - Au)$; the two hidden targets are $G_2(u) = G_0(u) + \lambda h_\delta(-u_3^2)$ and $G_1(u) = G_0(u) + \lambda h_\delta(-u_3)$. Setting A uses $(\delta, \lambda, p_s, p_t, w) = (0.45, 0.60, (-0.15, 0.35), (0.20, -0.30), (0.70, 1.10))$; setting B uses $(1.10, 1.25, (0.50, 0.10), (-0.10, -0.45), (-0.90, 0.65))$. The factorable case is perturbed along $(1, 0, 0)$, and the two hidden cases along the normalized source-fiber direction $(1, 0, -1)$. The seven radii are $10^{-1}, 3\times10^{-2}, 10^{-2}, 3\times10^{-3}, 10^{-3}, 3\times10^{-4}, 10^{-4}$.

We fit the reported log–log slopes on the five radii at or below 10^{-2} that remain above floating-point resolution. In the factorable case, the exact nonlinear map has the error in the second column of Table 10; the adjacent slope concerns its fixed linearization, whose leading error is quadratic.

Table 10: Held-out pseudo-Huber boundary checks. The first result is the error of the exact nonlinear factor. The remaining columns report slopes fitted to the fixed-linearization error; their predicted local orders are 2, 2, and 1.

		Fixed-linearization slope		
Setting	Exact-factor error	Factorable	Covered	Missing direction
A	1.11×10^{-16}	**2.000590**	**2.000000**	**0.999977**
B	1.11×10^{-16}	**2.001046**	**2.000000**	**0.999996**

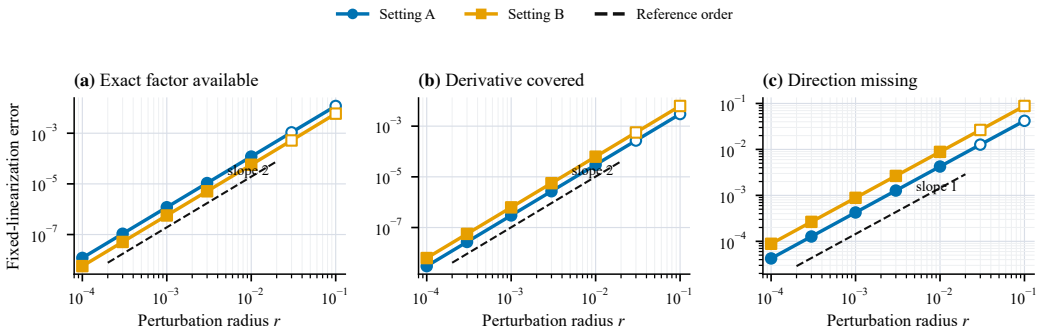


Figure 3: Fixed-linearization errors for the three non-affine controls. Each panel compares the two prespecified settings A and B over the seven recorded perturbation radii. Filled markers denote observations used in the reported slope fits; open markers denote the remaining observations. Dashed lines indicate reference slopes, not error bounds. Exact nonlinear-factor errors are reported separately in Table 10.

Both held-out settings follow the predicted orders in Figure 3. Equal source-record pairs also give identical target quantities in the factorable case, whereas the second-order and first-order counterexamples produce nonzero target differences.

Perturbation sensitivity. Additional CIFAR-100 perturbations complement the bounds in the stability analysis. Each entry in Table 11 is a median over five randomized image partitions of the same official CIFAR-100 test split. Checkpoint perturbations are normalized relative to the true endpoint displacement; independent Gaussian pre- and post-checkpoint directions are normalized in Frobenius norm and each receives $\sigma/\sqrt{2}$ times that displacement norm. Feature perturbations use a separate normalized Gaussian direction on the non-intercept columns, scaled relative to the frozen feature matrix while keeping the intercept fixed.

Table 11: CIFAR-100 perturbation results, medians over five randomized image partitions. Errors are relative target-update errors.

Family	Condition	Runs	Median error
Checkpoint	relative perturbation 10^{-3}	5	0.00492
Checkpoint	relative perturbation 10^{-2}	5	0.01328
Feature	relative perturbation 10^{-3}	5	0.00503
Feature	relative perturbation 10^{-2}	5	0.01475

The checkpoint and feature errors increase with the applied perturbation. These tests concern checkpoint and feature perturbations, not arbitrary distribution shift.

D.3 UT Construction and Reuse

Method names encode how a source is chosen. *Oracle single-source UT* selects the lower error after formal evaluation and is used only as a diagnostic. *Validation-selected single-source UT* selects a source on independent development data. *UT: source A/B* fixes one source without selection. The tables keep these three selection procedures distinct.

D.3.1 Target-Native Construction

Full-parameter construction. At $n = 800$, each architecture is the target in three partitions, with the other two providing corrected head records. Individual feature rank is 769 and pooled rank is 800 in all nine target–partition cases. Each source record retains 28,453 head coordinates. Table 12 also checks complete source updates, including coordinates outside the recorded heads. Across 18 routes, its median error is 5.42×10^{-5} and its maximum is 1.25×10^{-4}.

Table 12: Full-parameter Pet construction and consistency checks at $n = 800$. Panel A contains the nine target cases used in the main text; every pooled source-head record has row rank 800. Panel B checks the same identity along 18 independent source routes. Errors are ratios.

A. Target construction

Target	Partition 1	Partition 2	Partition 3	**Median**
ConvNeXt-Tiny	6.048×10^{-5}	6.582×10^{-5}	5.151×10^{-5}	6.048×10^{-5}
Swin-T	4.681×10^{-5}	2.272×10^{-5}	4.167×10^{-5}	4.167×10^{-5}
ViT-B/32	1.053×10^{-4}	1.195×10^{-4}	1.212×10^{-4}	1.195×10^{-4}

B. Additional source-route checks; six routes per architecture

Source	All coordinates	Retained (%)	Median e_θ	Max e_θ
ConvNeXt-Tiny	27,848,581	0.102	5.32×10^{-5}	1.25×10^{-4}
Swin-T	27,547,807	0.103	3.65×10^{-5}	4.41×10^{-5}
ViT-B/32	87,483,685	0.0325	1.18×10^{-4}	1.20×10^{-4}

The full-row-rank pooled head record determines the supervision contrast before the target computes its own full-parameter update.

Rank-deficient construction. The study rotates the same targets over three label-blind partitions at each $n \in \{400, 800, 1600\}$, with disjoint functional-evaluation inputs. Each model updates its final encoder block and head. The individual/pooled feature ranks are $400/400$, $769/800$, and $769/1537$ at the three sizes. Thus the pooled record is rank deficient at $n = 1600$. Active target counts are 4.79M, 7.12M, and 7.12M for ConvNeXt, Swin, and ViT.

Table 13 reports single-source UT, post-update outputs, probability changes, and no update against the common block-plus-head target. Head alignment is listed separately because it transports only head-weight and bias changes before being scored against the same reference; it is not THESEUS. The two fixed-source rows keep source identity unchanged across partitions. For the three target columns, source 1/source 2 are Swin/ViT, ConvNeXt/ViT, and ConvNeXt/Swin, respectively. Every single-source and pooled UT row uses the same source-normalized probability-constrained objective and ADMM stopping rule; only the retained source set changes.

The pooled sample-space nullspace has dimension 63 in each case. One true target-head direction and 15 fixed random target directions are extended to the feasible probability boundary in both signs. Across 144 directions, the median sampled target separation is 6.13×10^{-4}, while the largest source residual is 9.55×10^{-18}. The true direction is used only for evaluation. These finite probes demonstrate feasible ambiguity and its scale along the sampled directions; they do not upper-bound the diameter of the full feasible fiber.

Table 13: Rank-deficient Pet construction and feasible-set geometry at $n = 1600$. Panels A and B report median e_θ against the final-block-plus-head reference. Panel C reports target-head geometry over fixed directions and partitions; its errors therefore have a different target scope.

A. Matched block + head controls

Method	ConvNeXt	Swin	ViT
Probability-constrained UT	**0.00360**	**0.00242**	**0.00609**
UT: fixed source 1	1.6458	1.3726	3.5650
UT: fixed source 2	1.2012	0.9172	3.3145
Oracle single-source UT	1.2012	0.9172	3.1133
Uncovered direct rule	6.0188	4.6959	14.6988
Post-update output	0.8649	0.8768	0.9155
Probability change	0.9914	0.9952	0.9899
No update	1.0000	1.0000	1.0000

B. Restricted diagnostic

Method	Scope	ConvNeXt	Swin	ViT
Head alignment	Head only	0.9840	0.9908	0.9790

C. Sampled directional separation; normalized by soft-target dimension

Target	Median sampled sep.	Max sampled sep.	Direct error	Constrained error
ConvNeXt	1.65×10^{-4}	3.08×10^{-4}	2.906	**0.001432**
Swin	1.36×10^{-4}	2.73×10^{-4}	2.612	**0.001107**
ViT	1.72×10^{-4}	3.19×10^{-4}	3.619	**0.001493**

At $n = 1600$, the direct rule has large errors, whereas the probability-constrained rule has small errors on these events. Panel C also shows that supervision is not uniquely determined: distinct feasible signals have small effects along the probed target directions.

Compatible-ViT transport comparison. The experiment uses `open_clip_torch` 3.3.0 (Ilharco et al., 2021). The two sources are `ViT-B-16` with the `openai` and `laion400m_e32` checkpoints; the wider target is `ViT-B-16-plus-240` with `laion400m_e32`. All methods are scored on the final Transformer block and 37-way Pet head. UT reads two corrected source-head records, event inputs, and the required model states. THESEUS transports complete source updates using its feature-map procedure, fitted on 64 separate trainval images. A disjoint 800-example development update selects the LAION source and scale 0.1 by relative target-update error from $\{0, 0.1, \ldots, 5.0\}$; formal partitions are not used for selection.

Pooled feature ranks are 400, 800, and 1025 at $n = 400, 800, 1600$. UT uses the direct branch at the first two sizes and the probability-constrained branch at $n = 1600$, where the sample-space nullity is 575. Table 14 places fidelity and information access in one comparison. The single-source oracle is an evaluation-only reference.

Table 14: Compatible-ViT fidelity and information access. Panel A reports the three-partition median e_θ; Panel B records the unequal information and compatibility conditions under which the methods are evaluated.

A. Update fidelity

Method	$n = 400$	$n = 800$	$n = 1600$
UT (reported branch)	3.47×10^{-5}	9.31×10^{-5}	**0.00833**
Direct rule	3.47×10^{-5}	9.31×10^{-5}	0.1195
Oracle single-source UT	8.32×10^{-5}	0.2007	0.3398
THESEUS (validation-selected source)	0.9951	0.9951	0.995
THESEUS (two-source average)	0.9965	0.9965	0.9965
Post-update output teacher	2.7920	2.7975	2.8845
No update	1.0000	1.0000	1.0000

The reported UT row uses the direct branch at $n = 400$ and $n = 800$, and the probability-constrained branch at $n = 1600$.

B. Information access and applicability

Method	Source record	Extra target-side information	Status
UT	Corrected head records + states	Event inputs; no labels	Evaluated
THESEUS	Complete source updates	64 alignment images	Evaluated
TransFusion	Base/tuned checkpoints	Target-base rebasing	Shape-incompatible
GradFix	Task vector	Labeled target examples	Width-incompatible
Output teacher	Post-update outputs	Event inputs	Evaluated

One-step cross-domain construction. ConvNeXt-Large and Swin-V2-Base (Liu et al., 2022b) supply corrected head records for a ViT-L/16 target on 345-class DomainNet Real→Quickdraw. Five repetitions contain nested 690/1380/2070-example Quickdraw events; each event is one full-batch SGD step, with its labels hidden from UT and a disjoint test set.

ConvNeXt/Swin/pooled ranks are $690/690/690$, $1380/1024/1380$, and $1536/1024/2070$ at the three sizes in every repetition. All controls in Table 15 use the same target scope; the knowledge-distillation (KD) controls fit the indicated source endpoint.

Table 15: One-step DomainNet at $n = 2070$. The e_θ and held-out e_F columns are medians over five paired repetitions; NLL and ΔNLL are means.

	Update fidelity		*Secondary task metric*	
Method	e_θ	Held-out e_F	NLL	ΔNLL
No update	1.0000	1.0000	9.1275	0.0000
Output-change transfer	0.7078	0.7032	8.8001	0.3274
Head-checkpoint KD	0.2785	0.2683	8.3541	0.7734
Full-post KD	0.4335	0.4241	8.4797	0.6478
UT: Swin	0.0198	0.0042	8.1615	0.9660
UT: ConvNeXt	0.0137	0.0029	8.1614	0.9662
Pooled UT	1.90×10^{-4}	4.12×10^{-5}	8.1614	0.9662
Direct supervision	0	0	8.1614	0.9661

Pooling improves update fidelity over either source, while the strongest single source already attains the same NLL at the displayed precision.

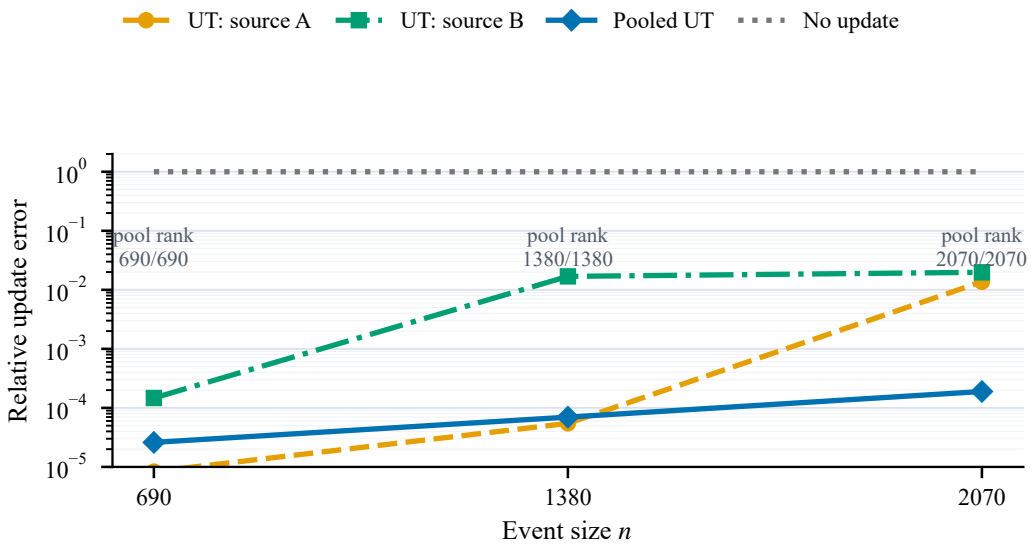


Figure 4: One-step DomainNet update fidelity across event sizes. The pooled ranks are 690/690, 1380/1380, and 2070/2070; pooling remains full row rank as individual records become rank deficient.

D.3.2 Persistent Records

We separate capture and replay into different processes on Pet and DomainNet. At capture, the corrected pooled measurement is compiled into the (U_r, Z) representation of Appendix C; replay receives the compiled record, event inputs, metadata, and current target state.

Pet uses an 800-example event with pooled rank 800, and DomainNet uses a 2,070-example event with pooled rank 2,070. The replay process does not load a source checkpoint, optimizer state, or RNG state. Labels are used only by the direct-supervision reference.

Table 16: **Source-independent compiled replay.** Raw and compiled UT are evaluated separately against the same direct-supervision target update. Times are wall-clock seconds for the corresponding capture or cold-replay process.

Dataset	n/rank	Compile time	Cold replay	Raw / reference	Compiled / reference	Compiled / raw
Pet	800/800	0.347	0.0248	5.415×10^{-5}	5.449×10^{-5}	1.051×10^{-5}
DomainNet	2070/2070	1.366	0.1202	1.906×10^{-4}	1.907×10^{-4}	4.758×10^{-6}

The FP32 compiled records occupy 2.38 MiB on Pet and 17.71 MiB on DomainNet.

The DomainNet compiled-versus-raw error is below the prespecified 10^{-5} auxiliary threshold. Pet gives 1.051×10^{-5}, exceeding that threshold by 5.1%. Only DomainNet meets this numerical-equivalence threshold.

D.3.3 Historical Replay

Selective CIFAR archive. Ten disjoint CIFAR-100 events are first processed by MobileNetV3-small and ResNet-18. Each fine-100 event is retained through the official-20 projection and compiled before the archive is sealed. Future targets are instantiated after capture.

The first future target is a whole-network EfficientNet-B0. Three fixed repetitions use the official-20 target, while two controls change either the target requirement or the retained grouping. A later whole-network ViT-B/32 is then evaluated from the same sealed CIFAR archive.

Table 17: **Selective CIFAR historical replay.** All targets are instantiated after capture and update all trainable parameters. Replay reads the archived event inputs and records, but no source training state or event labels.

Archived information	Future target	Events	Final $E_{10} \downarrow$	Result
Official-20, matched	Official-20 EfficientNet-B0	10	$(2.03 \pm 0.10) \times 10^{-5}$	Matches reference
Same archive	Fine-100 EfficientNet-B0	10	1.81 ± 0.50	Wrong target
Random-20	Official-20 EfficientNet-B0	10	3.626 ± 0.058	Wrong information
Same sealed archive	ViT-B/32	10	2.25×10^{-5}	Matches reference

The matched EfficientNet trajectory remains near 10^{-5} throughout all ten events. The fine-100 target and random-20 archive instead give order-one errors: the original fine supervision has not been recovered, and the choice of 20-way grouping matters.

Longer DomainNet archive. We next capture 20 disjoint, class-balanced Quickdraw events with 2,070 examples per event. ConvNeXt-Large and Swin-V2-Base evolve continuously with AdamW during capture. Their individual retained ranks are 1,536 and 1,025, while the pooled record has rank $2070/2070$ on every event. Future ViT-L/16 and EfficientNet-B0 targets are prepared only after capture.

Each historical event is replayed either once, giving 20 target updates, or three times, giving 60 target updates. Pooled UT is compared with the strongest single-source record and post-logit distillation; direct supervision defines the reference trajectory.

Table 18: **Twenty-event DomainNet historical replay.** E_t is the final trajectory error. Task columns show UT and direct-supervision values in the same order.

Future target	Replay	Best single E_t	Post-logit KD E_t	**Pooled UT** E_t	Acc. UT / ref.	NLL UT / ref.
ViT-L/16	20	0.03248	0.4872	2.039×10^{-5}	6.7826 / 6.7826	5.128058 / 5.128061
ViT-L/16	60	0.1100	0.8386	3.609×10^{-5}	15.0725 / 15.0725	4.035524 / 4.035532
EfficientNet-B0	20	0.05738	0.6255	1.852×10^{-5}	8.2319 / 8.2319	4.798766 / 4.798767
EfficientNet-B0	60	0.1266	0.8821	3.108×10^{-5}	34.4928 / 34.4928	2.817769 / 2.817769

Across all 20 events, the median source-measurement errors are 3.096×10^{-7} for ConvNeXt and 3.008×10^{-7} for Swin, and the pooled rank remains $2070/2070$. All four replays use the archived records without a source checkpoint, optimizer or RNG state, or event labels.

Output-change transfer is also evaluated and is worse than the strongest single-source record in all four settings.

D.3.4 Additional Application and Temporal Controls

Validation-selected use. Construction fidelity does not by itself decide whether an event should be used. We therefore return to the compatible-ViT models and target scope from Appendix D.3.1 and compare the original target head with a head fitted on 256 fixed training examples. The sources, update events, target backbone, and calibration data remain fixed; THESEUS maps are refitted for each target head. Held-out NLL selects the application scale before three formal partitions are evaluated at $n = 800$ and 1600.

In the preceding fidelity partitions, direct supervision raises median held-out NLL from 0.2514 to 0.2857 and 0.2867; UT yields 0.2857 and 0.2869.

Table 19 combines the two target-head starts. Raw error is computed before scaling; $\Delta\mathrm{NLL} = \mathrm{NLL}_{\text{no update}} - \mathrm{NLL}_{\text{applied}}$. A zero applied step does not imply an exact raw direction. The event-aligned THESEUS variant is reported only for the original head.

Table 19: Validation-selected Pet application from two target-head starts. Raw e_θ is reported as mean $\pm$ sample standard deviation over three formal partitions. The selected scale α comes from calibration data; ΔNLL and accuracy are partition means.

n	Method	Raw e_θ	α	ΔNLL	Acc. (%)
Original target head					
800	No update	1.00000 ± 0.00000	0	0	91.753
800	Output KD	0.94521 ± 0.00384	0	0	91.753
800	Validation-selected single-source UT	0.25048 ± 0.01354	0	0	91.753
800	**Pooled UT**	$(9.82 \pm 0.21) \times 10^{-5}$	**0**	0	**91.753**
800	THESEUS (64-image)	1.31700 ± 0.01205	2	-0.000631	91.791
800	THESEUS (event-aligned)	1.29573 ± 0.01437	0.75	-0.000196	91.868
800	Direct supervision	0	0	0	91.753
1600	No update	1.00000 ± 0.00000	0	0	91.753
1600	Output KD	0.94254 ± 0.00828	0	0	91.753
1600	Validation-selected single-source UT	0.46084 ± 0.00300	0	0	91.753
1600	**Pooled UT**	0.00882 ± 0.00038	**0**	0	**91.753**
1600	THESEUS (64-image)	1.29345 ± 0.00743	2	-0.000394	91.753
1600	THESEUS (event-aligned)	1.26917 ± 0.01070	0.75	-0.000135	91.791
1600	Direct supervision	0	0	0	91.753
Target head fitted on 256 examples; only methods evaluated from this start are shown					
800	No update	1.00000 ± 0.00000	0	0	77.062
800	**Pooled UT**	$(9.19 \pm 0.47) \times 10^{-5}$	**1.5**	0.187678	**78.519**
800	THESEUS (64-image)	1.42070 ± 0.04325	0	0	77.062
800	Direct supervision	0	1.5	0.187680	78.519
1600	No update	1.00000 ± 0.00000	0	0	77.062
1600	**Pooled UT**	0.00822 ± 0.00032	**1.5**	0.215023	**79.171**
1600	THESEUS (64-image)	1.42470 ± 0.01174	0	0	77.062
1600	Direct supervision	0	1.5	0.216112	79.210

For the original head, calibration selects a zero step for both UT and direct supervision; for the fitted head, it selects a step of 1.5 for both. THESEUS selects the opposite action in each case. This experiment tests whether the UT-constructed update inherits the action selected from the same independent calibration data and held-out NLL criterion as the supervised reference. These two starts do not establish a general test of event utility.

Reusing one recorded event. ConvNeXt-Large and Swin-V2-Base provide corrected source-head records for a ViT-L/16 target on five 2,070-example DomainNet Quickdraw events. Each target updates its final block and classifier for ten steps, recomputing the gradient at every state while keeping the recorded event fixed; labels are used only by the supervised reference.

Output-change transfers source logit changes; head-checkpoint KD distills the recorded head endpoint; full-post KD distills the complete post-event source. Independent calibration events select the settings listed in Table 24. The one-step study in Appendix D.3.1 uses target rate 0.03, whereas fixed-event reuse selects 0.3; their first-step values are therefore not directly comparable.

To test whether reuse is specific to the first source update, we also record adjacent source transitions at training depths 5 and 20. At depth d, UT reads only $(d-1) \to d$ and its matching pre-state; the target restarts from the same initial state and takes ten steps. Independent calibration events select rates, source weights, and KD temperatures.

Table 20: DomainNet fixed-event reuse across source training depth. Panel A reports complete endpoints for the depth-1 transition; Panel B tests later adjacent transitions at target step 10. Entries marked $\pm$ are means and sample standard deviations over five events. The pooled step-1 error is the recorded aggregate mean, and direct supervision is the exact zero-error reference. E_t is a ratio.

A. Source depth 1

Method	Accuracy (%)	NLL	Brier	E_t
Target step 1				
No update	2.123 ± 0.000	9.128 ± 0.000	1.1990 ± 0.0000	1.00000 ± 0.00000
Output change	0.506 ± 0.289	13.879 ± 1.854	1.8150 ± 0.1945	0.70683 ± 0.00551
Head-checkpoint KD	3.667 ± 0.047	7.547 ± 0.007	1.0371 ± 0.0007	0.09684 ± 0.00146
Full-post KD	3.674 ± 0.061	7.530 ± 0.007	1.0367 ± 0.0008	0.09522 ± 0.00141
UT: ConvNeXt	3.958 ± 0.036	7.362 ± 0.007	1.0341 ± 0.0006	0.01367 ± 0.00013
UT: Swin	3.962 ± 0.035	7.364 ± 0.007	1.0341 ± 0.0007	0.01975 ± 0.00017
Pooled UT	3.971 ± 0.030	7.361 ± 0.007	1.0341 ± 0.0007	1.90×10^{-4}
Direct supervision	3.971 ± 0.030	7.361 ± 0.007	1.0341 ± 0.0007	0
Target step 10				
No update	2.123 ± 0.000	9.128 ± 0.000	1.1990 ± 0.0000	1.00000 ± 0.00000
Output change	3.088 ± 0.350	7.758 ± 0.208	1.0819 ± 0.0329	0.89408 ± 0.04084
Head-checkpoint KD	4.170 ± 0.084	5.758 ± 0.003	0.9972 ± 0.0003	0.65156 ± 0.00647
Full-post KD	4.167 ± 0.072	5.760 ± 0.003	0.9971 ± 0.0003	0.65014 ± 0.00652
UT: ConvNeXt	7.677 ± 0.109	5.146 ± 0.007	0.9833 ± 0.0004	0.10399 ± 0.00094
UT: Swin	7.606 ± 0.083	5.153 ± 0.006	0.9836 ± 0.0004	0.14994 ± 0.00107
Pooled UT	7.755 ± 0.099	5.138 ± 0.007	0.9830 ± 0.0004	0.00144 ± 0.00001
Direct supervision	7.755 ± 0.100	5.138 ± 0.007	0.9830 ± 0.0004	0

B. Later adjacent source transitions; target step 10

Depth	Method	Accuracy (%)	NLL	E_{10}
5	Head-checkpoint KD	4.158 ± 0.086	5.765 ± 0.003	0.65269 ± 0.00655
5	Full-post KD	4.146 ± 0.075	5.765 ± 0.003	0.65181 ± 0.00657
5	**Pooled UT**	7.755 ± 0.098	5.138 ± 0.007	0.00198 ± 0.00003
5	Direct supervision	7.755 ± 0.100	5.138 ± 0.007	0
20	Head-checkpoint KD	4.210 ± 0.099	5.767 ± 0.003	0.64267 ± 0.00630
20	Full-post KD	4.207 ± 0.107	5.767 ± 0.003	0.64248 ± 0.00630
20	**Pooled UT**	7.758 ± 0.096	5.138 ± 0.007	0.00266 ± 0.00003
20	Direct supervision	7.755 ± 0.100	5.138 ± 0.007	0

At target step 10, pooled UT improves accuracy over head-checkpoint KD by 3.55–3.60 percentage points across source depths 1, 5, and 20. ConvNeXt alone already captures most of the task gain; adding Swin primarily reduces trajectory error.

Text-domain fixed-event reuse. BERT-base (Devlin et al., 2019) and RoBERTa-base (Liu et al., 2019) supply events of the same 800 raw texts, processed by their own fixed tokenizers, for a DistilBERT target (Sanh et al., 2019). The target updates its final encoder block and four-class head. Five formal events are fixed before evaluation on the separate official test set. UT and supervision use rate 0.003; independently calibrated head-checkpoint KD uses the same rate, temperature 16, and equal source weights. Output-change and full-post KD are not evaluated in this text experiment.

At cutoff $10^{-10}\sigma_{\max}$, both individual feature ranks are 769. The pooled rank is 800 in events 0–3 and 799 in event 4, which uses the constrained solver. Table 21 reports aggregate results and branches; per-event errors appear in Table 26.

Table 21: AG News fixed-event reuse. Panel A reports means ± sample standard deviations over five events. Panel B identifies the pooled branch in each event; both individual source ranks are 769 and the direct-supervision error is zero.

A. Aggregate ten-step results

Method	Accuracy (%)	NLL	E_{10}
No update	84.750 ± 0.000	0.973 ± 0.000	1.00000 ± 0.00000
Head-checkpoint KD	80.692 ± 0.577	0.534 ± 0.009	4.03180 ± 0.52464
UT: BERT	85.437 ± 0.102	0.756 ± 0.027	0.07585 ± 0.01071
UT: RoBERTa	85.432 ± 0.094	0.759 ± 0.028	0.06895 ± 0.00876
Pooled UT	85.434 ± 0.124	0.762 ± 0.028	$(3.783 \pm 0.305) \times 10^{-6}$
Direct supervision	85.434 ± 0.124	0.762 ± 0.028	0

B. Pooled branch by event

Event	Pooled rank	Branch	**Pooled UT** E_{10}
0	800	Direct	3.483×10^{-6}
1	800	Direct	4.243×10^{-6}
2	800	Direct	3.652×10^{-6}
3	800	Direct	3.605×10^{-6}
4	799	Constrained	3.933×10^{-6}

The lower NLL of head-checkpoint KD does not imply higher update fidelity: its ten-step trajectory error exceeds four and its accuracy falls below the unchanged target, whereas pooled UT follows the supervised trajectory to 3.78×10^{-6}. NLL is therefore a task metric, not an update-fidelity measure.

Rank-deficient fixed-event reuse. To test persistent undercoverage, we reuse the mixed-architecture $n = 1600$ records from Appendix D.3. For each target and partition, direct and probability-constrained representatives are inferred once, then reused for ten updates of the final block and classifier. Table 22 reports target-wise median displacement errors; all nine cases appear in Table 27. Reference and constrained task metrics coincide at the displayed precision and are listed once in Table 29. Because E_t is normalized by the reference cumulative displacement, a smaller E_t need not mean that absolute error has decreased.

Table 22: Rank-deficient Pet fixed-event reuse. Entries are target-wise medians over three partitions; reference errors are zero.

Target	Direct E_1	Direct E_{10}	**Constr.** E_1	**Constr.** E_{10}
ConvNeXt	6.01881	7.28259	**0.00360**	**0.00311**
Swin	4.69590	11.89433	**0.00243**	**0.00238**
ViT	14.69874	27.06637	**0.00609**	**0.00559**

Figures 5 and 6 show trajectory errors and task metrics, respectively.

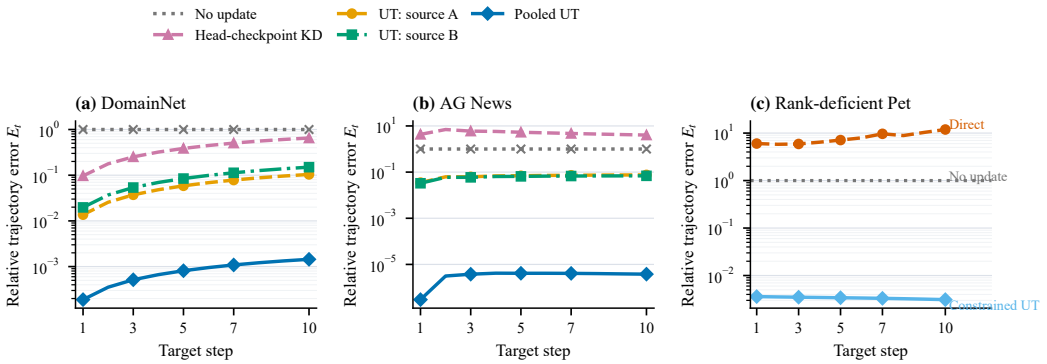


Figure 5: Fixed-event reuse. DomainNet and AG News show means over five events; Pet shows pointwise medians over nine target–partition cases. Panels (a)–(b) share the upper legend, while panel (c) labels its two rules directly. Vertical axes are log-scaled ratios and the zero-error reference is omitted.

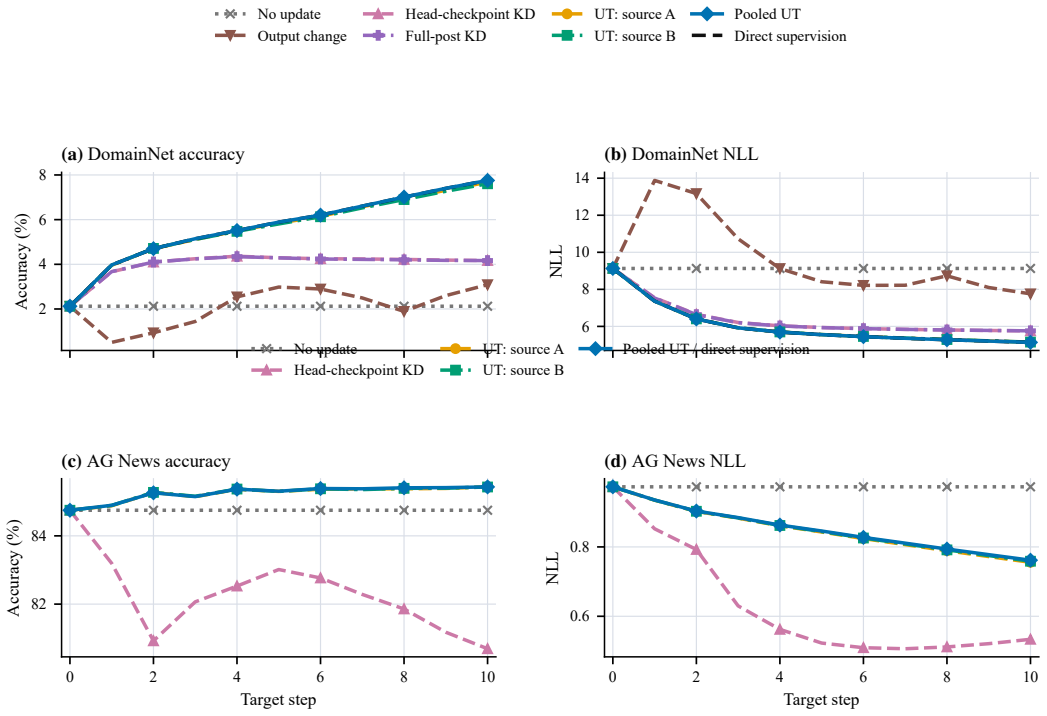


Figure 6: Target accuracy and NLL during fixed-event reuse. Curves are means over five fixed events, with a separate legend for each dataset. Pooled UT and direct supervision coincide at the plotted precision where combined. Task curves are secondary diagnostics; trajectory fidelity is reported separately in Figure 5.

Following fresh events. Using the same model families and final-block-plus-head target scope, five indexed 2070-example events arrive in order. Sources and their AdamW states advance once per event and provide fresh corrected records; the UT and direct-supervision target paths each continue from their own preceding state. Event supervision generates the source transitions and the reference path but is not given to UT. Every pooled record has rank 2070.

Table 23: One five-event DomainNet trajectory. All pooled ranks are 2070; E_t is measured from the common initial target. UT and direct-supervision NLL and accuracy coincide at the reported precision.

Event	E_t	NLL (UT / reference)	Accuracy (UT / reference, %)
1	8.81×10^{-6}	6.957030 / 6.957030	4.130 / 4.130
2	5.49×10^{-6}	6.546478 / 6.546478	4.275 / 4.275
3	4.35×10^{-6}	6.248759 / 6.248759	4.022 / 4.022
4	4.15×10^{-6}	6.099495 / 6.099495	3.391 / 3.391
5	3.95×10^{-6}	5.906452 / 5.906452	3.630 / 3.630

The unchanged target has NLL 7.396379 and accuracy 3.681%. The five constructed updates track the supervised trajectory closely, while the final accuracy remains below that of the unchanged model.

Table 24: Selected fixed-event settings. Source weights are ordered ConvNeXt/Swin for DomainNet and BERT/RoBERTa for AG News. These are the selected settings, not the complete calibration grids.

Method	Target LR	Other settings
DomainNet		
Output change	0.1	Scale 64; weights (0.5, 0.5)
Head-checkpoint KD	0.3	Swin only; temperature 16
Full-post KD	0.3	Swin only; temperature 16
UT: single / pooled	0.3	No additional scaling
Direct supervision	0.3	Reference trajectory
AG News		
Head-checkpoint KD	0.003	Equal source weights; temperature 16
UT: single / pooled	0.003	Fixed representative
Direct supervision	0.003	Reference trajectory

Table 25: Fitted-head Pet application by partition. Baseline denotes no update and zero-scaled THESEUS (64-image); their raw errors differ.

		Accuracy (%)			NLL		
Part.	n	Baseline	UT	Direct sup.	Baseline	UT	Direct sup.
1	800	76.410	**77.330**	77.330	1.027656	**0.843498**	0.843495
1	1600	76.410	**78.136**	78.136	1.027656	**0.833085**	0.832159
2	800	76.410	**78.136**	78.136	1.053984	**0.879976**	0.879975
2	1600	76.410	**78.251**	78.251	1.053984	**0.851501**	0.850262
3	800	78.366	**80.092**	80.092	0.957478	**0.752608**	0.752606
3	1600	78.366	**81.128**	81.243	0.957478	**0.709462**	0.708360

Selected settings and per-event results.

Table 26: AG News per-event fidelity. Every event has 800 rows; the pooled rank and branch appear in Table 21. All entries are E_{10} ratios.

Event	**Pooled UT**	UT: BERT	UT: RoBERTa	Head KD
0	3.483×10^{-6}	0.07088	0.06304	3.22372
1	4.243×10^{-6}	0.08032	0.06132	4.14821
2	3.652×10^{-6}	0.09254	0.06673	4.35133
3	3.605×10^{-6}	0.06621	0.07033	3.85517
4	3.933×10^{-6}	0.06929	0.08335	4.58056

Table 27: Complete Pet rank-deficient reuse errors. Reference errors are zero for every target–partition case; all displayed values are ratios.

Target	Part.	Direct E_1	Direct E_{10}	**Constr.** E_1	**Constr.** E_{10}
ConvNeXt	0	5.89083	8.09848	**0.00360**	**0.00311**
ConvNeXt	1	6.01881	6.88058	**0.00348**	**0.00298**
ConvNeXt	2	7.44766	7.28259	**0.00383**	**0.00335**
Swin	0	4.69590	12.51112	**0.00243**	**0.00238**
Swin	1	5.33694	11.89433	**0.00245**	**0.00242**
Swin	2	4.01869	10.33039	**0.00237**	**0.00232**
ViT	0	13.49984	27.06637	**0.00607**	**0.00559**
ViT	1	15.19141	22.16462	**0.00609**	**0.00543**
ViT	2	14.69874	32.54570	**0.00614**	**0.00565**

Table 28: AG News per-event task metrics. Pooled UT and direct supervision coincide only at the displayed task-metric precision.

Event	Method	Acc.$_1$ (%)	Acc.$_{10}$ (%)	NLL$_1$	NLL$_{10}$
0	Direct sup. / pooled UT	84.934	85.368	0.928	0.720
0	UT: BERT	84.921	85.408	0.927	0.714
0	UT: RoBERTa	84.934	85.342	0.928	0.719
0	Head-checkpoint KD	82.579	81.289	0.890	0.524
1	Direct sup. / pooled UT	84.237	85.553	0.966	0.770
1	UT: BERT	84.211	85.539	0.965	0.762
1	UT: RoBERTa	84.237	85.539	0.966	0.771
1	Head-checkpoint KD	83.487	80.539	0.832	0.535
2	Direct sup. / pooled UT	85.250	85.250	0.927	0.786
2	UT: BERT	85.224	85.276	0.927	0.776
2	UT: RoBERTa	85.237	85.329	0.927	0.784
2	Head-checkpoint KD	82.987	81.145	0.882	0.529
3	Direct sup. / pooled UT	85.000	85.487	0.927	0.748
3	UT: BERT	85.000	85.461	0.926	0.746
3	UT: RoBERTa	85.000	85.447	0.926	0.740
3	Head-checkpoint KD	83.105	80.658	0.852	0.533
4	Direct sup. / pooled UT	85.053	85.513	0.929	0.784
4	UT: BERT	85.053	85.500	0.928	0.780
4	UT: RoBERTa	85.053	85.500	0.928	0.780
4	Head-checkpoint KD	83.816	79.829	0.806	0.548
All	No update	84.750	84.750	0.973	0.973

Table 29: Pet per-partition task metrics for rank-deficient reuse. Reference and constrained entries share the displayed task metrics, not their trajectory errors.

Part.	Method	$\mathrm{Acc.}_1$ (%)	$\mathrm{Acc.}_{10}$ (%)	NLL_1	NLL_{10}
ConvNeXt					
0	Direct sup. / constrained UT	93.717	92.992	0.218	0.202
0	Direct	93.668	41.566	0.219	2.722
1	Direct sup. / constrained UT	94.345	93.282	0.198	0.194
1	Direct	94.442	54.519	0.196	2.545
2	Direct sup. / constrained UT	93.910	93.088	0.219	0.203
2	Direct	93.668	89.512	0.214	0.378
Swin					
0	Direct sup. / constrained UT	93.233	92.943	0.212	0.204
0	Direct	93.330	29.241	0.214	4.428
1	Direct sup. / constrained UT	93.813	93.862	0.190	0.185
1	Direct	93.910	84.244	0.194	0.915
2	Direct sup. / constrained UT	93.862	93.572	0.187	0.183
2	Direct	93.862	89.947	0.189	0.311
ViT					
0	Direct sup. / constrained UT	91.155	90.913	0.292	0.283
0	Direct	91.155	16.191	0.301	6.261
1	Direct sup. / constrained UT	91.542	90.768	0.285	0.276
1	Direct	91.445	69.454	0.285	2.303
2	Direct sup. / constrained UT	91.687	91.493	0.289	0.273
2	Direct	91.493	3.190	0.288	40.643

D.4 Experimental Scope, Assumptions, and Reporting Conventions

Correspondence and state assumptions. The present operators assume the same ordered examples and either shared supervision coordinates or a known label map represented in the measurement operator. Approximate example matching, unknown or unaligned label correspondences, and training-mode BatchNorm would require the measurement model to include the corresponding alignment or state. Optimizer and realized random state are retained whenever they affect the recorded transition.

Coverage and target scope. The deep-model experiments retain measurements from the source classification heads, while the target parameter set varies across experiments. In the $n = 800$ Pet experiment, the pooled source feature matrix has full row rank, so it determines S before the target gradient is computed and all trainable target parameters participate. The deep selective CIFAR intervention and the CIFAR mini-batch AdamW stream also use whole-network EfficientNet-B0 targets; the sealed CIFAR archive is replayed by whole-network EfficientNet-B0 and ViT-B/32 targets. The rank-deficient Pet construction and the one-step, fixed-event, fresh-event, and historical DomainNet experiments use the stated final-block-plus-classifier scope. With a rank-deficient retained record, coverage must be checked for the chosen target parameter set. General certification for rank-deficient records and full-parameter deep targets remains open. Controlled full-batch experiments are used where the rank transition must be manipulated explicitly; this is an experimental design choice rather than a restriction of the one-event formulation. The probability-constrained experiments use one-hot or mildly smoothed targets; higher-entropy supervision may leave a wider feasible set.

Implementation conventions. Numerical choices were fixed before formal evaluation; sensitivity checks are stated with the corresponding experiment. Table 30 collects the conventions needed to interpret the reported ranks and updates.

Reporting conventions. Medians summarize the controlled CIFAR/Sensorless and MSE/Pet partition studies and the one-step DomainNet error metrics; the DomainNet NLL columns use means. DomainNet and AG News fixed-event summaries use means and sample standard deviations over five events, not independently pretrained models. The Pet rank-deficient table uses medians over

Table 30: Implementation and numerical conventions used by the controlled and deep-model experiments.

Setting	Preprocessing / precision	Rank cutoff	Key convention
Mixed Pet	Official pretrained weights; ImageNet resize/crop 256/224; FP32	—	37-class heads; standardization fitted on trainval
CIFAR-100 rank sweep	Heads fitted in evaluation mode; dimensions 576/512/1280	$10^{-9}\sigma_{\max}$	Shared intercept; 0.9/0.1 smoothing; full-batch accumulation
Deep selective CIFAR	Source heads and whole-network target in evaluation mode; FP32	$10^{-9}\sigma_{\max}$	0.1 smoothing; official map or PCG64-seeded balanced map
CIFAR mini-batch stream	Twenty disjoint 128-example events; FP32 AdamW	—	Whole-network target; continuously advancing model and optimizer states
Pet–MSE	FP32 pretrained prefix; FP64 trainable tail	$10^{-8}\sigma_{\max}$; stable at 10^{-10}	Continuous four-output target
Pet, rank deficient	Records scaled by $1/\|X_m\|_2$	$10^{-10}\sigma_{\max}$	Undercoverage persists at cutoff 10^{-8}
Constrained solver	FP64 alternating-direction iterations	—	$\gamma = 0.1$; residual tolerance 10^{-7}; at most 20,000 iterations
One-step DomainNet construction	Full-batch vanilla SGD	$10^{-6}\sigma_{\max}$; stable at 10^{-8}	ConvNeXt/Swin/ViT rates 0.03/0.003/0.03; no momentum or weight decay

three partitions per target; its trajectory figure shows pointwise medians over nine target–partition cases.