SCOUT: MEASURING WHEN CROSS-MODEL TRANSFER HELPS FEDERATED LIGHTWEIGHT LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-model transfer can improve lightweight learners, yet the utility of an individual intervention can depend on receiver state and update scale. We study this dependence in federated learning through SCOUT, a receiver-centric matched-intervention protocol comparing a candidate-transfer branch with a matched NoTransfer control from the same pre-intervention state. Across 75 audited states, FL-normalized transfer has positive mean utility but negative observed utility in 21 states. Changing only the update scale reverses utility sign in 29/75 states (38.7%) and makes Fixed-step preferable in 24/75 states (32.0%). A retrospective statewise operator oracle adds 4.88×10^{-5} mean bidirectional $\Delta\mathrm{MRR}$ over the best fixed operator (22.2%). On 50 canonical states, eight pre-intervention features with Bayesian Ridge and regularized logistic models do not reliably predict FL-normalized utility or harm under seed-grouped held-out evaluation; relative operator preference is not tested. Five matched 20-round repeated-transfer policy pairs have positive final held-out effects, while one-step audits separately contain negative observed utilities. These results distinguish intervention utility, retrospective choice value, single-operator predictability, and trajectory-level outcome.

1 INTRODUCTION

Foundation models provide reusable multimodal representations for resource-constrained learners (Radford et al., 2021). In federated learning (FL) (McMahan et al., 2017), however, cross-model transfer is applied to a moving target: the lightweight receiver evolves across communication rounds under heterogeneous client participation and private data. Knowledge exchange and pretrained-model assistance are therefore increasingly used when direct parameter averaging is insufficient (He et al., 2020; Zhu et al., 2021; Chen et al., 2024). Yet source-model quality alone does not determine what the current receiver gains from a transfer; teacher–student capacity and representation gaps can make a stronger teacher less useful to a particular learner (Cho & Hariharan, 2019; Mirzadeh et al., 2020).

Most existing foundation–lightweight methods ask *how* knowledge should be adapted, aligned, filtered, or transferred. Negative transfer is usually treated as a method-level problem to be mitigated by a better transfer mechanism (Wang et al., 2019; Zhang et al., 2023), while transferability estimators assess model–task or representation–task suitability (Nguyen et al., 2020; You et al., 2021; Huang et al., 2022). We instead take the *individual transfer intervention* as the unit of analysis. A mechanism with positive mean utility may still harm a particular receiver state, and two controlled update scales can produce different outcomes at that same state. This distinction matters in FL because receiver states are generated by changing client samples, heterogeneous local updates, aggregation, and optimization history rather than by a single stationary training process.

For receiver state S_r and candidate operator $\mathcal{O}$, SCOUT forks the same pre-intervention executable state into Transfer and matched NoTransfer branches under an identical downstream contract. Their performance difference isolates the utility delivered by the intervention from ordinary FL progress, separating *source quality* from *receiver utility*. Our central question is: *what utility does a candidate cross-model transfer operator deliver at the current receiver state, and how well can evidence available before execution predict it?* Because the same receiver state can respond differently to different operators, we also ask how much retrospective value exists in statewise operator choice, while keeping that oracle analysis separate from prospective prediction.

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.

Across 75 matched states, FL-normalized transfer has positive mean utility but negative observed utility in 21 states. Changing only the update scale reverses utility sign in 29/75 cases (38.7%) and makes the lower-mean Fixed-step operator preferable in 24/75 states (32.0%). A retrospective operator oracle adds 4.88×10^{-5} mean $\Delta\mathrm{MRR}$ over the best fixed operator (22.2%). A separate canonical-50 analysis finds limited seed-held-out prediction of FL-normalized utility and harm, while all five audited 20-round repeated-transfer pairs have positive final held-out effects. These observations motivate three contributions:

- **A receiver-centric intervention audit.** Matched Transfer/NoTransfer branches isolate a candidate intervention from ordinary federated progress and define utility from the receiving learner's perspective.
- **Replicated state–operator heterogeneity and retrospective choice value.** Canonical and external audits reproduce beneficial-majority/harmful-tail behavior, reveal strict sign reversals across controlled operators, and quantify statewise oracle headroom without treating it as a deployable selector.
- **Bounded prospective predictability and horizon separation.** Seed-grouped held-out evaluation shows limited prediction of FL-normalized utility/harm for the evaluated features and models, while one-step intervention effects are analyzed separately from 20-round policy outcomes.

2 Related Work

Heterogeneous and foundation-assisted federated knowledge transfer. Heterogeneous FL often exchanges knowledge rather than directly averaging incompatible parameters, including ensemble distillation, group knowledge transfer, data-free distillation, and prototype sharing (Lin et al., 2020; He et al., 2020; Zhu et al., 2021; Tan et al., 2022). Recent work also studies pretrained/server-assisted, embedding-based, cluster-aware, and cross-modal knowledge transfer (Chen et al., 2024; Zhang et al., 2024; Mora et al., 2024; Le et al., 2025; Wang et al., 2025; Xia et al., 2026). Relational, similarity-preserving, and contrastive distillation can reduce dependence on coordinate-aligned teacher/student features (Park et al., 2019; Tung & Mori, 2019; Tian et al., 2020), while teacher–student gaps can make stronger teachers less useful (Cho & Hariharan, 2019; Mirzadeh et al., 2020). These works primarily optimize a final model. SCOUT instead measures the receiver-side effect of a particular intervention against an exactly matched NoTransfer control.

Transferability, negative transfer, and selective distillation. LEEP, LogME, and TransRate estimate pretrained-representation suitability (Nguyen et al., 2020; You et al., 2021; Huang et al., 2022); negative-transfer studies show that source knowledge can reduce target performance (Wang et al., 2019; Zhang et al., 2023), and ranking-specific evidence cautions that classification-oriented estimators need not align with ranking objectives (Bai et al., 2024). Selective distillation chooses samples, targets, predictions, or teachers (Wang et al., 2021; Lee et al., 2022; Liu et al., 2023; Ding et al., 2024). SCOUT is complementary rather than a replacement for these methods: it audits complete interventions at evolving receiver states, quantifies retrospective state–operator structure, and separately tests held-out prediction for a fixed operator. We therefore do not treat oracle headroom as evidence that a deployable selector has been learned.

3 Foundation-to-Lightweight Transferability

3.1 Federated retrieval setting

Consider K clients with private paired image–text datasets $\{\mathcal{D}_i\}_{i=1}^{K}$. At communication round r, the server maintains a lightweight multimodal retriever S_r and a frozen foundation model F. A subset of clients performs local optimization and communicates only protocol-permitted summaries and model updates; raw client samples remain local. Client sampling, heterogeneous local updates, aggregation, and optimization history create the evolving receiver states studied here.

Operationally, S_r is the executable state at the pre-intervention fork, captured after FL aggregation/evidence construction but before relation correction. The checkpoint stores the student trainable

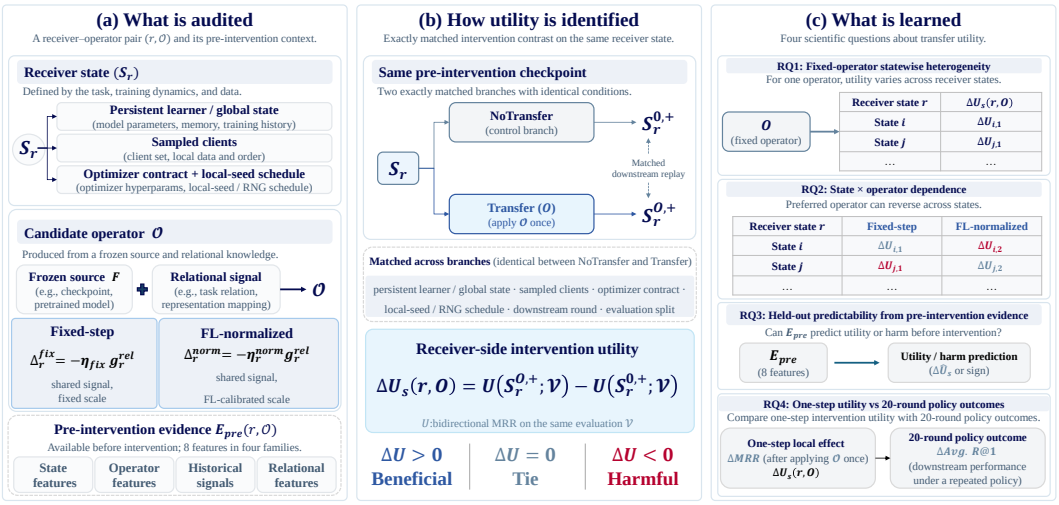


Figure 1: **Receiver-centric matched-intervention analysis in SCOUT. (a)** A receiver state, candidate operator, and evidence available before execution define the audited pair. **(b)** Transfer and NoTransfer branch from the same executable state and share the downstream contract; their post-replay difference defines receiver-side intervention utility. **(c)** Repeated contrasts support state–operator, prospective-prediction, and one-step versus policy-level analyses.

state, server global state, persistent FL/PFL state, RNG state, seed, round, and sampled-client IDs, all restored before replay. Figure 1 summarizes the receiver–operator pair, matched contrast, and scientific questions; all empirical conclusions concern this federated retrieval setting.

3.2 Controlled transfer operators

Let $\mathcal{O} \in \Omega$ denote a candidate foundation-to-lightweight transfer operator. On a server-public calibration set, the lightweight receiver and the raw and safety-selected foundation states produce normalized image/text embeddings and cross-modal similarity matrices A_r^{S}, $A^{\mathrm{F},raw}$, and $A^{\mathrm{F},safe}$; the foundation state is selected on protocol-defined safe-validation data without raw client samples. For each I2T/T2I query, we take the receiver's $k = 8$ hardest negatives and retain the negative with the largest *positive* foundation-induced increase in positive-versus-negative margin. Let $\mathcal{P}_r$ be the resulting bidirectional preference set. For nonempty $\mathcal{P}_r$ and $p \in \mathcal{P}_r$, the receiver margin is $m_p^{\mathrm{S}} = A_{q,q}^{\mathrm{S}} - A_{q,n}^{\mathrm{S}}$ for I2T and $A_{q,q}^{\mathrm{S}} - A_{n,q}^{\mathrm{S}}$ for T2I. The relation objective and candidate direction are

$$\mathcal{L}_{\mathrm{rel}}(S_r) = \frac{1}{|\mathcal{P}_r|} \sum_{p \in \mathcal{P}_r} \mathrm{softplus}\big(-m_p^{\mathrm{S}}/\tau\big), \qquad g_r^{\mathrm{rel}} = \nabla_{\theta_r} \mathcal{L}_{\mathrm{rel}}, \tag{1}$$

with $\tau = 0.07$; if $\mathcal{P}_r$ is empty, we set $g_r^{\mathrm{rel}} = 0$. Let g_r^{ret} be the gradient of the ordinary symmetric image–text retrieval loss on the same calibration set. The shared gate is active iff $\|g_r^{\mathrm{rel}}\|_2 > 0$ and $\cos(g_r^{\mathrm{ret}}, g_r^{\mathrm{rel}}) > 0$. Both operators therefore share preference construction, relation loss, compatibility gate, trainable receiver parameters, and intervention locus; only update scale differs. Let Δ_r^{FL} be the ordinary federated update already available at the fork. The two operators are

$$\Delta_r^{\mathrm{fix}} = -\eta_{\mathrm{fix}}\, g_r^{\mathrm{rel}}, \qquad \eta_{\mathrm{fix}} = 2 \times 10^{-5}, \tag{2}$$

$$\Delta_r^{\mathrm{norm}} = -\eta_r^{\mathrm{norm}}\, g_r^{\mathrm{rel}}, \qquad \eta_r^{\mathrm{norm}} = \frac{\rho \|\Delta_r^{\mathrm{FL}}\|_2}{\|g_r^{\mathrm{rel}}\|_2 + \epsilon}, \qquad \rho = 0.05. \tag{3}$$

Here ϵ is a numerical stabilizer. The frozen configuration uses up to 128 public calibration pairs, and $\rho = 0.05$ because the target ratio and cap both equal 0.05. Fixed-step uses a state-independent step, whereas FL-normalized targets correction norm $0.05\|\Delta_r^{\mathrm{FL}}\|_2$ and fails closed for nonpositive FL-update norm. No post-intervention utility sets either scale. When the gate is active, the corrections are collinear, so utility-sign changes reflect receiver response to scale rather than different transfer directions.

3.3 PAIRED INTERVENTION UTILITY

Comparing consecutive FL rounds would confound transfer with ordinary optimization. We therefore fork at the checkpoint captured after aggregation/evidence construction and before relation correction, restrict evidence to quantities already available there, and replay the same downstream contract:

$$S_r^{0,+} = \mathsf{NoTransfer}(S_r; \xi_r), \tag{4}$$

$$S_r^{\mathcal{O},+} = \mathsf{Transfer}(S_r, F, \mathcal{O}; \xi_r), \tag{5}$$

where ξ_r fixes the restored student/global/PFL state, sampled-client IDs, RNG and local-seed schedule, optimizer contract, evaluation split, and downstream round across branches. Receiver-side transfer utility is

$$\Delta U_{\mathrm{S}}(r, \mathcal{O}) = U(S_r^{\mathcal{O},+}; \mathcal{V}) - U(S_r^{0,+}; \mathcal{V}), \tag{6}$$

where $\mathcal{V}$ is the matched intervention-evaluation set. The primary realized utility is dense bidirectional MRR computed from post-replay ranks. Positive, zero, and negative ΔU_{S} are termed beneficial, tie, and harmful as descriptive signs on this fixed evaluation set; they are not per-state significance tests and do not by themselves establish practical importance. This receiver-centric target is deliberately distinct from source-model quality: a strong foundation model may supply useful knowledge, but the measured quantity is the utility actually delivered to the current receiver.

3.4 PROSPECTIVE PREDICTION TARGETS

Let $E_{\mathrm{pre}}(r, \mathcal{O})$ denote information available before execution. Conceptually, state- and operator-conditional transferability is

$$T^{\star}(E, \mathcal{O}) = \mathbb{E}[\Delta U_{\mathrm{S}}(r, \mathcal{O}) \mid E_{\mathrm{pre}}(r, \mathcal{O}) = E], \tag{7}$$

or, for sign identification,

$$p^{\star}(E, \mathcal{O}) = \Pr(\Delta U_{\mathrm{S}}(r, \mathcal{O}) > 0 \mid E_{\mathrm{pre}}(r, \mathcal{O}) = E). \tag{8}$$

These are conceptual prediction targets over receiver states, not quantities identified from the finite audit by assumption. The reported prediction experiment concerns $\mathcal{O}_{\mathrm{norm}}$ only. Predicting relative operator preference would require a different target, such as $\Delta U_{\mathrm{S}}(s, \mathcal{O}_{\mathrm{fix}}) - \Delta U_{\mathrm{S}}(s, \mathcal{O}_{\mathrm{norm}})$; that target is not evaluated here.

4 SCOUT: MEASURING AND ANALYZING TRANSFER UTILITY

SCOUT operationalizes the formulation above by cloning the full pre-intervention receiver state before a candidate cross-model action and replaying the downstream execution contract identically across Transfer and NoTransfer. The official final test set is inaccessible to intervention construction and matched-intervention evaluation. The resulting unit of analysis is a transfer *intervention*, not merely a transfer *method*: a rule may have positive average utility while producing beneficial, harmful, or neutral outcomes at particular receiver states.

4.1 DENSE RECEIVER UTILITY

A coarse endpoint metric can hide small changes in retrieval ordering. We retain per-query ranks for both image-to-text and text-to-image retrieval and define

$$\Delta\mathrm{MRR}_{r,\mathcal{O}} = \frac{1}{2}\left[\mathrm{MRR}_{\mathrm{I2T}}^{\mathcal{O}} - \mathrm{MRR}_{\mathrm{I2T}}^{\mathsf{NoTransfer}} + \mathrm{MRR}_{\mathrm{T2I}}^{\mathcal{O}} - \mathrm{MRR}_{\mathrm{T2I}}^{\mathsf{NoTransfer}}\right]. \tag{9}$$

We use Eq. (9) as the primary realized intervention utility. In the external audit, the comparison scalar is the bidirectional Avg. R@1 difference recorded on a 5×10^{-4} grid; dense ranks retain finer query-level ordering information. The purpose of this denser metric is measurement resolution, not to turn a small nonzero effect into a claim of statistical significance or robustness to resampling.

4.2 STATE–OPERATOR COMPARISON AND ORACLE HEADROOM

For a matched receiver-state set $\mathcal{S}$ and the two transfer operators $\Omega = \{\mathcal{O}_{\text{fix}}, \mathcal{O}_{\text{norm}}\}$, excluding NoTransfer, the best fixed operator is

$$\mathcal{O}^{\star}_{\text{fixed}} = \arg\max_{\mathcal{O}\in\Omega} \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \Delta U_{\text{S}}(s, \mathcal{O}), \tag{10}$$

whereas a retrospective statewise operator oracle obtains

$$\overline{\Delta U_{\text{S}}}_{\text{op-oracle}} = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \max_{\mathcal{O}\in\Omega} \Delta U_{\text{S}}(s, \mathcal{O}). \tag{11}$$

A safe oracle additionally admits the zero-utility NoTransfer action:

$$\overline{\Delta U_{\text{S}}}_{\text{safe-oracle}} = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \max\left(0, \max_{\mathcal{O}\in\Omega} \Delta U_{\text{S}}(s, \mathcal{O})\right). \tag{12}$$

Both are retrospective ceilings over the audited states and evaluated action set. They use realized outcomes and therefore do not estimate deployable selector performance or unseen-state population utility. Importantly, the safe-minus-operator difference measures only the incremental value of withholding *after* perfect choice between the two transfer operators; it does not quantify the value of transfer-versus-withhold for a fixed operator.

4.3 PRE-INTERVENTION EVIDENCE

RQ3 freezes eight quantities available before relation correction:

$$E_{\text{pre}} = [r/20,\ \Delta_F,\ \|\Delta_r^{\text{FL}}\|_2,\ R_{\text{rel}},\ \bar{m},\ \bar{u},\ \bar{\sigma}_m,\ N_{\text{rank}}].$$

Here Δ_F is the pre-transfer foundation-selector gain (with selected-minus-raw fallback), R_{rel} is mean accumulated rank-evidence reliability, and $\bar{m}$, $\bar{u}$, and $\bar{\sigma}_m$ average the I2T/T2I margin mean, uncertainty mean, and margin-standard-deviation summaries; N_{rank} is the amount of accumulated rank evidence. Candidate relation-update statistics, SafeCommit variables, post-intervention quantities, and final-test information are excluded. These features predict FL-normalized utility/harm only, not relative operator preference.

Strict held-out protocol. Continuous utility uses Bayesian Ridge under seed-grouped LOSO, with median imputation and standardization fitted on training seeds only. For harm estimation, each state's soft target is the fraction of 5,000 paired query-bootstrap resamples with $\Delta\text{MRR} < 0$; an ℓ_2 logistic model ($C = 1$) uses the same eight features. Nested seed-grouped LOSO holds out one seed (five states); inner folds choose the transfer/veto threshold by minimizing mean oracle regret, then harmful-selected count, then transfer rate. AUROC uses outer-fold harm probabilities directly and is threshold-free.

5 EXPERIMENTS

The primary setting is federated multimodal retrieval on MS COCO (Lin et al., 2014) and Flickr30K (Young et al., 2014) with 20 clients, four sampled per round, a frozen CLIP ViT-L/14 foundation model, and a CLIP ViT-B/32 receiver (Radford et al., 2021). For severe non-IID partitioning, samples are binned by text-length deciles and distributed across clients using Dirichlet $\alpha = 0.1$. We audit 75 matched states: 50 canonical COCO states and an independently constructed 25-state cohort with 15 COCO and 10 Flickr30K states. External states use COCO seeds 2034–2036 and Flickr30K seeds 3034–3035 at rounds $4, 8, 12, 16, 19$, with 1000 aligned retrieval queries per branch.

For every utility, NoTransfer is forked from the same pre-intervention state under the contract above. Fixed-step and FL-normalized share the relation signal, source state, gate, transfer locus, and downstream execution and differ only in scale. All 75 states match exactly across the two operator analyses using full-precision rank artifacts. The official final test is isolated from intervention construction and matched-branch evaluation and is used only for the terminal policy comparison in RQ4.

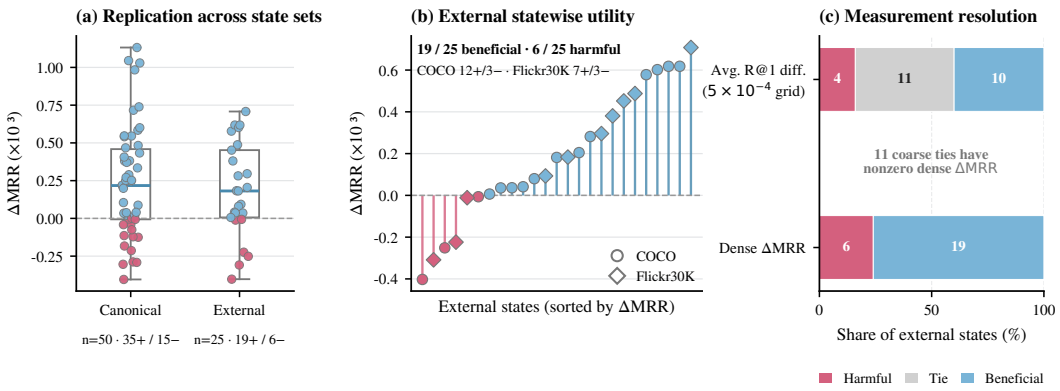


Figure 2: **Replication and measurement resolution. (a)** Dense bidirectional ΔMRR on canonical and external states. **(b)** External-state utility sorted by ΔMRR (COCO circles; Flickr30K diamonds). **(c)** Bidirectional Avg. R@1 differences recorded on a 5×10^{-4} grid versus dense ΔMRR; all 11 coarse ties have nonzero dense values. Signs are descriptive, not per-state significance tests.

Table 1: **Replication of intervention-level transfer utility.** FL-normalized transfer is evaluated with dense bidirectional ΔMRR. Counts and means are descriptive; states within a seed are dependent.

Cohort	States	Beneficial	Harmful	Mean ΔMRR	Median ΔMRR
Canonical	50	35 (70.0%)	15 (30.0%)	2.361×10^{-4}	2.180×10^{-4}
External	25	19 (76.0%)	6 (24.0%)	1.872×10^{-4}	1.816×10^{-4}
Combined	**75**	**54 (72.0%)**	**21 (28.0%)**	$\mathbf{2.198\times10^{-4}}$	$\mathbf{1.993\times10^{-4}}$

RQ1 uses the canonical 50 and external 25 states for FL-normalized utility/replication; RQ2 evaluates both operators on all 75 states. RQ3 uses only the canonical 50: external states are not pooled into fitting, threshold selection, or scoring. Because five receiver states can share one optimization seed, state-level counts and means are descriptive rather than IID/binomial estimates, and RQ3 holds out complete seed groups.

5.1 RQ1: How Heterogeneous Is Realized Transfer Utility?

We first ask whether a fixed transfer rule has stable receiver-side utility. On the canonical 50-state cohort, FL-normalized transfer is beneficial in 35 states and harmful in 15, with mean dense bidirectional $\Delta\text{MRR} = 2.361 \times 10^{-4}$ and median 2.180×10^{-4}. The independent external cohort contains 19 beneficial and 6 harmful states, with mean 1.872×10^{-4} and median 1.816×10^{-4} (Fig. 2(a), Table 1). Thus, positive mean utility over the audited states coexists with a persistent harmful tail; the positive mean should not be read as statewise dominance.

The external cohort reproduces both signs on both datasets: COCO contributes 12 beneficial and 3 harmful states, while Flickr30K contributes 7 beneficial and 3 harmful states (Fig. 2(b)). Because Flickr30K contributes only 10 states, these dataset-specific proportions are used as qualitative replication rather than precise population harm-rate estimates.

Figure 2(c) isolates a measurement issue from the transfer conclusion itself. The bidirectional Avg. R@1 difference, recorded on a 5×10^{-4} grid, yields 10 beneficial, 4 harmful, and 11 exact ties in the 25-state external cohort. Dense bidirectional ΔMRR yields 19 beneficial and 6 harmful states with no exact ties; all 11 coarse ties have nonzero dense values. We therefore use dense ΔMRR to resolve finer receiver-side ranking changes, while treating its sign as descriptive rather than as evidence that each small change is statistically or practically significant.

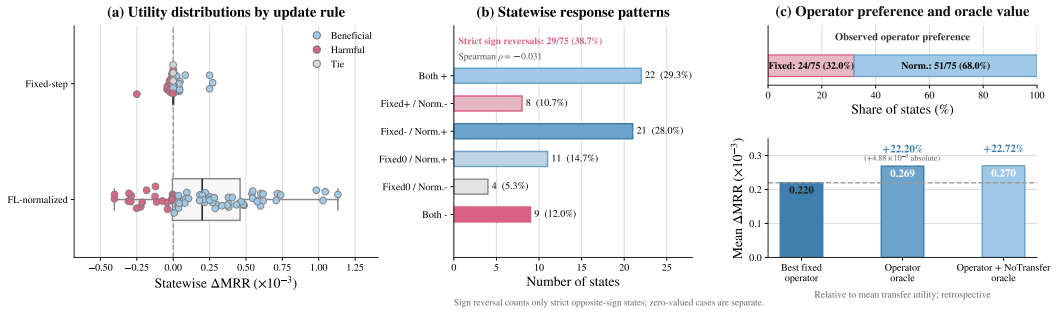


Figure 3: **State–operator dependence and oracle value. (a)** Matched utility distributions for two controlled update-scale rules. **(b)** Strict sign reversal occurs in 29/75 states (38.7%); zero-valued cases are separate. **(c)** Fixed-step is preferred in 24/75 states and FL-normalized in 51/75. The retrospective operator oracle adds 4.88×10^{-5} mean ΔMRR (22.2%) over the best fixed operator; adding NoTransfer raises this to 22.7%. Percentages are relative to mean transfer utility, not receiver MRR.

Table 2: **State–operator structure and oracle headroom.** Matched results over 75 receiver states. Oracle percentages are relative to the best fixed mean transfer utility; ρ and r denote Spearman and Pearson association, respectively.

Measure	**Fixed-step**	**FL-normalized**	**Statewise / Oracle**
Beneficial / harmful / tie	30 / 30 / 15	54 / 21 / 0	–
Mean ΔMRR	8.72×10^{-6}	$\mathbf{2.198\times10^{-4}}$	–
Median ΔMRR	0	$\mathbf{1.993\times10^{-4}}$	–
Preferred states	**24/75 (32.0%)**	**51/75 (68.0%)**	–
Sign reversal	–	–	**29/75 (38.7%)**
Cross-operator association	–	–	$\rho = -0.031,\ r = +0.032$
Best fixed mean	–	2.198×10^{-4}	–
Operator oracle	–	–	2.686×10^{-4} **(+22.20%)**
Safe oracle	–	–	2.697×10^{-4} **(+22.72%)**

5.2 RQ2: Is Transferability Jointly State- and Operator-Dependent?

RQ1 establishes receiver-state heterogeneity for a fixed operator but does not show whether response also depends on the update scale. Fixed-step and FL-normalized are therefore evaluated on the same 75 pre-intervention receiver states, at the same transfer locus and against the same matched NoTransfer control.

Changing only the applied update scale changes observed receiver utility. Fixed-step is approximately centered at zero (mean $\Delta\text{MRR} = 8.72 \times 10^{-6}$), whereas FL-normalized has higher mean utility (2.198×10^{-4}). Yet 29/75 states (38.7%) strictly reverse sign across the two operators, and statewise utilities have near-zero association (Spearman $\rho = -0.031$, Pearson $r = +0.032$). Sign reversals occur in both cohorts (16/50 canonical; 13/25 external), so the pattern is not confined to one audit set. Because the underlying relation direction is shared when the gate is active, these results show that collinear parameter corrections at different scales can induce different retrieval outcomes; lower mean utility does not imply uniform statewise inferiority. The 15 Fixed-step ties are retained as exact observed zeros, but the aggregate results do not identify whether any particular zero arises from gating or from unchanged retrieval ranks.

The lower-mean Fixed-step operator nevertheless yields higher utility in 24/75 states (32.0%), versus 51/75 for FL-normalized. The retrospective operator oracle raises mean utility from 2.198×10^{-4} to 2.686×10^{-4}, an absolute increment of 4.88×10^{-5} and a 22.2% increase relative to the best fixed mean transfer utility. Allowing NoTransfer in addition raises the oracle mean only to 2.697×10^{-4} (22.7% above the best fixed operator). The latter comparison supports only a narrow statement: once the better of the two transfer operators is already known in hindsight, adding the option to

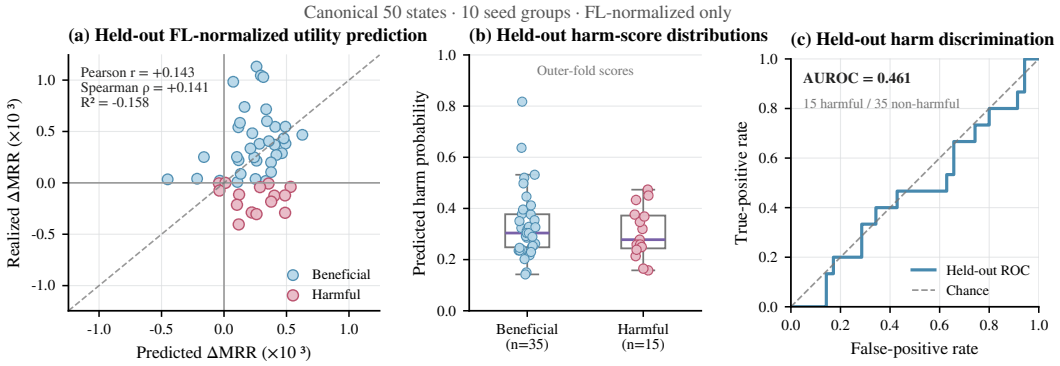


Figure 4: **Held-out FL-normalized utility prediction on 50 canonical states.** Seed-grouped LOSO evaluates eight pre-intervention features. **(a)** Pearson $r = 0.143$, Spearman $\rho = 0.141$, and $R^2 = -0.158$. **(b)** Outer-fold harm scores overlap strongly. **(c)** AUROC $= 0.461$ for 15 harmful versus 35 non-harmful states. These single-operator results do not test relative operator preference or establish an information-theoretic limit.

withhold provides a small additional gain on these audited states. It does not measure the value of transfer-versus-withhold for a fixed operator.

5.3 RQ3: Can Pre-Intervention Evidence Predict FL-Normalized Utility?

RQ3 is deliberately narrower than operator selection: it asks whether the evaluated pre-intervention features and model families predict FL-normalized utility or harm. The frozen canonical 50-state cohort is scored only through seed-grouped held-out predictions; the independent 25-state cohort is excluded from fitting and scoring, so the analysis should not be described as external predictor validation.

Figure 4(a) shows weak held-out point estimation: Pearson $r = 0.143$, Spearman $\rho = 0.141$, and $R^2 = -0.158$. The negative R^2 indicates that the reported predictor does not outperform the corresponding mean-style squared-error reference on these held-out states; it is not evidence that every form of pre-intervention information is uninformative. Harm prediction is similarly weak. Outer-fold scores for beneficial and harmful states overlap substantially, and threshold-free harm discrimination is AUROC $= 0.461$ for 15 harmful versus 35 non-harmful states (Fig. 4(b,c)).

These results therefore establish limited prospective predictive performance for the *evaluated* eight-feature representation and the specified Bayesian-Ridge/logistic models. They do not directly test the relative preference target in RQ2 and cannot quantify how much of the retrospective operator-oracle headroom a different selector might recover. A separate frozen rollback diagnostic reaches a related but distinct boundary: it commits 49/50 interventions, prevents 0/15 observed harmful cases, and has harm-score AUROC 0.472, despite exact restoration of the pre-intervention state when rollback occurs. Mechanical reversibility is thus an execution property rather than evidence of prospective harm recognition.

5.4 RQ4: How Do One-Step Audits Relate to 20-Round Outcomes?

Intervention-level and trajectory-level utility are different estimands. The former compares Transfer and NoTransfer branches from the same receiver state after a matched downstream replay. The latter compares complete repeated-transfer and never-transfer training policies, whose earlier actions induce different future receiver-state sequences. RQ4 therefore places the 20-round policy comparison alongside the one-step audits without treating their effects as interchangeable.

Figure 5(a) shows round-wise policy-level Avg. R@1 differences between repeated FL-normalized transfer and matched NeverTransfer training. These differences are non-monotonic and can change sign over the 20-round trajectories, so a positive terminal outcome does not imply that every intermediate state or intervention was locally beneficial. Panel (b) evaluates a different object on the protocol-defined final held-out test: all five endpoints are positive, with COCO differences of $+0.50$,

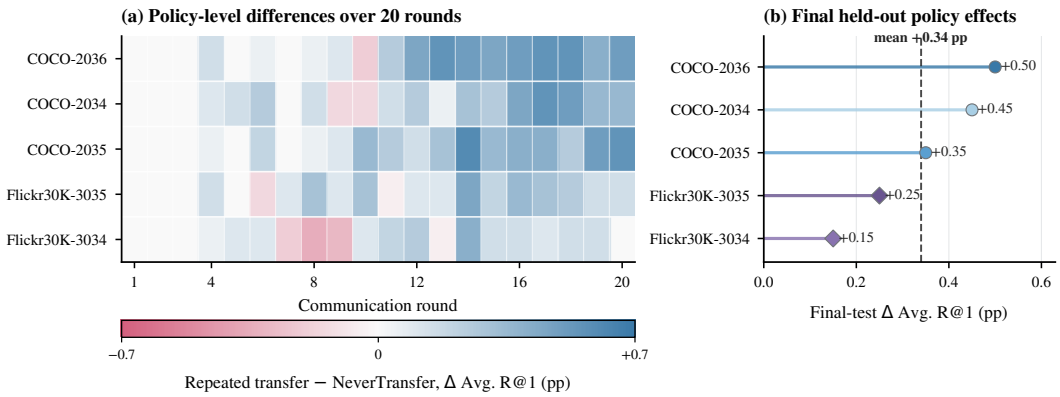


Figure 5: **Twenty-round policy differences and final held-out effects. (a)** Round-wise Avg. R@1 differences for five repeated FL-normalized minus NeverTransfer trajectories; these are policy-level, not isolated one-step effects. **(b)** Final held-out effects are positive for all five pairs (mean $+0.34$ pp; median $+0.35$ pp). Training-time and final-test roles remain separate.

$+0.45$, and $+0.35$ percentage points and Flickr30K differences of $+0.25$ and $+0.15$ points (mean $+0.34$, median $+0.35$ pp).

The one-step audits use dense bidirectional MRR, whereas the trajectory comparison uses Avg. R@1 and separate evaluation roles. The supported conclusion is therefore coexistence at the study level: harmful one-step audit outcomes can be observed in the same study in which repeated-transfer policies finish above their matched NeverTransfer trajectories. These results neither attribute the terminal gains to particular harmful interventions nor imply that harmful local effects are necessary for, or generally redeemed by, later gains.

6 Limitations and Discussion

Our empirical claims are scoped to CLIP ViT-L/14→ViT-B/32 transfer on COCO and Flickr30K, two controlled update-scale rules, and receiver states sampled from a limited set of optimization seeds. States sharing a seed are dependent, so state counts, sign proportions, means, and oracle values are reported descriptively rather than as independent population estimates. The intervention effects are also small in absolute MRR units; dense measurement improves resolution but does not by itself establish per-state statistical or practical significance. The prospective analysis concerns FL-normalized transfer, eight pre-intervention features, and two model families; it does not test relative operator preference or establish an impossibility result. Finally, the five positive 20-round endpoints show policy-level coexistence with negative one-step audits, not a causal path from harmful interventions to later benefit. Broader operators, model pairs, datasets, and independently constructed prospective prediction cohorts remain important tests of generality.

7 Conclusion

SCOUT studies cross-model transfer from the receiving learner's perspective through matched interventions rather than average method performance alone. Across 75 audited states, changing only update scale reverses observed utility sign in 38.7% of cases, while a retrospective statewise operator oracle adds 4.88×10^{-5} mean ΔMRR, or 22.2% relative to the best fixed operator. At the same time, the evaluated pre-intervention features and models provide limited seed-held-out prediction of FL-normalized utility and harm and do not test relative operator preference. Five matched 20-round repeated-transfer policies finish with positive held-out effects; separately, the intervention audits contain negative one-step utilities. Together, **in the evaluated setting**, these results separate four questions that are easily conflated: whether transfer is beneficial on average, how utility varies across receiver state and operator, whether that utility is predictable before execution, and how one-step effects relate to long-horizon policy outcomes.

AI USE STATEMENT

In this work, generative AI tools were used for language editing and drafting support, literature discovery and organizational assistance, feedback on experimental design and result interpretation, code-related assistance, and creation or refinement of scientific figures. All AI-assisted text, analyses, code changes, equations, citations, and figures were reviewed and verified by the authors. Generative AI was not used to generate experimental data or replace author verification of technical claims. The authors take full responsibility for the final content of the work.

REPRODUCIBILITY STATEMENT

The main text states the matched-intervention estimand, fork timing/state, relation-preference construction and loss, compatibility gate, both update-scale operators, frozen eight-feature held-out protocol, dense-MRR utility, cohort roles, oracle definitions, rollback diagnostic, and 20-round terminal comparison. The relation generator and gate define the audited intervention but are held fixed rather than claimed as SCOUT contributions.

REFERENCES

Jun Bai, Zhuofan Chen, Zhenzi Li, Hanhua Hong, Jianfei Zhang, Chen Li, Chenghua Lin, and Wenge Rong. Leveraging estimated transferability over human intuition for model selection in text ranking. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12356–12374. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.688.

Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. FedDAT: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11285–11293, 2024. doi: 10.1609/aaai.v38i10.29007.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4794–4802, 2019.

Zixiang Ding, Guoqing Jiang, Shuai Zhang, Lin Guo, and Wei Lin. How to trade off the quantity and capacity of teacher ensemble: Learning categorical distribution to stochastically employ a teacher for distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17915–17923, 2024. doi: 10.1609/aaai.v38i16.29746.

Chaoyang He, Murali Annavaram, and Salman Avestimehr. Group knowledge transfer: Federated learning of large CNNs at the edge. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9201–9225. PMLR, 2022.

Huy Q. Le, Minh N. H. Nguyen, Chu Myaet Thwal, Yu Qiao, Chaoning Zhang, and Choong Seon Hong. Fedmekt: Distillation-based embedding knowledge transfer for multimodal federated learning. *Neural Networks*, 183:107017, 2025. doi: 10.1016/j.neunet.2024.107017.

Dongkyu Lee, Zhiliang Tian, Yingxiu Zhao, Ka Chun Cheung, and Nevin Zhang. Hard gate knowledge distillation – leverage calibration for robust and reliable language model. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9793–9803. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.665.

Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Min Liu, Yu Bao, Chengqi Zhao, and Shujian Huang. Selective knowledge distillation for non-autoregressive neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13246–13254, 2023. doi: 10.1609/aaai.v37i11.26555.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5191–5198, 2020. doi: 10.1609/aaai.v34i04.5963.

Alessio Mora, Irene Tenison, Paolo Bellavista, and Irina Rish. Knowledge distillation in federated learning: A practical guide. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 8188–8196, 2024. doi: 10.24963/ijcai.2024/905.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305. PMLR, 2020.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8432–8440, 2022. doi: 10.1609/aaai.v36i8.20819.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6456–6466. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.504.

Xianda Wang, Yaqi Qiao, Duo Wu, Chenrui Wu, and Fangxin Wang. Cluster based heterogeneous federated foundation model adaptation and fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21269–21277, 2025. doi: 10.1609/aaai.v39i20.35426.

Zirui Wang, Zihang Dai, Barnabas Poczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019.

Yuankun Xia, Hui Wang, and Yufeng Zhou. Fedclip-distill: Heterogeneous federated cross-modal knowledge distillation for multi-domain visual recognition. *Knowledge-Based Systems*, 337: 115383, 2026. doi: 10.1016/j.knosys.2026.115383.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 2021.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. An upload-efficient scheme for transferring knowledge from a server-side pre-trained generator to clients in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12109–12119, 2024.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2023. doi: 10.1109/JAS.2022.106004.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12878–12889. PMLR, 2021.