WHAT DOES CONTINUAL RETRIEVAL RETAIN? A CANDIDATE-PROVENANCE STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual vision-language learning aims to acquire new retrieval capabilities while preserving those learned earlier. Yet evaluating each task within its own gallery leaves an important question open: do conclusions drawn from task-local evaluation still hold when candidates from multiple domains compete in a shared gallery? We evaluate CLIP ViT-B/32 under sequential fine-tuning and Mod-X across five domains at pre-task, post-task, and final checkpoints, using fixed task-local (Local) and five-domain (Shared) galleries. On locally assembled splits, queries, benchmark-defined positives, and candidate identities remain fixed across checkpoints. Across both configurations and retrieval directions, the same later updates produce 0.51–1.14 percentage points more R@10 loss in Shared than Local, averaged equally over the four tasks with subsequent training. We also follow queries successful at R@10 in both galleries immediately after a task. Among their final Shared failures, approximately 29–32% still succeed in Local when within-domain proportions are averaged equally for each configuration and direction. Under a fixed budget of 512 added negatives, the expected R@10 loss differs by source and can increase or decrease after later updates. These findings show that task-local evaluation alone can miss losses of previously successful retrievals under shared candidate competition. Assessing what continual retrieval retains therefore requires considering the candidate environment: how much post-task success survives subsequent updates, and which queries remain successful when domains share a search space.

1 INTRODUCTION

Continual image-text learning aims to acquire new retrieval capabilities while preserving earlier ones. Retrieval success also depends on the collection being searched. A query may remain successful in its task-local gallery after later training but fail when candidates from other domains enter the same ranked list. We ask whether conclusions about acquisition and retention drawn from a task-local gallery (Local) remain valid in a shared gallery (Shared).

More candidates can lower recall even when the model is unchanged. What matters for continual learning is whether later model updates widen or narrow that Local–Shared difference. A gap may already exist before a task, shrink when the task is learned, and widen during later updates. Final performance may remain above its initial level while still falling below the level reached immediately after learning the task. To interpret retention, we follow the same queries across both candidate environments and these distinct learning stages.

We track CLIP ViT-B/32 learning five domains through sequential fine-tuning or Mod-X. Each gallery retains the same candidate identities across checkpoints, so every model state faces the same candidates within that environment. We measure gains immediately after each task and losses after later updates, follow the same queries to see which successes survive, and control the number and source of added candidates.

For each learner and retrieval direction, we average R@10 loss from post-task to final over the four tasks with later training. For the same updates, **mean loss is 0.51–1.14 percentage points greater in Shared than in Local** (Table 2). All four primary simultaneous confidence intervals support an additional Shared loss. At R@1, both environments lose performance after later updates, but the confidence intervals do not establish an additional loss in Shared. Paired query histories show what

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.

the averages conceal: some queries successful in both galleries immediately after learning a task remain successful only in Local after later updates. Following this same post-task cohort identifies which subsequent Shared failures a task-local score misses, with substantial variation across domains. On a domain-equal basis, queries successful in both galleries before the task contribute more lost Shared coverage than queries newly successful in Shared during it.

Equal-budget controls show that the expected effect of added candidates varies by source and checkpoint. For the limited queries supporting score-matched replacement, the R@10 other-versus-same-source contrast changes little after later updates.

For a system returning a shortlist from a combined collection, query survival matters alongside average score changes. Task-local retention scores alone do not identify which previously successful queries still retrieve a known relevant result under heterogeneous competition. The five-domain study measures this difference during later updates; a separate out-of-stream SketchScene case examines when candidate replacement also reverses a model comparison. Retention evaluation should therefore specify the post-task checkpoint used as the reference, the gallery being searched, and the retrieval cutoff K.

2 Related Work

Continual vision-language retrieval. CLIP learns a shared image–text representation through contrastive pretraining (Radford et al., 2021). Mod-X preserves cross-modal relations using off-diagonal information (Ni et al., 2023). DKR rectifies transferred knowledge (Cui et al., 2024), while C-CLIP addresses multimodal acquisition and retention (Liu et al., 2025). CompoReAlign studies primitive–composition alignment (Xiao et al., 2026). DARAD evaluates an expanding all-seen gallery and separately compares adjacent models on a fixed historical gallery (Chen et al., 2026). For each learner and task, we evaluate both fixed galleries at the pre-task, post-task, and final checkpoints and track the same queries. We ask which conclusions about acquisition and retention agree across environments.

Measuring what is retained. Evaluation choices expose different aspects of continual learning. De Lange et al. (2023) show that task-end evaluation can miss temporary losses during training. Davari et al. (2022) distinguish observed task performance from information accessible through a newly fitted linear probe on frozen features. In recommendation, Krichene & Rendle (2020) show that sampled-candidate metrics need not preserve full-ranking model comparisons, even in expectation. We study retrieval success under a fixed retrieval rule while varying candidate competition and the checkpoint used as the reference for measuring change. The resulting query histories distinguish losses of pre-task success from losses of success acquired during the task. Our sampling calculations evaluate specified candidate environments; full Local and Shared performance is measured directly.

3 Study Design: Measuring Retrieval under Specified Candidate Environments

Queries, positives, and galleries. For an image or text query q, let $P(q)$ be its benchmark-defined direct positives and s_t the similarity at checkpoint t. In a gallery G containing $P(q)$, the rank of the highest-scoring positive is

$$r_t(q; G) = 1 + \sum_{x \in G \setminus P(q)} \mathbf{1}\left[s_t(q, x) > \max_{p \in P(q)} s_t(q, p) \right]. \tag{1}$$

Here, the candidate environment is the gallery searched; provenance records a candidate's benchmark source or sources. A blocker is a benchmark-negative ranked above the best positive; exact ties favor the positive. Reported R@K is 100 times the mean of $\mathbf{1}[r_t \le K]$: at least one annotated positive is among the first K results. MRR averages $1/r_t$ and is reported after multiplication by 100. *Local* contains candidates from the query's domain deduplicated by identity; *Shared* is the union of all five domains, also deduplicated by identity. For every retained query in all ten domain–direction groups, both environments use identical queries and protocol-allowed direct positive matches. Adding Shared negatives therefore cannot improve an individual rank at a fixed checkpoint.

Each gallery, query set and direct-positive set is reused across checkpoints (Figure 1). Shared contains the same five-domain candidate identities throughout training, including candidates from tasks not yet learned. It stays fixed rather than growing as tasks arrive. The two environments differ in size and initial difficulty. Keeping candidate identities fixed allows us to measure how model updates change the Local–Shared performance gap. Appendix A distinguishes the all-seen gallery and contrasts the candidate designs (Table 4), including recorded ranks and sampling expectations.

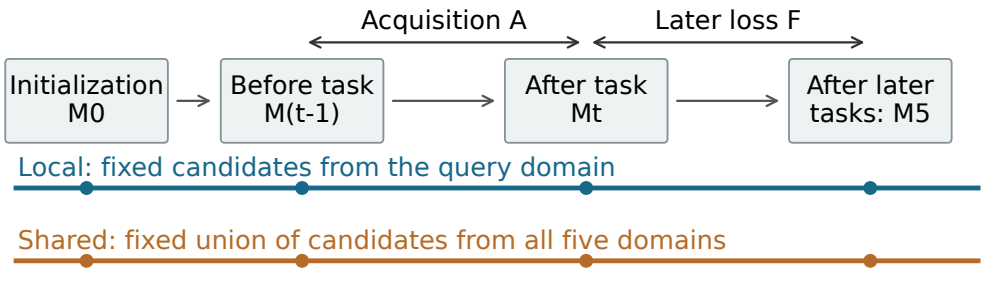

Figure 1: Evaluation checkpoints and fixed candidate galleries. Acquisition compares the checkpoints immediately before and after a task; retention compares its post-task and final checkpoints. Local and Shared each keep candidate identities fixed while the model changes. Initialization is a separate reference.

Acquisition and retention use different references. For domain d at task position τ_d, M_0 is the common pretrained initialization and M_t ($t = 1, \ldots, 5$) the checkpoint saved after task t. Thus $M_{\tau_d - 1}$ is pre-task, M_{τ_d} post-task, and M_5 final. Let $R^e_{d,t}$ be performance for domain d in environment $e \in \{L, S\}$, and let $T = 5$ be the final checkpoint. We define acquisition A and net gain over initialization Z by

$$A^e_d = R^e_{d,\tau_d} - R^e_{d,\tau_d - 1}, \qquad Z^e_d = R^e_{d,T} - R^e_{d,0}. \tag{2}$$

For tasks with later updates ($\tau_d < T$), retention loss is $F^e_d = R^e_{d,\tau_d} - R^e_{d,T}$. Positive A/Z denote improvement; positive F denotes loss. In general, $A - F \neq Z$ because performance immediately before a task need not equal performance at initialization. With competition gap $D_{d,t} = R^L_{d,t} - R^S_{d,t}$, the extra retention loss under Shared is

$$X_d = F^S_d - F^L_d = D_{d,T} - D_{d,\tau_d}, \qquad \tau_d < T. \tag{3}$$

Thus X is the post-task-to-final change in the Local–Shared gap, not the gap itself. We average A/Z equally over five domains and F/X over the first four. The final task has no subsequent retention observation and is marked N/A, rather than contributing a zero. Recall levels are percentages and changes percentage points (pp); MRR$\times$100 changes are index points.

Data and executed configurations. The task stream is COCO $\rightarrow$ Flickr30K $\rightarrow$ IAPR TC-12 $\rightarrow$ FashionGen $\rightarrow$ RSICD (Lin et al., 2014; Young et al., 2014; Grubinger et al., 2006; Rostamzadeh et al., 2018; Lu et al., 2017). The local export contains 41,285 unique images and 44,183 unique texts. The evaluation splits are locally assembled. Some normalized evaluation texts reappear in training data; scene-level and pretraining independence remain unverified. Positives follow direct retained associations, including FashionGen product-level associations. Components group observations for resampling and internal validation, but do not add positive pairs by transitive closure. Appendix A gives exclusions, text reuse, positive multiplicity, and tokenization scope.

The main study evaluates eleven states: the shared Frozen CLIP ViT-B/32 initialization and five checkpoints each for SeqFT and Mod-X. Both configurations train for five epochs per task. Retained code specifies weight-decay defaults of 0.2 (SeqFT) and 0.1 (Mod-X), but the historical run-time values could not be independently verified. Comparisons concern these two executed configurations, whose settings and encoding paths are specified in Appendix A.1.

Candidate controls answer different questions. The natural-source controls add or remove an entire source pool. FLC512 retains full Local and adds 512 negatives from one external source, measuring expected recall under uniform sampling without replacement. Small-budget controls instead use a sampled Local base. Score-constrained replacement fixes a query-specific background and exchanges equal-sized blocks matched at the learner's post-task checkpoint. For each source, we include only queries for which the prescribed candidate construction succeeds under all three seeds; this supported population differs from the full native population. Each design addresses a distinct comparison, with its own candidate environment and query population (Section 6; Table 4).

Inference. Our primary analysis jointly evaluates eight X contrasts over the complete retained query populations, termed native: the combinations of two configurations, two retrieval directions, and R@1/R@10. The original analysis uses 10,000 component-bootstrap resamples of 17,615 positive-graph components, sharing weights across all eight contrasts to construct approximate simultaneous 95% intervals. Each draw recomputes domain-specific query denominators before averaging four tasks. Uncertainty is conditional on the executed models and galleries; it does not cover training seeds, task orders, or arbitrary gallery variation. Other metrics, paths, source analyses, and selection summaries are descriptive unless stated otherwise. Protocol timing, dependence limits, and saved-draw reconstruction appear in Appendix A.2.

4 What Is Acquired and Retained?

Both environments show average learning gains followed by later losses. Table 1 shows positive average acquisition and net gain despite later losses. For SeqFT i2t R@1, Local acquisition and net gain average 9.57 and 9.48 pp across five domains; post-task-to-final loss averages 6.02 pp over four tasks with later training, versus 6.57 pp in Shared. The primary environment contrast compares these losses: substantial R@1 forgetting in both environments coexists with a much smaller difference between them.

Table 1: Acquisition A (post-task minus pre-task), later loss F (post-task minus final), and net gain Z (final minus initialization), in pp. A and Z average five domains equally; F averages the four tasks with later training equally. L: Local; S: Shared.

Metric	Model	Dir.	A_L	A_S	F_L	F_S	Z_L	Z_S
R@1	SeqFT	i2t	9.57	9.44	6.02	6.57	9.48	9.70
R@1	SeqFT	t2i	11.68	11.69	6.81	7.00	10.53	11.46
R@1	Mod-X	i2t	9.37	9.47	4.40	4.52	10.52	11.27
R@1	Mod-X	t2i	9.19	9.32	4.44	4.62	10.42	11.44
R@10	SeqFT	i2t	14.77	14.65	3.30	4.45	14.73	15.66
R@10	SeqFT	t2i	16.84	17.60	5.13	6.11	16.08	18.16
R@10	Mod-X	i2t	13.80	13.93	2.74	3.26	15.21	16.72
R@10	Mod-X	t2i	14.79	15.89	3.73	4.54	15.91	18.08

A large final Local–Shared gap can coexist with improvement over initialization. Figure 2 reports results for all five domains in both retrieval directions. Its lower panel follows SeqFT Flickr30K t2i R@10. Immediately before learning Flickr (M1), Local/Shared recall is 95.56/86.92%; immediately afterward (M2), it is 96.72/88.46%. Task acquisition is therefore 1.16/1.54 pp. Subsequent updates reduce recall to 93.68/82.70%: losses of 3.04/5.76 pp, with an extra Shared loss $X = 2.72$ pp.

The final Local–Shared gap ($D_5 = 10.98$ pp) far exceeds its post-task widening ($X = 2.72$ pp). The final gap is also smaller than the initial gap of 15.18 pp, and final recall in both environments remains above Frozen. The M0–M2 improvement includes the preceding COCO update, so it cannot all be attributed to learning Flickr. FashionGen t2i R@10 has an exact zero gap at initialization, post-task, and final despite post-task losses of 4.04 pp (SeqFT) and 4.69 pp (Mod-X) in each environment. Appendix B reports all stages and all four metrics.

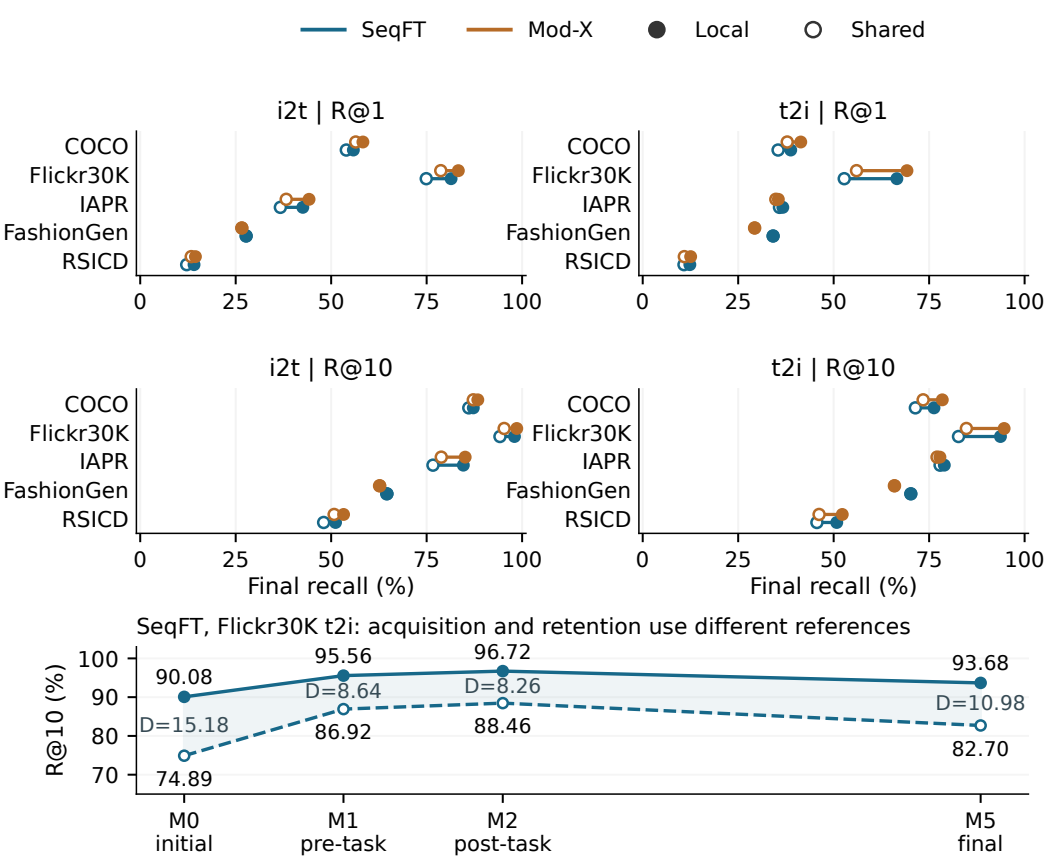

Figure 2: **Top:** final absolute recall for all five domains; filled/open markers denote Local/Shared. **Bottom:** a measured task-aligned trajectory, with M0 as a separate initial reference. D is the Local–Shared gap (pp); solid/dashed lines denote Local/Shared. Values and differences are calculated before rounding.

Extra retention loss is supported at R@10. All four R@10 primary intervals exclude zero; their point estimates range from 0.51 to 1.14 pp (Table 2). All four R@1 extra-loss estimates are also positive, but their intervals include zero. The additional loss at R@1 remains uncertain. These intervals do not establish equivalent Local and Shared R@1 retention losses. Support at R@10 but not R@1 also does not establish that the two effects differ. All R@10 point contrasts remain positive when any one domain is omitted. Magnitudes still depend on composition: excluding Flickr reduces t2i X from 0.978 to 0.396 pp for SeqFT and from 0.818 to 0.437 pp for Mod-X.

Table 2: Extra Shared versus Local post-task-to-final loss, $X = F_S - F_L$ (pp); positive values mean greater Shared loss. Approximate simultaneous 95% component-bootstrap intervals are constructed jointly for all eight contrasts.

Model	Dir.	R@1 X	Simultaneous interval	R@10 X	Simultaneous interval
SeqFT	i2t	0.551	[-0.106, 1.208]	1.145	[0.681, 1.609]
SeqFT	t2i	0.197	[-0.226, 0.620]	0.978	[0.619, 1.336]
Mod-X	i2t	0.128	[-0.453, 0.709]	0.514	[0.085, 0.944]
Mod-X	t2i	0.172	[-0.208, 0.553]	0.818	[0.493, 1.143]

Native endpoint ordering across galleries. Complete-population endpoint rankings agree across galleries for every domain, direction and metric, including Frozen CLIP. Internal selection can differ (Appendix C.3).

5 Which Successes Survive Later Updates?

Successes present in both environments after a task can survive later updates in Local while disappearing in Shared. We follow each query from post-task to final evaluation to distinguish this loss from gains confined to Local. At a fixed threshold, a query succeeds in both environments (C), in Local only (V), or in neither (W). Shared-only success is excluded by the positive-invariant nested galleries. Within each domain, post-task-to-final paths give

$$X_d = 100\big[P(C\to V) + P(W\to V) - P(V\to C) - P(V\to W)\big]. \tag{4}$$

The first term loses Shared success while retaining Local success; the second gains Local success only. Both enlarge the gap. Shared recovery ($V \to C$) and losing Local success ($V \to W$) both shrink it; the latter does not restore Shared success. Thus the same net change can combine different behavioral outcomes. For SeqFT t2i at R@1, the four-domain mean $C \to V$ contribution is 1.766 pp of the original query populations, yet net X is 0.197 pp after the signed paths are combined (Table 7 in Appendix C). Figure 3 separates the R@10 net contrast into its four contributions, then follows Shared-specific losses back to the queries' pre-task success states.

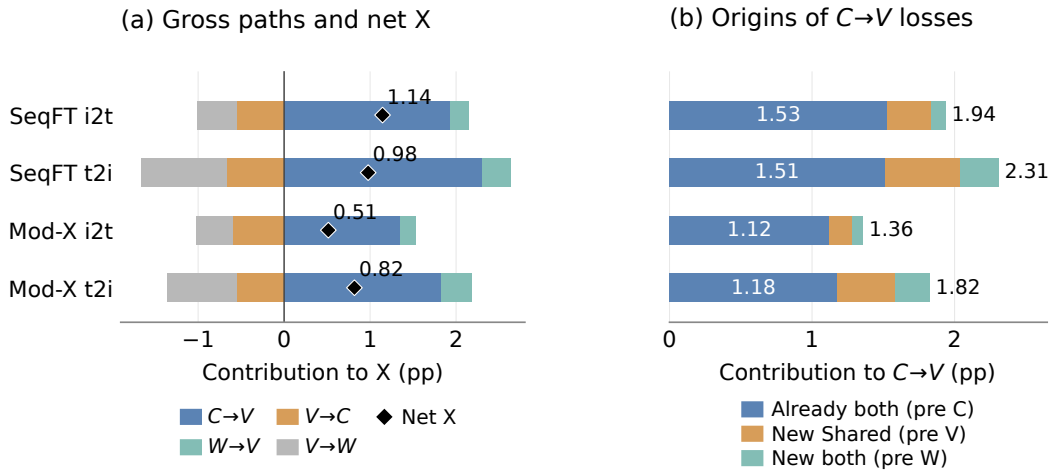

Figure 3: R@10 query histories. Left: four post-task-to-final flow contributions and their net X. Right: the $C \to V$ loss partitioned by pre-task state. Values are means over four domains, each normalized by its original query population (pp), not conditional cohort rates. All eight R@1/R@10 decompositions appear in Appendix C.

Following the same post-task successes. We follow queries successful in both galleries immediately after each task to the final checkpoint. At R@10, the final Local–Shared retention gap is 1.54–2.72 pp across learners and directions, after averaging each of the four eligible domains equally (Table 3). Among their final Shared failures, a domain-equal mean of 28.67–31.81% remain successful locally; FashionGen t2i has none, although some queries fail in both galleries. These conditional rates complement the original-population net contrast X.

Existing and newly acquired successes contribute differently. For SeqFT t2i R@10, the 2.308 pp original-population $C \to V$ loss comprises 0.272 pp newly successful in both galleries, 0.524 pp newly successful in Shared after Local, and 1.512 pp already successful in both before the task. Across all eight domain-averaged summaries, pre-existing successes contribute more than both new groups combined. Within the post-task cohort, however, the newly Shared-successful group has a higher final-V rate in 14 of 16 R@10 domain comparisons; both rates are zero in the other two (Appendix C.2). These denominators differ from Table 3.

Table 3: R@10 retention of post-task common successes (post-C). Coverage uses all queries; final Local/Shared success and their gap use the post-C cohort. The last column uses only final Shared failures within that cohort: $100n_{CV}/(n_{CV} + n_{CW})$, where CV is Local-only success and CW failure in both galleries. Each within-domain rate is averaged equally across four tasks; gaps are computed before rounding.

Model	Dir.	All queries	Within post-C cohort			Shared failures within post-C
		Post-C (%)	Local (%)	Shared (%)	Gap (pp)	Local-only (%)
SeqFT	$i \to t$	84.81	93.97	91.76	2.20	28.90
SeqFT	$t \to i$	81.66	91.86	89.14	2.72	28.67
Mod-X	$i \to t$	84.27	94.75	93.21	1.54	30.40
Mod-X	$t \to i$	79.84	92.82	90.65	2.17	31.81

Rank thresholds admit different loss paths. With b_E external blockers, $r_S = r_L + b_E$; Local success survives in Shared when $b_E \leq K - r_L$. Thus R@1 $C \to V$ needs a new blocker. At R@10, even with an unchanged number of external blockers, worsening Local ranks can cause previously successful queries to fail in Shared. This route contributes 0.033 pp of SeqFT t2i's 2.308 pp $C \to V$ mass. Table 10 partitions all paths; blocker counts need not preserve competitor identities or scores.

6 Candidate Sources, Exposure, and Score Constraints

The query paths identify which annotated successes cross a rank threshold. We next examine the candidate pools around those paths through source additions, fixed-budget exposure, and score-constrained replacement.

A source's effect depends on the surrounding competitors. Adding COCO to Local at SeqFT M5 loses 6.38 pp of Flickr t2i R@10 (319/4,999 queries); removing it from Shared recovers only 4.04 pp (202/4,999). The source is the same but the surrounding competitors differ (Figure 4). At R@10, individually insufficient sources jointly induce failures in 79 of 110 checkpoint–domain–direction conditions, affecting 0–1.680% of queries; this cannot occur at R@1, where one blocker suffices. The largest fraction already occurs at Frozen Flickr t2i. Later changes require following sources across checkpoints.

The FashionGen update raises SeqFT Flickr t2i single-add losses for all four sources (COCO 0.98, IAPR 0.70, FashionGen 0.84, RSICD 0.26 pp). In Mod-X Flickr i2t, COCO's loss instead falls 0.50 pp and FashionGen's stays zero (Appendix D.1).

Source and checkpoint differences persist at a fixed budget. We compute expected recall for complete Local augmented with 512 unique eligible negatives drawn uniformly without replacement from one external source. This design, termed FLC512, is evaluated from query-specific blocker counts without sampling realized galleries. Define the addition loss $\ell_t = R_{L,t} - \mathbb{E}[R_{\mathrm{FLC},t}]$ and its later change $X_{\mathrm{FLC}} = \ell_5 - \ell_\tau = F_{\mathrm{FLC}} - F_L$. Across all 64 R@10 retention contrasts, 39 are positive, 12 negative and 13 exactly zero, spanning -0.06436 to $+0.57250$ pp (Figure 5). With the budget fixed, expected addition loss still differs across sources and can either increase or decrease after later updates.

For SeqFT COCO t2i, RSICD addition loss rises from 0.214 to 0.787 pp, overtaking Flickr (0.516 to 0.591 pp) and IAPR (0.492 to 0.581 pp); FashionGen remains lower (0.016 to 0.032 pp). Thus the expected effect of the same 512-negative exposure depends on both source and checkpoint. Table 18 gives all four source comparisons; Appendix F specifies the identity/count construction.

With a sampled Local base and equal source blocks, other-source negatives are less damaging in all 744 R@1 contrasts at each budget, but not uniformly at R@10. At 64/64 and 128/128, respectively, 701/930 and 587/930 R@10 cells saturate at $\geq 99\%$ (Appendix E).

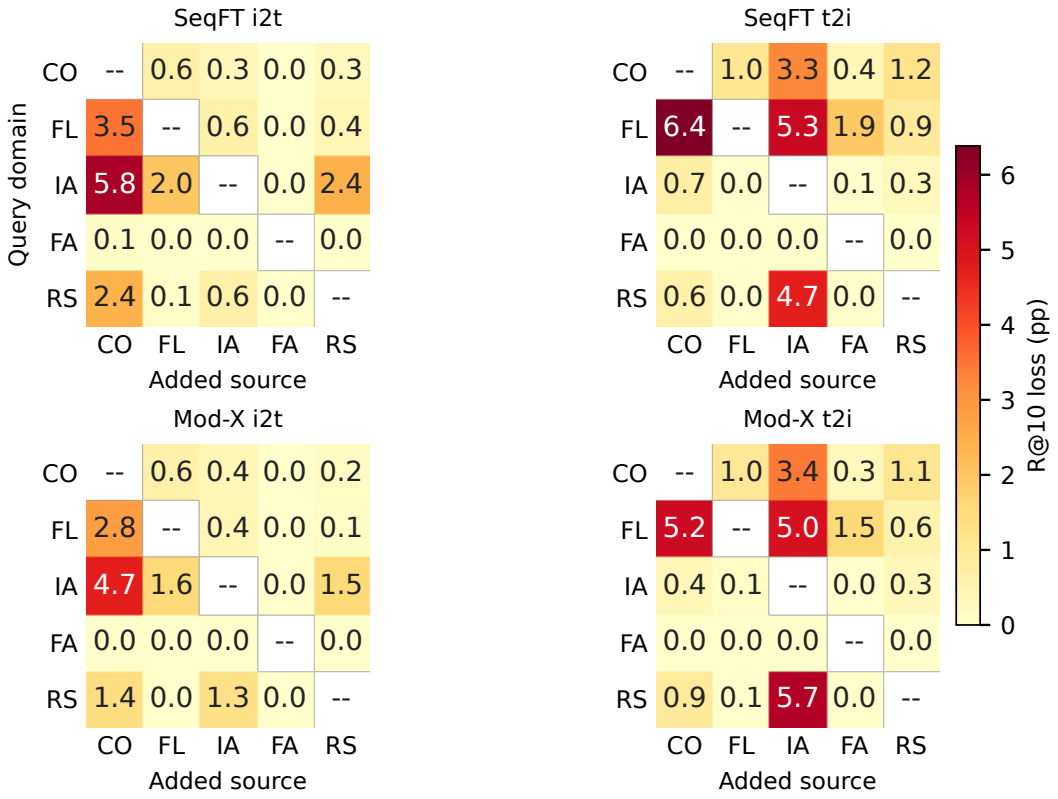

Figure 4: Final R@10 loss (pp) from adding one complete source to Local. Rows are query domains; columns are sources (CO/FL/IA/FA/RS in stream order). Natural pool sizes differ; diagonal entries are N/A and rounded 0.0 need not be exact zero. Appendix D pairs single additions with last removals and Frozen results.

Retaining complete Local instead yields seven-fraction exposure curves: all eight domain-equal extra-loss curves rise here, but intermediate points are sampling expectations, not realized query paths or members of the endpoint inferential family (Figure 12 in Appendix E).

Small source-contrast changes under score-constrained replacement. For COCO/IAPR, we exchange 64-negative blocks against a common query-specific Local background. Disjoint same-source blocks A/B and external block O have matched post-task triplet score spans $\leq .005$; U is random from the external pool. All positives remain. Candidate identities remain fixed from their post-task construction through M5. We track $\Theta = \Delta_5 - \Delta_\tau$, where $\Delta_t = R_{O,t} - (R_{A,t} + R_{B,t})/2$; extra loss has the opposite sign.

On these limited populations, four-source mean R@10 Θ ranges from -0.07331 to $+0.21104$ pp; three usable simultaneous intervals include zero, four comparisons fail the preset query/component support requirements, and one is degenerate. At each evaluated checkpoint, the replacement conditions—including the random control—rarely disagree on R@10 success. Small contrast changes cannot be credited to matching alone or taken as small absolute retention loss. Appendix F.2 gives support, disagreement and lower-threshold results.

Scores move while source contrasts change little. Fixed-identity records distinguish positive-score changes, best-positive switches and relative negative-score movement. For SeqFT COCO i2t/Flickr, cross-minus-same negative-score change is $\Gamma = -0.00481$ cosine despite little R@10 source-contrast change. Fixed-positive change is -0.02106, switching gain $+0.00312$, and mean absolute relative drift 0.01527 (cross) versus 0.01618 (same). Triplets and seeds are averaged within queries, then queries equally; Appendix F.3 gives all sources and distributions.

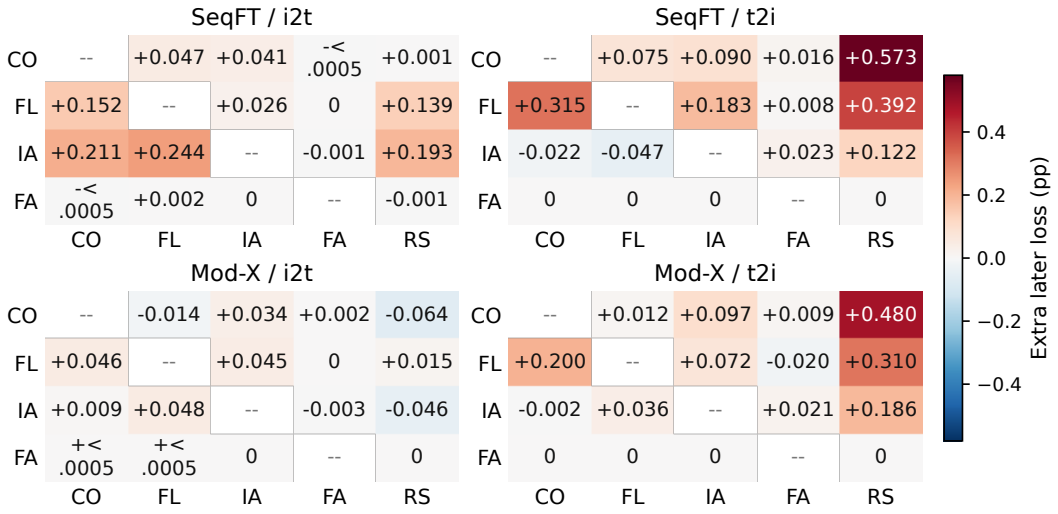

Figure 5: R@10 extra retention loss X_{FLC} (pp) for full Local plus 512 negatives. Rows: query domains; columns: added sources, coded CO/FL/IA/FA/RS in stream order. Positive/negative values indicate increasing/decreasing addition loss from post-task to final. Exact zero is 0; $\pm<.0005$ denotes signed nonzero magnitudes below 0.0005 pp; other values use three decimals. Dashes are N/A. The last task has no later update. These finite-pool expectations have no new intervals or realized query paths.

7 Discussion and Conclusion

Retrieval depth specifies the success being retained. R@1 requires an annotated positive first; R@10 accepts one within ten. Both measure retained access to annotated matches, without validating unannotated relevance or full-shortlist ordering.

Complementary cases address separate questions. For held-out SketchScene-to-COCO queries excluded from adaptation, frozen-score-constrained replacement at $N = 512$ supports i2t Frozen-versus-adapted reversals at R@5/R@10 (post-hoc 32-contrast family), but not R@1 or t2i. This contrasts with unchanged native five-domain ordering. Appendix G gives absolute values, partly supported original R@1 predictions and cross/sham precision.

Fixed-role mixtures and common-target adaptation address composition and continued adaptation separately (Appendices E.1 and H). All five historical starts improve over their own re-encoded starts; two seeds repeat only the short target run, not the historical trajectories.

Behavioral evidence and open mechanisms. Rank paths and equal-count controls reveal changes beyond candidate count. Small R@10 source-contrast changes coexist with score movement on the limited matched populations. Overfitting and altered capacity for later learning remain open explanations.

Scope and implementation provenance. The evidence concerns one task order, two executed configurations, benchmark-defined relevance and locally assembled splits with some training-text reuse. Component intervals condition on these models and galleries, without covering all shared-negative dependence. The numerical check does not certify full-gallery rank stability (Appendix I).

Conclusion. In the evaluated trajectories, the same later updates produce greater mean R@10 loss in Shared than Local. Some post-task common successes, including pre-existing successes, later fail only in Shared. Other query paths reduce the net contrast without necessarily restoring those successes. Task-local scores and net contrasts alone therefore do not identify which queries survive heterogeneous competition. Retention claims need a post-task reference, candidate gallery, retrieval depth, and paired query outcomes.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery and language polishing. All cited references were independently checked by the authors against the original sources. AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final manuscript, including its claims, citations, analyses, and conclusions.

REPRODUCIBILITY STATEMENT

We provide anonymized supplementary materials containing evaluation source code, reconstruction scripts, and saved numerical inputs for the reported results, figures, and tables. Appendix A and the experimental appendices detail the settings and evaluation protocols; the supplementary README explains the scope of reconstruction from saved outputs. Raw datasets and pretrained model weights are not redistributed.

REFERENCES

Xi Chen, Xu Chen, Xiangyang Jia, Wei Wang, Xu Zhang, and Zhenyuan Sun. DARAD: Dual adapters and ranking-aware distillation for continual remote sensing image-text retrieval. *arXiv preprint arXiv:2608.06059*, 2026. URL https://arxiv.org/abs/2608.06059.

Pinaki Nath Chowdhury, Aneeshan Sain, Ayan Kumar Bhunia, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. FS-COCO: Towards understanding of freehand sketches of common objects in context. In *European Conference on Computer Vision*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136680245.pdf.

Zhenyu Cui, Yuxin Peng, Xun Wang, Manyu Zhu, and Jiahuan Zhou. Continual vision-language retrieval via dynamic knowledge rectification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11704–11712, 2024. doi: 10.1609/aaai.v38i10.29054. URL https://ojs.aaai.org/index.php/AAAI/article/view/29054.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16712–16721, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Davari_Probing_Representation_Forgetting_in_Supervised_and_Unsupervised_Continual_Learning_CVPR_2022_paper.html.

Matthias De Lange, Gido M. van de Ven, and Tinne Tuytelaars. Continual evaluation for lifelong learning: Identifying the stability gap. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zy350cRstc6.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage 2006: Language Resources for Content-Based Image Retrieval, held in conjunction with LREC 2006*, 2006. URL https://thomas.deselaers.de/publications/papers/grubinger_lrec06.pdf.

Walid Krichene and Steffen Rendle. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020. URL https://research.google/pubs/on-sampled-metrics-for-item-recommendation/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. URL https://arxiv.org/abs/1405.0312.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-CLIP: Multimodal continual learning for vision-language model. In *International Conference on Learning Representations*,

2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/72fb9ab442fc60b7ae5d53bf6b478273-Abstract-Conference.html.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *arXiv preprint arXiv:1712.07835*, 2017. URL https://arxiv.org/abs/1712.07835.

Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, and Qi Tian. Continual vision-language representation learning with off-diagonal information. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26129–26149. PMLR, 2023. URL https://proceedings.mlr.press/v202/ni23c.html.

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. URL https://proceedings.mlr.press/v139/radford21a.html.

Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. URL https://arxiv.org/abs/1806.08317.

Canran Xiao, Tianxiang Xu, SiYuan Ma, Yiyang Jiang, Haoyu Gao, and Yuhan Wu. Reversible primitive–composition alignment for continual vision–language learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8f0d446441a938d9de420a8ab8d7fd36-Abstract-Conference.html.

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. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

zoheb. Sketch scene descriptions: zoheb/sketch-scene dataset card. Hugging Face dataset repository, n.d. URL https://huggingface.co/datasets/zoheb/sketch-scene. Accessed 2026-09-09; local cached revision 22ffd55109e12e1b82003a93e40fee0298e985a3. Card attributes FS-COCO; local train/test export is not an upstream official split.

A EVALUATION POPULATIONS, IMPLEMENTATION, AND INFERENCE

Reading guide. This appendix specifies the evaluation populations, executed configurations, and primary inference. Appendix B gives complete trajectories; Appendix C gives query paths and their denominators. Source and candidate controls appear in Appendices D–F. The separate SketchScene and common-target cases follow in Appendices G and H. Appendix I defines the numerical checks and their limits; Appendix J locates the complete companion records.

Table 4: Candidate designs and controls. All preserve their complete specified positives. "Fixed" means across compared checkpoints. Count equality is between conditions for a given query, not across domains. P: positives; C: common background.

Design	Background / count control	Identity / score anchor	Result type
Local / Shared	Full Local; L/S sizes differ	Fixed; no matching	Recorded ranks
Source subsets	Full Local; subset sizes differ	Fixed; no matching	Exact blocker counts
Small-budget C	Same sampled base; equal self/off additions	Fixed seeds; score-independent	Recorded ranks
FLC512	Full Local; 512 per external source	Fixed pools; no sampled identity set	Hypergeometric expectation
Four-arm replacement	Common $P + C$; equal size $\|L\| - 64$	Fixed; each learner's post-task scores, span $\leq .005$	Recorded ranks and scores
Uniform arrival	Full Local; fraction of external pool	Fixed pools; no score matching	Expected recall
Fixed-role mixture	Full Local; 1,024 negatives	Roles fixed at M3	PMF expectation
Sketch replacement	Separate queries; $N = 512$	Fixed; Frozen M0 summaries	Recorded case ranks

Tasks, checkpoints, and units. The fixed task order is COCO, Flickr30K, IAPR-TC12, Fashion-Gen, and RSICD. The eleven physical states comprise one common frozen CLIP ViT-B/32 initialization and five checkpoints for each of SeqFT and Mod-X. At every checkpoint, Local contains a query domain's canonical candidates; Shared contains the unique union of all five domains, including domains not yet trained. Both galleries remain fixed across checkpoints. The 60-cell Seen-Shared export instead grows with the seen domains and is descriptive: it combines model and gallery changes and does not replace this comparison.

Let $\tau(d)$ be domain d's task position and $R^g_{t,d}$ its score at M_t in environment $g \in \{L, S\}$. The task-aligned quantities are (with F/X defined only for $\tau(d) < 5$)

$$\begin{aligned} A^g_d &= R^g_{\tau(d),d} - R^g_{\tau(d)-1,d}, & F^g_d &= R^g_{\tau(d),d} - R^g_{5,d}, \\ Z^g_d &= R^g_{5,d} - R^g_{0,d}, & D_{t,d} &= R^L_{t,d} - R^S_{t,d}, \\ X_d &= F^S_d - F^L_d = D_{5,d} - D_{\tau(d),d}. \end{aligned} \tag{5}$$

Acquisition A and net change Z average all five domains equally; retention F and its environment difference X average the first four. RSICD has no subsequent update and is outside the retention estimand; its entry is N/A rather than a zero observation. Negative F means improvement after the task. In general $Z \neq A - F$: A starts at the pre-task model, whereas Z starts at M_0. The 220 model–domain–direction–environment cells each have R@1/5/10 and exact first-positive MRR. Recalls are percentages and their differences pp; $100\times$MRR is an index and its differences are index points.

Local data and identity. The evaluation uses local held-out exports, not a certification that every export equals an upstream official split. COCO has 82,783 training images and a 5,000-image val2014-derived test export; the original Karpathy split file is not bound. Flickr30K has 30,014 training and 1,000 test images from the available DKR annotations, rather than its compatibility training pool; its upstream Karpathy file is likewise not independently bound. IAPR uses TITLE and DESCRIPTION and a filename-sorted 15,000-image training partition: the rule is code-supported, but its historical invocation is missing. FashionGen uses 260,490 training and 32,528 validation images, with validation serving as local test and product IDs as native groups. Its available code defaults to `input_description` with a fallback, but the historical parameter invocation is not

certified. RSICD has 9,828 training and 1,093 test images; a Karpathy-style conversion does not establish independent underlying content.

Raw test exports have 5,000/1,000/4,999/32,528/1,093 image rows in task order. Native-group/exact-file exclusion removes one COCO and 3,202 FashionGen rows before canonicalization. Its closure has 732 contaminated groups (731 initial and one induced); it uses identity and file hashes, not caption or statistical component edges. Separately, raw tests contain 132 exact duplicate rows. These units differ, so 3,203 excluded rows do not explain every reduction in Table 5. All recorded image files were present. Perceptual duplicates, differently encoded copies, and semantic overlap are outside this exact-identity check.

Table 5: Canonical test populations. Positive multiplicities are minimum–maximum direct legal positives per query. Reused training texts count normalized test texts also occurring in training, not overlapping image identities.

Domain	Unique images	Unique texts	$\lvert P\rvert$ i2t	$\lvert P\rvert$ t2i	Reused train texts
COCO	4999	24825	4–6	1–8	977
Flickr30K	1000	4999	5–5	1–2	16
IAPR	4970	4266	1–2	1–10	10
FashionGen	29223	6763	1–2	1–16	212
RSICD	1093	3330	1–5	1–17	238

Canonical text uses Unicode NFKC, case folding, and whitespace collapse; the tokenizer representative is the UTF-8-minimum original string. Positives retain direct annotation edges and native grouping, including FashionGen product identity. Statistical components are resampling units, not arbitrary complete bipartite positive graphs. The export has 41,285 images, 44,183 texts, 67,861 direct positive edges, and 17,615 statistical components. The largest component has 70 nodes; none crosses domain identities. All ten domain/direction groups have 100% positive-invariant coverage between Local and Shared, so native and positive-invariant populations coincide here. This does not certify exhaustive semantic positives or CLIP pretraining disjointness. Repeated training captions are retained; image overlap does not follow from text reuse.

Evaluation restores the end token after truncation at 77 tokens. There are 614 truncated canonical texts and one tokenizer-collision group of two distinct texts. The canonical source counts are COCO 0, Flickr30K 1, IAPR 55, FashionGen 558 and RSICD 0 (Table 28). Candidate counts and positive multiplicities remain heterogeneous. Source comparisons describe these actual populations, not an isolated effect of dataset names.

A.1 Executed configurations and ranking

Both adapted trajectories use CLIP ViT-B/32, five epochs per task, batch size 128, AdamW with learning rate 10^{-6}, betas $(0.9, 0.999)$ and epsilon 10^{-8}. Task steps are 3,230/1,170/585/10,175/380. The different retained training entry points specify weight-decay defaults of 0.2 for SeqFT and 0.1 for Mod-X. The recorded commands do not explicitly set this parameter, and historical runtime dependency bytes remain unbound; these are retained defaults, not independently certified runtime values. The comparison concerns these executed configurations, not an isolated loss-term ablation. Available code resets the optimizer and cosine scheduler per task, uses 20% warmup (646/234/117/2,035/76 steps), accumulation one, and a seed-setting call of zero. These settings are code-supported rather than a recovered historical random-state log.

All model parameters train, with logit scale clamped to $[0, 4.6052]$. SeqFT disables LoRA and auxiliary loss. Mod-X has no first-task distillation; later tasks use the previous checkpoint as a frozen evaluation-mode teacher, without EMA, and coefficient 20 on sum-reduced bidirectional KL from cosine logits. Teacher-error handling replaces affected target rows by detached current predictions. The image pipeline uses bicubic resize, a 224-pixel center crop, RGB conversion and CLIP normalization. Text processing and its verification are documented in Appendix I. Float32 tensor storage does not certify historical GPU arithmetic, TF32 or cuDNN settings.

Each checkpoint encodes both queries and candidates; no stale index is paired with an updated query encoder. The main caches use FP16 model encoding followed by float32 normalization and stored

representations; ranking uses normalized float32 cosine scores. AMP and TF32 are disabled on the recorded evaluation path. Float32 stored arrays do not imply all model arithmetic was FP32. Budget controls reuse scores for common items. With legal positives P_q, the best-positive rank is

$$r_q^g = 1 + \#\{c \in G_g \setminus P_q : s(q,c) > \max_{p \in P_q} s(q,p)\}. \tag{6}$$

Tied negatives do not outrank a positive; no epsilon alters this strict ordering. A separate wrong-top-item provenance convention breaks source ties by query-domain membership, then canonical ID. It does not change rank. There is one task order and one trajectory per adapted configuration; the common Frozen state is not two independent runs. Local stage results and Frozen metadata were available during exploration before the unified protocol was locked. The study is not a fresh preregistered test, and unfinished methods or diagnostics do not count as results.

A.2 Statistical family and numerical scope

The primary family contains eight X contrasts: two configurations, two directions, and R@1/R@10. The original analysis draws 10,000 multinomial weight vectors over 17,615 global components (PCG64 seed 2026091701), sharing weights across checkpoints, directions, metrics and domains. It recomputes each domain's weighted query denominator before the equal four-domain mean. For centered, standard-error-scaled draws, the higher 95th percentile of the maximum absolute deviation is $c = 2.7268582152$; intervals are $\widehat{X}_j \pm c\widehat{\mathrm{se}}_j$. A zero domain denominator or degenerate standard error would invalidate certification, not silently remove a draw; neither occurred. The complete primary table appears once in the main text. All four R@10 intervals exclude zero positively; all four R@1 intervals include zero. This is neither R@1 equivalence nor a test that the metrics differ. Secondary results do not enter that family.

Table 6: Complete descriptive leave-one-domain-out sensitivity. Recall entries are pp; MRR entries are $100\times$MRR index points. Overlapping omission averages are not independent replications or additional significance tests.

Model	Dir.	Metric	All four	Omit COCO	Omit Flickr	Omit IAPR	Omit Fashion
SeqFT	i2t	R@1	0.551	0.681	0.068	0.728	0.727
SeqFT	i2t	R@5	0.984	1.198	0.545	0.889	1.303
SeqFT	i2t	R@10	1.145	1.420	0.893	0.762	1.505
SeqFT	i2t	MRR	0.728	0.875	0.266	0.810	0.962
SeqFT	t2i	R@1	0.197	0.272	0.030	0.224	0.263
SeqFT	t2i	R@5	0.578	0.656	0.270	0.614	0.771
SeqFT	t2i	R@10	0.978	1.024	0.396	1.186	1.303
SeqFT	t2i	MRR	0.380	0.463	0.106	0.444	0.507
Mod-X	i2t	R@1	0.128	0.104	0.104	0.137	0.167
Mod-X	i2t	R@5	0.141	0.128	0.122	0.142	0.174
Mod-X	i2t	R@10	0.514	0.586	0.319	0.471	0.681
Mod-X	i2t	MRR	0.279	0.271	0.174	0.307	0.366
Mod-X	t2i	R@1	0.172	0.173	0.056	0.230	0.230
Mod-X	t2i	R@5	0.564	0.539	0.326	0.635	0.757
Mod-X	t2i	R@10	0.818	0.763	0.437	0.981	1.090
Mod-X	t2i	MRR	0.336	0.339	0.141	0.416	0.448

Component uncertainty is conditional on these trajectories and candidate pools. It does not cover training randomness, sampled task orders/domains, all dependence through shared negatives, or research and model selection. All leave-one-domain-out R@10 contrasts remain positive; their magnitudes vary, especially when Flickr30K is omitted (Table 6).

Saved-data comparisons check schemas, row order, IDs, integers, booleans, text and missingness exactly. Float tolerances are 10^{-10} absolute with zero relative tolerance and 5×10^{-13} for probabilities. The original eight intervals reuse archived draws. `code/REPRODUCTION_SCOPE.csv` separates earlier reconstruction from current checks and states each module's limits. Neither is training or raw-image encoding.

B COMPLETE ABSOLUTE TRAJECTORIES AND TASK-ALIGNED ENDPOINTS

The plots below retain every domain, checkpoint, direction and metric. Both learner trajectories start from the common Frozen model. Curves use per-domain query means; summary quantities retain their equal-domain weights. Exact pre-task, post-task and final R@1/R@10 values are supplied in `manuscript/tables/compact_endpoints_r1.tex` and `manuscript/tables/compact_endpoints_r10.tex`, indexed through `extended_results/README.md`. The same index locates all four-metric $A/F/Z/X$ and initial/net tables, including undefined last-task retention.

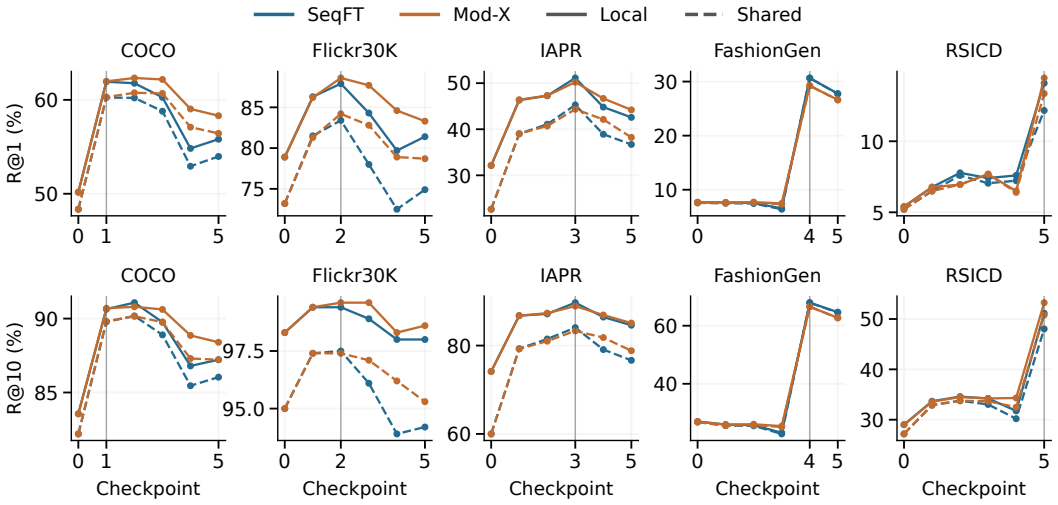

Figure 6: Complete i2t R@1 and R@10 trajectories (%). Local and Shared are fixed at every checkpoint; the query domain's task separates acquisition from subsequent retention.

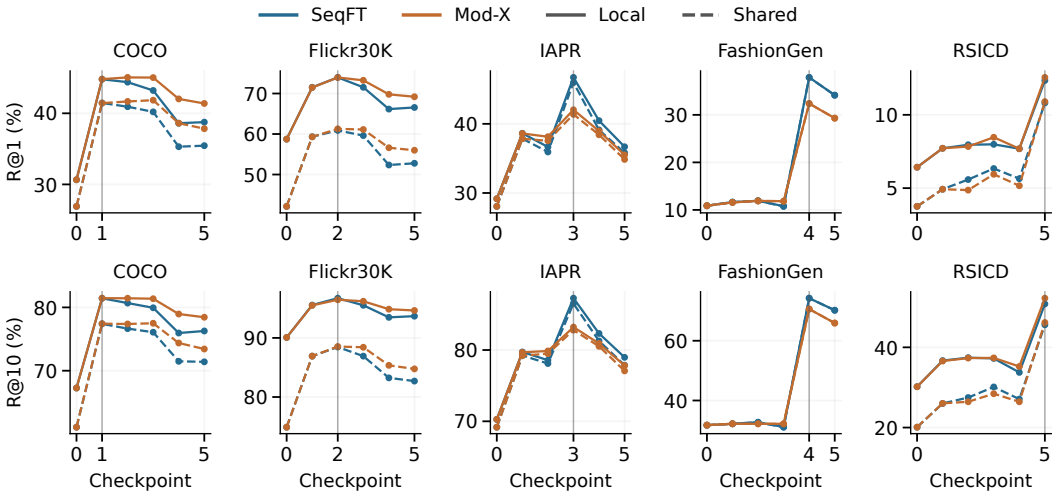

Figure 7: Complete t2i R@1 and R@10 trajectories (%) with the same fixed environments and task references. Domain populations differ in size.

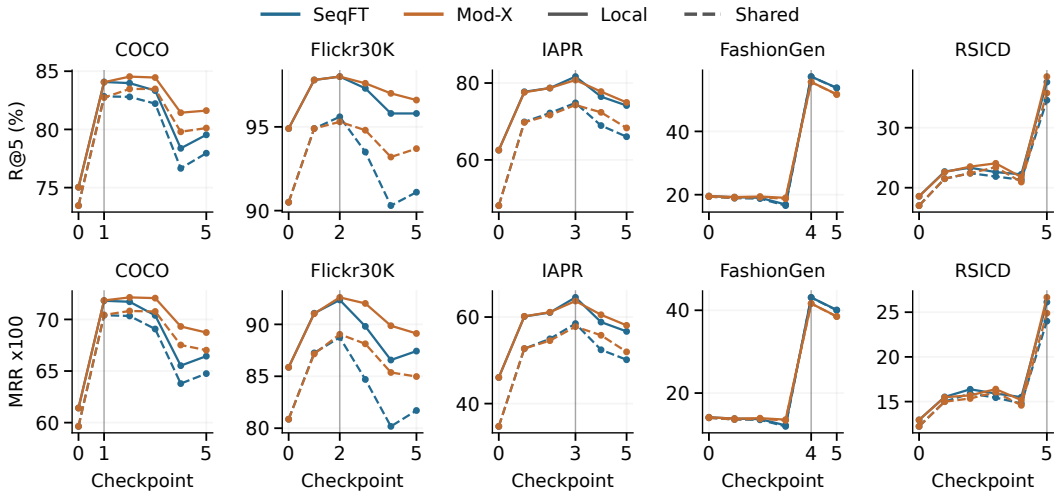

Figure 8: Complete descriptive i2t R@5 (%) and MRR trajectories. The latter are $100\times$MRR index values, not percentages of successful queries.

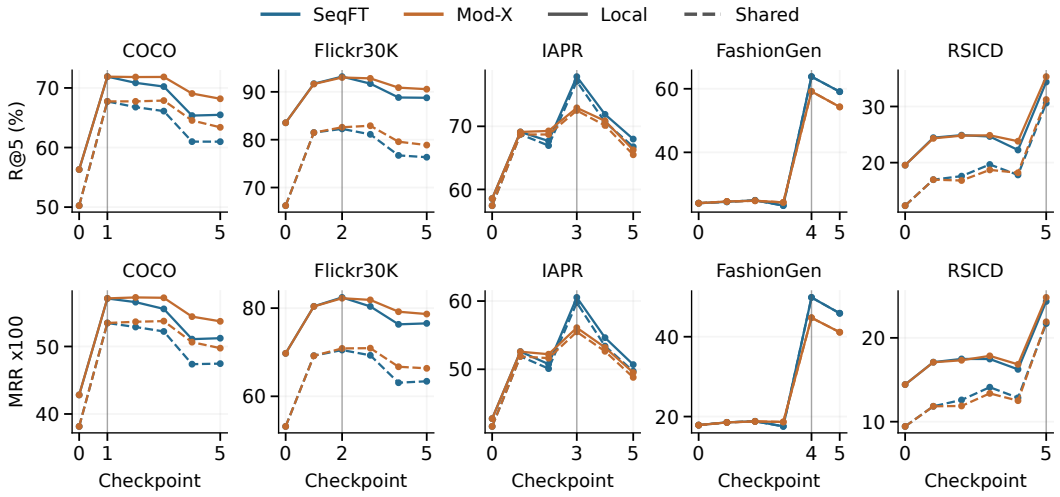

Figure 9: Complete descriptive t2i R@5 and MRR trajectories, with the metric-specific units of Figure 8.

C QUERY PATHS AND INTERNAL MODEL SELECTION

At threshold K, write $C = (1, 1)$ for success in both environments, $V = (1, 0)$ for Local-only success, and $W = (0, 0)$ for failure in both. Shared contains Local with unchanged positives, so Shared-only success is absent. Actual post-task-to-final joints give

$$X = 100\big[\Pr(C \to V) + \Pr(W \to V) - \Pr(V \to C) - \Pr(V \to W)\big]. \tag{7}$$

The reported contributions use original-domain query denominators, multiply by 100, then average four domains. They are not cohort-conditional rates. In particular, $V \to W$ shrinks the environment gap through loss of Local success, not Shared recovery.

Table 7: All eight four-flow decompositions in original-population pp. The final three columns partition $C \to V$ by pre-task state: "New both" is WCV, "New to S" is VCV, and "Already both" is CCV. Rounding can leave a final-digit sum difference.

Model	Dir.	K	CV	WV	VC	VW	X	New both	New to S	Already both
SeqFT	i2t	R@1	2.008	0.403	0.843	1.017	0.551	0.126	0.297	1.585
SeqFT	i2t	R@10	1.935	0.215	0.551	0.455	1.145	0.099	0.306	1.530
SeqFT	t2i	R@1	1.766	0.748	0.785	1.532	0.197	0.248	0.226	1.292
SeqFT	t2i	R@10	2.308	0.329	0.661	0.999	0.978	0.272	0.524	1.512
Mod-X	i2t	R@1	1.403	0.411	0.883	0.804	0.128	0.099	0.341	0.963
Mod-X	i2t	R@10	1.355	0.178	0.599	0.421	0.514	0.070	0.161	1.124
Mod-X	t2i	R@1	1.486	0.624	0.757	1.181	0.172	0.194	0.170	1.122
Mod-X	t2i	R@10	1.824	0.355	0.552	0.809	0.818	0.240	0.405	1.179

C.1 FINAL RETENTION OF POST-TASK COMMON SUCCESSES

For each domain, the post-C cohort contains queries successful in both galleries immediately after its task. At final evaluation, state C retains success in both galleries, V retains it only in Local, and W loses it in both. The conditional Local–Shared retention gap is $100n_{CV}/n_{\text{post}C}$ pp; among the cohort's Shared failures, the percentage still successful in Local is $100n_{CV}/(n_{CV} + n_{CW})$. These are domain-specific ratios before four-domain equal averaging. The conditional Shared-failure fraction cannot be reconstructed by dividing displayed macro means; the retention gap equals the difference of macro Local and Shared rates before rounding. They describe final-checkpoint outcomes, not success at every intervening checkpoint, and have no new confidence intervals.

Table 8 gives the R@1 counterpart to the main-text R@10 Table 3. Table 9 shows all sixteen R@10 domain rows, including the cohort counts and the denominator of the Local-only share among Shared failures. The full eight macro rows and all 32 R@1/R@10 domain rows are indexed in `figure_data/b1_posttask_cohort/README.md` and reconstructed from the archived 27-state joint counts.

Table 8: R@1 final retention of the post-task common-success cohort, averaged equally over four domains as in Table 3. Post-C uses all domain queries as its denominator; Local, Shared and Gap use the post-C cohort. Local-only is $100n_{CV}/(n_{CV} + n_{CW})$, conditional on a final Shared failure within that cohort. Each ratio is computed within domain before averaging.

Model	Dir.	All queries	Within post-C cohort			Shared failures within post-C
		Post-C (%)	Local (%)	Shared (%)	Gap (pp)	Local-only (%)
SeqFT	$i \to t$	54.89	80.46	77.19	3.27	14.83
SeqFT	$t \to i$	46.55	77.02	73.82	3.19	14.11
Mod-X	$i \to t$	54.52	83.43	81.02	2.41	13.63
Mod-X	$t \to i$	44.12	80.44	77.71	2.73	16.18

Table 9: R@10 cohort counts and Shared failures for all sixteen eligible domain–configuration–direction groups. $n_{\text{post}C}$ counts queries successful in both galleries immediately after the task; n_{CV} and n_{CW} count those ending in Local-only success and failure in both galleries. The last column divides n_{CV} by $n_{CV} + n_{CW}$, the cohort's final Shared failures. The two FashionGen t2i zeros have positive denominators. RSICD has no later update and is ineligible. The four within-domain percentages, rather than pooled counts, produce each summary in Table 3.

Model	Dir.	Domain	$n_{\text{post}C}$	n_{CV}	n_{CW}	$\frac{100n_{CV}}{n_{CV}+n_{CW}}$ (%)
SeqFT	$i \to t$	COCO	4488	39	292	11.78
SeqFT	$i \to t$	Flickr30K	975	22	15	59.46
SeqFT	$i \to t$	IAPR TC-12	4178	232	310	42.80
SeqFT	$i \to t$	FashionGen	19840	27	1719	1.55
SeqFT	$t \to i$	COCO	19215	666	1645	28.82
SeqFT	$t \to i$	Flickr30K	4422	291	77	79.08
SeqFT	$t \to i$	IAPR TC-12	3691	31	425	6.80
SeqFT	$t \to i$	FashionGen	5021	0	539	0.00
Mod-X	$i \to t$	COCO	4490	39	218	15.18
Mod-X	$i \to t$	Flickr30K	974	17	8	68.00
Mod-X	$i \to t$	IAPR TC-12	4141	145	236	38.06
Mod-X	$i \to t$	FashionGen	19441	7	1872	0.37
Mod-X	$t \to i$	COCO	19220	610	1090	35.88
Mod-X	$t \to i$	Flickr30K	4426	215	38	84.98
Mod-X	$t \to i$	IAPR TC-12	3532	23	338	6.37
Mod-X	$t \to i$	FashionGen	4775	0	602	0.00

For SeqFT COCO t2i R@10, 4,053 of 24,825 queries newly acquire Local success. Of these, 215 follow WCV: 5.305% of the cohort, but 0.866 pp of the original population. The cohort has 404 final-V queries, so that entire state cannot be called WCV. FashionGen t2i WCV is zero for both learners at R@1 and R@10. The archive includes all 27 three-time-point paths in 40 groups (1,080 rows), all pairwise transitions and the 40-row cohort table. The saved-count reconstruction covers S0 R@1/R@10 paths, not every original A–D rank-bin analysis. Table 10 partitions all $C \to V$ mass by the post-to-final external-blocker change. It aggregates the existing 9,720-row path/stratum table; all three columns sum to the original $C \to V$ contribution without changing its population or adding inference.

Table 10: Complete $C \to V$ partition by external-blocker count change, in original-population pp averaged equally over four domains. The last three columns sum to Total before rounding. These are contributions, not conditional query fractions.

Configuration	Dir.	K	Total $C \to V$	$\Delta b_E > 0$	$\Delta b_E = 0$	$\Delta b_E < 0$
SeqFT	i2t	1	2.0080	2.0080	0.0000	0.0000
SeqFT	i2t	10	1.9351	1.9167	0.0076	0.0109
SeqFT	t2i	1	1.7661	1.7661	0.0000	0.0000
SeqFT	t2i	10	2.3077	2.2716	0.0330	0.0030
Mod-X	i2t	1	1.4029	1.4029	0.0000	0.0000
Mod-X	i2t	10	1.3554	1.3045	0.0359	0.0150
Mod-X	t2i	1	1.4862	1.4862	0.0000	0.0000
Mod-X	t2i	10	1.8243	1.7704	0.0451	0.0089

Most R@10 $C \to V$ mass has more external blockers. Smaller nonzero portions have unchanged or fewer external blockers; all such queries have more Local blockers. At R@1 these latter routes are impossible: Local must keep zero blockers while Shared acquires at least one. For a Local-successful query, $K - r_L$ is the total number of external blockers it can tolerate; if already Shared-successful, the additional allowance is $K - r_S$. These exact rank relations identify observed routes, not the cause of the different interval outcomes at R@1 and R@10. Unchanged counts need not

mean unchanged identities or scores. The archive retains the full 3×3 Local/external change partition and every domain denominator.

C.2 Success-history contributions and conditional failure fractions

The pre-existing cohort contains queries with C at both pre-task and post-task evaluation. The newly Shared-successful cohort has pre-task state W or V and post-task state C. For each cohort, the denominator counts all its queries and the numerator counts final state V: Shared success is lost while Local success remains. Final state W is excluded from the numerator. Thus these fractions describe one particular post-task failure path, conditional on having succeeded in both environments immediately after the task; they are not loss rates over all pre-task successes.

Table 11 is a descriptive reconstruction from the existing 27-cell joint tables, with no new inferential family. All 32 defined domain–configuration–direction–threshold groups sum to their original query populations. Every newly Shared-successful denominator agrees with the saved cohort table. Small groups, including three and six Flickr i2t queries at R@10, are retained.

Table 11: Within-cohort $C \to V$ fractions. "Existing" means pre/post C/C; "New Shared" means pre W or V, post C. Each count is final-V queries divided by its cohort size; rates are percentages. Domain abbreviations follow stream order. The last task has no later update and is omitted.

Configuration	Direction	Domain	Pre-existing success		New Shared success	
			n_V/n_C	Rate (%)	n_V/n_C	Rate (%)
R@1						
SeqFT	i2t	COCO	30/2,162	1.39	20/850	2.35
SeqFT	i2t	Flickr30K	36/784	4.59	5/50	10.00
SeqFT	i2t	IAPR	106/1,799	5.89	38/451	8.43
SeqFT	i2t	FashionGen	2/1,431	0.14	8/7,522	0.11
SeqFT	t2i	COCO	109/5,825	1.87	200/4,465	4.48
SeqFT	t2i	Flickr30K	220/2,782	7.91	51/262	19.47
SeqFT	t2i	IAPR	14/1,405	1.00	3/556	0.54
SeqFT	t2i	FashionGen	0/538	0.00	0/2,024	0.00
Mod-X	i2t	COCO	18/2,163	0.83	21/851	2.47
Mod-X	i2t	Flickr30K	16/801	2.00	8/41	19.51
Mod-X	i2t	IAPR	94/1,944	4.84	26/260	10.00
Mod-X	i2t	FashionGen	0/1,786	0.00	5/6,754	0.07
Mod-X	t2i	COCO	96/5,828	1.65	205/4,458	4.60
Mod-X	t2i	Flickr30K	191/2,891	6.61	28/172	16.28
Mod-X	t2i	IAPR	12/1,548	0.78	3/217	1.38
Mod-X	t2i	FashionGen	0/580	0.00	0/1,610	0.00
R@10						
SeqFT	i2t	COCO	27/3,975	0.68	12/513	2.34
SeqFT	i2t	Flickr30K	20/969	2.06	2/6	33.33
SeqFT	i2t	IAPR	177/3,929	4.50	55/249	22.09
SeqFT	i2t	FashionGen	5/6,280	0.08	22/13,560	0.16
SeqFT	t2i	COCO	287/14,651	1.96	379/4,564	8.30
SeqFT	t2i	Flickr30K	220/4,273	5.15	71/149	47.65
SeqFT	t2i	IAPR	21/3,286	0.64	10/405	2.47
SeqFT	t2i	FashionGen	0/1,934	0.00	0/3,087	0.00
Mod-X	i2t	COCO	21/3,975	0.53	18/515	3.50
Mod-X	i2t	Flickr30K	16/971	1.65	1/3	33.33
Mod-X	i2t	IAPR	123/4,003	3.07	22/138	15.94
Mod-X	i2t	FashionGen	0/7,061	0.00	7/12,380	0.06
Mod-X	t2i	COCO	236/14,652	1.61	374/4,568	8.19
Mod-X	t2i	Flickr30K	166/4,320	3.84	49/106	46.23
Mod-X	t2i	IAPR	19/3,370	0.56	4/162	2.47
Mod-X	t2i	FashionGen	0/2,010	0.00	0/2,765	0.00

At R@10 the new cohort's fraction is higher in 14 of 16 cells and equal in the two FashionGen t2i cells, where both fractions are zero. At R@1 it is higher in 12, equal in two and lower in two.

Meanwhile, pre-existing successes contribute most of the $C \to V$ original-population mass in all eight four-domain summaries in Table 7. This aggregate is domain-equal, rather than a pooled count ratio. The two observations coexist because loss contribution depends on cohort size as well as the within-cohort fraction; they do not establish an intrinsic effect of success age.

C.3 Selection agreement and signed transfer differences

Five fixed global-component folds are assigned round-robin after ordering components by the SHA256 of `wace-v1-selection-20260917` and identity. Four folds select a model; the held-out fold evaluates it. Selection averages queries within domains, then five domains equally; held-out aggregation uses query counts within domains before that equal-domain mean. This is component-fold validation, not leave-one-domain-out. No required domain fold is empty. The main pool is SeqFT/Mod-X M_5; the auxiliary pool adds Frozen. Ties favor SeqFT, then Mod-X, then Frozen. For route $g \to h$, the reported difference is held-out performance of the model selected in h minus that of the model selected in g, both selected using the other folds. It can be negative: neither selection is a target-test oracle.

Table 12: Complete internal selection results. Recall differences are pp; MRR differences are index points. Changed-fold counts use metric order R@1/R@5/R@10/MRR, each out of five. Both pools and both routes are retained.

Pool	Dir.	Route	R@1	R@5	R@10	MRR×100	Changed folds
Adapted	i2t	L to S	0.000	0.000	0.000	0.000	0/0/0/0
Adapted	i2t	S to L	0.000	0.000	0.000	0.000	0/0/0/0
Adapted	t2i	L to S	-0.159	-0.107	-0.111	-0.144	2/1/1/2
Adapted	t2i	S to L	0.145	0.052	0.113	0.159	2/1/1/2
+ Frozen	i2t	L to S	0.000	0.000	0.000	0.000	0/0/0/0
+ Frozen	i2t	S to L	0.000	0.000	0.000	0.000	0/0/0/0
+ Frozen	t2i	L to S	-0.159	-0.107	-0.111	-0.144	2/1/1/2
+ Frozen	t2i	S to L	0.145	0.052	0.113	0.159	2/1/1/2

All i2t selections agree. The t2i differences are small and signed, with one or two changed folds. Native adapted-endpoint comparisons show no strict Local/Shared sign reversal, either within a domain or in the five-domain mean. Thus larger X does not imply widespread reversal of endpoint selection. Full fold, component/domain-fold and model-comparison records are indexed; these internal comparisons are descriptive, not deployment regret guarantees.

D Source effects across backgrounds and training updates

Each source-subset environment retains complete Local and adds one of the 16 subsets of the other four sources. Unique unions use exact disjoint provenance-bitmask buckets, not sums of overlapping membership counts. The archive contains all 1,760 environments, 7,040 absolute metric rows, 14,080 all-background marginals and eight full eleven-state R@1/R@10 maps. Single-add loss is $R(L) - R(L \cup u)$; last-removal recovery is $R(S \setminus u) - R(S)$. The main figure and the two figures below cover all query/source pairs in their stated conditions. Natural pool sizes differ; these are not effects per negative or difficulty-matched source comparisons. Diagonals are inapplicable, rounded 0.0 need not be exact zero, and marginal effects are not an additive attribution of the full environment gap.

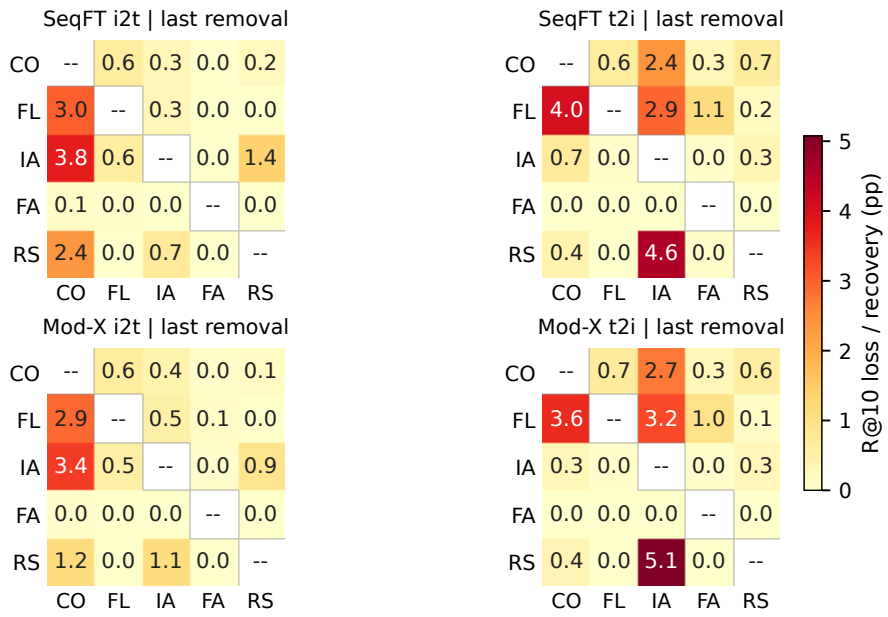

Figure 10: Complete final-checkpoint R@10 last-removal recovery in pp, for both learners and directions. Rows are query domains; columns are removed sources. This full-Shared background differs from single addition to Local.

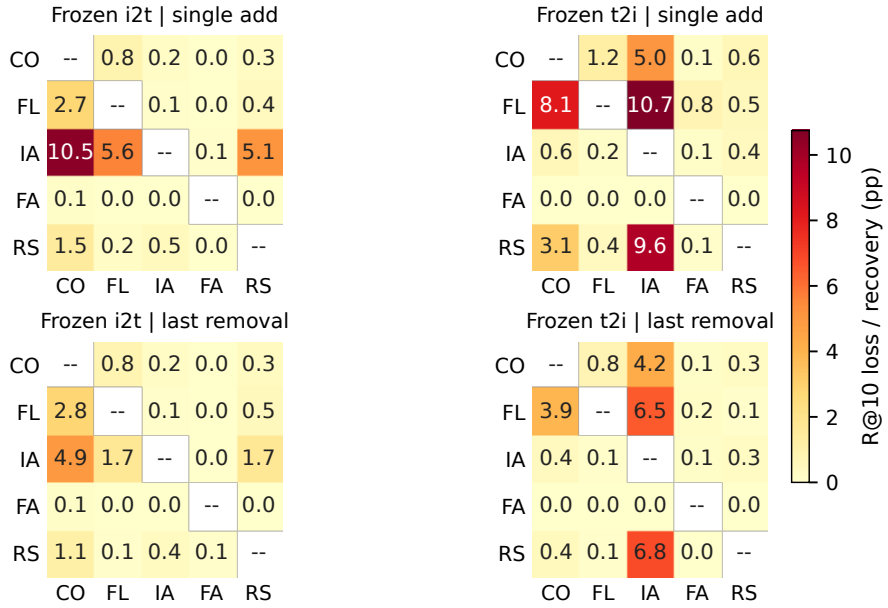

Figure 11: The common Frozen R@10 reference: all query/source pairs, both directions, and both single-add loss and last-removal recovery (pp). It is one physical initialization, not a separate baseline per learner.

D.1 Background dependence and event-aligned changes

The number of newly failed queries depends on background in 936 of 1,320 recall source roles. At SeqFT M_5 on Flickr30K t2i, COCO's R@1 marginal ranges from 6.2613 to 9.3019 pp and IAPR's R@10 marginal from 2.9406 to 5.3211 pp. Joint-only threshold failure is absent at R@1 by

the rank rule, but occurs in 81/110 R@5 and 79/110 R@10 cells: individually insufficient blocker sets can jointly cross K. This is ranking-threshold dependence, not identification of a nonlinear representation mechanism.

For single-add loss $L_t(d \leftarrow u)$, define $J_j(d \leftarrow u) = L_j(d \leftarrow u) - L_{j-1}(d \leftarrow u)$. Positive J means the same source becomes more damaging after update j. All 1,600 event rows retain 20 directed pairs, two learners, two directions, five updates and four metrics. Query roles are old ($\tau(d) < j$, 640 rows), current (320), and future (640). Source roles distinguish just-trained, previously trained and future sources. All signs remain; rows share queries and checkpoints and are not independent runs. Summing five updates recovers $L_5 - L_0$ for each of 320 fixed pair/model/direction/metric groups (maximum recorded residual 1.11×10^{-16}).

Table 13: Complete old-query R@10 role summaries. Mean J is pp; records are equally weighted pair/update observations, not queries or runs. Positive, negative and exact-zero counts are retained.

Model	Dir.	Candidate-source role	Mean J	Records	$+/-/0$
SeqFT	i2t	Newly trained	-0.0789	10	2/5/3
SeqFT	i2t	Previously trained	+0.2659	20	13/4/3
SeqFT	i2t	Future	+0.1524	10	6/2/2
SeqFT	t2i	Newly trained	+0.1035	10	6/3/1
SeqFT	t2i	Previously trained	+0.2033	20	12/5/3
SeqFT	t2i	Future	+0.1099	10	8/2/0
Mod-X	i2t	Newly trained	-0.0260	10	4/3/3
Mod-X	i2t	Previously trained	+0.1482	20	10/7/3
Mod-X	i2t	Future	-0.0161	10	2/5/3
Mod-X	t2i	Newly trained	+0.1417	10	5/4/1
Mod-X	t2i	Previously trained	+0.1329	20	12/4/4
Mod-X	t2i	Future	-0.0019	10	4/6/0

For old i2t queries the just-trained source's mean R@10 J is negative for both configurations; t2i means are positive. During the FashionGen update, old Flickr30K t2i R@10 loss increases for all four sources: COCO/IAPR/FashionGen/ RSICD changes are 0.9802/0.7001/0.8402/0.2601 pp for SeqFT and 0.6001/0.7802/0.4801/0.0400 pp for Mod-X. Conversely, during the IAPR update, SeqFT Flickr30K t2i loss from IAPR decreases from 6.7814 to 5.6411 pp at R@1 and 4.4009 to 4.0208 pp at R@10. Competition therefore need not track only the just-trained domain or change in one direction. Full 128-role and 1,600-event tables remain indexed. These observed updates do not identify training causality, semantic mechanisms, or loss of plasticity.

Joint-source failure has a query denominator. At R@10, the 110 condition-specific affected fractions have a median of 0.080% and range from 0 to 1.680%. This is an unweighted distribution over related conditions, not a pooled-query prevalence or a set of independent replicates. Table 14 retains every domain and direction at Frozen and the two final models, including zeros. Frozen has the largest observed fraction; adaptation is not required for this phenomenon. The complete 110-row table is included in the review materials, with the original denominator and the structural R@1 zeros kept separate.

Table 14: R@10 failures requiring external sources jointly, when no single external source causes failure alone. Entries are query count / original domain–direction population (%), not conditional cohort rates or loss differences.

Query domain	Direction	Frozen	SeqFT M_5	Mod-X M_5
COCO	i2t	5/4,999 (0.100%)	5/4,999 (0.100%)	4/4,999 (0.080%)
COCO	t2i	138/24,825 (0.556%)	106/24,825 (0.427%)	132/24,825 (0.532%)
Flickr30K	i2t	4/1,000 (0.400%)	2/1,000 (0.200%)	3/1,000 (0.300%)
Flickr30K	t2i	84/4,999 (1.680%)	63/4,999 (1.260%)	59/4,999 (1.180%)
IAPR-TC12	i2t	55/4,970 (1.107%)	27/4,970 (0.543%)	24/4,970 (0.483%)
IAPR-TC12	t2i	2/4,266 (0.047%)	4/4,266 (0.094%)	2/4,266 (0.047%)
FashionGen	i2t	3/29,223 (0.010%)	1/29,223 (0.003%)	0/29,223 (0.000%)
FashionGen	t2i	0/6,763 (0.000%)	0/6,763 (0.000%)	0/6,763 (0.000%)
RSICD	i2t	2/1,093 (0.183%)	2/1,093 (0.183%)	0/1,093 (0.000%)
RSICD	t2i	10/3,330 (0.300%)	9/3,330 (0.270%)	6/3,330 (0.180%)

E CANDIDATE BUDGETS, ARRIVAL, AND FIXED-ROLE MIXTURES

Score-independent equal-budget controls. Every condition retains all positives, B same-domain base negatives and B additional negatives, for $B = 128$ or 64. Total size is $|P_q| + 2B$: equal across conditions for each query, but potentially different across queries. Self-add is disjoint from the base; external additions exclude query-domain members; every negative pool excludes global positives. Three seeds use canonical IDs, a hashed namespace and PCG sampling without scores or Frozen-hardness gates. The smaller budget is a nested prefix of the larger. There are 300 manifests, 1,860 condition cells over 31 domain/checkpoint roles, and 372 base-only cells.

The contrast is other-source minus self-add performance. Positive means the external negatives are less damaging than the same number of additional own-source negatives, not that adding them improves on the base. Every R@1/MRR contrast is positive, whereas R@5/R@10 include negative and zero contrasts (Table 15). For Frozen IAPR i2t, RSICD's mean R@10 contrast is $-1.831/-1.348$ pp at budgets 128/64; Flickr30K changes from $+0.322$ to -0.074 pp. The ranking of sources depends on the budget.

Table 15: Complete non-diagonal contrast signs. Each row has 744 dependent source/seed/state condition records. Exact zeros are not rounded sign bins.

Base/add budget	Metric	Negative	Zero	Positive
128	R@1	0	0	744
128	R@5	30	72	642
128	R@10	60	75	609
128	MRR	0	0	744
64	R@1	0	0	744
64	R@5	40	95	609
64	R@10	83	213	448
64	MRR	0	0	744

These small galleries can be nearly saturated at R@10. Among 930 controlled condition cells per budget (including self-add), 587 at $B = 128$ and 701 at $B = 64$ have R@10 at least 99%; 167 and 331 respectively equal 100%. The denominator is 930 absolute cells, not 744 non-diagonal contrasts. Saturation limits this control's discrimination and does not establish equivalence. Table 16 gives both absolute counts. The archive retains all 7,440 metric contrast rows, absolute cells and task-aligned results; the present signs are descriptive.

Table 16: Small-budget R@10 saturation. Only controlled cells are counted; base-only rows are excluded. Shared queries, states and seeds make these dependent condition records, not independent experiments.

Base/add negatives	Controlled cells	R@10 $\geq 99\%$	R@10 $= 100\%$
128/128	930	587	167
64/64	930	701	331

For task-aligned C, $\delta_t = R_{\text{off},t} - R_{\text{self},t}$ and $X_C = \delta_{\text{post}} - \delta_{\text{final}} = F_{\text{off}} - F_{\text{self}}$. The complete 1,536-row table retains both budgets, three seeds and four metrics. At budget 128, four-source/three-seed mean R@10 X_C ranges from -0.3713 to $+0.0333$ pp across sixteen learner–domain–direction groups. C's positive extra-loss sign is opposite to matched Θ in Appendix F.2. Absolute C entries are fractions; Recall differences are pp and MRR$\times$100 differences are index points.

Uniform arrival while retaining complete Local. The separate arrival calculation adds $m = \lfloor \lambda N_{\text{ext}} \rfloor$ unique external candidates uniformly without replacement, with $\lambda \in \{0, .01, .05, .10, .25, .50, 1\}$. For Local blocker count b_L and B_{ext} blockers in the external pool,

$$H \sim \text{Hypergeom}(N_{\text{ext}}, B_{\text{ext}}, m), \qquad \mathbb{E}[\text{R@}\,K_q] = \Pr(H \leq K - 1 - b_L). \tag{8}$$

The per-query R@K here is a zero–one indicator; displayed recall is 100 times its query-mean expectation. The probability is zero if Local already fails at K. Recall averages per-query probabilities, not a CDF of average blockers. The 770 expected environments yield 2,310 R@1/5/10 rows; no expected MRR is supplied. Endpoints recover Local/Shared exactly. At fixed state expected recall cannot increase with arrival; differences across states need not be monotone.

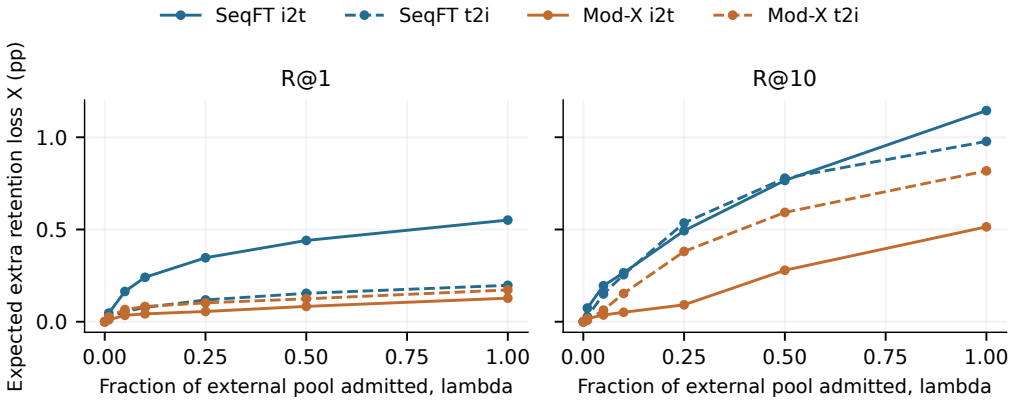

Figure 12: Expected extra retention loss at all seven prescribed arrival fractions. All eight macro curves increase in these data; monotonicity of a difference of losses is not required by the sampling rule. Only the original endpoint family has simultaneous intervals.

External-pool coverage differs from combined-gallery exposure:

$$\rho_{\text{all}} = \frac{m}{|G_L| + m}, \qquad \rho_{\text{neg},q} = \frac{m}{|G_L| - |P_q| + m}. \tag{9}$$

The former includes positives in its denominator; neither generally equals λ. Complete final R@1/R@10 absolute curves are retained in `manuscript/figures/wace/dose_absolute_r1.pdf` and `manuscript/figures/wace/dose_absolute_r10.pdf`. `extended_results/README.md` also indexes all checkpoint curves, task-aligned means and the 2,170-row expected-exposure file. These are finite-pool expectations, not realized query counts or joint paths such as $C \to V$.

E.1 FIXED-ROLE MIXTURES

These calculations keep complete Local and add exactly 1,024 negatives for COCO/Flickr30K/IAPR queries, both directions, and Frozen plus SeqFT/Mod-X M_3/M_5. Roles are fixed at M_3: excluding the query domain, two sources among the first three tasks are "seen," and FashionGen/RSICD are "future." Those labels remain fixed at M_0 and M_5; future-at-M_3 does not mean untrained at M_5. At future shares $\pi = 0, .25, .5, .75, 1$, each seen source gets 512/384/256/128/0 negatives and each future source 0/128/256/384/512. Capacity checks precede calculation. Convolving source hypergeometric blocker PMFs yields 150 environments and 450 expected R@1/5/10 entries. Terms 0–9 are retained without renormalizing truncated mass. There is no expected MRR or realized joint-transition calculation.

Define $\Psi_t = \mathbb{E}[R_t(\pi = 1)] - \mathbb{E}[R_t(\pi = 0)]$: positive values favor the future-only mixture at this fixed budget and role assignment. The complete 36-combination table (pp of expected recall), including every sign and threshold, is retained in `manuscript/tables/wace/s3_interactions.tex`. At R@10, t2i $\Psi_5 - \Psi_3$ is negative on COCO and IAPR but positive on Flickr30K; i2t changes are all small and positive. Future sources are therefore not uniformly harder in these comparisons. These results are descriptive and add no inferential family; the design does not isolate an effect of prior training or source labels.

All six exposure groups are retained in `manuscript/tables/wace/s3_exposure.tex`. They are fractions in $[0, 1]$; negative-only exposure averages query-specific ratios rather than using an average positive count. The original mixture field ρ denotes ρ_{all}. The local metadata derivation preserves it and adds $\rho_{\text{neg},q}$ from actual positive counts, without changing recall or quotas. `extended_results/README.md` indexes these tables and the complete absolute, Ψ, interaction, quota and check records. Both mixture and arrival expectations condition on saved blocker counts: they describe ranking consequences of quantity and composition, not the representation changes that produced the scores.

F EQUAL EXPOSURE AND SCORE-CONSTRAINED SOURCE COMPARISONS

F.1 FULL LOCAL PLUS 512 EXTERNAL NEGATIVES

FLC512 retains all identities in Local and all direct positives. For query q and source u, its eligible external pool is the unique identity set $E_u(q) = G_u \backslash (G_L \cup P(q))$. A candidate belonging to multiple sources is counted once through its exact provenance mask; masks containing the query domain are excluded. The saved identity graph records the query order, positive counts, source masks and component IDs for the count arrays. All forty domain–direction–source pools have capacity for 512 negatives.

Write $b_L = r_L - 1$, $N_u = |E_u|$, and B_u for external negatives strictly above the best positive. For $Z_u \sim \mathrm{Hypergeom}(N_u, B_u, 512)$, expected recall is $100\,\mathrm{mean}_q \Pr(Z_u \le K - 1 - b_L)$. Each mean uses the full original domain–direction query population, including Local failures. The calculation uses each query's integer counts. Zero additions recover Local, and the entire source pool recovers the single-add result, using exact identities and positives.

The fixed budget yields 440 distribution environments and 1,320 R@1/5/10 values across eleven physical states. They are expectations under sampling without replacement, not 440 independently sampled galleries. The 240 task-aligned rows contain 192 defined retention contrasts; the last task has no later update. With addition loss $\ell_t = R_{L,t} - \mathbb{E}[R_{\mathrm{FLC},t}]$, $X_{\mathrm{FLC}} = \ell_5 - \ell_\tau$ uses the same task reference as native X, but a different candidate design. No new interval is assigned. An expectation over sampled galleries does not give an observed joint query path.

Table 17: Complete R@10 X_{FLC} (pp), retaining all 64 contrasts. Columns identify added sources; "0 (exact)" distinguishes exact zeros from rounded nonzero values. There are 39 positive, 12 negative and 13 exact-zero entries. RSICD is a candidate source but has no query-retention row.

Config.	Query	Dir.	COCO	Flickr	IAPR	Fashion	RSICD
SeqFT	COCO	i2t	–	+0.046696	+0.040621	-0.000001	+0.000685
SeqFT	COCO	t2i	–	+0.075415	+0.089696	+0.016294	+0.572500
SeqFT	Flickr	i2t	+0.151827	–	+0.026073	0 (exact)	+0.139017
SeqFT	Flickr	t2i	+0.314877	–	+0.182966	+0.007687	+0.391640
SeqFT	IAPR	i2t	+0.211185	+0.243978	–	-0.000981	+0.192720
SeqFT	IAPR	t2i	-0.022388	-0.047029	–	+0.022995	+0.122483
SeqFT	Fashion	i2t	-0.000042	+0.001788	0 (exact)	–	-0.000539
SeqFT	Fashion	t2i	0 (exact)	0 (exact)	0 (exact)	–	0 (exact)
Mod-X	COCO	i2t	–	-0.014286	+0.033558	+0.001514	-0.064356
Mod-X	COCO	t2i	–	+0.011711	+0.096698	+0.009385	+0.480439
Mod-X	Flickr	i2t	+0.045699	–	+0.044754	0 (exact)	+0.014746
Mod-X	Flickr	t2i	+0.200414	–	+0.072410	-0.019829	+0.309982
Mod-X	IAPR	i2t	+0.008811	+0.047666	–	-0.003298	-0.045747
Mod-X	IAPR	t2i	-0.002430	+0.036322	–	+0.020634	+0.186208
Mod-X	Fashion	i2t	+0.000340	+0.000350	0 (exact)	–	0 (exact)
Mod-X	Fashion	t2i	0 (exact)	0 (exact)	0 (exact)	–	0 (exact)

Table 18: All four sources for SeqFT COCO t2i R@10 (pp). Post is M1 and final is M5. Addition loss compares full Local with its 512-negative expectation, preserving the full query denominator.

Added source	Post loss	Final loss	X_{FLC}
Flickr	0.516045	0.591460	0.075415
IAPR	0.491747	0.581443	0.089696
Fashion	0.016047	0.032340	0.016294
RSICD	0.214308	0.786809	0.572500

F.2 FOUR ARMS WITH A COMMON BACKGROUND AND A POST-TASK SCORE ANCHOR

This protocol-locked post-hoc extension covers COCO and IAPR, both learners and directions, all four external sources, and candidate seeds 0/1/2. For each domain–direction, the first 1,024 canonical IDs ordered by a fixed namespace hash are requested before matching; no failed query is replaced. The learner's own post-task checkpoint anchors matching (M1 for COCO; M3 for IAPR); M5 measures later change. Candidate identities and feasibility masks were sealed before these final-score outcomes were read. Earlier aggregate results had informed this post-hoc design.

Let P be complete direct positives and L^- all Local negatives. The proposal pool H takes the 256 negatives closest in anchor score to the tenth-highest Local negative, with hash tie-breaking. A deterministic sorted-window procedure seeks 64 disjoint triplets (a, b, o) with score span at most .005. It orders Local pairs by score/hash, assigns the first unused eligible external identity in the score interval, and uses a fixed hash coin for the A/B assignment. Failed queries are not refilled from outside H, and the tolerance is not relaxed. The external pool excludes Local and all positives. The random arm U samples 64 unique identities from the same pool without score selection and may overlap O.

With $C = L^- \setminus (A \cup B)$, the four galleries are $P + C + A$, $P + C + B$, $P + C + O$ and $P + C + U$. Each has $|G_L| - 64$ candidates. They share a query-specific background rather than one deployment gallery. Identities are fixed after construction; different learner anchors can produce different identities. Full Local ranks are retained as a pressure reference. A query enters Q_* only if the prescribed candidate construction succeeds under all three seeds for that source. Query means first average seeds, then queries; four-source means average the four source-specific populations equally. They are not comparisons on a common four-source query set.

$$\Delta_t = R_{O,t} - \tfrac{1}{2}(R_{A,t} + R_{B,t}), \qquad \Theta = \Delta_5 - \Delta_\tau, \qquad F_O - \tfrac{1}{2}(F_A + F_B) = -\Theta. \tag{10}$$

The random-arm counterpart and signed $B - A$ identity contrast are descriptive. The eight R@10 inference slots are separate from the original five-domain family. Each source must have at least 100 supported queries and 30 components. Original inference shares component weights over 3,520 components, with 10,000 saved draws and eight-slot Bonferroni percentile coordinates .003125/.996875. Three nonfinite draws in an unsupported slot remain missing, without deletion or replacement. Eligibility and interval disposition follow the original rules even when some slots are unsupported.

Table 19: All eight matched R@10 slots (pp). A zero point estimate or a degenerate slot is not an equivalence interval. All three usable approximate simultaneous intervals include zero, including endpoint zero. These intervals do not apply to native X, individual sources or secondary metrics.

Config.	Query	Dir.	Θ (pp)	Interval / disposition
SeqFT	COCO	i2t	+0.2110	Insufficient support
SeqFT	COCO	t2i	-0.0731	[-0.3482, +0.0000]
SeqFT	IAPR	i2t	+0.0257	[+0.0000, +0.1247]
SeqFT	IAPR	t2i	+0.0000	Insufficient support
Mod-X	COCO	i2t	+0.0000	Insufficient support
Mod-X	COCO	t2i	-0.0733	[-0.3442, +0.0000]
Mod-X	IAPR	i2t	+0.0000	Degenerate; no interval
Mod-X	IAPR	t2i	+0.0000	Insufficient support

Support and failure have different meanings. There are 11,416 supported requests out of 32,768; requests repeat queries across sources and configurations. Of 21,352 unsupported requests, 20,839 (97.5974%) have a necessary capacity bound below 64 for the fixed seed-0 proposal pool. This rules out three-seed support under that construction, not all possible matchers. The remaining 513 lack such a certificate. A saved bounded search of 256 fixed failures found no witness; fourteen of those still lack a capacity certificate. No optimality claim follows. Four-source common intersections for SeqFT are 0/26/127/14 queries (COCO i2t/t2i, IAPR i2t/t2i), and for Mod-X 0/26/129/4. The analysis uses the specified source-wise populations, not these smaller intersections.

Table 20: Complete source-specific matched coverage and R@10 Θ (pp). Each row requests 1,024 queries. Comp. counts supported positive-graph components; Cert. counts unsupported requests with a seed-0 necessary capacity certificate; Unres. counts remaining unsupported requests. They are not retrieval errors.

Config.	Query	Dir.	Source	$\lvert Q_* \rvert$	Comp.	Cert.	Unres.	Θ_{10}
SeqFT	COCO	i2t	Flickr	231	230	789	4	+0.2886
SeqFT	COCO	i2t	IAPR	19	19	1005	0	+0.0000
SeqFT	COCO	i2t	Fashion	8	8	1015	1	+0.0000
SeqFT	COCO	i2t	RSICD	120	119	900	4	+0.5556
SeqFT	COCO	t2i	Flickr	401	374	601	22	+0.0000
SeqFT	COCO	t2i	IAPR	769	708	249	6	+0.0000
SeqFT	COCO	t2i	Fashion	349	334	599	76	+0.0000
SeqFT	COCO	t2i	RSICD	342	316	639	43	-0.2924
SeqFT	IAPR	i2t	COCO	975	935	32	17	+0.0513
SeqFT	IAPR	i2t	Flickr	707	680	307	10	+0.0000
SeqFT	IAPR	i2t	Fashion	129	125	875	20	+0.0000
SeqFT	IAPR	i2t	RSICD	968	931	35	21	+0.0517
SeqFT	IAPR	t2i	COCO	383	383	631	10	+0.0000
SeqFT	IAPR	t2i	Flickr	83	83	938	3	+0.0000
SeqFT	IAPR	t2i	Fashion	84	84	932	8	+0.0000
SeqFT	IAPR	t2i	RSICD	283	283	727	14	+0.0000
Mod-X	COCO	i2t	Flickr	231	230	789	4	+0.0000
Mod-X	COCO	i2t	IAPR	19	19	1005	0	+0.0000
Mod-X	COCO	i2t	Fashion	8	8	1015	1	+0.0000
Mod-X	COCO	i2t	RSICD	120	119	900	4	+0.0000
Mod-X	COCO	t2i	Flickr	403	376	601	20	+0.0000
Mod-X	COCO	t2i	IAPR	769	708	251	4	+0.0000
Mod-X	COCO	t2i	Fashion	349	334	598	77	+0.0000
Mod-X	COCO	t2i	RSICD	341	315	641	42	-0.2933
Mod-X	IAPR	i2t	COCO	957	917	52	15	+0.0000
Mod-X	IAPR	i2t	Flickr	705	685	305	14	+0.0000
Mod-X	IAPR	i2t	Fashion	133	129	876	15	+0.0000
Mod-X	IAPR	i2t	RSICD	947	910	51	26	+0.0000
Mod-X	IAPR	t2i	COCO	270	270	749	5	+0.0000
Mod-X	IAPR	t2i	Flickr	48	48	974	2	+0.0000
Mod-X	IAPR	t2i	Fashion	45	45	971	8	+0.0000
Mod-X	IAPR	t2i	RSICD	220	220	787	17	+0.0000

Why replacement conditions rarely disagree at R@10. All 32 source-wise R@10 post-task contrasts are zero. Averaging each group's query–seed records, then all 32 groups and two times equally, the three matched arms disagree on success in 0.032152%; including the random arm raises this to 0.093496%. These group averages are not a pooled fraction of independent queries. Background-forced failure is 9.4850% and all-matched-arm success 90.4829%. Full Local failure becoming same-A success averages only 0.01308 pp on the same supports, providing little support for widespread new saturation caused by removing A/B. In 99.7239% of matched triplets, all three negatives lie on the same side of the best positive. Matching near the tenth-highest negative therefore need not place candidates near the score boundary that determines R@10 success.

Supports can also favor easier queries: for SeqFT COCO i2t with Flickr negatives, requested post-task Local R@10 is 91.21%, versus 97.40% on Q_*. This selection difference concerns that source comparison; it need not hold for every source. At R@10, the source-averaged mean absolute query contribution equals $|\Theta|$: small means do not hide large opposing flows on these supports. Lower thresholds need not behave identically: Mod-X IAPR i2t R@5 has macro $\Theta = 0.000138$ pp but mean absolute contribution 0.026261 pp. The numeric archive includes all R@1/5/10/MRR values, seed supports and signed/unsigned flows.

Table 21: All four arms and full Local on each source's Q_*: post/final R@10 (%), with three seeds averaged. O is matched and U random; all comparisons within a row share positives, background and count. Between-row populations differ. Full-precision values and all four metrics are included in the candidate archive.

Config.	Query	Dir.	Source	Full Local	A	B	O	U
SeqFT	COCO	i2t	Flickr	97.40/94.37	97.40/94.52	97.40/94.52	97.40/94.81	97.40/94.81
SeqFT	COCO	i2t	IAPR	94.74/94.74	94.74/94.74	94.74/94.74	94.74/94.74	94.74/94.74
SeqFT	COCO	i2t	Fashion	100.00/100.00	100.00/100.00	100.00/100.00	100.00/100.00	100.00/100.00
SeqFT	COCO	i2t	RSICD	95.83/91.67	95.83/92.22	95.83/91.67	95.83/92.50	95.56/92.50
SeqFT	COCO	t2i	Flickr	85.04/80.05	85.04/80.05	85.04/80.05	85.04/80.05	85.04/79.97
SeqFT	COCO	t2i	IAPR	80.49/77.11	80.49/77.11	80.49/77.11	80.49/77.11	80.41/77.07
SeqFT	COCO	t2i	Fashion	89.68/84.53	89.68/84.53	89.68/84.53	89.68/84.53	89.68/84.53
SeqFT	COCO	t2i	RSICD	77.78/75.73	77.78/75.73	77.78/75.73	77.78/75.44	77.68/75.24
SeqFT	IAPR	i2t	COCO	90.87/86.46	90.87/86.50	90.87/86.53	90.87/86.56	90.87/86.46
SeqFT	IAPR	i2t	Flickr	90.95/89.25	90.95/89.25	90.95/89.25	90.95/89.25	90.95/89.20
SeqFT	IAPR	i2t	Fashion	92.25/93.80	92.25/93.80	92.25/93.80	92.25/93.80	92.25/93.80
SeqFT	IAPR	i2t	RSICD	89.67/85.74	89.67/85.85	89.67/85.74	89.67/85.85	89.57/85.74
SeqFT	IAPR	t2i	COCO	91.12/85.12	91.12/85.12	91.12/85.12	91.12/85.12	91.04/85.12
SeqFT	IAPR	t2i	Flickr	98.80/90.36	98.80/90.36	98.80/90.36	98.80/90.36	98.80/90.36
SeqFT	IAPR	t2i	Fashion	100.00/96.43	100.00/96.43	100.00/96.43	100.00/96.43	100.00/96.43
SeqFT	IAPR	t2i	RSICD	92.93/88.34	92.93/88.34	92.93/88.34	92.93/88.34	92.93/88.22
Mod-X	COCO	i2t	Flickr	97.40/95.24	97.40/95.24	97.40/95.24	97.40/95.24	97.40/95.24
Mod-X	COCO	i2t	IAPR	94.74/94.74	94.74/94.74	94.74/94.74	94.74/94.74	94.74/94.74
Mod-X	COCO	i2t	Fashion	100.00/100.00	100.00/100.00	100.00/100.00	100.00/100.00	100.00/100.00
Mod-X	COCO	i2t	RSICD	95.83/95.83	95.83/95.83	95.83/95.83	95.83/95.83	95.56/95.83
Mod-X	COCO	t2i	Flickr	84.86/81.89	84.86/81.89	84.86/81.89	84.86/81.89	84.86/81.64
Mod-X	COCO	t2i	IAPR	80.49/79.19	80.49/79.19	80.49/79.19	80.49/79.19	80.41/78.98
Mod-X	COCO	t2i	Fashion	89.68/86.82	89.68/86.82	89.68/86.82	89.68/86.82	89.68/86.82
Mod-X	COCO	t2i	RSICD	78.01/78.01	78.01/78.01	78.01/78.01	78.01/77.71	77.91/77.91
Mod-X	IAPR	i2t	COCO	91.43/87.88	91.43/87.88	91.43/87.88	91.43/87.88	91.33/87.77
Mod-X	IAPR	i2t	Flickr	92.34/90.21	92.34/90.21	92.34/90.21	92.34/90.21	92.25/90.07
Mod-X	IAPR	i2t	Fashion	95.49/95.49	95.49/95.49	95.49/95.49	95.49/95.49	95.49/95.49
Mod-X	IAPR	i2t	RSICD	90.39/87.12	90.39/87.12	90.39/87.12	90.39/87.12	90.39/87.05
Mod-X	IAPR	t2i	COCO	88.52/83.33	88.52/83.33	88.52/83.33	88.52/83.33	88.52/83.33
Mod-X	IAPR	t2i	Flickr	100.00/91.67	100.00/91.67	100.00/91.67	100.00/91.67	100.00/90.97
Mod-X	IAPR	t2i	Fashion	100.00/100.00	100.00/100.00	100.00/100.00	100.00/100.00	100.00/100.00
Mod-X	IAPR	t2i	RSICD	94.09/89.55	94.09/89.55	94.09/89.55	94.09/89.55	94.09/89.39

F.3 Fixed-identity score changes and retrieval consequences

For a triplet define $\gamma = [s_O - (s_A + s_B)/2]_5 - [s_O - (s_A + s_B)/2]_\tau$. The positive cancels in this relative negative contrast, while both query and candidate encoders can change. Γ_q averages the 64 triplets, then the three fixed seeds. Source means and quantiles summarize queries equally on Q_*. Absolute within-query drift first takes each triplet's absolute change before averaging; it differs from $|\Gamma_q|$. The same-source quantity replaces $s_O - (s_A + s_B)/2$ with $s_B - s_A$.

For anchor-best positive p_τ, best-positive change separates into the fixed-identity score change and a nonnegative switching gain:

$$\max_p s_5(q,p) - \max_p s_\tau(q,p) = [s_5(q,p_\tau) - s_\tau(q,p_\tau)] + [\max_p s_5(q,p) - s_5(q,p_\tau)]. \tag{11}$$

Actual ranks always use the current best positive, not the anchor identity alone. Negative order-statistic changes similarly distinguish a fixed anchor-boundary identity from reordering; the latter term can be negative for $K > 1$. All these changes are cosine units, not pp. They compare checkpoints and must not be confused with the same-checkpoint numeric-path differences in Appendix I.2.

All 32 source-specific score summaries, including query quantiles and both absolute-drift measures, are indexed in `extended_results/README.md`. Entries use cosine units; rounded zeros do not assert exact zero.

For the SeqFT COCO i2t/Flickr example, $\Gamma = -0.004811$ coexists with fixed-positive change -0.021059 and switching gain $+0.003123$. The signed same-source mean is near zero under symmetric A/B assignment, but mean absolute within-query drift is 0.016182 versus 0.015271 for cross. Similar absolute drift does not show that provenance matters more than identity, or explain a small signed contrast. Scores can move without crossing a Top-K threshold. Here the off-versus-same R@10 contrast changes little, which does not imply small absolute retention losses. The archive includes all populations, quantiles, lower-threshold disagreements and boundary identities.

G SEPARATE OUT-OF-STREAM SKETCHSCENE CASE

Here SketchScene names the local export from `zoheb/sketch-scene`, whose dataset card attributes its source to FS-COCO (zoheb, n.d.; Chowdhury et al., 2022); its local split is not certified as an upstream official split. This earlier probe uses 2,000 SketchScene image queries and 1,885 normalized text queries, three candidate-composition seeds, and fixed $N = 512$. SketchScene is never trained; this is out-of-stream behavior during the five-domain adaptation stream, not acquisition or forgetting of a SketchScene task. It compares disjoint same-source negative sets (same1 and identity-control same2) and COCO negatives (off), preserving queries and legal positives. Frozen negative-score summary matching is not full difficulty equivalence. Historical pools, B/16 diagnostics and the source-pair extension have separate populations; none is treated as another held-out replication.

What is held fixed in the replacement. Each query retains all direct legal positives, leaving $512 - |P_q|$ negative slots. Frozen M_0 normalized float32 cosine scores define an off-source support interval between its 0.001 and 0.999 quantiles (inclusive). Same-source negatives within that interval form the reference sampling pool. If too few remain, the code expands to the complete off-source score range; it fails if this is still insufficient. All 11,655 recorded query–seed constructions in the six held-out manifests used the first branch. This population must not be joined to the historical 800/674-query exploratory pool.

Seeds 0/1/2 with base namespace 260903 select same1 deterministically. Each reference score is matched to the nearest unused off-source negative; sorted-score search breaks equal-distance choices to the left. Legal positive identities are excluded. Same2 instead matches to remaining same-source identities after excluding all positives and same1 negatives, with a sham-specific random comparator. Candidates are sealed before adapted evaluation and reused at every checkpoint; the later same2 extension is descriptive, not an independent confirmation.

Matching gates first compute each query's negative-score mean and p90/p95/p99, then average the absolute same/off discrepancy of each statistic across queries. The mean discrepancy must beat random matching and be at most 0.005; each tail-statistic discrepancy is at most 0.01, in cosine units. These are neither quantiles of matched-pair discrepancies nor per-query worst-case bounds. The N=128 maximum-negative gate does not apply to N=512. Changing candidate identities changes content and score support, so the construction is not a pure source-label intervention.

Two inferential questions, kept separate. For the original R@1 trajectory define $\Delta_t = R_{t,\mathrm{off}} - R_{t,\mathrm{same1}}$ and $\Theta = \Delta_5 - \Delta_0$. The four predictions were positive i2t and negative t2i amplification. Both i2t predictions receive support; neither t2i prediction does (Table 22). Original nominal intervals drew component weights separately by direction, breaking their cross-direction dependence. The later locked post-hoc sensitivity uses common weights over 1,885 legal-positive components, 10,000 draws (seed 26090904), query-equal random denominators and higher-quantile maximum standardized deviations. Its critical values are 2.468940 for four endpoints and 2.924282 for the separate 24-value trajectory family. These families are not joined to each other or to the five-domain primary family. Full stage absolute performance, source contrasts, intervals and rank-bin joins are retained in the archive, including unsupported predictions.

Table 22: All four original R@1 endpoint contrasts (pp). Current intervals are the post-hoc shared-component approximate simultaneous 95% sensitivity; original bounds retain nominal status. Predicted signs are unchanged.

Learner	Dir.	Θ	Original nominal CI	Current four-family CI	Pred.	Supported
SeqFT	i2t	4.300	[3.094, 5.506]	[3.081, 5.519]	+	Yes
Mod-X	i2t	5.050	[3.836, 6.264]	[3.815, 6.285]	+	Yes
SeqFT	t2i	0.813	[-1.184, 2.811]	[-1.194, 2.820]	−	No
Mod-X	t2i	-0.884	[-2.867, 1.098]	[-2.882, 1.114]	−	No

The later CPU float32/batch-64 endpoint path compares one physical Frozen B/32 state with SeqFT/Mod-X M_5 on the same identities. Its separate post-hoc family has 32 adapted-minus-Frozen comparisons, common query-equal component weights, 10,000 draws (seed 26091101), higher-quantile critical value 2.932345 and no artificial epsilon for constant columns. Supported fixed-metric sign reversals occur at i2t R@5/R@10 for both learners; i2t R@1 remains below Frozen, Mod-X i2t MRR reverses only as a point estimate, and t2i has no corresponding supported reversal. Same2 is outside this family.

Table 23 gives the absolute values underlying the highlighted reversals: the adapted endpoints are below Frozen in same1 and above it in off at both metrics. Three seed values are averaged within each query before averaging queries equally. The complete 72-row endpoint table retains both directions, all four metrics and same2; it uses the same CPU path and does not mix the earlier GPU trajectory records.

Table 23: Absolute i2t recall in the separate held-out case (2,000 queries). These are point estimates; the corresponding differences and simultaneous intervals belong to the separate 32-contrast family below.

	R@5 (%)		R@10 (%)	
Model	same1	off	same1	off
Frozen	28.73	28.75	38.97	38.97
SeqFT M_5	21.95	32.72	31.42	43.42
Mod-X M_5	25.20	35.73	35.27	48.10

Table 24: Complete 32-contrast adapted-minus-Frozen family. Each row gives two conditions and their approximate simultaneous 95% intervals. Recall differences are pp; MRR differences are $100\times$MRR index points.

State	Dir.	Metric	same1 − M0	Simultaneous CI	off − M0	Simultaneous CI
SeqFT	i2t	R@1	-9.583	[-12.061, -7.106]	-5.283	[-7.883, -2.684]
SeqFT	i2t	R@5	-6.783	[-9.818, -3.748]	+3.967	[+0.747, +7.186]
SeqFT	i2t	R@10	-7.550	[-10.845, -4.255]	+4.450	[+1.024, +7.876]
SeqFT	i2t	MRR×100	-8.364	[-10.647, -6.080]	-1.516	[-3.893, +0.862]
Mod-X	i2t	R@1	-7.650	[-10.037, -5.263]	-2.600	[-5.102, -0.098]
Mod-X	i2t	R@5	-3.533	[-6.459, -0.607]	+6.983	[+3.932, +10.034]
Mod-X	i2t	R@10	-3.700	[-6.911, -0.489]	+9.133	[+5.865, +12.402]
Mod-X	i2t	MRR×100	-5.879	[-8.051, -3.707]	+1.397	[-0.847, +3.642]
SeqFT	t2i	R@1	-10.027	[-12.428, -7.625]	-9.213	[-11.468, -6.958]
SeqFT	t2i	R@5	-6.083	[-8.649, -3.517]	-14.306	[-17.182, -11.430]
SeqFT	t2i	R@10	-5.411	[-8.271, -2.551]	-16.746	[-19.859, -13.634]
SeqFT	t2i	MRR×100	-8.388	[-10.462, -6.314]	-10.835	[-12.843, -8.827]
Mod-X	t2i	R@1	-7.657	[-9.924, -5.390]	-8.541	[-10.676, -6.406]
Mod-X	t2i	R@5	-3.307	[-5.743, -0.871]	-13.740	[-16.487, -10.993]
Mod-X	t2i	R@10	-2.688	[-5.324, -0.052]	-16.428	[-19.413, -13.443]
Mod-X	t2i	MRR×100	-5.760	[-7.689, -3.831]	-10.127	[-12.006, -8.248]

G.1 Reference loss, rank changes, and the identity control

Source contrasts do not measure absolute change. In the original trajectory, $H = (1, 0)$ denotes same1-only success and $SW = (0, 0)$ failure under both; $H \to SW$ improves the off-minus-same1 contrast by losing reference success. Table 25 shows its contribution to the positive part P_+ of temporal change, with negative part P_- and $\Theta = P_+ + P_-$. A path that changes the source contrast by two units contributes twice its query mass; contribution and query percentage must therefore be distinguished. Actual M0–M5 rank-bin joints also separate losing rank one from moving beyond rank ten, without treating bins as semantic abilities. All eight learner/direction/condition joints remain available, rather than inferring paths from endpoint marginals.

Table 25: Reference-success loss in all four original endpoint changes. $H \to SW$ mass is a query percentage; $P_+/P_-/\Theta$ are pp. The share column divides this path's contribution by P_+.

Learner	Dir.	$H \to SW$ mass (%)	P_+ (pp)	P_- (pp)	Share of P_+ (%)	Θ (pp)
SeqFT	i2t	0.03	+6.28	-1.98	0.53	+4.30
SeqFT	t2i	5.91	+7.87	-7.06	75.06	+0.81
Mod-X	i2t	0.03	+6.83	-1.78	0.49	+5.05
Mod-X	t2i	5.45	+7.13	-8.01	76.43	-0.88

For a within-state replacement, rescue R_u and harm H_u are percentages of the full query population; $\Delta = R_u - H_u$ (pp) and unsigned transition mass $T = R_u + H_u$ (%). Risk-set conditional rescue/harm have different denominators and cannot be subtracted. Both cross and sham use the same current same1 reference on one CPU path. Sham changes many query outcomes: its adapted net effects are small because recovery and harm roughly cancel, not because identity replacement has no effect.

Table 26: Complete cross/sham R@1 net contrasts and transition mass, including Frozen. Both replacements share same1; rounded $R_u - H_u$ may differ in the last digit from rounded Δ. No new tests are added.

State	Dir.	Replacement	Δ (pp)	R_u (%)	H_u (%)	T (%)
Frozen M_0	i2t	Cross	0.00	0.03	0.03	0.07
Frozen M_0	i2t	Sham	-0.17	0.22	0.38	0.60
Frozen M_0	t2i	Cross	-8.91	0.05	8.97	9.02
Frozen M_0	t2i	Sham	+0.02	0.41	0.39	0.80
SeqFT M_5	i2t	Cross	+4.30	6.25	1.95	8.20
SeqFT M_5	i2t	Sham	-0.08	2.48	2.57	5.05
SeqFT M_5	t2i	Cross	-8.10	1.36	9.46	10.82
SeqFT M_5	t2i	Sham	+0.23	3.36	3.13	6.49
Mod-X M_5	i2t	Cross	+5.05	6.80	1.75	8.55
Mod-X M_5	i2t	Sham	+0.20	2.93	2.73	5.67
Mod-X M_5	t2i	Cross	-9.80	1.17	10.96	12.13
Mod-X M_5	t2i	Sham	-0.09	3.36	3.45	6.81

Matching precision differs. Across three seeds, cross/sham mean absolute negative-mean discrepancies are $(5.85\text{–}6.09) \times 10^{-7}$ versus $(1.45\text{–}1.53) \times 10^{-5}$ for i2t, and $(1.87\text{–}1.88) \times 10^{-3}$ versus $(1.23\text{–}1.35) \times 10^{-5}$ for t2i. Their mean absolute p99-score discrepancies are $(4.85\text{–}5.06) \times 10^{-5}$ versus $(3.99\text{–}4.06) \times 10^{-4}$, and $(3.02\text{–}3.06) \times 10^{-3}$ versus $(4.03\text{–}4.30) \times 10^{-4}$, respectively, in cosine units. These are averages of within-query discrepancies, not per-query maxima. All six means/tails and original gate decisions remain indexed; the N=128 maximum-score gate is not retroactively added to N=512 matching.

The internal comparison with five component folds also depends on the eligible pool. With Frozen included, all i2t R@5/R@10 same1 selections are Frozen and off selections Mod-X; with adapted endpoints only, that difference disappears. The adapted-only t2i same1→off signed differences are $-0.619/-0.707$ pp for R@5/R@10, respectively. They are not clipped to nonnegative "regret."

Both pools, all routes and 720 fold-level choices remain indexed. Frozen's fitness for newly learned tasks is not tested by this out-of-stream probe.

G.2 Residual competition, identity, and replay limits

Let $m_{0,g,s}(q)$ be Frozen maximum-negative minus best-positive score. The common residual is $E(q) = \max_s \max_{g\in\{\text{off},\text{same2}\}} |m_{0,g,s}(q) - m_{0,\text{same1},s}(q)|$, with $Q_\tau = \{q : E(q) \le \tau\}$. All nine existing coordinates $0, .0001, .0005, .001, .0025, .005, .01, .02, \infty$ remain; membership is fixed across models and metrics. At .005, i2t retains 1,918 queries (95.9%). Its four R@5/R@10 adapted-minus-Frozen reversal signs persist descriptively: SeqFT same1/off are $-6.900/+3.806$ and $-7.803/+4.432$ pp; Mod-X values are $-3.754/+6.917$ and $-3.962/+9.124$ pp. No subset inherits full-population intervals. The strictest nonempty coordinate retains only 17 i2t queries. The maximum i2t cross residual is .001315087, below .005; all 82 exclusions at .005 therefore arise through sham residuals (maximum .017473966), not a removed cross-source tail. First-negative residual control does not equate rank-five/rank-ten competition or post-update difficulty.

For within-query contrast d_q, original-population contributions are $C_{\text{in}} = 100|Q|^{-1}\sum_{q\in Q_\tau} d_q$ and $C_{\text{out}} = 100|Q|^{-1}\sum_{q\notin Q_\tau} d_q$; the conditional retained mean is C_{in}/c_τ, where $c_\tau = |Q_\tau|/|Q|$. Adapted $\Xi = \Theta_{\text{cross}} - \Theta_{\text{sham}}$ differs from original Θ, while Frozen Δ_0 is a within-state source contrast. At .005, i2t retained/complement Ξ contributions are 4.167/0.050 pp for SeqFT and 4.617/0.067 pp for Mod-X. The same coordinate retains 54.80% of t2i queries: contributions are $-2.865/+3.466$ and $-3.378/+2.599$ pp. Frozen t2i's -8.912467 pp full contrast is concentrated in higher-residual queries: at .01, 68.4350% coverage contributes -0.671972 pp and its complement -8.240495 pp. These descriptive decompositions narrow the interpretation; they do not establish residual causation or source-label effects. All 432 subset comparisons and nine thresholds remain in the numeric index, including 48 undefined comparison entries with empty retained subsets; empty conditional means are undefined although their population contribution is zero. Units are already pp/index points after subtracting displayed-scale scores, without a further factor of 100.

The local query-to-reference-photo mapping remains unrecovered. Overlap counts for the upstream reference-photo population are not local query leakage rates, and unknown overlap is not zero. Historical identity checks do not establish perceptual or semantic disjointness. The recovered support-branch input has 11,655 original rows and a six-row summary, bound to its original manifest hashes; it is not reconstructed from a PDF table. The archive preserves that input and its existing reconstruction entry. The 66-cell N=128 extension retains 12 passes and 54 failures; failed adapted effects remain unmeasured. Historical concentration uses pooled-seed counts by default; B/16, equal-seed and tie-sensitive variants retain their different identities in the archive.

The historical strict re-encoding checks, identified as A16 in the archive, failed feature parity and subsequently failed unchanged retrieval tolerances; these remain failed checks. The H/C/R diagnostic separates candidate construction from Frozen-cache substitution: H uses the historical candidate manifest, C reconstructs candidates from the original Frozen cache in the diagnostic runtime, and R repeats that reconstruction using the replacement Frozen SketchScene cache. H→C changed 1,098 of 4,422 query/seed candidate sets without changing tested adapted R@1 correctness; aggregate equality is not stimulus or rank equivalence. A later, separately locked raw-input check covers only Frozen M_0, SeqFT M_5 and Mod-X M_5 B/32 endpoints. Twelve feature matrices (7,000 images and 26,715 texts per state) have recorded equal complete-matrix hashes and zero feature difference; 104,895 query-condition rows agree under the paired scorer (54 cells, or 34,965 wide records, not independent queries). This does not certify all stages, cross-environment replay or training. Eight earlier GPU-to-CPU first-positive-rank differences, none changing R@1 correctness, remain recorded. Current local lightweight checks read these records; they do not re-encode the omitted raw images or erase the earlier failures.

H COMMON-TARGET ADAPTATION FROM HISTORICAL CHECKPOINTS

Question and design. This separate diagnostic adapts five historical starts—M_0, SeqFT M_1/M_4 and Mod-X M_1/M_4—to RSICD. None has encountered RSICD in the original task stream; pre-training overlap is unknown. Each start receives 256 updates with seeds 0 and 1, sharing each seed's image order and caption choices. All branches use symmetric one-pair CLIP loss, without distillation or automatic multiple-positive treatment of repeated batch semantics. Labels identify historical weights; the target-training algorithm is shared, and the M_0-initialized endpoint is no longer frozen.

A reset AdamW optimizer updates all parameters: learning rate 10^{-6}, weight decay 0.1, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, batch 128 and accumulation one. The 256-step schedule uses 51 warmup steps followed by cosine decay. Training and encoding use float32 without AMP or TF32, with the original state reloaded after model construction. The inherited logit scale is trainable, with its logarithm clamped to $[0, 4.6052]$. CLIP224 preprocessing and length-77 tokenization preserve EOT.

Evaluation identity. At updates 0, 32, 64, 128 and 256, each model re-encodes the candidate universe for 1,093 i2t and 3,330 t2i RSICD queries in full Local and five-domain Shared galleries, using the original direct-positive graph. Gains use each start's own re-encoded update zero, shared between seeds. The seeds repeat only the short target adaptation, not the historical trajectories. No task follows RSICD, so this changing environment gap is not the main study's retention contrast X.

Table 27: Common-target R@10 (%). Each cell is update $0 \rightarrow 256$, with the final value averaged over two diagnostic seeds. Starts, gains and final levels are distinct quantities; the labels denote weight histories. No confidence intervals are assigned.

Historical start	i2t Local	i2t Shared	t2i Local	t2i Shared
M_0	$29.00 \rightarrow 49.45$	$27.17 \rightarrow 45.84$	$30.09 \rightarrow 47.13$	$20.06 \rightarrow 35.65$
SeqFT M_1	$33.76 \rightarrow 50.73$	$33.03 \rightarrow 48.58$	$36.61 \rightarrow 49.64$	$25.98 \rightarrow 42.45$
SeqFT M_4	$31.75 \rightarrow 49.59$	$30.38 \rightarrow 47.12$	$33.75 \rightarrow 49.23$	$27.00 \rightarrow 43.60$
Mod-X M_1	$33.58 \rightarrow 50.73$	$32.85 \rightarrow 48.54$	$36.58 \rightarrow 49.64$	$25.98 \rightarrow 42.46$
Mod-X M_4	$34.22 \rightarrow 51.05$	$32.48 \rightarrow 48.49$	$35.23 \rightarrow 50.02$	$26.43 \rightarrow 43.27$

What improves, and what this leaves open. Every run improves endpoint R@1/5/10 and MRR in both directions and environments over its own update zero. Later starts do not consistently perform worse. However, one target and a short shared recipe do not establish preserved plasticity: initial competence, historical training budget and inherited state differ. Table 27 reports starting and final levels; Figure 13 shows every scheduled R@10 checkpoint. Three R@1 segments decrease in the full records, so learning is not uniformly monotone.

An environment-dependent branch comparison. For target-adapted SeqFT histories, mean t2i R@10 is 49.64% Local / 42.45% Shared from M_1, versus 49.23% / 43.60% from M_4. The two seeds' $M_4 - M_1$ differences are $-0.571/-0.240$ pp in Local and $+0.811/+1.502$ pp in Shared. This opposite ordering is descriptive; no new inferential family tests it. For R@10 it already exists at update zero (-2.853 and $+1.021$ pp), so the short adaptation cannot be said to have created it. These new branches are separate from the native endpoint pool and Sketch comparisons. The archive reports all M_1/M_4 comparisons across parent configurations, directions, metrics, seeds and updates 0/256, including cases without opposite ordering.

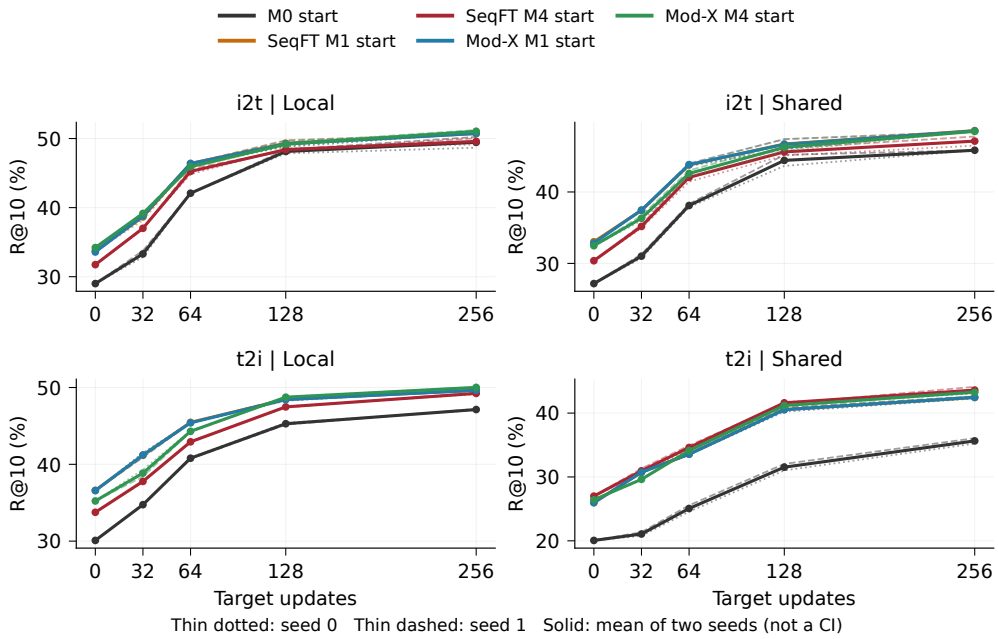

Figure 13: All five historical starts, both directions and both candidate environments under the common RSICD target. The horizontal axis uses actual update counts. Thin dotted/dashed curves are seeds 0/1; solid curves are their means. These are observed curves, not confidence bands. The supplement also gives complete R@1, R@5, MRR and fixed-loss curves.

Fixed loss and material scope. Fixed train and held-out diagnostics each use 256 images in two batches of 128. Losses are reported at the current learned scale and fixed scale $1/0.07$, separately from training-batch loss. Every endpoint loss decreases, but intermediate increases occur: SeqFT M_4's two-seed mean fixed-scale held-out values are $3.8826, 3.9129, 3.9112, 3.7806, 3.7309$. These batch losses do not substitute for full-gallery retrieval or identify a mechanism.

The review package's `diagnostics/common_target_adaptation/README.md` indexes 800 absolute metric rows, 160 endpoint changes, two loss logs and all branch comparisons. Query IDs, components and positive counts agree with the reference graph. The protocol lock is available, but the local export lacks every evaluation's complete positive-ID list and completion metadata, full training streams, fixed-diagnostic sample identities and training/encoding implementation. Local checks therefore rebuild saved results without verifying every executed positive set, replaying training or re-encoding images. These results do not isolate an encoder-specific cause.

I TEXT PROCESSING AND NUMERICAL VERIFICATION SCOPE

I.1 TOKENIZATION AND TEXT POPULATIONS

Training and evaluation retain the end-of-text token when truncating to the 77-token context. Table 28 distinguishes canonical evaluation texts from training annotations and the expected frequency of truncation under the caption-selection rule.

Table 28: Different text populations must not be combined. Eval texts are retained canonical representatives; their truncation counts sum to 614. Training unique-text and row counts refer to retained annotations. Potential % is row-weighted expected truncation under the retained caption-selection rule, not observed historical batches or performance loss.

Source	Eval texts	Trunc.	Train rows	Unique texts	Trunc.	Potential %
COCO	24825	0	82783	400962	0	0.0000
Flickr	4999	1	30014	149588	2	0.0013
IAPR	4266	55	15000	14802	42	0.3000
Fashion	6763	558	260490	59127	5141	10.3428
RSICD	3330	0	9828	15097	0	0.0000

FashionGen's 10.3428% training estimate follows uniform caption selection within each annotation row. It is neither unique-text prevalence nor the rate in the canonical evaluation population after exclusions. The common-target diagnostic also preserves EOT. Input profiles are included in the archive.

I.2 SIX-CHECKPOINT BOUNDED RAW-INPUT COMPARISON

The server check compares SeqFT and Mod-X M1/M3/M5 CPU encodings against saved CUDA representations. CPU models use `torch.float16`, float32 normalization, image batch 4 and text batch 16. This bounded comparison is not an FP32 replay and does not isolate device, precision, batching or preprocessing effects. The original absolute score tolerance is 10^{-5}; ranking uses strict comparisons without an epsilon.

Table 29: All six raw-input checks completed. The 256 query–state–domain–direction records contain 64 distinct query IDs, with candidate subsets of 139–407 identities. Completion does not establish numerical equivalence.

State	Pairs	$\lvert\Delta s\rvert > 10^{-5}$	Records	Rank changes
modx_m1	5744	5043	32	0
modx_m3	5904	5254	32	1
modx_m5	23104	20530	64	0
seqft_m1	5728	5041	32	0
seqft_m3	6384	5703	32	0
seqft_m5	24000	21539	64	0
Total	70864	63110	256	1

Of 70,864 score pairs, 63,110 exceed tolerance; the maximum absolute error is 0.0004043877. One subset rank changes from 3 to 4, changing reciprocal rank from $1/3$ to $1/4$. R@1/5/10 flags and best-positive identities remain unchanged in all 256 records. Every sampled record succeeds at R@10, with maximum rank seven, so these subsets do not test full-gallery Top-10 boundaries. The local verifier reconstructs errors, margins, blockers and ranks from saved scores; it does not re-encode images or establish numerical equivalence.

Exact scope of the changed record. The changed record is an IAPR i2t query at Mod-X M3 with one legal positive. Query, positive and crossing negative all have IAPR-only provenance; the negative belongs to a different statistical component. The identity records are indexed in `extended_results/README.md`.

Table 30: The changed pair in its 230-candidate subset. Scores and margins use cosine units. The negative score falls, but the positive falls further; this is a same-checkpoint numeric-path comparison, not a training-stage mechanism.

Quantity	Saved CUDA	CPU FP16 path
Best-positive score	0.2897466421	0.2896497250
Negative score	0.2897353172	0.2896571457
Positive-minus-negative margin	$+1.1324883 \times 10^{-5}$	$-7.4207783 \times 10^{-6}$
Subset best-positive rank	3	4
R@1 / R@5 / R@10	0 / 1 / 1	0 / 1 / 1

The query belongs to the original 4,970-query IAPR i2t population. Its saved Local/Shared ranks are 13/13 at M2, 10/10 at M3 and 81/84 at M5: failure, acquisition and later failure in both environments. Thus the saved post-task full-gallery rank lies exactly at the R@10 boundary. Local contains 4,266 texts and Shared 44,183, compared with 230 candidates in the raw subset. Saved complete scalar rows reproduce the M3/M5 ranks without fresh scoring.

Matching succeeds for COCO and Flickr across three seeds, but fails for FashionGen and RSICD. In all six supported constructions, the crossing negative is in the common Local background. All four post-task arm ranks are 10; final ranks are 80 or 81, and all 48 saved ranks reproduce. The raw subset contains only 2/64 complete triplets per construction. It therefore cannot establish all matched residuals or full-gallery rank changes: an error below .005 does not verify matching decisions, and subset rank 3–4 does not imply a one-position change in the complete gallery. No fresh/cached mixture is used.

Relationship to other checks. These six-state checks update the earlier partial raw snapshots without changing FLC512, matched analyses, the native eight contrasts, separate Sketch endpoints or the historical strict re-encoding failures (A16; Appendix G). A separate 1,062-record cached-score check found identical ranks and maximum error 2.62047×10^{-7}; it tests a different numerical level. The portable entry in `diagnostics/history_numeric_fact_check/` verifies saved scalar inputs without rescoring, encoding or training. Evaluation and reconstruction code is available, but code availability does not establish completion of a numerical check. The inspected records do not contain a complete, consistent fresh encoding and full-gallery reranking of all five-domain model states; the full-gallery numerical impact therefore remains unmeasured.

J Complete evidence and reconstruction index

The anonymous supplement maps figures and tables to their inputs and code in `code/PAPER_ARTIFACT_MAP.csv`; `code/EVIDENCE_ARCHIVE_INDEX.csv` indexes the packaged evidence. It includes native counts, query-path and cohort records, FLC512 inputs, retained four-arm ranks and statistical inputs, separate Sketch/support-branch records, common-target curves and logs, and bounded raw-check scores. Complete four-arm candidate identities and selected-score arrays are not included in this minimal archive.

`code/REPRODUCTION_SCOPE.csv` and the supplementary README distinguish saved-record reconstruction and plotting from raw-input encoding and training. Raw datasets, full feature caches and model weights are not redistributed. The separate full-gallery comparison records in `diagnostics/full_gallery_replay/` concern post-task and final query roles for the four retention tasks. They are distinct from the bounded CPU comparison in Appendix I.2; the local reconstruction does not repeat either raw-input encoding procedure. The original inference families are unchanged.