READABLE FACTORS, UNRESOLVED COMBINATIONS: SUPPORT-AWARE CORRECTION FOR WARM-START RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A warm item can combine factors a user has seen separately but never together, even when that combination occurs elsewhere in training. Standard warm-start metrics mix these cases with repeated combinations, while tests of globally unseen combinations study a different setting. We define a pre-cutoff evaluation cohort and use an independently fitted audit to test whether pair features add held-out ranking value beyond a frozen ranker and the individual factors. On H&M fashion and M3L movies, the audit finds a positive increment from purified pair features; pair permutations and equal-capacity random controls are near zero. We introduce Support-Aware Compositional Factor Residual (SCFR), which removes declared lower-order variation from pair features and uses train-only pairwise-logistic curvature to shrink weakly supported comparison directions. One correction scores every eligible warm-catalog item. The analysis states when a target comparison is identifiable and locally stable and gives a conditional squared-risk result for shrinkage when signal aligns with support. On the two confirmatory anchors, SCFR improves target NDCG@20 over the strongest matched-input control selected on validation by +0.0074 [0.0015, 0.0081] and +0.0083 [0.0014, 0.0094], corresponding to relative gains of 20.1% and 6.7%, while passing prespecified overall and non-target non-inferiority checks. Amazon Sports replicates the target-gain direction; controlled anti-alignment and weak-support tests show where the advantage vanishes or reverses.

1 INTRODUCTION

Even when a user and an item are both warm, the item can combine factors the user has seen separately but never together. Suppose the user has seen red shirts and blue jackets before cutoff T. A red jacket is warm, each factor is familiar to this user, and red jackets occur elsewhere in training. We ask where a later red-jacket positive ranks among all eligible warm items. The evaluation cohort identifies such held-out positives; every method still scores the same full warm catalog. Aggregate warm-start metrics mix these positives with repeated combinations and cannot answer this user-local question.

Existing work supplies useful ingredients but evaluates different settings. Multimodal and content-aware recommenders learn richer item representations (Zhang et al., 2025; Li et al., 2026c), and explicit-cross models can express factor interactions (Wang et al., 2021). Constrained recommendation and broader compositional-generalization studies often consider attribute queries, set constraints, or combinations absent from global training support (Mahajan et al., 2025; Dasgupta et al., 2025). These evaluations do not directly test whether a globally supported pair that is new to one user adds ranking value beyond a frozen ranker and the same factor main effects. Figure 1 shows the full-catalog decision and the independent audit of pair value that precedes our method.

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.

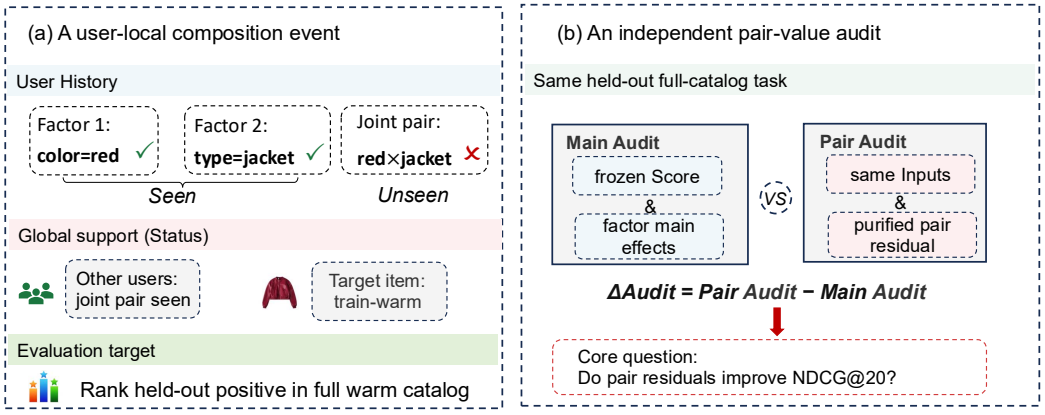

Figure 1: **The user-local composition decision.** A warm item's factor pair is supported elsewhere in training but new to this user. Two independently fitted audit heads with and without purified pair residuals use the same audit data, budget, and held-out full catalog; the audit data are disjoint from SCFR training. Cohort labels do not filter candidates or enter serving-time scoring.

We first test whether this decision contains a pair-specific residual, without using SCFR. Cohort rules are frozen from pre-cutoff data; the audit is fitted on data disjoint from SCFR. On the predeclared H&M and multimodal MovieLens (M3L) anchors, the clean cohorts contain 127,536 (11.5%) and 16,899 (2.0%) evaluation positives across 1,291 and 182 factor-pair cells. Applying the frozen protocol to Amazon Sports gives 84,774 (6.7%) evaluation positives across 855 factor-pair cells as an external replication. With the same audit data and budget, the independently fitted head with purified pair features improves target NDCG@20 over the main-only head by +0.0029 [0.0005, 0.0048] and +0.0039 [0.0007, 0.0071] on the anchors and +0.0035 [0.0006, 0.0051] on Amazon (Table 1). Pair permutations and equal-capacity random controls are near zero. Matched-seen point effects are smaller, although the direct target-minus-matched intervals do not consistently exclude zero. The clean cohorts cover only 2.0%–11.5% of evaluation positives. The audit is predictive evidence for this restricted slice of full-catalog rankings, not an average gain over ordinary warm traffic or a causal effect.

The audit points to two estimation requirements. First, a raw pair feature can repeat lower-order content: "red × jacket" must be separated from red and jacket main effects (Park et al., 2025; Ferrere et al., 2026). Second, pair occurrence alone cannot tell whether training comparisons constrain the target-versus-competitor direction needed for ranking (Chang et al., 2026; Atanasov et al., 2025).

Support-Aware Compositional Factor Residual (SCFR) addresses these requirements in order. It projects pair features off the declared lower-order comparison span, then uses train-only ranking-loss curvature in a regularized fit so weakly constrained directions shrink more. One fitted correction scores every warm-catalog candidate; cohort membership is neither an input nor a serving-time gate. The theory separates identification and local stability from a conditional squared-risk result for directional shrinkage; full-catalog NDCG improvement is tested empirically.

In the predeclared primary test, SCFR improves target NDCG@20 over the strongest matched-input control selected on validation by +0.0074 [0.0015, 0.0081] and +0.0083 [0.0014, 0.0094] on both confirmatory anchors, while passing overall and complement non-inferiority checks against the calibrated frozen backbone. Amazon provides a secondary same-protocol replication of +0.0066 [0.0011, 0.0078]. Second backbones and later cutoffs preserve the target-effect direction; controlled anti-alignment, weak-support, factor-noise, and nonlinear tests show where the advantage vanishes or reverses. We contribute:

1. a past-only warm-start cohort and a disjoint audit of its prevalence, dispersion, pair value, and specificity;
2. a catalog-wide correction combining lower-order separation with curvature-weighted shrinkage;

3. conditional identification, stability, and alignment results tested by matched controls, interventions, and explicit failures.

2 Related Work

Warm-start representation and ranking. Collaborative and sequential methods improve warm user states (He et al., 2020; Kang & McAuley, 2018; Zhang et al., 2025); multimodal methods strengthen item representations and fusion (Zhou & Shen, 2023; Xu et al., 2025b; Li et al., 2026c). A representation can also encode content that a recommender does not use in ranking (Pomo et al., 2025). These works supply strong performance context, but their usual evaluation does not separately measure the value of a globally supported conjunction that is new to one user. We therefore freeze the ranker and control the factor main effects. Appendix C.2 reports the complete same-protocol frontier and documents eligibility for systems with different input or feedback contracts (Jin et al., 2026; Li et al., 2026b;d).

Constrained recommendation and compositional generalization. Set-theoretic recommendation studies explicit attribute constraints (Dasgupta et al., 2025); broader compositional work studies controlled task structures, distribution shifts, or combinations absent from global training support (Mahajan et al., 2025; Chang et al., 2026; Lippl & Stachenfeld, 2025; Liang et al., 2025). Two-stage embedding tests further separate factor accessibility from unseen-combination generalization (Guo et al., 2025). We use these works as neighboring principles and, where stated, as adapted controls—not as native baselines for our task. Our setting keeps the user, item, factors, and pair globally warm. Only the conjunction is new to this user. Evaluation uses the original full-catalog rank rather than an attribute query or a target-specific candidate set.

Interactions and estimability. Explicit-cross networks add conjunction capacity but need not separate it from lower-order content (Wang et al., 2021). Functional ANOVA provides that reference under dependent inputs (Park et al., 2025; Ferrere et al., 2026), while interaction inference and correlated-sample ridge characterize related uncertainty (Little et al., 2025; Atanasov et al., 2025). SCFR puts these ingredients in the ranking comparison space: the pair block is purified against the lower-order comparison span, and loss curvature determines which remaining comparison directions can support a correction. Section 3 next defines the user-local decision and tests for pair value before the correction is introduced.

3 Problem Setup and Pre-Method Audit

Target decision. At cutoff T, let $\mathcal{H}_u(< T)$ be user u's history and $\mathcal{C}_T$ the eligible warm catalog. A held-out positive (u, i, T) enters the *broad* cohort when u and i are warm, every selected factor of i occurred separately in $\mathcal{H}_u(< T)$, and their conjunction did not. The *clean* cohort also requires the pair to meet frozen occurrence and diversity thresholds across other training users, items, or families. H&M uses product type $\times$ color, M3L uses exact-two-genre pairs, and Amazon Sports uses deepest-nonempty-category $\times$ source-namespaced brand/store (H&M Group, 2022; Spillo et al., 2026; Hou et al., 2024). Cohort membership defines the relevance mask for evaluating the same full-catalog rankings and for validation selection. It never changes candidates, correction-training data, or the serving-time score. Appendix B.1 gives the support policy and broad-to-clean funnel. This is an evaluation stratum, not a new recommendation task. H&M and M3L are the two predeclared confirmatory anchors; Amazon applies the definition frozen before final-test inspection as an external replication and does not alter their primary decision.

Factor interface and information control. A cross-fitted head maps each frozen item representation to calibrated factor predictions z_i. A past-only state $q_u(T)$ aggregates the same predictions over $\mathcal{H}_u(< T)$. Metadata may supervise heads inside training folds and define evaluation strata, but SCFR, the audit, and all matched controls receive only predictions from training-only heads (out of fold for training items). This representation-only contract keeps the comparisons controlled; it does not claim that metadata is unavailable in every product system. We therefore report direct metadata separately as a headroom diagnostic. No compared correction receives future events, cohort membership, or an unrestricted ID bypass.

Separately fitted audit. On a split disjoint from SCFR training and selection, $\widehat{f}_m$ contains the calibrated frozen score, prescribed nuisances, and factor-main features; $\widehat{f}_{mp}$ is independently refitted with the same data and budget plus residualized pair features. For held-out full-catalog queries Z and audit-fit data A_f, the query-macro NDCG@20 estimand is

$$H_{\text{int}}^{\text{fit}} = \mathbb{E}[\mathcal{M}_{20}^{\text{qm}}(\widehat{f}_{mp}; Z) - \mathcal{M}_{20}^{\text{qm}}(\widehat{f}_m; Z) \mid A_f]. \tag{1}$$

Matched-seen controls are held-out events whose pair was already seen by that user; they are balanced on activity, time, popularity, factor frequency, item/family attributes, and base margin. Pair permutations, within-label shuffles, and equal-dimensional random control factors test capacity, frequency, and leakage explanations. The audit freezes only the lower-order *specification*; its fitted projection and coefficients are never reused by SCFR. A positive result is predictive evidence for this fitted pair head, not a causal or oracle function-class claim. Section 4 builds a separate correction from correction-training comparisons under the declared feature specification.

4 Support-Aware Compositional Factor Residual

SCFR is a residual correction to a frozen ranker. It is designed around three requirements: isolate what factor main effects do not explain; measure support in the comparison directions used by the loss; and score the full catalog without knowing the evaluation cohort. The deployed score is

$$s(u,i) = b_{\text{cal}}(u,i) + \widetilde{\psi}(u,i)^\top \widehat{\theta}, \qquad \widehat{\theta} = F_\eta \widehat{\beta}, \tag{2}$$

Here b_{cal} is a calibrated frozen-backbone score; $\widetilde{\psi}$ and F_η are the purified feature and support map defined below. Personalization enters through predicted factors and a pre-cutoff user state, not through free per-user coefficients. Figure 2 shows feature construction, training fit, and final catalog scoring.

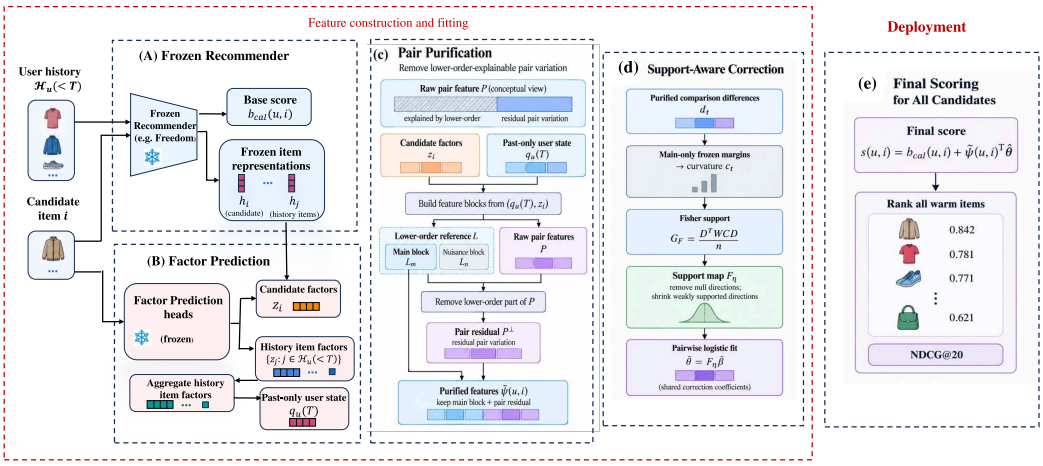

Figure 2: **SCFR data flow.** The frozen ranker supplies item representations and a base score. Predicted factors summarize the user's pre-cutoff history and represent each candidate. Training comparisons fit pair purification and a curvature-weighted shared correction. At scoring time, the fitted transforms construct purified features for each candidate; these combine with the frozen base score and shared correction weights to rank every warm item.

4.1 Past-only factor state and pair purification

For each training item, cross-fitting predicts factor family k from its frozen representation h_i as $z_i^{(k)} = g_{k,\text{OOF}(i)}(h_i)$. Held-out items use an ensemble of the training-only fold heads. Train-fitted IDF weights, normalization N, and projection V_R summarize the user's prefix:

$$n_{u,c}(T) = \sum_{j \in \mathcal{H}_u(<T)} z_{j,c}, \qquad q_u(T) = N(\log(1 + n_u(T)) \odot \text{idf}) V_R. \tag{3}$$

We build two fixed dictionaries from (q_u, z_i): $L_m(u,i)$ contains factor-main coordinates, and $P(u,i)$ contains raw factor-pair coordinates. The lower-order reference $L = [L_m; L_n]$ also includes predeclared past-only nuisance terms. Neither dictionary uses the held-out cohort label. On a correction-training reference set, each row of D_L and D_P is the positive-minus-negative difference of the corresponding features, and W_0 contains frozen sampling weights. We fit

$$B = (D_L^\top W_0 D_L)^+ D_L^\top W_0 D_P, \quad P_\perp = P - B^\top L, \quad \widetilde{\psi} = N_{\mathrm{blk}}[L_m; P_\perp]. \tag{4}$$

Here N_{blk} is fitted after residualization. The comparison residual is $D_{P_\perp} = D_P - D_L B$: a red–jacket coordinate survives only if its pairwise variation is not already explained by red, jacket, and the declared nuisances. This is a statistical decomposition, not a causal or unique semantic one (Park et al., 2025; Ferrere et al., 2026). The audit declares the lower-order specification, but SCFR refits B on correction training; no audit coefficient or outcome enters the method.

Proposition 1 (Lower-order separation)**.** *With the exact Moore–Penrose inverse in Equation (4), let $D_{P_\perp} = D_P - D_L B$. On the projection reference set, $D_L^\top W_0 D_{P_\perp} = 0$. Moreover, if a pair-feature direction $D_P a$ lies in the lower-order column span on the positive-weight rows, then $D_{P_\perp} a = 0$ there.*

Thus the residual block contains only comparison variation left after the declared main and nuisance span in the exact reference projection. A rank-truncated numerical inverse guarantees separation only from its retained lower-order singular subspace; Appendix A.2 states the corresponding residual explicitly. We test, rather than assume, that this reference decomposition remains useful at later cutoffs.

4.2 Support from the loss, not from occurrence counts

For a correction-training comparison $t = (u_t, i_t^+, i_t^-)$, define $d_t = \widetilde{\psi}(u_t, i_t^+) - \widetilde{\psi}(u_t, i_t^-)$ and let m_t be the frozen-base margin. A main-only correction fitted on correction training supplies a preliminary margin $\bar{m}_t$. We freeze it before fitting any pair coefficient, so the pair model cannot alter the geometry used to regularize itself. Pairwise logistic curvature is $c_t = \sigma(\bar{m}_t)\sigma(-\bar{m}_t)$. Using the same sampler and fixed weights W as the objective, let D stack the row vectors $d_t^\top$. SCFR constructs

$$X = C^{1/2} W^{1/2} D, \quad G_{\mathrm{F}} = X^\top X / n = V \,\mathrm{diag}(g_j) V^\top, \qquad F_\eta = V \,\mathrm{diag}\left(\sqrt{\frac{g_j}{g_j + \eta}}\right) V^\top. \tag{5}$$

Here $C = \mathrm{diag}(c_t)$ uses a train-frozen numerical floor. A comparison with a saturated preliminary margin carries little local curvature even if its factor pair occurs many times. Raw cell counts ignore which features differ between the positive and negative items; an unweighted Gram keeps that direction but treats saturated and uncertain margins equally. G_{F} keeps both comparison direction and local loss curvature, and exact null directions map to zero. SCFR then solves

$$\widehat{\beta} = \arg\min_\beta \frac{1}{n} \sum_t w_t \log\left(1 + e^{-(m_t + d_t^\top F_\eta \beta)}\right) + \lambda_2 \|\beta\|_2^2. \tag{6}$$

For $g_j > 0$, the equivalent penalty on the deployed coefficient is $\lambda_2 \theta^\top (I + \eta G_{\mathrm{F}}^+)\theta$ with $\theta \in \mathrm{range}(G_{\mathrm{F}})$. Equivalently, direction j receives weight $(g_j + \eta)/g_j$. Exact null directions are excluded, weakly supported directions are shrunk most, and $\eta = 0$ recovers isotropic ridge on the supported subspace. This is why three superficially similar controls test different choices: count shrinkage discards comparison direction; $C = I$ discards loss curvature; and $\eta = 0$ discards support-dependent shrinkage.

The two operations are complementary. Purification identifies comparison variation beyond the declared lower-order reference. Fisher geometry measures support for directions in the resulting main-and-residual dictionary. Either operation alone leaves one of these questions unanswered.

4.3 Fit once, rank every candidate

Training and selection have four stages. *(1) Interface:* cross-fit item factors and build each user state from the prefix only. *(2) Decomposition:* fit B and block normalization on correction-training comparisons. *(3) Geometry and estimation:* freeze the correction-training main-only margins,

construct G_{F} from the objective's sampler and weights, and fit candidate corrections. *(4) Selection:* choose the configuration on V_{sel}, apply the ordinary-traffic guardrails once on V_{chk}, and then freeze B, N_{blk}, F_η, and $\widehat{\beta}$. At serving time, Equation (2) is applied to every warm-catalog candidate; neither cohort membership nor a support threshold routes scoring.

With cached factor features, building G_{F} costs $O(nd^2)$ and its one-time eigendecomposition costs $O(d^3)$ for n comparisons and correction dimension d. Deployment adds one $O(d)$ dot product per candidate and no second backbone pass. Algorithmic details, frozen data roles, and numerical conventions are in Appendix A.

4.4 What is determined, stable, and beneficial?

The sampled weights and clipped curvatures are positive, so X is a nonsingular row rescaling of the active comparison design D and has the same row and null spaces. A margin-based loss depends on θ through $D\theta$; we can therefore state identifiability in the Fisher geometry X. For a held-out comparison direction v and a reference correction $\theta^\star$, the target is $\tau_v = v^\top \theta^\star$. Let $P_X = X^+X$, $r(v) = \|(I - P_X)v\|_2$, and $a_{\mathrm{F}}(v) = \sqrt{v^\top G_{\mathrm{F}}^+ v}$. We report the scale-free diagnostics $r_{\mathrm{rel}} = r/(\|v\|_2 + 10^{-12})$ and $a_{\mathrm{rel}} = a_{\mathrm{F}}/(\|v\|_2 + 10^{-12})$ on real data; neither fits SCFR nor defines the cohort.

The direction v is a ranking contrast—what the target must gain relative to a competing item—not a pair-occurrence indicator. Consequently, r_{rel} asks whether training comparisons span that contrast, while a_{rel} asks how strongly the spanned component is supported.

Proposition 2 (Loss-level identifiability)**.** *Every $h \in \mathrm{null}(X)$ leaves all training margins, and hence any margin-based empirical loss, unchanged. Therefore τ_v is determined by the training margins if and only if $v \in \mathrm{row}(X)$.*

Occurrence elsewhere is consequently insufficient: it need not place the target comparison in the observed row space.

Corollary 1 (Local logistic stability)**.** *If the penalized logistic Hessian from $P_X\theta^\star$ to a minimum-norm stationary point $\widehat{\theta}$ dominates μG_{F} on $\mathrm{row}(X)$, target error is bounded by a null-space term proportional to $r(v)$ plus a supported estimation term proportional to $a_{\mathrm{F}}(v)/\mu$; Appendix A.9 gives the exact gradient-norm form.*

Identifiability is binary, but stability depends on the strength of support in the target direction.

Identifiability and stability alone do not establish that directional shrinkage helps. In a fixed-design squared-loss analysis, spectral direction j has shrinkage $h_j = g_j^2/(g_j^2 + \lambda g_j + \lambda\eta)$. Suppose transformed noise is homoscedastic and the target comparison covariance commutes with G_{F}; let t_j be its nonnegative weights in a common eigenbasis. At a finite, interior oracle ridge optimum λ_R, define $p_j \propto t_j/(g_j + \lambda_R)^3$ over $g_j > 0$.

Proposition 3 (Local signal–support alignment)**.** *Under these fixed-design, homoscedastic, and common-eigenbasis conditions, if $\mathrm{Cov}_p(g_j, (\theta_j^\star)^2) > 0$, some sufficiently small $\eta > 0$ improves target squared risk over oracle isotropic ridge at λ_R. A negative covariance gives the opposite local derivative.*

The three results answer different questions. Row-space membership determines whether the target is identified; Fisher leverage controls its local sensitivity. The alignment condition has a practical reading: preferentially shrinking weak directions helps when predictive signal lies mainly in strong directions, and can hurt when signal lies in weak ones. The last result is a squared-risk mechanism analysis, not dominance over arbitrary generalized ridge. No theorem turns it into an NDCG guarantee; full-catalog ranking improvement is tested empirically. Proofs, design perturbation, dependent-comparison covariance, and top-K scope appear in Appendix A.

These quantities also determine the experimental tests. A large r_{rel} predicts unresolved directions, a large a_{rel} predicts instability among determined directions, and the alignment proposition predicts a sign reversal under anti-alignment. RQ3 tests the first two on a later cutoff and the third in a controlled intervention; RQ4 exposes the corresponding weak-support, factor-noise, and nonlinear failure regions.

5 EXPERIMENTS

The experiments follow the claim chain through four questions. RQ1 asks whether the cohort contains a pair-specific residual before SCFR is fitted. RQ2 tests target improvement and overall and non-target guardrails. RQ3 tests the proposed support mechanism. RQ4 checks transfer, explicit failures, and cost.

5.1 PROTOCOL AND DECISIONS

Data and evaluation. We use temporal splits of H&M (article_id, product_code, product type×color) and M3L-10M (movieId, exact-two genres, official MPNet-plot + ViT-poster), with Amazon Reviews'23 Sports as an external replication (H&M Group, 2022; Spillo et al., 2026; Hou et al., 2024). H&M forms one unordered, multi-positive query per user–day from the strictly pre-day state. M3L and Amazon map interactions to canonical items, distinguish positive history (rating ≥ 4) from all-rating observed-history exclusion, and use fixed 90-day and 365-day user windows. Every method ranks the same warm catalog (Pereira et al., 2025). Query-macro NDCG@20 (equal weight per query) is primary. Target, overall, and complement (non-target) metrics reuse the same ranks. H&M and M3L must each pass target superiority and both guardrails for the primary claim; Amazon cannot change that decision. Exact keys, deduplication, window anchors, catalogs, and sensitivities are in Appendix B.1.

Fair comparison and inference. We keep six data roles disjoint: base/interface, audit, correction, V_{sel}, V_{chk}, and final test. The published frontier provides performance context across collaborative, graph, sequential, multimodal, and e-commerce recommenders; Appendix C.2 gives the complete frozen pool. The matched-input pool tests the mechanism: factor-main, ridge, MLP, gate, MoE, and refitted main-only heads share SCFR's score, factor interface, state, sampler, and trial budget. DCNv2-head, LambdaMART-head, ANOVA-TPNN-Rank, BoxRank, and CRM-Rank are our task adaptations (Wang et al., 2021; Burges, 2010; Park et al., 2025; Dasgupta et al., 2025; Mahajan et al., 2025). V_{sel} freezes each winner; the strongest non-privileged matched winner defines superiority, and V_{chk} applies both guardrails once. Primary backbones are FREEDOM and MENTOR; Amazon uses MENTOR only for replication. Appendix C.2 gives all models, recent-method eligibility decisions, and second backbones.

At the primary cutoff, effects are paired on queries across five model seeds and use 2,000 multiway cluster bootstraps. Earlier and later cutoffs are reported separately as robustness environments. For both anchors, the two-sided 95% confidence interval for the paired target effect must lie above zero. Overall and complement NI require the one-sided 5th percentile above $-\delta_d$, where δ_d is frozen before test as 0.5% of validation-only base NDCG@20, rounded to 10^{-4}. This rule yields $\delta_{\text{H\&M}} = 0.0004$, $\delta_{\text{M3L}} = 0.0011$, and $\delta_{\text{Amazon}} = 0.0006$. Full clustering, ordered-environment, and multiplicity rules are in the appendix.

5.2 RQ1: COHORT AND SEPARATELY FITTED RESIDUAL

The target cohort is dispersed, and the independently fitted audit detects a positive pair-specific residual on both anchors (Table 1). Target-minus-matched effects are positive in mean but not uniformly separated from zero; permutations and random controls stay near zero. In detail, the two anchor cohorts contain 127,536 (11.5%) and 16,899 (2.0%) evaluation positives across 1,291 and 182 factor-pair cells; the Amazon replication contains 84,774 (6.7%) positives across 855 cells. The full-minus-main effects are +0.0029 [0.0005, 0.0048] and +0.0039 [0.0007, 0.0071] on the anchors, with +0.0035 [0.0006, 0.0051] on the Amazon replication. This supports a pair-specific fitted residual under the declared interface, not universal difficulty or causality.

Table 1: **Past-only cohort and separately fitted audit.** Effects are paired NDCG@20 differences with two-sided 95% intervals. Complete controls and factor/availability audits are in Appendix B.1.

Domain	Clean N (%)	Broad→clean	Full−main	Target−matched	Permuted pair
H&M	127,536 (11.5%)	71.1%	+.0029 [.0005,.0048]	+.0015 [−.0003,.0029]	+.0002 [−.0007,.0011]
M3L	16,899 (2.0%)	61.0%	+.0039 [.0007,.0071]	+.0023 [−.0006,.0049]	−.0004 [−.0018,.0015]
Amazon Sports	84,774 (6.7%)	68.3%	+.0035 [.0006,.0051]	+.0013 [.0004,.0030]	+.0003 [−.0008,.0012]

5.3 RQ2: One rule improves the target and passes guardrails

Table 2 places SCFR beside the highest compatible published rerun (descriptive context) and the strongest matched-input control frozen on V_{sel}. This control shares the base score, factor interface, past-only state, sampler, and trial budget. The winners are the constructed Factor-MoE head, SCFR-main (independently refitted), and the constructed Factor-MoE head on H&M, M3L, and Amazon. Table 3 tests the joint decision: SCFR beats the matched winner and passes both ordinary-traffic guardrails on each confirmatory anchor. Appendix C.2 reports both complete comparison pools.

Table 2: **Published context and matched-control frontier.** Untouched-test target NDCG@20. The published column is the descriptive maximum of compatible reruns (ties listed jointly). The matched control (ours) shares SCFR's frozen base and factor interface and was chosen on V_{sel} before test.

Domain	Frozen backbone	Highest published rerun	V_{sel} matched (ours)	SCFR
H&M	FREEDOM (.0318)	PRISM (.0333)	Factor-MoE-head (.0369)	**.0443**
M3L	MENTOR (.1088)	DHMRec (.1110)	SCFR-main (.1246)	**.1329**
Amazon Sports (rep.)	MENTOR (.0528)	PGL/PRISM (.0536)	Factor-MoE-head (.0549)	**.0615**

Both confirmatory superiority intervals exclude zero, and all four primary NI margins are positive. Relative target gains are 20.1%, 6.7%, and 12.0%; these are target-cohort effects, not aggregate improvements. Amazon overall has paired mean +0.0008 and one-sided LCB −0.000563. The prespecified rounded bound (0.000600) gives NI margin +0.000037; the unrounded bound (0.000553) gives −0.000010. This secondary NI decision is boundary-sensitive; the alternative margin does not establish an overall-score decline. SCFR-minus-refitted-main intervals and absolute scores appear in Appendix C.2.

Table 3: **Primary decision and distribution checks.** Upper: target superiority versus the V_{sel} matched winner; NI versus the frozen backbone, $m_{\text{NI}} = \text{LCB} + \delta$ (positive passes). H&M/M3L are confirmatory; Amazon is a rounding-sensitive replication. Lower: three prespecified paired aggregation views (complete results in Appendix C.2).

Domain	Matched winner	Winner score	SCFR	Δ [95% CI]	Overall m_{NI}	Comp. m_{NI}
H&M	Factor-MoE-head	.0369	.0443	+.0074 [.0015,.0081]	+.0003	+.0002
M3L	SCFR-main	.1246	.1329	+.0083 [.0014,.0094]	+.0008	+.0006
Amazon Sports (rep.)	Factor-MoE-head	.0549	.0615	+.0066 [.0011,.0078]	+.000037	+.0002

Aggregation	H&M effect	M3L effect	Amazon effect
Query-macro	+0.0074 [0.0015, 0.0081]	+0.0083 [0.0014, 0.0094]	+0.0066 [0.0011, 0.0078]
Cell-macro	+0.0041 [0.0018, 0.0068]	+0.0056 [0.0018, 0.0099]	+0.0035 [0.0013, 0.0059]
Validation-locked worst decile	+0.0016 [0.0002, 0.0031]	+0.0019 [0.0003, 0.0037]	+0.0013 [0.0002, 0.0025]

The cell-macro and validation-locked worst-decile effects remain positive, although their magnitudes are smaller than the query-macro effects. All aggregation views are in Appendix C.2.

Table 4 reports the decisive method ablations.

Table 4: **Decisive ablations.** Deltas are from full SCFR on target NDCG@20; the complete table is in Appendix E.

Variant	H&M	Δ	M3L	Δ
Full SCFR	**0.0443**	–	**0.1329**	–
Main block only	0.0368	−0.0075	0.1246	−0.0083
Raw pair, without purification	0.0374	−0.0069	0.1237	−0.0092
Unweighted comparison Gram ($C = I$)	0.0397	−0.0046	0.1274	−0.0055
Purified isotropic ridge ($\eta = 0$)	0.0352	−0.0091	0.1195	−0.0134
Count-based directional shrinkage	0.0369	−0.0074	0.1214	−0.0115

Table 4 shows that raw pairs lose ground when lower-order content remains in the pair block. Isotropic ridge ($\eta = 0$) and count shrinkage also lose ground after purification. Relative to the unweighted comparison Gram, the Fisher-weighted design improves target NDCG@20 by +0.0046 [0.0029, 0.0057], +0.0055 [0.0033, 0.0071], and +0.0047 [0.0023, 0.0062] on H&M, M3L, and Amazon Sports; all three paired intervals exclude zero. The larger purification and shrinkage contrasts support the two design steps. The $C = I$ contrast also changes Gram scale and does not isolate curvature alone; Appendix E gives the full pool.

5.4 RQ3: Conditional mechanism, not universal gain

RQ3 combines a real-data association with a controlled intervention. For later-cutoff comparisons, r_{rel} and a_{rel} are computed against the earlier-cutoff training design. After adjusting for count, popularity, activity, time, factor confidence, and feature norm, they add held-out predictive value for later-cutoff fixed-corrector error. On H&M, standardized coefficients for r_{rel} and a_{rel} are +0.127 [0.044, 0.213] and +0.163 [0.072, 0.256], adding 0.009 and 0.016 held-out R^2; on M3L they are +0.103 [0.007, 0.202] and +0.138 [0.032, 0.247], adding 0.007 and 0.011. Figure 3a shows this association. In the fixed-count intervention, panel b reverses the SCFR–ridge ordering between positive and negative alignment as conditioning worsens.

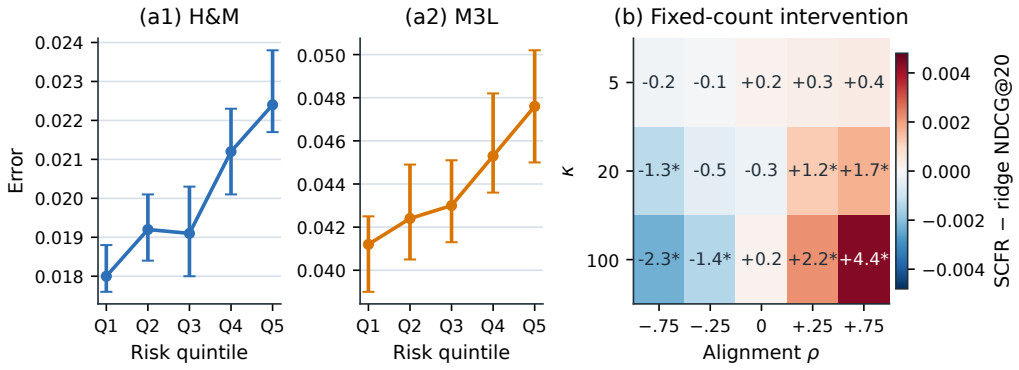

Figure 3: **Conditional support boundary.** (a1–a2) Later-cutoff fixed-corrector error by validation-locked risk quintile (association). (b) Fixed-count SCFR-minus-ridge NDCG@20 (10^{-3} units within cells); stars mark 95% intervals excluding zero across generators.

5.5 RQ4: Transfer, failure regions, and cost

Secondary tests largely preserve the positive target-effect direction. With PRISM as a second backbone, effects are +0.0027 [0.0008, 0.0049] (H&M) and +0.0030 [0.0004, 0.0056] (M3L); later cutoffs, broader catalogs, and a top-K-aligned loss mostly preserve direction (Yang et al., 2025). Some secondary intervals cross zero; Amazon earlier-cutoff complement NI margin is −0.0002.

Failure cases locate the boundary: weak support attenuates gains, and Factor-MLP wins on a nonlinear H&M generator (SCFR minus MLP: −0.0017 [−0.0030, −0.0002]). SCFR adds one train-only eigendecomposition and 5.9% to 7.6% inference overhead relative to full-catalog ranking with cached backbone representations. Full results are in Appendix E.

6 Conclusion and Limitations

This work studies how a globally warm combination that is new to one user ranks within the unchanged warm catalog. A disjoint audit finds held-out value in pair features beyond individual factors. SCFR then separates lower-order variation and shrinks poorly supported comparison directions, improving target ranking on both confirmatory anchors while passing overall and complement guardrails; Amazon replicates the gain direction. The study uses hand-specified factors, and the gains weaken with interface error or nonlinear interactions. The conditional theory does not guarantee NDCG, and online effects remain untested.

AI USE STATEMENT

Generative AI tools assisted literature review, research ideation, method and experiment design, mathematical claims and proof drafts, result interpretation, and manuscript writing, figures, and checking scripts. They did not run the reported experiments or supply their measurements. The authors reviewed the AI-assisted material and take responsibility for the final claims and artifacts.

REPRODUCIBILITY STATEMENT

Sections 3–5 and the appendices describe the method and proofs, data and cohort construction, comparison and selection protocols, reported configurations, uncertainty analysis, and robustness checks. The review submission does not include experimental code; we plan to release the implementation and execution instructions after acceptance.

REFERENCES

Alexander Atanasov, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Risk and cross validation in ridge regression with correlated samples. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 1887–1930, 2025.

Christopher J. C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, 2010.

Hoyeon Chang, Jinho Park, Hanseul Cho, Sohee Yang, Miyoung Ko, Hyeonbin Hwang, Seungpil Won, Dohaeng Lee, Youbin Ahn, and Minjoon Seo. Characterizing pattern matching and its limits on compositional task structures. In *International Conference on Learning Representations*, 2026.

Zhaoqi Chen, Wanni Xu, Yunfeng Zhang, Yawei Hou, Zhenyu Wen, and Cong Wang. DeCoRec: Decoupled collaborative refinement for multi-modal sequential recommendations. In *Proceedings of the ACM International Conference on Multimedia*, pp. 6163–6172, 2025. doi: 10.1145/3746027.3755475.

Shib Sankar Dasgupta, Michael Boratko, and Andrew McCallum. A geometric approach to personalized recommendation with set-theoretic constraints using box embeddings. In *Proceedings of the International Conference on Machine Learning*, 2025.

Baptiste Ferrere, Nicolas Bousquet, Fabrice Gamboa, Jean-Michel Loubes, and Joseph Muré. Exact functional ANOVA decomposition for categorical inputs models. In *Proceedings of the International Conference on Machine Learning*, 2026.

Zhijin Guo, Chenhao Xue, Zhaozhen Xu, Hongbo Bo, Yuxuan Ye, Janet B. Pierrehumbert, and Martha Lewis. Quantifying compositionality of classic and state-of-the-art embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22130–22146, 2025. doi: 10.18653/v1/2025.findings-emnlp.1206.

Ruining He and Julian McAuley. VBPR: Visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

H&M Group. H&M personalized fashion recommendations. Kaggle competition dataset, 2022. URL https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation, 2024.

Liwei Jin, Dan Luo, Lixin Zou, Chenliang Li, Xiangyang Luo, Xixun Lin, Liming Dong, and Yifan Zhang. Scaling collaborative filtering with multimodal contrastive fine-tuning. In *Proceedings of the ACM Web Conference*, pp. 6505–6516, 2026. doi: 10.1145/3774904.3792523.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *Proceedings of the IEEE International Conference on Data Mining*, 2018.

Changhong Li, Zhiqiang Guo, Guohui Li, Zhong Yang, and Chuhang Hong. SGP4SR: Separated-modality guided user preference learning for multimodal sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15054–15062, 2026a. doi: 10.1609/aaai.v40i18.38528.

Haodong Li, Lianyong Qi, Weiming Liu, Fan Wang, Chong Li, Shengye Pang, Wenwen Gong, Yanwei Xu, Xiaoxiao Chi, Yang Zhang, and Xiaokang Zhou. Subspace-aware graph construction and contrastive alignment for multimodal recommendation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15099–15107, 2026b. doi: 10.1609/aaai.v40i18.38533.

Jiuqiang Li and Hongjun Wang. Multi-view semantic contrastive alignment for multimodal recommendation. In *Proceedings of the ACM Web Conference*, pp. 5941–5952, 2026. doi: 10.1145/3774904.3792192.

Yutong Li, Xinyi Zhang, Peijie Sun, Letian Sha, and Zhongxuan Han. PRISM: Personalized recommendation via information synergy module. In *Proceedings of the ACM Web Conference*, pp. 6021–6032, 2026c. doi: 10.1145/3774904.3792230.

Zilong Li, Jia Zhu, Chenglei Huang, Zhangze Chen, Hanghui Guo, Guoqing Ma, and Jianxia Ling. Capturing dynamic user interests under modality imbalance for multimodal sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15198–15206, 2026d. doi: 10.1609/aaai.v40i18.38544.

Yahong Lian, Haotian Tian, Chunyao Song, and Tingjian Ge. SiMGR: Sign-aware multimodal graph recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15225–15233, 2026. doi: 10.1609/aaai.v40i18.38547.

Qiyao Liang, Daoyuan Qian, Liu Ziyin, and Ila R. Fiete. Compositional generalization via forced rendering of disentangled latents. In *Proceedings of the International Conference on Machine Learning*, 2025.

Samuel Lippl and Kimberly Stachenfeld. When does compositional structure yield compositional generalization? a kernel theory. In *International Conference on Learning Representations*, 2025.

Camille Little, Lili Zheng, and Genevera Allen. iLOCO: Distribution-free inference for feature interactions, 2025.

Weihai Lu and Li Yin. DMMD4SR: Diffusion model-based multi-level multimodal denoising for sequential recommendation. In *Proceedings of the ACM International Conference on Multimedia*, pp. 6363–6372, 2025. doi: 10.1145/3746027.3755861.

Divyat Mahajan, Mohammad Pezeshki, Charles Arnal, Ioannis Mitliagkas, Kartik Ahuja, and Pascal Vincent. Compositional risk minimization. In *Proceedings of the International Conference on Machine Learning*, 2025.

Seokhun Park, Insung Kong, Yongchan Choi, Chanmoo Park, and Yongdai Kim. Tensor product neural networks for functional ANOVA model. In *Proceedings of the International Conference on Machine Learning*, 2025.

Bruno L. Pereira, Alan Said, and Rodrygo L. T. Santos. On the reliability of sampling strategies in offline recommender evaluation. In *Proceedings of the ACM Conference on Recommender Systems*, 2025. doi: 10.1145/3705328.3748086.

Claudio Pomo, Matteo Attimonelli, Danilo Danese, Fedelucio Narducci, and Tommaso Di Noia. Do recommender systems really leverage multimodal content? a comprehensive analysis on multimodal representations for recommendation, 2025. Accepted as a full research paper at CIKM 2025.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2009.

Giuseppe Spillo, Alessandro Petruzzelli, Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. Binge watch: Reproducible multimodal benchmark datasets for large-scale movie recommendation on MovieLens-10M and 20M. In *Proceedings of the ACM Conference on Recommender Systems*, 2026. doi: 10.1145/3773078.3831852.

Gu Tang, Jinghe Wang, Bo Jiang, Ze Zhao, Jianping Zhou, Xiaoying Gan, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. TargetMR: Learning modality target for multimodal recommendation. In *Proceedings of the ACM Web Conference*, pp. 6457–6468, 2026. doi: 10.1145/3774904.3792466.

Ruoxi Wang, Rakesh Shivanna, Derek Z. Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed H. Chi. DCN V2: Improved deep and cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference*, 2021.

Cong Xu, Yunhang He, Jun Wang, and Wei Zhang. STAIR: Manipulating collaborative and multimodal information for e-commerce recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12899–12907, 2025a. doi: 10.1609/aaai.v39i12.33407.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith C.-H. Ngai. MENTOR: Multi-level self-supervised learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12908–12917, 2025b. doi: 10.1609/aaai.v39i12.33408.

Weiqin Yang, Jiawei Chen, Shengjia Zhang, Peng Wu, Yuegang Sun, Yan Feng, Chun Chen, and Can Wang. Breaking the top-k barrier: Advancing top-k ranking metrics optimization in recommender systems. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3542–3552, 2025. doi: 10.1145/3711896.3736866.

Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Mind individual information! principal graph learning for multimedia recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13096–13105, 2025. doi: 10.1609/aaai.v39i12.33429.

Xiaohan Zhan, Yuliang Shi, Jihu Wang, Shijun Liu, Fanyu Kong, and Zhiyong Chen. DHMRec: Collaboration-guided multimodal disentanglement and hierarchical fusion for recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16244–16252, 2026. doi: 10.1609/aaai.v40i19.38661.

Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. LATTICE: Latent structure mining for multimodal recommendation. In *Proceedings of the ACM International Conference on Multimedia*, 2021.

Shengzhe Zhang, Liyi Chen, Dazhong Shen, Chao Wang, and Hui Xiong. Hierarchical time-aware mixture of experts for multi-modal sequential recommendation. In *Proceedings of the ACM Web Conference*, pp. 3672–3682, 2025. doi: 10.1145/3696410.3714676.

Xin Zhou and Zhiqi Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the ACM International Conference on Multimedia*, 2023.

Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference*, pp. 845–854, 2023. doi: 10.1145/3543507.3583251.

A METHOD AND THEORY DETAILS

Supplement roadmap. This appendix expands the construction and theoretical scope of Section 4. Appendix B corresponds to the setting and RQ1; Appendix C to the comparison contract and RQ2; Appendix D to RQ3; Appendix E to RQ4; and Appendix F summarizes reproducibility information and maps every assumption to a falsification check. The main paper contains every result needed for its primary conclusions; the supplement provides definitions, proofs, complete comparison pools, stress tests, and negative evidence.

A.1 DATA ROLES AND FROZEN MAPS

The implementation uses six non-interchangeable temporal roles: base/interface training, audit fitting, correction training, V_{sel}, V_{chk}, and final test. A map fitted in one role is frozen before it is applied in a later role. In particular, the backbone, factor heads, calibrators, IDF weights, state projection, lower-order projection, block normalization, curvature-weighted comparison Gram matrix, and correction coefficients are never re-estimated from V_{chk} or the final test. The audit and correction roles share only the declared feature specification: their fitted heads, projection matrices, coefficients, and sampled comparisons are distinct. Thus the separately fitted audit cannot pretrain or tune SCFR. Within the audit role, the fit partition A_f and held-out evaluation queries Z are also disjoint; Table 5 groups them because neither partition supplies a fitted object to the correction role.

Table 5: Non-interchangeable data roles and their permitted influence. A later role may consume a frozen output from an earlier role, but no decision flows backward.

Role	Permitted fitting or decision	Frozen output
Base/interface	Backbone, factor heads, calibration, state map	Scores, out-of-fold factors, interface transforms
Audit fit/eval	Main-only and main-plus-pair diagnostic heads	Audit effects and specificity controls only
Correction train	Projection, normalization, comparison geometry, model coefficients	One candidate model per configuration
V_{sel}	Hyperparameters and one winner per family	Family winners and direct-comparator identity
V_{chk}	One-shot overall and complement guardrails	Recorded pass/fail outcome; no fallback
Final test	No fitting or selection	Tables, figures, and paired uncertainty estimates

Items used to supervise a factor head are partitioned by item-side blocks and time. These cross-fitting blocks do not define the H&M-specific product-family eligibility rule. Each training item is represented by an out-of-fold prediction from a head that excludes its block. Validation and test items are scored by the ensemble of training-only fold heads. The same stored prediction is used whenever an item appears in a history or as a candidate. This convention avoids a candidate-side oracle while keeping the serving interface deterministic.

A.2 LOWER-ORDER SEPARATION AND NUMERICAL CONVENTION

Let $D_L \in \mathbb{R}^{n_0 \times d_L}$ and $D_P \in \mathbb{R}^{n_0 \times d_P}$ contain lower-order and pair-feature differences on the frozen projection reference set. For fixed diagonal $W_0 \succeq 0$, define

$$\begin{aligned} B &= (D_L^\top W_0 D_L)^+ D_L^\top W_0 D_P, \\ P_\perp(u,i) &= P(u,i) - B^\top L(u,i). \end{aligned} \tag{7}$$

For the exact Moore–Penrose inverse, put $A = W_0^{1/2} D_L$ and $Z = W_0^{1/2} D_P$. Then $B = A^+ Z$ is the minimum-Frobenius-norm weighted least-squares solution and the weighted residual is $Z - AB = (I - AA^+)Z$. In numerical computation a relative singular-value tolerance is selected on training data and frozen. If U_r contains the retained left singular vectors, the truncated solution instead has

weighted residual $(I - U_r U_r^\top)Z$: it is orthogonal to the *retained* subspace, but need not be orthogonal to the full column span of A. The statement concerns pairwise differences under the reference measure; it neither identifies causal interactions nor assigns a unique semantic decomposition under dependent support.

Proof of Proposition 1. For the exact inverse, $A^\top(I - AA^+)Z = 0$, which is the displayed normal equation $D_L^\top W_0 D_{P_\perp} = 0$. If $D_P a = D_L b$ on the positive-weight rows, then Za lies in $\mathrm{col}(A)$ and $(I - AA^+)Za = 0$. Every positive-weight entry of $D_{P_\perp} a$ is therefore zero. For the truncated inverse, these two conclusions hold with the *retained* singular subspace substituted for $\mathrm{col}(A)$; a lower-order direction carried only by discarded singular vectors may remain. More precisely, if $A = U_r \Sigma_r V_r^\top + A_{\mathrm{disc}}$, then $A^\top(Z - AB_\tau) = A_{\mathrm{disc}}^\top Z$, which states the possible full-span normal-equation error without pretending it is exactly zero. Distribution shift can additionally weaken out-of-reference relevance and is addressed by later-cutoff and alternative-sampler checks.

Block normalization is fitted after residualization. It standardizes the main and residual-pair blocks separately and never rotates one block into the other. All mechanism-matched spectral controls receive the identical normalized dictionary. Flexible same-information controls instead receive the same raw predicted factors and user state, so they retain their native internal transforms rather than being forced into SCFR's basis. The spectral map acts on the joint normalized main and residual-pair dictionary. Its eigenvectors can mix the two blocks when the weighted comparison Gram has cross-block terms. Thus the projection proposition describes the reference residual before spectral shrinkage; it does not assert that the final coefficient is a uniquely attributable pair component.

A.3 Equivalent penalty on the supported subspace

The spectral map is not an extra feature transform whose effect disappears after reparameterization; together with the isotropic penalty on β, it defines a direction-dependent penalty on the deployed coefficient θ.

Lemma 1 (Penalty equivalence)**.** *Let $G = V \,\mathrm{diag}(g_j) V^\top$ be positive semidefinite and $F_\eta = V \,\mathrm{diag}(\sqrt{g_j/(g_j+\eta)})V^\top$ for $\eta > 0$. Under $\theta = F_\eta \beta$, optimization over β is equivalent to optimization over $\theta \in \mathrm{range}(G)$ with penalty*

$$\lambda_2 \sum_{j: g_j > 0} \frac{g_j + \eta}{g_j} \theta_j^2, \qquad \theta_j = v_j^\top \theta, \tag{8}$$

and the constraint $\theta_j = 0$ whenever $g_j = 0$.

Proof. For $g_j > 0$, the change of variables gives $\beta_j = \sqrt{(g_j + \eta)/g_j}\,\theta_j$; for $g_j = 0$, the corresponding coordinate of $F_\eta \beta$ is identically zero. Substitution into $\lambda_2 \|\beta\|_2^2$ proves Equation (8). Equivalently, the penalty is $\lambda_2 \theta^\top (I + \eta G^+)\theta$ on $\mathrm{range}(G)$. Thus exact null directions are excluded and weakly supported directions receive larger shrinkage than strong ones. At $\eta = 0$, the continuous extension is isotropic ridge restricted to the supported subspace; positive η adds the inverse-support term.

A.4 Curvature-weighted comparison support

The main method uses a train-only local-curvature design rather than raw row counts. Let $\bar{m}_t$ be the margin from the frozen backbone plus the correction-training main-only fit and let $c_t = \sigma(\bar{m}_t)\sigma(-\bar{m}_t)$. After applying a numerical floor $c_{\min}$ chosen without final-test access, define

$$X = C^{1/2} W^{1/2} D, \qquad G_{\mathrm{F}} = X^\top X / n, \qquad C = \mathrm{diag}(c_t). \tag{9}$$

For the pairwise logistic loss, $D^\top W C D / n$ is the empirical Fisher block at the frozen preliminary margin. Rows with saturated margins carry less local curvature even when their occurrence counts are large. Setting $C = I$ recovers the unweighted comparison-Gram ablation; replacing the spectrum by cell counts gives the count-shrinkage control. The preliminary fit, C, G_{F}, and its eigenspace are frozen before pair-correction fitting. This ordering is substantive: if the pair head set its own curvature, it could reshape the support measure used to penalize itself, making the geometry circular and the $C = I$ comparison uninterpretable.

A.5 Training and selection procedure

For each domain and frozen backbone, the procedure is:

1. construct out-of-fold factor predictions and the past-only user state;
2. fit the lower-order projection and block normalization on training-only comparisons;
3. fit the correction-training main-only offset, freeze its curvature weights, and construct $X = C^{1/2}W^{1/2}D$ and $G_{\mathrm{F}} = X^\top X/n$ from the same sampler and weights used by the objective;
4. for every declared method family, train the pre-specified candidate configurations with an equal validation-trial allowance;
5. select one family winner on V_{sel} using target NDCG@20 and the frozen tie-break rule;
6. identify the strongest non-oracle direct family winner, including the independently refitted main-only family;
7. evaluate each selected model once on V_{chk}; record rather than repair a failed warm-start check; and
8. freeze every identity and configuration before producing the final-test export used by all tables and prose.

A.6 Stability to an estimated comparison design

The comparison Gram matrix is empirical. The following deterministic result quantifies the sensitivity of its squared spectral filter; it does not remove bias from a misspecified factor interface or sampler.

Lemma 2 (Design perturbation)**.** *For positive semidefinite $G, \widehat{G}$ and $\eta > 0$, define $S_\eta(G) = G(G+\eta I)^{-1} = F_\eta(G)^2$. Then*

$$\|S_\eta(\widehat{G}) - S_\eta(G)\|_2 \le \frac{\|\widehat{G} - G\|_2}{\eta}. \tag{10}$$

Proof. Because $S_\eta(G) = I - \eta(G+\eta I)^{-1}$, the resolvent identity gives

$$S_\eta(\widehat{G}) - S_\eta(G) = \eta(\widehat{G}+\eta I)^{-1}(\widehat{G} - G)(G+\eta I)^{-1}.$$

Both inverse norms are at most $1/\eta$, which yields the bound. The $1/\eta$ dependence makes the bias–stability role of η explicit and motivates the sampler and factor-noise sensitivities in Appendix E.

A.7 Proof of target-comparison identifiability

Write $P_X = X^+X$, the orthogonal projector onto $\mathrm{row}(X)$. If $v \notin \mathrm{row}(X)$, then $v_0 = (I - P_X)v \neq 0$. For any scalar c, $X(\theta^\star + cv_0) = X\theta^\star$, so the two coefficients produce identical training margins under squared, logistic, BPR, and any other margin-based empirical loss. Yet $v^\top(\theta^\star + cv_0) = v^\top\theta^\star + c\|v_0\|^2$, so the target is not determined. Conversely, if $v = X^\top a$, then $v^\top\theta^\star = a^\top X\theta^\star$ and is determined by the margins. This proves Proposition 2 without a squared-loss assumption.

A fixed-design concentration bound also gives a useful special case. If $\widehat{\theta} = X^+\widetilde{y}$ and $\widetilde{y} = X\theta^\star + \varepsilon$, then

$$v^\top(\widehat{\theta} - \theta^\star) = v^\top X^+\varepsilon - v^\top(I - P_X)\theta^\star. \tag{11}$$

The second term is bounded by $\|(I - P_X)v\|\,\|\theta^\star\|$. The first is a weighted sum of conditionally independent σ-sub-Gaussian coordinates with squared coefficient norm $v^\top(X^\top X)^+v$. Thus, with probability at least $1 - \delta$,

$$|v^\top(\widehat{\theta} - \theta^\star)| \le r(v)\|\theta^\star\|_2 + \sigma\sqrt{v^\top(X^\top X)^+v}\sqrt{2\log(2/\delta)}. \tag{12}$$

The result follows from the standard sub-Gaussian tail inequality and a union over the two tails.

A.8 Derivative behind the alignment proposition

Condition on a fixed transformed design X with $G_{\mathrm{F}} = X^\top X/n$ and homoscedastic independent noise of variance σ^2 in $y = X\theta^\star + \varepsilon$. For a target comparison v, write $G_{\mathrm{tar}} = \mathbb{E}[vv^\top]$. Assume G_{tar} and G_{F} can be simultaneously orthogonally diagonalized, and let $t_j \geq 0$ be the resulting target weights. With $\nu = \sigma^2/n$, the target squared risk of the support-restricted estimator is

$$R(\lambda, \eta) = \sum_{j:g_j>0} t_j \left[(1-h_j)^2(\theta_j^\star)^2 + \frac{\nu h_j^2}{g_j} \right] + R_{\mathrm{null}}, \qquad R_{\mathrm{null}} = \sum_{j:g_j=0} t_j(\theta_j^\star)^2. \tag{13}$$

The null term is independent of η. In general, off-diagonal target covariances generate cross-bias terms, and dependent or heteroscedastic noise changes the variance term; the covariance criterion below does not apply without these assumptions. At $\eta = 0$, $h_j = g_j/(g_j+\lambda)$ and $\partial h_j/\partial\eta = -\lambda/(g_j+\lambda)^2$. Differentiating the target risk gives

$$\left.\frac{\partial R}{\partial \eta}\right|_{(\lambda,0)} = -2\lambda \sum_j \frac{t_j}{(g_j+\lambda)^3} \left(\nu - \lambda(\theta_j^\star)^2\right), \tag{14}$$

At a finite interior minimizer λ_R of the ridge risk, and with at least one target weight $t_j > 0$ on $g_j > 0$,

$$\sum_j \frac{t_j g_j}{(g_j+\lambda_R)^3} \left(\nu - \lambda_R(\theta_j^\star)^2\right) = 0. \tag{15}$$

Put $a_j = t_j/(g_j+\lambda_R)^3$ and $p_j = a_j/\sum_k a_k$. Substituting Equation (15) into Equation (14) gives the exact identity

$$\left.\frac{\partial R}{\partial \eta}\right|_{(\lambda_R,0^+)} = -\frac{2\lambda_R^2 \sum_j a_j}{\mathbb{E}_p[g_j]} \operatorname{Cov}_p\!\left(g_j, (\theta_j^\star)^2\right). \tag{16}$$

Here $\mathbb{E}_p[g_j] > 0$. Positive covariance implies a negative derivative and improvement for some sufficiently small $\eta > 0$ over the oracle ridge risk; negative covariance yields local deterioration at λ_R, while zero covariance yields no first-order ordering. This is a conditional squared- risk statement, not dominance over arbitrary generalized ridge, a logistic guarantee, or an NDCG guarantee.

A.9 Logistic objective, dependent comparisons, and ranking scope

Let $S = \mathrm{row}(X)$, P_X be its orthogonal projector, and $\theta_R^\star = P_X\theta^\star$. Write $Q_S(\vartheta) = L_{\log}(\vartheta) + \lambda\vartheta^\top R\vartheta$ restricted to $\vartheta \in S$, where R is the induced SCFR penalty. Let $\widehat{\theta} \in S$ be a stationary point and define the projected gradient $g_R = P_X\nabla L_{\log}(\theta_R^\star) + 2\lambda R\theta_R^\star \in S$. By the integral mean-value identity *within* S,

$$\bar{H}_S(\widehat{\theta} - \theta_R^\star) = -g_R, \qquad \bar{H}_S = \int_0^1 P_X\nabla^2 Q_S(\theta_R^\star + t(\widehat{\theta} - \theta_R^\star))P_X\, dt \,\big|_S. \tag{17}$$

If $\bar{H}_S \succeq \mu G_{\mathrm{F}}$ on S, Cauchy–Schwarz in the G_{F} geometry gives

$$|v^\top(\widehat{\theta} - \theta^\star)| \leq r(v)\|\theta^\star\|_2 + \frac{\sqrt{v^\top G_{\mathrm{F}}^+ v}}{\mu}\|(G_{\mathrm{F}}^+)^{1/2} g_R\|_2. \tag{18}$$

This proves Corollary 1. The first term is non-identifiability; the second separates target leverage, local curvature, stochastic score error, and penalty bias. It is a local statement, not a global logistic or NDCG guarantee.

For clustered comparisons, let H be the population penalized Hessian and J the covariance of clustered score contributions. Under standard local regularity and fixed-feature conditions, the correction has sandwich covariance $H^{-1}JH^{-1}$. User, item-side, and temporal dependence enter J through multiway cluster aggregation; they are not treated as independent pair rows. This distinction is also central to ridge analysis with correlated samples (Atanasov et al., 2025).

Cross-fitted predicted factors introduce measurement error. The supplement's direct-metadata-versus-predicted and controlled-noise experiments assess this interface gap rather than assuming it away. Finally, reduced squared or logistic pairwise risk does not imply increased NDCG. The primary evidence is therefore full-catalog NDCG, and a top-K-aligned objective is used only as a robustness check. No theoretical statement in the paper guarantees NDCG improvement.

A.10 A CONDITIONAL BRIDGE TO TOP-K RANKING

The fixed-design results concern comparison scores, whereas the evaluation is top-K ranking. A margin condition supplies only a conditional bridge between them.

Proposition 4 (Top-K set stability)**.** *For query q, let $s_q^\star(i)$ be ideal catalog scores and let $\gamma_q = s_{q,(K)}^\star - s_{q,(K+1)}^\star > 0$ be the boundary gap. If $\max_{i \in \mathcal{C}_T} |\widehat{s}_q(i) - s_q^\star(i)| < \gamma_q/2$, then $\widehat{s}_q$ and $s_q^\star$ select the same top-K item set.*

Proof. For every ideal top-K item i and every excluded item j, $s_q^\star(i) - s_q^\star(j) \geq \gamma_q$. Perturbing both scores by less than $\gamma_q/2$ preserves the strict ordering across the boundary.

The proposition does not preserve order within the selected set. Equality of graded NDCG additionally requires analogous pairwise margins for items with different relevance gains. Because neither margin condition is assumed uniformly, NDCG improvement remains an empirical claim rather than a theorem (Yang et al., 2025).

A.11 COMPUTATIONAL COMPLEXITY

Let n be the number of sampled training comparisons and d the normalized correction dimension. Given cached factor features, building comparison rows costs $O(nd)$, forming G_{F} costs $O(nd^2)$, and its one-time eigendecomposition costs $O(d^3)$. Each dense optimization epoch costs $O(nd)$; streamed rows require $O(d^2)$ geometry memory in addition to model and minibatch storage. At inference, SCFR adds one $O(d)$ dot product per candidate to cached backbone scores. It does not alter the eligible catalog, invoke a target-specific gate, or require a second backbone pass.

B DATA, COHORT, AND AUDIT PROTOCOL

B.1 DATASET, COHORT, AND CONCENTRATION DETAILS

Table 6 reports the frozen primary-cutoff scale. "Broad $\rightarrow$ clean" applies every validity, availability, and occurrence–diversity filter without consulting an audit or SCFR outcome. Broad and clean counts come from one logged cohort builder, and percentages are recomputed from those counts rather than entered independently.

Frozen support policy. The two confirmatory releases are H&M Personalized Fashion Recommendations and M3L-10M; Amazon Reviews'23 Sports_and_Outdoors is a secondary, same-protocol external replication and cannot change the two-anchor decision (H&M Group, 2022; Spillo et al., 2026; Hou et al., 2024). The following rules are applied before any audit or SCFR outcome is read.

For H&M, an item is article_id, its variant family is product_code, and the ordered factor schema is product_type_no $\times$ colour_group_code; display names are never used as join keys. All events for one user on one calendar day form an unordered multi-positive query. Every candidate on that day is scored from the state ending strictly before the day, repeated user–article events collapse to one relevance label, and the state is updated only after the whole query. A target is excluded if its article, factor pair, or product-code family already appears in that user's prefix. Support witnesses also exclude the target product-code family.

For M3L, the release is exactly M3L-10M, the item key is canonical movieId, and the main cohort contains movies with exactly two genres, represented by one unordered genre pair. The main representation uses the release's MPNet plot and ViT poster features; audio and trailer features are outside the main input budget. Ratings ≥ 4 define positive history and relevance. Crucially, every earlier rating, including a lower rating, enters the observed-history exclusion for the movie and all its genre pairs; lower ratings are never sampled as negatives. A query is one non-overlapping, epoch-anchored 90-day user window, ranked once from its pre-window state with no within-window update. The all-pairs-unseen multi-genre cohort is a prespecified coverage extension, not a replacement for the exact-two main cohort.

The released M3L-10M table already contains 9,031 movies in the all-modality overlap (and 9,409,884 ratings). Our 6,042 count is the downstream protocol-eligible subset after the temporal, warmness, exact-two-genre, and support filters; it is not a claimed modality-overlap count.

For Amazon, Unicode text is NFKC-normalized, trimmed, and whitespace-collapsed. Each review asin is first mapped to parent_asin; all deduplication and history checks happen only after this mapping. The factors are the deepest nonempty category and a source-namespaced seller label: nonempty details[Brand] becomes `brand::value`, otherwise nonempty store becomes `store::value`; namespaces are never merged. An item missing either category or both seller fields stays in the ranked catalog but is ineligible as a target/support witness and receives no factor correction. Within a user window, all reviews mapping to one parent_asin collapse to one label, positive iff any rating is ≥ 4; lower ratings remain observed-history events, never negatives. A query is one non-overlapping, epoch-anchored 365-day user window. The reported replication does not filter on verified-purchase status, and the verified-purchase flag is not used to define any reported cohort.

In every domain, the broad cohort requires a warm user and item, separately user-familiar constituents, and a conjunction absent from the complete pre-query observed history. The clean cohort additionally requires $n_{\text{event}}, n_{\text{user}}, n_{\text{item}} \geq h$ with validation-locked $h = 8$, at least three temporal blocks, and no single user or item contributing more than 25% of witness mass. H&M additionally requires $n_{\text{family}} \geq h$ and applies the same cap to product-code families; we do not invent an analogue of product family for M3L or Amazon. Candidate catalogs, feature availability, factor vocabularies, ties, window anchors, and cutoffs are fixed before final-test evaluation and shared by every compared method.

Table 6: Dataset and target-cohort statistics under the primary cutoff.

Domain	Eval. positives	Eval. user–windows	Warm users	Items
H&M	1,112,291	93,005	71,237	34,527
M3L	854,138	12,947	12,042	6,042
Amazon Sports	1,263,596	58,611	73,384	41,973

Domain	Broad targets	Clean targets	Retention	Clean share	Pair cells
H&M	179,415	127,536	71.1%	11.5%	1,291
M3L	27,682	16,899	61.0%	2.0%	182
Amazon Sports	124,157	84,774	68.3%	6.7%	855

Queries and users are intentionally different units. "Warm users" counts unique eligible users over the frozen prefix and evaluation horizon; "Eval. user–windows" counts query instances pooled across the fixed evaluation windows. One user may therefore contribute multiple queries, while a warm user with no eligible positive in a window contributes none.

Table 7: Absolute clean-cohort coverage. "Window / blocks" gives the fixed evaluation-window width and the number of nonempty clean time blocks. H&M families are product-code families; no analogous field is declared for M3L or Amazon Sports.

Domain	Queries	Users	Items	Families	Window / blocks
H&M	72,805	47,428	22,624	13,228	30 days / 14
M3L	9,361	8,633	3,952	–	90 days / 12
Amazon Sports	44,130	42,205	28,024	–	365 days / 10

Table 8: Complete separately fitted interaction audit. Effects are paired NDCG@20 differences with two-sided 95% intervals. Permuted pairs and random factors preserve capacity while breaking the intended interaction signal.

Domain	Full−main	Target−matched
H&M	+0.0029 [0.0005, 0.0048]	+0.0015 [−0.0003, 0.0029]
M3L	+0.0039 [0.0007, 0.0071]	+0.0023 [−0.0006, 0.0049]
Amazon Sports	+0.0035 [0.0006, 0.0051]	+0.0013 [0.0004, 0.0030]

Domain	Permuted pair	Random controls
H&M	+0.0002 [−0.0007, 0.0011]	+0.0004 [−0.0005, 0.0013]
M3L	−0.0004 [−0.0018, 0.0015]	+0.0005 [−0.0008, 0.0017]
Amazon Sports	+0.0003 [−0.0008, 0.0012]	+0.0001 [−0.0006, 0.0013]

Table 9: One logged support funnel. "Locally unseen" is the broad cohort; "diversified" is the clean cohort. Counts can only decrease from left to right.

Domain	Eligible positives	Factor-familiar	Locally unseen	Globally supported	Diversified
H&M	1,112,291	534,229	179,415	161,428	127,536
M3L	854,138	302,136	27,682	23,404	16,899
Amazon Sports	1,263,596	513,251	124,157	98,941	84,774

Table 10: Clean factor-pair-cell concentration. Parentheses report shares of clean targets, except in the last column, where they report cell ESS as a percentage of active cells.

Domain	Clean targets	Active cells	Largest cell	Largest 10	Cell ESS
H&M	127,536	1,291	2,988 (2.34%)	9,582 (7.51%)	605.21 (46.88%)
M3L	16,899	182	1,321 (7.82%)	3,802 (22.50%)	77.03 (42.32%)
Amazon Sports	84,774	855	2,595 (3.06%)	8,841 (10.43%)	382.25 (44.71%)

For cell counts n_c, we use $\mathrm{ESS}_{\mathrm{cell}} = (\sum_c n_c)^2 / \sum_c n_c^2$ and normalize by the number of active cells. The largest factor-pair cell contains 2.3%, 7.8%, and 3.1% of clean targets on H&M, M3L, and Amazon Sports; the largest ten cells contain 7.5%, 22.5%, and 10.4%. Cell-level effective sample sizes are 46.9%, 42.3%, and 44.7% of active cell counts. User-, item-, and time-block-level effective sample sizes remain at 61%/54%/69%, 58%/51%/66%, and 60%/53%/68% of their corresponding counts. These diagnostics prevent the number of nominal cells from being mistaken for dispersed evidence. All displayed shares are derived from the counts in Tables 6 and 10.

The primary eligible item set contains 6,042 of the 9,031 released M3L-10M movies (66.9%). This denominator is the released M3L-10M item count, not a modality-overlap count: the release already provides the frozen plot/poster feature resources used here. The stagewise funnel separately attributes losses to valid factor labels, complete-prefix history, warm temporal eligibility, and the exact-two-genre definition. We repeat the main comparison with item-level inverse-retention weighting and on the all-pairs-unseen extension. Prefix assertions verify that target factors, combination novelty, support witnesses, and catalog eligibility use no event later than the query start.

Table 11: Anchor-domain cohort-definition and support-transfer sensitivities. Entries are SCFR minus the validation-frozen matched-input comparator on target NDCG@20. The primary row is confirmatory; the remaining rows are descriptive, and intervals that cross zero are retained rather than filtered.

Check	H&M effect [95% CI]	M3L effect [95% CI]
Primary exact-pair protocol	+0.0074 [0.0015, 0.0081]	+0.0083 [0.0014, 0.0094]
Leave target item/family out of support witnesses	Primary by construction	+0.0024 [−0.0001, 0.0044]
Coarse parent ontology	+0.0026 [0.0001, 0.0042]	+0.0019 [−0.0004, 0.0047]
Semantic-neighbor factors	+0.0017 [−0.0009, 0.0036]	+0.0023 [−0.0006, 0.0049]
All-pairs-unseen multi-genre coverage	–	+0.0020 [−0.0003, 0.0048]
Eligible-item inverse-retention weighting	–	+0.0026 [−0.0002, 0.0045]

The support-transfer row removes the evaluated item from every M3L support witness before recomputing eligibility. H&M already excludes the entire target product-code family from support witnesses in the primary protocol, which is strictly stronger than removing only the target article; its cell is therefore "primary by construction." M3L has no declared variant-family key, so its reported cell is item-out rather than an invented family-level analogue. The ontology rows replace exact labels with frozen parent or semantic-neighbor mappings. The M3L all-pairs-unseen row broadens beyond exact-two-genre items but requires every constituent pair to be absent from the user's prefix. The inverse-retention row targets the released-item population before factor/history/warm filters and is not used to redefine the primary estimand.

Table 12: Matched-seen balance audit. Maximum absolute standardized mean difference (SMD) is computed over the declared matching covariates; retention is the fraction of target events with an accepted match.

Domain	Max SMD before	Max SMD after	Retention	Target−matched audit
H&M	0.42	0.06	79.2%	+0.0015 [−0.0003, 0.0029]
M3L	0.37	0.07	74.5%	+0.0023 [−0.0006, 0.0049]
Amazon Sports	0.36	0.05	82.8%	+0.0013 [0.0004, 0.0030]

Table 13: Independently fitted availability-model interaction audit. Entries are target-cohort NDCG@20 increments with paired two-sided 95% intervals.

Domain	Full − main interaction increment
H&M	+0.0031 [0.0005, 0.0059]
M3L	+0.0045 [0.0007, 0.0083]
Amazon Sports	+0.0036 [0.0009, 0.0066]

This availability diagnostic is fitted after the cohort is locked, on a data role disjoint from the base audit and SCFR training. It neither defines cohort membership nor supplies features, weights, or thresholds to SCFR; it tests whether the residual can be explained solely by which item-factor interfaces are observable.

C Factor Interface and Comparison Frontier

C.1 Predicted-factor interface

Table 14: Held-out factor-head quality. ECE uses the training-frozen binning rule. Clean-cohort values are reported separately because interface error can concentrate on the target population.

Head	Overall macro-F1	Overall ECE	Clean macro-F1	Clean ECE
H&M product type	0.782	0.021	0.759	0.034
H&M color	0.746	0.032	0.687	0.041
H&M joint pair	0.679	0.025	0.643	0.038
M3L genre marginal	0.699	0.039	0.674	0.045
M3L pair inclusion	0.649	0.041	0.637	0.044
Amazon leaf category	0.716	0.028	0.715	0.039
Amazon brand/store	0.672	0.037	0.658	0.043
Amazon joint pair	0.730	0.033	0.696	0.038

Macro-F1 is computed over the training-supported vocabulary, and ECE uses the same frozen binning rule for the overall and clean populations. Direct metadata is interpreted as privileged-label headroom under our representation-only contract, not as a mechanism-matched baseline or a deployability claim. Within-label shuffles preserve label frequency while breaking the item-specific interface; equal-dimensional random controls preserve capacity but not semantic alignment.

Table 15: Factor-interface headroom on target NDCG@20. Direct metadata is a privileged-label diagnostic under the declared representation-only contract, not a mechanism-matched baseline; the gap quantifies how much remaining error could plausibly come from factor prediction.

Domain	Predicted-factor SCFR	Direct-metadata SCFR	Headroom
H&M	0.0443	0.0482	+0.0039
M3L	0.1329	0.1358	+0.0029
Amazon Sports	0.0615	0.0638	+0.0023

C.2 Published frontier and complete matched-control pool

Model provenance. The published recommender baselines are Popularity, BPR, LightGCN, SASRec, VBPR, LATTICE, BM3, FREEDOM, MENTOR, PGL, STAIR, HM4SR, DeCoRec, DMMD4SR, DHMRec, MSCA, PRISM, and SGP4SR (Rendle et al., 2009; He et al., 2020; Kang & McAuley, 2018; He & McAuley, 2016; Zhang et al., 2021; Zhou et al., 2023; Zhou & Shen, 2023; Xu et al., 2025b; Zhang et al., 2025; Yu et al., 2025; Xu et al., 2025a; Chen et al., 2025; Lu & Yin, 2025; Zhan et al., 2026; Li & Wang, 2026; Li et al., 2026c;a). They are retrained and re-evaluated here; no score is copied from its source paper. DHMRec is an independent paper-based reimplementation rather than a claimed official-code run. TargetMR (Tang et al., 2026) is cited and registered below but is not assigned a quantitative row because the available artifact does not freeze its required host backbone and native-versus-feature-only input contract. DCNv2-head, LambdaMART-head, ANOVA-TPNN-Rank, BoxRank, and CRM-Rank adapt a published component to our frozen correction interface (Wang et al., 2021; Burges, 2010; Park et al., 2025; Dasgupta et al., 2025; Mahajan et al., 2025); the suffix marks that these are our adaptations, not the original systems. The remaining named correction heads (factor-main, empirical-Bayes, group-DRO, ridge/PCR, raw concatenation, Factor-MLP, Factor-Gate, Factor-MoE, and independently refitted SCFR-main) are constructed controls from this study. This provenance distinction is orthogonal to the inferential role below.

The submission-cutoff eligibility rule is simple: an included method must emit one score for every item in the same eligible catalog, without future information or an expanded input budget. Table 16 records the recent-method decisions. The representative pool and every exclusion are fixed before final-test inspection; publication date alone neither guarantees inclusion nor justifies silent omission.

Table 16: Recent-baseline eligibility record. "Context only" denotes relevant work that cannot enter the common quantitative frontier under the available pre-test contract.

Candidate	Decision	Frozen reason
PGL	Included	Public implementation; single-domain warm item scoring and the same catalog contract.
STAIR	Included	Public implementation; multimodal e-commerce fusion can be retrained without additional test-time inputs.
TargetMR	Context only; no quantitative row	The method is backbone-dependent and its native target-object/selector pipeline changes the input contract. Our pre-test protocol did not freeze its host backbone and preprocessing mode, so we cite it without assigning a retrospectively labelled quantitative result.
DeCoRec	Included	Public implementation; compatible past-only multimodal sequential scoring under the common catalog.
DMMD4SR	Included	Public implementation; diffusion-based sequential scoring emits one score per item under the frozen feature budget.
DHMRec	Included	Independent paper-faithful reimplementation; collaboration-guided multimodal disentanglement and hierarchical fusion satisfy the warm-item, single-domain catalog contract.
SiMGR	Excluded from equal-budget pool	Requires native signed feedback; H&M purchases provide no corresponding negative-feedback channel.
RecCLIP	Excluded from equal-budget pool	Native cross-domain/cold-start scaling assumes heterogeneous auxiliary data beyond the single-domain warm-start contract.
SCALE	Excluded from equal-budget pool	Expands item/user semantics with LLM-derived inputs absent from the frozen feature budget.
DuAF-MAT	Excluded from equal-budget pool	Targets missing-modality and modality-imbalance sequential settings; the complete-modality main protocol does not instantiate that contract.

Eligibility-record citations: Yu et al. (2025); Xu et al. (2025a); Tang et al. (2026); Chen et al. (2025); Lu & Yin (2025); Zhan et al. (2026); Lian et al. (2026); Jin et al. (2026); Li et al. (2026b;d).

Table 17: Comparison contract. The confirmatory mechanism comparison is the matched-control column; published recommenders establish task difficulty but are not used to identify the SCFR mechanism.

Dimension	SCFR and matched controls (ours)	Published frontier
Frozen input	Same base scores, out-of-fold factors, past-only state	Method-native representations and objectives
Ranking task	Same eligible catalog, positives, queries, and full-catalog metrics	Same temporal split and full-catalog evaluation
Training signal	Same comparison sampler, weights, and legal information	Method-native objective; the same frozen text/image features for architectures that consume those modalities
Selection	Equal 30-trial allowance per family on V_{sel}	Same 30-trial cap with native search spaces logged; descriptive context
Serving rule	One score for every candidate; no cohort gate	Same catalog contract; not a mechanism-matched correction
Inferential role	Strongest non-oracle family winner defines superiority comparator	Descriptive performance frontier only

Table 18 places the frozen backbones among published collaborative, sequential, and multimodal recommenders. These numbers answer a performance-context question, not the mechanism question, because representations, objectives, and compute differ. Every included release uses the same temporal split and full-catalog evaluation; methods that cannot satisfy this contract are cited but excluded with a reason in the eligibility record. The compact main-paper published entry is mechanically derived from the complete table as the highest observed score among these 18 compatible reruns and is explicitly descriptive. We do not treat this maximum as confirmatory model selection or as an unqualified state-of-the-art claim over every possible recommender.

Table 18: Published recommender frontier on the two predeclared mechanism anchors. Target and overall columns are NDCG@20. Every displayed score is a same-protocol rerun of one of the 18 contract-compatible methods; no value is copied from a source paper.

Method	H&M target	H&M overall	M3L target	M3L overall
Popularity	0.0180	0.0775	0.0729	0.1883
BPR	0.0268	0.0833	0.0992	0.2145
LightGCN	0.0291	0.0862	0.1079	0.2130
SASRec	0.0310	0.0882	0.1080	0.2146
VBPR	0.0298	0.0869	0.1073	0.2144
LATTICE	0.0320	0.0877	0.1081	0.2158
BM3	0.0316	0.0879	0.1080	0.2156
FREEDOM	0.0318	0.0884	0.1073	0.2162
MENTOR	0.0312	0.0871	0.1088	0.2148
PGL	0.0323	0.0883	0.1103	0.2167
STAIR	0.0329	0.0879	0.1101	0.2160
HM4SR	0.0325	0.0878	0.1099	0.2153
DeCoRec	0.0330	0.0875	0.1101	0.2156
DMMD4SR	0.0331	0.0880	0.1105	0.2159
DHMRec	0.0332	0.0888	0.1110	0.2168
MSCA	0.0327	0.0885	0.1097	0.2152
PRISM	0.0333	0.0885	0.1102	0.2160
SGP4SR	0.0329	0.0881	0.1106	0.2169

Table 19: Amazon Sports external-replication frontier. Values are target and overall full-catalog NDCG@20 under the same temporal split; this table cannot change the two-anchor confirmatory decision. PGL and PRISM tie for the highest target score at the displayed four-decimal precision.

Method	Target	Overall
Popularity	0.0315	0.0845
BPR	0.0439	0.1023
LightGCN	0.0479	0.1068
SASRec	0.0506	0.1090
VBPR	0.0497	0.1082
LATTICE	0.0513	0.1097
BM3	0.0518	0.1104
FREEDOM	0.0513	0.1099
MENTOR	0.0528	0.1105
PGL	0.0536	0.1105
STAIR	0.0530	0.1109
HM4SR	0.0531	0.1104
DeCoRec	0.0526	0.1109
DMMD4SR	0.0534	0.1106
DHMRec	0.0531	0.1108
MSCA	0.0533	0.1103
PRISM	0.0536	0.1109
SGP4SR	0.0533	0.1105

Table 20: Complete information-matched comparison on H&M. Every correction head uses the frozen FREEDOM score. Underlining marks the frozen strongest non-oracle direct comparator; bold marks SCFR.

Method	Target NDCG@20	Target Recall@20	Overall NDCG@20	Complement NDCG@20
Frozen reference				
Calibrated base	0.0318	0.0685	0.0884	0.0893
Constructed matched controls (ours)				
Additive factor-main ridge	0.0332	0.0706	0.0878	0.0893
Empirical-Bayes factor/pair reranker	0.0342	0.0719	0.0870	0.0887
Pseudo-composition group-DRO	0.0337	0.0719	0.0873	0.0895
Purified isotropic ridge	0.0356	0.0746	0.0876	0.0893
PCR/TSVD residual	0.0360	0.0758	0.0872	0.0890
Validation-frozen generalized ridge	0.0366	0.0762	0.0873	0.0891
Raw factor concatenation head	0.0358	0.0751	0.0872	0.0892
Factor-MLP-head	0.0360	0.0757	0.0876	0.0892
Factor-Gate-head	0.0361	0.0760	0.0875	0.0891
Factor-MoE-head	0.0369	0.0772	0.0874	0.0889
Published-component adaptations (ours)				
DCNv2-head	0.0366	0.0767	0.0875	0.0894
LambdaMART-head	0.0359	0.0754	0.0873	0.0890
ANOVA-TPNN-Rank	0.0363	0.0763	0.0874	0.0892
BoxRank	0.0357	0.0751	0.0873	0.0890
CRM-Rank	0.0352	0.0747	0.0872	0.0888
Proposed family				
SCFR-main, independently refitted	0.0368	0.0763	0.0876	0.0893
SCFR	**0.0443**	**0.0832**	0.0889	0.0891

Table 21: Complete information-matched comparison on M3L. Every correction head uses the frozen MENTOR score. Underlining marks the validation-frozen SCFR-main comparator; bold marks SCFR.

Method	Target NDCG@20	Target Recall@20	Overall NDCG@20	Complement NDCG@20
Frozen reference				
Calibrated base	0.1088	0.1846	0.2148	0.2191
Constructed matched controls (ours)				
Additive factor-main ridge	0.1131	0.1901	0.2150	0.2192
Empirical-Bayes factor/pair reranker	0.1159	0.1942	0.2147	0.2188
Pseudo-composition group-DRO	0.1155	0.1936	0.2150	0.2192
Purified isotropic ridge	0.1192	0.1997	0.2149	0.2190
PCR/TSVD residual	0.1203	0.2014	0.2145	0.2186
Validation-frozen generalized ridge	0.1237	0.2075	0.2149	0.2188
Raw factor concatenation head	0.1206	0.2024	0.2145	0.2189
Factor-MLP-head	0.1229	0.2055	0.2151	0.2193
Factor-Gate-head	0.1228	0.2057	0.2145	0.2186
Factor-MoE-head	0.1239	0.2069	0.2150	0.2192
Published-component adaptations (ours)				
DCNv2-head	0.1238	0.2065	0.2146	0.2189
LambdaMART-head	0.1221	0.2041	0.2147	0.2189
ANOVA-TPNN-Rank	0.1234	0.2062	0.2149	0.2190
BoxRank	0.1213	0.2029	0.2144	0.2189
CRM-Rank	0.1209	0.2026	0.2148	0.2191
Proposed family				
SCFR-main, independently refitted	0.1246	0.2089	0.2152	0.2198
SCFR	**0.1329**	**0.2150**	0.2153	0.2194

Table 22: Complete information-matched comparison on Amazon Sports. Every correction head uses the frozen MENTOR score. Underlining marks the frozen strongest non-oracle direct comparator; bold marks SCFR.

Method	Target NDCG@20	Target Recall@20	Overall NDCG@20	Complement NDCG@20
Frozen reference				
Calibrated base	0.0528	0.0996	0.1105	0.1132
Constructed matched controls (ours)				
Additive factor-main ridge	0.0530	0.1009	0.1104	0.1131
Empirical-Bayes factor/pair reranker	0.0533	0.1018	0.1101	0.1129
Pseudo-composition group-DRO	0.0534	0.1012	0.1104	0.1132
Purified isotropic ridge	0.0542	0.1026	0.1103	0.1131
PCR/TSVD residual	0.0543	0.1028	0.1101	0.1129
Validation-frozen generalized ridge	0.0547	0.1035	0.1103	0.1128
Raw factor concatenation head	0.0542	0.1027	0.1103	0.1130
Factor-MLP-head	0.0546	0.1033	0.1103	0.1132
Factor-Gate-head	0.0546	0.1032	0.1103	0.1130
Factor-MoE-head	0.0549	0.1038	0.1102	0.1129
Published-component adaptations (ours)				
DCNv2-head	0.0548	0.1036	0.1103	0.1131
LambdaMART-head	0.0543	0.1029	0.1101	0.1129
ANOVA-TPNN-Rank	0.0548	0.1036	0.1104	0.1133
BoxRank	0.0540	0.1025	0.1103	0.1129
CRM-Rank	0.0539	0.1023	0.1102	0.1129
Proposed family				
SCFR-main, independently refitted	0.0543	0.1031	0.1104	0.1132
SCFR	**0.0615**	**0.1089**	0.1113	0.1135

The final winner audit is mechanical: for each family, the selected V_{sel} configuration is frozen; the largest target NDCG@20 on V_{sel} among all non-oracle direct family winners, including SCFR-main, supplies the comparator identity. Final-test values cannot change that identity. If displayed validation values tie, the unrounded validation metric and frozen tie-break rule determine the winner. This audit prevents the main text from naming a convenient but weaker comparator.

Table 23: Complete decision record. H&M and M3L are the confirmatory intersection–union anchors; Amazon Sports is explicitly an external replication. The upper panel reports target superiority and SCFR minus the independently refitted main-only control; the lower panel reports non-inferiority against the calibrated frozen backbone. A positive $\mathrm{LCB}+\delta$ margin passes. Extra precision is retained for the Amazon overall row because its decision is within 10^{-4} of the boundary.

Domain	Target comparator	Comparator	SCFR	Superiority [95% CI]	SCFR−main
H&M	Factor-MoE-head	0.0369	0.0443	+0.0074 [0.0015, 0.0081]	+0.0075 [0.0048, 0.0085]
M3L	SCFR-main	0.1246	0.1329	+0.0083 [0.0014, 0.0094]	+0.0083 [0.0014, 0.0094]
Amazon Sports (rep.)	Factor-MoE-head	0.0549	0.0615	+0.0066 [0.0011, 0.0078]	+0.0072 [0.0056, 0.0079]

Domain	Population	Effect	One-sided LCB	δ	$\mathrm{LCB}+\delta$
H&M	Overall	+0.0005	−0.0001	0.0004	+0.0003
H&M	Complement	−0.0002	−0.0002	0.0004	+0.0002
M3L	Overall	+0.0005	−0.0003	0.0011	+0.0008
M3L	Complement	+0.0003	−0.0005	0.0011	+0.0006
Amazon Sports (rep.)	Overall	+0.0008	−0.000563	0.000600	+0.000037
Amazon Sports (rep.)	Complement	+0.0003	−0.0004	0.0006	+0.0002

The validation-only frozen-base NDCG@20 values used to set the NI bounds are 0.0887, 0.2149, and 0.1106 for H&M, M3L, and Amazon Sports, respectively. Applying the pre-specified 0.5% rule and rounding the bound to 10^{-4} yields $\delta = 0.0004/0.0011/0.0006$. The Amazon overall LCB is -0.000563, so its strict margin under the predeclared rounded bound is $+0.000037$. Using 0.5% of the displayed validation base without rounding would instead give a bound of 0.000553 and a margin of -0.000010. The reported replication decision therefore depends on the prespecified rounding convention and is not evidence of a comfortable overall NI margin.

Table 24: Across-seed dispersion. Means are first paired within seed; the paired delta is not reconstructed from independently resampled marginal means. These standard deviations are descriptive, while Table 23 gives clustered uncertainty.

Domain	SCFR mean ± sd	Comparator mean ± sd	Paired Δ mean ± sd	Seeds
H&M	0.0443 ± 0.0008	0.0369 ± 0.0009	+0.0074 ± 0.0008	5
M3L	0.1329 ± 0.0019	0.1246 ± 0.0014	+0.0083 ± 0.0012	5
Amazon Sports	0.0615 ± 0.0012	0.0549 ± 0.0008	+0.0066 ± 0.0009	5

Table 25: SCFR-minus-frozen-strongest target NDCG@20 under complementary aggregation rules. Entries are paired effects with two-sided 95% intervals; the worst decile is locked on validation rather than selected on the test.

Aggregation	H&M effect	M3L effect	Amazon effect
Query-macro	+0.0074 [0.0015, 0.0081]	+0.0083 [0.0014, 0.0094]	+0.0066 [0.0011, 0.0078]
Positive-event-weighted	+0.0035 [0.0017, 0.0053]	+0.0036 [0.0011, 0.0062]	+0.0027 [0.0009, 0.0045]
User-macro	+0.0037 [0.0017, 0.0058]	+0.0041 [0.0011, 0.0073]	+0.0029 [0.0009, 0.0049]
Cell-macro	+0.0041 [0.0018, 0.0068]	+0.0056 [0.0018, 0.0099]	+0.0035 [0.0013, 0.0059]
Validation-locked worst decile	+0.0016 [0.0002, 0.0031]	+0.0019 [0.0003, 0.0037]	+0.0013 [0.0002, 0.0025]

D Extended Results and Mechanism Diagnostics

D.1 Comparison-design diagnostics

The real-data outcome is a *fixed-corrector error*. A correction fitted at the earlier cutoff is frozen; at the later cutoff, query q receives

$$e_q = \frac{1}{|\mathcal{D}_q|} \sum_{(i^+,i^-)\in\mathcal{D}_q} \log\bigl(1+\exp[-(m_q + d_q^\top \widehat{\theta}_{\text{early}})]\bigr), \tag{19}$$

where comparisons are sampled from the same full eligible catalog with training-frozen popularity strata. Query summaries are averaged before user/time resampling. The control model contains occurrence count, item popularity, user activity, time, factor confidence, and $\|v\|$. Table 26 reports standardized additions of r_{rel} and a_{rel} defined in Section 4.4. The diagnostics never enter training, cohort selection, or the displayed risk quintiles' validation lock.

For the empirical a_{rel}, take the residual-feature difference v of each held-out positive–negative pair after frozen feature normalization. Diagonalize the correction-training Gram matrix as $G_{\text{F}} = V\,\text{diag}(g_j)V^\top$, let $\mathcal{J}_+$ denote the active spectrum under the training-frozen numerical tolerance, and put $z = V^\top v$. The computed diagnostic is

$$a_{\text{rel}}(v) = \frac{\bigl(\sum_{j\in\mathcal{J}_+} z_j^2/g_j\bigr)^{1/2}}{\|v\|_2 + 10^{-12}}. \tag{20}$$

We compute this value per pair and average it within query before the RQ3 association regression. Larger values indicate greater inverse-Fisher leverage within the retained training eigenspace, hence weaker support there. The separate r_{rel} captures exact out-of-span components, which a_{rel} alone cannot diagnose. Propositions in Section 4.4 use the mathematical Moore–Penrose inverse; this finite-precision diagnostic uses the frozen active spectrum.

Table 26: Held-out association between comparison-design diagnostics and later-cutoff fixed-corrector error. Effects are standardized coefficients with cluster-bootstrap 95% intervals; ΔR^2 is evaluated on held-out folds.

Added diagnostic	H&M effect	H&M ΔR^2	M3L effect	M3L ΔR^2
r_{rel} beyond controls	+0.127 [0.044, 0.213]	+0.009	+0.103 [0.007, 0.202]	+0.007
a_{rel} beyond controls	+0.163 [0.072, 0.256]	+0.016	+0.138 [0.032, 0.247]	+0.011

Table 27: Later-cutoff fixed-corrector error by validation-locked design-risk quintile.

Quintile	H&M error [95% CI]	M3L error [95% CI]
Q1 (lowest risk)	0.0180 [0.0176, 0.0188]	0.0412 [0.0390, 0.0425]
Q2	0.0192 [0.0184, 0.0201]	0.0424 [0.0405, 0.0449]
Q3	0.0191 [0.0180, 0.0203]	0.0430 [0.0413, 0.0451]
Q4	0.0212 [0.0201, 0.0223]	0.0453 [0.0436, 0.0482]
Q5 (highest risk)	0.0224 [0.0217, 0.0238]	0.0476 [0.0450, 0.0502]

D.2 Controlled signal–support intervention

The generator is deliberately separate from the real-data association above. It changes only conditioning and signal–support alignment, so it is the mechanism test for Proposition 3; the real-data diagnostic remains associational. The controlled generator uses dimension $d = 32$, 12,000 training comparisons, 2,000 target queries, 100 candidates per query, and Gaussian noise $\sigma = 0.35$. Across cells it fixes $\mathrm{tr}(G)$, the nonzero interaction-signal norm, target-query count, and noise scale. It varies only the supported-design condition number κ and the rank correlation ρ between signal energy and support eigenvalues. Each cell averages 200 independent generator instances; the table reports the mean SCFR-minus-ridge NDCG@20 difference and a two-sided 95% interval across instances.

Table 28: Controlled SCFR-minus-ridge NDCG@20. Positive values favor SCFR; anti-alignment gives the predicted reversal.

κ	$\rho = -0.75$	$\rho = -0.25$	$\rho = 0$
5	−0.0002 [−0.0008, 0.0005]	−0.0001 [−0.0006, 0.0006]	+0.0002 [−0.0004, 0.0006]
20	−0.0013 [−0.0019, −0.0006]	−0.0005 [−0.0011, 0.0005]	−0.0003 [−0.0007, 0.0006]
100	−0.0023 [−0.0037, −0.0019]	−0.0014 [−0.0021, −0.0008]	+0.0002 [−0.0003, 0.0007]

κ	$\rho = 0.25$	$\rho = 0.75$
5	+0.0003 [−0.0002, 0.0008]	+0.0004 [−0.0003, 0.0008]
20	+0.0012 [0.0003, 0.0019]	+0.0017 [0.0005, 0.0026]
100	+0.0022 [0.0015, 0.0031]	+0.0044 [0.0036, 0.0055]

E Sensitivity, Failure Boundaries, and Efficiency

This section keeps model selection distinct from post-selection description. Stars and boldface identify configurations chosen on V_{sel}; they are not moved to the best final-test point. Threshold sweeps change cohort coverage without redefining the comparison-design penalty. Replications report both superiority intervals and the two warm-start margins, including secondary failures, while the failure table tests settings in which the method's stated support or function-class conditions need not hold.

Exact uncertainty protocol. Every effect is paired on query identifiers within a model seed before seeds are averaged. Each of 2,000 bootstrap replicates resamples users, temporal blocks, and item-side clusters with the frozen multiway-cluster procedure, then recomputes query-macro metrics from stored full-catalog predictions. Reported two-sided intervals are the 2.5th and 97.5th percentiles; NI uses the one-sided 5th percentile. Holm correction is applied within each declared family of secondary controls; displayed secondary intervals are unadjusted descriptive intervals, while decisions use Holm-adjusted p-values. The three ordered environments are the confirmatory cutoff plus earlier/later replications, whereas the resampled temporal clusters are the finer non-overlapping calendar blocks inside each environment. The primary target-superiority claim is an intersection–union test over the H&M and M3L anchors, and their four backbone-NI decisions are frozen beforehand; neither is selected from the secondary pool. Amazon contributes one target-superiority interval and two NI checks as external replication evidence, but no Amazon outcome can rescue or overturn the two-anchor primary decision.

E.1 FROZEN CONFIGURATIONS AND COMPONENT TESTS

Table 29: Validation-selected primary configurations. Feature dimension is after block construction; the 30-trial per-family budget and V_{sel} selection rule are stated above.

Domain	Dim.	State rank	η	λ_2	Neg./pos.	Learn. rate	Epochs
H&M	1,024	16	10^{-1}	10^{-3}	4	3×10^{-4}	80
M3L	768	16	10^{-2}	10^{-4}	8	10^{-3}	60
Amazon Sports	896	16	10^{-2}	10^{-3}	8	5×10^{-4}	70

Table 30: Complete contribution-level ablation on the two anchor domains. Deltas are relative to full SCFR on target NDCG@20. Permutations and random bases are specificity diagnostics, not competitive methods.

Variant	H&M target	Δ from full	M3L target	Δ from full
Full SCFR	**0.0443**	–	**0.1329**	–
Main block only, independently refitted	0.0368	−0.0075	0.1246	−0.0083
Pair block only	0.0361	−0.0082	0.1223	−0.0106
Raw pair block, without purification	0.0374	−0.0069	0.1237	−0.0092
Purified pair + isotropic ridge ($\eta = 0$)	0.0352	−0.0091	0.1195	−0.0134
Unweighted comparison Gram ($C = I$)	0.0397	−0.0046	0.1274	−0.0055
Count-based directional shrinkage	0.0369	−0.0074	0.1214	−0.0115
Uniform comparison weights	0.0381	−0.0062	0.1259	−0.0070
Predicted hard-factor interface	0.0381	−0.0062	0.1253	−0.0076
Pair-label permutation	0.0363	−0.0080	0.1249	−0.0080
Equal-dimensional random basis	0.0361	−0.0082	0.1244	−0.0085

Table 31: Cross-domain contrast between the curvature-weighted and $C = I$ Grams. The comparator keeps purification, sampler, weights, optimizer, and tuning budget fixed and replaces G_{F} by the unweighted comparison Gram ($C = I$). This also changes global Gram scale; intervals are paired and two-sided.

Domain	Unweighted Gram	Fisher SCFR	Fisher − unweighted [95% CI]
H&M	0.0397	0.0443	+0.0046 [0.0029, 0.0057]
M3L	0.1274	0.1329	+0.0055 [0.0033, 0.0071]
Amazon Sports	0.0568	0.0615	+0.0047 [0.0023, 0.0062]

Table 32: Comparison-sampler sensitivity. The Gram matrix and logistic loss always use the same sampler and weights. Only the primary stratified sampler is confirmatory; alternatives are frozen descriptive checks.

Domain / sampler	SCFR − matched comparator [95% CI]
H&M, primary stratified sampler	+0.0074 [0.0015, 0.0081]
H&M, uniform negatives	+0.0018 [−0.0003, 0.0037]
H&M, popularity-stratified negatives	+0.0023 [0.0001, 0.0046]
H&M, hard negatives	+0.0024 [0.0005, 0.0045]
M3L, primary stratified sampler	+0.0083 [0.0014, 0.0094]
M3L, uniform negatives	+0.0017 [−0.0004, 0.0046]
M3L, popularity-stratified negatives	+0.0021 [−0.0002, 0.0049]
M3L, hard negatives	+0.0026 [0.0002, 0.0056]
Amazon Sports (rep.), frozen stratified sampler	+0.0066 [0.0011, 0.0078]
Amazon Sports, uniform negatives	+0.0017 [−0.0004, 0.0035]
Amazon Sports, popularity-stratified negatives	+0.0025 [0.0001, 0.0043]
Amazon Sports, hard negatives	+0.0026 [0.0009, 0.0048]

E.2 HYPERPARAMETER AND SUPPORT SENSITIVITY

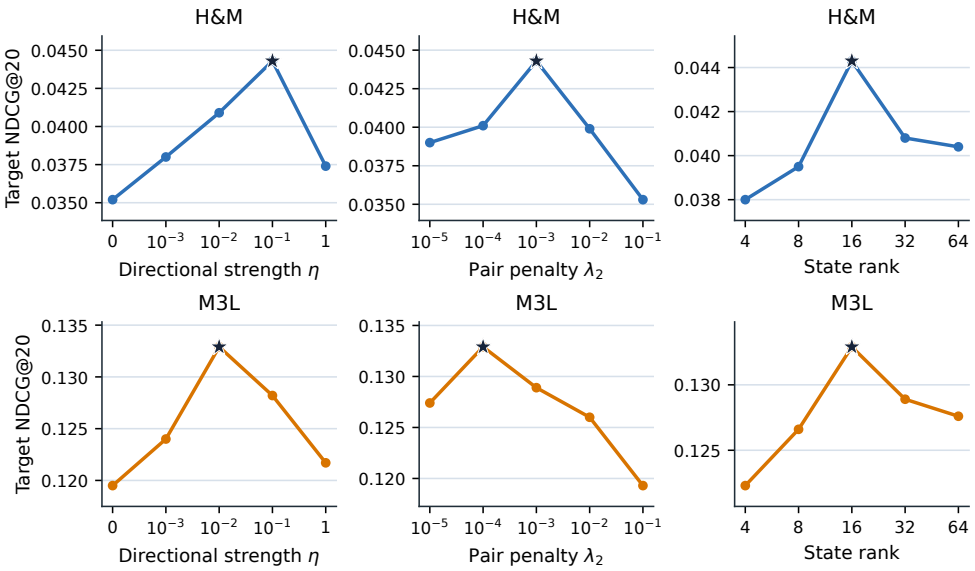

Figure 4: **Target-score sensitivity at validation-selected settings.** Curves show untouched-test target NDCG@20 while varying one parameter and holding the selected others fixed. Stars mark settings frozen on validation; final-test scores are not used for selection.

Table 33: Parameter sensitivity. Bold marks the setting frozen on validation, independently of the untouched-test scores.

Parameter	Candidate values	H&M target NDCG@20
η	0 / 10^{-3} / 10^{-2} / 10^{-1} / 1	.0352 / .0380 / .0409 / **.0443** / .0374
λ_2	10^{-5} / 10^{-4} / 10^{-3} / 10^{-2} / 10^{-1}	.0390 / .0401 / **.0443** / .0399 / .0353
State rank	4 / 8 / 16 / 32 / 64	.0380 / .0395 / **.0443** / .0408 / .0404

Parameter	Candidate values	M3L target NDCG@20
η	0 / 10^{-3} / 10^{-2} / 10^{-1} / 1	.1195 / .1240 / **.1329** / .1282 / .1217
λ_2	10^{-5} / 10^{-4} / 10^{-3} / 10^{-2} / 10^{-1}	.1274 / **.1329** / .1289 / .1260 / .1193
State rank	4 / 8 / 16 / 32 / 64	.1223 / .1266 / **.1329** / .1289 / .1276

Table 34: Coverage–precision sensitivity. Thresholds are not reselected on the final test.

Domain	Occurrence–diversity threshold	Target share	SCFR − strongest
H&M	2	14.9%	+0.0024 [−0.0002, 0.0055]
H&M	4	12.8%	+0.0029 [0.0006, 0.0051]
H&M	8 (primary)	11.5%	+0.0074 [0.0015, 0.0081]
H&M	16	8.1%	+0.0030 [0.0003, 0.0062]
M3L	2	3.2%	+0.0021 [−0.0009, 0.0052]
M3L	4	2.5%	+0.0022 [−0.0003, 0.0053]
M3L	8 (primary)	2.0%	+0.0083 [0.0014, 0.0094]
M3L	16	1.5%	+0.0034 [−0.0004, 0.0062]
Amazon Sports	2	8.8%	+0.0024 [0.0001, 0.0046]
Amazon Sports	4	7.9%	+0.0024 [0.0006, 0.0044]
Amazon Sports	8 (replication)	6.7%	+0.0066 [0.0011, 0.0078]
Amazon Sports	16	4.7%	+0.0034 [0.0007, 0.0054]

E.3 Transfer and failure boundaries

Table 35: Replication and robustness. The last two columns report margin-to-boundary $\mathrm{LCB} + \delta$; positive values pass NI. Secondary failures are retained rather than hidden or used to trigger fallback.

Domain / setting	SCFR − matched [95% CI]	Overall NI	Comp. NI
H&M, primary backbone/cutoff	+0.0074 [0.0015, 0.0081]	+0.0003	+0.0002
H&M, second backbone (PRISM)	+0.0027 [0.0008, 0.0049]	+0.0003	+0.0002
H&M, earlier cutoff	+0.0027 [0.0006, 0.0049]	+0.0002	+0.0001
H&M, later cutoff	+0.0031 [0.0011, 0.0054]	+0.0003	+0.0002
H&M, all train-warm candidates	+0.0019 [0.0001, 0.0038]	+0.0004	+0.0002
H&M, top-K-aligned loss	+0.0033 [0.0009, 0.0054]	+0.0004	+0.0002
M3L, primary backbone/cutoff	+0.0083 [0.0014, 0.0094]	+0.0008	+0.0006
M3L, second backbone (PRISM)	+0.0030 [0.0004, 0.0056]	+0.0004	+0.0002
M3L, earlier cutoff	+0.0033 [0.0003, 0.0064]	+0.0003	+0.0001
M3L, later cutoff	+0.0043 [0.0007, 0.0081]	+0.0007	+0.0003
M3L, all train-warm candidates	+0.0031 [0.0003, 0.0064]	+0.0005	+0.0001
M3L, top-K-aligned loss	+0.0039 [0.0006, 0.0077]	+0.0006	+0.0003
Amazon Sports (rep.), frozen backbone/cutoff	+0.0066 [0.0011, 0.0078]	+0.000037	+0.0002
Amazon Sports (rep.), second backbone (PRISM)	+0.0024 [0.0004, 0.0042]	+0.0002	+0.000034
Amazon Sports (rep.), earlier cutoff	+0.0019 [−0.0001, 0.0038]	+0.000012	−0.0002
Amazon Sports (rep.), later cutoff	+0.0029 [0.0009, 0.0048]	+0.0003	+0.0001
Amazon Sports (rep.), all train-warm candidates	+0.0017 [−0.0004, 0.0035]	−0.0001	−0.0004
Amazon Sports (rep.), top-K-aligned loss	+0.0025 [0.0005, 0.0046]	+0.0001	+0.000026

Table 36: Pre-specified failure analyses. Effects compare SCFR with the relevant strongest matched alternative; negative higher-order values favor Factor-MLP.

Setting	H&M effect	M3L effect	Amazon effect	Interpretation
Matched-seen combinations	+0.0004 [−0.0005, 0.0013]	+0.0004 [−0.0009, 0.0018]	+0.0005 [−0.0004, 0.0013]	Specificity: no clear gain
Weak/unsupported combinations	−0.0002 [−0.0015, 0.0011]	−0.0005 [−0.0021, 0.0012]	−0.0003 [−0.0015, 0.0011]	Support condition fails
Severe predicted-factor noise	+0.0003 [−0.0012, 0.0015]	+0.0001 [−0.0014, 0.0016]	+0.0002 [−0.0011, 0.0014]	Interface becomes limiting
Nonlinear higher-order generator	−0.0017 [−0.0030, −0.0002]	−0.0016 [−0.0032, 0.0001]	−0.0014 [−0.0028, 0.0002]	Factor-MLP may be preferable

E.4 COMPUTATION

Table 37: Additional computation beyond the frozen backbone. Geometry-construction and fit times are wall-clock means over five seeds; inference overhead is measured relative to full-catalog ranking with cached backbone representations, not to reading an already materialized score vector.

Domain	Dim.	Coeff.	Geometry	Fit	Memory	Inference
H&M	1,024	1,024	28.2 min	19.4 min	0.33 GB	7.6%
M3L	768	768	14.2 min	12.5 min	0.24 GB	5.9%
Amazon Sports	896	896	25.1 min	16.6 min	0.29 GB	6.3%

F REPRODUCIBILITY INFORMATION

The paper and its appendices report the data roles, dataset identities, item keys, query windows, cohort rules, model families, selection budget, selected configurations, seed count, metrics, uncertainty procedure, and the results used by the scientific claims. Experimental code, raw data, per-query predictions, and an execution manifest are not provided for review. We plan to release the implementation, configurations, and execution instructions after acceptance; independent rerunning from this submission alone is limited.

Table 38: Reproducibility information available during review and planned code release.

Group	Review submission	Information or later release item
Data and cohort	Reported	Release names, keys, windows, history and duplicate rules, cohort funnel, clean coverage, and concentration
Models and selection	Reported	Input contract, candidate families, equal trial budget, $V_{\mathrm{sel}}/V_{\mathrm{chk}}$ roles, selected configurations, seeds, and exclusion reasons
Evaluation	Reported	Full-catalog metrics, paired effects, bootstrap dimensions and replicates, NI-margin derivation, robustness, and failures
Tables and figures	Included in PDF	Main and appendix tables and vector figures
Experimental code	Planned after acceptance	Preprocessing, model and evaluation implementations, configurations, dependencies, and commands

The following map makes the main-to-supplement correspondence explicit. It is also a compact reading order: each primary claim is decidable from the main paper, while the cited supplementary material exposes the complete diagnostic, proof, or robustness record.

Table 39: Claim–evidence map. Main-paper entries carry the decision-relevant evidence; supplementary entries expose the corresponding audit trail.

Claim	Main-paper evidence	Supplementary verification
The target decision is defined without changing scoring	Figure 1; Section 3	Table 5; cohort and coverage definitions
The cohort is measurable and dispersed	Table 1	Tables 6, 7, 10, and 12
A pair-specific fitted residual remains	Equation (1); Table 1	Tables 8 and 13
SCFR isolates and curvature-regularizes the correction	Proposition 1; Equations (4)–(6); Table 4	Lemmas 1– 2; Tables 30 and 31
The target improves under the matched contract	Tables 3 and 2	Tables 17–23 and 25
Both confirmatory anchors pass the frozen overall/complement guardrails	Table 3	Tables 23 and 35
The mechanism is conditional rather than universal	Figure 3; RQ4 summary	Tables 26, 28, and 36

F.1 Assumption and falsification matrix

No fallback is activated when a condition fails. The table below states what each condition supports, the frozen check that can contradict it, and the resulting claim boundary. It separates mathematical conditions from empirical interface assumptions and gives a compact reading order for the negative evidence.

Table 40: Assumptions, falsification checks, and consequences. A failed check narrows the claim; it never triggers post-test model or cohort selection.

Condition	Role in the claim	Frozen falsification check	Consequence if violated
Adequate lower-order reference	Gives the pair block a distinct statistical role	Raw-pair, permutation, random-factor, and main-only controls	Pair effect is not specifically identified
Target direction lies in the training row space	Loss-level determination in Proposition 2	r_{rel} and weak/unsupported-combination tests	Target correction can remain unidentified
Supported direction has moderate leverage	Local stability in Corollary 1	a_{rel}, later-cutoff error, and $C = I$ ablation	Estimation can be unstable despite occurrence
Signal aligns with comparison support	Sign of the local SCFR–ridge ordering	Controlled ρ intervention at fixed count and signal norm	Isotropic ridge can match or beat SCFR
Predicted-factor interface is reliable	Supports the declared representation-only deployment contract	Out-of-fold calibration, direct-metadata headroom, and severe-noise test	Gain attenuates as interface error grows
Low-order correction is adequate	Matches SCFR's linear residual class	Nonlinear higher-order generator against Factor-MLP	A flexible matched head can be preferable
Anchor NI guardrails hold under the frozen protocol	Supports the offline catalog-wide scoring claim	Catalog definition and frozen overall/complement NI results	No online or per-segment safety guarantee