THE NEXT RECOMMENDATION IS A DIRECTION: DISCOVERING HIGH-VALUE SEMANTIC OPPORTUNITIES BEYOND USER HISTORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recommenders usually choose the next item, but interest exploration and generative recommendation raise an earlier question: *where in the content space should the system go next?* We formulate next semantic direction selection: given a short user history, find a continuous semantic location whose nearby, previously unobserved content contains a high-value realization. Our hypothesis is that population feedback contains cross-content behavioral regularities that transfer information from observed responses to semantically different content. We introduce NORTH, a population-anchored framework for continuous direction search. Shared feedback learns behavioral representations; the user's history turns population opportunity into a personalized direction score. A learned policy starts from promising population anchors, decides where to search and whether to stay or move, and proposes continuous displacements. Each proposal is grounded in a fixed-size core of real items and assessed by the same direction scorer. Best-Realization Value measures the best observed response within this fixed realization budget. On the full and fully-observed dataset variants, KuaiRec-Full and KuaiRec-FO, NORTH exceeds population-only bank selection in all four evaluated settings. It leads the compared OPEN generators on KuaiRec-Full and KuaiRec-FO. Ablations support the use of history-conditioned proposals and multimodal displacements.

1 INTRODUCTION

Recommendation is moving beyond ranking a fixed list of known items. Interactive recommenders explore interests that are not yet visible in a user's history, while generative recommenders can produce semantic identifiers, repurpose existing content, or create content (Geng et al., 2022; Rajput et al., 2023; Wang et al., 2023). These capabilities raise a question upstream of realization: *what semantic direction is worth exploring for this user?* A direction can focus a retrieval query or provide a semantic target for a generator. The decision matters when only a small number of directions can be explored.

The difficulty is information. A short history covers only a small part of the content space. Semantic similarity tells us which items are close in content, but not which ones elicit related responses. A user can enjoy semantically distant topics, while two nearby items can receive very different feedback. Other users provide a way to bridge this gap: their responses connect items in the target user's history to content that the target user has not observed.

Population-level cross-content hypothesis. Population feedback contains predictive relationships across content locations. A target user's observed responses help identify which relationships apply to that user, allowing the system to estimate opportunity beyond the observed history. Figure 1 illustrates the separation: semantic geometry determines *where content is*; population feedback learns *how response information transfers*; the target history determines *which transfer is useful*.

Turning this hypothesis into a direction-search method requires more than a score over a predefined menu. A fixed bank can contain useful local directions, but a scorer restricted to that bank cannot propose a different realization core. At the other extreme, predicting a location from a short history

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.

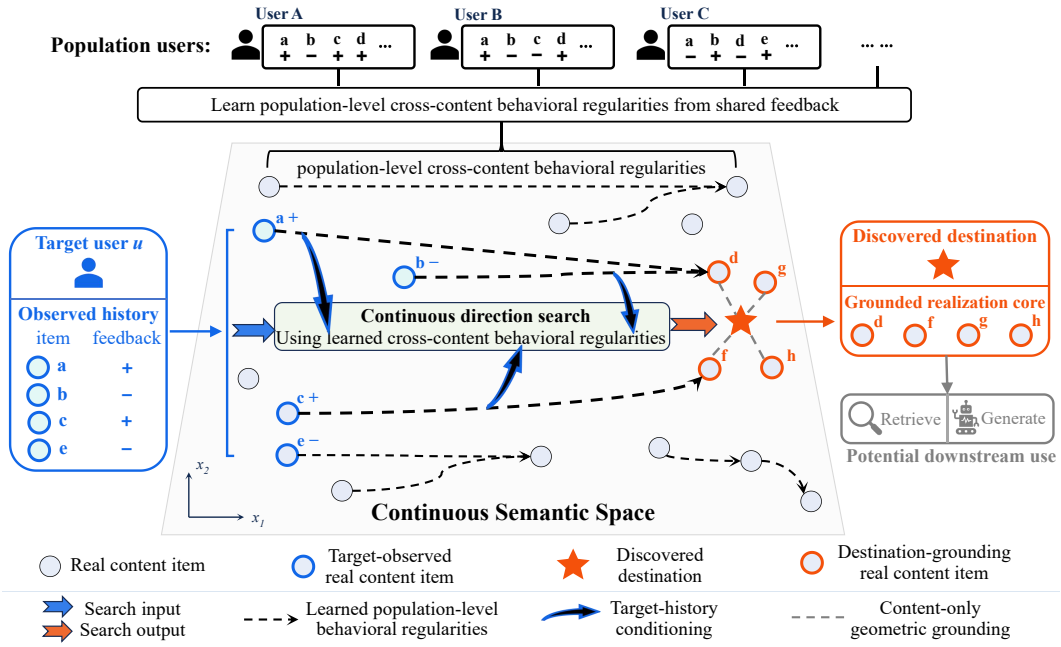


Figure 1: **From a target user's observed history to a new semantic direction.** Population feedback supplies cross-content behavioral regularities. The target user's observed history determines how that shared structure is used in continuous search; blue curves illustrate this conditioning. The orange star is a semantic location, and nearby real items form its fixed-size realization core. Retrieval and generation are potential downstream uses of the discovered direction.

without a population starting point discards valuable information about where promising content already lies.

We introduce NORTH, a population-anchored framework for cross-content direction search. First, a shared behavioral encoder places historical and candidate content in a common response space. Second, a personalized direction scorer combines a core's population opportunity with a history-dependent correction. Third, a search policy uses high-opportunity cores as anchors, reweights them for the user, and decides whether to retain an anchor or move to a new continuous position. Grounding each position in nearby real content makes its value observable. The same scorer then compares the realizations visited by the search.

We evaluate a direction using **Best-Realization Value (BRV)**: the best observed user response within a fixed-size realization core. This asks whether the direction contains a worthwhile realization, rather than whether every nearby item is equally good. Fixing the core size makes the realization budget comparable. The best-choice value of a bounded set has a precedent in recommendation-set utility (Viappiani & Boutilier, 2010); here it provides a measurable target for semantic direction search.

Our evaluation separates fixed-bank scoring from independent OPEN proposals on KuaiRec-Full and KuaiRec-FO. The bank scorer leads the listed ranking baselines. In OPEN search, NORTH exceeds population-only bank selection in all four settings and leads the compared generators on KuaiRec-Full and KuaiRec-FO.

The paper makes three contributions. **First**, we move the decision target from which item comes next to which semantic direction is worth exploring. The output is a continuous location grounded in a fixed-size real-content core, providing a target for both retrieval and content generation. **Second**, we posit that population-level cross-content behavioral regularities can transfer evidence from a observed history to unseen content, and realize this hypothesis through population-anchored, history-conditioned direction search. **Third**, we evaluate this decision with BRV on two population-training settings, separating fixed-bank scoring from independent OPEN proposals and testing the roles of proposal conditioning and multiple displacement modes.

2 Related Work

2.1 Interest Exploration

Contextual bandits select actions using reward estimates and uncertainty (Li et al., 2010; Auer et al., 2002; Agrawal & Goyal, 2013); NeuralUCB extends this approach to neural reward models (Zhou et al., 2020). UpliftRec estimates effects of category-exposure allocations for interest exploration (Chen et al., 2024). LLM-based interest exploration is especially close to our decision interface: Wang et al. (2024b) propose interests before retrieving items, and Wang et al. (2025) align proposals with user feedback. We also separate direction proposal from realization, but learn the proposal policy from population responses over content-supported continuous positions. Our offline decision uses one observed history; no new target-user feedback is revealed while the system searches.

2.2 Generative Recommendation

P5 casts recommendation tasks as language generation (Geng et al., 2022). TIGER generates semantic item identifiers (Rajput et al., 2023), while LETTER learns tokenizations that combine semantic and collaborative information (Wang et al., 2024c). TALLRec and LLaRA align language models with recommendation tasks and behavioral representations (Bao et al., 2023; Liao et al., 2024). DreamRec generates target-item representations through history-guided diffusion (Yang et al., 2023); PreferDiff integrates preference learning into diffusion-based recommendation (Liu et al., 2025). Their continuous outputs motivate the independent proposal comparison in Section 5.3.

GeneRec describes retrieval, repurposing, and content creation within a broader generative recommendation paradigm (Wang et al., 2023). More directly, Cheng et al. (2026) introduce Recommendation-as-Generation (RaG), unifying personalized video recommendation and generation at industrial scale. Shared semantic identifiers separate content and creative-style semantics, and video-generation agents use the inferred identifiers to plan and refine videos. NORTH addresses a complementary intermediate decision: selecting a continuous semantic target whose opportunity is grounded in existing real content. We evaluate that target before downstream realization, rather than training or evaluating an end-to-end video generator. Table 1 contrasts these interfaces.

2.3 Collaborative and Content-Aware Recommendation

Latent-factor and neural recommenders learn shared preference structure (Koren et al., 2009; Hu et al., 2008; He et al., 2017). BPR, LightGCN, and SASRec model pairwise preferences, user–item graphs, and sequential histories, respectively (Rendle et al., 2009; He et al., 2020; Kang & McAuley, 2018). Content-aware methods connect semantics to collaborative factors, from collaborative topic regression and deep content models to DropoutNet, CLCRec, and CARec (Wang & Blei, 2011; Wang et al., 2015; van den Oord et al., 2013; Volkovs et al., 2017; Wei et al., 2021; Wang et al., 2024a). NORTH uses this connection to predict core-level opportunity and propose locations beyond an enumerated bank.

Table 1: The direction-search decision compared with related recommendation interfaces. RaG directly creates videos; NORTH returns an intermediate semantic target supported by real content.

Paradigm	Primary output	Next operation
Contextual bandit	Candidate action	Observe feedback and update the decision rule
Item or sequential ranker	Item scores or next item	Rank or retrieve catalogue items
Content-aware collaborative model	Behavior-aware item representation	Score known or content-described items
Generative recommender	Identifier, representation, or content	Retrieve or realize a recommendation
LLM interest exploration	Interest description or cluster	Retrieve content for the proposed interest
RaG (Cheng et al., 2026)	Content/style semantic IDs and personalized videos	Plan, generate, and deliver video content
NORTH	Continuous location and realization core	Retrieve or create content around a population-supported personal direction

3 NEXT SEMANTIC DIRECTION SELECTION

Let $e_i \in \mathbb{R}^p$ be the content embedding of item i, and let $x_i \in \mathbb{R}^d$ be its frozen semantic coordinate. The target user's observed history is $H_u = \{(i_t, y_{ui_t})\}_{t=1}^{W}$, with observed item set $I(H_u)$. Here y denotes the response used for learning and evaluation; its preprocessing is specified in Appendix A. The task is to return a location $c_u \in \mathbb{R}^d$ and its real-content support, using H_u and training-population feedback.

Grounding a location. A content-only neighborhood $\mathcal{R}(c)$ defines the local semantic support of c. Its realization core contains the K nearest supported items:

$$C_K(c) = \mathrm{Nearest}_K(c; \mathcal{R}(c)), \qquad |C_K(c)| = K. \tag{1}$$

We use $K = 4$. Local support adapts to content density; its construction is given in Appendix B. A core is *external to the history* when $C_K(c) \cap I(H_u) = \varnothing$. Thus the output offers previously unobserved real content. Geometric grounding uses content coordinates, while feedback is used to learn and evaluate opportunity.

Best-realization value. For a core with observed outcomes for all members,

$$V_u(C) = \mathrm{BRV}(u, C) = \max_{i \in C} y_{ui}, \qquad |C| = K. \tag{2}$$

The fixed budget prevents a larger neighborhood from winning merely by containing more items. BRV measures the opportunity available within the core. A downstream realizer still has to identify or produce a good realization. Predicting a set's best outcome is different from taking the maximum of predicted item means: in general, $\max_i \mathbb{E}[Y_{ui} \mid H_u] \neq \mathbb{E}[\max_i Y_{ui} \mid H_u]$. This motivates a core-level learning target (Viappiani & Boutilier, 2010; Elmachtoub & Grigas, 2022).

Fixed-bank scoring and OPEN search. Let $\mathcal{B}$ be a bank of previously enumerated realization cores. Fixed-bank scoring returns a core in $\mathcal{B}$. OPEN search proposes continuous coordinates and grounds them through Equation (1), so the returned core can satisfy $C_K(c_u) \notin \mathcal{B}$. Bank membership is determined by the identities of the K items, not by whether the coordinate lies outside a spatial outline. Different coordinates inducing the same core are the same realizable action. We distinguish the realization budget K from the proposal budget S, the number of continuous search attempts.

4 NORTH: POPULATION-ANCHORED DIRECTION SEARCH

NORTH combines three components (Figure 2): cross-content behavioral learning, personalized direction scoring, and continuous direction search. Population feedback trains the first two; the resulting representations and scorer are frozen while the search policy is trained. At test time, the target user's observed history conditions both scoring and search.

4.1 CROSS-CONTENT BEHAVIORAL LEARNING

A shared encoder $q_i = f_\theta(e_i)$ maps content into behavioral response features. The semantic coordinate x_i determines locality, while q_i is trained to explain differences in user response. Historical and candidate items use the same encoder. For history design matrix Φ_H, with rows $[1; q_{i_t}]^\top$, and response residuals $r_H = y_H - m_H$, we obtain the user state by minimizing

$$v_u = \arg\min_v \ \|W_H^{1/2}(r_H - \Phi_H v)\|_2^2 + \|v\|_2^2. \tag{3}$$

The first-order condition is $(I + \Phi_H^\top W_H \Phi_H) v_u = \Phi_H^\top W_H r_H$. The matrix is positive definite, giving the unique fold-in solution

$$v_u = (I + \Phi_H^\top W_H \Phi_H)^{-1} \Phi_H^\top W_H r_H, \qquad \chi_u = [v_u; v_u \odot v_u; \mathrm{SiLU}(P_v v_u)]. \tag{4}$$

Here m_i is a training-population item mean and W_H contains history weights. Differentiation through the ridge solve lets population training shape the representation used for short-history adaptation (**?**). Both positive and negative residuals contribute to χ_u.

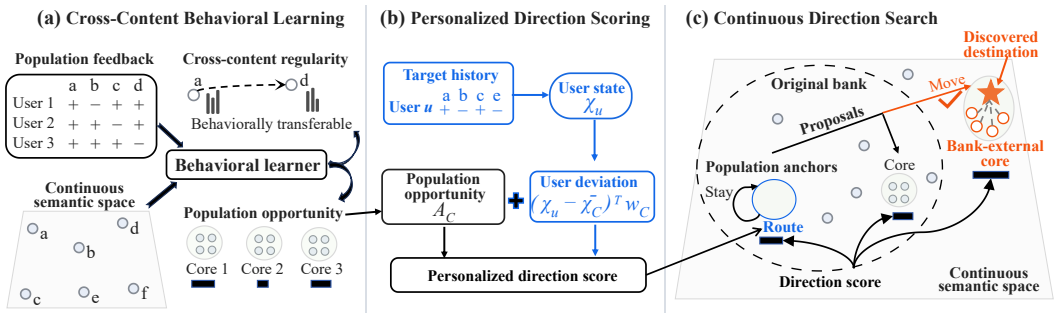


Figure 2: **Method overview.** (a) Shared feedback learns behavioral representations and provides population opportunity estimates. (b) The target history becomes a user state that adjusts each core's population opportunity. (c) A learned policy routes search from population anchors, stays or proposes a continuous move, and grounds proposals in real items. The frozen direction scorer evaluates the visited cores. The dashed bank outline schematically denotes enumerated core identities.

How response information transfers. Write $\psi_i = [1; q_i]$ and $v_u = [b_u; z_u]$. Substituting Equation (4) into the adapted item readout gives

$$\widehat{y}_{uj} = m_j + b_u + z_u^\top q_j = m_j + \sum_{t=1}^{W} a_{jt}(H_u)(y_{ui_t} - m_{i_t}), \tag{5}$$

$$a_{jH}^\top = \psi_j^\top (I + \Phi_H^\top W_H \Phi_H)^{-1} \Phi_H^\top W_H. \tag{6}$$

These weights depend on the learned response features and the observed history, not on semantic distance alone. The feature inner product $q_i^\top q_j$ can therefore link semantically distant content. Its predictive direction is determined by which content is observed and which destination is queried; we do not assume a causal relation or require the feature kernel itself to be asymmetric. This makes the cross-content hypothesis explicit: population training learns a shared response basis, and the target history determines how its residuals are transferred. The direction scorer below is supervised by complete core outcomes rather than by the maximum of these item predictions.

4.2 Personalized Direction Scoring

For core C, let D_C be training users with complete post-history outcomes for its members and for whom C is external. Its population opportunity and user-state center are

$$A_C = \frac{1}{|D_C|} \sum_{v \in D_C} V_v(C), \qquad \bar{\chi}_C = \frac{1}{|D_C|} \sum_{v \in D_C} \chi_v. \tag{7}$$

A permutation-invariant set encoder maps the real members to a direction representation (Zaheer et al., 2017):

$$w_C = g_\theta \left([\mathrm{Avg}_{i \in C}\, \phi_i; \max_{i \in C} \phi_i; \mathrm{Var}_{i \in C}\, \phi_i] \right). \tag{8}$$

Pooling is componentwise, and the output dimension matches χ_u. The personalized score is

$$Q(u, C) = A_C + \alpha(\chi_u - \bar{\chi}_C)^\top w_C. \tag{9}$$

The first term captures population opportunity. The second estimates whether the target history makes that core more or less valuable for this user. Centering preserves the population estimate over the same donor states:

$$\frac{1}{n_C} \sum_{v \in D_C} Q(v, C) = A_C + \alpha \left(\frac{1}{n_C} \sum_{v \in D_C} \chi_v - \bar{\chi}_C \right)^\top w_C = A_C. \tag{10}$$

Thus the model learns who should deviate from a direction's population reference, rather than relearning its intercept. A training query user's contribution is excluded from the population statistics used for that query; Appendix C gives the leave-one-out formula.

The primary training loss directly regresses the real core value:

$$L_{\mathrm{BRV}} = \mathbb{E}_{\mathrm{tr}}[(Q(u, C) - V_u(C))^2]. \tag{11}$$

An auxiliary loss fits value differences from a population-preferred reference core; item-response supervision grounds the behavior features, and correction regularization discourages unnecessarily large deviations. We average valid targets within each user and then average users. Appendix C.3 specifies the combined objective. The member-based score also applies to a newly grounded core: its representation and population statistics are computed from its real members and legal training donors, without fitting a new core-specific parameter.

4.3 Continuous Direction Search

Population anchors and routing. From eligible bank cores, we take high-opportunity cores, reduce heavy member overlap, and retain at most $J = 8$ anchors. Each anchor C_j has a representative coordinate c_j. The population-best core is retained. A routing network makes a bounded history-dependent adjustment to the population prior:

$$\pi_0(j) \propto \exp(A_{C_j}/\tau), \qquad \pi_\psi(j \mid H_u) \propto \pi_0(j) \exp r_\psi(u, j). \tag{12}$$

Its inputs include χ_u, the anchor coordinate and representation, its population opportunity, training support, and $Q(u, C_j)$. Thus population evidence provides starting points, while the history controls how search effort is allocated.

Stay or move. For a routed anchor, the policy predicts a stay probability $s_\psi(u, j)$. Staying returns c_j. The move distribution is a four-component diagonal Gaussian mixture,

$$q_\psi(\Delta \mid H_u, C_j) = \sum_{m=1}^{M} \omega_{ujm} \mathcal{N}(\Delta; \mu_{ujm}, \mathrm{diag}(\sigma^2_{ujm})), \qquad M = 4. \tag{13}$$

The weights sum to one and component variances are positive. A sampled displacement is rescaled by the anchor's local semantic scale:

$$\Delta c \sim q_\psi(\cdot \mid H_u, C_j), \qquad c' = c_j + \ell(c_j)\Delta c. \tag{14}$$

Multiple modes allow the policy to propose different promising neighborhoods instead of averaging their coordinates. Staying preserves a useful anchor when the history gives little reason to move. More formally, the joint proposal has a discrete stay mass and a continuous move density:

$$\begin{aligned} p_\psi(j, \mathrm{stay}, \Delta \mid H_u) &= \pi_\psi(j \mid H_u) s_\psi(u, j) \delta_0(\Delta), \\ p_\psi(j, \mathrm{move}, \Delta \mid H_u) &= \pi_\psi(j \mid H_u)(1 - s_\psi(u, j)) q_\psi(\Delta \mid H_u, C_j). \end{aligned} \tag{15}$$

Routing allocates search effort across anchors; staying and moving determine how that effort is used within an anchor's neighborhood.

Ground and score. Each proposed coordinate is grounded through Equation (1), checked for admissibility, and deduplicated by core identity. The visited set includes the anchor seeds, the population-best bank core, and the bank scorer's incumbent. Within $S = 72$ proposal attempts, the frozen Q compares the resulting cores and returns the highest-scoring one with its coordinate. This can produce a bank-external core. It can also retain a bank core when no visited alternative receives a higher score. The bank scan used to obtain the incumbent is counted separately from the proposal budget.

4.4 Learning the Search Policy

For training users, we sample positions around population anchors, ground them, and use complete training-user outcomes to compute $V_u(C)$. High-value actions receive larger target weights. An action records its anchor, stay/move decision, and displacement; its joint probability is $p_\psi(j, z, \Delta c \mid H_u)$. We train the policy by weighted negative log-likelihood with a population-prior penalty:

$$L_{\mathrm{search}} = -\mathbb{E}_u \sum_{j,k} \widetilde{\omega}_{ujk} \log p_\psi(j, z_{jk}, \Delta c_{jk} \mid H_u) + \lambda_{\mathrm{prior}} L_{\mathrm{prior}}. \tag{16}$$

For valid training actions, the unnormalized target mass is

$$\omega_{ujk} = \pi_0(j) b_{jk} \exp\left[\mathrm{clip}\left(\frac{V_u(C_{jk}) - A_{C_j}}{\tau_g}, -5, 5\right)\right], \qquad \tau_g = 0.1. \tag{17}$$

The reference mass b_{jk} allocates stay and move mass among the observed representatives; the normalized $\widetilde{\omega}$ sums to one across the user's valid actions, including different anchors. Hence the policy can learn *which* promising population direction to explore, not only where to move after an anchor has been fixed.

Without clipping, and with $\tau_g = \tau$, the exponential factors obey

$$\exp(A_{C_j}/\tau) \exp((V_u(C_{jk}) - A_{C_j})/\tau) = \exp(V_u(C_{jk})/\tau). \tag{18}$$

Thus population opportunity supplies the starting distribution while the observed personal difference adjusts it toward high realized utility. Clipping limits the influence of extreme target weights; it does not alter the BRV labels. The scorer and behavioral encoder remain frozen during this stage. At inference, proposals use the observed history and training-population information; held-out target outcomes are used only to measure the returned core. Appendix D specifies the reference masses, prior penalty, and masked likelihood. An unobserved action is not assigned a low-value label: routing supervision is normalized only across anchors with observed targets, and stay/move likelihood is used only when both categories have observations. This avoids treating missing feedback as evidence against a direction.

Training and deployment have distinct information roles. Population users provide the labels for learning the policy; a new target user provides only the short history used to form χ_u. Search evaluates proposals using the frozen scorer and training-population statistics, never newly observed rewards from the target user. It is therefore a single-decision, model-guided search procedure, not an online bandit that learns from target feedback at each proposal. The output is a coordinate–core pair: the coordinate describes where to realize content, and the core makes its opportunity measurable.

5 Experiments

5.1 Data and Evaluation

We use two population-training settings from KuaiRec (Gao et al., 2022). **KuaiRec-Full** uses 5,765 users outside the nearly fully-observed subset as training evidence. **KuaiRec-FO** uses 903 training users from that subset, with 226 development and 282 test users. Both settings evaluate on held-out KuaiRec-FO users, enabling measurement of complete realization cores. The Full-trained model is evaluated without further training on FO. We use history budgets $W \in \{20, 50\}$ and realization size $K = 4$.

The primary metric is mean BRV over evaluable users. Missing responses remain missing; an outcome is measurable only when all core members have feedback. Dataset processing, response transformation, local training support, and coverage accounting appear in Appendix A. The two settings share the target-user evaluation source but differ in population-training evidence.

5.2 Fixed-Bank Direction Scoring

Table 2 holds the eligible bank fixed within each setting. NORTH–Bank is the scorer-only variant: it scores the original bank without continuous proposals. This comparison tests the direction-level learning objective independently of search.

The baselines cover semantic selection, item ranking (BPR, LightGCN, SASRec), a single-decision NeuralUCB-Diag adaptation, UpliftRec, content–collaborative alignment (CARec and CLCRec), and LLM interest exploration. For item scorers, each core is represented by its member scores sorted in descending order; cores are compared lexicographically. HAdapt denotes conditioning on the available target history. Appendix E.1 details this common interface.

NORTH–Bank has the largest reported mean BRV in all four fixed-bank settings. Its margins over the strongest listed baseline are 0.0044, 0.0130, 0.0231, and 0.0572, in table order. CARec-HAdapt and CLCRec-HAdapt are close competitors, highlighting the usefulness of content–behavior

Table 2: **Fixed-bank direction scoring.** Mean BRV (↑) on the same eligible bank within each setting. NORTH–Bank scores the fixed bank without continuous search.

	KuaiRec-Full		KuaiRec-FO	
Method	$W = 20$	$W = 50$	$W = 20$	$W = 50$
Random	0.8485	0.8795	0.8393	0.8863
Semantic Similarity	0.8878	0.8987	0.8902	0.9023
BPR	1.2281	1.1881	1.2538	1.2889
LightGCN	1.2481	1.3074	1.3369	1.3327
SASRec	1.0317	1.1272	1.1730	1.1755
NeuralUCB-Diag	1.1845	1.2054	1.2837	1.2989
UpliftRec-LightGCN-HAdapt	1.2440	1.2956	1.3398	1.3287
CARec-HAdapt	1.3273	1.3213	1.3430	1.3303
CLCRec-HAdapt	1.3010	1.3474	1.3079	1.3082
LLM interest exploration	1.2048	1.2481	1.2527	1.2738
NORTH–Bank	**1.3317**	**1.3604**	**1.3661**	**1.3899**

Table 3: Independent OPEN proposals with a shared frozen scorer. Mean BRV (↑); every generator proposes its own destinations and uses the same real-content grounding and direction scorer Q. Bold marks the largest mean in each column.

	KuaiRec-Full		KuaiRec-FO	
Method	$W = 20$	$W = 50$	$W = 20$	$W = 50$
DreamRec-HAdapt + Q	1.3189	1.3375	1.3498	1.3597
PreferDiff-HAdapt + Q	1.3284	1.3406	1.3509	1.3713
NORTH	**1.3332**	**1.3453**	**1.3601**	**1.3774**

alignment. This comparison evaluates scoring on the same realizable actions. It retains the fixed-bank experiment's own evaluation protocol; its means are not subtracted from those of the separate OPEN runs to estimate a search increment.

5.3 INDEPENDENT OPEN DIRECTION SEARCH

Table 3 compares independent proposal mechanisms: DreamRec-HAdapt (Yang et al., 2023), PreferDiff-HAdapt (Liu et al., 2025), and NORTH. Each produces its own destinations from the permitted history and training evidence; no baseline receives NORTH's visited set. All use the same semantic realizer, admissibility conditions, and frozen direction scorer Q (denoted X in the implementation). HAdapt identifies the adaptation to the current history-to-direction interface, not a reproduction of the original next-item metric. The proposal-attempt budget is $S = 72$, distinct from the four-item realization budget.

NORTH achieves the highest mean BRV in all four settings. Under KuaiRec-Full, it reaches 1.3332 at $W = 20$ and 1.3453 at $W = 50$, exceeding PreferDiff-HAdapt by 0.0048 and 0.0047, respectively. Under KuaiRec-FO, NORTH obtains 1.3601 at $W = 20$ and 1.3774 at $W = 50$, with corresponding gains of 0.0092 and 0.0061 over PreferDiff-HAdapt. Relative to DreamRec-HAdapt, the gains are 0.0143, 0.0078, 0.0103, and 0.0177 in table order. These results show that the population-anchored policy consistently yields higher BRV than the two adapted diffusion-based proposal mechanisms under the shared evaluation interface.

Producing a continuous destination does not by itself imply discovery beyond the existing realization bank: a proposal may map back to a core whose members have already appeared in the bank. As defined in Section 3, bank-external discovery is determined by realized member identity rather than by the mere presence of a nonzero displacement in continuous semantic space.

5.4 ABLATING POPULATION SELECTION AND THE PROPOSAL POLICY

Table 4 compares the full policy with three variants. *Population-only bank* selects the eligible bank core with the largest A_C. *History-free proposal policy* removes χ_u and the personalized anchor score

Table 4: Ablations under the OPEN evaluation protocol. Mean BRV (↑). The history-free variant removes preference-history inputs only from the proposal policy; the final personalized scorer is retained.

	KuaiRec-Full		KuaiRec-FO	
Variant	$W = 20$	$W = 50$	$W = 20$	$W = 50$
Population-only bank	1.3158	1.3157	1.3561	1.3561
History-free proposal policy	1.3274	1.3265	1.3536	1.3672
Single-Gaussian displacement	1.3215	1.3426	1.3520	1.3689
Full NORTH	**1.3332**	**1.3453**	**1.3601**	**1.3774**

from proposal inputs while retaining the final personalized scorer. *Single-Gaussian displacement* replaces the four-component move distribution with one diagonal Gaussian, preserving routing, stay/move decisions, supervision, and search budget. Appendix E.3 gives the intervention boundaries.

The full method has the largest mean in all four ablation settings. Relative to population-only bank selection, its BRV is higher by 0.0174 and 0.0296 on Full, and by 0.0040 and 0.0213 on FO, for $W = 20$ and $W = 50$, respectively. These differences support using the complete history-conditioned direction-search system rather than population-only selection.

Removing history from the proposal policy lowers mean BRV by 0.0058 and 0.0188 on Full and by 0.0065 and 0.0102 on FO. Because the final scorer is unchanged, this comparison supports conditioning the proposal policy itself on the user, rather than relying on personalized scoring alone. Replacing the mixture with one Gaussian lowers the corresponding means by 0.0117, 0.0027, 0.0081, and 0.0085. Multiple displacement modes are useful in these comparisons, although the single-Gaussian policy remains competitive, particularly on Full at $W = 50$.

Scope of the evidence. The population-only contrast changes both personalized scoring and the proposal mechanism; it is not an isolated measurement of the benefit of moving away from identical seeds. The history-free and single-Gaussian controls isolate narrower policy choices. We therefore interpret the reported means as evidence about the complete system and these specified components, not as a guarantee that every new coordinate improves its parent direction. No statistical-significance claim is inferred from the four-decimal aggregate means. Appendix F examines the Full-trained results as an aggregate case study and illustrates how a returned direction can guide retrieval or content generation.

6 CONCLUSION

NORTH selects where to search next in semantic content space. Population feedback supplies behavioral structure and opportunity estimates; the target history personalizes those estimates and controls a stay-or-move policy. Continuous proposals are grounded in fixed-size real-content cores and evaluated by BRV. NORTH exceeds population-only bank selection in all four settings and leads the compared generators on KuaiRec-Full and KuaiRec-FO. The ablations support history-conditioned proposals and multiple displacement modes. The returned location and core provide an intermediate target for retrieval or generation, whose downstream performance remains a separate evaluation question.

AI USE STATEMENT

The authors used GPT-5.6 Sol (OpenAI) for two limited purposes during the preparation of this work: (1) writing assistance, specifically to polish the language and improve the clarity of author-written text, and (2) retrieval and discovery, specifically to help identify potentially relevant prior work and related literature. All substantive research contributions — including problem formulation, methodology, experiments, analysis, and interpretation — were conceived and conducted by the authors. The authors reviewed, verified, and take full responsibility for all content in the paper.

ETHICS STATEMENT

This work uses the released KuaiRec data and studies an offline direction-search task. Watch behavior is a response signal, not a direct measure of user welfare. A deployment should respect user control, privacy, and content-safety constraints when choosing or realizing a direction. Optimizing the best response in a small core should be considered alongside diversity and the risk of repeatedly steering users toward narrowly engaging content.

REPRODUCIBILITY STATEMENT

Section 4 specifies the representation, scorer, and search mechanism. Appendices A–D describe response measurement, content-only grounding, behavioral adaptation, policy supervision, and inference. Appendix E defines the baseline and ablation interfaces and records the method settings. The fixed-bank and OPEN experiments are reported separately.

REFERENCES

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM conference on recommender systems*, pp. 1007–1014, 2023.

J. Chen, W. Wang, C. Gao, P. Wu, J. Wei, and Q. Hua. Treatment effect estimation for user interest exploration on recommender systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. doi: 10.1145/3626772.3657736.

Yanhua Cheng, Bo Wang, Haotian Zhang, Xinyuan Gao, Zhihui Yin, Ben Xue, Yongzhi Li, Jieting Xue, Ye Ma, Minquan Wang, Jiahui Li, Tianyu Xu, Zhiqiang Liu, Xiao Lin, Shiyang Wen, Changcheng Li, Liu Liu, Quan Chen, Peng Jiang, and Kun Gai. Recommendation as generation: Unifying personalized video generation and recommendation at industrial scale, 2026. URL `https://arxiv.org/abs/2606.25496`. arXiv preprint arXiv:2606.25496.

A. N. Elmachtoub and P. Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

C. Gao, S. Li, W. Lei, J. Chen, B. Li, P. Jiang, X. He, J. Mao, and T.-S. Chua. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, 2022. doi: 10.1145/3511808.3557220.

S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 299–315, 2022. doi: 10.1145/3523227.3546767.

X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 173–182, 2017. doi: 10.1145/3038912.3052569.

X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648, 2020. doi: 10.1145/3397271.3401063.

Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 263–272, 2008. doi: 10.1109/ICDM.2008.22.

W.-C. Kang and J. McAuley. Self-attentive sequential recommendation. In *IEEE International Conference on Data Mining*, pp. 197–206, 2018. doi: 10.1109/ICDM.2018.00035.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010. doi: 10.1145/1772690.1772758.

J. Liao, S. Li, Z. Yang, J. Wu, Y. Yuan, X. Wang, and X. He. LLaRA: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1785–1795, 2024. doi: 10.1145/3626772.3657690.

Shuo Liu, An Zhang, Guoqing Hu, Hong Qian, and Tat-Seng Chua. Preference diffusion for recommendation. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=6GATHdOi1x`.

B. M. Marlin, R. S. Zemel, S. T. Roweis, and M. Slaney. Collaborative filtering and the missing at random assumption. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 267–275, 2007.

S. Rajput, N. Mehta, A. Singh, R. H. Keshavan, T. Vu, L. Heldt, L. Hong, Y. Tay, V. Q. Tran, J. Samost, M. Kula, E. H. Chi, and M. Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461, 2009.

T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1670–1679, 2016.

A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823, 2015.

A. van den Oord, S. Dieleman, and B. Schrauwen. Deep content-based music recommendation. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

P. Viappiani and C. Boutilier. Optimal bayesian recommendation sets and myopically optimal choice query sets. In *Advances in Neural Information Processing Systems*, volume 23, pp. 2352–2360, 2010.

M. Volkovs, G. Yu, and T. Poutanen. DropoutNet: Addressing cold start in recommender systems. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011. doi: 10.1145/2020408.2020480.

C. Wang, L. Yang, Z. Liu, X. Liu, M. Yang, Y. Liang, and P. S. Yu. Collaborative alignment for recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024a. doi: 10.1145/3627673.3679535.

H. Wang, N. Wang, and D.-Y. Yeung. Collaborative deep learning for recommender systems. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015. doi: 10.1145/2783258.2783273.

J. Wang, H. Lu, Y. Liu, H. Ma, Y. Wang, Y. Gu, S. Zhang, N. Han, S. Bi, L. Baugher, E. Chi, and M. Chen. LLMs for user interest exploration in large-scale recommendation systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024b. doi: 10.1145/3640457.3688161.

J. Wang, Y. Liu, Y. Sun, X. Ma, Y. Wang, H. Ma, Z. Su, M. Chen, M. Gao, O. Dalal, E. H. Chi, L. Hong, N. Han, and H. Lu. User feedback alignment for LLM-powered exploration in large-scale recommendation systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 996–1003, 2025. doi: 10.18653/v1/2025.acl-industry.70.

W. Wang, X. Lin, F. Feng, X. He, and T.-S. Chua. Generative recommendation: Towards next-generation recommender paradigm, 2023. arXiv preprint arXiv:2304.03516.

W. Wang, H. Bao, X. Lin, J. Zhang, Y. Li, F. Feng, S.-K. Ng, and T.-S. Chua. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024c.

Y. Wei, X. Wang, Q. Li, L. Nie, Y. Li, X. Li, and T.-S. Chua. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. doi: 10.1145/3474085.3475665.

Zhengyi Yang, Jiancan Wu, Zhicai Wang, Xiang Wang, Yancheng Yuan, and Xiangnan He. Generate what you prefer: Reshaping sequential recommendation via guided diffusion. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. Salakhutdinov, and A. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

D. Zhou, L. Li, and Q. Gu. Neural contextual bandits with UCB-based exploration. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11492–11502, 2020.

A Dataset Processing and Response Measurement

A.1 Population sources and target-user split

KuaiRec provides a large interaction log and a nearly fully-observed user–item subset (Gao et al., 2022). We use the names KuaiRec-Full and KuaiRec-FO to identify the source of population-training feedback. In the Full setting, 5,765 users outside the FO user set provide training evidence. In the FO setting, 903 users provide training evidence. Development and test use 226 and 282 held-out FO users, respectively. The source-user sets, target-user sets, and cached content coordinates are fixed before comparison.

Table 5: User-level split and shared evaluation interface. Counts denote the nominal split, before outcome-measurability filtering.

	KuaiRec-Full	KuaiRec-FO
Source-training users	5,765	903
Development users (FO)	226	226
Test users (FO)	282	282
Observed history budget W	$20, 50$	$20, 50$
Realization size K	4	4

The Full-trained behavioral model, direction scorer, and proposal policy are evaluated on target FO users without further training on FO feedback. Inferring a target state from its permitted history is adaptation of that user's state, not an additional population-training stage. Content features organize the catalogue; target-user held-out responses do not enter their construction.

A.2 Observed history and repeated exposure

For each user, the decision boundary separates the observed history from post-history outcomes. The history budget counts the observations supplied to the model. Repeated valid exposures can contribute repeated evidence to the history; a realization core, in contrast, always contains K distinct items. The observed prefix and the post-history outcome window remain separate throughout training and evaluation.

Let $\mathcal{E}^+_{ui}$ be the valid post-history events for a user–item pair. With event utility $g(r)$ defined below, the item outcome is

$$y_{ui} = \frac{1}{|\mathcal{E}^+_{ui}|} \sum_{e \in \mathcal{E}^+_{ui}} g(r_e), \qquad |\mathcal{E}^+_{ui}| > 0. \tag{19}$$

A pair with no valid event has no outcome. In an event-level history, the same item can have several observed residuals; these are kept within the history calculation, rather than treated as distinct real items at grounding time. Appendix C.1 shows the equivalent grouped, count-weighted ridge calculation.

A.3 Single-log response transformation

The response scale is based on the released watch ratio r. We use a single nonnegative log utility,

$$g(r) = \log\big(1 + \max\{r, 0\}\big). \tag{20}$$

The transformation is applied once to each valid event before repeated-event aggregation in Equation (19). A cache that already contains this utility is not transformed again. Missing or invalid records remain missing; they are not replaced by a zero response.

Why use this scale? BRV takes a maximum, so a very large response can have substantial influence on both a core target and its population average. The derivative $g'(r) = 1/(1 + r)$ for $r > 0$ reduces the influence of increasingly large ratios while preserving their ordering. The transformation retains a graded response rather than converting feedback into a binary label. It also puts item supervision, core supervision, and population opportunity on the same utility scale.

All subsequent quantities are computed from y: $V_u(C) = \max_{i \in C} y_{ui}$, A_C is the donor mean of $V_u(C)$, and the scorer fits $V_u(C)$ directly. There is no additional logarithm of BRV or of its residual, and no exponential back-transformation of the final score. The exponential in the routing prior or training weights converts scores into probabilities; it is not a reward-scale conversion.

For one observation per member, monotonicity gives

$$\max_{i \in C} g(r_{ui}) = g\left(\max_{i \in C} r_{ui}\right). \tag{21}$$

This identity does not justify interchanging the log with repeated-event averaging. In general, $\mathrm{Avg}_e\, g(r_e) \neq g(\mathrm{Avg}_e\, r_e)$. Nor does monotonicity make raw and log *population averages* identical in ranking. Reported BRV must therefore be interpreted on the documented utility scale.

A.4 Local donor support in KuaiRec-Full

For each core, define its source-training donor set by

$$D_C = \{v \in U_{\mathrm{tr}} : C \cap I(H_v) = \varnothing,\ O_{vi} = 1 \text{ for every } i \in C\}, \qquad n_C = |D_C|, \tag{22}$$

where O_{vi} indicates an observed post-history item outcome. The support check counts *distinct users*, not repeated events. We require $n_C \geq 10$ for the core-level population estimate. The same donor definition is used for scoring existing and newly grounded cores.

The purpose of this threshold is local estimation stability. With independent donor values of variance σ_C^2, the variance of their mean is σ_C^2/n_C. A core observed completely by one or two users provides a fragile population anchor. Requiring ten donors removes these extremely low-support cases while retaining a local, rather than whole-matrix, coverage requirement. The cutoff is a practical support requirement, not a universal confidence guarantee.

In KuaiRec-Full, the interaction matrix remains sparse. The rule identifies cores for which the population opportunity is estimable from that sparse source. It uses source-training observations and is separate from whether a test user's outcome is measurable. Logged exposure also affects which users enter D_C (Marlin et al., 2007; Schnabel et al., 2016; Swaminathan & Joachims, 2015); A_C is an average over these observed donors.

A.5 Outcome masks and coverage

For a returned core $\widehat{C}_u$, define

$$M_u = \prod_{i \in \widehat{C}_u} O_{ui}. \tag{23}$$

BRV is evaluated only when $M_u = 1$. The evaluator records the selected action before looking up the target outcomes. A method is not allowed to replace an unmeasurable action using the target's held-out responses. Missing outcomes are excluded from the value calculation, not counted as disliked content.

For methods a and b, paired effects use $U_{ab} = \{u : M_u^a = M_u^b = 1\}$. This definition distinguishes outcome coverage from value on a common cohort. The aggregate means in the main tables do not by themselves identify a paired confidence interval, because different selected cores can have different outcome masks.

B Semantic Geometry, Core Identity, and BRV

B.1 Content-only support

The frozen semantic coordinate is

$$x_i = P^\top (e_i - \bar{e}), \tag{24}$$

where the projection and centering are fitted to the permitted content representation. Let $\ell(c) > 0$ denote a content-only local scale. Following the local-scaling principle (Zelnik-Manor & Perona, 2004), the support can be written as

$$\mathcal{R}(c) = \{i : \|x_i - c\| \leq \rho(c)\}, \qquad \rho(c) = \min\{\ell(c), \rho_{\max}\}. \tag{25}$$

Grounding returns the nearest K distinct items in this support. A location with fewer than K supported members is rejected. The geometry is frozen during behavior learning, proposal training, and evaluation; response evidence is not used to move the content coordinates.

Externality is checked after grounding: the core must contain no item in $I(H_u)$. The source-training support requirement in Appendix A.4 is then applied. These checks should not be conflated with complete target feedback, which is an evaluation condition rather than a source of search information.

B.2 A NEW COORDINATE IS NOT NECESSARILY A NEW CORE

A core is identified by the sorted tuple of its K member item identifiers. Search proposals with the same tuple share one realizable action and one cached score. The original bank is a set of such identities,

$$\mathcal{B} = \{C^{(1)}, \ldots, C^{(L)}\}. \tag{26}$$

An OPEN discovery is bank-external when its grounded tuple is absent from $\mathcal{B}$. A small movement can change the nearest-neighbor set, while a larger movement can remain in the same set. Consequently, the dashed outline in Figure 2 is a schematic distinction between listed and unlisted cores, not a geometric boundary used by the algorithm.

For ordinary nearest-K grounding, and away from ties, the coordinates producing a given core satisfy

$$\|c - x_i\|^2 \le \|c - x_j\|^2 \quad \text{for all } i \in C,\ j \notin C. \tag{27}$$

Each comparison is a half-space constraint,

$$2c^\top (x_j - x_i) \le \|x_j\|^2 - \|x_i\|^2. \tag{28}$$

A support-radius check may further restrict eligible coordinates. Within a fixed core identity, both its observed BRV and its member-based score remain the same. Thus continuous coordinates provide a way to propose real-content sets; they do not make the observed objective smooth everywhere. Search samples coordinates and evaluates the grounded sets without differentiating through nearest-neighbor identities.

B.3 WHY A CORE-LEVEL TARGET DIFFERS FROM ITEM-SCORE POOLING

For any core member i, $\max_{j \in C} Y_{uj} \ge Y_{ui}$ pointwise. Taking the conditional expectation and then the maximum over i gives

$$\mathbb{E}[\max_{j \in C} Y_{uj} \mid H_u] \ge \max_{i \in C} \mathbb{E}[Y_{ui} \mid H_u]. \tag{29}$$

Equality need not hold. For example, let a two-item core have outcomes $(1, 0)$ or $(0, 1)$ with equal probability. Each item has mean $1/2$, but the core's best outcome is always 1. The distinction concerns the prediction target, even when item representations are expressive.

A fixed K is equally important: adding more items to a set cannot reduce its realized maximum. We therefore hold the number of members constant when comparing directions. BRV is a bounded-set opportunity measure, related to best-choice recommendation-set utility (Viappiani & Boutilier, 2010). Our offline search budget counts proposed locations; it does not reveal new target-user rewards during inference.

C BEHAVIORAL ADAPTATION AND DIRECTION-SCORE TRAINING

C.1 RIDGE TRANSFER AND REPEATED HISTORY ITEMS

Equation (3) has a unique solution because W_H is positive semidefinite and the identity contributes a strictly positive ridge penalty. For any nonzero vector z, $z^\top (I + \Phi_H^\top W_H \Phi_H) z = \|z\|^2 + \|W_H^{1/2} \Phi_H z\|^2 > 0$. The implementation can therefore solve the linear system without explicitly forming a matrix inverse. The transfer weights in Equation (6) depend on the full history design, not merely on a nearest historical item.

Suppose item i appears n_i times in a unit-weight history, with residuals $r_{i1}, \ldots, r_{in_i}$ and mean $\bar{r}_i$. Then

$$\sum_{j=1}^{n_i}(r_{ij} - \psi_i^\top v)^2 = n_i(\bar{r}_i - \psi_i^\top v)^2 + \sum_{j=1}^{n_i}(r_{ij} - \bar{r}_i)^2. \tag{30}$$

The second term is constant in v. An event-level history and a grouped history with count weights consequently give the same ridge solution when the supplied prefix is unchanged. This is a computational equivalence, not permission to replace an event budget with a different distinct-item budget.

C.2 Population centering and training-query exclusion

For a training query $u \in D_C$, its contribution is removed from the population moments:

$$A_C^{(-u)} = \frac{n_C A_C - V_u(C)}{n_C - 1}, \qquad \bar{\chi}_C^{(-u)} = \frac{n_C \bar{\chi}_C - \chi_u}{n_C - 1}. \tag{31}$$

When $u \notin D_C$, the moments are unchanged. These are leave-one-out population statistics; they do not make the source-trained representation an out-of-fold representation. Equation (10) applies to the common donor center from which the moments are formed.

C.3 Backbone training objective

Let $\mathbb{E}_{\mathrm{tr}}$ average valid targets within each user and then average users. Besides the BRV regression in Equation (11), let C_u^A be the population-preferred core in a user's training comparison set. When both outcomes are observed, the auxiliary loss is

$$L_{\mathrm{adv}} = \mathbb{E}_{\mathrm{tr}}\mathcal{H}_{0.5}\big([Q(u,C) - Q(u,C_u^A)] - [V_u(C) - V_u(C_u^A)]\big). \tag{32}$$

The reference is selected by population opportunity, not by the query user's best outcome. Item-response grounding and correction regularization are

$$L_{\mathrm{item}} = \mathbb{E}_{\mathrm{tr}}\mathcal{H}_1(\widehat{y}_{ui} - y_{ui}), \tag{33}$$

$$L_{\mathrm{corr}} = \mathbb{E}_{\mathrm{tr}}[(Q(u,C) - A_C)^2]. \tag{34}$$

Here $\mathcal{H}_\beta$ denotes the backbone's Huber-type regression penalty with transition scale β. The combined objective is

$$L_{\mathrm{backbone}} = L_{\mathrm{BRV}} + 0.2L_{\mathrm{adv}} + 0.2L_{\mathrm{item}} + \lambda_\delta L_{\mathrm{corr}} + \lambda_\theta\|\theta\|_2^2. \tag{35}$$

The training correction uses $\alpha = 1$. The resulting shared encoder and direction scorer are fixed before proposal-policy training. The policy stage does not optimize this backbone objective again.

D Search Supervision and Inference Details

D.1 Policy inputs and reference distribution

The per-anchor input is

$$\xi_{uj} = [\chi_u; c_j; w_{C_j}; A_{C_j}; \log(1 + n_{C_j}); Q(u, C_j)]. \tag{36}$$

Eligible bank cores are sorted by A_C. From the top 32, the anchor selector first admits cores with pairwise member Jaccard overlap at most 0.5; if fewer than eight are available, it relaxes this overlap restriction within the same shortlist. The population winner is retained. A representative coordinate is re-grounded to verify the corresponding core.

The policy uses two width-128 SiLU layers and separate output parameters for routing, stay probability, and a four-component diagonal Gaussian mixture. The route residual is $2\tanh\widetilde{r}_\psi$. The output layer is zero initialized, giving the population route prior and stay probability $s_0 = 0.25$. For a d-dimensional displacement, the reference component standard deviation is $\sigma_0 = 0.25/\sqrt{d}$. Reference means form two opposite pairs of fixed vectors, each of norm 0.04. Learned componentwise mean changes are bounded by $0.35/\sqrt{d}$, and standard deviations vary by at most a factor of two from σ_0. At inference, a sampled displacement with norm above 0.75 is radially rescaled to that norm. The policy likelihood is evaluated on the pre-rescaling Gaussian density; the norm limit is a deployment constraint, not an exact truncated-Gaussian likelihood.

D.2 Observed-action target masses

For each training user, let O_{uj} be the set of valid observed actions at anchor j. Let h_s indicate that the stay action has an observed outcome, and let h_m indicate that at least one move has an observed outcome. If N_m move representatives are observed, define

$$Z_{uj} = s_0 h_s + (1 - s_0) h_m, \qquad b_{jk} = \begin{cases} s_0 / Z_{uj}, & k \text{ is stay}, \\ (1 - s_0)/(N_m Z_{uj}), & k \text{ is an observed move}. \end{cases} \tag{37}$$

Equation (17) is normalized over all valid actions across the user's anchors. A missing observation is masked rather than assigned utility zero. Anchors without observed targets are omitted from the supervised routing normalization. If only stay or only move is observed at an anchor, its category log-probability is removed from the supervised likelihood; missing category evidence does not teach the model that the observed category wins. The prior penalty remains defined on all legal anchors.

Training coordinates are grounded before responses are attached, and repeated core identities are deduplicated within an anchor. The coordinate sampler includes the anchor and 48 local perturbation attempts, formed from 24 Gaussian directions and their negatives, at normalized scales 0.125, 0.25, and 0.5. Users with fewer than two observed action labels do not contribute a policy-training update. All support and population moments used for a training query exclude its own contribution. These targets form an observed-data learning objective; they do not correct all selection effects of logged exposure.

D.3 Prior regularization

Write $\pi_j = \pi_\psi(j \mid H_u)$, $s_j = s_\psi(u, j)$, and ω_{jm} for the learned mixture weights. The reference move weights are $1/M$, and component m has reference mean μ_m^0 and covariance $\Sigma_0 = \sigma_0^2 I$. Define

$$L_{\text{move},j} = \sum_m \omega_{jm} \left[\log(M\omega_{jm}) + D_{\text{KL}}\big(\mathcal{N}(\mu_{jm}, \Sigma_{jm}) \,\|\, \mathcal{N}(\mu_m^0, \Sigma_0)\big) \right]. \tag{38}$$

The policy regularizer is

$$L_{\text{prior}} = \mathbb{E}_u \left[D_{\text{KL}}(\pi \| \pi_0) + \sum_j \pi_j \left\{ D_{\text{KL}}(\text{Ber}(s_j) \| \,\text{Ber}(s_0)) + (1 - s_j) L_{\text{move},j} \right\} \right]. \tag{39}$$

The local term couples corresponding mixture components. It is the KL of the joint component/displacement representation and therefore an upper bound on the marginalized mixture KL, not its exact value. At zero initialization all components equal their references, so this penalty is zero. With $\lambda_{\text{prior}} = 0.1$, training uses AdamW for 40 epochs, learning rate 3×10^{-4}, weight decay 10^{-5}, batches of 32 users, and gradient-norm clipping at 1. Input features are standardized using source-training values, with a standard-deviation floor of 0.1.

D.4 Inference and computational boundary

Invalid, duplicate, and stay attempts count toward $S = 72$. A new coordinate that produces an already visited core does not create a new realizable action. The initial bank scan, population-statistic lookups, and geometric queries are additional costs. Keeping C_u^Q guarantees only that the maximum predicted score in the visited set is at least its original score; it does not guarantee an increase in the target user's realized BRV.

E Experimental Interfaces

E.1 Fixed-bank baselines

All methods in Table 2 choose among the same eligible bank cores within their setting. Random samples a core without response-based ranking; Semantic Similarity uses semantic history information. BPR and LightGCN use history-based new-user adaptation, while SASRec receives the causal observed sequence. For an item scorer $s_u(i)$, the adapter forms

$$\mathbf{s}_u(C) = \text{sort}_\downarrow\{s_u(i) : i \in C\}. \tag{40}$$

Algorithm 1 NORTH inference for one observed history

Require: History H_u; frozen geometry, encoder, and scorer Q; policy p_ψ; bank $\mathcal{B}$; TRAIN donor store; budget S

Infer χ_u from H_u using Equation (4).
Form the eligible external bank $\mathcal{B}_u$.
Select up to J population anchors with representative coordinates.
Find the population-best core C_u^A and bank-score incumbent C_u^Q.
Initialize $\mathcal{V}_u \leftarrow \{C_j\}_{j=1}^J \cup \{C_u^A, C_u^Q\}$ and retain coordinates.
for $t = 1, \ldots, S$ **do**
    Sample anchor j and a stay/move action from the joint policy.
    Set $c \leftarrow c_j$ for stay, or $c_j + \ell(c_j)\Delta c$ for move.
    Ground c to a real core $C = C_K(c)$.
    if C satisfies geometry, externality, and TRAIN-support conditions **then**
        Add its member identity to $\mathcal{V}_u$ if new; cache its coordinate and $Q(u, C)$.
    end if
end for
return $\widehat{C}_u = \arg\max_{C \in \mathcal{V}_u} Q(u, C)$ and its stored coordinate.

It compares cores lexicographically, first by the largest member score and then by subsequent scores to break ties. The evaluator uses the returned core's real outcomes, not these predicted item scores, to compute BRV.

NeuralUCB-Diag is adapted to a single decision using its diagonal uncertainty approximation; the experiment is not an online cumulative-regret evaluation. UpliftRec-LightGCN-HAdapt, CARec-HAdapt, and CLCRec-HAdapt use the available target history and the same item-to-core interface. The LLM interest-exploration baseline reads the observed history, content descriptions, and permitted TRAIN evidence to select a core. HAdapt identifies the common task interface rather than an exact reproduction of the original methods' industrial evaluation settings.

E.2 INDEPENDENT CONTINUOUS GENERATORS

DreamRec-HAdapt and PreferDiff-HAdapt independently generate continuous proposals conditioned on the permitted user history. Their outputs are adapted to the same frozen semantic-coordinate interface, then subjected to the common realizer, externality rule, and source-support requirement. The common frozen scorer Q, denoted X in the experiment implementation, selects among each generator's own grounded proposals. Consequently, Table 3 compares independently proposed opportunities under a shared value readout; it does not give the external generators NORTH's visited set, nor report their original next-item metrics as BRV. The proposal budget and real-content realization budget remain distinct.

The fixed-bank results and the independent OPEN runs are separate experiments. Their aggregate means are interpreted within each table rather than subtracted across tables as a search-only ablation. The current comparisons use the unfiltered catalogue scope.

E.3 ABLATION DEFINITIONS

Population-only bank. Select the eligible bank core with maximal A_C. This removes both personalized scoring and learned continuous proposals, so it tests the full system against a population reference rather than a single component.

History-free proposal policy. Train a separate policy without χ_u or the personalized anchor score as proposal inputs. The implementation masks the user-state features and replaces the personalized score input with A_C. Content geometry, core representation, population opportunity, support, and the final personalized scorer are retained. History-based externality is still required by the task; this control is not an entirely history-free recommendation system.

Single-Gaussian displacement. Use one diagonal Gaussian for moving instead of a mixture of four. Anchor selection, history conditioning, stay/move decisions, training supervision, proposal budget,

and final scorer follow the full method. This tests the displacement-distribution choice without replacing the direction-value model.

E.4 FIXED SETTINGS

Table 6: Settings of the population-anchored search policy.

Setting	Value
Realization size K	4
History budget W	20 or 50
Maximum population anchors J	8
Population shortlist size	32
Continuous proposal attempts S	72
Move mixture components M	4
Covariance family	Diagonal
Routing / target temperature	0.1 / 0.1
Target exponent clipping	$[-5, 5]$
Reference stay probability	0.25
Prior-penalty coefficient	0.1
Policy training epochs	40
Optimizer / learning rate	AdamW / 3×10^{-4}
Batch size / weight decay	32 / 10^{-5}
Random seed	42
Backbone during policy training	Frozen
Distinct TRAIN donors per core	$n_C \geq 10$

F DOWNSTREAM USE

F.1 THE RETURNED DIRECTION AS A DOWNSTREAM INTERFACE

The method returns a coordinate–core pair $(\widehat{c}_u, \widehat{C}_u)$. Its support radius and permitted content metadata can accompany this pair. The coordinate defines a semantic target; the core supplies concrete examples of what that location means. Its predicted score is an opportunity estimate, not a guarantee for a newly retrieved or generated item.

Retrieval example. A retrieval system can query its content index near $\widehat{c}_u$, remove previously observed or otherwise inadmissible items, and let its own ranker choose among the remaining realizations. For an allowed retrieval catalogue $\mathcal{I}_{\text{serve}}$, the first stage can be written as

$$\mathcal{R}_u = \text{Nearest}_L(\widehat{c}_u; \mathcal{I}_{\text{serve}} \setminus I(H_u)) \,. \tag{41}$$

Here L is a downstream retrieval-pool size, not the experimental realization size K. The policy has specified *where to retrieve*; the downstream ranker still decides *which item to deliver*.

Generation example The metadata of one returned core describes short, hands-on science demonstrations. A downstream system could turn the direction into the following brief:

> Create an original short science-demonstration video near the supplied semantic target. Use the accompanying core descriptions to identify the topic and presentation constraints, without copying their scripts or footage. Produce a concise script and shot plan, then check topic alignment and content safety before rendering.

A system such as RaG (Cheng et al., 2026) learns a direct connection between interest representations and video generation; NORTH instead supplies a content-grounded direction that such a downstream process could use.