ROBUST RECOMMENDATION UNDER UNKNOWN USER INTENT VIA WORST-OBJECTIVE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern recommender systems must simultaneously serve multiple user objectives, such as relevance, engagement, diversity, novelty, and serendipity. A common solution is to optimize a fixed objective or a user-specific scalarization of several objectives. However, even a personalized scalarization is typically fixed for a user, while the user's immediate intent can change across visits and even within a session. A policy that is excellent for the dominant intent can therefore fail severely when a less frequent intent becomes active. We study recommendation under an unknown, time-varying intent distribution and propose Objective-Diverse Recommendation (ODR), a robust learning principle that directly optimizes the weakest objective for each user rather than committing to a fixed scalarization. Our formulation views the unknown intent distribution as an adversary over a set of candidate objectives and reduces to a max–min ranking objective. We provide a policy-gradient procedure that, at each update and for each context, selects the currently worst-performing objective as the reward, thereby estimating the gradient of the max–min objective without requiring the true intent distribution at deployment. Synthetic experiments with relevance, diversity, novelty, and serendipity objectives, together with experiments on the KuaiRec full-feedback benchmark, show that the proposed approach maintains stable worst-objective performance across substantial shifts in the intent mixture. Its advantage becomes more substantial as the number of plausible intents increases. These results suggest that robust optimization over plausible objectives can be effective even when the user's active intent distribution is unknown and not explicitly estimated.

1 INTRODUCTION

Recommender systems have traditionally been optimized and evaluated through accuracy- or engagement-oriented metrics, despite long-standing evidence that accuracy alone does not characterize the quality of a recommendation experience (Herlocker et al., 2004; McNee et al., 2006; Zangerle & Bauer, 2022). A user may want a highly relevant item in one moment, broad exploration in another, and novel or serendipitous content in yet another. This observation has motivated a large literature on list diversification, novelty, serendipity, calibration, and other beyond-accuracy objectives (Carbonell & Goldstein, 1998; Ziegler et al., 2005; Zhang & Hurley, 2008; Lathia et al., 2010; Ge et al., 2010; Vargas & Castells, 2011; Kaminskas & Bridge, 2016; Steck, 2018).

Most existing approaches nevertheless make one of two commitments. The first is to use a common trade-off between relevance and beyond-accuracy objectives for all users. The second, increasingly common in industrial multi-objective recommendation, is to personalize the scalarization itself, for example by learning user-specific weights in a reward such as $a\, r_{\mathrm{rel}} + b\, r_{\mathrm{div}} + c\, r_{\mathrm{nov}}$ (Zheng & Wang, 2021; Jeunen et al., 2024; Huang et al., 2024). The latter is substantially more flexible, but it still tends to commit each user to one preferred trade-off. This is problematic when user intent changes over time. Recent work on intent-aware recommendation explicitly emphasizes that the same user's goal can vary across visits and within sessions (Jannach & Zanker, 2024; Bhattacharya & Lamkhede, 2022; Oh et al., 2024). Large-scale industrial systems similarly combine specialized recommendation surfaces or increasingly construct entire pages end-to-end (Gomez-Uribe & Hunt, 2015; Hsiao et al., 2025; Wang et al., 2026). More importantly, personalization across users does not address objective variation within a user, whose preferred recommendation criterion may change across visits or moments.

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.

We ask a different question: *How should a ranking policy be learned when the set of plausible user intents is known, but their conditional distribution is unknown and may arbitrarily change over time?* Rather than estimate a single intent mixture and optimize its expectation, we focus on avoiding catastrophic underperformance for any plausible intent. This is especially relevant when a severe mismatch may create dissatisfaction or abandonment. We observe this asymmetry in an analysis of real-world behavioral data collected from a production content recommendation platform. As shown in Figure 1, churn increases sharply at high levels of dissatisfaction (i.e., a high share of videos abandoned shortly after playback begins, used here as a behavioral proxy for poor recommendation outcomes). This observational asymmetry motivates a shift in emphasis from further improving already satisfactory outcomes toward preventing highly dissatisfying ones, which are associated with a much sharper increase in churn. This is consistent with prior work showing that optimizing aggregate interaction outcomes can differ fundamentally from optimizing long-term user retention (Kishimoto et al., 2026; Wu et al., 2017). In particular, improving poor user outcomes can contribute more to retention than further improving already satisfactory outcomes (Kishimoto et al., 2026).

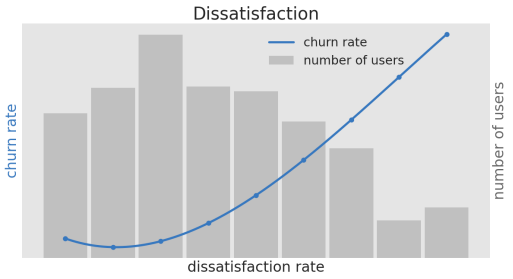

Figure 1: Relationship between dissatisfaction and subsequent churn in a production content recommendation platform. Dissatisfaction is the share of valid non-trailer watches abandoned shortly after playback begins. A user active during a given period is counted as churned if they have no activity in the subsequent period. The blue line shows churn rate and the gray bars show the number of users in each bin. Churn increases sharply at high dissatisfaction levels.

This asymmetry motivates our Objective-Diverse Recommendation (ODR), which prioritizes the worst-performing plausible intent for each context, directly targeting severe intent-level failures rather than allocating additional optimization pressure to objectives that already perform well. Specifically, for a context x, ranking policy π, and intent z, let $g(x, \pi; z) \in [0, 1]$ denote a higher-is-better intent-specific ranking score. If the distribution $p(z \mid x)$ is completely unknown, then minimizing the worst expected failure $(1 - g(x, \pi; z))$ over all distributions on the intent set is equivalent to maximizing the score of the worst intent as follows.

$$\min_{\pi} \mathbb{E}\left[\sup_{p(z|x)} \sum_{z} p(z \mid x)\big(1 - g(x, \pi; z)\big)\right] = \min_{\pi} \mathbb{E}\left[\max_{z}\big(1 - g(x, \pi; z)\big)\right]$$
$$= 1 - \max_{\pi} \mathbb{E}\left[\min_{z} g(x, \pi; z)\right].$$

This objective favors a ranking that balances multiple intents when specializing in a single intent leads to poor performance on other intents. For example, a relevance-only ranking may score $(0.9, 0.2)$ on relevance and diversity, while a diversity-only ranking scores $(0.5, 0.8)$. A mixed ranking scoring $(0.8, 0.6)$ is preferred because its minimum score, 0.6, exceeds 0.2 and 0.5. Our main contributions are as follows.

- We formalize *objective diversity within a user*: recommendation under a set of plausible but distributionally unknown intents, where robustness is measured by the worst intent-specific ranking performance for each user.
- We derive a max–min objective that does not require the true intent distribution and distinguish it from fixed or user-specific linear scalarization.
- We give a practical policy-gradient procedure that repeatedly selects the worst-performing metric for each context and updates the policy using that metric as the reward.
- We evaluate the proposed approach in controlled synthetic environments and on KuaiRec, showing stable worst-objective performance across intent-mixture shifts and increasing gains as the number of plausible intents grows.

2 Related Work

The need to move beyond predictive accuracy in recommender systems has long been recognized (Herlocker et al., 2004; McNee et al., 2006). Maximal Marginal Relevance introduced a canonical relevance–diversity trade-off (Carbonell & Goldstein, 1998), followed by work on topic and list diversification (Ziegler et al., 2005; Zhang & Hurley, 2008). Subsequent research has studied temporal diversity (Lathia et al., 2010), coverage, serendipity, and rank-aware novelty/diversity (Ge et al., 2010; Vargas & Castells, 2011), with surveys synthesizing these and broader recommender objectives (Kaminskas & Bridge, 2016; Alhijawi et al., 2022; Zangerle & Bauer, 2022). Calibration aligns recommendation lists with target interest distributions (Steck, 2018), while performative approaches can encourage diverse content supply over time (Eilat & Rosenfeld, 2023). Our work differs by introducing diversity in the *optimization pressure across objectives*, rather than only across items or toward a fixed target distribution, to protect against an unknown active intent.[1]

Intent-aware recommendation recognizes that a user's goal can vary across visits or sessions (Jannach & Zanker, 2024). Prior work includes in-session adaptation in Netflix Search (Bhattacharya & Lamkhede, 2022) and hierarchical multi-task models for predicting latent session intent (Oh et al., 2024). Industrial systems combine specialized models and business objectives (Gomez-Uribe & Hunt, 2015; Hsiao et al., 2025), while recent generative systems construct entire homepages conditioned on user and request context (Wang et al., 2026). These approaches infer or condition on context and intent; ODR instead addresses the complementary regime where active intent remains uncertain and intent prediction may be unavailable or wrong.

Multi-objective recommendation commonly uses scalarization, multi-task learning, Pareto methods, or learned policies over scalarization parameters (Zheng & Wang, 2021; Jeunen et al., 2024; Huang et al., 2024) to trade off engagement, revenue, fairness, diversity, and related goals. Our distinction is between *optimizing an expected or selected trade-off* and *protecting the worst plausible intent*. Even an oracle scalarization using the true average intent distribution can favor a dominant intent while poorly serving a minority intent. ODR therefore changes the optimization target rather than merely changing how scalarization weights are selected.

Our setting also connects to offline policy learning, where logged interactions require correcting selection bias. Counterfactual risk minimization and inverse-propensity methods provide foundations for policy learning (Swaminathan & Joachims, 2015; Schnabel et al., 2016), including slate estimators (Swaminathan et al., 2017). Deep models can learn from logged bandit feedback (Joachims et al., 2018), with top-K off-policy corrections (Chen et al., 2019) and reinforcement-learning decompositions such as SlateQ (Ie et al., 2019). ODR is orthogonal to the off-policy estimator, requiring only intent-specific policy-value estimates appropriate to the logging regime.

Finally, our max–min perspective relates to distributionally robust and worst-group learning. DRO protects against adverse distributions or groups rather than minimizing average risk (Duchi & Namkoong, 2021); worst-group objectives address repeated loss disparity (Hashimoto et al., 2018), and group DRO targets worst-case generalization (Sagawa et al., 2020). Robust policy evaluation and learning have likewise been studied for offline contextual bandits under distribution shift (Si et al., 2020; Yang et al., 2023). ODR differs in the adversarial dimension: it ranges over *latent user objectives/intents for the same context*, rather than demographic groups or covariate distributions, yielding an objective-space robustness criterion.

An extended discussion of these connections, including recent max–min multi-objective reinforcement learning, is provided in Appendix A.

3 Problem Formulation

3.1 Contextual ranking with latent intent

Let $x \in \mathcal{X}$ denote the observable recommendation context, including user features, history, and request/session features. Let $\mathcal{A}(x)$ be the candidate item set and let $y = (a_1, \ldots, a_K)$ denote a length-K ranking without replacement. A stochastic ranking policy $\pi_\theta(y \mid x)$ is parameterized by θ.

[1] We use *active intent* to denote the latent objective that is operative at a particular recommendation moment.

We assume a finite set of plausible intents $\mathcal{Z} = \{1, \ldots, M\}$. Each intent z is associated with a higher-is-better raw ranking metric $g_z^{\mathrm{raw}}(x, y)$. Examples include relevance/nDCG, category coverage, novelty, serendipity, watch ratio, or watch probability. Its raw policy value is

$$G_z^{\mathrm{raw}}(x; \pi_\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[g_z^{\mathrm{raw}}(x, y) \right].$$

Because max–min optimization compares objectives directly, we normalize every metric by its improvement rate over the reference policy π_0 (e.g., a logging policy or a currently deployed production policy), i.e.,

$$G_z(x; \pi_\theta) = \frac{G_z^{\mathrm{raw}}(x; \pi_\theta) - G_z^{\mathrm{raw}}(x; \pi_0)}{|G_z^{\mathrm{raw}}(x; \pi_0)|}.$$

Equivalently, the normalized ranking reward is $\widetilde{g}_z(x, y) = \big(g_z^{\mathrm{raw}}(x, y) - G_z^{\mathrm{raw}}(x; \pi_0)\big) / \big(|G_z^{\mathrm{raw}}(x; \pi_0)|\big)$, so that $G_z(x; \pi_\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[\widetilde{g}_z(x, y)]$.

The active intent is latent. At context x, it follows an unknown conditional distribution $p(z \mid x)$. The conventional expected-utility objective would be

$$J_{\mathrm{avg}}(\pi_\theta; p) = \mathbb{E}_x \left[\sum_{z \in \mathcal{Z}} p(z \mid x) G_z(x; \pi_\theta) \right].$$

This is appropriate when $p(z \mid x)$ is accurately known and stable. **The key distinction in our setting is that the set of plausible intents is known and meaningful, but their context-dependent mixture is unknown and potentially time-varying.** Consequently, J_{avg} cannot be reliably specified, increasing the risk of severe intent-level failures.

3.2 Worst-intent risk

Define intent-specific failure as $1 - G_z(x; \pi_\theta)$ after normalization. With no trustworthy information about the intent mixture, we consider the full probability simplex Δ_M as the ambiguity set, where q_z denotes the probability assigned to intent z by a candidate intent distribution $q \in \Delta_M$:

$$V(\pi_\theta) = \mathbb{E}_x \left[\sup_{q \in \Delta_M} \sum_{z \in \mathcal{Z}} q_z \left(1 - G_z(x; \pi_\theta)\right) \right].$$

Lemma 1 (simplex adversary). For any fixed x and policy π_θ,

$$\sup_{q \in \Delta_M} \sum_z q_z \left(1 - G_z(x; \pi_\theta)\right) = 1 - \min_z G_z(x; \pi_\theta).$$

The equality follows because a linear function over the simplex is maximized at a vertex, placing all mass on the largest failure (the smallest utility). Thus minimizing worst-case failure is equivalent to

$$\pi^\star \in \arg\max_{\pi_\theta} \mathbb{E}_x \left[\min_{z \in \mathcal{Z}} G_z(x; \pi_\theta) \right].$$

The formulation is deliberately conservative. It is most appropriate when a rare but severe intent mismatch matters more than a small reduction in the dominant objective. This choice is consistent with the asymmetry observed in Figure 1, where severe dissatisfaction is associated with a much sharper increase in churn than the reduction in churn associated with further increases in satisfaction.

3.3 Evaluation by relative improvement

To compare heterogeneous ranking metrics, we evaluate each objective using the same context-level normalized score $G_z(x; \pi)$ introduced above. For interpretability, we report the average improvement of each individual objective,

$$\mathrm{Imp}_z(\pi) = \mathbb{E}_x \left[G_z(x; \pi) \right].$$

Our primary robustness metric, however, evaluates the worst objective separately for each context before averaging, i.e.,

$$\mathrm{WorstImp}(\pi) = \mathbb{E}_x \left[\min_{z \in \mathcal{Z}} G_z(x; \pi) \right].$$

This distinction is important because the identity of the worst-performing objective may vary across contexts. In particular, $\mathrm{WorstImp}(\pi) \leq \min_{z \in \mathcal{Z}} \mathrm{Imp}_z(\pi)$, since the minimum is taken before averaging on the left-hand side. Consequently, all objectives can exhibit favorable aggregate improvements while some contexts still suffer from a poorly served intent.

4 Proposed Method

4.1 From fixed scalarization to worst-objective policy gradients

A fixed scalarization optimizes $J_{\mathrm{lin}}(\theta) = \mathbb{E}_x \left[\sum_z w_z(x) G_z(x; \pi_\theta)\right]$, where $w_z(x)$ may be global or user-specific. Once w is fixed, optimization pressure can remain concentrated on objectives with large weights even when another objective performs poorly. Our ODR instead directly optimizes

$$J_{\min}(\theta) = \mathbb{E}_x \left[\min_{z \in \mathcal{Z}} G_z(x; \pi_\theta)\right].$$

Appendix B provides a minimax interpretation of $J_{\min}$ and shows that even oracle scalarization with the true intent frequencies can sacrifice a rare intent that the max–min objective protects.

At each optimization loop, ODR estimates the normalized utility of every metric for each sampled context and selects the currently worst-performing metric

$$z^\star(x; \theta) \in \arg\min_{z \in \mathcal{Z}} G_z(x; \pi_\theta).$$

The selected metric is then used as the reward for the policy-gradient update at that context. Intuitively, every update asks *which plausible user intent is the current policy serving least well for this context?* The policy is then improved specifically with respect to that intent. Because this selection is context-specific, different contexts within the same optimization loop may be updated using gradients from different objectives.

When the minimizing intent is unique, the gradient of the pointwise minimum follows the active objective of $\nabla_\theta \min_z G_z(x; \pi_\theta) = \nabla_\theta G_{z^\star(x;\theta)}(x; \pi_\theta)$. Appendix B provides formal minimax, gradient-characterization, worst-metric estimation-error, regularization technique, and finite-sample analyses for this robust objective and update rule.

4.2 Policy-gradient estimator

For a discrete ranking policy, let $y \sim \pi_\theta(\cdot \mid x)$ be the sampled ranking and let $z^\star = z^\star(x; \theta)$ be the worst metric under the current policy. We define the policy-gradient reward to be the reference-policy-normalized value of that metric, $r(x, y) = \widetilde{g}_{z^\star}(x, y)$. A score-function estimator is

$$\nabla_\theta G_{z^\star}(x; \pi_\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[(\widetilde{g}_{z^\star}(x, y) - b(x)) \nabla_\theta \log \pi_\theta(y \mid x)\right],$$

where $b(x)$ is an optional variance-reduction baseline (Swaminathan & Joachims, 2015; Schnabel et al., 2016; Swaminathan et al., 2017; Chen et al., 2019).

For a minibatch $\mathcal{B}$, one optimization loop estimates

$$\widehat{\nabla_\theta J_{\min}} = \frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} \widehat{\nabla_\theta G_{z_i^\star}}(x_i; \pi_\theta), \quad z_i^\star \in \arg\min_z \widehat{G}_z(x_i; \pi_\theta),$$

and updates θ in this direction. Thus this update directly estimates the gradient of $\mathbb{E}_x[\min_z G_z(x; \pi_\theta)]$ at every loop rather than introducing a separately learned or smoothed objective weighting.

Why ODR produces objective-diverse rankings. Suppose relevance-specialized rankings achieve high relevance but low diversity, whereas exploration-heavy rankings do the reverse. A fixed scalarization can remain specialized when its dominant weight is large. Under ODR, the neglected objective becomes the selected reward for subsequent updates, so its policy gradient receives direct optimization pressure. The policy is therefore pushed toward rankings that raise the minimum utility, potentially assigning different objective emphases to different positions in a slate.

Inference. Worst-metric selection is used only during training. At serving time ODR uses the learned policy π_θ like a standard ranking model and does not require the active intent or its distribution.

Algorithm 1 Worst-objective policy-gradient training for Objective-Diverse Recommendation (ODR)

Require: Logged data $\mathcal{D}$; intent set $\mathcal{Z}$; ranking policy π_θ; learning rate η

Initialize policy parameters θ
for $t = 1, \ldots, T$ **do**
  Sample a minibatch $\mathcal{B} = \{x_i\}_{i=1}^{B}$ from $\mathcal{D}$
  for each $x_i \in \mathcal{B}$ **do**
    for each $z \in \mathcal{Z}$ **do**
      Estimate normalized utility $\widehat{G}_{i,z} \leftarrow \widehat{G}_z(x_i; \pi_\theta, \mathcal{D})$
    end for
    Select the worst metric $z_i^\star \leftarrow \arg\min_{z \in \mathcal{Z}} \widehat{G}_{i,z}$
    Sample/evaluate $y_i \sim \pi_\theta(\cdot \mid x_i)$ and set reward $r_i \leftarrow \widetilde{g}_{z_i^\star}(x_i, y_i)$
    Estimate $\widehat{h}_i \leftarrow (r_i - b_i)\nabla_\theta \log \pi_\theta(y_i \mid x_i)$ (with the required off-policy correction)
  end for
  Estimate $\widehat{\nabla_\theta J_{\min}} \leftarrow \frac{1}{B}\sum_{i=1}^{B} \widehat{h}_i$
  Update $\theta \leftarrow \theta + \eta\, \widehat{\nabla_\theta J_{\min}}$
end for
return π_θ

Relationship to personalized reward design. A learned reward design such as $\pi(a, b, c \mid \text{user})$ chooses coefficients for a scalar reward and then trains a recommender against that reward (Jeunen et al., 2024). ODR instead redefines the reward at every optimization loop and context as the currently worst-performing metric. Thus, ODR uses a more adaptive reward design in the sense that the optimization target can change across contexts and training loops according to the currently worst-served plausible intent, rather than remaining tied to a selected scalar trade-off.

5 Synthetic Experiment

Experimental design. We first construct a controlled environment in which optimizing a single relevance objective can severely damage discovery-oriented objectives. There are 10 actions, ranking length $K = 5$, four categories, six-dimensional contexts sampled from a standard Gaussian, and a default logged sample size of $1{,}000$. The four intent-specific metrics are relevance (nDCG), category coverage, novelty, and serendipity. Items are divided into two broad types. *Relevance* items have high relevance but are concentrated in one category and tend to have lower discovery scores. *Discovery* items have more moderate relevance but are spread across categories and tend to have higher novelty and serendipity. This construction creates an explicit tension between exploiting currently relevant items and maintaining discovery quality.

The intent set is $\mathcal{Z} = \{\text{rel}, \text{div}, \text{nov}, \text{ser}\}$. We parameterize the context-independent intent distribution by $\rho \in [0, 1]$ as $p(\text{Rel}) = 0.9 - \rho/2$, $p(\text{Div}) = 0.1$, and $p(\text{Novel}) = p(\text{Seren}) = \rho/4$. Thus, increasing ρ shifts probability mass from the relevance intent toward the novelty and serendipity intents, while keeping the diversity-intent probability fixed. Small ρ therefore represents relevance-heavy demand, whereas large ρ represents more discovery-oriented demand. Importantly, the proposed method does not observe ρ.

Baselines. We compare five methods under the same neural policy class and hyperparameters.[2]

1. **Primary-only**: maximize DCG and ignore diversity, novelty, and serendipity.
2. **Equal scalarization**: maximize the unweighted mean of the four metrics without intent-distribution information.
3. **Oracle intent-weighted scalarization**: maximize a weighted sum using the true intent mixture. This baseline has information unavailable to ODR and is therefore intentionally favorable.
4. **Constrained optimization**: maximize DCG subject to the non-DCG metrics not falling below the reference-policy reference.

[2]The Appendix provides the exact neural architecture, ranking distribution, optimizer, train/validation/test split, number of random seeds, and hyperparameter-search protocol.

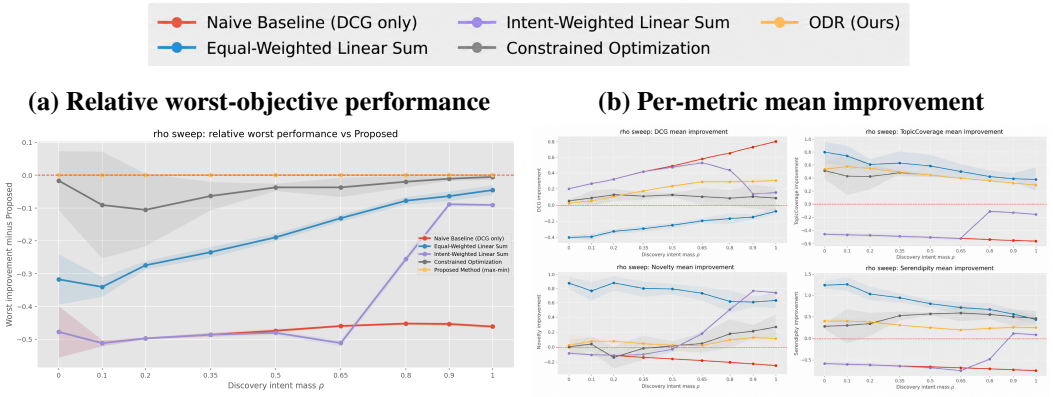

Figure 2: Synthetic results across different intent mixtures controlled by ρ. (a) Relative worst-objective performance, defined as $\mathrm{WorstImp}(\pi_{\mathrm{ODR}}) - \mathrm{WorstImp}(\pi)$ for each comparison method; positive values indicate an advantage for ODR, and the zero line represents parity with ODR. (b) Mean improvement of individual metrics.

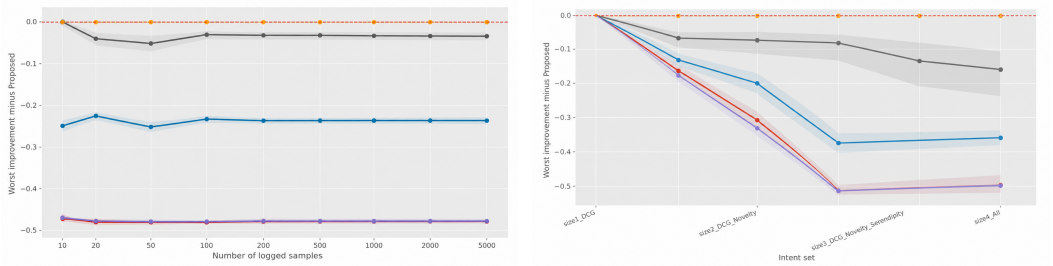

Figure 3: Robustness analyses in the synthetic environment. Relative performance is defined as $\mathrm{WorstImp}(\pi_{\mathrm{ODR}}) - \mathrm{WorstImp}(\pi)$, so positive values indicate an advantage for ODR. (a) Relative worst-objective performance as the number of logged samples varies. (b) Relative worst-objective performance as the number of plausible intents increases.

5. **ODR**: optimize the robust objective without access to the intent mixture.

Results across intent mixtures. Across the ρ sweep, ODR achieves higher worst-objective performance than the baselines across a broad range of the examined intent mixtures, as indicated by the positive relative gaps in Figure 2(a). Figure 2(b) further illustrates the underlying objective-level trade-offs. The primary-only and equal-scalarization policies exhibit opposite trade-offs between relevance and discovery-oriented objectives in relevance-heavy regimes. More specifically, the primary-only policy drives category coverage, novelty, and serendipity below the reference policy, whereas equal scalarization improves these discovery metrics at the cost of negative DCG improvement. The oracle intent-weighted scalarization also remains specialized in relevance-heavy regimes, achieving strong DCG while leaving several discovery metrics poorly served. Constrained optimization avoids such severe collapses and is therefore the closest competitor, but still yields lower worst-objective performance than ODR. These results illustrate that even knowing the *average* intent mixture is not equivalent to protecting all plausible intents. ODR instead maintains a more balanced profile across the individual objectives while protecting the weakest one. Notably, these advantages are obtained without ODR observing the underlying intent-mixture parameter ρ, and without a collapse in overall metric performance.

Data size and number of intents. We vary the amount of logged data while keeping the policy class fixed in Figure 3 (a). The relative gaps remain generally small between ODR and constrained optimization as the amount of logged data varies, with ODR retaining a modest advantage in all the examined settings. In Figure 3 (b), we grow the true intent set from one objective (DCG) to two (adding novelty), three (adding serendipity), and four (adding category coverage). When there is only one intent, all methods solve essentially the same problem. As additional intents are introduced, the relative advantage of ODR becomes more substantial, which is consistent with the hypothesis that fixed scalarizations become increasingly brittle as objective uncertainty expands.

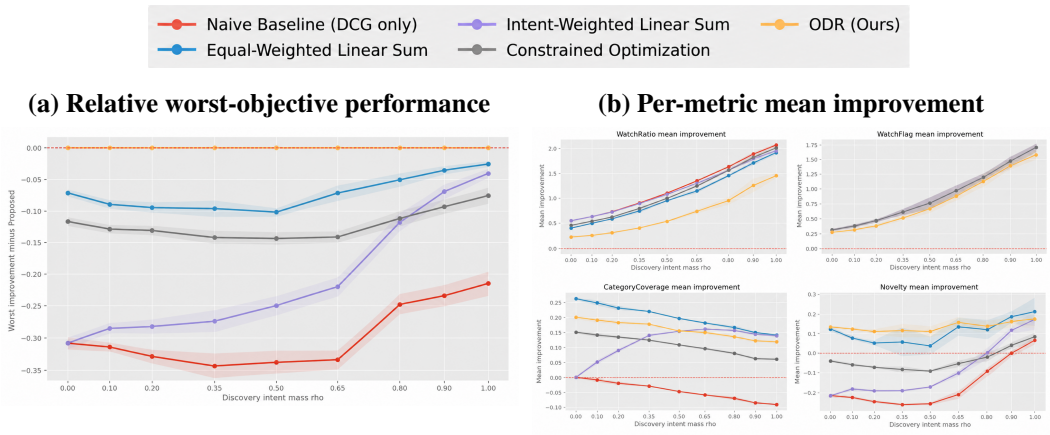

Figure 4: KuaiRec results across different intent mixtures controlled by ρ. (a) Relative worst-objective performance, defined as $\mathrm{WorstImp}(\pi_{\mathrm{ODR}}) - \mathrm{WorstImp}(\pi)$; positive values indicate that ODR better protects the per-context worst objective, while the zero line denotes parity with ODR. (b) Mean improvement of the individual WatchRatio, WatchFlag, CategoryCoverage, and Novelty metrics.

6 Real-world Experiment

KuaiRec setup We next evaluate on KuaiRec (Gao et al., 2022), using its fully observed small user–item matrix. We retain $1{,}411$ users and uniformly subsample 500 short videos from $3{,}327$ candidate items; the ranking length is $K = 5$. The main KuaiRec experiment is a full-feedback candidate-ranking benchmark. Rather than fitting an end-to-end neural contextual ranking model, we construct the same finite set of candidate rankings for every method and select among them according to each method's objective.

We define four intents using the observed feedback and logged exposures:

- **Watch ratio**: nDCG@K using the clipped and rescaled watch ratio as the item utility.
- **Watch flag**: nDCG@K using the indicator that the raw watch ratio is above zero.
- **Diversity**: normalized category coverage in the top K.
- **Novelty**: nDCG@K using the negative log marginal exposure frequency estimated from the generated training logs, with Laplace smoothing and min–max normalization.

For a given ρ, the simulated logging policy uses $p(\mathrm{ratio}) = 0.9 - \rho/2$, $p(\mathrm{flag}) = 0.1$, and $p(\mathrm{div}) = p(\mathrm{nov}) = \rho/4$ to combine watch ratio, watch flag, category rarity, and item rarity. Independent Gumbel noise with scale 0.12 is added to the resulting item scores before taking the top K items.

We use the same five baseline families as in the synthetic experiment, replacing DCG-only with watch-ratio-only optimization. All methods select from exactly the same candidate rankings; ODR selects the ranking with the largest worst improvement over an independently generated logging-policy ranking as a reference. Its selection objective does not use ρ, although the common candidate pool contains a ranking generated from the true intent vector for all methods.

Results across intent mixtures. The real-data results in Figure 4 (a) reproduce the central synthetic finding. Across the ρ range, the relative worst-objective gaps show that ODR outperforms the competing objectives over a broad range of intent mixtures. The primary-only method predictably performs strongly on watch ratio but can harm category coverage and novelty. Equal scalarization improves discovery-oriented metrics but can sacrifice watch-oriented quality. The intent-weighted scalarization changes with the true mixture yet still exhibits regimes in which its worst-objective performance falls below ODR. ODR reasonably trades some peak performance on the easiest objective for protection against severe degradation on the weakest one.

The average-metric curves in Figure 4 (b) are also stable, suggesting that the method does not merely equalize objectives at a uniformly poor level. Although max–min optimization can be overly

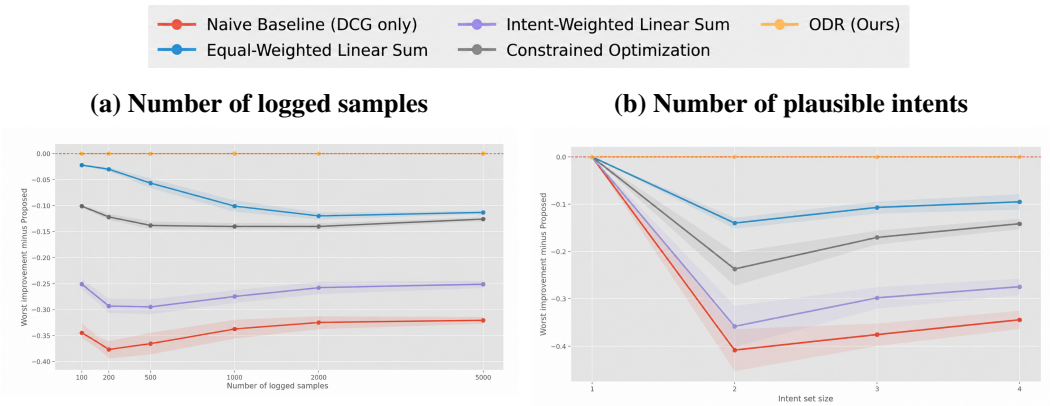

Figure 5: Robustness analyses on KuaiRec. Relative performance is defined as $\mathrm{WorstImp}(\pi_{\mathrm{ODR}}) - \mathrm{WorstImp}(\pi)$, with positive values indicating an advantage for ODR. (a) Relative worst-objective performance as the amount of logged data varies. (b) Relative worst-objective performance as the number of plausible intents increases.

conservative, the stable average-metric curves indicate that the improved worst-objective performance does not come with a broad degradation in average quality of each metric.

Logged-data size and intent-set size. When varying the number of logged samples in Figure 5 (a), the relative gap between ODR and constrained optimization remains modest, while ODR retains an advantage in the studied regimes. We then grow the true intent set from one to four objectives in Figure 5 (b): WatchRatio; WatchRatio + Novelty; WatchRatio + Novelty + CategoryCoverage; and finally all four objectives including WatchFlag. As in the synthetic experiment, all methods are effectively equivalent in the single-intent case, yielding relative gaps near zero, while ODR's advantage clearly increases as objective uncertainty grows.

What the experiment does and does not establish. KuaiRec provides a useful full-feedback testbed, but it is still an offline benchmark. The experiment establishes robustness across multiple plausible recommendation objectives under controlled shifts in their mixture. It does not by itself prove that max–min metric robustness causally reduces user churn or increases subscription revenue. A production A/B test with retention-sensitive outcomes would be required for that stronger claim.

7 Conclusion

We introduced objective diversity as a new recommendation principle for users whose active intent is unknown. Rather than learn a fixed global or personalized scalarization, ODR optimizes a robust max–min objective over a set of plausible intents. The resulting training rule repeatedly uses the currently weakest metric for each context as the policy-gradient reward and requires no intent-distribution estimate at serving time. Across the synthetic and KuaiRec experiments, ODR maintains more stable worst-objective performance than the single-objective, intent-weighted, and constrained baselines.

Several limitations remain. First, ODR assumes that the intent set itself is correctly specified; robustness to a misspecified or open-ended intent set remains unresolved. Second, the full-simplex adversary is deliberately conservative and may sacrifice unnecessary utility when reliable intent-frequency information is available. Ambiguity sets centered on estimated intent distributions could interpolate between expectation and worst-case robustness. Third, our current evaluation uses offline ranking metrics as intent proxies. Direct evidence on satisfaction and retention requires online experimentation and longitudinal outcomes. Finally, richer page-level policies, in which different rows or positions optimize different objectives, are a natural extension in modern end-to-end homepage construction (Wang et al., 2026).

AI USE STATEMENT

Generative AI tools were used for language editing, wording refinement, and title brainstorming. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content.

REFERENCES

Bushra Alhijawi, Arafat Awajan, and Salam Fraihat. Survey on the objectives of recommender systems: Measures, solutions, evaluation methodology, and new perspectives. *ACM Computing Surveys*, 55(5):1–38, 2022. doi: 10.1145/3527449.

Moumita Bhattacharya and Sudarshan Lamkhede. Augmenting netflix search with in-session adapted recommendations. *arXiv preprint arXiv:2206.02254*, 2022.

Woohyeon Byeon, Giseung Park, Jongseong Chae, Amir Leshem, and Youngchul Sung. Multi-objective reinforcement learning with max-min criterion: A game-theoretic approach. *arXiv preprint arXiv:2510.20235*, 2025.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998. doi: 10.1145/290941.291025.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 456–464, 2019. doi: 10.1145/3289600.3290999.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021. doi: 10.1214/20-AOS2004.

Itay Eilat and Nir Rosenfeld. Performative recommendation: Diversifying content via strategic incentives. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9082–9103. PMLR, 2023.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng 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*, pp. 540–550, 2022. doi: 10.1145/3511808.3557220.

Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, pp. 257–260, 2010. doi: 10.1145/1864708.1864761.

Carlos A. Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, 6(4):1–19, 2015. doi: 10.1145/2843948.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938. PMLR, 2018.

Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1): 5–53, 2004. doi: 10.1145/963770.963772.

Ko-Jen Hsiao, Yesu Feng, and Sudarshan Lamkhede. Foundation model for personalized recommendation. Netflix Technology Blog, 2025. Published March 21, 2025.

Xu Huang, Jianxun Lian, Hao Wang, Hao Liao, Defu Lian, and Xing Xie. A data-centric multi-objective learning framework for responsible recommendation systems. In *Proceedings of the ACM Web Conference 2024*, pp. 3129–3138, 2024. doi: 10.1145/3589334.3645324.

Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Tushar Chandra, and Craig Boutilier. Slateq: A tractable decomposition for reinforcement learning with recommendation sets. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 2592–2599, 2019. doi: 10.24963/ijcai.2019/360.

Dietmar Jannach and Markus Zanker. A survey on intent-aware recommender systems. *ACM Transactions on Recommender Systems*, 3(2):1–32, 2024. doi: 10.1145/3700890.

Olivier Jeunen, Jatin Mandav, Ivan Potapov, Nakul Agarwal, Sourabh Vaid, Wenzhe Shi, and Aleksei Ustimenko. Multi-objective recommendation via multivariate policy learning. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pp. 712–721, 2024. doi: 10.1145/3640457.3688132.

Thorsten Joachims, Adith Swaminathan, and Maarten de Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018.

Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10598–10632. PMLR, 2022.

Marius Kaminskas and Derek Bridge. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems*, 7(1), 2016. doi: 10.1145/2926720.

Woosung Kim, Jinho Lee, Jongmin Lee, and Byung-Jun Lee. Fairdice: Fairness-driven offline multi-objective reinforcement learning. *arXiv preprint arXiv:2506.08062*, 2025.

Ren Kishimoto, Rikiya Takehi, Koichi Tanaka, Masahiro Nomura, Riku Togashi, Yoji Tomita, and Yuta Saito. Beyond match maximization and fairness: Retention-optimized two-sided matching. In *International Conference on Learning Representations*, 2026.

Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. Temporal diversity in recommender systems. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 210–217, 2010. doi: 10.1145/1835449.1835486.

Sean M. McNee, John Riedl, and Joseph A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, pp. 1097–1101, 2006. doi: 10.1145/1125451.1125659.

Sejoon Oh, Moumita Bhattacharya, Yesu Feng, and Sudarshan Lamkhede. Intentrec: Predicting user session intent with hierarchical multi-task learning. *arXiv preprint arXiv:2408.05353*, 2024.

Giseung Park, Woohyeon Byeon, Seongmin Kim, Elad Havakuk, Amir Leshem, and Youngchul Sung. The max-min formulation of multi-objective reinforcement learning: From theory to a model-free algorithm. *arXiv preprint arXiv:2406.07826*, 2024.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten 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. PMLR, 2016.

Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust policy evaluation and learning in offline contextual bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8884–8894. PMLR, 2020.

Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 154–162, 2018. doi: 10.1145/3240323.3240372.

Adith Swaminathan and Thorsten 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. PMLR, 2015.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miroslav Dudík, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, pp. 109–116, 2011. doi: 10.1145/2043932.2043955.

Lequn Wang, Jiangwei Pan, Fengdi Che, and Linas Baltrunas. Genpage: Towards end-to-end generative homepage construction at netflix. *arXiv preprint arXiv:2606.31031*, 2026.

Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1927–1936, 2017.

Zhouhao Yang, Yihong Guo, Pan Xu, Anqi Liu, and Animashree Anandkumar. Distributionally robust policy gradient for offline contextual bandits. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6443–6462. PMLR, 2023.

Eva Zangerle and Christine Bauer. Evaluating recommender systems: Survey and framework. *ACM Computing Surveys*, 55(8):1–38, 2022. doi: 10.1145/3556536.

Mi Zhang and Neil Hurley. Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pp. 123–130, 2008. doi: 10.1145/1454008.1454030.

Yong Zheng and David Xuejun Wang. Multi-objective recommendations: A tutorial. *arXiv preprint arXiv:2108.06367*, 2021.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, pp. 22–32, 2005. doi: 10.1145/1060745.1060754.

A Extended Related Work

A.1 Beyond-accuracy recommendation and intent variation

Classical recommender-system evaluation has long emphasized that predictive accuracy alone is insufficient (Herlocker et al., 2004; McNee et al., 2006). This motivated diversification and beyond-accuracy objectives such as topical diversity, novelty, serendipity, and coverage (Carbonell & Goldstein, 1998; Ziegler et al., 2005; Zhang & Hurley, 2008; Lathia et al., 2010; Ge et al., 2010; Vargas & Castells, 2011; Kaminskas & Bridge, 2016). Calibration provides a related perspective by matching recommendation distributions to a target user-interest distribution (Steck, 2018). These methods mainly diversify the items appearing in a list or optimize a pre-specified trade-off. ODR instead treats the objective itself as uncertain: the relevant target may change across moments for the same user.

Intent-aware recommendation models explicitly attempt to infer the user's current goal from session or interaction context (Jannach & Zanker, 2024; Bhattacharya & Lamkhede, 2022; Oh et al., 2024). Such approaches and ODR are complementary. Intent prediction is useful when the current intent is identifiable, whereas ODR addresses residual uncertainty or misspecification by protecting all plausible intents.

A.2 Multi-objective recommendation and scalarization

Multi-objective recommendation commonly reduces a vector of metrics to a scalar reward through fixed, learned, or personalized weights (Zheng & Wang, 2021; Jeunen et al., 2024; Huang et al., 2024). Scalarization is attractive because it reduces the problem to standard single-objective learning and can encode known business preferences. However, a weighted expectation answers a different question from worst-intent robustness. Proposition 2 shows that even the exact intent-frequency weights can favor a policy that completely fails a low-probability intent.

The distinction is particularly relevant to personalized reward design. A meta-policy may choose user-specific coefficients (a, b, c) and subsequently optimize the resulting scalar reward. This captures heterogeneity between users, but the resulting coefficients can remain fixed within a user. ODR instead allows the active optimization target to vary during learning according to whichever intent-specific metric is currently weakest for a context.

A.3 Max–min multi-objective reinforcement learning

Recent MORL work studies nonlinear welfare criteria, including max–min fairness. Park et al. formulate MORL using a max–min objective and develop theory and a model-free algorithm (Park et al., 2024). Subsequent work gives a game-theoretic treatment with convergence, iteration-complexity, and sample-complexity analysis (Byeon et al., 2025). FairDICE studies nonlinear welfare optimization in offline multi-objective RL using distribution correction (Kim et al., 2025).

Our setting differs in three respects. First, the objective index corresponds to a latent user intent that may change for the same context/user population. Second, our metrics are heterogeneous ranking objectives such as relevance, coverage, novelty, serendipity, watch ratio, and watch flag, so logging-policy improvement normalization is central to making the minimum meaningful. Third, our proposed optimizer directly chooses the currently worst objective per context and applies its policy gradient, leading to the active-objective gradient characterization in Theorem 1.

A.4 Distributionally robust learning and worst-group objectives

ODR is conceptually related to distributionally robust optimization, which optimizes performance under an ambiguity set of distributions (Duchi & Namkoong, 2021). Group DRO similarly focuses learning pressure on the worst-performing predefined group (Sagawa et al., 2020), while repeated-loss formulations have used worst-case objectives to protect latent subpopulations (Hashimoto et al., 2018). The important semantic difference is the adversarial dimension: ODR places the simplex over intent/objective indices inside each recommendation context.

Offline contextual-bandit work has studied robust policy evaluation and learning under deployment distribution shift (Si et al., 2020; Yang et al., 2023); doubly robust methods improve statistical

efficiency for distributionally robust OPE/OPL (Kallus et al., 2022). Finite-sample robust regret has also been analyzed for offline RL (Zhou et al., 2021). These results are complementary to Theorem 2.

A.5 Slates, counterfactual learning, and long-term objectives

Learning ranking/slate policies from logged feedback requires accounting for selection bias and action combinatorics. Counterfactual risk minimization and inverse-propensity approaches provide a foundation for policy learning from logged bandit data (Swaminathan & Joachims, 2015; Schnabel et al., 2016), including slate-specific off-policy evaluation (Swaminathan et al., 2017) and top-K policy corrections (Chen et al., 2019). Deep models can also be trained directly from logged bandit feedback (Joachims et al., 2018). The ODR objective is estimator-agnostic: ODR requires an estimate of each intent-specific policy value and a policy gradient for the selected worst objective.

Finally, the motivating downstream concern is long-term dissatisfaction and retention rather than only immediate interaction maximization. Industrial recommender systems already optimize a mixture of engagement and longer-term goals (Gomez-Uribe & Hunt, 2015; Hsiao et al., 2025), and retention-oriented matching research highlights that maximizing short-term matching outcomes can differ from reducing user abandonment (Kishimoto et al., 2026). Our experiments do not directly estimate retention effects; they test the more proximal hypothesis that avoiding severe intent-specific recommendation failures can be achieved through objective-space robustness.

B Theoretical Analysis

This appendix develops the theoretical interpretation of the proposed objective and the policy-gradient update used in our implementation. The analysis is intentionally modular: the first results concern the objective itself and do not depend on a particular off-policy estimator, while the final finite-sample result is stated in terms of a generic uniform estimation error. This separation lets the theory remain valid when the estimator used in the experiments is replaced by another consistent OPE/OPL estimator.

B.1 Setup and logging-policy normalization

Let $x \in \mathcal{X}$ denote a recommendation context, $y \in \mathcal{Y}(x)$ a ranking, and $\pi_\theta(y \mid x)$ a stochastic ranking policy. Let $\mathcal{Z} = \{1, \dots, M\}$ denote the finite set of plausible user intents. Each intent z is associated with a raw higher-is-better ranking metric $g_z^{\mathrm{raw}}(x, y)$. The raw policy value is

$$G_z^{\mathrm{raw}}(x; \pi_\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[g_z^{\mathrm{raw}}(x, y) \right].$$

Because the objectives have different natural scales, all objectives are converted to improvement rates relative to the logging policy π_0. Specifically, we define

$$G_z(x; \pi_\theta) = \frac{G_z^{\mathrm{raw}}(x; \pi_\theta) - G_z^{\mathrm{raw}}(x; \pi_0)}{|G_z^{\mathrm{raw}}(x; \pi_0)| + \epsilon_0},$$

where $\epsilon_0 > 0$ prevents numerical instability when the logging-policy value is close to zero. Equivalently, define the normalized ranking-level reward

$$\widetilde{g}_z(x, y) = \frac{g_z^{\mathrm{raw}}(x, y) - G_z^{\mathrm{raw}}(x; \pi_0)}{|G_z^{\mathrm{raw}}(x; \pi_0)| + \epsilon_0},$$

so that $G_z(x; \pi_\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[\widetilde{g}_z(x, y)]$.

Our robust objective is

$$F(\pi_\theta) = \mathbb{E}_{x \sim p(x)} \left[\min_{z \in \mathcal{Z}} G_z(x; \pi_\theta) \right].$$

Thus, the proposed method maximizes the expected improvement rate of the worst-served plausible intent for each context.

B.2 MINIMAX INTERPRETATION UNDER AN UNKNOWN INTENT DISTRIBUTION

The first result explains why the minimum over intents is the appropriate robust objective when no reliable information about the intent mixture is available.

Proposition 1 (distribution-free minimax equivalence). For any fixed context x and policy π, let Δ_M denote the probability simplex over $\mathcal{Z}$. Then

$$\min_{q \in \Delta_M} \sum_{z \in \mathcal{Z}} q_z G_z(x;\pi) = \min_{z \in \mathcal{Z}} G_z(x;\pi).$$

Consequently,

$$\max_{\pi} \mathbb{E}_x \left[\min_{q \in \Delta_M} \mathbb{E}_{z \sim q} \left[G_z(x;\pi) \right] \right] = \max_{\pi} \mathbb{E}_x \left[\min_{z} G_z(x;\pi) \right].$$

Proof. For any $q \in \Delta_M$, a convex combination cannot be smaller than its smallest component, hence $\sum_z q_z G_z(x;\pi) \geq \min_z G_z(x;\pi)$. Equality is achieved by placing all mass on any minimizer $z^\star \in \arg\min_z G_z(x;\pi)$. Taking the outer expectation and maximization gives the second statement. □

For any true but unknown intent distribution $p^\star(z \mid x)$,

$$\mathbb{E}_{z \sim p^\star(\cdot|x)}[G_z(x;\pi)] \geq \min_z G_z(x;\pi),$$

and therefore

$$\mathbb{E}_{x,z \sim p^\star}[G_z(x;\pi)] \geq F(\pi).$$

Hence $F(\pi)$ is a distribution-free lower bound on expected intent-conditioned performance that is valid for every conditional intent distribution supported on $\mathcal{Z}$.

B.3 WHY EVEN AN ORACLE WEIGHTED SUM CAN FAIL

A natural alternative is to estimate the intent distribution and optimize its expectation. The next construction shows that the limitation is not only estimation error: even the exact intent distribution can prefer a policy with arbitrarily poor performance on a rare intent.

Proposition 2 (oracle scalarization can sacrifice a rare intent). Consider two intents with true probabilities $p(z = 1) = 1 - \rho$ and $p(z = 2) = \rho$, where $\rho \in (0, 1)$. Suppose the policy class contains policies A and B with normalized objective vectors

$$(G_1(A), G_2(A)) = (1, 0), \qquad (G_1(B), G_2(B)) = (b, b),$$

for some $b \in (0, 1 - \rho)$. Then the oracle expected-utility objective selects A, while the max–min objective selects B:

$$(1-\rho)G_1(A) + \rho G_2(A) = 1 - \rho > b = (1-\rho)G_1(B) + \rho G_2(B),$$

but

$$\min_z G_z(A) = 0 < b = \min_z G_z(B).$$

Proof. The inequalities follow directly from $b < 1 - \rho$ and $b > 0$. □

The separation can be large. As $\rho \to 0$, choose b just below $1 - \rho$. The oracle scalarization then chooses a policy with worst-intent value 0 even though a balanced policy exists whose worst-intent value is arbitrarily close to 1. This formalizes the distinction between knowing the average intent mixture and protecting every plausible intent.

B.4 GRADIENT CHARACTERIZATION OF THE WORST-OBJECTIVE UPDATE

Let

$$z^\star(x, \theta) \in \arg\min_{z \in \mathcal{Z}} G_z(x; \pi_\theta).$$

Assumption 1 (regularity). For every z, $G_z(x;\pi_\theta)$ is continuously differentiable in θ for almost every x, and its gradient is dominated by an integrable function in a neighborhood of the parameter values considered. In addition, the minimizing objective $z^\star(x,\theta)$ is unique for almost every x.

Theorem 1 (gradient of the robust objective). Under Assumption 1,

$$\nabla_\theta F(\pi_\theta) = \mathbb{E}_x\left[\nabla_\theta G_{z^\star(x,\theta)}(x;\pi_\theta)\right].$$

Proof. For a fixed x with a unique minimizer, the pointwise function $\theta \mapsto \min_z G_z(x;\pi_\theta)$ agrees locally, to first order, with the active branch $G_{z^\star(x,\theta)}(x;\pi_\theta)$. Hence its gradient equals the gradient of the active objective. The domination condition allows differentiation and expectation to be interchanged. □

The theorem gives a direct justification for the implemented algorithm: at every training loop and for every sampled context, identify the currently worst-performing metric and use that metric as the policy-gradient reward. With exact objective values, this is stochastic gradient ascent on F.

Corollary 1 (policy-gradient form). Assume that the support of $\pi_\theta(\cdot \mid x)$ does not depend on θ and that differentiation under the ranking expectation is valid. Then

$$\nabla_\theta G_{z^\star}(x;\pi_\theta) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\widetilde{g}_{z^\star}(x,y)\nabla_\theta \log \pi_\theta(y \mid x)\right].$$

For any baseline $b(x)$ independent of the sampled ranking action,

$$\nabla_\theta G_{z^\star}(x;\pi_\theta) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[(\widetilde{g}_{z^\star}(x,y) - b(x))\,\nabla_\theta \log \pi_\theta(y \mid x)\right].$$

Thus, for a minibatch $\{x_i\}_{i=1}^B$, the update has the form

$$\widehat{h}_t = \frac{1}{B}\sum_{i=1}^{B}(\widetilde{g}_{\widehat{z}_i}(x_i,y_i) - b_i)\,\nabla_\theta \log \pi_\theta(y_i \mid x_i), \qquad \theta_{t+1} = \theta_t + \eta_t \widehat{h}_t,$$

where $\widehat{z}_i$ is the estimated worst metric for context x_i. When learning from logged data, the ranking-level gradient term is combined with the appropriate off-policy correction.

Remark on ties. If several objectives attain the same minimum, F is generally nonsmooth. We use an almost-sure uniqueness assumption in the theorem because it matches the generic case. We break ties deterministically in the experiments; an alternative is to average gradients over the active set.

B.5 Robustness to errors in identifying the worst metric

Let $\widehat{G}_z(x;\pi)$ be the estimate used to identify the worst objective and define

$$z^\star = \arg\min_z G_z(x;\pi), \qquad \widehat{z} = \arg\min_z \widehat{G}_z(x;\pi).$$

Proposition 3 (near-worst selection under uniform metric error). If

$$\max_{z\in\mathcal{Z}}\left|\widehat{G}_z(x;\pi) - G_z(x;\pi)\right| \le \varepsilon,$$

then

$$G_{\widehat{z}}(x;\pi) \le \min_z G_z(x;\pi) + 2\varepsilon.$$

Proof. Let $z^\star$ be a true minimizer. Then

$$G_{\widehat{z}} \le \widehat{G}_{\widehat{z}} + \varepsilon \le \widehat{G}_{z^\star} + \varepsilon \le G_{z^\star} + 2\varepsilon.$$

□

Corollary 2 (exact worst-metric recovery under a margin). Let $G_{(1)}(x;\pi) \le G_{(2)}(x;\pi) \le \cdots$ denote the sorted objective values and define

$$\Delta(x,\pi) = G_{(2)}(x;\pi) - G_{(1)}(x;\pi).$$

If $\Delta(x,\pi) > 2\varepsilon$, then $\widehat{z} = z^\star$.

B.6 FINITE-SAMPLE GUARANTEE FOR ROBUST POLICY SELECTION

Let

$$F(\pi) = \mathbb{E}_x[\min_z G_z(x;\pi)], \qquad \widehat{F}(\pi) = \mathbb{E}_x[\min_z \widehat{G}_z(x;\pi)].$$

Let $\pi^\star \in \arg\max_{\pi\in\Pi} F(\pi)$ and $\widehat{\pi} \in \arg\max_{\pi\in\Pi} \widehat{F}(\pi)$.

Theorem 2 (robust oracle inequality). Suppose that, with probability at least $1-\delta$,

$$\sup_{\pi\in\Pi} \max_{z\in\mathcal{Z}} \left|\widehat{G}_z(\pi) - G_z(\pi)\right| \le \varepsilon_n.$$

Then, on the same event,

$$F(\widehat{\pi}) \ge F(\pi^\star) - 2\varepsilon_n.$$

Proof. The minimum operator is 1-Lipschitz with respect to the ℓ_∞ norm, hence $|\widehat{F}(\pi) - F(\pi)| \le \varepsilon_n$ uniformly over Π. Therefore

$$F(\pi^\star) \le \widehat{F}(\pi^\star) + \varepsilon_n \le \widehat{F}(\widehat{\pi}) + \varepsilon_n \le F(\widehat{\pi}) + 2\varepsilon_n.$$

Rearranging proves the result. □

Corollary 3 (finite policy class, bounded full-feedback case). For intuition, consider a finite policy class Π and suppose that for every policy and objective the per-context normalized value is i.i.d. and lies in an interval of width B. If $\widehat{G}_z(\pi)$ is the sample mean over n independent full-feedback contexts, then with probability at least $1-\delta$,

$$F(\widehat{\pi}) \ge F(\pi^\star) - B\sqrt{\frac{2\log(2|\Pi||\mathcal{Z}|/\delta)}{n}}.$$

Proof sketch. Hoeffding's inequality followed by a union bound over $|\Pi||\mathcal{Z}|$ policy-objective pairs gives

$$\varepsilon_n \le B\sqrt{\frac{\log(2|\Pi||\mathcal{Z}|/\delta)}{2n}}.$$

Substituting into Theorem 2 yields the stated result. □

The finite-class corollary is deliberately simple and is not claimed as the exact bound for our neural policy class. For neural policies learned from logged bandit feedback, the relevant ε_n must incorporate function-class complexity, overlap/propensity terms, and the specific OPE/OPL estimator. Existing offline robust policy-learning work provides estimator-specific finite-sample analyses under such assumptions (Si et al., 2020; Yang et al., 2023; Kallus et al., 2022; Zhou et al., 2021). Thus, any uniform bound on the intent-specific values directly yields the corresponding guarantee for the objective-diverse max–min criterion through Theorem 2.

B.7 SCOPE AND LIMITATIONS OF THE THEORETICAL CLAIMS

The theory above establishes four points: (i) the max–min objective is the exact full-simplex minimax objective under unknown intent frequencies; (ii) oracle linear scalarization can still sacrifice rare intents; (iii) the proposed worst-metric policy-gradient rule is the gradient of the robust objective when the active worst metric is uniquely identified; and (iv) uniform objective-estimation error translates into a robust policy-value guarantee.

The analysis does not claim global convergence of deep nonconvex policy optimization, nor does it claim a causal guarantee on retention. Recent max–min multi-objective reinforcement-learning work develops stronger convergence or sample-complexity results in structured MORL settings (Park et al., 2024; Byeon et al., 2025). Our contribution is complementary: per-context latent intent uncertainty in recommendation, logging-policy-normalized heterogeneous ranking objectives, and a simple active-worst-objective policy-gradient update.

Table 1: Synthetic action catalog used in the experiments.

Action	Type	Category	Base relevance	Novelty	Serendipity
0	relevance	0	1.00	0.20	0.18
1	relevance	0	0.95	0.16	0.14
2	relevance	0	0.90	0.12	0.11
3	relevance	0	0.85	0.85	0.07
4	relevance	0	0.80	0.80	0.04
5	discovery	1	0.35	0.95	0.90
6	discovery	1	0.21	0.85	0.75
7	discovery	2	0.60	0.08	0.80
8	discovery	3	0.43	0.05	0.65
9	discovery	2	0.37	0.10	0.70

C Experimental Implementation Details

This section describes the experimental setup and optimization procedures. We distinguish two experimental optimizers because the synthetic and KuaiRec experiments optimize the robust objective differently. The synthetic experiments train a neural contextual ranking policy by backpropagating through differentiable soft-ranking metrics. The main KuaiRec experiments use the fully observed KuaiRec small matrix to construct and score a finite set of candidate rankings and then select, for each user, the candidate maximizing each method's objective. Thus, the KuaiRec experiment is a full-feedback candidate-policy-selection benchmark rather than a neural policy-gradient training run. This distinction is important for reproducibility and for interpreting the theoretical analysis in Appendix B.

C.1 Shared objective and reporting convention

Let $M_z(\pi; x)$ denote the hard evaluation value of metric z for policy/ranking π in context x, and let $M_z(\pi_0; x)$ denote the corresponding value under the logging policy. We report per-context improvement relative to the logging policy. In the synthetic experiments the denominator is

$$\mathrm{Imp}_z(\pi; x) = \frac{M_z(\pi; x) - M_z(\pi_0; x)}{\max(|M_z(\pi_0; x)|, 10^{-8})},$$

while the KuaiRec experiments use the more conservative floor

$$\mathrm{Imp}_z(\pi; x) = \frac{M_z(\pi; x) - M_z(\pi_0; x)}{\max(|M_z(\pi_0; x)|, 0.05)}.$$

The primary robust statistic is the average per-context worst improvement,

$$\mathrm{WorstImp}(\pi) = \mathbb{E}_x \left[\min_{z \in \mathcal{Z}_{\mathrm{active}}} \mathrm{Imp}_z(\pi; x) \right].$$

The reported mean-performance statistic is the average improvement across the active metrics. Unless otherwise noted, curves aggregate five independent repetitions and the shaded region is ± 1 empirical standard deviation across repetitions.

C.2 Synthetic experiment

C.2.1 Environment and item catalog

The synthetic setting uses 10 actions, ranking length $K = 5$, four categories, six-dimensional contexts, 1,000 logged contexts by default, and 3,000 independently generated test contexts. Context vectors are sampled i.i.d. from $\mathcal{N}(0, I_6)$.

The catalog is shown in Table 1. In addition to the four evaluation attributes, the policy/model features include a manually specified popularity feature $(0.98, 0.95, 0.92, 0.88, 0.84, 0.30, 0.25, 0.20, 0.16, 0.12)$ and a one-hot item identity vector.

The context-dependent relevance of item a is

$$r_a(x) = \text{clip}\left(r_a^{\text{base}} + x^\top w_a,\, 0.02,\, 1\right), \qquad w_a \sim \mathcal{N}(0, 0.015^2 I_6).$$

The context coefficient vectors are sampled once when an environment instance is initialized. Topic gain is defined using a uniform topic-preference vector:

$$g_{a,t}^{\text{topic}}(x) = r_a(x)\, \mathbf{1}\{c(a) = t\}/4.$$

Novelty and serendipity are context-independent catalog attributes in this experiment.

C.2.2 Synthetic logging policy and logged feedback

For every logged/test context, the logging policy first samples an instantaneous intent according to

$$p(\text{Rel}) = 0.9 - \rho/2, \qquad p(\text{Div}) = 0.1, \qquad p(\text{Novel}) = p(\text{Seren}) = \rho/4.$$

The default intent-specific ranking templates inherited by the catalog environment are

$$\begin{aligned} \text{Rel} &: (0, 1, 2, 5, 6), \\ \text{Div} &: (0, 6, 7, 8, 5), \\ \text{Novel} &: (5, 6, 7, 8, 9), \\ \text{Seren} &: (5, 6, 7, 8, 9). \end{aligned}$$

With probability 0.15, the template is replaced by an exploratory top-K ranking. The exploratory item score depends on the sampled intent and combines relevance, novelty, serendipity, and/or a coverage proxy; independent Gaussian noise with standard deviation 0.05 is then added before taking the top five items.

For each exposed item, we record the vector of component labels (relevance, four topic-gain components, novelty, and serendipity) and adds independent Gaussian label noise with standard deviation 0.02, followed by clipping to $[0, 1]$. Since a logged context exposes $K = 5$ items, $n_{\log}$ contexts produce $5n_{\log}$ supervised item-level component observations.

C.2.3 Direct component model

The synthetic offline learner uses a direct regression model rather than IPS/SNIPS/DR. Pair features consist of the six-dimensional context, 18 item features, and 24 context–topic interaction features, for a total of 48 input dimensions. The 18 item features are four category indicators, popularity, novelty, base relevance, base serendipity, and a 10-dimensional item one-hot vector.

A multi-output ridge regression model is fit using `StandardScaler()` followed by `Ridge(alpha=1.0)`. The fitted model predicts the full component vector for every context–item pair. These predictions are clipped to $[0, 1]$ and are used to construct the differentiable training metrics. Thus, in the synthetic experiment the OPE/OPL component is a regression-based direct model; no propensity-score correction is used.

C.2.4 Neural ranking policy

All five synthetic methods use the same contextual scorer. For each context–item pair, the network concatenates the six context features with the 18 item features and applies $24 \to 64 \to 64 \to 1$, with ReLU nonlinearities after the two hidden layers. The resulting scalar item scores define the final hard ranking by sorting all 10 actions and taking the top $K = 5$.

Training uses a differentiable sequential soft-ranking relaxation. At each of the five positions, a softmax with temperature 0.20 is applied to item logits while previously selected mass is multiplicatively downweighted. Soft DCG, novelty, and serendipity accumulate position discounts $1/\log_2(k+1)$; soft category coverage tracks, for each category, the probability of not having appeared in previous positions. All metric values are normalized by their corresponding ideal denominators.

The optimizer is AdamW with learning rate 3×10^{-3} and weight decay 10^{-4}. The batch size is 128 (or the entire logged set when $n_{\log} < 128$), and we perform 80 minibatch optimization updates. Each update samples a fresh minibatch uniformly from the logged contexts, with replacement only when the dataset is smaller than the batch. Gradients are clipped to Euclidean norm 5.0.

C.2.5 SYNTHETIC TRAINING OBJECTIVES

For a minibatch context x_i, let $\widehat{m}_{i,z}$ be the differentiable metric value computed from predicted components, and let $\widehat{m}_{i,z}^{\log}$ be the hard metric of the logged ranking evaluated on the same predicted components. The training improvement is

$$\widehat{I}_{i,z} = \frac{\widehat{m}_{i,z} - \widehat{m}_{i,z}^{\log}}{\max(\widehat{m}_{i,z}^{\log}, 10^{-8})}.$$

The methods optimize the following minibatch objectives:

- **Primary only:** mean soft DCG, $\frac{1}{B}\sum_i \widehat{m}_{i,\mathrm{DCG}}$.
- **Equal-weight linear sum:** arithmetic mean of the active soft metric values.
- **Intent-weighted linear sum:** $\frac{1}{B}\sum_i \sum_z p_\rho(z)\widehat{m}_{i,z}$, with the true intent weights supplied to this oracle baseline.
- **Constrained optimization:** mean soft DCG with batch-level guardrails for the three non-DCG metrics. Let $v_j = [\frac{1}{B}\sum_i(\widehat{m}_{i,j}^{\log} - \widehat{m}_{i,j})]_+$. The optimized objective is

  $$J_{\mathrm{constr}} = \overline{m}_{\mathrm{DCG}} - \sum_{j=1}^{3} \lambda_j v_j - 100 \sum_{j=1}^{3} v_j^2.$$

  The dual variables start at zero, are updated after each policy step as $\lambda_j \leftarrow \mathrm{clip}(\lambda_j + 5v_j, 0, 100)$, and are shared across minibatches.
- **ODR:** the mean per-context worst improvement,

  $$J_{\mathrm{ODR}} = \frac{1}{B}\sum_{i=1}^{B} \min_{z \in \mathcal{Z}_{\mathrm{active}}} \widehat{I}_{i,z}.$$

In the synthetic experiments, we optimize ODR by direct backpropagation through the differentiable minimum and soft-ranking metrics. The autograd gradient follows the currently active worst metric for each context (up to ties), which is the differentiable-surrogate counterpart of the active-worst-objective gradient characterized in Appendix B.4.

C.2.6 SYNTHETIC EVALUATION PROTOCOL

Evaluation uses 3,000 independently sampled contexts per repetition. The learned policy is converted to a deterministic hard ranking by sorting its item logits. Each evaluation metric is computed from the *true* environment components rather than the ridge predictions. The reference logging ranking is independently sampled for the same test context. DCG, novelty, and serendipity are normalized discounted cumulative gains using the ideal top-K denominator; category coverage is the number of distinct covered categories divided by $\min(K, 4)$.

For each context, improvements are computed against its logging-policy ranking and the minimum improvement across active metrics is recorded. Exact ties among worst metrics are split evenly only for the diagnostic "worst metric fraction" plots.

The main sweep settings are:

- intent-mixture sweep: $\rho \in \{0, 0.1, 0.2, 0.35, 0.5, 0.65, 0.8, 0.9, 1.0\}$ with $n_{\log} = 1{,}000$;
- logged-data-size sweep: $n_{\log} \in \{100, 200, 500, 1000, 2000, 5000\}$ at $\rho = 0.35$;
- five repetitions for every condition;
- 3,000 test contexts per repetition.

Table 2: Modal top-5 rankings and objective-wise improvements in representative relevance-heavy ($\rho = 0$) and discovery-heavy ($\rho = 1$) settings. R and D denote relevance- and discovery-oriented items, respectively. The shaded cell indicates the worst-performing objective for each method.

Method	Ranking	R/D composition	Share	DCG impr.	Coverage impr.	Novelty impr.	Serendipity impr.
(a) Relevance-heavy setting ($\rho = 0$)							
Primary-only	0–1–2–3–4	R–R–R–R–R	0.787	0.201	-0.455	-0.067	-0.552
Equal scalarization	5–6–3–9–7	D–D–R–D–D	0.138	-0.407	0.730	0.881	1.298
Oracle intent-weighted	0–1–2–3–4	R–R–R–R–R	0.802	0.201	-0.455	-0.067	-0.552
Constrained optimization	0–5–3–1–7	R–D–R–R–D	0.075	0.049	0.430	-0.027	0.233
ODR	0–5–7–3–1	R–D–D–R–R	0.491	-0.019	0.634	0.107	0.527
(b) Discovery-heavy setting ($\rho = 1$)							
Primary-only	0–1–2–3–4	R–R–R–R–R	0.833	0.805	-0.570	-0.218	-0.703
Equal scalarization	5–6–3–4–7	D–D–R–R–D	0.151	-0.094	0.404	0.622	0.472
Oracle intent-weighted	5–3–6–4–0	D–R–D–R–R	0.595	0.142	-0.134	0.755	0.123
Constrained optimization	5–7–6–3–4	D–D–D–R–R	0.089	0.136	0.291	0.295	0.406
ODR	5–7–0–1–3	D–D–R–R–R	0.617	0.311	0.291	0.119	0.265

C.2.7 Synthetic intent-set-size and learning-curve experiments

The intent-set-size experiment uses $\rho = 0$, $n_{\log} = 1{,}000$, five repetitions, and the following active/evaluated objective sets:

$$
\begin{aligned}
&|\mathcal{Z}| = 1 : \{\mathrm{DCG}\},\\
&|\mathcal{Z}| = 2 : \{\mathrm{DCG}, \mathrm{Novelty}\},\\
&|\mathcal{Z}| = 3 : \{\mathrm{DCG}, \mathrm{Novelty}, \mathrm{Serendipity}\},\\
&|\mathcal{Z}| = 4 : \{\mathrm{DCG}, \mathrm{TopicCoverage}, \mathrm{Novelty}, \mathrm{Serendipity}\}.
\end{aligned}
$$

The oracle scalarization weights are obtained by restricting the $\rho = 0$ weight vector $(0.9, 0.1, 0, 0)$ to the active set and renormalizing. Consequently, the size-2 and size-3 oracle scalarizations place all their weight on DCG; the size-4 case uses weights $(0.9, 0.1, 0, 0)$.

The synthetic learning-curve diagnostic trains ODR for 80 updates at $\rho \in \{0, 1\}$, using $n_{\log} = 1{,}000$, $3{,}000$ test contexts, and five repetitions. Hard test metrics are recorded at initialization and after every optimization update.

subsectionQualitative Analysis of Modal Rankings

To better understand how the objective-level trade-offs in Figure 2 translate into actual ranking behavior, we inspect the modal top-5 ranking produced by each method under two representative intent mixtures, $\rho = 0$ and $\rho = 1$. Recall that smaller ρ corresponds to relevance-heavy demand, whereas larger ρ shifts more intent mass toward novelty and serendipity. In Table 2, R and D denote relevance- and discovery-oriented items, respectively. We additionally report the objective-wise improvements associated with each modal ranking to illustrate which objectives are sacrificed by different ranking strategies.

Under the relevance-heavy setting ($\rho = 0$), both the primary-only and oracle intent-weighted policies produce an all-relevance ranking (R–R–R–R–R). While this preserves DCG, it drives category coverage, novelty, and serendipity below the reference policy; for the oracle intent-weighted policy, the corresponding improvements are -0.455, -0.067, and -0.552. Equal scalarization exhibits the opposite behavior, selecting four discovery items and substantially improving the discovery-oriented metrics, but at the cost of a large negative DCG improvement (-0.407). ODR instead produces a mixed ranking (R–D–D–R–R), retaining relevance-oriented positions while introducing discovery items. As a result, its DCG degradation is small (-0.019), while coverage, novelty, and serendipity remain improved relative to the reference policy.

The ranking compositions also change as the intent mixture becomes more discovery-oriented ($\rho = 1$). The primary-only policy remains fully specialized toward relevance and therefore continues to suffer large reductions in discovery metrics, including a serendipity improvement of -0.703. Equal scalarization shifts toward discovery items but still incurs negative DCG improvement (-0.094), while the oracle intent-weighted policy mixes relevance and discovery items but leaves category coverage below the reference policy (-0.134). ODR also increases the presence of discovery-

oriented items, producing D–D–R–R–R, while keeping all four objective improvements positive in this representative ranking.

Overall, these examples illustrate the ranking-level mechanism behind objective diversity. Policies based on a fixed or mixture-weighted scalarization can specialize toward one side of the relevance–discovery trade-off, causing another objective to deteriorate sharply. ODR instead tends to combine relevance- and discovery-oriented items within the same slate and changes this composition as the intent environment shifts. This qualitative behavior is consistent with the more stable worst-objective performance observed across the full ρ sweep in Figure 2.

C.3 KuaiRec full-feedback experiment

C.3.1 Data preprocessing and candidate set

We use the small-matrix interaction data and item-category metadata provided by KuaiRec. Candidate videos are sampled once, uniformly without replacement, from all video IDs appearing in the small matrix. The number of candidate items is 500 and the fixed candidate-sampling seed is 20260731. All users remaining after the item filter are retained; after this filtering, 1,411 users remain.

For each retained user–item pair, the raw watch ratio is transformed to

$$w_{ui} = \mathrm{clip}(\mathrm{watch_ratio}_{ui}, 0, 5)/5,$$

and the watch flag is

$$f_{ui} = \mathbf{1}\{\mathrm{watch_ratio}_{ui} > 0\}.$$

The category file contains a string representation of a list in column `feat`; we parse this list and use its first element as the item's category. Missing/unparseable categories are assigned -1, after which all observed category IDs (including -1 if present) are remapped to contiguous indices.

Two auxiliary rarity features are computed. Item rarity is one minus min–max normalized mean watch score:

$$r_i^{\mathrm{item}} = 1 - \frac{\overline{w}_i - \min_j \overline{w}_j}{\max_j \overline{w}_j - \min_j \overline{w}_j},$$

and category rarity is min–max normalized $-\log$ category prevalence. These two features are used to construct logging/candidate scores; the evaluated novelty metric is defined separately from logged exposure frequencies.

C.3.2 Simulated logging policy

For a given ρ, the synthetic logging mixture on top of the real KuaiRec feedback matrix is

$$p(\mathrm{ratio}) = 0.9 - \rho/2, \qquad p(\mathrm{flag}) = 0.1, \qquad p(\mathrm{div}) = p(\mathrm{nov}) = \rho/4.$$

The deterministic component of the logging score for user u and item i is

$$s_{ui} = p(\mathrm{ratio}) w_{ui} + p(\mathrm{flag}) f_{ui} + p(\mathrm{div}) r_i^{\mathrm{cat}} + p(\mathrm{nov}) r_i^{\mathrm{item}}.$$

Independent Gumbel noise with location 0 and scale 0.12 is added to every user–item score, and the top K items form the logged ranking. When generating $n_{\log}$ logged contexts, users are sampled uniformly with replacement from the retained user set.

C.3.3 Novelty and evaluation metrics

Novelty is estimated from the *training logged rankings*. Let c_i be the number of exposures of candidate item i across all logged top-K positions. With Laplace smoothing constant 1,

$$\widehat{p}_i = \frac{c_i + 1}{\sum_j c_j + 500}, \qquad \nu_i = -\log \widehat{p}_i.$$

The vector ν is then min–max normalized to $[0, 1]$. Thus, we define novelty using marginal logged-exposure rarity rather than the context-conditional quantity $-\log \pi_0(a \mid x)$.

For $K = 5$, position discounts are $d_k = 1/\log_2(k+1)$. WatchRatio and WatchFlag are nDCG-style normalized discounted gains:

$$M_{\mathrm{WR}}(u,y) = \frac{\sum_{k=1}^{K} d_k w_{u,y_k}}{\max_{|S|=K} \sum_{k=1}^{K} d_k w_{u,S_k}}, \qquad M_{\mathrm{WF}}(u,y) = \frac{\sum_{k=1}^{K} d_k f_{u,y_k}}{\max_{|S|=K} \sum_{k=1}^{K} d_k f_{u,S_k}}.$$

Novelty uses the same normalized discounted-gain form with ν_i and a global ideal top-K novelty denominator. Category coverage is

$$M_{\mathrm{Cov}}(u,y) = \frac{|\{c(y_1), \ldots, c(y_K)\}|}{\min(K, N_{\mathrm{cat}})}.$$

All KuaiRec improvements use denominator floor 0.05:

$$\mathrm{Imp}_z(u) = \frac{M_z(u,y) - M_z(u,y^{\log})}{\max(|M_z(u,y^{\log})|, 0.05)}.$$

C.3.4 Candidate-ranking search

The main KuaiRec experiment does not fit a neural ranking model. Instead, it constructs a common finite candidate set for every method and selects the candidate maximizing that method's objective for each user.

Candidate rankings are generated from four item-level score components: WatchRatio, WatchFlag, category rarity, and logged-exposure novelty. We evaluate the following 14 fixed scalarization vectors:

$$(1,0,0,0),\ (0,1,0,0),\ (0,0,1,0),\ (0,0,0,1),$$
$$(0.25,0.25,0.25,0.25),\ (0.5,0.2,0.15,0.15),\ (0.35,0.15,0.25,0.25),$$
$$(0.2,0.2,0.3,0.3),\ (0.15,0.15,0.35,0.35),\ (0.6,0.1,0.2,0.1),$$
$$(0.6,0.1,0.1,0.2),\ (0.4,0.3,0.1,0.2),\ (0.4,0.1,0.3,0.2),\ (0.4,0.1,0.2,0.3).$$

The true intent vector p_ρ is appended as a 15th scalarization vector. For each vector, we construct three rankings: ordinary top-K by scalar score, and two greedy diversity-enhanced rankings with unseen-category bonuses 0.18 and 0.35. This produces $15 \times 3 = 45$ candidate rankings per user.

The diversity-enhanced generator proceeds sequentially. At each position it adds a fixed bonus to every item whose category has not yet appeared in the partial ranking, excludes already selected items, and chooses the highest-scoring remaining item.

All 45 candidates are evaluated on the fully observed WatchRatio/WatchFlag matrix, category labels, and log-derived novelty score before method-specific selection. Thus, the KuaiRec experiment evaluates the objective-selection principle within a fixed candidate-policy class; it does not estimate an unconstrained ranking policy from bandit feedback.

C.3.5 KuaiRec method objectives

Let $m_{c,u,z}$ be the hard metric value of candidate c for user u, and $m^{\log}_{u,z}$ the metric of the independently sampled logging ranking.

- **WatchRatio only:** choose $\arg\max_c m_{c,u,\mathrm{WR}}$.
- **Equal-weight linear sum:** choose the candidate maximizing the arithmetic mean of the active raw metric values.
- **Intent-weighted linear sum:** restrict p_ρ to the active objective set, renormalize it, and maximize the corresponding weighted mean of raw metric values.
- **Constrained optimization:** for active non-WatchRatio metrics $\mathcal{N}$, maximize

$$m_{c,u,\mathrm{WR}} - 5 \sum_{z \in \mathcal{N}} \left[m^{\log}_{u,z} - m_{c,u,z} \right]_+^2.$$

- **ODR:** maximize the worst logging-policy improvement among active metrics,

$$\arg\max_c \min_{z \in \mathcal{Z}_{\mathrm{active}}} \frac{m_{c,u,z} - m^{\log}_{u,z}}{\max(|m^{\log}_{u,z}|, 0.05)}.$$

The same 45 candidate rankings are used for all methods, so differences arise only from the objective used to choose among them.

C.3.6 KuaiRec experiment grids and repetitions

The main KuaiRec experiments use ranking length $K = 5$ and five repetitions:

- **Intent-mixture sweep:** $\rho \in \{0, 0.1, 0.2, 0.35, 0.5, 0.65, 0.8, 0.9, 1.0\}$ with $n_{\log} = 1{,}000$.
- **Logged-data-size sweep:** $n_{\log} \in \{100, 200, 500, 1000, 2000, 5000\}$ at $\rho = 0.35$.
- **Intent-set-size sweep:** $\rho = 0.35$, $n_{\log} = 1{,}000$, with

$$
\begin{aligned}
&|\mathcal{Z}| = 1 : \{\mathrm{WatchRatio}\},\\
&|\mathcal{Z}| = 2 : \{\mathrm{WatchRatio}, \mathrm{Novelty}\},\\
&|\mathcal{Z}| = 3 : \{\mathrm{WatchRatio}, \mathrm{CategoryCoverage}, \mathrm{Novelty}\},\\
&|\mathcal{Z}| = 4 : \{\mathrm{WatchRatio}, \mathrm{WatchFlag}, \mathrm{CategoryCoverage}, \mathrm{Novelty}\}.
\end{aligned}
$$

There is no conventional train/validation/test split in this full-feedback benchmark. The 500 candidate items are fixed once. Within each repetition, $n_{\log}$ training contexts are generated by sampling users with replacement and are used to estimate novelty. Evaluation then uses all retained users exactly once, with an independent stochastic logging ranking as the per-user reference.

C.3.7 Ranking-length diagnostic

As a sensitivity analysis, we repeat the experiment for $K \in \{1, 2, 3\}$ using the same 500-item candidate set, ρ grid, $n_{\log} = 1{,}000$, and five repetitions. For each K, all ideal denominators and the category-coverage normalization are recomputed. Together with the main $K = 5$ experiment, these runs provide ranking-length sensitivity checks.

C.3.8 KuaiRec learning-curve diagnostic

The learning-curve diagnostic uses a separate optimization setup from the main candidate-ranking experiment. It uses a tabular soft-policy proxy with one trainable logit per user–item pair. Logits are initialized from the true intent-weighted combination of WatchRatio, WatchFlag, category rarity, and item rarity, plus Gaussian noise with standard deviation 0.02. The optimizer is Adam with learning rate 0.18 for 80 updates. At each update, a softmax over the 500 item logits is used to compute differentiable proxy metrics and the minimum logging-policy improvement is maximized. Hard top-K rankings are evaluated after every update. Curves are produced for $\rho \in \{0, 1\}$ with five repetitions. These learning curves should be viewed as an optimization diagnostic, not as the optimizer used for the main KuaiRec comparison.

D Additional Experimental Analyses

D.1 Sensitivity to Ranking Length

We further examine whether the effectiveness of ODR depends on the ranking length K. This analysis is particularly relevant to our motivation because objective diversity can be expressed not only through the choice of items but also through their allocation across multiple positions in a recommendation slate. When $K = 1$, the policy has limited flexibility to simultaneously serve multiple conflicting objectives, whereas a longer ranking can allocate different positions to items favored by different objectives.

Figure 6 shows that the qualitative advantage of ODR is preserved across different ranking lengths in both the synthetic and KuaiRec experiments. The advantage is relatively limited for very short rankings and becomes more visible as K increases. This behavior is consistent with the mechanism underlying objective diversity: with a single recommendation position, satisfying one objective can directly exclude items favored by another objective, while a longer slate provides more degrees of freedom for balancing heterogeneous objectives.

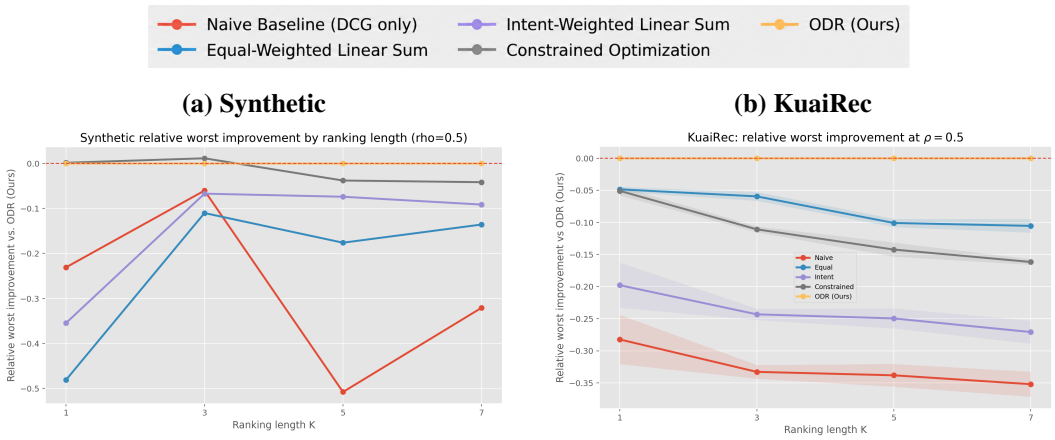

Figure 6: Sensitivity to ranking length K. For each $K \in \{1, 2, 3\}$, we vary the intent-mixture parameter ρ and compare the worst-objective performance of each baseline relative to ODR. The proposed method serves as the zero reference in the relative-performance panels.

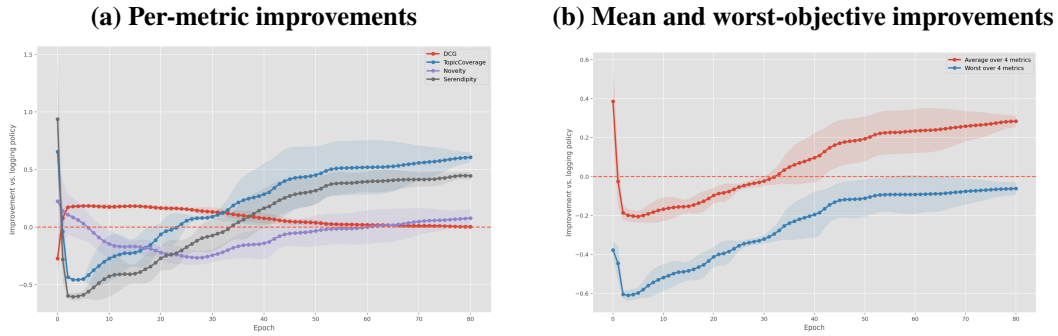

Figure 7: Learning dynamics of ODR in the synthetic environment at $\rho = 0$. (a) Improvements of the four individual objectives relative to the logging policy. (b) Mean improvement over the four objectives and per-context worst-objective improvement.

The results also suggest that the benefit of robust objective optimization is not specific to the default ranking length used in the main experiments. In particular, the relative worst-objective curves indicate that several baselines can substantially underperform ODR for particular values of ρ even when K is changed. At the same time, the magnitude of the advantage depends on both K and the intent mixture, indicating that ranking length interacts with the degree of objective conflict rather than simply producing a uniform improvement.

D.2 Learning Dynamics of ODR

We further inspect the optimization dynamics of ODR to understand how the max–min objective redistributes performance across competing objectives during training. We focus on $\rho = 0$, the relevance-heavy end of the intent-mixture spectrum, and report both the individual objective improvements and the mean and per-context worst-objective improvements.

Figure 7 illustrates the balancing behavior induced by the max–min objective. During the initial optimization phase, several discovery-oriented objectives deteriorate substantially, while DCG improves. Subsequent updates reverse this imbalance: topic coverage and serendipity recover strongly, novelty also improves, and part of the initial DCG gain is gradually traded away. Correspondingly, after an initial drop, the per-context worst-objective improvement steadily recovers throughout training, while the mean improvement eventually becomes positive. These trajectories are consistent with the active-worst-objective interpretation of ODR: optimization pressure does not remain concentrated on a fixed objective but shifts as the weakest objective changes during learning.

(a) Per-metric improvements **(b) Mean and worst-objective improvements**

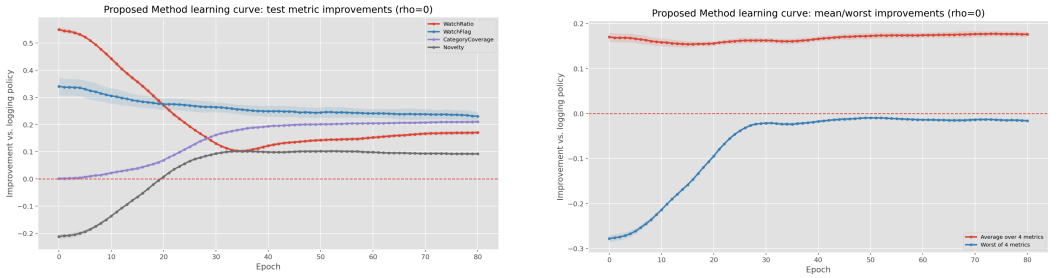

Figure 8: Learning dynamics of the ODR optimization diagnostic on KuaiRec at $\rho = 0$. (a) Improvements of WatchRatio, WatchFlag, CategoryCoverage, and Novelty relative to the logging policy. (b) Mean improvement over the four objectives and per-context worst-objective improvement.

Notably, the final per-context worst-objective improvement remains below zero even though the aggregate improvements of the individual objectives are approximately nonnegative. This is not contradictory because the robust metric takes the minimum *within each context* before averaging. Different contexts can have different weakest objectives, so $\mathbb{E}_x[\min_z G_z(x;\pi)]$ can be strictly smaller than $\min_z \mathbb{E}_x[G_z(x;\pi)]$.

Figure 8 shows a similar but smoother balancing pattern in the KuaiRec diagnostic. Novelty begins substantially below the logging-policy reference and improves steadily during optimization, while category coverage also increases from approximately zero. At the same time, the initially large WatchRatio improvement decreases, indicating that ODR trades part of the performance of an already strong objective for improvements in weaker ones. WatchFlag remains positive throughout training. As a result, the per-context worst-objective improvement moves from a substantially negative value toward zero, while the mean improvement over the four objectives remains positive and relatively stable.

Together, these learning curves provide a direct view of the mechanism underlying ODR. The method does not maximize all objectives simultaneously or preserve the initially strongest objective; rather, it reallocates optimization pressure toward poorly served objectives, improving the robust value while maintaining favorable average performance. The KuaiRec curves are intended as an optimization diagnostic rather than the training trajectory of the main KuaiRec comparison, which uses the finite candidate-ranking procedure described in Appendix C.

E Broader Impact

Recommender systems increasingly mediate access to information, entertainment, and commercial opportunities. ODR is motivated by reducing severe recommendation failures when a user's active intent differs from the objective emphasized by the current recommendation strategy. By protecting the worst-performing plausible intent, the method may make recommendation more robust to short-term changes in user needs. However, our experiments evaluate robustness in intent-specific ranking metrics and do not establish a causal improvement in user satisfaction, retention, or other long-term outcomes.

The benefits of ODR depend critically on the choice of the intent set $\mathcal{Z}$. The method protects only objectives explicitly included in this set, and deciding which objectives to include is itself a design choice. A system may therefore be robust with respect to relevance, diversity, or novelty while remaining insensitive to other dimensions of user welfare. Moreover, the max–min objective can be conservative: protecting unlikely intents may reduce performance on a dominant intent. Practical deployments should therefore validate the semantics and normalization of protected metrics and, when reliable intent-frequency information is available, consider less conservative ambiguity sets.

Changing the balance among relevance, diversity, novelty, and related objectives can also alter exposure across items and content providers. Such changes may improve discovery but can redistribute

traffic or economic opportunities in ways not captured by user-side ranking metrics. Our current formulation does not explicitly model these provider-side or ecosystem effects.

Finally, robustness across intents should not be conflated with demographic fairness, content safety, or information quality. ODR protects user objectives rather than demographic groups, and these additional concerns should be addressed through explicit constraints, objectives, or safety mechanisms when relevant. Overall, we view ODR as a mechanism for reducing one specific source of recommendation brittleness, while broader deployment decisions still require careful consideration of objective design and downstream effects.