

Measuring What Matters: Consistency and Compactness in Evaluation of Counterfactual Explanations

AMIR REZA MOHAMMADI, University of Innsbruck, Innsbruck, Austria

ANDREAS PEINTNER, University of Innsbruck, Innsbruck, Austria

MICHAEL MUELLER, University of Innsbruck, Innsbruck, Austria

EVA ZANGERLE, Universität Innsbruck, Innsbruck, Austria

Explainability in recommender systems (RS) remains a pivotal challenge. Counterfactual explanations have emerged as a particularly actionable paradigm, offering intuitive “what-if” reasoning. However, their evaluation lacks principled standards. Current metrics primarily assess whether explanations change the top-ranked recommendation, overlooking two fundamental aspects of explanation quality. First, evaluation results are frequently *inconsistent*, as metrics are tightly coupled to the underlying recommender’s performance. Second, explanations are rarely assessed for *compactness*—whether changes are sufficiently small to remain interpretable. Oversized counterfactuals may technically succeed but fail to provide practical insight.

In this work, we advocate for a holistic evaluation perspective centered on *consistency* and *compactness*. We systematically analyze how extending evaluation beyond top-1 to top-k recommendations improves metric stability and reduces dependence on recommender fluctuations. In parallel, we introduce compactness-aware evaluation criteria that quantify the minimality of counterfactual modifications. Through extensive experiments across multiple datasets and models, we demonstrate that jointly considering these dimensions yields more reliable assessments. Our findings expose key factors driving metric instability, highlight the trade-off between effectiveness and explanation size, and provide practical guidelines toward standardized, faithful, and compact evaluation of counterfactual explanations in recommender systems.

CCS Concepts: • **Information systems** → **Recommender systems**; **Evaluation of retrieval results**; • **Human-centered computing** → *HCI design and evaluation methods*.

Additional Key Words and Phrases: Recommender Systems, Counterfactual Explanations, Evaluation, Consistency, Compactness

1 Introduction

As the demand for transparency and interpretability in AI continues to grow [4, 17], eXplainable Artificial Intelligence (XAI) has emerged as a critical research domain. Within XAI, counterfactual explanations (CE) [12] have become powerful tools for advancing model interpretability. By illustrating how small, targeted changes to the input can lead to different model outputs, CEs provide intuitive and actionable insights into model behavior, fostering user trust and supporting accountability requirements [26]. While foundational research on CEs has primarily focused on tabular and image domains [24, 39, 55], recent work has increasingly adapted these methods to recommender systems (RS) [16, 43, 47, 59, 61].

Despite growing interest, the evaluation of CEs in RS remains underdeveloped [7]. Unlike prediction accuracy, explanation quality lacks universally accepted benchmarks, resulting in heterogeneous experimental setups and

Authors’ Contact Information: Amir Reza Mohammadi, University of Innsbruck, Innsbruck, Austria; e-mail: amir.reza@uibk.ac.at; Andreas Peintner, University of Innsbruck, Innsbruck, Austria; e-mail: a.peintner@gmx.net; Michael Mueller, University of Innsbruck, Innsbruck, Tyrol, Austria; e-mail: michael.m.mueller@uibk.ac.at; Eva Zangerle, Universität Innsbruck, Innsbruck, Tirol, Austria; e-mail: eva.zangerle@uibk.ac.at.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

© 2026 Copyright held by the owner/author(s).

ACM 2770-6699/2026/8-ART

<https://doi.org/10.1145/3839564>

inconsistent evaluation practices [34]. Recent meta-analyses further reveal that CE evaluation spans multiple dimensions—ranging from faithfulness and plausibility to minimality and robustness—yet these aspects are rarely assessed jointly or systematically. In particular, two fundamental challenges remain insufficiently addressed: *consistency* and *compactness* of counterfactual explanations.

Consistency in Evaluation. A typical CE pipeline in RS first trains a recommender model (e.g., MF) on a dataset (e.g., ML-1M), and subsequently applies a CE method to generate input perturbations that alter the model’s recommendations. However, because counterfactual explainers aim to manipulate the input to induce output changes, their measured effectiveness inherently depends on the underlying recommender. As shown in [34], variations in recommender quality can substantially alter the relative performance ranking of CE methods. This dependency can be illustrated through a thought experiment: if a recommender behaves randomly, arbitrary perturbations would frequently change its output, misleadingly suggesting that a CE method is highly effective. More critically, empirical evidence shows that the *relative ranking* of CE methods varies with recommender performance. That is, a method E_1 that outperforms E_2 under one recommender configuration may no longer do so under a slightly modified or better-performing model. We define *consistency* as the ability of an evaluation protocol to preserve the relative ranking of CE methods under minor changes in recommender performance. Formally, if E_1 outperforms E_2 when explaining recommender R_1 , then under a perturbed recommender R_2 with comparable predictive quality, E_1 should remain superior in relative terms, even if absolute scores shift. The absence of such stability undermines the validity of current benchmarking practices and complicates fair comparison across studies. A key source of this inconsistency lies in the predominant focus on the top-1 recommendation. Borrowed largely from vision and GNN explanation paradigms [23, 50], this practice overlooks a defining property of RS: their inherently ranked outputs. Recent findings indicate that top-1 predictions are particularly volatile and sensitive to minor score perturbations [31], especially in moderate-performance regimes. Consequently, evaluation protocols centered solely on top-1 changes amplify randomness and distort explanation assessment.

To mitigate this issue, we hypothesize that extending evaluation beyond the top-1 item to the top- k recommendations ($1 \leq k \leq 5$) improves consistency. By aligning evaluation with the ranked nature of RS outputs, top- k -based assessment reduces volatility and yields more stable comparisons across CE methods.

Compactness of Explanations. Beyond consistency, explanation *compactness* is crucial for practical interpretability. Compactness refers to the minimality of a counterfactual explanation—i.e., how many user–item interactions must be modified to achieve the desired outcome. In RS, counterfactuals often consist of removing or altering multiple historical interactions. However, large counterfactual sets are cognitively demanding and rarely actionable for end-users. While prior work has occasionally reported explanation size as a secondary statistic, its systematic impact on evaluation remains largely unexplored. Importantly, explanation effectiveness may vary across different size regimes: certain methods may excel at identifying a single highly influential interaction, whereas others require broader perturbations to induce changes. Ignoring this dimension risks favoring methods that produce overly large yet superficially effective counterfactuals. To address this gap, we analyze CE performance under varying granularity levels and controlled explanation sizes. By explicitly examining how performance evolves as progressively more interactions are removed, we disentangle precision in importance ranking from brute-force perturbation strategies. This perspective allows us to assess not only whether a method succeeds, but *how efficiently* it achieves the counterfactual outcome.

Contributions. In this paper, we provide a systematic investigation of consistency and compactness in CE evaluation for recommender systems. We:

- Analyze the instability of existing evaluation protocols and formalize consistency as a core requirement for reliable benchmarking.

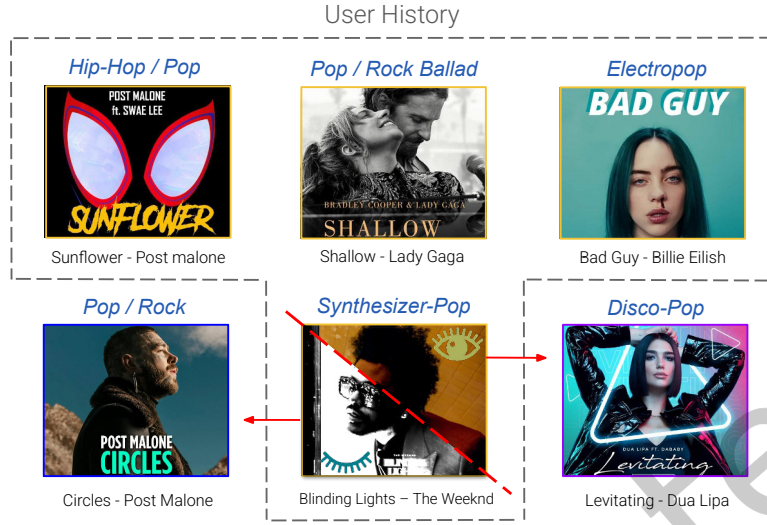


Fig. 1. Illustrative example of a counterfactual explanation (CE) in a music recommender system. Given the user's interaction history, the system recommends *Levitating*. The CE identifies *Blinding Lights* as a critical interaction whose removal would change the recommendation to *Circles*.

- Demonstrate that extending evaluation from top-1 to top- k recommendations substantially improves ranking stability across CE methods.
- Introduce a compactness-aware evaluation perspective that examines explanation effectiveness across different size regimes.
- Conduct extensive experiments across two recommender models, six state-of-the-art CE methods, and three widely used datasets.

2 Related Work

The field of interpretable machine learning has witnessed substantial progress in recent years [16, 18, 43, 52]. Early research primarily focused on intrinsically interpretable models, aiming to design recommendation architectures whose internal representations are directly human-understandable [40, 41]. In parallel, post-hoc explainability emerged as a complementary paradigm, treating recommenders as black boxes and generating explanations retrospectively without altering the underlying model [25, 42, 54].

Recent large-scale analyses highlight that, despite rapid methodological advances, evaluation practices in XAI remain fragmented. A systematic review of more than 300 papers reports that a substantial fraction relies solely on anecdotal evidence, while structured quantitative evaluation is often limited to a small subset of explanation properties [37]. The survey proposes twelve conceptual properties (Co-12), including *Correctness*, *Compactness*, and *Consistency*, and argues that explainability is inherently multi-dimensional. However, most works evaluate only one or two properties in isolation. In recommender systems—where outputs are inherently ranked and explanations are tightly coupled to user interaction histories—such partial evaluation becomes particularly problematic. Our work builds upon this perspective by systematically analyzing consistency and compactness in the specific context of counterfactual explanations for RS.

Explanation Paradigms in Recommender Systems

Explainable AI for recommender systems has gained increasing attention due to its relevance for transparency, trust, and regulatory compliance [8, 22]. Existing approaches can broadly be categorized into model-specific, post-hoc, and feature- or knowledge-driven methods.

Model-specific approaches embed interpretability into the recommendation architecture itself. Early work on explainable matrix factorization aligned latent factors with interpretable components [2, 3, 27]. More recent models such as ProtoMF [33] and AI2V++ [14] employ prototypes or attention mechanisms to highlight influential interactions. While effective, these methods are tightly coupled to specific architectures.

Post-hoc approaches aim to explain predictions without modifying the recommender. General-purpose attribution methods such as LIME [45] and SHAP [32] estimate feature importance via local surrogates or Shapley values [20]. Influence-based techniques, including FIA [11] and ACCENT [52], adapt influence functions to quantify the contribution of historical interactions. Similarity-based heuristics such as Jaccard [5] or Cosine similarity [48] provide lightweight importance estimates. However, these approaches are predominantly correlational and do not explicitly model ranking sensitivity.

Another line of work leverages side information such as textual reviews, aspects, or knowledge graphs. Aspect-based models like the Explicit Factor Model (EFM) [60] align latent dimensions with interpretable attributes, while path-based recommenders justify predictions through semantic graph paths [10, 28]. Recent work questions the faithfulness of attention mechanisms and proposes causal path reasoning alternatives [29]. Although expressive, these methods require rich auxiliary data and are not applicable in pure interaction-only settings.

Counterfactual Explanations for Recommendation

Counterfactual explanations (CEs) describe how recommendations would change under hypothetical modifications to user or item data [54]. Due to their actionability and causal interpretation, CEs have become a prominent post-hoc paradigm in RS.

Early CE approaches relied on heuristic perturbations or search procedures to identify influential interactions. Reinforcement learning-based frameworks such as CERec [57] and CFairER [56] reduce the combinatorial search space by learning policies over attributes or graph paths, often with fairness objectives. Conversational recommenders integrate counterfactual reasoning to refine explanations via user feedback [58]. Algorithmic recourse approaches such as RecRec [53] focus on actionable changes that alter item rankings from the perspective of content providers.

Model-agnostic frameworks such as LXR [6] improve scalability by learning explanation masks through self-supervised objectives, avoiding explicit search over perturbations. While LXR establishes a strong baseline for interaction-level CEs, it performs best under relatively large perturbation budgets and does not explicitly account for stability across the ranked output list.

Evaluation of Counterfactual Explanations

Despite methodological advances, evaluation practices for CEs remain limited. Most works assess explanation quality based on whether the top-1 recommendation can be flipped [6, 46, 51, 61], thereby neglecting the inherently list-wise nature of recommender outputs. Recent empirical findings demonstrate that CE evaluation can be inconsistent: variations in recommender performance may change the relative ranking of explanation methods [34, 35]. This instability challenges the reliability and reproducibility of existing benchmarks.

In contrast, our work introduces a list-wise evaluation perspective that explicitly incorporates multiple top-ranked items. By aligning evaluation with the ranked structure of recommender outputs and systematically analyzing explanation compactness across perturbation budgets, we move toward more consistent and principled benchmarking of counterfactual explanations in recommender systems.

Algorithm 1: Consistent Evaluation of CE Methods

Input: Explainer e , recommender f_θ , test users $\mathcal{U}$, rank threshold $T \in \{5, 10, 20\}$, top- k recommendations tk

Output: Evaluation score POS-P@T

Load recommender parameters $f_\theta^{(c)}$;

foreach rank threshold $T \in \{5, 10, 20\}$ **do**

 Initialize accumulators: POS_total $\leftarrow 0$;

foreach user $u \in \mathcal{U}$ **do**

$\mathbf{x}_u \leftarrow$ user's interaction vector ;

$\mathcal{R}_u^k \leftarrow$ top- k recommendations from $f_\theta^{(c)}(\mathbf{x}_u)$;

foreach $tk \in \mathcal{R}_u^k$ **do**

$i_{tk}^* \leftarrow tk$ -th item in $\mathcal{R}_u^k$;

$CE \leftarrow e(\mathbf{x}_u, i_{tk}^*)$

 Generate positive sequence $\{\mathbf{x}_u^{(1)}, \dots\}$ by removing most relevant items from CE ;

 rank_pos $\leftarrow \text{rank}(i^*; \mathbf{x}_u^{(t)})$;

 POS_total += $\mathbb{1}[\text{rank}_{\text{pos}} > k]$;

end

end

 POS-P@T $_{(k)} \leftarrow \text{POS_total} / (|\mathcal{U}| \cdot k)$

end

return All POS-P@T $_{(k)}$

3 Consistency

To investigate the impact of extending the evaluation from top-1 items to top- k items, we propose the following approach.

Preliminaries. Let $\mathcal{U}$ denote the set of users and $\mathcal{I}$ the set of items. Each user $u \in \mathcal{U}$ is represented by a binary interaction vector $\mathbf{x}_u \in \{0, 1\}^{|\mathcal{I}|}$, where $\mathbf{x}_u[i] = 1$ indicates that user u has interacted with item i . A recommender model f_θ maps this input to a predicted score vector, with $f_\theta(\mathbf{x}_u)[i]$ representing the predicted relevance of item i . The RS generates a ranked list of recommendations for each user. We denote the top- k recommended items as $\mathcal{R}_u^k = \{i_1^*, i_2^*, \dots, i_k^*\}$, where i_j^* is the j -th item in the ranked recommendation list. These are the items for which we seek to evaluate CEs.

3.1 Counterfactual Explanation Metrics

Given a counterfactual explainer e , we obtain a relevance ranking over user history items with respect to each item i_j^* in $\mathcal{R}_u^k$. Based on this relevance, we iteratively perturb the user input vector $\mathbf{x}_u$ by removing one item at a time, yielding a sequence $\mathbf{x}_u^{(t)}$ of perturbed inputs. Following the literature [6, 34], we compute the following metrics:

Positive Perturbation (POS-P@T) This metric quantifies how quickly the top- k recommended items fall below the rank threshold T during perturbation. For user u , the POS-P@T score is:

$$\text{POS-P@T}_u = \sum_{j=1}^k \sum_{t=1}^{|\mathbf{x}_u|} \mathbb{I} \left[\text{rank}(i_j^*; \mathbf{x}_u^{(t)}) > T \right] \quad (1)$$

where $\text{rank}(i_j^*; \mathbf{x}_u^{(t)})$ is the rank of i_j^* in the output of f_θ given the perturbed input $\mathbf{x}_u^{(t)}$ and the indicator function $\mathbb{I}[\cdot]$ checks whether the item still appears within the rank threshold T . The key difference between this formulation and the literature [6, 34] lies in computing the inner value over the top- k items ($\sum_{j=1}^k$) rather than only the top-1 item.

The final POS-P@T score is the average over all users:

$$\text{POS@T} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \text{POS-P@T}_u \quad (2)$$

We have detailed the computation process of this metric in Algorithm 1.

Negative Perturbation (NEG-P@T) Analogous to the positive perturbation setting, we define the NEG-P@K metric to evaluate the robustness of the top- m recommended items under negative perturbations. Here, instead of removing the most important items, we iteratively remove the least important items from the user's interaction vector $\mathbf{x}_u$, again based on the importance ranking returned by the counterfactual explainer e . This metric evaluates how well each top- k items ($1 \leq k \leq 5$) maintain their position within the T rank threshold under negative perturbation, where less relevant items are removed:

$$\text{NEG-P@T}_u = \sum_{j=1}^k \sum_{t=1}^{|\mathbf{x}_u|} \mathbb{I} \left[\text{rank}(i_j^*; \mathbf{x}_u^{(t)}) \leq T \right] \quad (3)$$

where $\mathbf{x}_u^{(t)}$ now represents the user interaction vector after removing the t -th *least important* item, and $\text{rank}(i_j^*; \mathbf{x}_u^{(t)})$ is the resulting rank of the explained item and the overall NEG-P@T score is:

$$\text{NEG@T} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \text{NEG-P@T}_u \quad (4)$$

NDCG Perturbations (NDCG-P) uses the Normalized Discounted Cumulative Gain (NDCG) in a positive perturbation test. It evaluates how quickly the explained item drops in ranking as the most important user data is gradually removed. For each step m ,

$$\text{NDCG-P}(m) = \frac{1}{\log_2 (1 + \text{rank}(\mathbf{x}_u^{\text{pos}}(m)))}. \quad (5)$$

3.2 Consistency Metrics for Ranking Evaluation

To quantify the stability and consistency of explanation method rankings across recommender systems with varying performance levels, we employ two complementary metrics: Spearman's rank correlation coefficient and rank variance. These metrics collectively provide a comprehensive assessment of ranking reliability, capturing different aspects of consistency from global rank-order agreement to position-specific stability and variance in ranking positions.

3.2.1 Spearman's Rank Correlation Coefficient. Spearman's rank correlation coefficient (ρ) [49] measures the monotonic relationship between two rankings, providing a nonparametric assessment of rank-order agreement that is robust to outliers and does not assume linear relationships. For two sets of rankings R_i and R_j corresponding to different performance levels of a recommender system, Spearman's ρ is calculated as:

$$\rho = 1 - \frac{6 \sum_{k=1}^n d_k^2}{n(n^2 - 1)} \quad (6)$$

where n is the number of explanation methods being ranked, and d_k represents the difference between the ranks assigned to the k -th explanation method in the two rankings: $d_k = R_i(k) - R_j(k)$. The coefficient ranges from -1 to $+1$, where $\rho = 1$ indicates perfect positive correlation (identical rankings), $\rho = 0$ indicates no correlation, and $\rho = -1$ indicates perfect negative correlation (completely reversed rankings). In our consistency analysis, we compute pairwise Spearman correlations between all pairs of performance levels and report the mean correlation across all pairs. Higher Spearman correlation values indicate that explanation methods maintain more consistent relative rankings regardless of the underlying recommender's performance, suggesting that the evaluation metric reliably identifies explanation quality independent of model-specific characteristics.

3.2.2 Rank Variance. Rank variance [15] quantifies the degree of fluctuation in the ranking position of each explanation method across different recommender performance levels. For each explanation method m , we compute the variance of its ranking positions across all p performance levels:

$$\text{Var}(m) = \frac{1}{p} \sum_{i=1}^p (R_i(m) - \bar{R}(m))^2 \quad (7)$$

where $R_i(m)$ denotes the rank of method m at performance level i , and $\bar{R}(m) = \frac{1}{p} \sum_{i=1}^p R_i(m)$ represents the mean rank of method m across all performance levels. The overall rank variance is then computed as the mean variance across all explanation methods:

$$\text{Rank Variance} = \frac{1}{n} \sum_{m=1}^n \text{Var}(m) \quad (8)$$

Lower rank variance values indicate greater stability, as explanation methods maintain similar positions across performance levels. A rank variance of zero would indicate that all methods retain identical rankings across all performance levels, representing perfect stability. In contrast, high rank variance suggests that method rankings are highly sensitive to recommender performance characteristics, potentially indicating evaluation metrics that conflate explanation quality with model-specific artifacts. Rank variance provides a complementary perspective to Spearman correlation by focusing on absolute position changes rather than relative rank-order preservation.

4 Compactness

Compactness evaluates the minimality of an explanation, focusing on how few features or changes are required to achieve the desired counterfactual outcome. This aligns with the Occam's Razor Principle [9], which suggests that when two explanations are equally effective, the simpler one is preferred. Formally, let x_u represent the set of items in the interaction history for each user $u \in \mathcal{U}$. Let E be an explanation method that outputs a subset $R_u \subseteq x_u$, where $|R_u|$, representing the items to be removed from x_u to alter the recommendation. Then the compactness metric S_E is defined as:

$$S_E(\tau) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} |R_u|, \quad \text{subject to } |R_u| \leq \tau \quad (9)$$

where τ is the threshold for size of explanation. This metric represents the average number of changes required across all test users, subject to the constraint that no more than τ changes are allowed per user. For S_E , lower values indicate better performance, showing that explanation method is able to make fewer changes to history to provide CE.

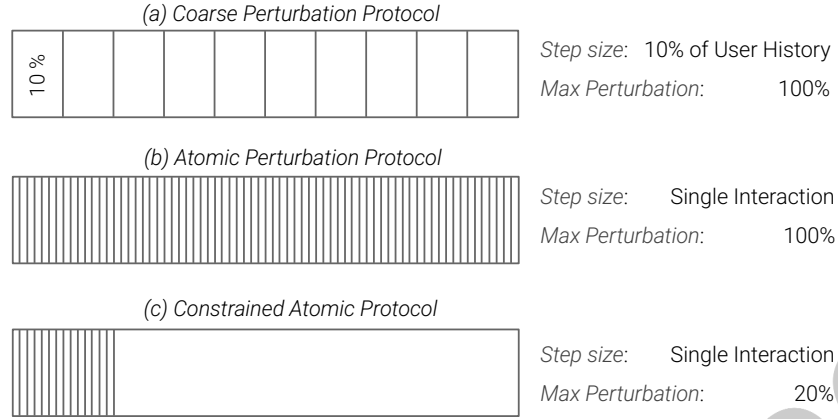


Fig. 2. Illustration of the evaluation protocols. Each rectangle represents the full interaction history of a user.

4.1 Evaluation Protocols

To analyze the behavior of different CE methods based on the perturbation size and limit, we have defined three evaluation protocols as follows:

Coarse Perturbation. We begin by evaluating all methods under the widely adopted *binned evaluation* protocol [6], where user interactions are removed in coarse percentage increments (typically 10%) until the recommendation outcome changes (Figure 2(a)). This protocol remains the de facto standard for benchmarking CE methods and is therefore included to ensure comparability with prior work. However, the coarse protocol fundamentally limits the interpretability of results. Since all perturbations within a bin are aggregated, the protocol assigns identical scores to explanations that succeed after removing a single interaction and those that require removing up to 10% of the user history. As a result, it fails to discriminate between explanations that are *precise and minimal* and those that are *coarse but eventually effective*. This limitation is particularly problematic for CEs, whose core objective is to identify the smallest and most plausible change that alters the recommendation. Moreover, large, percentage-based removals substantially distort the user profile and can easily push the recommender into out-of-distribution regions. In such cases, a ranking change may reflect global profile corruption rather than the faithful identification of truly influential interactions. Consequently, binned evaluation alone is insufficient to assess explanation faithfulness or practical plausibility. For this reason, we next turn to atomic and constrained evaluation protocols that explicitly expose how quickly and how efficiently different methods induce counterfactual ranking changes.

Atomic perturbation. In this approach we evaluate CE quality under protocols that more closely reflect the intended use of CE: inducing a change in the *top-1* recommendation through minimal and plausible modifications to the user interaction history. Unlike binned evaluation, these protocols expose how precisely an explainer identifies the most influential interactions and how robust it remains under realistic perturbation budgets. In the Atomic protocol, interactions are removed one at a time according to their explanation scores (Figure 2(b)). This setting enables a stepwise analysis of ranking sensitivity and directly measures the quality of the learned importance ordering. Methods that correctly prioritize the most critical interactions cause rapid degradation of ranking metrics in the early steps, whereas less precise methods require multiple removals before any substantial effect is observed.

Constrained Atomic perturbation. While the atomic protocol isolates precision, it does not explicitly constrain the extent of profile modification. To assess plausibility under realistic conditions, we further restrict the maximum number of removable interactions to 20% of the user history (Figure 2(c)). This constrained setting reflects practical counterfactual scenarios in which explanations must succeed through limited changes.

5 Experiments and Results

In this section, we describe the setup and results of our extensive experiments. Given our goal of focusing on consistency and compactness of counterfactual explanations for recommender systems, we establish the following research questions:

- RQ1 Evaluation Consistency:** To what extent does extending the evaluation horizon from the top-1 item to the top- k recommendations improve the stability of counterfactual explainer rankings across recommenders of varying performance?
- RQ2 Quantifying Stability:** How effective are list-wise aggregation strategies (e.g., mean or weighted ranking) at reducing rank variance and improving correlation compared to standard top-1 metrics?
- RQ3 Impact of Compactness:** How does the effectiveness of counterfactual methods evolve when evaluated under strictly controlled explanation size regimes (compactness), and does the relative ranking of methods change as the size budget varies?
- RQ4 Minimality vs. Effectiveness:** How do different explanation methods perform under constrained perturbation budgets (minimality) versus unconstrained settings, and what trade-offs exist between granular precision and global explanation success?

5.1 Experimental Setup

5.1.1 Evaluation Setup. We use the MovieLens 1M (ML-1M) [19], Yahoo! Music [13] and Pinterest [21] datasets following the literature [6, 34], focusing on collaborative filtering models based on implicit user-item interactions. To simulate implicit ratings in ML-1M, following [6], we retained only ratings of 3.5 or higher and included users and items with at least two ratings. This resulted in 575,128 ratings from 6,037 users across 3,381 items. For the Yahoo! Music dataset, we retained ratings of 70 or higher, and following [6] we took a random sample of 486,744 ratings from 19,155 users for 9,362 items from the Pinterest dataset. We split each dataset into training and testing (80/20) subsets using user identifiers.

For each user, we generate a ranked list of recommendations and evaluate the explanation of the top- k recommendations. For each k value, we use four levels of recommender performance by training the recommender based on Hit Rate and saving model checkpoints after 25%, 50%, 75%, and full training, to simulate recommendations of various quality levels. We evaluate 24 configurations (3 dataset x 2 Recommenders x 4 performance levels), repeat our experiments three times, and report the average performance. We share our dataset processing scripts and the source code.¹

5.1.2 CE Methods. Our evaluation includes a range of CE methods, from established baselines to recent advancements. Specifically, we tested two similarity-based approaches, two traditional explainers, and two state-of-the-art methods:

Jaccard [5]/*Cosine* [48] Similarity generate explanations by comparing an item to those in the user’s history using Jaccard [5] or Cosine [48] similarity metrics based on co-interacted users.

LIME-RS [38] is a modified version of the LIME framework designed specifically for recommender systems. It produces explanations by approximating the complex recommendation model with a locally interpretable linear

¹<https://github.com/dbis-uibk/CE4RS-Eval>

model, capturing relationships within a neighborhood around the user’s data. We applied LIME-RS to the user’s historical interactions, carefully adjusting the number of samples based on performance on the validation set.

SHAP (SHapley Additive exPlanations) [61] quantifies the influence of individual features on a model’s predictions. By leveraging game theory, it assigns a value to each feature according to its contribution to the final output. SHAP, though effective, is computationally intensive due to the exponential number of possible perturbations in a user’s history.

ACCENT (Action-based Counterfactual Explanations for Neural Recommenders for Tangibility) [52] is a CE framework that extends influence functions to provide actionable, model-agnostic insights for neural recommenders. It extends techniques originally designed for latent factor models, making them applicable to a broader class of neural recommender systems.

LXR [6], the state-of-the-art CE approach, employs self-supervised learning to generate explanation masks that highlight critical user data without perturbations using a gradient learning approach. We note that the recently proposed DiceRec[36] method was not yet available at the time this research was conducted and therefore is not included in our comparisons.

Our study includes two recommenders: (1) *Matrix Factorization (MF)*, which decomposes the user-item interaction matrix into lower-dimensional latent factors [1]. Recent reproducibility studies indicate that MF continues to achieve competitive performance [44]. We used the implementation from [6] for easier comparison. (2) *Variational Autoencoder (VAE)*, a generative model that encodes and decodes data while learning a probabilistic latent representation [30]. VAE-based models have demonstrated strong performance in collaborative filtering tasks. In our experiments, we adopt a VAE-based recommender with an architecture similar to that of [30].

Based on our hypothesis that extending evaluations to include lower-ranked recommendations—such as the top-3 items—enhances consistency, we incorporate lower-ranked items in the evaluation. Particularly, we compute performance separately for top-1, top-2, top-3, and higher-ranked recommendations, then aggregate these performance values to obtain a comprehensive assessment. For instance, computing the value for top-3 involves selectively masking data from the user’s history to ensure that the third-highest recommended item drops to a lower rank (rank ≥ 4). This method ensures that the evaluation captures the influence of multiple recommendation ranks rather than relying solely on the highest-ranked item, leading to a more stable comparison of CE methods.

5.2 Consistency Experiments

To investigate the impact of extending the evaluation from top-1 items to top- k items, we evaluate six representative CE methods using the POS, NEG, and NDCG-P metrics, analyzing their performance across four thresholds of recommendation length (top- k) and four levels of recommender performance. First in section 5.2.1 we apply the top- k approach and check the performance through ranking improvement and visualizations. Then in section 5.2.2 we use the metrics to formally evaluate the consistency and evaluate different aggregation approaches for top- k items.

5.2.1 Ranking Consistency (RQ1). As shown in fig. 3, evaluations based solely on the top-1 recommendation (fig. 3(a)) exhibit significant fluctuations across recommenders, particularly when transitioning from the lowest-performing recommender (TP25%) to TP50%. However, when extending the evaluation to the top-2 recommendations (Fig. fig. 3(b)), we can already observe more consistency, especially for high-performing recommenders (TP50%–TP100%). Following this, setting the threshold to include the top-5 recommendations (Fig. fig. 3(c)) stabilizes the evaluation outcomes for the highest performing recommenders (TP50%–TP100%). We observed a similar pattern in other configurations, albeit with different thresholds. For the MF recommender, the ranking of methods remained stable even at smaller k values across all datasets, in contrast to the VAE. An example of this observation is presented in table 1, which shows the performance of CE methods across three levels of

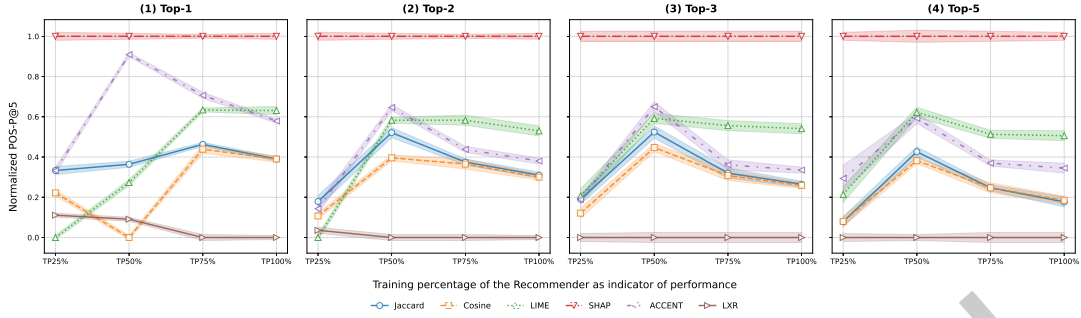
Table 1. *CE methods performance for Explaining the MF Recommender on the Yahoo! Music Dataset based on top-3 ranks.* The best results in each metric are highlighted in **bold**, and the second-best results are *underlined*. Arrows next to the metrics indicate performance direction: a downward arrow (↓) signifies that lower values are better, while an upward arrow (↑) signifies that higher values are better.

Setting	Method	T = 5		T = 10		T = 20		NDCG-P ↓
		POS ↓	NEG ↑	POS ↓	NEG ↑	POS ↓	NEG ↑	
TP50%	Jaccard	0.767	0.837	0.874	0.921	0.932	0.957	0.641
	Cosine	0.766	0.836	0.876	0.920	0.935	0.956	0.598
	LIME-RS	0.779	0.840	0.882	0.924	0.941	0.960	0.647
	SHAP	0.813	0.820	0.915	0.902	0.964	0.936	0.695
	ACCENT	0.751	0.830	0.856	0.910	0.918	0.948	0.662
	LXR	0.743	0.861	0.829	0.929	0.894	0.968	0.584
TP75%	Jaccard	0.507	0.751	0.623	0.815	0.692	0.858	0.528
	Cosine	0.502	0.753	0.618	0.817	0.684	0.861	0.516
	LIME-RS	0.488	0.715	0.544	0.798	0.603	0.846	0.555
	SHAP	0.631	0.665	0.712	0.767	0.781	0.809	0.580
	ACCENT	0.459	0.731	0.538	0.800	0.589	0.852	0.536
	LXR	0.447	0.764	0.529	0.827	0.574	0.872	0.485
TP100%	Jaccard	0.463	0.759	0.512	0.795	0.561	0.829	0.463
	Cosine	0.457	0.776	0.507	0.801	0.553	0.836	0.445
	LIME-RS	0.476	0.765	0.527	0.793	0.579	0.822	0.476
	SHAP	0.618	0.616	0.666	0.663	0.721	0.701	0.520
	ACCENT	0.453	0.692	0.491	0.748	0.528	0.781	0.482
	LXR	0.431	0.792	0.471	0.811	0.509	0.851	0.431

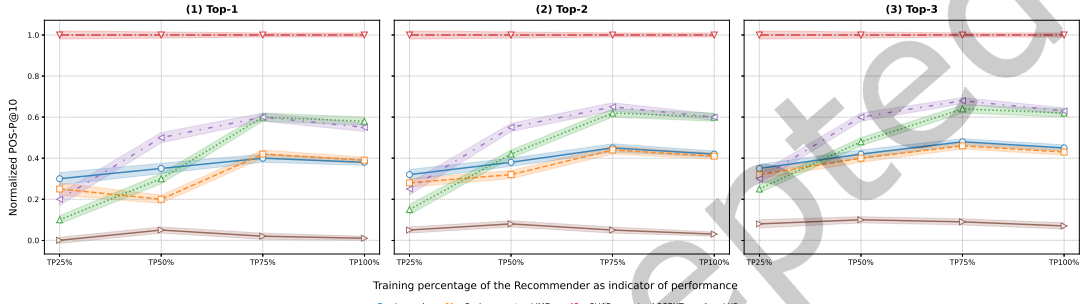
recommender performance (determined by training percentage). The results, based on top-3 values, demonstrate the consistency of the comparative rankings among the methods. The observed consistency when considering only the top-3 items stems from the simplicity of the MF model, which results in a more straightforward recommendation task and lower performance variability among recommenders, ultimately leading to less variability in explainers. Additionally, for the NEG metric, we found that a lower k , specifically $k = 2$, consistently yielded stable performance across all recommender configurations. Overall, this evidence highlights the importance of considering list-wise recommendations in CE evaluations. Adopting such an approach not only enhances evaluation consistency but also ensures better alignment with the inherent characteristics of recommendation systems.

Importantly, our findings show that evaluation consistency is not uniformly affected across all metrics. POS, which measures how quickly the explained item drops out of the top- K list under positive perturbations, is particularly sensitive to the quality of the recommender. In contrast, NEG, which checks whether the explained item remains within the top- K under less relevant perturbations, proves to be more stable even when the recommender is weaker. These results suggest that the choice of metric—and the value of k —should be treated as a tunable hyperparameter, influenced by both the dataset and the recommender architecture.

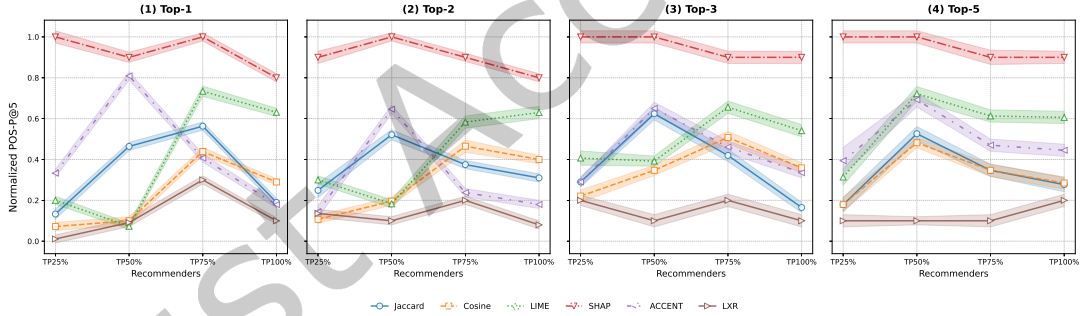
Moreover, the use of multiple performance checkpoints for the recommender further reinforces the robustness of our evaluation framework. By capturing performance across training stages (25-100%), we ensure that the evaluation of explainers is not biased by any single fixed model state. This approach not only provides a more



(a) VAE recommender on the ML-1M dataset



(b) MF recommender on Pinterest dataset



(c) VAE recommender on Yahoo! Music dataset

Fig. 3. Comparison of CE methods based on (a,c) $POS@5 \downarrow$ (lower value is the better) and (b) $POS@10 \downarrow$ across four performance levels. The figure shows the impact of going beyond top-1 (a) and considering top- k (2-4) recommendations on improving consistency when evaluating CE models. To facilitate clearer comparisons, the values are normalized using min-max normalization, and shading is used to represent the variance in the results.

comprehensive view of the explainer's effectiveness but also reveals how susceptible existing metrics are to fluctuations in model quality.

5.2.2 Consistency Metrics (RQ2). To formally quantify the stability of evaluation protocols, we analyzed the Spearman rank correlation coefficient (ρ) and Rank Variance (RV) across different recommender performance levels. Table 2 presents these consistency metrics for the Pinterest dataset using the VAE model, comparing the

standard top-1 baseline against our proposed top- k aggregation strategies (Mean, Weighted Linear, and Weighted Exponential).

The results provide strong empirical evidence that extending evaluation beyond the single top-ranked item significantly enhances consistency. The baseline approach (top-1 Only) consistently exhibits high Rank Variance and low Spearman correlation across metrics. For instance, in the POS@5 metric, the baseline shows a high variance of 0.933 and a weak correlation of $\rho = 0.146$. This confirms that evaluating explanation quality based solely on the top-1 item is highly volatile and susceptible to minor fluctuations in the underlying recommender’s performance.

In contrast, aggregating results over the top- k recommendations (specifically top-1-to-5) yields substantial improvements. Using the Mean aggregation strategy for POS@5 drastically reduces Rank Variance to 0.125 and increases the Spearman correlation to 0.943. This indicates that while individual top-1 rankings may shift due to model noise, the collective quality of explanations across the top-5 items remains stable.

Comparing the aggregation strategies, we observe that while the Mean approach offers the most significant stability gains for POS metrics, the Weighted approaches (Linear and Exponential) also consistently outperform the baseline. This supports our hypothesis that a list-wise evaluation perspective—whether equally weighted or rank-discounted—provides a more reliable benchmark for counterfactual explainers than the traditional top-1 protocol.

Table 2. Progressive top- k Aggregation Analysis for Pinterest Dataset with VAE Model

Metric	Aggregation	top-1 Only		top-1-to-2		top-1-to-3		top-1-to-4		top-1-to-5	
		ρ	RV	ρ	RV	ρ	RV	ρ	RV	ρ	RV
POS@5	Baseline	0.933	0.146								
	Mean			0.791	0.458	0.791	0.458	0.943	0.125	0.971	0.063
	Weighted Linear			<u>0.838</u>	<u>0.354</u>	0.791	0.458	0.791	0.458	0.933	0.146
	Weighted Exp			0.800	0.438	0.791	0.458	0.838	0.354	0.971	0.063
POS@10	Baseline	0.876	0.271								
	Mean			0.876	0.271	0.886	0.250	0.886	0.250	0.857	0.313
	Weighted Linear			0.962	0.083	<u>0.886</u>	<u>0.250</u>	<u>0.886</u>	<u>0.250</u>	<u>0.886</u>	<u>0.250</u>
	Weighted Exp			<u>0.914</u>	<u>0.188</u>	0.886	0.250	0.886	0.250	0.886	0.250
NEG@5	Baseline	0.743	0.563								
	Mean			0.933	0.146	0.733	0.583	0.771	0.500	0.819	0.396
	Weighted Linear			<u>0.876</u>	<u>0.271</u>	0.743	0.563	<u>0.857</u>	<u>0.313</u>	0.848	0.333
	Weighted Exp			0.876	0.271	0.743	0.563	0.848	0.333	0.933	0.146
NEG@10	Baseline	0.762	0.521								
	Mean			0.848	0.333	0.733	0.583	<u>0.876</u>	<u>0.271</u>	0.943	0.125
	Weighted Linear			<u>0.791</u>	<u>0.458</u>	0.791	0.458	0.791	0.458	0.800	0.438
	Weighted Exp			0.791	0.458	0.791	0.458	0.762	0.521	0.819	0.396
NDCG	Baseline	0.810	0.417								
	Mean			0.857	0.313	0.886	0.250	0.857	0.313	0.829	0.375
	Weighted Linear			0.848	0.333	0.857	0.313	0.857	0.313	0.857	0.313
	Weighted Exp			<u>0.857</u>	<u>0.313</u>	0.857	0.313	<u>0.857</u>	<u>0.313</u>	<u>0.857</u>	<u>0.313</u>

Note: ρ = Spearman correlation, RV = Rank Variance.

Best values are in bold, second-best values are underlined. Baseline represents top-1 only performance.

Higher ρ indicates better consistency; lower RV indicates better stability.

5.3 Compactness Experiments

To assess the impact of compactness, we conduct an experiment to evaluate the performance of CE methods in different explanation size thresholds.

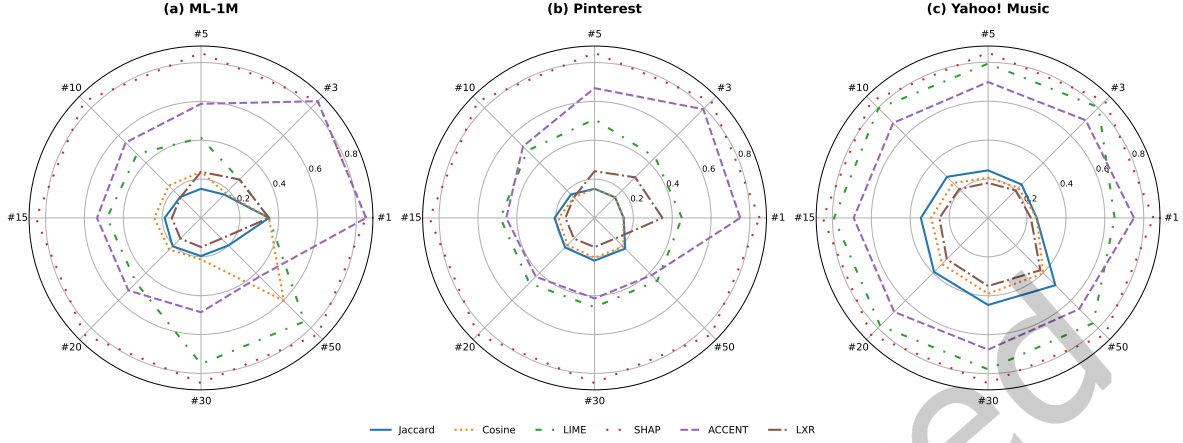


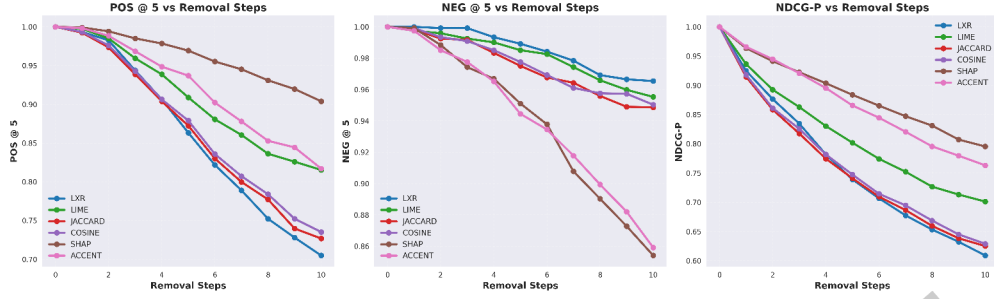
Fig. 4. Performance of CE methods across three datasets based on explanation compactness metric (S_E) using VAE recommender. The evaluation is conducted over eight explanation sizes, providing a comparative analysis of the methods. To facilitate clearer comparisons, the values are normalized using Min-max normalization. The results highlight dataset-specific performance variations, reflecting the effectiveness of each CE method on specific compactness levels.

5.3.1 Compactness Metric (RQ3). As shown in Figure 4, we compare the performance of various CE methods across all datasets using the S_E metric. These results indicate that performance is notably influenced by the chosen explanation size.

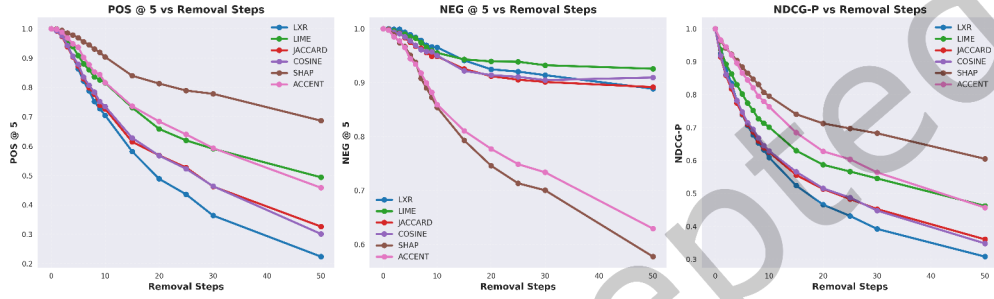
Key findings in this experiment include: (i) incorporating the compactness metric provides deeper insights into how each CE method leverages information. For instance, *LXR* exhibits superior performance when explanations are extended in length (Fig. 4a, b), suggesting that it effectively uses more extensive feature sets. In contrast, similarity-based approaches such as Cosine and Jaccard excel at shorter explanation lengths, indicating their capacity to produce concise, focused counterfactuals. This strength stems from their focus on localized feature relationships, which inherently favor compactness. (ii) In the Yahoo! Music dataset (Fig. 4c), the ranking of CE methods remains relatively stable across varying explanation sizes. Under such conditions, it may be prudent to prioritize computational efficiency over marginal performance gains. For example, if *LXR* surpasses other methods by a small margin across all explanation lengths, selecting a computationally lighter approach like Cosine Similarity could be advantageous, particularly when resource constraints are a factor. These findings underscore the importance of considering compactness as a fundamental metric in the evaluation of CE methods. By accounting for both the interpretability and resource requirements associated with explanation size, it becomes possible to make more informed and efficient decisions in the recommendation process.

5.4 Compactness Evaluation Protocols (RQ4)

Atomic perturbation. Figure 5 illustrates the performance trends under the atomic perturbation protocol with (a) 0-10 and (b) 0-50 removal steps, where we allow the removal of items until the recommendation changes, up to the full history. In this unconstrained setting, *LXR* consistently emerges as the most effective method over the long run. As shown in the POS@5 and NDCG-P plots, the *LXR* curve descends rapidly as the perturbation budget increases, indicating that it successfully identifies the set of interactions necessary to flip the recommendation, even if that set is relatively large. This aligns with *LXR*'s design as a global, learned method capable of capturing complex feature interactions over the entire user history.



(a) Atomic Perturbation protocol with 0-10 removal steps.



(b) Atomic Perturbation protocol with 0-50 removal steps.

Fig. 5. Comparative performance trends of various methods under Atomic Perturbation protocol for user history removal on the *ML-1M* dataset based on *VAE Recommender*: For both metrics, lower values are better.

Constrained atomic perturbation. However, practical interpretability often demands that counterfactuals be minimal; users are rarely interested in removing half their history to change a recommendation. To assess this, we examine the Constrained Atomic protocol (Figure 6, which restricts changes to a maximum of 20% of the user history).

This constrained view reveals a critical trade-off. While LXR excels at identifying *sufficient* sets of interactions over a large budget, it is often outperformed by simpler methods like Jaccard, Cosine, or LIME in the early, low-budget stages. For example, in the initial steps (0–10 removals) of the POS@5 and NEG@5 plots, similarity-based and local methods often induce a faster drop in the metric (where lower is better) than LXR. This suggests that while learned explainers are powerful, they may lack the “precision” required for compact counterfactuals. Heuristic and local methods, by focusing on immediate item-item similarities, are often more efficient at identifying the single most critical interaction to remove first. Consequently, for applications requiring highly compact explanations (e.g., “remove just one item”), traditional methods may still offer superior utility compared to complex state-of-the-art learners.

5.5 Discussion: Toward a Standardized Evaluation Workflow for Counterfactual Explanations

Beyond demonstrating the benefits of consistency- and compactness-aware evaluation, our findings have important implications for how future counterfactual explanation methods should be assessed and compared. Existing studies often report a single effectiveness score obtained from a specific recommender configuration and based solely on the top-1 recommendation. As our results show, such evaluations can produce unstable rankings of explanation methods because they are highly sensitive to recommender performance variations.

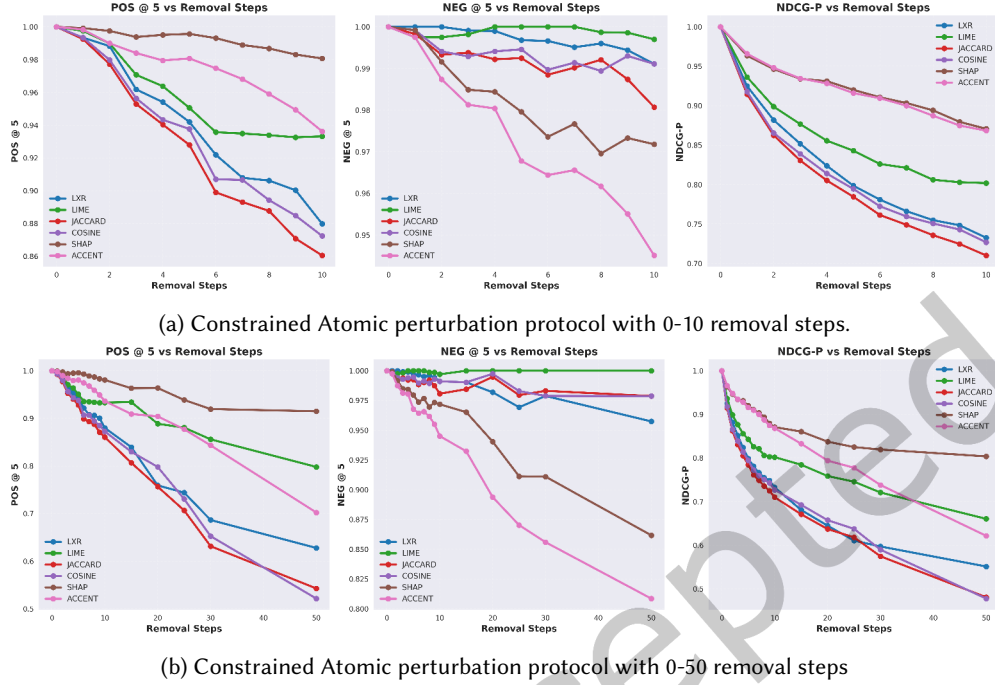


Fig. 6. Comparative performance trends of various methods under Constrained Atomic perturbation protocol for user history removal on the *ML-1M* dataset based on *VAE Recommender*: For both metrics, lower values are better.

Consequently, conclusions regarding the superiority of one method over another may not generalize across datasets, recommenders, or training conditions. To address this issue, we advocate a standardized evaluation workflow that explicitly incorporates both consistency and compactness considerations.

First, researchers should evaluate CE methods across multiple recommender performance levels rather than relying on a single trained model. This can be achieved by evaluating explainers on recommender checkpoints of varying quality, as performed in our experiments. The resulting rankings should then be analyzed using consistency metrics such as Spearman’s rank correlation and Rank Variance. High correlation and low variance indicate that an evaluation protocol reliably preserves the relative ordering of explanation methods despite fluctuations in recommender quality.

Second, evaluation should move beyond the top-1 recommendation and consider multiple items in the recommendation list. Our experiments demonstrate that aggregating results across the top- k recommendations substantially improves ranking stability. Therefore, we recommend reporting both traditional top-1 results and aggregated top- k evaluations. Based on our findings, top-5 aggregation provides a strong balance between stability and computational overhead, although the optimal value of k may depend on the recommender architecture and dataset characteristics.

Third, explanation effectiveness should be evaluated jointly with compactness. Reporting only the final success rate of a counterfactual explanation can obscure important differences between methods. Some approaches achieve strong performance only after large portions of the user profile are modified, while others are capable of inducing recommendation changes with only a few carefully selected interactions. Evaluating performance under

multiple explanation-size budgets enables a more nuanced comparison between globally effective and locally efficient methods.

Furthermore, our compactness experiments reveal that the relative ranking of CE methods can change depending on the perturbation budget. For example, learning-based methods such as LXR perform strongly when larger modification budgets are available, whereas similarity-based methods often exhibit superior performance under highly constrained settings. Consequently, evaluating methods under a single perturbation budget may provide an incomplete picture of their practical utility.

Finally, practitioners should select evaluation metrics according to the intended application. Applications that prioritize minimal user interventions should emphasize constrained compactness evaluations, whereas applications focusing on broader recommendation sensitivity may benefit from unconstrained evaluation protocols. Similarly, if robustness across recommender configurations is important, consistency metrics should be considered a primary evaluation criterion rather than a secondary analysis.

Based on these observations, we recommend that future studies report four complementary dimensions: (i) effectiveness metrics (POS, NEG, and NDCG-P), (ii) consistency metrics (Spearman correlation and Rank Variance), (iii) compactness scores across multiple explanation budgets, and (iv) performance under constrained perturbation settings. Such a protocol enables fairer comparisons between CE methods and provides a more complete characterization of explanation quality than existing top-1-only evaluation approaches. By jointly considering effectiveness, consistency, and compactness, researchers can obtain evaluations that are both more reliable and more representative of real-world explanatory requirements.

6 Conclusion and Future Work

In this work, we revisited the evaluation standards for counterfactual explanations in recommender systems, identifying two critical gaps in current practices: the instability of metrics tied to recommender performance (consistency) and the lack of emphasis on explanation size (compactness).

Through extensive experiments across multiple datasets and models, we demonstrated that the prevailing practice of evaluating only the top-1 recommendation is inherently volatile. We introduced a top- k evaluation protocol, showing that extending assessment to the top-5 recommendations and aggregating results (via mean or weighted strategies) yields significantly higher rank correlation and lower variance. This ensures that benchmarks reflect the true quality of the explanation method rather than the noise of the underlying recommender.

Furthermore, our analysis of compactness revealed that explanation effectiveness is highly sensitive to the perturbation budget. While sophisticated learning-based methods like LXR perform best when allowed larger modification budgets, our Constrained Atomic evaluation showed that simpler, similarity-based methods often excel at generating minimal, “short” counterfactuals. This finding advocates for a multi-faceted evaluation approach: future benchmarks should explicitly measure performance at different granularity levels to distinguish between methods that are globally effective versus those that are locally efficient.

Ultimately, we argue that for counterfactual explanations to be trusted and adopted, their evaluation must be as robust and user-centric as the recommendations they seek to explain. By adopting consistency and compactness as core evaluation pillars, the community can move toward more reproducible and practically relevant explanation standards.

References

- [1] Mohamed Hussein Abdi, George Onyango Okeyo, and Ronald Waweru Mwangi. 2018. Matrix Factorization Techniques for Context-Aware Collaborative Filtering Recommender Systems: A Survey. *Comput. Inf. Sci.* 11, 2 (2018), 1–10.
- [2] Behnoush Abdollahi and Olfa Nasraoui. 2016. Explainable Matrix Factorization for Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Companion Volume*. ACM, 5–6. doi:10.1145/2872518.2889405

- [3] Behnoush Abdollahi and Olfa Nasraoui. 2017. Using explainability for constrained matrix factorization. In *Proceedings of the eleventh ACM conference on recommender systems*. 79–83.
- [4] Saugat Aryal and Mark T. Keane. 2024. Even-Ifs from If-Onlys: Are the Best Semi-factual Explanations Found Using Counterfactuals as Guides?. In *Case-Based Reasoning Research and Development - ICCBR 2024, Merida, Mexico, July 1-4, 2024, Proceedings*, Vol. 14775. Springer, 33–49. doi:10.1007/978-3-031-63646-2_3
- [5] Sujoy Bag, Sri Krishna Kumar, and Manoj Kumar Tiwari. 2019. An efficient recommendation generation using relevant Jaccard similarity. *Inf. Sci.* 483 (2019), 53–64.
- [6] Oren Barkan, Veronika Bogina, Liya Gurevitch, Yuval Asher, and Noam Koenigstein. 2024. A Counterfactual Framework for Learning and Evaluating Explanations for Recommender Systems. In *WWW '24*. ACM, 3723–3733.
- [7] Christine Bauer, Eva Zangerle, and Alan Said. 2024. Exploring the Landscape of Recommender Systems Evaluation: Practices and Perspectives. *Trans. Recomm. Syst.* 2, 1 (2024), 11:1–11:31. doi:10.1145/3629170
- [8] Mustafa Bilgic and Raymond J Mooney. 2005. Explaining recommendations: Satisfaction vs. promotion. In *Beyond personalization workshop, IUI*, Vol. 5. 153.
- [9] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. 1987. Occam's razor. *Information processing letters* 24, 6 (1987), 377–380.
- [10] Tong Chen, Hongzhi Yin, Guanhua Ye, Zi Huang, Yang Wang, and Meng Wang. 2020. Try this instead: Personalized and interpretable substitute recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 891–900.
- [11] Weiyu Cheng, Yanyan Shen, Linpeng Huang, and Yanmin Zhu. 2019. Incorporating Interpretability into Latent Factor Models via Fast Influence Analysis. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Anchorage, AK, USA) (KDD '19)*. Association for Computing Machinery, New York, NY, USA, 885–893. doi:10.1145/3292500.3330857
- [12] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* (2017).
- [13] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. 2012. The Yahoo! Music Dataset and KDD-Cup '11. In *Proceedings of KDD Cup 2011 competition, San Diego, CA, USA, 2011 (JMLR Proceedings, Vol. 18)*. JMLR.org, 8–18. <http://proceedings.mlr.press/v18/dror12a.html>
- [14] Keren Gaiger, Oren Barkan, Shir Tsiptory-Samuel, and Noam Koenigstein. 2023. Not all memories created equal: Dynamic user representations for collaborative filtering. *Ieee Access* 11 (2023), 34746–34763.
- [15] Yasser Ganjisaffar, Rich Caruana, and Cristina Videira Lopes. 2011. Bagging gradient-boosted trees for high precision, low variance ranking models. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*. Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft (Eds.). ACM, 85–94. doi:10.1145/2009916.2009932
- [16] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. 2020. PRINCE: Provider-side Interpretability with Counterfactual Explanations in Recommender Systems. In *WSDM '20*. ACM, 196–204.
- [17] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 9273–9282. <https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html>
- [18] Amirata Ghorbani and James Y. Zou. 2020. Neuron Shapley: Discovering the Responsible Neurons. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. <https://proceedings.neurips.cc/paper/2020/hash/41c542dfe6e4fc3deb251d64cf6ed2e4-Abstract.html>
- [19] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (dec 2015), 19 pages.
- [20] Sergiu Hart and Andreu Mas-Colell. 1989. Potential, value, and consistency. *Econometrica: Journal of the Econometric Society* (1989), 589–614.
- [21] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 173–182. doi:10.1145/3038912.3052569
- [22] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 241–250.
- [23] Zexi Huang, Mert Kusan, Sourav Medya, Sayan Ranu, and Ambuj K. Singh. 2023. Global Counterfactual Explainer for Graph Neural Networks. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM 2023, Singapore, 27 February 2023 - 3 March 2023*. ACM, 141–149. doi:10.1145/3539597.3570376
- [24] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. 2024. Text-to-Image Models for Counterfactual Explanations: A Black-Box Approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 4757–4767.

- [25] Vassilis Kaffes, Dimitris Sacharidis, and Giorgos Giannopoulos. 2021. Model-Agnostic Counterfactual Explanations of Recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation UMAP 2021*. ACM, 280–285. doi:10.1145/3450613.3456846
- [26] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: from Counterfactual Explanations to Interventions. In *FAccT '21*. ACM, 353–362.
- [27] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
- [28] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601* (2021).
- [29] Yicong Li, Xiangguo Sun, Hongxu Chen, Sixiao Zhang, Yu Yang, and Guandong Xu. 2024. Attention Is Not the Only Choice: Counterfactual Reasoning for Path-Based Explainable Recommendation. *IEEE Trans. Knowl. Data Eng.* 36, 9 (2024), 4458–4471. doi:10.1109/TKDE.2024.3373608
- [30] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 WWW 2018, Lyon, France, April 23–27, 2018*. ACM, 689–698. doi:10.1145/3178876.3186150
- [31] Yang Liu, Alan Medlar, and Dorota Glowacka. 2023. What We Evaluate When We Evaluate Recommender Systems: Understanding Recommender Systems' Performance using Item Response Theory. In *RecSys '23*. ACM, 658–670.
- [32] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*. 4765–4774. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- [33] Alessandro B. Melchiorre, Navid Rekabsaz, Christian Ganhör, and Markus Schedl. 2022. ProtoMF: Prototype-based Matrix Factorization for Effective and Explainable Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems (Seattle, WA, USA) (RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 246–256. doi:10.1145/3523227.3546756
- [34] Amir Reza Mohammadi, Andreas Peintner, Michael Müller, and Eva Zangerle. 2024. Are We Explaining the Same Recommenders? Incorporating Recommender Performance for Evaluating Explainers. In *RecSys '24*. ACM, 1113–1118.
- [35] Amir Reza Mohammadi, Andreas Peintner, Michael Müller, and Eva Zangerle. 2025. Beyond Top-1: Addressing Inconsistencies in Evaluating Counterfactual Explanations for Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys 2025, Prague, Czech Republic, September 22–26, 2025*, Mária Bielíková, Pavel Kordík, Markus Schedl, Marco de Gemmis, Sole Pera, Rodrigo Alves, Olivier Jeunen, and Vito Ostuni (Eds.). ACM, 515–520. doi:10.1145/3705328.3748028
- [36] Amir Reza Mohammadi, Andreas Peintner, Michael M. Müller, and Eva Zangerle. 2026. Minimal-Perturbation Counterfactuals through Guided Denoising Diffusion for Recommender Systems Explanation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2026, Melbourne/VIC Australia, July 20–24, 2026*, Alistair Moffat, Falk Scholer, Hannah Bast, Marc Najork, and Min Zhang (Eds.). ACM, 1371–1381. doi:10.1145/3805712.3809584
- [37] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* 55, 13s (2023), 295:1–295:42. doi:10.1145/3583558
- [38] Caio Nóbrega and Leandro Balby Marinho. 2019. Towards explaining recommendations through local surrogate models. In *ACM/SIGAPP '19*. ACM, 1671–1678.
- [39] Amir Hossein Ordibazar, Omar Hussain, and Morteza Saberi. 2021. A Recommender System and Risk Mitigation Strategy for Supply Chain Management Using the Counterfactual Explanation Algorithm. In *Service-Oriented Computing - ICSOC 2021 Workshops - AIOps, STRAPS, AI-PA and Satellite Events, Dubai, United Arab Emirates, November 22–25, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 13236)*, Hakim Hacid, Monther Aldwairi, Mohamed Reda Bouadjenek, Marinella Petrocchi, Noura Faci, Fatma Outay, Amin Beheshti, Lauritz Thamsen, and Hai Dong (Eds.). Springer, 103–116. doi:10.1007/978-3-031-14135-5_8
- [40] Andreas Peintner, Amir Reza Mohammadi, and Eva Zangerle. 2023. SPARE: Shortest Path Global Item Relations for Efficient Session-based Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023*. ACM, 58–69. doi:10.1145/3604915.3608768
- [41] Andreas Peintner, Amir Reza Mohammadi, and Eva Zangerle. 2025. Efficient Session-based Recommendation with Contrastive Graph-based Shortest Path Search. *ACM Trans. Recomm. Syst.* 3, 4, Article 46 (April 2025), 24 pages. doi:10.1145/3701764
- [42] Bastian Pfeifer, Arne Gevaert, Markus Loecher, and Andreas Holzinger. 2025. Tree smoothing: Post-hoc regularization of tree ensembles for interpretable machine learning. *Inf. Sci.* 690 (2025), 121564. doi:10.1016/J.INS.2024.121564
- [43] Niloofar Ranjbar, Saeedeh Momtazi, and MohammadMehdi Homayoonpour. 2024. Explaining recommendation system using counterfactual textual explanations. *Mach. Learn.* 113, 4 (2024).
- [44] Steffen Rendle, Walid Krichene, Li Zhang, and John R. Anderson. 2020. Neural Collaborative Filtering vs. Matrix Factorization Revisited. In *RecSys '20*. ACM, 240–248.
- [45] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016*. ACM, 1135–1144. doi:10.1145/2939672.2939778

- [46] Javier Del Ser, Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Francisco Herrera, Anna Saranti, and Andreas Holzinger. 2024. On generating trustworthy counterfactual explanations. *Inf. Sci.* 655 (2024), 119898. doi:10.1016/J.INS.2023.119898
- [47] Shahrzad Shashaani. 2024. Explainability in Music Recommender System. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*, Tommaso Di Noia, Pasquale Lops, Thorsten Joachims, Katrien Verbert, Pablo Castells, Zhenhua Dong, and Ben London (Eds.). ACM, 1395–1401. doi:10.1145/3640457.3688028
- [48] Ramni Harbir Singh, Sargam Maurya, Tanisha Tripathi, Tushar Narula, and Gaurav Srivastav. 2020. Movie recommendation system using cosine similarity and KNN. *International Journal of Engineering and Advanced Technology* 9, 5 (2020), 556–559.
- [49] Eulalia Szmidt and Janusz Kacprzyk. 2010. The Spearman rank correlation coefficient between intuitionistic fuzzy sets. In *5th IEEE International Conference on Intelligent Systems, IS 2010, 7-9 July 2010, University of Westminster, London, UK*. IEEE, 276–280. doi:10.1109/IS.2010.5548399
- [50] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. 2022. Learning and Evaluating Graph Neural Network Explanations based on Counterfactual and Factual Reasoning. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*. ACM, 1018–1027. doi:10.1145/3485447.3511948
- [51] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual Explainable Recommendation. In *CIKM '21*. ACM, 1784–1793.
- [52] Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. 2021. Counterfactual Explanations for Neural Recommenders. In *SIGIR '21*. ACM, 1627–1631.
- [53] Sahil Verma, Ashudeep Singh, Varich Boonsanong, John P. Dickerson, and Chirag Shah. 2023. RecRec: Algorithmic Recourse for Recommender Systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 4325–4329. doi:10.1145/3583780.3615181
- [54] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *CoRR* abs/1711.00399 (2017). arXiv:1711.00399 <http://arxiv.org/abs/1711.00399>
- [55] Arran Zeyu Wang, David Borland, and David Gotz. 2025. Beyond Correlation: Incorporating Counterfactual Guidance to Better Support Exploratory Visual Analysis. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 776–786. doi:10.1109/TVCG.2024.3456369
- [56] Xiangmeng Wang, Qian Li, Dianer Yu, Qing Li, and Guandong Xu. 2024. Counterfactual Explanation for Fairness in Recommendation. *ACM Trans. Inf. Syst.* 42, 4 (2024), 106:1–106:30. doi:10.1145/3643670
- [57] Xiangmeng Wang, Qian Li, Dianer Yu, Qing Li, and Guandong Xu. 2024. Reinforced Path Reasoning for Counterfactual Explainable Recommendation. *IEEE Trans. Knowl. Data Eng.* 36, 7 (2024), 3443–3459. doi:10.1109/TKDE.2024.3354077
- [58] Dianer Yu, Qian Li, Xiangmeng Wang, Qing Li, and Guandong Xu. 2024. Counterfactual Explainable Conversational Recommendation. *IEEE Trans. Knowl. Data Eng.* 36, 6 (2024), 2388–2400. doi:10.1109/TKDE.2023.3322403
- [59] Yi Yu, Kazunari Sugiyama, and Adam Jatowt. 2025. Domain Counterfactual Data Augmentation for Explainable Recommendation. *ACM Trans. Inf. Syst.* 43, 3, Article 58 (Feb. 2025), 30 pages. doi:10.1145/3711856
- [60] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 83–92.
- [61] Jinfeng Zhong and Elsa Negre. 2022. Shap-enhanced counterfactual explanations for recommendations. In *SAC '22*. ACM, 1365–1372.

Received 17 February 2026; revised 22 June 2026; accepted 31 July 2026