Refining Fidelity Metrics for Explainable Recommendations

Mikhail Baklanov
Tel Aviv University
Tel Aviv, Israel

Veronika Bogina
Tel Aviv University
Tel Aviv, Israel

Yehonatan Elisha
Tel Aviv University
Tel Aviv, Israel

Yahlly Schein
Tel Aviv University
Tel Aviv, Israel

Liron Allerhand
Tel Aviv University
Tel Aviv, Israel

Oren Barkan
The Open University of Israel
Tel Aviv, Israel

Noam Koenigstein
Tel Aviv University
Tel Aviv, Israel

Abstract

Counterfactual evaluation provides a promising framework for assessing explanation fidelity in recommender systems, but perturbation metrics adapted from computer vision suffer three key limitations: (1) they conflate explaining and contradictory features, (2) they average over entire user histories instead of prioritizing concise, high-impact explanations, and (3) they use fixed-percentage perturbations, leading to inconsistencies across users.

We introduce refined counterfactual metrics that focus on the most relevant explaining features, exclude contradictory elements, and assess fidelity at a fixed explanation length, ensuring a more consistent and interpretable evaluation. Our code is at: https://github.com/DeltaLabTLV/FidelityMetrics4XRec

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

Recommender Systems, Explanations, Counterfactual Evaluation

ACM Reference Format:
Mikhail Baklanov, Veronika Bogina, Yehonatan Elisha, Yahlly Schein, Liron Allerhand, Oren Barkan, and Noam Koenigstein. 2025. Refining Fidelity Metrics for Explainable Recommendations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730242

1 Introduction

Explanations enhance transparency in recommender systems, yet ensuring their **fidelity**—the degree to which they accurately reflect the model's decision-making process—remains challenging. Fidelity is crucial for building user trust, detecting biases, and debugging recommendation models [8, 12, 17, 18, 35]. However, most existing evaluations prioritize **user perception** over factual correctness [27, 34], so explanations may appear plausible while misrepresenting the model's actual reasoning [32].

A recent counterfactual framework [4, 19] introduced **perturbation-based fidelity metrics**, adapting saliency-map techniques from computer vision and NLP [2, 3, 9, 14, 29] to recommender systems by systematically removing user data and measuring changes in recommendation scores. While this approach moves toward correctness-based evaluation, it suffers three key drawbacks: *(1)* it considers **all** user data rather than focusing on concise explanations; *(2)* it does not differentiate between **supporting** vs. **contradictory** features; *(3)* it relies on fixed-percentage perturbation steps, which vary across users with different amounts of data.

To address these shortcomings, we propose refined counterfactual metrics that concentrate on the most influential K_e features, excluding contradictory features, and evaluating explanations at a fixed length. Furthermore, beyond theoretical improvements, our refined metrics align with real-world explainability needs by allowing practitioners to evaluate explanations at a fixed granularity, ensuring consistency and interpretability across different recommendation settings. This is particularly valuable for user-facing applications where explanation length must be controlled due to UI constraints and cognitive load.

2 Limitations of Perturbation Metrics in Fidelity Evaluation

To evaluate the fidelity of explanations in recommender systems, Barkan et al. [4] introduced a set of perturbation metrics inspired by heatmap evaluations in computer vision [29]. These metrics operate within a counterfactual framework where user data is gradually removed in fixed-percentage steps, and the resulting changes in the recommendation score or ranking of the explained item are tracked. The final AUC score is computed as the integral over all perturbation steps, quantifying the sensitivity of the recommendation to these removals. For a formal definition of these metrics, we refer the reader to [4].

While this approach provides an important step toward fidelity-aware evaluation, we identify several key limitations that make it unsuitable for assessing explanations in recommender systems.

2.1 Perturbation Metrics Do Not Ensure Concise Explanations

Unlike in computer vision, explanations in recommender systems must be concise, as users expect only a small subset of key features [1, 35]. Perturbation metrics, however, average over all

SIGIR '25, July 13–18, 2025, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730242

features, treating important and irrelevant elements equally. Since lengthy explanations overwhelm users [15], our refined metrics focus only on the top K_e most influential user features, aligning evaluation with real-world usability constraints.

2.2 Perturbation Metrics Do Not Differentiate Supporting vs. Contradictory Features

Another key limitation is that perturbation-based AUC fidelity evaluation treats all features equally, failing to differentiate between those that support a recommendation and those that actively suppress it. In reality, some features negatively impact the recommendation process, contradicting the intended explanation. For example, in implicit-feedback collaborative filtering [5–7, 11, 13, 23, 28, 30], a user who watches both horror and romance movies may receive a horror recommendation, while their history of romance movies acts as a suppressing factor, reducing the likelihood of receiving horror recommendations. Since perturbation metrics treat all user data equally, they conflate genuinely supporting features with suppressing ones, distorting fidelity measurements.

Empirical Evidence of Contradictory Elements. Figure 1a shows the AUC curve of the *POS-P*@20 metric from [4] for several explanation methods, applied to an implicit-feedback MF [10, 24] model using users' historical interactions. The lowest value occurs after masking approximately 70% of the most relevant user data. Beyond this point, removing additional data counterintuitively improves the recommended item's rank.

Intuitively, removing **all** user data should further degrade recommendation confidence. However, the figure reveals that masking the remaining data improves the item's rank. This indicates that some user-features act as *contradictory* elements—features that actively *suppress* the recommended item rather than explain it. Such features negatively impact the model's score and distort fidelity assessments.

A similar pattern appears in other perturbation-based metrics that remove explaining features, such as *NDCG-P* (Fig. 1c) and *DEL-P* (Fig. 1g) from [4]. By failing to exclude contradictory elements, these metrics misrepresent fidelity. Our refined metrics address this limitation by focusing only on the most explaining features, thus avoiding the contradictory elements.

2.3 Perturbation Metrics Depend on User Data Size

Perturbation-based metrics compute the AUC curve using steps of a fixed *percentage* of a user's history. Unlike images, where feature spaces are relatively uniform, user profiles in recommendation systems vary widely in size. A user with thousands of interactions loses much more information per step than a user with only a few interactions, leading to inconsistent evaluation granularity.

Moreover, while explanations must be concise, the optimal explanation length depends on the specific recommender system, its UI constraints, and users' cognitive capacity. Our refined metrics address this by evaluating fidelity at multiple fixed lengths, ensuring consistent assessment across users with varying data sizes. This flexibility allows system operators to tailor the maximum explanation length to their application needs, balancing transparency with usability.

3 Refined Fidelity Metrics

The limitations discussed in Sec. 2 highlight the need for *counterfactual metrics* that: (i) evaluate only the top K_e most relevant user features, (ii) exclude contradictory features, and (iii) avoid percentage-based perturbations in favor of fixed-length explanation steps. In this section, we propose a refined set of metrics designed to provide a more accurate and interpretable assessment of fidelity-aware explanations.

Let f be a recommender model that computes affinity scores for a given user u based on their personal data vector $\mathbf{x}_u$. Following Barkan et al. [4], we focus on the case of implicit-feedback collaborative filtering [23, 28, 30]. Accordingly, $\mathbf{x}_u \in \{0,1\}^{|\mathcal{V}|}$ is a binary vector indicating the historical items consumed by u (i.e., the user features).

The recommender assigns a vector of affinity scores $f(\mathbf{x}_u) \in \mathbb{R}^{|\mathcal{V}|}$, where each entry represents the predicted affinity of user u for an item in $\mathcal{V}$. Specifically, the affinity score assigned to item y is denoted by $f(\mathbf{x}_u)_y$.

The *rank* of item y in the recommendation list for user u is defined as:

$$\text{rank}_f^y(\mathbf{x}_u) = 1 + \sum_{i \in \mathcal{V} \setminus \{y\}} \mathbb{1}\left[f(\mathbf{x}_u)_i > f(\mathbf{x}_u)_y\right], \tag{1}$$

where $\mathbb{1}[\cdot]$ is the indicator function. A lower rank indicates a higher recommendation priority.

Counterfactual User Vectors. We define two complementary counterfactual user vectors to assess explanation fidelity:

- **Removed Explanations Vector.** A vector where the top K_e most explaining elements in $\mathbf{x}_u$ are removed (set to zero):
$$\mathbf{x}_u^{\setminus K_e} = \mathbf{x}_u \circ (\mathbf{1} - \mathbf{m}_{K_e}), \tag{2}$$
where $\mathbf{m}_{K_e}$ is a binary mask selecting the top K_e most explaining elements, $\mathbf{1}$ is an all-ones vector, and $\circ$ denotes the Hadamard (element-wise) product.
- **Retained Explanations Vector.** A vector that retains only the top K_e explaining elements and sets all other entries to zero:
$$\mathbf{x}_u^{K_e} = \mathbf{x}_u \circ \mathbf{m}_{K_e}. \tag{3}$$

3.1 Refined Metrics

Our refined metrics are designed to directly evaluate the fidelity of explanations by focusing only on the most explaining user features thus ignoring misleading contradictory features. Unlike perturbation-based methods, which aggregate across all features, our approach provides a finer-grained view of how explanations contribute to model decision-making.

We thus propose the following metrics to evaluate the fidelity of explanations in recommender systems:

1. Positive Perturbations at K_r and K_e (POS@K_r, K_e). This metric refines the *POS-P@K* metric from [4]. It measures whether the explained item drops out of the top K_r recommendations when the top K_e explaining features are removed:

$$POS@K_r, K_e = \mathbb{1}\left[\text{rank}_f^y(\mathbf{x}_u^{\setminus K_e}) \leq K_r\right]. \tag{4}$$

If $\text{rank}_f^y(\mathbf{x}_u^{\setminus K_e}) > K_r$, it implies that removing the top-K_e most crucial features has caused the explained item to drop below the top-K_r recommended items. Hence, **lower values** indicate **higher fidelity**.

2. Counterfactual Discounted Cumulative Gain at K_e (CDCG@K_e). This metric refines the *NDCG-P@K* metric from [4], and captures how severely the ranking of the recommended item degrades when critical features are excluded.

CDCG@K_e is defined as:

$$CDCG@K_e = \sum_{i=1}^{|\mathcal{V}|} \frac{\mathbb{1}\left[\text{rank}_f^y(\mathbf{x}_u^{\setminus K_e}) = i\right]}{\log_2(i+1)}, \tag{5}$$

Lower values indicate a stronger negative impact on the explained item's rank after removing top-K_e features, implying **higher fidelity**.

3. Insertion at K_e (INS@K_e). This metric refines *INS-P* from [4]. It evaluates how the recommender's confidence *increases* as the most important explaining features are gradually *added* to an initially empty vector, simulating how explanatory power is restored:

$$INS@K_e = \frac{f\left(\mathbf{x}_u^{K_e}\right)_y}{f\left(\mathbf{x}_u\right)_y}. \tag{6}$$

Higher values indicate that reintroducing the top-explaining features significantly boosts the recommeded item's score, highlighting **higher fidelity**.

4. Deletion at K_e (DEL@K_e). This metric refines *DEL-P* from [4]. It measures how the recommender's confidence *declines* when top-explaining features are removed, quantifying the reduction in recommendation strength due to feature removal:

$$DEL@K_e = \frac{f\left(\mathbf{x}_u^{\setminus K_e}\right)_y}{f\left(\mathbf{x}_u\right)_y}. \tag{7}$$

Lower values imply that removing critical features substantially weakens the explained item's score, demonstrating **higher fidelity**.

Finally, we chose not to refine *NEG-P@K* from [4], as negative perturbations primarily target the least explaining and often contradictory features, making them less relevant for fidelity assessment.

4 Empirical Evaluation and Insights

To assess the impact of our refined metrics, we compare them against the original perturbation metrics from Barkan et al. [2]. We conduct experiments on multiple implicit-feedback recommendation models—Matrix Factorization (MF) [24], Variational Autoencoders (VAE) [25], and Neural Collaborative Filtering (NCF) [21] — using datasets such as MovieLens 1M (ML1M) [20], Yahoo!Music [16], and Pinterest [22]. Due to space constraints, we report only ML1M results here; additional experiments on Yahoo!Music and Pinterest (as well as other recommender architectures) can be found in our Git repository.

We evaluate multiple explanation methods, including similarity-based approaches (Jaccard, Cosine), classical post-hoc methods (LIME [31], SHAP [26]), and counterfactual methods (ACCENT [33], LXR [4]).

Recognizing that users can process only a limited number of explanations, we demonstrate our evaluation metrics using the top-5 most explaining features ($1 \leq K_e \leq 5$). Since the optimal explanation length is application-specific, we leave its choice to system operators, while ensuring that our refined metrics faithfully assess fidelity across different K_e values.

4.1 Comparing Perturbation-Based and Refined Metrics

Our refined metrics offer a more structured and fine-grained evaluation of explanation fidelity.

First, although all metrics aim to capture fidelity, they are not fully correlated. For instance, Cosine outperforms LIME on *POS*@20, 5 (Fig. 1b), but LIME performs better on *CDCG*@5 (Fig. 1d). LXR is consistently superior on *INS*@K_e (Fig. 1f), yet ties with LIME on *DEL*@1 (Fig. 1h). These discrepancies, which are obscured by the original perturbation metrics, highlight how different explainers emphasize different aspects of fidelity.

Second, the refined metrics exhibit *monotonic behavior*, unlike perturbation metrics, which are prone to artifacts from contradictory features. By focusing only on the top-K_e explaining features, our metrics avoid relying on contradictory features, resulting in smoother, monotonic evaluation curves.

Third, our refined metrics expose finer-grained distinctions. In Figure 1b, LXR is not the top explainer for *POS*@20, 1 and *POS*@20, 2; its advantage appears only for $K_e \geq 3$. The original *POS-P*@20 metric (Fig. 1a) fails to capture this nuance, incorrectly suggesting that LXR is uniformly superior. This insight is crucial for applications that require very short explanations, such as when $K_e \leq 2$. Similar distinctions emerge for *CDCG*@1 and *DEL*@1. Our refined metrics thus help practitioners better match each explainer's strengths to platform constraints (e.g., short explanations for mobile versus richer explanations for expert users).

5 Implications and Final Remarks

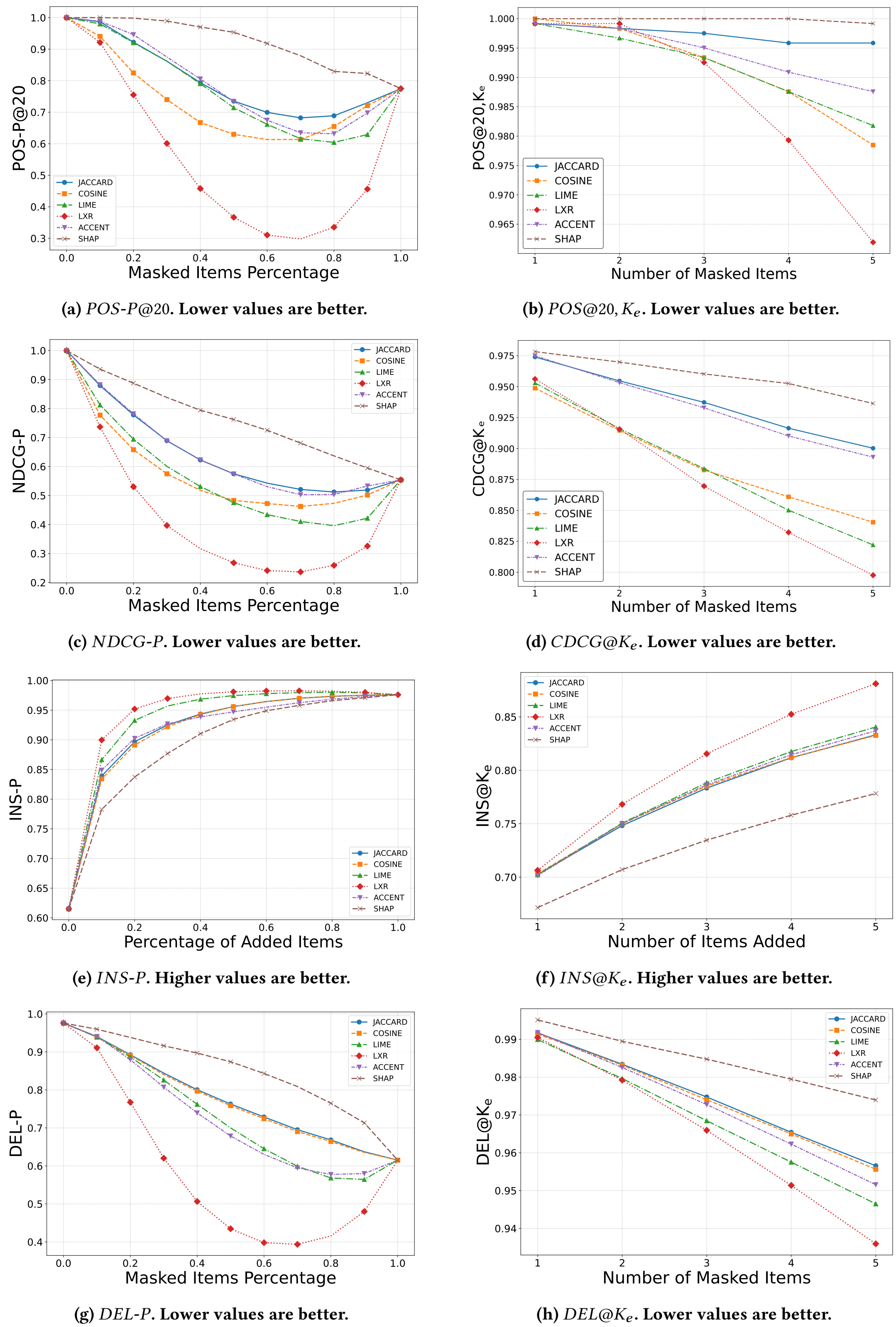
Our results highlight the limitations of perturbation-based fidelity evaluation and demonstrate the advantages of our refined approach. By focusing on the most relevant (supporting) features, we filter out contradictory elements, resulting in a more faithful assessment of explainability.

Unlike prior methods that average over entire user histories, our approach enables *targeted* fidelity evaluation at varying levels of granularity. This flexibility allows system operators to balance user-interface constraints with model transparency. For instance, if an application can display only a few explanations, operators may favor methods that excel at lower K_e values. Conversely, if a larger interface or expert users are targeted, methods that perform best at higher K_e can be selected.

We hope our refined metrics will foster more precise, practical evaluation of explainable recommendation systems and serve as a foundation for future research into fidelity-aware explanation methods.

6 Acknowledgement

This work was supported by the Israeli Science Foundation (ISF grant 2243/20) and by the Ministry of Innovation, Science & Technology, Israel.



(a) *POS-P@20*. **Lower values are better.**

(b) $POS@20, K_e$. **Lower values are better.**

(c) *NDCG-P*. **Lower values are better.**

(d) $CDCG@K_e$. **Lower values are better.**

(e) *INS-P*. **Higher values are better.**

(f) $INS@K_e$. **Higher values are better.**

(g) *DEL-P*. **Lower values are better.**

(h) $DEL@K_e$. **Lower values are better.**

Figure 1: Comparison of explanation fidelity metrics. The original perturbation metrics from Barkan et al. [4] (left) are contrasted with our refined metrics (right). Results are shown for an MF-based [24] recommender trained on the ML1M dataset [20].

References

[1] Mikhail Baklanov. 2024. CEERS: Counterfactual Evaluations of Explanations in Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1323–1329.

[2] Oren Barkan, Yuval Asher, Amit Eshel, Yehonatan Elisha, and Noam Koenigstein. 2023. Learning to explain: A model-agnostic framework for explaining black box models. *Proceedings of the IEEE International Conference on Data Mining* (2023), 944–949.

[3] Oren Barkan, Yuval Asher, Amit Eshel, Noam Koenigstein, et al. 2023. Visual explanations via iterated integrated attributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2073–2084.

[4] Oren Barkan, Veronika Bogina, Liya Gurevitch, Yuval Asher, and Noam Koenigstein. 2024. A Counterfactual Framework for Learning and Evaluating Explanations for Recommender Systems. *Proceedings of the ACM Web Conference* (2024), 3723–3733.

[5] Oren Barkan, Avi Caciularu, Ori Katz, and Noam Koenigstein. 2020. Attentive item2vec: Neural attentive user representations. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3377–3381.

[6] Oren Barkan, Avi Caciularu, Idan Rejwan, Ori Katz, Jonathan Weill, Itzik Malkiel, and Noam Koenigstein. 2020. Cold item recommendations via hierarchical item2vec. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 912–917.

[7] Oren Barkan, Avi Caciularu, Idan Rejwan, Ori Katz, Jonathan Weill, Itzik Malkiel, and Noam Koenigstein. 2021. Representation learning via variational bayesian networks. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 78–88.

[8] Oren Barkan, Yonatan Fuchs, Avi Caciularu, and Noam Koenigstein. 2020. Explainable recommendations via attentive multi-persona collaborative filtering. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 468–473.

[9] Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. 2021. Grad-sam: Explaining transformers via gradient self-attention maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2882–2887.

[10] Oren Barkan, Ori Katz, and Noam Koenigstein. 2020. Neural attentive multiview machines. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3357–3361.

[11] Oren Barkan, Idan Rejwan, Avi Caciularu, and Noam Koenigstein. 2020. Bayesian Hierarchical Words Representation Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3871–3877.

[12] Oren Barkan, Tom Shaked, Yonatan Fuchs, and Noam Koenigstein. 2024. Modeling users' heterogeneous taste with diversified attentive user profiles. *User Modeling and User-Adapted Interaction* 34, 2 (2024), 375–405.

[13] Shay Ben-Elazar, Gal Lavee, Noam Koenigstein, Oren Barkan, Hilik Berezin, Ulrich Paquet, and Tal Zaccai. 2017. Groove radio: A bayesian hierarchical model for personalized playlist generation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 445–453.

[14] Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 782–791.

[15] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2019), 765–774.

[16] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. 2012. The yahoo! music dataset and kdd-cup'11. In *Proceedings of KDD Cup 2011*. PMLR, 3–18.

[17] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 69–78.

[18] Keren Gaiger, Oren Barkan, Shir Tsipory-Samuel, and Noam Koenigstein. 2023. Not all memories created equal: Dynamic user representations for collaborative filtering. *Ieee Access* 11 (2023), 34746–34763.

[19] Liya Gurevitch, Veronika Bogina, Oren Barkan, Yahlly Schein, Yehonatan Elisha, and Noam Koenigstein. 2025. LXR: Learning to eXplain Recommendations. *ACM Transactions on Recommender Systems* 1, 1 (2025). https://doi.org/10.1145/3732292

[20] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[21] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. International World Wide Web Conferences Steering Committee, 173–182.

[22] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. *Proceedings of the 26th International Conference on World Wide Web* (2017), 173–182.

[23] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*. Ieee, 263–272.

[24] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[25] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*. 689–698.

[26] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[27] Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. 2016. A live-user study of opinionated explanations for recommender systems. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 256–260.

[28] Maryam Khanian Najafabadi and Mohd Naz'ri Mahrin. 2016. A systematic literature review on the state of research and practice of collaborative filtering technique and implicit feedback. *Artificial intelligence review* 45, 2 (2016), 167–201.

[29] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421* (2018).

[30] Steffen Rendle. 2021. Item recommendation from implicit feedback. In *Recommender Systems Handbook*. Springer, 143–171.

[31] Marco Tulio 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*. 1135–1144.

[32] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.

[33] Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. 2021. Counterfactual explanations for neural recommenders. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1627–1631.

[34] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable recommendation via multi-task learning in opinionated text data. *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (2018), 285–294.

[35] Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval* 14, 1 (2020), 1–101.