JEECS (Journal of Electrical Engineering and Computer Sciences)
Vol. 11, No. 1, June 2026, pp. 9-22
e-ISSN: 2579-5392 p-ISSN: 2528-0260

Review-Grounded Explainable Recommendation with Faithfulness Evaluation on Amazon Reviews

Xiaohan Chang [1], Yifei Lu [2], Ziliang Samuel Zhong [3]

[1]Computer Science, University of Connecticut, 352 Mansfield Rd, Storrs, CT 06269, USA

[2]Computer Science, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA

[3]New York University, New York, NY 10012, USA

Article Info

Article history:
Received: 12 January 2026
Revised: 27 March 2026
Accepted: 15 April 2026

Keyword:
Amazon reviews
Evidence extraction
Explainable recommendation
Faithfulness evaluation
Review-grounded justification

Abstract

Review text can support explainable recommendations, but many recommender systems still optimize ranking accuracy without providing verifiable textual evidence, or they attach post-hoc explanations whose faithfulness to the model is unclear. This study addresses the lack of a reproducible evaluation setting that jointly measures recommendation quality and whether extracted review evidence actually supports model scoring. We propose Review-Grounded eXplainable Recommender (RGXRec), a lightweight hybrid method that combines interaction signals and TF-IDF review similarity, and we evaluate it on the Luxury Beauty and Video Games subsets of the Amazon Review Data. The pipeline includes rating thresholding, iterative 5-core pruning, chronological leave-one-out splitting, ranked recommendation, extractive evidence generation, and faithfulness evaluation. We compare RGXRec with popularity, metadata-graph KNN, SVD-MF, and ReviewSim using NDCG@K, Recall@K, MRR, evidence coverage, ROUGE-1, sentiment agreement, and a term-attribution faithfulness score. On Luxury Beauty, RGXRec achieves the best ranking performance, reaching NDCG@10 of 0.3606 and outperforming the strongest single-view baseline. On Video Games, collaborative and metadata signals remain stronger for ranking, but RGXRec preserves competitive accuracy while providing non-zero review-grounded faithfulness that interaction-only baselines cannot offer. These findings show that review-grounded recommendation should be evaluated on both ranking quality and explanation faithfulness.

Corresponding author:
Xiaohan Chang, xhchang06@yahoo.com

DOI: https://doi.org/10.54732/jeecs.v11i1.2

This is an open access article under the CC–BY license.

1. Introduction

Recommender systems are central to modern e-commerce, yet accurate ranking alone is not sufficient when users and stakeholders also need to understand why an item is recommended. In review-rich platforms, product reviews provide fine-grained preference cues and can therefore serve as a natural source of evidence for recommendation justification [1], [2], [3]. Most practical recommenders, however, still focus primarily on predictive effectiveness through interaction-based latent factors [4], neighborhood signals [5], or sparse text representations [6], [7]. These methods are effective for top-N recommendation and are commonly evaluated with ranking metrics such as NDCG, Recall, and MRR [8]. However, a strong ranking result does not guarantee that an attached explanation is faithful to the actual scoring process.

Explainable recommendation research has addressed this issue from multiple directions. Survey work has summarized the main goals of transparency, persuasiveness, and scrutability in explainable recommendation [9]. More recently, model-agnostic work has proposed reusable pipelines for faithfully

explaining recommendation rankings [10]. Review analysis resources such as sentiment lexicons [11] and review-grounded models such as HFT, DeepCoNN, and NARRE have further shown that review text can support both preference modeling and human-readable recommendation justification [12], [13], [14]. A recent review-based recommender for rating prediction also reported that review properties and sequential review information can improve both accuracy and interpretability [15]. In addition, sampled top-N evaluation remains a standard benchmarking protocol for implicit recommendation [16], while recent surveys have emphasized the importance of evaluating the "why" of recommendations in a systematic manner [17].

Recent studies continue to strengthen this line of work. Counterfactual explanation has been investigated for fairness diagnosis in recommendation [18]. Contrastive learning has been used to improve the faithfulness and factuality of generated explanations [19]. At a broader level, recent surveys on review-based recommender systems and review-based explainable recommendation highlight persistent challenges in representation learning, sparsity, evaluation, and transparency [20], [21]. In parallel, review-enhanced graph neural models such as IReGNN attempt to reduce review sparsity while preserving explainability [22]. However, two research gaps remain. First, many existing approaches either rely on complex generation architectures or evaluate explanations mainly through plausibility-oriented measures, making it difficult to verify whether the highlighted evidence truly accounts for the model score. Second, fully reproducible pipelines that jointly report ranking performance, overlap with held-out user reviews, and model-level faithfulness are still limited. Faithfulness research has repeatedly shown that explanations should be evaluated by whether removing or isolating the highlighted evidence changes model behavior, rather than by readability alone [23], [24].

To address these gaps, this paper proposes Review-Grounded eXplainable Recommender (RGXRec), a lightweight hybrid model that combines collaborative latent factors and TF-IDF review similarity. The method produces a ranked list together with extractive evidence consisting of shared aspect terms and supporting review sentences. In addition, we introduce a lightweight term-attribution faithfulness metric for the TF-IDF component and conduct an accuracy-faithfulness sensitivity analysis with respect to the hybrid weight. The study is implemented in an end-to-end pipeline on two Amazon Review Data subsets. After rating thresholding, 5-core pruning, and chronological leave-one-out splitting, we compare RGXRec against popularity, metadata-graph KNN, SVD-MF, and ReviewSim. The paper reports ranking metrics, explanation-quality metrics, and model-level faithfulness metrics, followed by failure analysis and a concise discussion of what the results imply for review-grounded explainable recommendation.

2. Research Methodology

2.1 End-to-end experimental pipeline

Figure 1 summarizes the complete workflow adopted in this study. The pipeline starts from Amazon review and metadata collection, followed by interaction binarization, iterative 5-core pruning, chronological train/validation/test splitting, review document construction, model training, ranked recommendation, extractive evidence generation, and joint evaluation of ranking accuracy and explanation faithfulness. This dedicated subsection is included to clarify how recommendation and explanation stages are connected in a single reproducible end-to-end setting.

2.2 Datasets, preprocessing, and descriptive statistics

Datasets and preprocessing. We use the Amazon Review Data (2018 release) with review text and item metadata including the also-bought/also-viewed edges [1], [2], [3]. We select two categories that differ in size and domain: Luxury Beauty (dense niche catalog) and Video Games (large entertainment catalog). Each raw review record provides a user identifier, an item identifier (ASIN), a timestamp, a numerical rating, and free-text review fields. Each metadata record provides the item title and the also_buy/also_view lists. To align the task with top-N implicit recommendation, we treat a review as a positive interaction when its rating is overall ≥ 4, which is a common binarization for Amazon data in the recommender literature. We then apply iterative 5-core pruning (users and items must have at least five remaining positive interactions) to guarantee enough history for leave-one-out evaluation. Table 1 reports the resulting dataset sizes. Table 2 reports the metadata graph statistics after restricting edges to items that remain in the recommendation set.

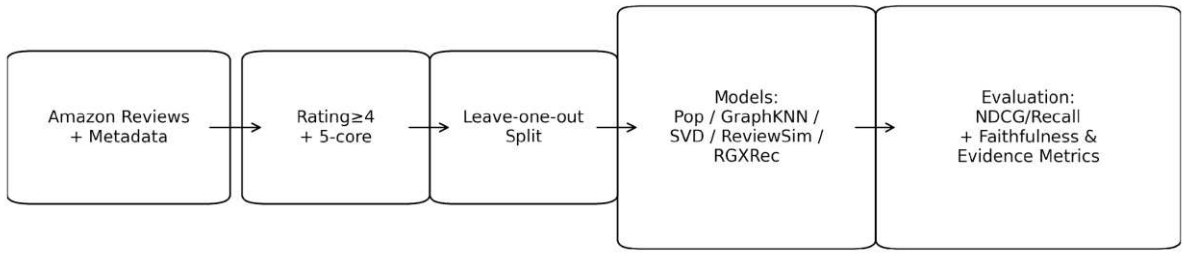

Figure 1. End-to-end experimental pipeline. The workflow consists of data acquisition, interaction binarization, 5-core pruning, chronological splitting, model training, ranked recommendation, extractive evidence generation, and explanation evaluation

Table 1. Dataset statistics after rating thresholding (overall ≥ 4) and iterative 5-core pruning

	Users	Items	Interactions	Density	AvgInter/User	AvgInter/Item	AvgReview Len
Luxury_Beauty	2382.0000	1047.0000	21911.0000	0.0088	9.1986	20.9274	96.2541
Video_Games	36869.0000	12851.0000	321753.0000	0.0007	8.7269	25.0372	115.2348

Table 2. Product metadata graph statistics (also-buy/also-view) restricted to items in the recommendation set

	Also Buy Edges	Also View Edges	Total Edges	Avg Degree
Luxury_Beauty	6157.000	4147.000	6992.000	13.356
Video_Games	248991.000	175179.000	330743.000	51.474

Table 3. Leave-one-out split and sampled evaluation protocol (99 negatives per user)

	Train Interactions	Val Interactions	Test Interactions	Negatives Per User	Evaluation Candidates Per User
Luxury_Beauty	17147	2382	2382	99	100
Video_Games	248015	36869	36869	99	100

Within each dataset, we perform a chronological leave-one-out split per user: the most recent positive interaction is assigned to the test set, the second-most recent to the validation set, and all earlier positives to the training set. This protocol is widely used for implicit top-N evaluation and simulates forecasting the next item from a user's historical behavior [16]. Table 3 summarizes the split sizes and evaluation candidate construction.

2.3. Task formulation.

Let U be the set of users and I the set of items. The training data consists of implicit positive interactions (u, i) derived from reviews with overall ≥ 4. For each user u, the goal is to rank candidate items so that the held-out test item i* appears near the top. We report NDCG@K and Recall@K for K ∈ {5,10,20} as standard ranking metrics [8], as well as mean reciprocal rank (MRR).

2.4. Compared recommenders.

We compare five models that cover common signals available in the Amazon data:

1) Pop: ranks items by training-set popularity counts.
2) MetaGraphKNN: uses the also-bought/also-viewed product graph. For a user u with training history H(u), we score a candidate item i by the number of graph neighbors of items in H(u) that connect to i (a simple neighborhood counting heuristic inspired by item-based collaborative filtering [5])
3) SVD-MF: an interaction-only latent factor model. We build a sparse user–item matrix X where X(u,i)=1 for training interactions and apply TruncatedSVD to obtain user factors P and item factors Q. The collaborative score is defined as:

$$s_{cf}(u,i) = (P_u, Q_i) \quad (1)$$

4) ReviewSim: a review-only model. We build a user document d_u by concatenating sentences from the user's training reviews, and an item document d_i by concatenating sentences from the item's training reviews. We compute TF–IDF vectors t_u and t_i. The text-based score is

$$s_{text}(u,i) = cos(t_u, t_i) \quad (2)$$

Table 4. Overview of compared recommenders, the signals they use, and whether model-level textual faithfulness is defined

	Uses Interactions	Uses Reviews	Uses Metadata Graph	Personalized Explanation	Faithfulness Evaluated
Pop	Yes	No	No	No	N/A
MetaGraphKNN	Yes	No (only for post-hoc)	Yes	Partially	N/A
SVD-MF	Yes	No	No	No (post-hoc)	0 by definition
ReviewSim	No	Yes	No	Yes	Yes
RGXRec	Yes	Yes	Optional (not used in this implementation)	Yes	Yes (scaled by 1-α)

5) RGXRec (proposed): a hybrid model that combines collaborative and textual evidence. The final ranking score is

$$s(u,i) = \propto s_{cf}(u,i) + (1-\propto)s_{text}(u,i) \quad (3)$$

Unless otherwise stated, α=0.75. This value is selected from the accuracy–faithfulness sensitivity analysis reported in Table 11 and Figures 8–9. Table 4 summarizes which signals each model consumes and whether it can be evaluated for model-level textual faithfulness.

2.5. Review-grounded explanation generation.

For each recommended item i for user u, we generate an extractive explanation consisting of (i) a short list of shared aspect terms and (ii) supporting evidence sentences. The aspect terms are selected as follows: we take the top TF–IDF terms from the user vector t_u and item vector t_i and select up to m=5 shared terms that maximize the summed TF–IDF weight. When the intersection is empty, we fall back to the item's top terms. Given the aspect terms, we retrieve evidence sentences from (a) the user's training reviews and (b) the item's training reviews by selecting the sentences that contain the largest number of aspect terms (ties are broken in favor of shorter sentences). We use 1 user sentence and 2 item sentences. Figure 2 illustrates the overall RGXRec architecture and where explanations are computed.

2.6. Explanation quality metrics.

We evaluate explanations using the held-out test review text as a reference because it provides a user-written description of why the user liked the item (the test interaction itself is a positive review). We compute:

1) Evidence Coverage: the fraction of top-K terms (K=20) from the held-out review that appear in the extracted evidence text (after the same tokenization/stopwording as TF–IDF).
2) ROUGE-1 F1: unigram overlap F1 between the evidence text and the held-out review.
3) Sentiment Agreement: whether the evidence text and held-out review have the same polarity sign, computed from the Hu–Liu opinion lexicon [11].

2.7 Faithfulness metric and hybrid-weight sensitivity analysis.

Model faithfulness asks whether the explanation evidence corresponds to the features actually used by the recommender. For ReviewSim, the TF–IDF cosine score decomposes additively across terms:

$$s_{text}(u,i) = \sum_{t \in V} t_u(t) \cdot t_i(t) \quad (4)$$

For the extracted aspect set A(u,i), the attributed score mass is defined as

$$mass(u,i) = \sum_{t \in A(u,i)} t_u(t) \cdot t_i(t) \quad (5)$$

The model-level textual faithfulness ratio is then

$$Faithfulness(u,i) = mass(u,i)/s_{text}(u,i) \quad (6)$$

For RGXRec, only the textual component contributes to textual faithfulness; therefore, the model-level score is

$$Faithfulness_{\text{RGXRec}}(u,i) = (1-\propto) \cdot mass(u,i)/s_{text}(u,i) \quad (7)$$

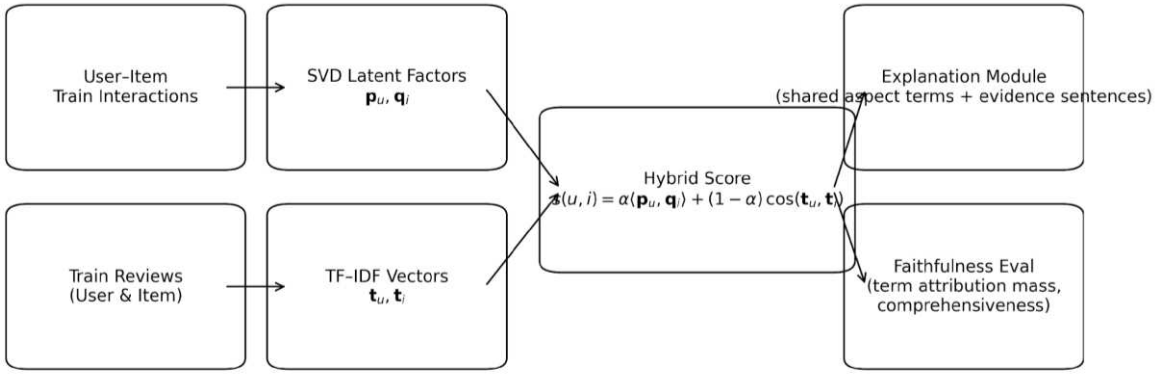

Figure 2. Architecture of RGXRec. Collaborative latent factors and review similarity are fused to produce the final ranking score, after which shared aspect terms, evidence sentences, and textual faithfulness are computed

Table 5. Hyperparameters and deterministic settings used in all experiments

No.	Component	Setting
0	Rating threshold	overall ≥ 4 (implicit positive interactions)
1	K-core	Iterative 5-core on users and items after thresholding
2	Split	Leave-one-out by timestamp: last=Test, second-last=Val, rest=Train
3	Evaluation	Sampled ranking with 99 negatives per user; report NDCG@{5,10,20}, Recall@{5,10,20}, MRR
4	TF-IDF (LB)	stop_words=english, max_features=15000, min_df=2
5	TF-IDF (VG)	stop_words=english, max_features=20000, min_df=5
6	Doc construction (LB)	max_reviews_per_user=30, max_reviews_per_item=30, max_sentences=50
7	Doc construction (VG)	max_reviews_per_user=20, max_reviews_per_item=20, max_sentences=50
8	SVD	TruncatedSVD: n_components=64, n_iter=3, random_state=2026
9	RGXRec α	0.75 (selected to balance accuracy and faithfulness)
10	Aspects per explanation	m=5 shared terms (user–item overlap)
11	Evidence sentences	1 user sentence + 2 item sentences containing aspects
12	Seed	2026 (all sampling operations)

This term-attribution metric is lightweight because it does not require retraining or perturbation loops. To make the accuracy–faithfulness trade-off explicit, we further vary α in {0, 0.25, 0.5, 0.75, 1.0} and report the corresponding results in Table 11 and Figures 8–9. We also report Comprehensiveness Drop as the average mass(u,i), which equals the drop in s_text when removing all aspect terms. Finally, we report a Sufficiency Similarity computed using only the aspect terms, which measures whether the extracted aspects alone form a coherent similarity signal. For RGXRec, only the (1−α) portion of the score depends on text, so the model-level faithfulness and comprehensiveness drop are scaled by (1−α). For Pop, MetaGraphKNN, and SVD-MF, the input does not include text, so textual faithfulness is 0 by definition; any review-based explanation attached to these models is post-hoc and does not change their scoring.

2.8 Implementation details.

Table 5 lists the hyperparameters and deterministic settings used in all experiments, including the random seed for negative sampling and sentence subsampling. All reported metrics are computed on the full test sets with the fixed evaluation protocol.

Because full-catalog ranking is computationally expensive on large item sets, we use sampled evaluation with a fixed number of negatives per user, which is standard in recommender benchmarking [16]. For each user u, we construct a candidate set C(u) consisting of the held-out positive test item and 99 negatives sampled uniformly from items the user never interacted with in train/validation/test. We use a single random seed (2026) to ensure that the same candidate sets are used across all methods. Each recommender produces a score for every candidate in C(u), and candidates are ranked by descending score. The position of the held-out item yields MRR and contributes to NDCG@K and Recall@K.

Formally, let rank_u denote the 1-indexed position of the held-out item for user u. The evaluation metrics are written separately from the paragraph text as follows:

$$MRR = \frac{1}{|U_{test}|}\sum_{u}\frac{1}{rank_u} \quad (8)$$

$$Recall@K = \frac{1}{|U_{test}|}\sum_{u}1[rank_u \le K] \quad (9)$$

$$NDCG@K = \frac{1}{|U_{test}|}\sum_{u}1[rank_u \le K]/log2(rank_u + 1) \quad (10)$$

Because each user has exactly one positive test item in the sampled candidate set, Recall@K is equivalent to hit rate in this setup.

We construct user and item documents from training reviews only. We first split each review into sentences using punctuation-based rules, then concatenate sentences up to a maximum of 50 sentences per document. To limit compute and avoid domination by prolific users or popular items, we subsample at most 30 reviews per user and 30 reviews per item in Luxury Beauty, and at most 20 reviews per user and 20 reviews per item in Video Games. This design yields a compact representation that is sufficient for TF–IDF similarity while keeping the pipeline scalable.

We compute TF–IDF using standard inverse document frequency weighting and L2 normalization, which makes cosine similarity equivalent to a dot product [7]. We remove English stopwords and cap the vocabulary size to 15,000 (Luxury Beauty) or 20,000 (Video Games). We also apply a minimum document frequency threshold (min_df) to remove extremely rare tokens that would not generalize across users/items. These choices are reported in Table 5 and are held fixed across all experiments. Because the representation is sparse and non-negative, the term-level contribution to cosine similarity is always non-negative, enabling the exact faithfulness decomposition used in Section F.

The evidence selection step is deterministic given the aspect term set. For each candidate evidence sentence, we compute a match score equal to the number of aspect terms appearing as substrings (case-insensitive). We select the top-scoring sentence(s), breaking ties by preferring shorter sentences to improve readability. This produces an extractive explanation that is easy to audit: every aspect term in the explanation can be located in the cited sentences. Importantly, for ReviewSim and RGXRec, the aspect terms are derived directly from the TF–IDF vectors used to compute s_text, which is why faithfulness can be evaluated exactly.

To measure whether evidence is consistent with the user's held-out review, we estimate sentiment polarity from a fixed opinion lexicon. We use the positive and negative word lists distributed with the Hu–Liu sentiment lexicon and compute polarity as (|pos|−|neg|)/|tokens| [11]. Sentiment agreement is the fraction of test cases where the evidence polarity and held-out review polarity have the same sign. Although this lexicon-based score is coarse, it is fully reproducible and does not require additional supervised training.

3. Results and Discussions

3.1 Results

3.1.1. Recommendation accuracy

Tables 6 and 7 report the sampled top-N ranking results on the test sets. Overall, the strongest baseline differs by domain. In Luxury Beauty, ReviewSim performs strongly, indicating that review language captures stable product attributes (e.g., scent, texture, skin sensitivity) that generalize from training history to the next purchase. In Video Games, MetaGraphKNN and SVD-MF dominate ReviewSim, which suggests that behavioral co-occurrence and latent collaborative structure are more informative than sparse textual overlap for predicting the next game.

RGXRec improves the overall ranking on Luxury Beauty. Specifically, RGXRec reaches NDCG@10 of 0.3606, exceeding both ReviewSim (0.2962) and SVD-MF (0.2766). On Video Games, RGXRec yields NDCG@10 of 0.2341, which is below MetaGraphKNN and SVD-MF but above the popularity baseline, showing that incorporating review evidence does not collapse ranking performance even when the text-only signal is weak. Figure 3 and Figure 4 visualize NDCG@10 across methods. The results highlight that a single model family is not universally best across categories; hybridization can be beneficial when text and interactions provide complementary information.

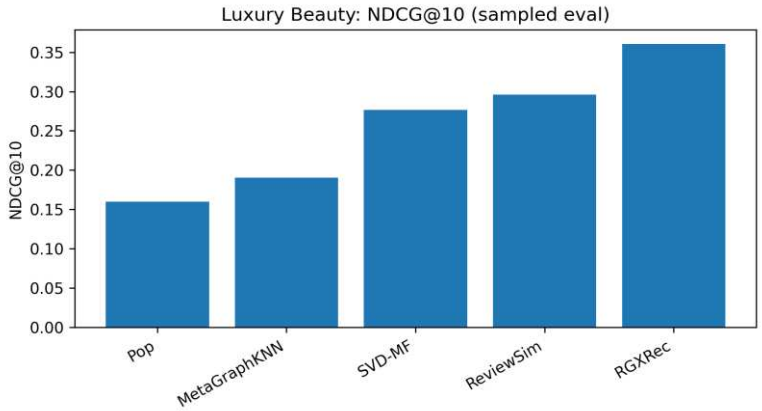

Figure 3. Luxury Beauty: NDCG@10 for all methods (sampled evaluation)

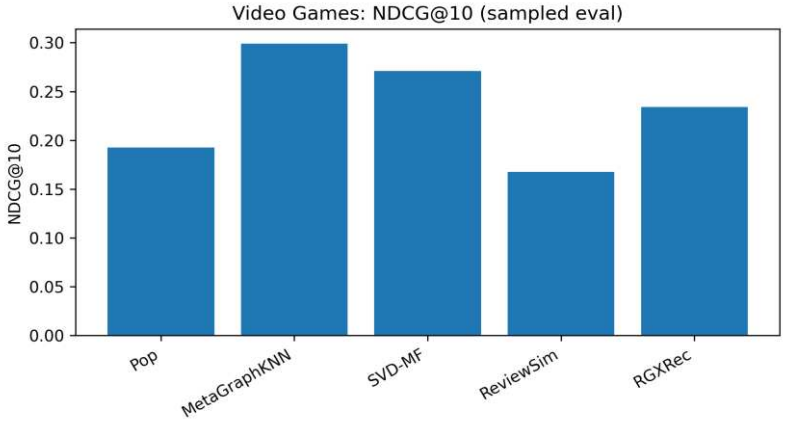

Figure 4. Video Games: NDCG@10 for all methods (sampled evaluation)

Table 6. Recommendation performance on Luxury Beauty (test set; 99 negative samples per user)

	NDCG@5	NDCG@10	NDCG@20	Recall@5	Recall@10	Recall@20	MRR
Pop	0.1334	0.1600	0.1976	0.1923	0.2758	0.4232	0.1452
MetaGraph KNN	0.1702	0.1902	0.2110	0.2427	0.3052	0.3879	0.1721
SVD-MF	0.2606	0.2766	0.2947	0.3006	0.3501	0.4228	0.2693
ReviewSim	0.2721	0.2962	0.3207	0.3577	0.4324	0.5306	0.2699
RGXRec	0.3394	0.3606	0.3811	0.4135	0.4790	0.5600	0.3381

Table 7. Recommendation performance on Video Games (test set; 99 negative samples per user)

	NDCG@5	NDCG@10	NDCG@20	Recall@5	Recall@10	Recall@20	MRR
Pop	0.1534	0.1926	0.2299	0.2370	0.3585	0.5067	0.1633
MetaGraph KNN	0.2823	0.2991	0.3157	0.3662	0.4176	0.4840	0.2760
SVD-MF	0.2354	0.2711	0.3046	0.3266	0.4372	0.5697	0.2387
ReviewSim	0.1380	0.1675	0.2011	0.1999	0.2917	0.4258	0.1514
RGXRec	0.2021	0.2341	0.2662	0.2781	0.3772	0.5051	0.2099

3.1.2. Explanation quality and faithfulness

Tables 8 and 9 report explanation metrics. Evidence Coverage and ROUGE-1 evaluate alignment to the held-out review, while Faithfulness and Comprehensiveness Drop quantify model-level dependence on the extracted evidence. In both datasets, personalized extractive explanations (SVD-MF post-hoc, ReviewSim, RGXRec) achieve higher coverage and ROUGE than item-only explanations (Pop and MetaGraphKNN post-hoc). This reflects that including a user evidence sentence tends to introduce preference words that also appear in the user's held-out review.

Faithfulness differentiates models that truly use review text from those that do not. ReviewSim achieves average faithfulness of 0.3715 on Luxury Beauty and 0.2993 on Video Games, meaning that the extracted aspect terms capture a substantial fraction of the TF–IDF similarity mass. RGXRec scales this faithfulness by (1−α)=0.25 due to the hybrid scoring, yielding faithfulness of 0.0929 (Luxury Beauty) and 0.0748 (Video Games). Interaction-only models have zero textual faithfulness by definition because their inputs do not include review text. Figure 5–7 visualize coverage and faithfulness. These findings illustrate a key practical point: a post-hoc review quote can look relevant (high coverage) even when it is not causally connected to the recommender's score. Therefore, reporting both coverage-style metrics and faithfulness metrics is necessary for evidence-based explainable recommendation [17], [18], [23], [24].

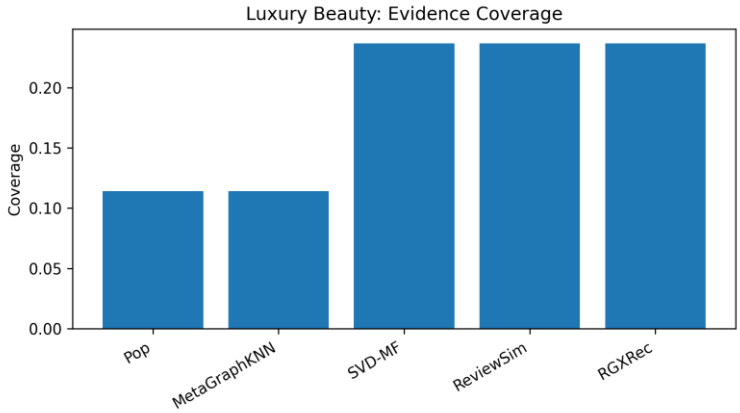

Figure 5. Luxury Beauty: evidence coverage across explanation strategies

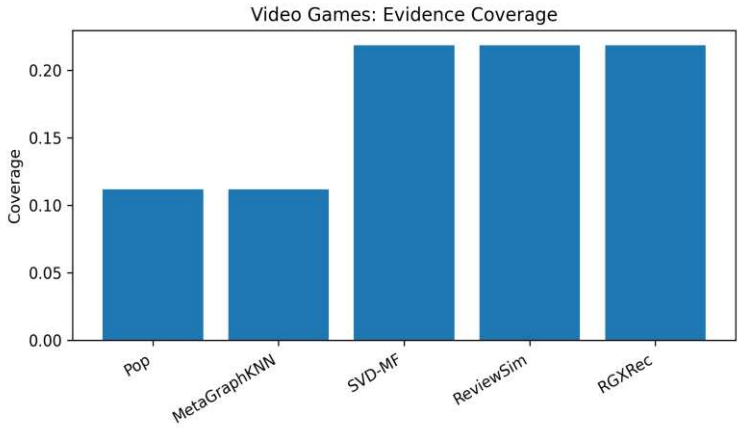

Figure 6. Video Games: evidence coverage across explanation strategies

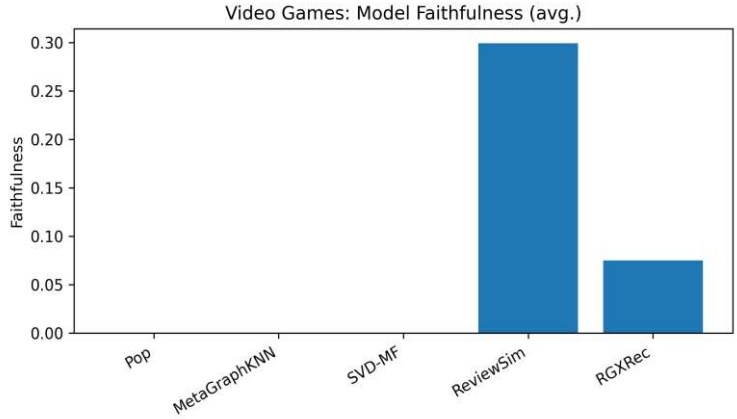

Figure 7. Video Games: average model faithfulness (textual) across methods

Table 8. Explanation quality and faithfulness on Luxury Beauty (test set)

	Coverage	ROUGE1-F1	SentAgree	Faithfulness	CompDrop	SuffSim
Pop	0.1143	0.1553	0.8799	0.0000	0.0000	0.0000
MetaGraphKNN	0.1143	0.1553	0.8799	0.0000	0.0000	0.0000
SVD-MF	0.2369	0.2347	0.8984	0.0000	0.0000	0.6362
ReviewSim	0.2369	0.2347	0.8984	0.3715	0.0690	0.6362
RGXRec	0.2369	0.2347	0.8984	0.0929	0.0172	0.6362

Table 9. Explanation quality and faithfulness on Video Games (test set)

	Coverage	ROUGE1-F1	SentAgree	Faithfulness	CompDrop	SuffSim
Pop	0.1119	0.1171	0.8267	0.0000	0.0000	0.0000
MetaGraphKNN	0.1119	0.1171	0.8267	0.0000	0.0000	0.0000
SVD-MF	0.2186	0.1507	0.8908	0.0000	0.0000	0.5755
ReviewSim	0.2186	0.1507	0.8908	0.2993	0.0348	0.5755
RGXRec	0.2186	0.1507	0.8908	0.0748	0.0087	0.5755

3.1.3. Ablation and hyperparameter sensitivity

Table 10 isolates the contribution of collaborative signals and review signals by comparing SVD-MF (α=1), ReviewSim (α=0), and the hybrid RGXRec (α=0.75). In Luxury Beauty, the hybrid improves both NDCG@10 and Recall@10, indicating complementary information between interactions and review language. In Video Games, the collaborative component dominates, but the hybrid retains competitive accuracy while enabling review-grounded evidence and non-zero faithfulness. To make the accuracy–faithfulness trade-off explicit, Table 11 and Figure 8–9 report performance as α varies. As α increases, ranking accuracy approaches the interaction-only model while textual faithfulness decreases linearly because less weight is placed on the review similarity component. We select α=0.75 as a balanced operating point: it is near-optimal for Luxury Beauty and yields substantially more faithfulness than α=1 on Video Games with a moderate accuracy cost.

Table 10. Ablation of collaborative and review components (test sets)

	NDCG@10	Recall@10	MRR
Luxury_Beauty:ReviewSim	0.2962	0.4324	0.2699
Luxury_Beauty:SVD-MF	0.2766	0.3501	0.2693
Luxury_Beauty:RGXRec	0.3606	0.4790	0.3381
Video_Games:ReviewSim	0.1675	0.2917	0.1514
Video_Games:SVD-MF	0.2711	0.4372	0.2387
Video_Games:RGXRec	0.2341	0.3772	0.2099

Table 11. Accuracy–faithfulness trade-off as a function of α. Video Games results are computed on a fixed 5,000-user subset for efficiency, whereas Luxury Beauty uses the full test set

	NDCG@10	Recall@10	Faithfulness_model
('Luxury_Beauty', 0.0)	0.2962	0.4324	0.3715
('Luxury_Beauty', 0.25)	0.3449	0.4626	0.2786
('Luxury_Beauty', 0.5)	0.3588	0.4769	0.1857
('Luxury_Beauty', 0.75)	0.3606	0.4790	0.0929
('Luxury_Beauty', 1.0)	0.2766	0.3501	0.0000
('Video_Games(subset)', 0.0)	0.1677	0.2874	0.2993
('Video_Games(subset)', 0.25)	0.1839	0.3102	0.2245
('Video_Games(subset)', 0.5)	0.2054	0.3352	0.1497
('Video_Games(subset)', 0.75)	0.2375	0.3790	0.0748
('Video_Games(subset)', 1.0)	0.2725	0.4368	0.0000

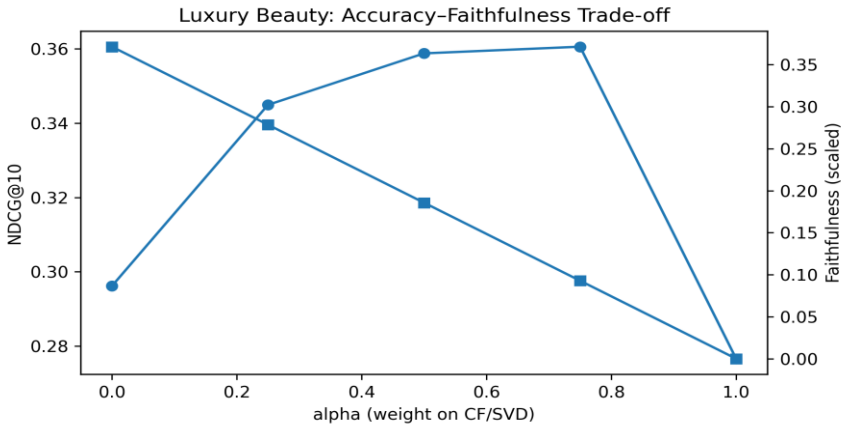

Figure 8. Luxury Beauty: NDCG@10 and scaled faithfulness versus α

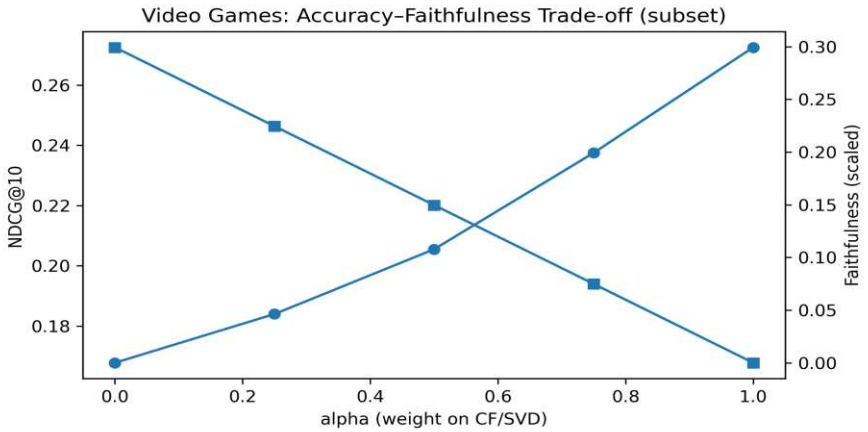

Figure 9. Video Games: NDCG@10 and scaled faithfulness versus α (subset)

3.1.4. Failure analysis, qualitative examples, and runtime.

To understand where review-grounded explanations struggle, we categorize RGXRec test cases into four groups: (i) NoSharedAspects, where the user's TF–IDF top terms and the item's TF–IDF top terms do not overlap (so the extracted aspects fall back to item-only terms); (ii) LowCoverage (<0.1), where the extracted evidence covers few of the held-out review's top terms; (iii) LowFaith (<0.2), where the extracted aspects capture little of the text similarity mass; and (iv) Other, which includes well-covered and faithful cases. Figure 10 shows the failure type distribution for Video Games.

The dominant failure mode in Video Games is NoSharedAspects. This is consistent with the weak performance of the text-only ReviewSim baseline: many users and games do not share distinctive vocabulary in short review snippets, so extractive overlap explanations become less personalized. In contrast, Luxury Beauty contains more standardized attribute language (e.g., "smell", "moisturizing", "sensitive"), which produces stronger overlap and higher coverage. This observation suggests that domain-specific review style affects both ranking and explanation. Figure 11 shows a high-coverage and high-faithfulness example where the extracted aspect terms align with both the user evidence sentence and the item evidence sentences, and the held-out review repeats these terms. Figure 12 shows a low-coverage example where the evidence text does not reflect the held-out review's primary topics, illustrating that simple lexical overlap can miss paraphrases and implicitly stated reasons. These examples motivate future extensions that use semantic matching or aspect extraction beyond surface terms.

Table 12 reports measured preprocessing and training times for the main pipeline components. The review document construction and TF–IDF fitting are fast relative to SVD training on the larger Video Games

set, and the exact term-level faithfulness computation scales linearly with the number of test users because it avoids re-running the recommender under perturbations.

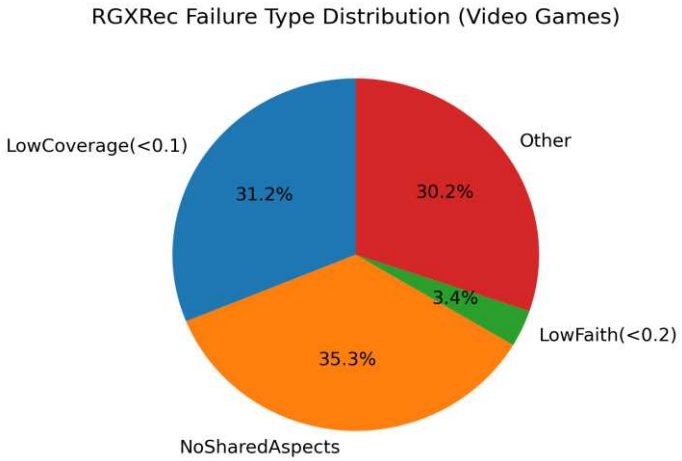

Figure 10. RGXRec failure type distribution on Video Games

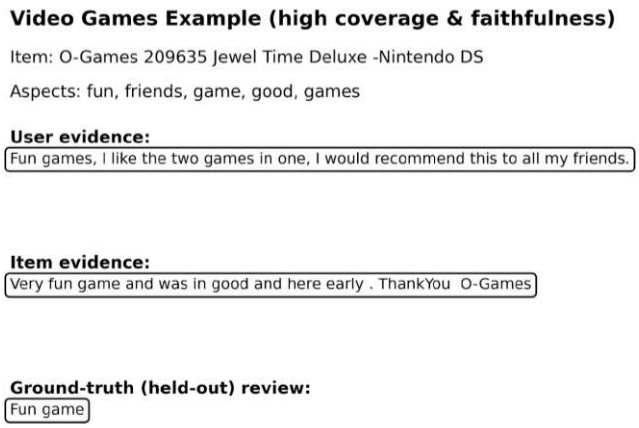

Figure 11. Example explanation with high coverage and faithfulness (Video Games)

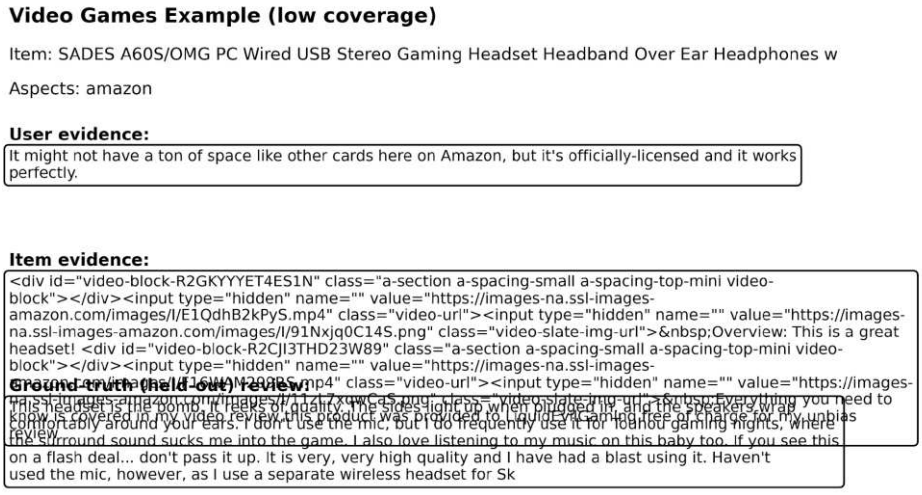

Figure 12. Example explanation with low evidence coverage (Video Games)

Table 12. Measured runtime (seconds) for key pipeline stages on the two datasets

	DocBuild_s	TFIDFfit_s	SVDfit_s	ExplainEval_s
Luxury_Beauty	0.429	0.745	29.745	1.917
Video_Games	6.604	9.369	40.433	31.361

3.2 Discussion

The results show that review-grounded signals are not equally useful across domains. In Luxury Beauty, review language repeatedly describes stable product aspects such as scent, texture, and skin sensitivity, so lexical overlap captures meaningful user–item alignment and the hybrid model improves both NDCG and Recall. In Video Games, user vocabulary is more diverse and the metadata graph is denser, which explains why collaborative and graph-based signals remain stronger for ranking. A second observation is that explanation relevance and explanation faithfulness should be reported separately. Personalized evidence sentences can improve coverage and ROUGE, but interaction-only models still have zero textual faithfulness because their scores do not depend on review text. RGXRec therefore provides a more defensible explanation setting than post-hoc quoting: even when its faithfulness is lower than ReviewSim due to the collaborative component, the textual evidence still corresponds to a defined portion of the hybrid score.

The sensitivity analysis further clarifies the role of α. Lower α values increase review-grounded faithfulness but may reduce ranking quality, whereas higher α values favor ranking accuracy at the cost of textual explainability. The selected α=0.75 offers a practical compromise because it keeps competitive recommendation performance while preserving non-zero model-level faithfulness.

This study has two main limitations. First, the extractive explanation module relies on lexical overlap and may miss paraphrases or multi-word aspects. Second, the experiments use sampled ranking with one positive item per user. Even with these limitations, the pipeline is reproducible, computationally light, and suitable as a baseline for future extensions using richer text encoders or stronger collaborative models.

4. Conclusion

This paper presented RGXRec, a review-grounded hybrid recommender that combines collaborative latent factors with TF-IDF review similarity and produces extractive evidence from reviews. The main contribution is an end-to-end evaluation setting that reports ranking quality together with evidence coverage, ROUGE-1, sentiment agreement, and an exact term-attribution faithfulness score for the text component.

Experimental results on Luxury Beauty and Video Games show that the usefulness of review-grounded explanation is domain dependent. RGXRec achieved the strongest overall results in Luxury Beauty and maintained competitive ranking with non-zero textual faithfulness in Video Games, where collaborative and metadata signals were stronger. Overall, the study shows that explainable recommendation should not be evaluated by human-readable evidence alone; the explanation must also be tied to the scoring mechanism. The proposed pipeline can therefore serve as a reproducible baseline for future work on more expressive yet faithful review-aware recommenders.

References

[1] J. Ni, J. Li, and J. McAuley, "Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects," *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 188–197, 2019, doi: 10.18653/V1/D19-1018.

[2] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015, doi: 10.1145/2766462.2767755.

[3] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 2015-Augus, pp. 785–794, 2015, doi:

10.1145/2783258.2783381/SUPPL_FILE/P785.MP4.
[4] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009, doi: 10.1109/MC.2009.263.
[5] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," *Proceedings of the 10th International Conference on World Wide Web, WWW 2001*, pp. 285–295, 2001, doi: 10.1145/371920.372071/ASSET/CFB8B952-6F16-43A6-8125-16F950D0D3E3/ASSETS/371920.372071.FP.PNG.
[6] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011, doi: 10.1137/090771806.
[7] J. E. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," *Proc. 1st Instructional Conf. Machine Learning*, 2003.
[8] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002, doi: 10.1145/582415.582418.
[9] Y. Zhang and X. Chen, "Explainable Recommendation: A Survey and New Perspectives," *Foundations and Trends® in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020, doi: 10.1561/1500000066.
[10] Z. Xu, H. Zeng, J. Tan, Z. Fu, Y. Zhang, and Q. Ai, "A Reusable Model-agnostic Framework for Faithfully Explainable Recommendation and System Scrutability," *ACM Transactions on Information Systems*, vol. 42, no. 1, Aug. 2023, doi: 10.1145/3605357/ASSET/C83B5D09-928A-4E72-9F0D-369B55A11851/ASSETS/IMAGES/LARGE/TOIS-2022-0270-ALGO1.JPG.
[11] M. Hu and B. Liu, "Mining and summarizing customer reviews," *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177, 2004, doi: 10.1145/1014052.1014073.
[12] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," *RecSys 2013 - Proceedings of the 7th ACM Conference on Recommender Systems*, pp. 165–172, 2013, doi: 10.1145/2507157.2507163.
[13] L. Zheng, V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation," *WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, pp. 425–433, 2017, doi: 10.1145/3018661.3018665.
[14] C. Chen, M. Zhang, Y. Liu, and S. Ma, "Neural attentional rating regression with review-level explanations," *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, pp. 1583–1592, 2018, doi: 10.1145/3178876.3186070.
[15] J. Lei, C. Zhu, S. Yang, J. Wang, and Y. X. Yu, "Influence of Review Properties in the Usefulness Analysis of Consumer Reviews: A Review-Based Recommender System for Rating Prediction," *Neural Processing Letters*, vol. 55, no. 8, pp. 11035–11054, 2023, doi: 10.1007/S11063-023-11363-5/METRICS.
[16] P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-N recommendation tasks," *RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems*, pp. 39–46, 2010, doi: 10.1145/1864708.1864721/SUPPL_FILE/RECSYS2010-28092010-04-01.MOV.
[17] X. Chen, Y. Zhang, and J.-R. Wen, "Measuring 'Why' in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation," 2022, Accessed: May 12, 2026. [Online]. Available: https://arxiv.org/abs/2202.06466v1.
[18] X. Wang, Q. Li, D. Yu, Q. Li, and G. Xu, "Counterfactual Explanation for Fairness in Recommendation," *ACM Transactions on Information Systems*, vol. 42, no. 4, 2024, doi: 10.1145/3643670/ASSET/C7962D24-6CA7-4311-BE93-1FB81B3C5BFD/ASSETS/IMAGES/LARGE/TOIS-2023-0217-F07.JPG.
[19] H. Zhuang, W. Zhang, W. Chen, J. Yang, and Q. Z. Sheng, "Improving Faithfulness and Factuality with Contrastive Learning in Explainable Recommendation," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 1, p. 23, 2024, doi: 10.1145/3653984/ASSET/D22C586E-C379-4A9A-8EA7-20B7C1BB6B2A/ASSETS/IMAGES/LARGE/TIST-2023-07-0395-F06.JPG.
[20] E. Hasan, M. Rahman, C. Ding, J. X. Huang, and S. Raza, "Review-based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives," *ACM Computing Surveys*, vol. 58, no. 1, p. 41, 2025, doi: 10.1145/3742421/ASSET/60B50CED-0752-4C0C-B0DC-627881A7F49F/ASSETS/IMAGES/LARGE/CSUR-2024-0435-F05.JPG.
[21] ZhouYao, WangHaonan, HeJingrui, and WangHaixun, "Review-Based Explainable

Recommendations: A Transparency Perspective," *ACM Transactions on Recommender Systems*, vol. 3, no. 3, pp. 1–20, 2025, doi: 10.1145/3701762.

[22] Q. Hao, C. Wang, Y. Xiao, and W. Zheng, "IReGNN: Implicit review-enhanced graph neural network for explainable recommendation," *Knowledge-Based Systems*, vol. 311, p. 113113, 2025, doi: 10.1016/J.KNOSYS.2025.113113.

[23] J. DeYoung *et al.*, "ERASER: A Benchmark to Evaluate Rationalized NLP Models," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020, doi: 10.18653/V1/2020.ACL-MAIN.408.

[24] Q. Lyu, M. Apidianaki, and C. Callison-Burch, "Towards Faithful Model Explanation in NLP: A Survey," *Computational Linguistics*, vol. 50, no. 2, pp. 657–723, 2024, doi: 10.1162/COLI_A_00511.