A EXPLANATION RANKING POLICIES

To examine the validity and effectiveness of our proposed three metrics, we borrow the insight of discriminative power[6,5,7] for the meta-evaluation of IR metrics, which verify a metric through its ability to differentiate two systems. Hence, our initial attempt for the validity test of the proposed metrics is to implement several explanation generation methods that intuitively can optimize three metrics, respectively. Then we probe whether the metrics can identify the optimization method from other baseline methods. Additionally, we measure each optimization method in terms of all three evaluation perspectives to investigate the comprehensiveness of the overall evaluation framework. In this section, we introduce all implemented explanation generation methods.

A.1 Problem Setup

The post-hoc manner of explanation generation on Meituan platform leads us to frame the explanation generation as an **explanation ranking task** proposed in [1,2]. The formulation of the explanation ranking task is as follows:

$$Top(u, i, C_{u,i}, K) = \underset{E_{u,i} \subseteq C_{u,i}, |E_{u,i}|=K}{\operatorname{argmax}} \sum_{e \in E_{u,i}} \hat{r}_{u,i,e} \tag{1}$$

Each ranking policy π features different method of calculating $\hat{r}_{u,i,e}$, resulting in different rankings of candidate explanations.

A.2 Policies for Usefulness Optimization

The usefulness metric is related to users' interaction (e.g. click and purchase) with triplets (u, i, e). To predict CTR/CVR for usefulness, we employ a SOTA Tensor Factorization method: Bayesian Personalized Explanation Ranking (BPER)[2] for usefulness optimization.

- **CD (Canonical Decomposition)**: CD predicts score $\hat{r}_{u,i,e}$ by summing over the element-wise multiplication of the latent factors of user p_u, item q_i and explanation o_e:

$$\hat{r}_{u,i,e} = (p_u \odot q_i)^\top o_e = \sum_{k=1}^{d} p_{u,k}.q_{i,k}.o_{e,k} \tag{2}$$

  The BPR loss for optimization is:

$$\min_{\theta} \sum_{u \in U} \sum_{i \in I_u} \sum_{e \in E_{u,i}} \sum_{e' \in E \setminus E_{u,i}} -\ln \sigma(\hat{r}_{u,i,e} - \hat{r}_{u,i,e'}) \tag{3}$$

- **PITF (Pairwise Interaction Tensor Factorization [3])**: Due to data sparsity, CD may be ineffective for explanation ranking. Thus, Rendle et al. [3] suggested converting CD's tensor decomposition into two simultaneous matrix factorizations of the user-explanation interaction matrix:

$$\hat{r}_{u,i,e} = p_u^\top o_e^U + q_i^\top o_e^I = \sum_{k=1}^{d} p_{u,k}.o_{e,k}^U + \sum_{k=1}^{d} q_{i,k}.o_{e,k}^I \tag{4}$$

  PITF shares the same loss function as CD (Equation (9)).

– **BPER (Bayesian Personalized Explanation Ranking [2])**: To bring more flexibility to PITF, Li et al.[2] proposed BPER, which features two more bias terms b^U, b^I and a hyper-parameter μ to balance between the two types scores for users and items.

$$\begin{aligned} \hat{r}_{u,e} &= p_u^\top o_e^U + b_e^U \\ \hat{r}_{i,e} &= q_i^\top o_e^I + b_e^I \\ \hat{r}_{u,i,e} &= \mu \hat{r}_{u,e} + (1-\mu)\hat{r}_{i,e} \end{aligned} \tag{5}$$

The BPR loss for BPER is defined as follows:

$$\min_\theta \sum_{u \in U} \sum_{i \in I_u} \sum_{e \in E_{u,i}} \left[\sum_{e' \in E \setminus E_u} -\ln \sigma(\hat{r}_{u,e} - \hat{r}_{u,e'}) + \sum_{e'' \in E \setminus E_i} -\ln \sigma(\hat{r}_{i,e} - \hat{r}_{i,e''}) \right] \tag{6}$$

Since the usefulness perspective involves two types of interactions (click and purchase), for each tensor factorization technique, we will train two models based on click and purchase, respectively. Then, we combine the prediction of two models to optimize the usefulness metric.

Baseline ranking policies

– **Random**: Randomly select top-k explanations from the candidate set $C_{u,i}$ for each (u, i).
– **General CTR Statistics based**: Rank the explanations according to the overall CTR on different explanation types calculated in the logged data.
– **General CVR Statistics based**: Rank the explanations according to the overall CVR on different explanatin types calculated in the logged data.

A.3 Polices for Fidelity Optimization

The intuition behind evaluating fidelity is to measure the correlation between explanations and system predictions, leading us to adopt LIME[4], a feature importance-based method, to enhance explanation fidelity.

LIME for estimating the score $\hat{r}_{u,i,e}$ Our LIME-based Policy uses LIME as a means for computing $\hat{r}_{u,i,e}$. The numerical score is correlated with the influence of the explanation feature $x_{u,i,e}$ learned by LIME to explain the prediction $y_{u,i}$ from online system. Following the procedure of LIME, we implement the LIME-based ranking policy as follows:

– Convert **candidate explanation** $C_{u,i}$ for each (u, i) in recommendation list p into categorical features. Combine these with other item features like average price and delivery distance to form feature vector $x_{u,i}$. Discretize continuous features and represent them as one-hot vectors $x'_{u,i} \in \{0,1\}^{|D|}$.
– Different from perturbation-based generation of synthetic sample z' around x' in the original LIME [4], we use other samples in the same ordered list p as the neighborhood of each $x'_{u,i}$. And the prediction on each sample by the original system f is based on its rank in p.

– Based on the constructed training dataset, we get a local interpretable model $g_{u,i}$ for each $(u,i) \in p$ using the same optimization method as LIME:

$$\underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) \tag{7}$$

– Estimate $r_{u,i,e}$ by the absolute weight of the feature $x^{'}_{u,i,e}$ in the model $g_{u,i}$:

$$\hat{r}_{u,i,e} = s(x^{'}_{u,i,e}, g_{u,i}) \tag{8}$$

Baseline ranking policies

– **Random**: Randomly select top-k explanations from the candidate set $C_{u,i}$ for each (u,i).
– **Per impression avg Rank Statistics based** : The fidelity metric measures how well selected explanations can reconstruct the original ranking of the recommended list, which leads us to develop a straightforward statistics-based method. For each ordered recommended list p, we compute the average rank of the valid (u,i) pair for a candidate explanation in p. This practice is supported by our observation from the Meituan platform that some explanation types (e.g. Leaderboard based and User Interaction History based) are only legitimate for those top-ranked (u,i). The estimation of the score $\hat{r}_{u,i,e}$ is as follows, where $p_{u,i}$ denotes the ordered list contains (u,i):

$$\hat{r}_{u,i,e} = \frac{\sum_{(u_j,i_j) \in p_{u,i}} I(e \in C_{u_j,i_j}) Rank(u_j, i_j | p_{u,i})}{\sum_{(u_j,i_j) \in p_{u,i}} I(e \in C_{u_j,i_j})} \tag{9}$$

A.4 Polices for Generalizability Optimization

To achieve strong generalizability, we aim to construct a globally interpretable proxy model for each user u to capture their consistent preferences:

– For each user u, the global model is trained on all their own time-ordered set of historical recommended lists P_u on the platform.
– We transform candidate explanations $\{e_{u,i} \in C_{u,i} \mid (u,i) \in P_u\}$ into categorical features, combining with other features of (u,i) to obtain feature vectors $x_{u,i} \in X_u$. Pseudo-labels $y_{u,i}$ are assigned based on the rank of (u,i) in its corresponding ranked list p.
– A proxy linear model l_u is trained by minimizing the squared error between predicted and actual ranks:

$$w_u = \underset{w \in W}{\operatorname{argmin}} \sum_{(u,i) \in P_u} (l_u(x_{u,i}; w) - y_{u,i})^2 \tag{10}$$

– Using the learned weights w_u, we estimate $r_{u,i,e}$ by taking the absolute value of the weight corresponding to the feature index j of explanation e in $x_{u,i}$: $\hat{r}_{u,i,e} = \|w_{u,j}\|$.

Baseline ranking policies

- **Random**: Randomly select top-k explanations from the candidate set $C_{u,i}$ for each (u, i).
- **Per User Avg Rank Statistics based** : The similar Kendall's τ metric for generalizability as fidelity metric brings us to a similar practice as the Per Impression Avg Rank Statistics based method for optimizing fidelity. However, we use per-user statistics instead of per-impression statistics to meet the generalization requirement. The formal estimation of the score $\hat{r}_{u,i,e}$ is as follows:

$$\hat{r}_{u,i,e} = \frac{\sum_{p \in P_u \wedge (u,i) \in p} \sum_{(u_j,i_j) \in p} I(e \in C_{u_j,i_j})(Rank(u_j, i_j|p))}{\sum_{p \in P_u \wedge (u,i) \in p} \sum_{(u_j,i_j) \in p} I(e \in C_{u_j,i_j})} \tag{11}$$

B User Study Design

We build an annotation platform to simulate the online food recommendation scenario. We first sample impressions (i.e. recommendation lists) from the real historical logs of Meituan and generate explanations with different explanation ranking policies. Then we ask participants to act like real users on the food recommendation platform to evaluate the explanations generated by different policies in terms of effectiveness, transparency, persuasiveness, trust, and satisfaction proposed in a 5-level Likert Scale. In the end, we conduct a post-study interview to investigate the reason for the participants' evaluation.

B.1 Setup

Data Preparation We sample a small dataset from the collected Meituan Dataset to simulate the impressions (i.e. recommendation lists) on the annotation platform for each participant. We first randomly sample a set of users who have at least five historical impressions in the dataset. Then, for each sampled user, we sample five impressions and sort them chronologically.

Participant Recruitment We recruit 26 participants (57.7% female; 42.3% male) from different majors (economics, agriculture, law, statistics, computer science, et.al) at our university with an online SNS. Each participant is paid 9 dollars. This study is conducted in Chinese, and all the participants are native speakers. All the participants declare that they have basic computer operation skills and experience using online food recommendation platforms.

B.2 Procedure

The overall procedure of our user study is depicted in Figure 1. During the experiment, each participant needs to complete two sessions. In each session, the participant plays the role of a user on a food recommendation platform and rates the explanations generated by different methods for the same set of impressions. To be more specific, each session consists of 4 rounds, and each round

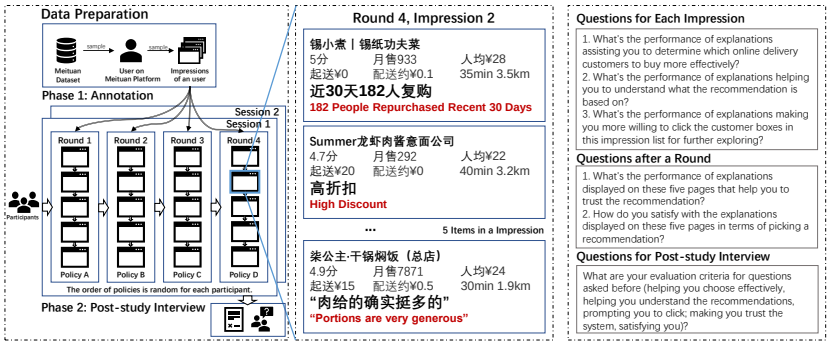

Fig. 1. The left panel gives an overview of the user study. The middle offers an example of the displayed impression. Each item in the impression is paired with basic information, such as the restaurant name and monthly sales. The original language is Chinese. We marked the English translations for the explanations in red. The right panel lists all the questions that we ask participants to answer in our study.

contains 5 impressions of the same Meituan user. In each round, we randomly choose one explanation ranking policy from four candidate policies: random policy and three best policies in terms of usefulness (i.e. BPER-based), fidelity (i.e. LIME-based), and generalizability (i.e. Interpretable proxy model-based). Then the selected policy generates explanations for each impression. The impressions used in each round of the same session are identical, except that the explanations for each impression are generated by different explanation ranking policies. After reviewing each impression, the participant is asked to evaluate the explanations in terms of persuasiveness, effectiveness, and transparency. The detailed questions are listed in the "Questions for each Impression" panel of Figure 1. Then, after evaluating all five *impressions* in each *round*, the participant answers two additional questions about goals of satisfaction and trust, as listed in the "Questions after a Round" panel of Figure 1. Finally, we conduct a post-study interview by asking each participant to write a summary of the criterion of their evaluation.

The average completion time of the annotation phase is 29 minutes and the average length of the rating criterion summary is 177 Chinese characters, which guarantees the quality of our study.

References

1. Li, L., Zhang, Y., Chen, L.: Extra: Explanation ranking datasets for explainable recommendation. In: Proceedings of the 44th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 2463–2469 (2021)
2. Li, L., Zhang, Y., Chen, L.: On the relationship between explanation and recommendation: Learning to rank explanations for improved performance. ACM Trans. Intell. Syst. Technol. **14**(2) (feb 2023)

3. Rendle, S., Schmidt-Thieme, L.: Pairwise interaction tensor factorization for personalized tag recommendation. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 81–90 (2010)
4. Ribeiro, M.T., Singh, S., Guestrin, C.: " 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. pp. 1135–1144 (2016)
5. Robertson, S.E., Kanoulas, E., Yilmaz, E.: Extending average precision to graded relevance judgments. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 603–610. SIGIR '10, Association for Computing Machinery, New York, NY, USA (2010)
6. Sakai, T.: Evaluating evaluation metrics based on the bootstrap. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 525–532. SIGIR '06, Association for Computing Machinery, New York, NY, USA (2006)
7. Smucker, M.D., Clarke, C.L.: Time-based calibration of effectiveness measures. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 95–104. SIGIR '12, Association for Computing Machinery, New York, NY, USA (2012)