A Appendix: 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 [32] in a 5-level Likert Scale. In the end, we conduct a post-study interview to investigate the reason for the participants' evaluation.

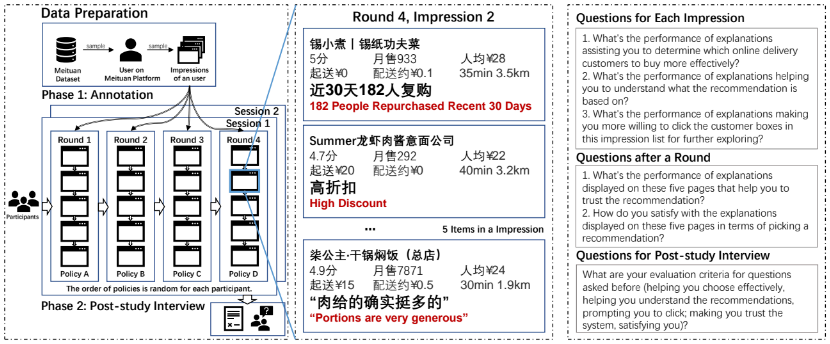
A.1 Setup

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.

A.2 Procedure

The overall procedure of our user study is depicted in Figure 4. 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 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 4. Then, after evaluating all five *impressions* in each *round*, the participant answers two additional questions about goals of satisfaction and trust, as listed in



锡小煮丨锡纸功夫菜
5分 月售933 人均¥28
起送¥0 配送约¥0.1 35min 3.5km
近30天182人复购
182 People Repurchased Recent 30 Days

Summer龙虾肉酱意面公司
4.7分 月售292 人均¥22
起送¥20 配送约¥0 40min 3.2km
高折扣
High Discount

…

柒公主·干锅焖饭（总店）
4.9分 月售7871 人均¥24
起送¥15 配送约¥0.5 30min 1.9km
“肉给的确实挺多的”
“Portions are very generous”

Questions for Each Impression

1. What's the performance of explanations assisting you to determine which online delivery customers to buy more effectively?
2. What's the performance of explanations helping you to understand what the recommendation is based on?
3. What's the performance of explanations making you more willing to click the customer boxes in this impression list for further exploring?

Questions after a Round

1. What's the performance of explanations displayed on these five pages that help you to trust the recommendation?
2. How do you satisfy with the explanations displayed on these five pages in terms of picking a recommendation?

Questions for Post-study Interview

What are your evaluation criteria for questions asked before (helping you choose effectively, helping you understand the recommendations, prompting you to click; making you trust the system, satisfying you)?

Fig. 4. 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.

the “Questions after a Round” panel of Figure 4. 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.