MUSE: A Multimodal Conversational Recommendation Dataset with Scenario-Grounded User Profiles

Zihan Wang[1], **Xiaocui Yang**[1], **Yongkang Liu**[1], **Shi Feng**[1,†]
Daling Wang[1], **Yifei Zhang**[1]
[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
2310744@stu.neu.edu.cn, yangxiaocui@cse.neu.edu.cn, misonsky@163.com
{fengshi, wangdaling, zhangyifei}@cse.neu.edu.cn

Abstract

Current conversational recommendation systems focus predominantly on text. However, real-world recommendation settings are generally multimodal, causing a significant gap between existing research and practical applications. To address this issue, we propose MUSE, the first multimodal conversational recommendation dataset. MUSE comprises 83,148 utterances from 7,000 conversations centered around the Clothing domain. Each conversation contains comprehensive multimodal interactions, rich elements, and natural dialogues. Data in MUSE are automatically synthesized by a multi-agent framework powered by multimodal large language models (MLLMs). It innovatively derives user profiles from real-world scenarios rather than depending on manual design and history data for better scalability, and then it fulfills conversation simulation and optimization. Both human and LLM evaluations demonstrate the high quality of conversations in MUSE. Additionally, fine-tuning experiments on three MLLMs demonstrate MUSE's learnable patterns for recommendations and responses, confirming its value for multimodal conversational recommendation. Our dataset and codes are available at https://github.com/wzhwzhwzh0921/Muse.

1 Introduction

Conversational recommendation (CR) (Lei et al., 2020) is an emerging research field. It leverages natural language to deliver personalized, context-aware suggestions for users. Unlike the traditional implicit recommendation paradigm (Jalili et al., 2018; Wang et al., 2019, 2021), CR emphasizes both the recommendation performance and the real-time dialogue with users. Some existing CR datasets supporting the research, such as Redial (Li et al., 2018) and TG-Redial (Zhou et al., 2020), are launched through crowdsourcing. Crowdworkers take on dual roles as users and rec-assistants, interacting with each other to generate CR data. Innovatively, Pearl (Kim et al., 2024) and LLM-Redial (Liang et al., 2024) harness the advanced capabilities of large language models (LLMs) (Chang et al., 2024) to fulfill the simulation of conversation.

Figure 1: Comparison of data cases from Redial and MUSE. Red denotes interactions about visual features, and green shows scenario-related content.

While these datasets have significant contributions, they still possess limitations. (1) These datasets are predominantly limited to textual modality. However, like other recommendation fields (Zhou et al., 2023), text-only data is insufficient to simulate the multisensory decision-making processes that characterize real-world shopping behaviors. Multimodal information is particularly crucial for visually driven fields, such as clothing and food. (2) Their scalability is limited. Given CR's data-driven nature, CR datasets are supposed to be scaled to include more conversations and a wider range of users and items (Liang et al., 2024), enabling the development of more comprehensive CR systems. Existing LLM-based methods have demonstrated the ability to scale conversation volume, which partially addresses the limitations of

†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 1027–1053
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Table 1: Comparison Between MUSE and some dialogue datasets. **Exp. Rec** means 'Expainable Recommendation.' **Conv. Scal.** and **U/I. Scal.** are the scalability of conversations and users/items

Datasets	#Dial.	#Utter.	Domains	Types	Exp. Rec	Conv. Scal.	U/I. Scal.	U. Profile	Modal
Multimodal Dialogue Datasets									
SURE (Long et al., 2023)	12K	223K	Fashion, furniture	Task-oriented.Dial	–	×	–	–	VR+Text
SIMMC1.0 (Crook et al., 2021)	13K	169K	Fashion, furniture	Task-oriented.Dial	–	×	–	–	VR+Text
SIMMC2.0 (Kottur et al., 2021)	11K	117K	Fashion, furniture	Task-oriented.Dial	–	×	–	–	VR+Text
IGC (Mostafazadeh et al., 2017)	4K	25K	Image concepts	Image-based QA	–	×	–	–	Image+Text
GuessWhat (De Vries et al., 2017)	155K	1.6M	Image concepts	Image-based QA	–	×	–	–	Image+Text
MMD (Saha et al., 2018)	150K	6M	Fashion	Conv.Search	–	×	–	–	Image+Text
MMCONV (Liao et al., 2021)	5.1K	39.7K	Travel	Conv.Search	–	×	–	–	Image+Text
Conversational Recommendation Datasets									
Redial (Li et al., 2018)	10K	182K	Movie	Conv.Rec	×	×	×	From human design	Text
OpenDialKG (Moon et al., 2019)	15K	91K	Movie, Book	Conv.Rec	×	×	×	From human design	Text
TG-Redial (Zhou et al., 2020)	10K	129K	Movie	Conv.Rec	×	×	×	From human design	Text
DuRecDial (Liu et al., 2020)	10.2K	156K	Movie, Music, Food	Conv.Rec	×	×	×	From human design	Text
INSPIRED (Hayati et al., 2020)	1K	35K	Movie	Conv.Rec	×	×	×	From human design	Text
Pearl (Kim et al., 2024)	57.2K	482K	Movie	Conv.Rec	✓	✓	×	From history data	Text
LLM-Redial (Liang et al., 2024)	47.6K	548K	Movie, Book, Sports	Conv.Rec	✓	✓	×	From history data	Text
MUSE	7K	83K	Cloth, Shoes, Jewelry	Conv.Rec	✓	✓	✓	From real-world scenarios	Image+Text

crowdsourcing datasets. However, due to their heavy reliance on user history data, their user and item coverage remains confined to historically collected data. Moreover, facing increasingly stringent privacy regulations (Voigt and Von dem Bussche, 2017; Regulation, 2016; Harding et al., 2019) and cold-start situations (Lam et al., 2008), these methods become hard to apply.

To tackle these challenges, we introduce MUSE, a **MU**ltimodal Conversational Recommendation Dataset with **S**c**E**nario-grounded user profiles. To the best of our knowledge, it is the first multimodal dialogue dataset specifically designed for CR tasks. MUSE is based on real-world products from the multimodal dataset, Amazon Cloth, Shoes, and Jewelry (Hou et al., 2024), and comprises a total of 7,000 multimodal conversations. Figure. 1 illustrates comparative case studies of conversations drawn from ReDial and MUSE. Motivated by the effectiveness of LLM-based data synthesis in previous works, a multi-agent framework powered by Multimodal LLMs (MLLMs) facilitates conversations in MUSE. The framework has three modules: Scenario-Grounded User Profile Generator, Simulated Conversation Generator, and Conversation Optimizer. Inspired by the understanding that user engagement in recommendations involves both preference-based interactions ("I like...") and scenario-grounded requirements (Paul et al., 2016) ("Need for wedding/sports..."), the Scenario-Grounded User Profile Generator adopts an innovative approach. It places roles within different real-world scenarios and identifies their diverse needs to match different suitable target items, resulting in scenario-grounded user profiles. The infinite diversity of real-world scenarios naturally brings the scalability of both users and items. The Simulated Conversation Generator utilizes an ensemble of interconnected sub-agents that synergistically leverage multimodal information for advanced dialogue simulation. The system integrates fine-grained multimodal characteristics into the conversation process and adds natural chit-chat parts that simulate human conversations. The Conversation Optimizer consists of a Rewriter and a Reviewer. The former amplifies dialogue diversity through both sentence/word variation and colloquial elements, and the latter filters out conversations that are not eligible, ensuring data quality in MUSE. As a result, MUSE offers diverse element-rich multimodal conversations and keeps remarkable scalability because of the framework. Table. 1 presents a comparison between MUSE and representative datasets, which are categorized into multimodal dialogue datasets and CR datasets.

We perform comprehensive assessments of MUSE conversations through both human evaluation and LLM analysis, evaluating them from both global and granular perspectives. Empirical results indicate that MUSE generates dialogues with exceptional fluency, diversity, depth of bilateral interaction, and multimodal coherence. To validate the utility of MUSE as a multimodal conversational recommendation (MCR) dataset, we conduct extensive evaluations on three representative open-source MLLMs under both zero-shot and fine-tuned configurations. The quantitative results demonstrate MUSE's capacity to facilitate reliable recommendation reasoning and response generation for CR, establishing its value as a benchmark dataset for MCR.

2 Related Work

2.1 Conversational Recommendation

Conversational recommendation (CR) research can be broadly divided into two categories (Jannach et al., 2021; Fu et al., 2020; Jannach and Chen, 2022). The first frames the task as a multi-step decision-making process, leveraging reinforcement learning to minimize the number of conversation rounds required to identify the target item (Deng et al., 2021; Zhang et al., 2022). The second prioritizes natural language communication, aiming to gain a deep understanding of user preferences through conversation and, in some cases, even influence those preferences (Li et al., 2018; Wang et al., 2022a; Ravaut et al., 2024). Our work centers on the latter. Since the inception of the Redial dataset (Li et al., 2018), numerous crowdsourcing datasets (Moon et al., 2019; Zhou et al., 2020; Liu et al., 2021; Hayati et al., 2020) have emerged, expanding the CR task across various data. Pearl (Kim et al., 2024) and LLMRedial (Liang et al., 2024) introduce innovative approaches with LLMs. However, these datasets are limited to plain text. MUSE marks a major breakthrough in the field as a multimodal conversational recommendation dataset.

2.2 LLM-Driven Data Synthesis

Large Language Models (LLMs) possess extensive and diverse world knowledge (Zhao et al., 2023), enabling them to comprehend and generate complex language with human-like proficiency. Therefore, LLMs have demonstrated remarkable potential in data synthesis (Ding et al., 2024; Wang et al., 2022b; Sahu et al., 2022; Liu et al., 2024), especially in generating conversational datasets (Abbasiantaeb et al., 2024; Kim et al., 2022). In the realm of conversational recommendation, two prominent approaches—Pearl (Kim et al., 2024) and LLM-Redial (Liang et al., 2024)—have gained recognition. While the two approaches enable increased dialogue quantity, they struggle to diversify user and item coverage due to their reliance on history data. MUSE ingeniously harnesses the vast world knowledge embedded in multimodal LLMs to craft user profiles and match target items, enabling full scalability in conversational recommendation datasets.

3 MUSE Construction

We construct MUSE, the first MCR dataset, which is built based on real-world clothing product information. In this section, we introduce the multi-agent framework behind MUSE for conversation synthesis, which is organized into three functional components as descriptions in Figure. 2: ① Scenario-Grounded User Profile Generator; ② Simulated Conversation Generator; ③ Conversation Optimizer. The main backbone LLM of this framework is `gpt-4o-mini`[†]; check Appendix A.5 for more MLLM settings.

3.1 Data Preprocess

To anchor MUSE conversations in real-world products, we utilize the Amazon Clothing, Shoes, and Jewelry dataset (Hou et al., 2024), which combines both textual and visual information. Using multimodal product information, we build a local product database to support subsequent product retrieval operations, where each product is accompanied by a main image and a text description. The detailed processing is documented in the Appendix A.1.

3.2 Scenario-Grounded User Profile Generator

In the real world, users' immediate purchasing decisions are not solely driven by personal interests but are significantly influenced by a multitude of external factors (Piligrimienė et al., 2020). Among these, scenario context plays a pivotal role, as different scenarios, such as buying a suit for a party or a T-shirt for the summer, drive diverse consumer needs, requiring tailored products to meet them. Note that these requirements can correspond to multimodal product features. Ideally, every user need can be traced back to a real-life scenario, just as suitable scenarios and users can be identified for any product, which naturally enables flexible scalability in both user and item. Building on this concept, we develop the Scenario-Grounded User Profile Generator, whose entire process can be divided into two steps, each assigned to a dedicated agent. The first step is to collect diverse real-world basic scenarios. In the second step, we situate users in various scenarios and match them with products that align with both the scenario requirements and their individual characteristics to get detailed user profiles to support the subsequent user simulator. Figure. 3 illustrates the whole workflow.

3.2.1 Basic Scenario Generation

Basic scenarios reflect real-world events that shape user shopping behavior. To capture a diverse range of such scenarios, the Basic Scenario Generator harnesses the expansive capabilities of LLMs. It begins with a set of seed scenarios related to clothing purchases, including but not limited to attending important occasions, meeting athletic needs, celebrating special dates, and purchasing gifts

[†]https://openai.com

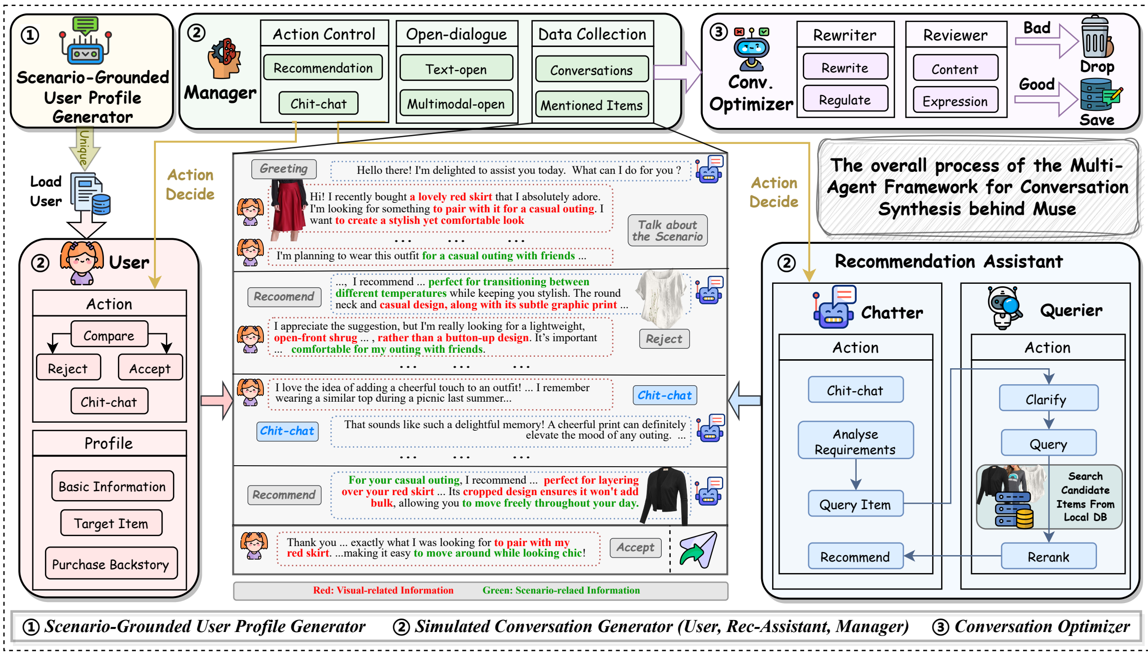

Figure 2: The multi-agent framework for synthesizing MCR data in MUSE.

for friends and family. Utilizing the self-instruct method (Wang et al., 2022b), we expand the scenarios. During this expansion, the BLEU (Papineni et al., 2002) metric is adopted to eliminate duplicates, ensuring that the collected real-world scenarios maintain their diversity. Ultimately, we identify 593 basic scenarios for MUSE.

3.2.2 User Profile Generation

In our design, a complete user profile consists of three key components: basic user information, target products, and the purchase backstory. To enhance the traditional user profile, we incorporate detailed driving information about the user's current shopping behavior, which we call the Purchase Backstory. This addition provides a more comprehensive explanation of the user's motivation based on scenario requirements, enabling more accurate role simulation (Chen et al., 2024). Specifically, we first generate basic user information, including demographic details such as age and occupation, and then match the user with a specific scenario and a target product. In this step, we utilize MLLMs to evaluate the rationality of the (user, scenario, product) combination from two perspectives: user-product matching and scenario-product matching and screen low-quality results. In the second step, the MLLMs generate a purchase backstory for each reasonable combination (user, scenario, product). Here, we also employ the BLEU metric to remove duplicates, ensuring the uniqueness of each purchase backstory. By integrating the above information, we create scenario-grounded user profiles.

3.3 Simulated Conversation Generator

This generator comprises three specialized agents: a User Simulator and a Rec-assistant Simulator, both dedicated to simulating user dialogues through iterative multimodal interactions, and a Manager that oversees the actions of the two simulators and collects data, as shown in Figure. 2.

3.3.1 User Simulator

The user simulator adopts scenario-grounded user profiles generated above for role-playing. Research by (Zhang et al., 2024; Wang et al., 2023a) demonstrates that advanced LLMs are highly effective in performing user simulation tasks. Among the factors in the provided user profiles, two stand out as having the greatest impact on CR. (1) **Scenario-based Requirements** represent the users' primary concerns and needs, which are seamlessly incorporated into open-dialogues. Open-dialogues represent the first few rounds of the conversation, in which rec-assistant learns some basic requirements from the user through question and answer, explained in Section 3.3.3. (2) **Target requirements** specify the desired features of the target item of the user. These requirements align with Pearl's methodology (Kim et al., 2024), equipping the user simulator with a clear framework to generate more precise and insightful feedback.

The user simulator is designed to perform two primary types of actions. (1) **Actions towards rec-**

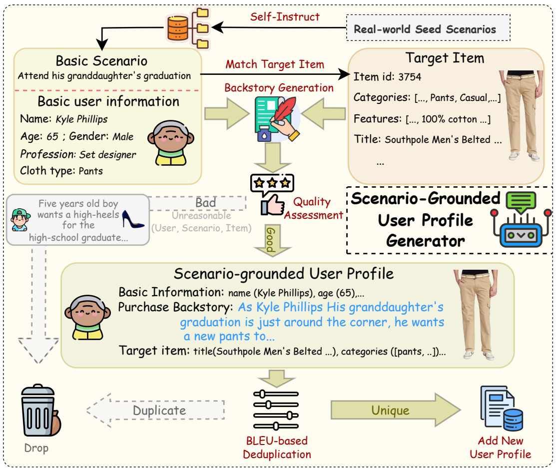

Figure 3: Workflow of the scenario-grounded user profile generator

ommended items. The user compares the recommended product's visual and textual information against his/her requirements to identify any discrepancies. If the product falls short of the user's needs, the simulator provides a logical justification for rejection. When the product aligns with requirements, it generates an acceptance response with appropriate appreciation. (2) **Chit-chat.** (Wang et al., 2023b; Liang et al., 2024) emphasize chit-chat is a vital component of natural human dialogue. Existing LLM-based datasets have largely overlooked it while MUSE acknowledges and incorporates this crucial aspect into conversations.

3.3.2 Rec-assistant Simulator

The rec-assistant simulator is composed of two sub-agents: **Chatter**, which specializes in user communication, and **Querier**, which handles recommendations and provides product information. This division is designed to align with the primary goals of conversational recommendations: delivering high-quality interactions and offering products that meet user expectations.

Chatter. Chatter prioritizes the quality of its responses and supports two main actions: recommendation and chit-chat. For recommendations, it prompts the Querier to provide a suitable product. Based on the contextual information, Chatter evaluates the compatibility between the product's multimodal information and the user's needs, using this alignment as a key selling point in its recommendation. Chit-chat focuses on responding to the user's casual dialogues, offering engaging content and emotion support for better user experiences.

Querier. Querier is responsible for finding products that meet the user's exposed requirements. First, the Querier analyzes the overall current conversation to craft the user's interests. Then, it clarifies the user's interests because in natural language expression, the user's preferences may be vague, and they need to be matched with the local database that focuses on attribute descriptions. For example, "need a quick-drying clothing" will be clarified as "need clothes made of polyester, modal... materials." Based on the clarified user needs, a rough retrieval is performed from local products, followed by LLM-powered reranking to identify the best-matched product. Chatter is provided with multimodal information of the best-matched product. If the round limit is reached, the Querier provides the user's target product to end the conversation.

3.3.3 Manager

The Manager's responsibilities encompass three key functions: (1) initiating the open-dialogues to start conversations, (2) orchestrating action control to regulate the exchanges between users and the rec-assistant in each round, and (3) performing data collection to document the conversation content.

Open-dialogue. Open-dialogues refer to some initial exchanges between the user and the system at the start of the recommendation process, primarily involving greetings and basic inquiries. It is commonly observed in artificial CR datasets (Li et al., 2018; Zhou et al., 2020), underscoring its significance as a feature of human conversations. Therefore, in MUSE, we integrate the simulation of two types open-dialogue to further imitate human conversations. One is the text-open, like those text-only datasets. The other is the multimodal-open, designed to accommodate users who need to express their needs with images. We constrain multimodal-open dialogues specifically to outfit-matching applications, which is a common concern in Clothing data (Lin et al., 2019; Xu et al., 2024). In the context of clothing datasets, outfit coordination represents a typical multimodal-open use case - such as "Please select a T-shirt to complement these pants" accompanied by an image. More details are in Appendix A.3.

Action Control. The Manager needs to guide the actions of the user and rec-assistant in rounds, mainly to control the distribution of chit-chat rounds and recommendation rounds. Different action monitoring strategies can be tailored to specific contexts. For example, "the longer the conversation, the less likely users are to engage in chit-chat," which is the approach we have adopted.

Data Collection. The process is straightforward: Manager records and organizes the conversation content in a structured format.

3.4 Conversation Optimizer

The conversation optimization system consists of two specialized agents—the Rewriter and Reviewer—who work in tandem to enhance conversation diversity and perform conversation quality assessments to screen low-quality conversations.

3.4.1 Rewriter

In the previous conversation generation process, to ensure the agents strictly followed the instructions in the prompt, maintain dialogue stability, and relieve LLM hallucinations, we set the MLLM model temperature for both the User and Rec-Assistant Simulators to 0.1. However, the setting results in the sentence structure and wording of the User and Rec-Assistant outputs being relatively repetitive. To introduce more diverse expressions in MUSE, we implement a Rewriter tasked with modifying the sentence structure and wording of the dialogue. An internal supervision mechanism is employed to ensure consistency in content, with particular emphasis on preserving the accuracy and immutability of product attribute information. Additionally, we incorporate probabilistic "use colloquial expressions" instructions into the prompt to generate more human-like responses. As a result, greater diversity in expressions is achieved while maintaining logical coherence. The details are in Appendix A.4.

3.4.2 Reviewer

After completing all conversation rounds, we evaluate the overall quality of the conversation to filter out low-quality products. To ensure a reliable assessment, we use three key indicators: content quality, logical fluency, and user consistency. These indicators collectively assess the conversation content, and a scoring strategy is applied to calculate the total score. The combined results are then used to evaluate and screen the conversations effectively.

4 Experiments

In this section, we present comprehensive experiments to validate the value of MUSE. Initially, we analyze various data parameters in MUSE along with the composition of its dialogue elements. Then, we assess the dialogue quality in MUSE using a two-fold evaluation methodology: overall conversation-level quality and utterance-level quality. Furthermore, to demonstrate the practical utility of our dataset for the conversational recommendation, we conduct experiments with three representative open-source MLLMs, evaluating their performance in both recommendation generation and response generation on MUSE.

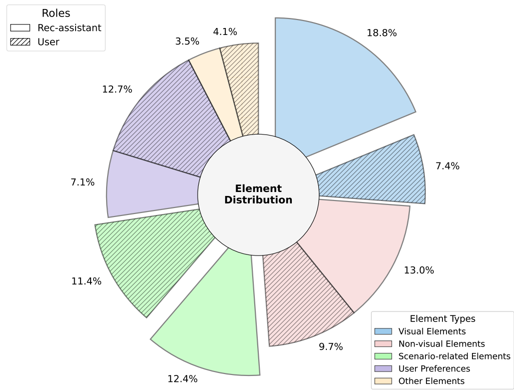

Figure 4: Distribution of dialogue Elements in MUSE

Table 2: Dataset Statistics of MUSE and other datasets.

Metric	MMCONV	Redial	INSPIRED	Pearl	MUSE
#Users	–	1.0K	1.0K	4.7K	7.0K
#Items	–	51.7K	1.7K	9.4K	13.7K
#Images	114K	–	–	–	13.7K
#4-Grams	230K	38K	140K	3.5M	2.3M
Distinct-3	0.24	0.27	0.55	0.09	0.30
Distinct-4	0.38	0.48	0.76	0.18	0.54
Avg.word/Turn	12.8	7.6	7.9	34.7	46.6

4.1 Statistics of MUSE

A comparative analysis of basic statistics between MUSE and three conversational recommendation datasets is shown in Table. 2. Both PEARL and MUSE demonstrate notably higher average word counts per conversation. It indicates that synthesized data from LLMs tend to produce more extensive expressions. Furthermore, the higher 4-grams (Loper and Bird, 2002) (nltk==3.9.1) confirms the presence of more distinctive content, more likely due to the detailed articulation of fine-grained product features. An intriguing observation emerges: PEARL exhibits notably low distinct-n (Li et al., 2015). Upon further investigation, we discover that while Pearl contains richer product content, it demonstrates excessive repetition in its dialogue patterns and phrasal expressions. The presence of rigid sentence patterns can substantially reduce conversational diversity and hinder the generalization ability of trained systems. MUSE overcomes this limitation by employing a Rewriter, which effectively diversifies sentence structures and lexical choices, thereby improving overall data quality. More discussion is placed in Appendix B.4.

Also, we employ `gpt-4o` to extract and clas-

Table 3: Comparison of different datasets across multiple evaluation metrics

Metrics	MMCONV	Redial	INSPIRED	PEARL	MUSE
Natural(0-2)	1.41	1.57	1.71	1.66	**1.85**
Logical(0-2)	1.41	1.60	1.62	1.78	**1.88**
Informative(0-2)	1.37	1.53	1.51	1.67	**1.80**
P-C Correlation(0-2)	1.68	1.92	1.83	1.50	**1.98**
I-T Correspondence(0-2)	1.35	-	-	-	**1.91**

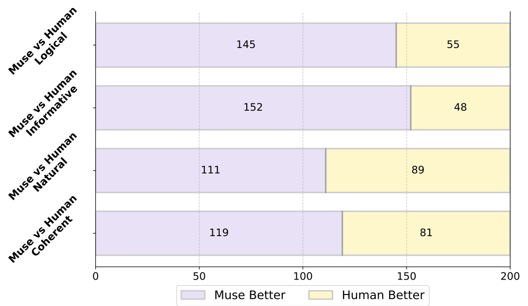

Figure 5: Utterance-level comparison: the quality between human responses and Muse's utterances,

sify dialogue elements across 5,000 dialogues in MUSE to conduct an in-depth analysis. Figure. 4 illustrates the distribution of main dialogue elements in the Muse dataset. The dialogue elements are systematically categorized based on both their functional types (Element Types) and participant sources (Roles). The analysis reveals that both users and Rec-assistants frequently incorporate visual and scenario-related elements in their exchanges, accounting for a substantial proportion of the dialogue content. This highlights the importance of these elements in enriching interactions and supporting communication within MUSE.

4.2 Conversation-level Evaluation

To evaluate the global quality of our dataset, we conduct comparative analyses against four representative datasets with similar characteristics to MUSE: MMCONV, Redial, Inspired and PEARL as Table. 3. Given the widespread adoption and proven effectiveness of LLMs in various evaluation tasks (Liu et al., 2023; Desmond et al., 2024), we select the LLM-based method for conversational-level evaluation. Furthermore, utilizing LLM evaluation offers a distinct advantage: given the unique content and structural characteristics of MUSE, conventional manual assessment for conversation-level quality could potentially introduce subjective biases. 200 conversations are randomly sampled from each dataset, and LLMs are employed to evaluate them on a scale of 0-2 across five dimensions: dialogue naturalness (Natural), logical coherence (Logical), information richness (Informative), product-context relevance (P-C Correlation), and image-text alignment (I-T Correspondence, specifically for multimodal datasets). Details are presented in Appendix D.4. The results demonstrate MUSE's superior performance across all five metrics. MUSE's high scores in naturalness and coherence establish the fundamental quality. The high informative score observed in MUSE and PEARL reflects a characteristic advantage of LLM-synthesized datasets. The strong P-C Correlation of MUSE confirms its suitability for recommendation tasks. Regarding I-T Correspondence, MUSE outperforms the typical multimodal dialogue dataset MMCONV (Liao et al., 2021), attributed to its richer integration of image-explanatory elements.

4.3 Utterance-level Evaluation

In order to evaluate quality at the utterance-level, we randomly select conversation contexts from MUSE and instruct annotators to generate responses based on these contexts. This process created "artificial utterances" for direct comparison with the original ones within the same contexts. Noting that by holding the same contexts and focusing solely on the quality of single-utterance responses, potential biases are significantly reduced. Therefore, we choose to apply manual judgment to perform an utterance-level evaluation to assess the quality of conversations in MUSE from more perspectives. Specifically, annotators are presented with masked paired responses, informed of the given context, and asked to perform anonymous 1:1 comparisons to determine which response is better across four perspectives: Logical, Informative, Natural, and Coherence (Context Coherence). Figure. 5 demonstrates that original dialogues in MUSE are superior quality compared to human-authored dialogues. Our observations suggest that this disparity arises from the LLM's ability to thoroughly interpret and present product attributes. All annotators are graduate students from our university with expertise in conversational recommendation tasks. We provide detailed task descriptions and a fair, anonymous evaluation environment as Appendix D.5. Three of them are responsible for utterance generation, while the other three conduct anonymous 1:1 evaluations.

4.4 Evaluation on Conversational Recommendation Task: Recommend

We investigate the applicability of MUSE for recommendation tasks by conducting recommendation experiments with MUSE on three open-source

Table 4: Recommendation performance of different models under zero-shot and fine-tuning settings

Setting	Recall@10	Recall@20	MRR@10	MRR@20
LlaVA-NEXT-Llama3-8B				
Zero-Shot	0.16	0.25	0.07	0.09
Finetune	0.25	0.37	0.13	0.16
Yi-VL-6B				
Zero-Shot	0.15	0.23	0.07	0.08
Finetune	0.25	0.35	0.12	0.14
Qwen-2-VL-7B				
Zero-Shot	0.20	0.33	0.12	0.13
Finetune	**0.34**	**0.45**	**0.22**	**0.24**

MLLMs: `Qwen2-VL-7B-Instruct`(Wang et al., 2024), `LlaVA-Next-LLaMA-8B`(Li et al., 2024), and `Yi-VL-6B`(Young et al., 2024). Models are tasked with generating queries based on multimodal contexts to recall items for recommendation in the current round. Then we use recall@n and mrr@n to evaluate the accuracy. The fine-tuned models underwent Low-rank adaptation Finetune (LoRA) (Hu et al., 2021) using 200 conversations as the setting in (Liang et al., 2024), with actual queries serving as the golden responses for training. Table. 4 contrasts the performance metrics. The consistent performance gains across all models validate our dataset's effectiveness, supporting previous findings (Bao et al., 2023b,a) that suggest LLMs require fine-tuning for recommendation tasks. Notably, the relative performance ranking among the three models remained consistent across both two settings, aligning with their respective rankings on OpenCompass's MMBench (Liu et al., 2025). This consistency validates our dataset's internal coherence of recommendation logic and demonstrates its discriminative power in differentiating model capabilities.

4.5 Evaluation on Conversational Recommendation Task: Response

Although LLMs excel at general conversation, their response efficacy as specialized rec-assistants warrants investigation. We evaluate three MLLMs in Section 4.4, comparing their performance before and after fine-tuning. The fine-tuning protocol implements a dual-mode approach: (1) For recommendation rounds, the model receives both contextual information and specific multimodal details of the product to be recommended in the current round to generate an appropriate recommendation response. (2) For standard conversation rounds, the model generates responses based solely on the dialogue context. As illustrated in Table. 5, fine-tuning significantly improved the alignment between LLM responses and our dataset patterns.

Table 5: Response performance of different models under zero-shot and fine-tuning settings (p-value).

Setting	BLEU-4	ROUGE-1	ROUGE-L	Distinct-4	Avg. Words
LlaVA-NEXT-Llama3-8B					
Zero-Shot	17.1	17.2	9.61	0.63	86.6
Finetune	44.0	37.8	27.2	0.64	52.6
Yi-VL-6B					
Zero-Shot	16.5	16.7	9.48	0.58	107.7
Finetune	44.1	37.7	27.1	0.63	54.9
Qwen-2-VL-7B					
Zero-Shot	41.1	35.3	23.5	0.69	72.3
Finetune	**46.8**	**42.7**	**31.5**	**0.71**	**48.7**

Table 6: Comparison between responses from zero-shot and finetuned models

	Zero-Shot	Finetune
Win Ratio	0.12	0.88

This improvement demonstrates two key findings: first, our dataset contains learnable response patterns for rec-assistant; second, the enhanced response diversity indicates that LLMs can generate more specific and varied outputs for conversational recommendation after training.

Additionally, we randomly sample 100 pairs of zero-shot and finetuned responses from different models and combine them for anonymous manual evaluation as Appendix D.5. As shown in the Table.6, responses generated from fine-tuned models are more favored by users. Interview feedback indicated that evaluators generally find fine-tuned responses better capture users' key interests.

5 Conclusion

Existing conversational recommendation (CR) research focuses solely on text, leaving a gap with real-world applications. MUSE, the first multimodal conversational recommendation dataset with 7,000 Clothing-related conversations, is introduced to bridge the gap. Validated by LLMs and humans, the conversations in MUSE are shown to be highly informative, fluent, and logically coherent. Through benchmark testing on several multimodal LLMs (MLLMs), we demonstrate that MUSE exhibits reliable recommendation and response logic, making it a valuable resource for CR research. The conversations in MUSE are automatically generated using a multi-agent framework powered by MLLMs, which leverages a scenario-grounded approach to create user profiles tailored to specific products and simulate realistic CR conversations. Addressing the scalability limitations of existing CR data synthesis methods, it holds the potential to expand MCR datasets to include a wider range of domains, users, and products in the future.

6 Limitation

The synthesis process of conversations in MUSE relies extensively on the powerful capabilities of multimodal large models (MLLMs). As a result, the data quality is inherently influenced by the model's capabilities. Due to cost constraints, we opt to use `gpt-4o-mini` as the primary model instead of the more powerful but more expensive `gpt-4o`. Similarly, because API calls for image processing are expensive and each conversation synthesis involves reading a large number of images, we are unable to scale the dataset to the size of pure text datasets like PEARL and LLM-Redial. In addition, while MUSE's conversations already include more rounds than some existing datasets and can be further extended based on specific settings, increasing the context length and the number of images can impact the response generation capabilities of LLMs. As a result, we do not pursue the synthesis of CR data with ultra-long dialogue rounds. In the future, we plan to explore prompt compression techniques to address this limitation.

7 Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62272092, No. 62172086).

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2024. Synthetic dialogue dataset generation using llm agents. *arXiv preprint arXiv:2401.17461*.

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yancheng Luo, Chong Chen, Fuli Feng, and Qi Tian. 2023a. A bi-step grounding paradigm for large language models in recommendation systems. *arXiv preprint arXiv:2308.08434*.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023b. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1007–1014.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Paul A Crook, Satwik Kottur, Seungwhan Moon, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2021. Situated interactive multimodal conversations (simmc) track at dstc9. In *AAAI DSTC9 Workshop*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1431–1441.

Michael Desmond, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M Johnson. 2024. Evalullm: Llm assisted evaluation of generative outputs. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 30–32.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. pages 1679–1705, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zuohui Fu, Yikun Xian, Yongfeng Zhang, and Yi Zhang. 2020. Tutorial on conversational recommendation systems. In *Proceedings of the 14th ACM conference on recommender systems*, pages 751–753.

Elizabeth Liz Harding, Jarno J Vanto, Reece Clark, L Hannah Ji, and Sara C Ainsworth. 2019. Understanding the scope and impact of the california consumer privacy act of 2018. *Journal of Data Protection & Privacy*, 2(3):234–253.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. *arXiv preprint arXiv:2009.14306*.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Mahdi Jalili, Sajad Ahmadian, Maliheh Izadi, Parham Moradi, and Mostafa Salehi. 2018. Evaluating collaborative filtering recommender algorithms: a survey. *IEEE access*, 6:74003–74024.

Dietmar Jannach and Li Chen. 2022. Conversational recommendation: A grand ai challenge. *AI Magazine*, 43(2):151–163.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Minjin Kim, Minju Kim, Hana Kim, Beong-woo Kwak, SeongKu Kang, Youngjae Yu, Jinyoung Yeo, and Dongha Lee. 2024. Pearl: A review-driven persona-knowledge grounded conversational recommendation dataset. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1105–1120. Association for Computational Linguistics.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*.

Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. 2021. Augpt: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. *arXiv preprint arXiv:2102.05126*.

Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. 2008. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, pages 208–211.

Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2020. Conversational recommendation: Formulation, methods, and evaluation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2425–2428.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9748–9758.

Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. 2024. LLM-REDIAL: A large-scale dataset for conversational recommender systems created from user behaviors with llms. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 8926–8939. Association for Computational Linguistics.

Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. Mmconv: An environment for multimodal conversational search across multiple domains. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 675–684. ACM.

Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten De Rijke. 2019. Explainable outfit recommendation with joint outfit matching and comment generation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1502–1516.

Yongkang Liu, Shi Feng, Daling Wang, Yifei Zhang, and Hinrich Schütze. 2023. Evaluate what you can't evaluate: Unassessable quality for generated response. *arXiv preprint arXiv:2305.14658*.

Yongkang Liu, Ercong Nie, Shi Feng, Zheng Hua, Zifeng Ding, Daling Wang, Yifei Zhang, and Hinrich Schütze. 2024. A unified data augmentation framework for low-resource multi-domain dialogue generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 162–177. Springer.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Durecdial 2.0: A bilingual parallel corpus for conversational recommendation. *arXiv preprint arXiv:2109.08877*.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1036–1049. Association for Computational Linguistics.

Yuxing Long, Binyuan Hui, Caixia Yuan, Fei Huang, Yongbin Li, and Xiaojie Wang. 2023. Multimodal recommendation dialog with subjective preference: A new challenge and benchmark. *arXiv preprint arXiv:2305.18212*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 845–854. Association for Computational Linguistics.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Multi-Linguality Multi-Functionality Multi-Granularity. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Justin Paul, Ashwin Modi, and Jayesh Patel. 2016. Predicting green product consumption using theory of planned behavior and reasoned action. *Journal of retailing and consumer services*, 29:123–134.

Žaneta Piligrimienė, Andželika Žukauskaitė, Hubert Korzilius, Jūratė Banytė, and Aistė Dovalienė. 2020. Internal and external determinants of consumer engagement in sustainable consumption. *Sustainability*, 12(4):1349.

Mathieu Ravaut, Hao Zhang, Lu Xu, Aixin Sun, and Yong Liu. 2024. Parameter-efficient conversational recommender system as a language processing task. *arXiv preprint arXiv:2401.14194*.

Protection Regulation. 2016. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679:2016.

Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 696–704. AAAI Press.

Gaurav Sahu, Pau Rodriguez, Issam H Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. *arXiv preprint arXiv:2204.01959*.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Cong Wang, Yifeng Zheng, Jinghua Jiang, and Kui Ren. 2018. Toward privacy-preserving personalized recommendation services. *Engineering*, 4(1):21–28.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, et al. 2023a. When large language model based agent meets user behavior analysis: A novel user simulation paradigm. *arXiv preprint arXiv:2306.02552*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Shoujin Wang, Longbing Cao, Yan Wang, Quan Z Sheng, Mehmet A Orgun, and Defu Lian. 2021. A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)*, 54(7):1–38.

Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z Sheng, and Mehmet Orgun. 2019. Sequential recommender systems: challenges, progress and prospects. *arXiv preprint arXiv:2001.04830*.

Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023b. Rethinking the evaluation for conversational recommendation in the era of large language models. *arXiv preprint arXiv:2305.13112*.

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022a. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1929–1937.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Hu Xu, Seungwhan Moon, Honglei Liu, Bing Liu, Pararth Shah, and Philip S. Yu. 2020. User memory reasoning for conversational recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5288–5308. International Committee on Computational Linguistics.

Yiyan Xu, Wenjie Wang, Fuli Feng, Yunshan Ma, Jizhi Zhang, and Xiangnan He. 2024. Difashion: Towards personalized outfit generation. *arXiv preprint arXiv:2402.17279*.

Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the long tail recommendation. *arXiv preprint arXiv:1205.6700*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, pages 1807–1817.

Yiming Zhang, Lingfei Wu, Qi Shen, Yitong Pang, Zhihua Wei, Fangli Xu, Bo Long, and Jian Pei. 2022. Multiple choice questions based multi-interest policy learning for conversational recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 2153–2162.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473*.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4128–4139. International Committee on Computational Linguistics.

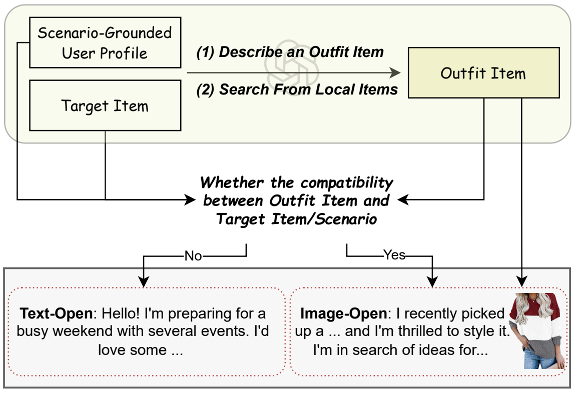

Figure 6: Generation of multimodal-open Open-Dialogues.

A Implementation Details

A.1 Data Preprocess

During data preprocessing, our primary task is to build a local database using the multimodal information of the products, which supports subsequent product retrieval. The local product database contains a subset of the Amazon Cloth, Shoes, and Jewelry products. The process is as follows. Due to the extensive volume of products in the initial dataset, we strategically reduce the product count and eliminate items with incomplete multimodal information to yield a refined dataset comprising 94,209 products. In the subsequent phase, we construct a local product database utilizing both the visual and text information of the product. Initially, we utilize MLLMs (`gpt-4o-mini`) to summarize the product's multimodal data, eliminating redundant details, marketing language, and other noise. This process ensures that the revised summary focuses on the fundamental and visual attributes of the product itself. Following this, we use the summary of products to establish a local product database with BGE-M3(Multi-Granularity, 2024). Alternatively, multimodal embedding models can be employed, which hold the advantage of retaining more of the original information; however, they also present the risk of introducing additional noise. Finally, 13,754 products in the local product database are used for the synthesis of MUSE data.

A.2 Product Distribution

To gain a deeper understanding of our dataset's composition, we visualize all the mentioned product types, as shown in the Figure. 7. On the right side of the figure, we display the proportions of the three main categories classified at the broadest

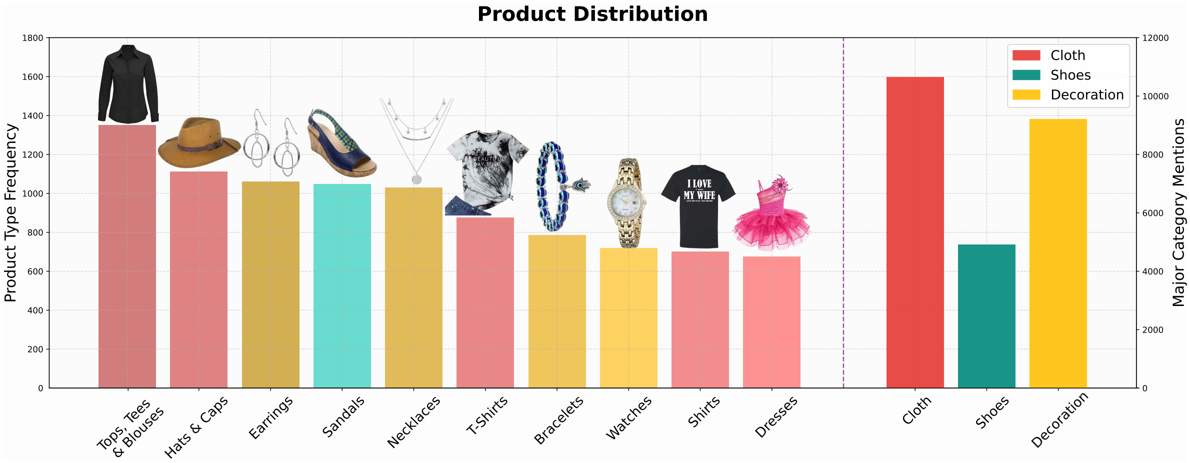

Figure 7: Distribution of mentioned products in MUSE.

level: Clothing, Shoes, and Decoration. (Note: The original dataset labels the third category as Jewelry, but we rename it Decoration after observing that items like wallets are also grouped into this category.) Meanwhile, the left side of the figure illustrates the frequency of the top-10 product types.

A.3 Details of the Open-dialogues

Our dataset incorporates two distinct approaches for conversation initiation. The first approach follows traditional crowdsourcing conversational recommendation systems, where users articulate their needs through text, which is ’text-open.‘ To integrate multimodal elements, we introduce a second approach that allows users to express their requirements through images, which is ’multimodal-open‘. Given the characteristics of our local product clothing dataset, we specifically align multimodal-open with the outfit problem. We frame an multimodal-open case as follows: a user who has already purchased an item of clothing (outfit item) seeks recommendations for a complementary item (target item) that both coordinates with the outfit item and aligns with the scenario. In this case, the user not only articulates the requirements verbally but also shares an image of the outfit item with the rec-assistant to ensure visual compatibility.

The workflow of the multimodal-open case can be seen in Figure. 6. Specifically, an MLLM (`gpt-4o-mini`) is employed to analyze outfit requirements based on the user’s target product/scenario, automatically generating descriptions of potential matching items. These descriptions are served as queries to search our product database. Since the retrieved local product may not perfectly match these generated descriptions - a commom occurence - the model performs an additional compatibility assessment. The secondary evaluation determines whether the retrieved local products can effectively coordinate with the user’s target item while fulfilling the intended scenario requirements. When the secondary evaluation is successful (with an approximate success rate of 47%), the conversation adopts an multimodal-open format, in which users express both their scenario-related requirements and outfit-item-driven needs, allowing the rec-assistant to identify suitable recommendations.

A.4 Architecture of Rewriter

During the simulated conversation generation, we employ a low-temperature setting for the MLLM to maintain process stability, though this constrains the diversity of expressions and sentence structures. To address this limitation, we develop a Rewriter system, as illustrated in Figure. 8. The system operates in two phases: first, using the LLM at high temperature to restructure and rephrase the original conversation, then supervising the output at low temperature to preserve semantic consistency with the source dialogue. Additionally, we probabilistically inject prompts that encourage the use of colloquial expressions, resulting in more authentic, conversation-like exchanges. This approach effectively brings diversity to sentences and keeps semantic fidelity.

A.5 Multimodal Large Language Model Settings for Synthesizing Data

Our data synthesis methodology fundamentally leverages the advanced capabilities of LLMs, necessitating the selection of a robust and reliable model as our foundation. Given our

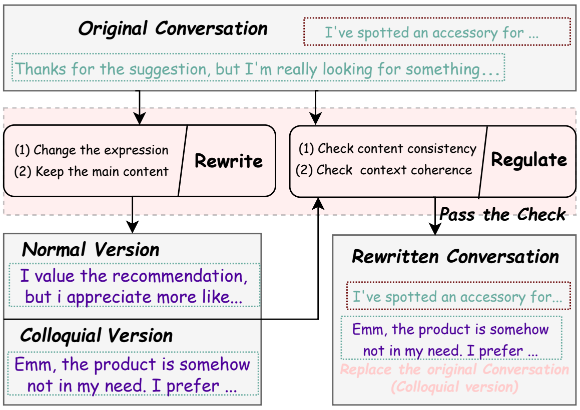

Figure 8: The basic workflow of Rewriter.

computational resource constraints, we adopt an API-based deployment strategy. After evaluating the performance-cost trade-offs, we select `gpt-4o-mini` as our primary Multi-modal Large Language Model (MLLM) foundation. Within our multi-agent framework for Scenario-Grounded User Profile Generator and Simulated Conversation Generator, we set the temperature parameter of `gpt-4o-mini` to 0.1-0.2 for agent role-playing. This low-temperature configuration, coupled with the model's robust instruction-following capabilities, ensures consistent and reliable agent behavior. For the Rewriter, we implement `Claude-3.5-haiku`[†] as the foundational model. This decision stems from our empirical observations that `gpt-4o-mini` exhibited limited syntactic diversity and lexical richness in dialogue generation. Even adjusting the temperature to higher values (0.8-0.9) failed to enhance output variability. And the validation operation in the Rewriter is based on `gpt-4o-mini`.

B Further Analysis

B.1 Discussion on Data Quality

Our framework actually has implemented five automatic screenings to ensure the stability of quality and diversity: (1) BLEU deduplication for basic scenarios. (2) Quality screening for the matching rationality of users, scenarios, and products. (3) BLEU deduplication for the generated purchase backstory. (4) Content consistency screening for the rewritten dialogue in the Rewriter. (5) Quality screening for a whole conversation by LLM. Considering the evaluation of the overall reliability of the public dataset, we also use (6) manual

[†]https://claude.ai

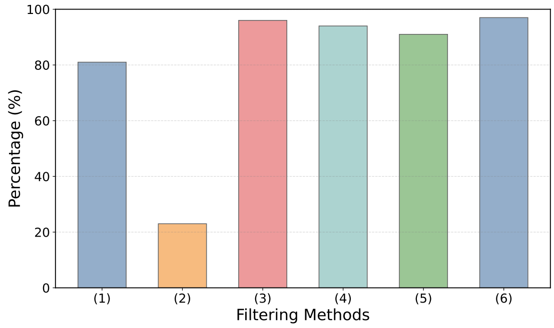

Figure 9: Comparison of pass probability of the six filtering methods.

Data	Muse	Muse (Replace)
Win Rate	0.68	0.32

Table 7: Comparison of Win Rates between Original Muse and Response-replaced Muse

screening, but it turn out that after the previous five screenings, the quality of the retained conversations was already quite guaranteed. Manual screening only filters out a small number of unqualified conversations discussed in the experimental section. Figure. 9 indicates the probability of passing different screening measures.

B.2 Extended Manual Verification

In the main text, we use manual verification to perform utterance-level comparisons, where humans generate responses based on the same context and compare them with the responses in MUSE. To provide more comprehensive manual comparisons, we introduce an additional comparison experiment here. In this experiment, we collect more human-generated replies and replace selected responses in the existing MUSE conversations. Specifically, we randomly select 50 complete conversations from MUSE and, for each conversation, replace 4 responses (two from users and two from the rec-assistant) with human-generated replies to create a conversation-level variant. Another user then judges which conversation—original or modified—is better and smoother overall. The results, shown in the Table. 7, show that even under these replacement conditions, the original conversations in MUSE are still preferred by human evaluators.

B.3 Cost Analysis

The multi-agent framework behind MUSE consists of four steps. The first step focuses on basic scene expansion, utilizing `gpt-4o-mini` to generate approximately 500 diverse real-world scenarios.

Given the concise nature of inputs and outputs in this step, the associated costs are negligible. While employing more advanced models like `gpt-4o` or `claude-3.5-haiku/sonnet` would increase expenses, they could potentially yield more diverse and nuanced scenes. The second step involves user-scenario-product matching to generate scenario-grounded user profiles, also powered by `gpt-4o-mini`. The input consists of a single image and some text and the output is not long. However, due to stringent quality control measures—including user-scenario-product alignment verification and BLEU-based deduplication—only 21.6% of generated user profiles meet the qualification standards. As a result, the average cost for producing a qualified user profile is approximately $0.009, though this can be reduced to $0.001 by utilizing product image descriptions as alternative inputs. The third component implements iterative conversation generation using `gpt-4o-mini`, which requires multiple image/text readings. The generation cost per complete MUSE conversation amounts to roughly $0.033. The final step encompasses conversation rewrite and quality assessment, as detailed in Appendix A.5. For conversation rewriting, we employ `claude-3.5-haiku` with high temperature settings. Since this phase processes plain text for both input and output, and considering the pass rate along with `gpt-4o-mini`'s quality supervision costs, the average expense for refining a complete conversation is $0.011. The dialogue quality review phase evaluates content and outputs scores at $0.008 per review. However, since image-text alignment is already verified during generation, we can optimize costs by screening text only, reducing the expense to $0.001 per review while maintaining quality standards.

B.4 Discussions of N-gram Diversity Patterns

The aforementioned analysis reveals that the Distinct-4 scores are significantly low in both the Pearl and Muse (without the Rewriter). This phenomenon can be attributed to the behavior of GPT-series models operating at low temperatures. While these models demonstrate strong instruction-following capabilities, they tend to generate dialogues with highly repetitive sentence patterns. To provide concrete evidence of this pattern repetition, we conduct a detailed analysis of the most frequent 4-grams in each dataset, which offers a clear visualization of this linguistic homogeneity.

Figure. 10, 11, and 12 present a comprehensive visualization of the 4-gram distribution in each dataset. The left panels display word clouds of the top 200 4-gram phrases, while the right panels show frequency distributions of the top 10 most recurring phrases. The word cloud visualizations clearly demonstrate that the majority of these 4-grams consist of repetitive sentence structures and fixed phrasal patterns. It substantiates our hypothesis that the low Distinct-4 scores are a direct result of this limited linguistic variability, as the prevalence of standardized sentence constructions inherently reduces 4-gram diversity. The analysis reveals a notable disparity in phrase frequency distributions. In both Pearl and Muse (without the Rewriter), the top 10 4-gram phrases exhibit disproportionately high frequencies relative to the total utterance count. However, following the application of Rewriter, we observe a substantial reduction in the frequency of these recurring phrases in Muse. This decline in repetitive patterns, coupled with the increased 4-gram diversity shown in Table 1, provides compelling evidence that the Rewriter successfully enhances linguistic variability by generating more diverse and sophisticated sentence structures. Additional diversity metrics are shown in Table. 8.

Metric	Redial	MMCONV	PEARL	Muse*	Muse
Distinct-2	0.09	0.08	0.03	0.06	0.09
Distinct-3	0.27	0.23	0.09	0.19	0.30
Distinct-4	0.46	0.39	0.18	0.37	0.54
2-gram Specificity	8.63	6.10	1.81	5.62	8.09
3-gram Specificity	24.30	17.50	6.77	18.44	27.20
4-gram Specificity	41.10	29.90	14.75	35.06	48.90

Table 8: Lexical diversity metrics across different conversational datasets. Muse* stands for Muse (without Rewriter).

B.5 Discussion of the Necessity of Images

While Muse presents both visual and textual product information in all product-related conversations, we investigate whether state-of-the-art multimodal models' caption capabilities could effectively convert visual information into textual descriptions for direct integration into conversations. We select 200 conversations, preserving the original textual components and substituting the visual elements with image descriptions generated by `gpt-4o`. We conduct an A/B test comparing human perception of two datasets: the original MUSE data (containing images) and the pure text version (with image descriptions). The results in Table. 9 decisively favor the original MUSE data with images, except

Data	Muse	Muse (Text)
Win Rate	0.97	0.03

Table 9: Comparison of Win Rates between Original Muse and Text-only Muse

in cases where discussions focus solely on product attributes. This preference can be attributed to two factors: first, humans process visual information more efficiently than text descriptions; second, even gpt-4o's image descriptions occasionally contain style interpretation inaccuracies. Additionally, text containing numerous visual features significantly increases dialogue length, making it more challenging for users to identify key information quickly.

B.6 Discussion about the Differences between Muse and Existing Multimodal Datasets

We further elaborate on the differences between MUSE and existing multimodal datasets, mainly about SURE and SIMMC (Long et al., 2023; Crook et al., 2021; Kottur et al., 2021). These datasets constrain conversations to specific VR scenarios and focus on multimodal interactions centered around recommendation tasks. They primarily emphasize spatial relationships in conversations, such as "Please introduce the red clothes just above the jeans," while providing limited content related to user preferences. Additionally, the products discussed in these conversations are restricted to a small selection available within the VR environment, and these products often lack detailed descriptions. The "conversational recommendation task" described in this paper refers to a conversation driven by user needs and interests. The goal is to identify products that meet user requirements from a large pool of options, based on the interest feedback provided by the user. Each product is accompanied by unique descriptions, including both images and text. This setup aligns with the task definition of traditional recommendation systems, such as collaborative filtering and sequential recommendation tasks, where the objective is to accurately identify products that match user preferences from massive datasets. The task scenario in MUSE mirrors real-world situations, such as engaging in a conversation with customer service while shopping online. Therefore, fundamental differences exist between datasets like MUSE and SURE-type VR datasets. MUSE aligns more closely with the definition of conversational recommendation tasks as outlined in existing datasets (Li et al., 2018; Zhou et al., 2020).

C Case Study

C.1 Scenario-Grounded User Profile Case

As demonstrated in Figure. 13, we present two scenario-grounded user profiles that exemplify how individual needs seamlessly align with specific scenarios. The inherent connection justifies our decision to incorporate real-world scenarios into the personality generation process.

C.2 Conversation Case

Figure. 14 and Figure. 15 illustrate two distinct conversational interactions within Muse, each corresponding to the scenarios described above. Figure. 14 demonstrates a dialogue initiated through the multimodal-open mechanism, while Figure. 15 displays a text-open interaction that includes chit-chat. The conversation content features comprehensive explainable recommendation factors, encompassing both contextual scene requirements and visual matching criteria.

D Prompt Template and Evaluation Setting

Compared with Pearl and LLM-Redial, our model incorporates more prompt templates for stricter process control. We present three important prompts here. Including scenario-grounded user profile generation, user-side and recommendation assistant-side prompts and the conversation-level evaluation. For more information, please check https://anonymous.4open.science/r/Muse-0086.

D.1 Scenario-Grounded User Profile Generator

The steps of the Scenario-Grounded User Profile Generator are divided into two steps as in Figure. 16. The first step is to screen the matched basic scenario, user basic profile, and target product. The second step is to generate the backstory of the user to demonstrate his/her preferences.

D.2 Actions of User

Users exhibit two primary actions: accepting a recommendation and rejecting a recommendation, with their corresponding prompt templates illustrated in Figure. 17. Additionally, users demonstrate two secondary action patterns: chit-chatting and comparing item (find reason for rejecting the item).

D.3 Actions of Rec-assistant

The recommendation assistant comprises two components: a Chatter and a Querier. Figure. 18 presents the prompts for two essential functions: the Querier's generation of basic product queries and the Chatter's recommendation process.

D.4 Setting for Conversation-level Evaluation

Here at Figure. 19, we present the evaluation prompt designed to assess conversation-level quality. The prompt instructs the large language model to conduct multi-dimensional scoring, providing detailed scoring criteria for each dimension with specific performance benchmarks. The quality of a multimodal conversational recommendation dataset can be effectively evaluated using the following five aspects. Dialogue Naturalness (Natural) measures how fluent and human-like the conversations are, ensuring realistic and engaging interactions. Logical Coherence (Logical) assesses whether responses align logically with previous dialogue turns, maintaining contextual consistency. Information Richness (Informative) evaluates the diversity and relevance of details provided, which is critical for generating meaningful and helpful recommendations. Product-Context Relevance (P-C Correlation) examines the alignment between recommended products and specific user scenarios, ensuring personalized and context-aware suggestions. Finally, Image-Text Alignment (I-T Correspondence) focuses on the consistency between visual and textual information, crucial for leveraging multimodal data effectively. These five aspects comprehensively cover linguistic, contextual, and multimodal dimensions, ensuring the dataset supports realistic, relevant, and high-quality conversational models. We employ multiple independent scoring rounds and average the scores to ensure reliability.

D.5 Setting for Manual Evaluation

For dialogue quality evaluation at the utterance-level, we employ a two-phase approach as shown in Figure. 20 and Figure. 21. In the first phase, we recruit annotators to generate artificial dialogues based on clear task instructions without any external prompts or interventions. In the second phase, we compile a test set containing the dialogue context, newly collected responses, and original dialogues from Muse. These are randomly shuffled to ensure unbiased assessment. We then conduct blind A/B testing with a separate group of annotators to get final results.

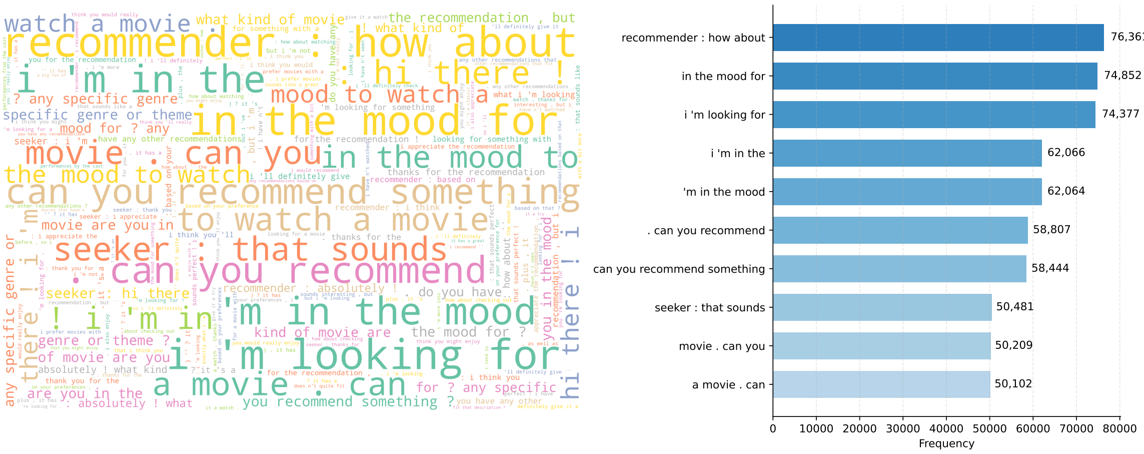

Figure 10: Word-cloud of top-200 4-grams in Pearl (left) and top-10 4-gram's frequency in Pearl (Right)

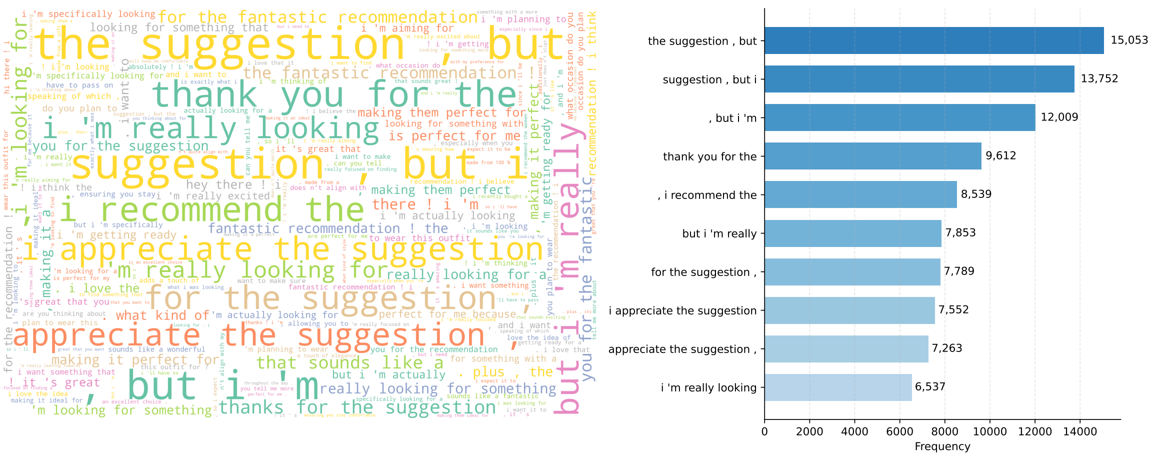

Figure 11: Word-cloud of top-200 4-grams in Muse without Rewriter (left) and top-10 4-gram's frequency in Muse without Rewriter (Right)

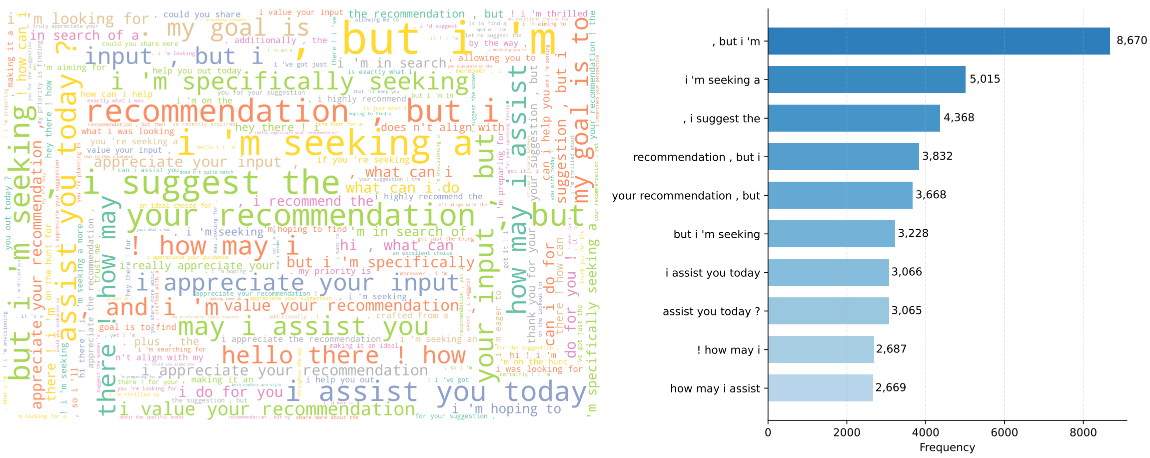

Figure 12: Word-cloud of top-200 4-grams in Muse with Rewriter (left) and top-10 4-gram's frequency in Muse with Rewriter (Right)

"**Basic Information**": "'Name: Leslie Robinson, Gender: female, Age: 35,Profession: Prison officer",

"**Purchase Backstory**": "Leslie Robinson sat at her kitchen table, sipping her morning coffee while scrolling through her phone. The holiday season was fast approaching, and she had promised herself to find the perfect gifts for her friends and family this year. As a prison officer, her days were often long and stressful, so she wanted to choose gifts that would bring joy and comfort to those she cared about. She remembered how her best friend, Sarah, had mentioned her love for cozy and festive clothing during their last coffee date. Sarah always had a knack for making the holidays feel special, and Leslie wanted to return the favor. Leslie thought about how much Sarah enjoyed dressing up for holiday gatherings, but also how she appreciated comfort during the busy season. She envisioned Sarah wearing something that was not only stylish but also easy to wear throughout the day, especially during family get-togethers. As she continued to browse, Leslie felt a wave of excitement at the thought of finding something that would make Sarah smile. She wanted to ensure that whatever she chose would fit well and allow for easy movement, as Sarah loved to dance and play with her kids during the holidays. With a determined spirit, Leslie began her online search, eager to find that perfect gift that would blend festive cheer with comfort and practicality.",

"**Target Item**": Title: Two Left Feet Women's Holiday Leggings. ...

"**Basic Information**": "Name: Timothy Barnett; Gender: male; Age: 40; Profession: Chief Strategy Officer"

"**Purchase Backstory**": "Timothy Barnett was buzzing with excitement as the date of the book launch approached. As the Chief Strategy Officer of a thriving start-up, he often mingled with influential authors and entrepreneurs at such events. This time, however, it was personal. The book was authored by his mentor, and he wanted to make a lasting impression while also supporting someone who had shaped his career. In the days leading up to the launch, Timothy found himself pondering how to present himself. He wanted to look professional yet trendy, an embodiment of his modern approach to business. He imagined himself surrounded by peers, exchanging business cards fluidly. However, the thought of his cards getting damaged or possibly even stolen filled him with dread. He needed something that would not only protect his cards but also reflect his stylish persona. \n\nAs he browsed through various online platforms, he stumbled upon a sleek, stainless steel card holder that promised to shield his cards from unwanted scanning. The idea of having RFID protection made him feel more secure, especially knowing that he would be in crowded spaces at the venue. The thought of confidently handing out his cards, knowing they were safeguarded, elevated his anticipation for the event. He could already picture himself networking effortlessly, showcasing both his credentials and his flair for contemporary style. Timothy felt this was exactly what he needed to enhance his experience at the book launch, and he decided it was time to make a purchase."

"**Target Item**":Title: Leopardd RFID Blocking Stainless Steel Card Holder Case (001) . ...

Figure 13: Two cases of the Scenario-Grounded User Profiles

Assistant: Hello! How may I assist you today?

User: I recently picked up a super comfy tunic, and I'm thrilled to style it. I'm in search of ideas for pieces that would enhance its chill vibe.

Assistant: What an exciting idea! A lively pattern would definitely inject some personality into your ensemble, complementing the solid-toned tunic beautifully and creating a visually appealing contrast."

User: Definitely! I'm currently exploring leggings with a festive, eye-catching holiday pattern. These vivid designs would create a stunning contrast when paired with the tunic.

Assistant: That's a delightful pairing! These festive printed leggings will undoubtedly inject some vibrant spirit into your comfortable tunic. Could you share the specific event or setting where you're planning to showcase this stylish ensemble?

User: I'm thinking of surprising Sarah with this festive ensemble for the holidays. My goal is to pick something that strikes a perfect balance between chic and cozy, allowing her to comfortably mingle with family and playfully dance with her little ones. It's crucial that the leggings feature a flexible waistband to ensure she can move around freely without any restrictions.

Assistant: If you are in search of the perfect pair to complement your tunic, the VIV Collection Plus Size Print Fashion Leggings are an excellent choice! Featuring a whimsical design of festive Christmas trees and ornaments, these leggings add a fun element to your look. Additionally, their flexible waistband ensures comfort, enabling your friend to enjoy holiday festivities and family gatherings with style and ease.

User: I see what you're getting at, but the festive design doesn't quite hit the mark for me. What I'm after are the Leggings with fun, adaptable vibe that would work perfectly for my friend's wardrobe, no matter the season.

Assistant: For a highly adaptable choice that will enhance your friend's cozy tunic, I suggest the Leggings Mania Regular/Plus Solid Buttery Soft High Waist Leggings. With their incredibly smooth texture and flexible waistband, these leggings provide exceptional comfort and mobility, ideal for family events and casual daily activities. Their versatile solid color enables effortless mixing and matching across various seasons!

User: Thanks for the suggestion, but I need to pass on those leggings since they don't provide any material details, which makes it hard to evaluate their quality. I need Leggings made from 100% polyester, really stand out for their superior softness and comfort, which is essential for what I'm looking for.

Assistant: Hey there! I've got a killer recommendation for your friend's wardrobe - the Two Left Feet Women's Holiday Leggings. These bad boys are super versatile and totally fun! Crafted from pure polyester, they rock a high-waist cut with this adorable quirky design that'll totally jazz up her tunic. The elastic waistband means maximum comfort, and trust me, they're soft as can be. Whether it's a family get-together or just a casual day out, these leggings will be her go-to piece all year long!

User: Thank you for the recommendation! I believe the Two Left Feet Women's Holiday Leggings are perfect for my friend. The quirky design adds a fun touch, and the 100% polyester material ensures they will be soft and comfortable for family gatherings and everyday wear. Plus, the high-waist and elastic waistband will provide the flexibility she needs to move around easily.

Figure 14: Text-open conversation case.

Assistant: Greetings! What assistance can I provide for you today?

User: I'm truly enthusiastic about my mentor's imminent book release and eager to connect with industry professionals. However, I'm concerned about potential risks to my professional networking materials, specifically the potential damage or theft of my business cards during the event.

Assistant: What an exciting idea! A lively pattern would definitely inject some personality into your ensemble, complementing the solid-toned tunic beautifully and creating a visually appealing contrast.

User: I'm seeking a business card design that effectively blends professional sophistication with contemporary style. My goal is to make a strong impression and feel self-assured when networking in a bustling environment.

Assistant: Let me suggest a sophisticated leather business card wallet that will elevate your professional image. This premium cowhide accessory offers ample storage for over 50 business cards, complemented by a practical ID compartment and extra slots for credit cards and cash. Its sleek design ensures your networking materials remain immaculate and easily accessible during your professional engagements."

User: I value your recommendation, but my preference leans towards an RFID-blocking card wallet. The aesthetic matters significantly to me - I'm seeking a slim, metallic accessory that not only shields my cards from potential electronic theft but also complements my style for the upcoming event.

Assistant: For the networking event ahead, I suggest the HAPTIME RFID Blocking Credit Card Holder. This elegant stainless steel accessory offers a chic appearance while safeguarding your cards against potential electronic theft. Its compact form factor allows you to carry multiple cards effortlessly, making it an ideal companion for professional networking situations. The integrated RFID blocking technology provides added security, enabling you to focus on making meaningful connections without worrying about card safety.

User: I'm captivated by the prospect of a streamlined, metallic card holder design! The memory of a tech conference floods back, where professionals flaunted their distinctive accessories that were both stylish and functional. It's fascinating how a compact item like a card holder can significantly enhance one's professional image during networking events. The sleek metallic aesthetic speaks volumes about personal branding, and the added RFID protection is a brilliant feature that offers peace of mind in bustling, crowded environments where security is paramount. Such an accessory isn't just about storage, it's a statement of sophistication and practicality.

Assistant: Absolutely, tech conferences are indeed a hub of creativity and style! The vibrant environment allows your individual flair to stand out while ensuring your belongings are safe. The HAPTIME RFID Blocking Credit Card Holder would perfectly enhance your outfit, providing both elegance and protection ideal for making a strong impression at your upcoming book launch without sacrificing security.

User: While I appreciate the recommendation, I'm specifically seeking an item with a contemporary and trendy design. The aesthetic appeal is crucial for me, and the suggested product's standard black appearance doesn't align with the sophisticated style I'm aiming for at the event.

Assistant: I recommend the Leopardd RFID Blocking Stainless Steel Card Holder Case for those seeking a contemporary and chic accessory. This sophisticated case boasts a streamlined, elegant appearance that elevates your personal style while offering top-tier protection against digital theft. Its sturdy build and convenient mechanism make it an ideal companion for professional gatherings, effectively safeguarding your cards from potential electronic scanning in busy settings.

User: I'm thrilled that the Leopardd RFID Blocking Stainless Steel Card Holder resonates so perfectly with my aesthetic and practical needs. Its elegant, streamlined profile not only complements my personal style but also offers robust security for my cards. The blend of contemporary design and utility makes it an exceptional choice for professional events like my mentor's book launch, ensuring I look sophisticated while keeping my cards well-protected.

Figure 15: Multimodal-open conversation case.

Scenario-User-Item Match

System Instruction:

Given a user's basic information, a scenario, and a product's text/image information.
Please judge:
1. Whether the product can match the user's basici information.
2. Whether the product can match the Scenario.

Output format:
1. If both criteria are satisfied: output 'Yes'.
2. If either criterion is not met: output 'No'.
3. No other output is permitted.

User Input:

Basic User Profile: **{user_profile}**
Scenario: **{scenario}**
Target Item: **<Image> {item_description}**

Purchase Backstory Generation

System Instruction:

You are scenarios generator for a consumer's purchase backstory.
Your goal is to create a scenario that could naturally lead to a online purchase.
without explicitly mentioning the actual item bought.
Remember, the scenario happened before the user's online purchase!
You will be given three information:
1. User information. 2. Basic Scenario for the backstory. 3. Information of the target item.
Use the provided information to craft a believable and engaging narrative.
The descriptions should:
1. An upcoming event (It can be a significant life event or a minor everyday occurrence.)
2. Include relevant contextual details such as scenarios, emotions.
3. Reveal the user's requirements for the product, but only towards two features.
4. Related to the revealed requirements.
Warning: Do not mention or describe the target item.
Please generate a backstory suits the case.

User Input:

Basic User Profile: **{user_profile}**
Scenario: **{scenario}**
Target Item: **<Image> {item_description}**

Figure 16: Some prompts of Scenario-Grounded driven User Profile Generator.

Accept

System Instruction:

You are now playing the role of a user engaged in a conversation with a rec-assistant. During this conversation, the rec-assistant successfully found a product that meets your needs.
Given two information:
1. The content of your conversation with the rec-assistant.
2. The image and detailed information of the product you finally accepted.
Based on this information, please response with a concluding statement expressing your feelings about the recommendation experience and your thoughts on the final product choice.
Your response should:
1. Ensure that you are speaking as the user.
2. Express gratitude for the recommendation system.
3. Explain why you think this product suits you.
4. Mention specific features of the product and relate them to your personal preferences.
Please ensure your response is natural and authentic.
The length of your response should be between 1-3 sentences.
Please generate your response.
Output only the personalized response.

User Input:

Conversation Context: **{conversation context}**
Target Item: **<Image> {item_description}**

Reject

System Instruction:

You are roleplaying as a user engaged in a conversation with a dialogue recommender system.
In a previous turn of the conversation (which may not be the immediately preceding turn due to potential chit-chat), the recommender system suggested a product to you.
You have already found a suitable reason to decline this recommendation, and this reason also includes some of your own needs or preferences.
Your task now is to respond to the recommender system, declining the suggested product while expressing your needs in a natural and engaging manner.
You have access to the following information:
1. The content of your conversation with the rec-assistant.
2. The specific reason for declining the recommendation, which includes some of your needs or preferences.
Your response should:
1. Avoid repetitive phrasing of rejections in history conversations! .
2. Decline the recommendation based the reasons you've been provided.
3. Reference relevant parts of the conversation history to maintain context and continuity.
4. The length is limited to 1-3 sentences.
Warning: Don't mention any brand or prize!
Now, based on the given reason for declining (which includes your needs) and the conversation history.
Craft a response that declines the current recommendation while expressing your needs and keeping the conversation flowing naturally.
Output only the response.

User Input:

Conversation Context: **{conversation context}**
Reject Reason: **{reject reason}**

Figure 17: Some prompts of User Simulator.

Analyze Requirements

System Instruction:
You are a helpful conversation analyzer.
Given a conversation between a user and a conversational recommendation system.
Based on the conversations, identify and summarize the core requirements and intentions of the user.
Pay special attention to:
1. The user's directly expressed needs or questions
2. The user's implicit needs or areas of interest
3. The user's emotional state and tone
4. The user's feedback on the system's responses
5. Any specific preferences, limitations, or criteria mentioned by the user
Clearly summary the user's main needs for the products.
Output only the summary.

User Input:
Conversation Context: **{conversation}**

Recmmend

System Instruction:
You are an advanced conversational recommendation assistant.
You are talking to a user and your task now is to recommend a product to the user.
Given two information:
1. The conversation context.
2. The image/text information of the product you need to recommend.
Your objective is to craft a product recommendation response that:
1. Seamlessly fits into the current conversation flow.
2. Demonstrates understanding of the user's preferences, needs, and previous interactions.
3. Highlights the most relevant features of the product based on what you know about the user.
4. Incorporates relevant details from the product's text description and visual elements.
5. Aiming for a length of 2-3 sentences.
Output only the the recommendation sentence for the product!

User Input:
Conversation Context: **{conversation}**
Recommended Item: **<Image> {item_description}**

Figure 18: Some prompts of Rec-assistant Simulator.

Score Conversation-level Quality

System Instruction:
You are a conversation score assistant.
Given a conversation.
Please score the conversation from four aspects,
scoring from 0 to 2 points (using 0.05 increments):

Aspect 1. Dialogue Naturalness (Natural)
2 score: Highly natural conversation flow with appropriate turn-taking and human-like expressions
1 score: Moderately natural with occasional awkward exchanges
0 score: Unnatural, robotic, or disjointed conversation \n

Aspect 2. Logical Coherence (Logical)
2 score: Perfect logical flow, each response directly relates to previous context
1 score: Generally coherent with minor logical gaps
0 score: Significant logical breaks or irrelevant responses \n

Aspect 3. Information Richness
2 score: Rich, detailed information with specific examples or explanations.
1 score: Basic information provided but lacks depth
0 score: Minimal or irrelevant information \n

Aspect 4. Product-Context Relevance
2 score: Discussion closely tied to product features and context
1 score: Partially relevant to product context
0 score: Little to no connection to product context \n
Please output the score of the four aspects with no reason.

#Optinal：
#Aspect 5 Image-Text Alignment (I-T Correspondence)
2.00: Perfect alignment between image content and dialogue
1.00: Moderate alignment with noticeable gaps
0.00: Minimal or no alignment between image and text

User Input:
Conversation: **{conversation}**

Figure 19: The evaluation prompt for LLM-based assessment of conversation-level quality

Utterance-level Evaluation (System Side)

Rules for Human (Generate Response):

For this response task, you will act as a conversational recommendation system for online shopping customer service.
Your task is to generate an appropriate response based on the following information:
1. The conversation context and required response type (recommend or chit-chat)
2. Product details including images and text descriptions (if applicable for recommend rounds)

Please generate your response following these criteria:
1. Analyze the conversation context thoroughly to understand the user's primary needs
2. Based on the response type:
   For recommendations: Highlight key product features and benefits
   For chit-chat: Provide emotional support while naturally redirecting attention to previously recommended items

Conversation Context: **{conversation}** and Action Type: **{action_type}**
Recommended Item: **<Image>{item_description}**

Rules for Human (1:1 Evaluation):

For this evaluation task, you will act as a user interacting with a conversational recommendation system.
You will be presented with:
1. A user profile
2. The existing conversation context between you and the system
3. Two alternative system responses

Please evaluate them following these criteria:
1. Review the user profile thoroughly to understand the user's background and needs
2. Examine the conversation context to identify expressed user requirements
3. Compare both responses and select the more appropriate one across four perspectives:

(1) Logical: The response follows clear reasoning and demonstrates sound judgment.
(2) Informative: The response provides sufficient, useful, and relevant information.
(3) Natural: The response flows smoothly with conversational language and tone.
(4) Coherent: The response maintains consistency with the conversation context.

Please indicate your choice with:
A: for the left response
B: for the right response
Conversation Context: **{conversation}**
User Profile: **{user_profile}**

Figure 20: The evaluation rules for manual evaluation of utterance-level quality

Utterance-level Evaluation (User Side)

Rules for Human (Generate Response):

For this response task, you will act as a user in online shopping and a chat with a conversational recommendation system.
You will be required to generate either a reject response or an accept response. based on the following information.
1. A user profile
2. The conversation context and required response type (reject or accept)
3. Recommended product details including images and text descriptions
4. Target product details including images and text descriptions (for accept)

Please generate your response following these criteria:
1. Analyze the conversation context thoroughly to understand the user's primary needs
2. Based on the response type:

For reject: Craft a response that declines the current recommendation while expressing your needs and keeping the conversation flowing naturally.

For accept: Please response with a concluding statement expressing your feelings about the recommendation experience and your thoughts on the final product choice.

User Profile: **{user_profile}**
Conversation Context: **{conversation}** and Action Type: **{action_type}**
Recommended Item: **<Image>{item_description} (Recommended)**
Target Item : **<Image>{item_description} (Target)**

Rules for Human (1:1 Evaluation):

For this evaluation task, you will act as a bystander, watching a conversation between a user and the system.
You will be presented with:
1. A user profile
2. The existing conversation context between you and the system
3. Two alternative user responses

Please evaluate them following these criteria:
1. Review the user profile thoroughly to understand the user's background and needs
2. Examine the conversation context to identify expressed user requirements
3. Compare both responses and select the more appropriate one across four perspectives:

(1) Logical: The response follows clear reasoning and demonstrates sound judgment.
(2) Informative: The response provides sufficient, useful, and relevant information.
(3) Natural: The response flows smoothly with conversational language and tone.
(4) Coherent: The response maintains consistency with the conversation context.

Please indicate your choice with:
A: for the left response
B: for the right response
Conversation Context: **{conversation}**
User Profile: **{user_profile}**

Figure 21: The evaluation rules for manual evaluation of utterance-level quality