VOGUE: A Multimodal Dataset for Conversational Recommendation in Fashion

David Guo*
University of Toronto
Toronto, Ontario, Canada
davidmy.guo@mail.utoronto.ca

Minqi Sun*
University of Waterloo
Waterloo, Ontario, Canada
maggie.sun@uwaterloo.ca

Yilun Jiang*
University of Waterloo
Waterloo, Ontario, Canada
yilun.jiang@uwaterloo.ca

Jiazhou Liang
University of Toronto
Toronto, Ontario, Canada
joe.liang@mail.utoronto.ca

Scott Sanner
University of Toronto
Toronto, Ontario, Canada
ssanner@mie.utoronto.ca

Abstract

Multimodal conversational recommendation has emerged as a promising paradigm for delivering personalized experiences through natural dialogue enriched by visual and contextual grounding. Yet, current multimodal conversational recommendation datasets remain limited: existing resources either simulate conversations, omit user history, or fail to collect sufficiently detailed feedback, all of which constrain the types of research and evaluation they support.

To address these gaps, we introduce Vogue, a novel dataset of 60 human-human dialogues in realistic fashion shopping scenarios. Each dialogue is paired with a shared visual catalogue, item metadata, user fashion profiles/histories, and post-conversation ratings from both Seekers and Assistants. This design enables rigorous evaluation of conversational inference, including not only alignment between predicted and ground-truth preferences, but also calibration against full rating distributions and comparison with explicit and implicit user satisfaction signals.

Our initial analyses of Vogue reveal distinctive dynamics of visually grounded dialogue, e.g., recommenders frequently recommend items simultaneously in feature-based groups, which creates distinct conversational phases bridged by Seeker critiques and refinements. Benchmarking Multimodal Large Language Models against human Recommenders shows that while MLLMs approach human-level alignment in aggregate, they exhibit systematic distribution errors in reproducing human ratings and struggle to generalize preference inference beyond explicitly discussed items. These findings establish Vogue as both a unique resource for studying multimodal conversational systems and a challenge dataset beyond the current recommendation capabilities of existing top-tier multimodal foundation models such as GPT-4o-mini, GPT-5-mini, and Gemini-2.5-Flash.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Information systems** → **Users and interactive retrieval**; • **Computing methodologies** → *Natural language processing*.

*These three authors contributed equally to this work.

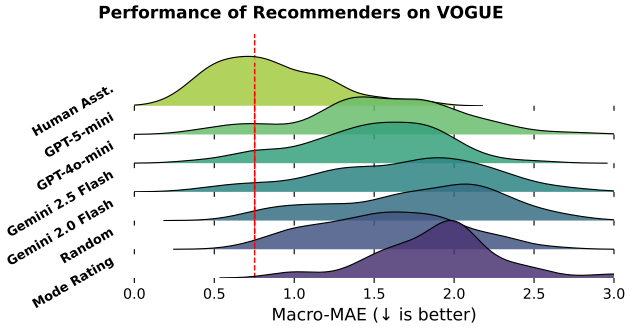


Figure 1: Using Vogue, we evaluate a variety MLLMs on their ability to model user preferences against Human Assistants and two baselines: randomly sampled rating, and mode (i.e. most common) rating. Our results show that state-of-the-art MLLM preference generalization and understanding across the board substantially underperforms human agents.

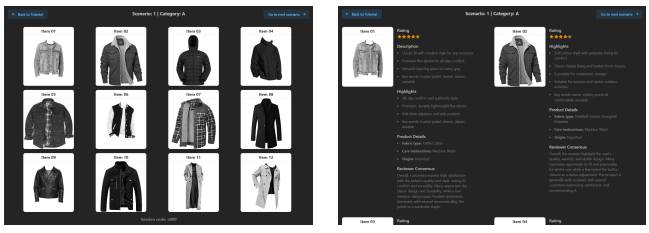
Keywords

Conversational Recommendation, Multimodal Dialogue, Dataset, Human-Computer Interaction, Visual Grounding, Preference Inference, Recommender Systems

Preprint version. This work has been submitted to an ACM venue for review.

1 Introduction

Recommender systems have become a central part of the modern Web, shaping how users navigate, consume, and engage with online content across e-commerce, media, and social platforms. As the Web



(a) Seeker User Interface **(b) Assistant User Interface**

Figure 2: Study User Interfaces used in our experiment. Full-size images are available in the Appendix.

continues to integrate multimodal interaction and human-centered personalization, Conversational Recommender Systems (CRS) have emerged as a expressive paradigm for enabling natural, context-aware user experiences [4, 7]. However, building effective CRS remains a significant challenge, owing to their inherently subjective natures, requisite multimodality & human-like conversational ability, and need for generalizable user preference understanding [5]. Among various domains, fashion shopping particularly exemplifies these challenges: preferences are highly subjective and often hinge on visual style and fashion sense, which are difficult to capture through text alone. These issues compound, making a comprehensive evaluation particularly challenging.

Existing conversational recommendation datasets often fall short in evaluating CRS on such difficulties, including the fashion shopping domain. Most are limited to (1) a single modality (typically text) [9, 11, 21], (2) simulated interactions that rely heavily on synthetic rather than organic human dialogues [12, 17, 19], (3) narrow domains such as film recommendation [9, 18], or (4) omit personal user profiles [2, 6, 8–10, 12, 17]. Additionally, Large Language Models (LLMs) and Multimodal LLMs (MLLMs) now serve as the backbone of many CRSs [20], offering strong capabilities for interpreting free-form dialogue and generating natural responses. However, existing datasets and metrics largely emphasize surface-level outcomes, such as the final recommended item, while neglecting the explicit item-level preference ratings and user satisfaction signals [2, 10, 19]. As a result, they fail to expose the limitations of MLLMs compared to human assistants and hinder robust, system-to-system comparison.

In this work we introduce **Vogue** (**V**isual-recommendation dial**O**gue with **G**rounded **U**ser **E**valuations), a dataset that addresses these gaps by featuring (1) multimodal grounding with both visual and verbal inputs, (2) authentic human–human dialogues in realistic fashion shopping settings (Fig. 2), (3) a broad and open-ended fashion domain, and (4) comprehensive user profiles capturing style history, preferences, and ratings of non-candidate items. Each conversation is visually grounded by a catalog of candidate items (including images and metadata) and annotated with a full post-conversation item ratings from both the Seeker (user) and Assistant (recommender), and subjective satisfaction metrics captured through post-task Likert surveys. This design enables evaluation along two complementary dimensions: item-level preference alignment and subjective satisfaction metrics. Each conversation is also tagged at the utterance level with a novel intent taxonomy specific to the fashion shopping domain. These measures move beyond accuracy alone, supporting richer and more detailed evaluations across a broad range of conversations.

By releasing Vogue, we aim to provide a practical benchmark (cf. Fig. 1) for studying multimodal conversational recommendation. Our dataset establishes an invaluable resource for designing recommendation agents as CRS move into multimodality, providing a foundation to study how dialogue, visual grounding, and user-centered evaluation interact in shaping effective systems.

Our contributions are threefold:

(1) **Open Dataset of Real Human Dialogues:** We introduce the first multimodal conversational recommendation dataset in the fashion shopping domain, encompassing organic human–human dialogues, associated item images and metadata, participant profiles, and detailed rating signals. The dataset provides utterance-level annotations of Assistant and Seeker intents across conversations.

(2) **Novel Analysis and Insights:** We provide a novel analysis for modeling dialogue dynamics, including stage-wise segmentation (Fig. 4) of conversations and intent-tagged utterance flows that reveal how recommendations and preferences evolve over time.

(3) **MLLM Benchmark Evaluation:** We benchmark MLLMs against humans on Vogue, revealing persistent challenges in calibration, generalization, and inference, and positioning it as a challenge dataset for advancing MLLM agents.

2 Related Work

To develop an effective end-to-end learning methodology for multimodal CRS, as well as to establish benchmark datasets for evaluation, studies have contributed meaningful datasets to support recommendation tasks in various domains, as listed in Tab. 1. However, existing conversational recommendation datasets are limited and fall short of meeting the needs of our study as below:

- **Lack of Fashion Datasets:** Most existing conversational recommendation datasets are not about fashion shopping. They usually focus on movies [6, 9, 21], while some cover restaurants, product search/recommendation or travel [1, 2, 10], which are not transferable to our study.
- **Lack of Explicit User Feedback:** The rating included along with the existing datasets serves only as an indicator of conversation quality rather than feedback on recommended items during conversations, as well as overall user satisfaction throughout the whole experience [1, 10], which hinders further evaluation of the recommendation quality based on dialogue data.
- **Lack of User History and Profiles:** Most existing multimodal conversational datasets either lack personal profiles [8, 12, 17], adapt them directly from interaction data [16], or generate them using LLMs [19]. This reflects a lack of a real user profile associated with the speaker. This limits the opportunity to develop stronger elicitation methods, since long-term preferences are missing [7].
- **Lack of Real User Dialogue:** For current multimodal conversational datasets targeting fashion shopping, they curated dialogue datasets using data generation methods (e.g., dialogue simulator, LLMs) to produce synthetic dialogue data [16, 19]. However, the deficiency of realistic conversational behaviors present in the synthetic datasets limits the opportunities to study dialogue patterns.

Thus, we believe there is a need for a multimodal conversational dataset supporting visual grounding, consisting of organic human-human dialogues linked to authentic personal profiles and ratings on recommended items, conversation qualities, and user satisfaction in general.

3 Dataset Construction

Based on previously described deficiencies of related work, we aim to curate a novel dataset that satisfies the following desiderata:

Table 1: We present a comparison of existing conversational datasets. Notably, Vogue introduces text and image multimodality, combined with real user profiles, real human dialogue, and comprehensive preference & user satisfaction feedback. [1] Profile is synthetic, created using scraped online data sources. [2] Profile completed by participant.

Name	Domain	Modality	User Profile	Dialogue Source	Satisfaction Rating	Preference Rating
DuRecDial 2.0 [11]	General	Text Only	Synthetic	Human–Human	✗	✗
ReDial [9]	Film	Text Only	None	Human–Human	✗	✓
TG-ReDial [21]	Film	Text Only	Scraped	Human–Machine	✗	✗
INSPIRED [6]	Film	Text Only	None	Human–Human	✗	✓
CRM [18]	Restaurant	Text Only	Scraped	Synthetic	✗	✗
MMConv [10]	Travel	Text + Real Images	None	Human–Human	✓	✗
MG-ShopDial [2]	Product Rec.	Text + Real Images	None	Human–Human	✗	✗
CoSRec-Curated [1]	Product Rec.	Text Only	Synthetic [1]	Synthetic	✗	✗
SIMMC 2.0 [8]	Fashion, Furniture	Text + Synthetic Images	None	Human–Human	✗	✗
SURE [12]	Fashion, Furniture	Text + Synthetic Images	None	Synthetic	✗	✗
MMD [17]	Fashion	Text + Real Images	None	Synthetic	✗	✗
MUSE [19]	Fashion	Text + Real Images	Synthetic	Synthetic	✗	✗
FashionRec [16]	Fashion	Text + Real Images	Scraped	Synthetic	✗	✗
Vogue	**Fashion**	**Text + Real Images**	**Self-reported [2]**	**Human–Human**	✓	✓

(1) **Multimodal grounding:** Combine text, images, and metadata to situation diaglogues in a realistic setting.
(2) **Fashion domain focus:** Address the lack of existing datasets centered on fashion shopping.
(3) **Explicit user feedback:** Provide comprehensive and quantitative item-level ratings as well as subjective satisfaction.
(4) **User history and profiles:** Include baseline preferences, rationale, and opinions to support personalization.
(5) **Organic human–human dialogues:** Collect real conversations rather than synthetic or simulated ones.

To meet these desiderata, we designed Vogue as six independent conversational recommendation trials between paired participants. We refer to the recommender as the **Assistant** and the user as the **Seeker**. Each trial centered on a contextualized fashion scenario (e.g., shoes for a rainy fall farm visit or a top for a winter office day), giving the Seeker a temporary decision-making goal.

In each scenario, participants shared a catalog of 12 candidate items drawn from Amazon.ca, spanning outerwear, layering pieces, and shoes. Each item included images and collated metadata (name, description, review score, review summary). To reduce cognitive load, descriptions and reviews were GPT-4o-mini summarized, while size, fit, color, and price were omitted, and participants were asked not to anchor on them.

Throughout the study, the catalog comprised 36 items (12 per scenario) plus 40 additional items used for fashion profiles. The Seeker's task was to articulate needs and select an item, while the Assistant guided preference elicitation and made recommendations.

3.1 Protocol

Our data collection protocol[1] proceeded in three steps:

(1) **Onboarding surveys:** All participants completed demographic forms and fashion entrance surveys, including a profile questionnaire and pre-ratings of 40 non-candidate items, which provided baseline preferences and served as user profiles, a feature rarely included in similar datasets. (Desiderata 1, 2, 4). The complete details for all surveys are in the `surveys` folder of the supplementary material[2].
(2) **Conversational trials:** In each of the six scenarios, the Seeker and Assistant discussed 12 candidate items, with the Assistant accessing full metadata and the Seeker seeing only images and the scenario description. Assistants were instructed not to use external resources and to elicit context through dialogue. Conversations continued until the Seeker indicated satisfaction with at least one item or rejected all options. (Desiderata 5). The complete details are in the `item_ratings` subfolder of Vogue[2].
(3) **Item rating surveys:** After each scenario, the Seeker provided ground-truth preference ratings of all 12 items on a 1 to 5 scale, while the Assistant attempted to predict these ratings. Participants were instructed to rate based on how likely they would be to purchase the item given the conversation and scenario, with 5 being "would purchase" and 1 being "would never purchase". Neither party saw the other's ratings. Both parties also completed a short feedback survey on the perceived quality of that conversation. (Desiderata 3)

3.2 Participants

We recruited 20 participants through informal outreach at a local University and among recent graduates in the surrounding area. Young adults were recruited in order to reflect a core demographic of fashion shopping and the adoption of conversational Assistants.

Participants were randomly assigned fixed roles as Seeker or Assistant for the entire session. To prepare Assistants for their role, we conducted a short coaching session, to familiarize them with the catalog of candidate items. Seeker participants received only the task scenarios beforehand to replicate real-world conditions, necessitating that Seekers articulate their needs without access to full product knowledge.

[1]The protocol received ethics approval from the University of Toronto Research Ethics Board (REB).

Table 2: Seeker intent, building on Cai and Chen [3], Lyu et al. [13]. We present them hierarchically, with higher-fidelity tags grouped under broader categories. 3-letter tag codes as used in the data release and analysis are also specified.

Category (Code)	Description (Seeker...)	Example
Ask for Recommendation		
Initial Query (IQ)	...asks for a recommendation in the first query	"Hey, I want to purchase one cloth..."
Continue (CON)	...asks for another recommendation	"...would you be giving me a Plan B instead?"
Provide		
Provide Context (PCT)	...provides background or situational information	"...I want to choose the right outerwear for a fall visit..."
Provide Explicit Preferences (PEP)	...specifies desired features or requirements	"...it has to be waterproof because..."
Provide Implicit Preference (PIP)	...implies preferences via comments or reactions	"Yeah, so I don't like Item11, Item10."
Refine Preference (RP)	...improves over-constrained/under-constrained preferences	"I don't need a hood necessarily..."
Answer (ANS)	...answers a question from the Assistant	"The weather is cool with a chance of light rain."
Acknowledgement (ACK)	...shows understanding towards a previous Assistant utterance	"Yeah."
Recommendation Rating		
Interest (INT)	...reacts positively to a recommendation without choosing it	"I do like Item08 and Item05, though."
Accept (ACT)	...accepts the recommended item, either explicitly or implicitly	"...I think Item02 would be a good fit for me."
Reject (RJT)	...rejects the recommended item, either explicitly or implicitly	"I just feel like Item03 and Item01 are too ugly."
Neutral Response (NR)	...does not indicate a decision with the current recommendations	"I see."
Inquire		
Inquire - Factual (IF)	...requires information regarding the recommendation	"Does it have a pocket that I can put my cell phone in?"
Inquire - Ask Opinion (IA)	...requires the Assistant's personal opinions	"...which one you think is more comfortable."
Critiquing		
Critique - Feature (CF)	...critiques a specific feature of the recommended item	"Item05, it's a bit long, I think..."
Critique - Compare (CC)	...requests comparison between a recommended item and another	"For Item23 and Item16, which one do you recommend?"
Explain (EXP)	...explains reasons for accepting a recommendation	"I like Item05 ...also it goes well with other pieces..."
Ask Clarification (AC)	...asks for clarification of something previously said	"Did you say Item13 as well?"
Others (OTH)	Greetings, gratitude expression, chit-chat utterance	"Hello"

3.3 Dialogue Intent

To better understand dialogue patterns with visual support in fashion shopping recommendations, we extended the holistic user and intent definitions from Lyu et al. [13] and also referred to the work of Cai and Chen on utterance-level labeling [3]. Since earlier studies focused on domains outside fashion shopping under conversational recommendation *without* visual grounding in conversations, we created new taxonomies on both sides to better categorize dialogue intents and allow for more thorough analysis. The taxonomies used in this study, designed for dialogue actions in a fashion shopping setting with shared visual catalogue, are listed in Tab. 2 & Tab. 3.

We conducted a two-phase labeling process to accurately annotate dialogue utterances for subsequent analysis. In the first phase, we developed an initial taxonomy list, primarily adapted from the work of Lyu et al. [13]. Next, we leveraged Gemini-2.5-Flash to streamline the tagging process. In the second phase, all tagged conversations are manually reviewed by two annotators to correct mislabeled utterances, and based on the collected dialogues, to refine and expand the taxonomy to better label the data. From these tags, we derive the intents we utilize in Sec. 4.1. For clarity, we primarily discuss overarching intent categories (e.g. *Provide*, or *Explain*) instead of specific tags.

3.4 Dataset Contents

Vogue[2] includes 60 dialogue transcripts (10 participant pairs × 6 scenarios each), transcribed and time-stamped at the turn level. Utterances are labeled with intent tags where appropriate. Transcripts are also annotated with the Seeker's final item choice and all items explicitly mentioned. Each scenario is matched to its catalogue. Items are released as utilized in the study with metadata.

All collected data (cf. Sec. 3.1) are available in Vogue[2], including fashion profile (onboarding) surveys (`fashion_profile` folder), item rating surveys from both Seeker and Assistant (`surveys` folder), and conversational transcripts (`conversations` subfolder of `data` folder). Major dataset statistics are listed in Tab. 4.

[2]The complete Vogue dataset is available at anonymous.4open.science/r/vogue-EF02

4 Data Analysis

In this section, we examine CRS dialogues from both a qualitative and quantitative perspectives, considering their overall structure, the alignment between Seeker and Assistant Preferences, and the performance of Modern MLLMs relative to the human Assistants.

Specifically, we focus our analysis with Vogue on three complementary aspects: **visual grounding**, which examines how shared images shape dialogue dynamics in inherently multimodal fashion scenarios; **preference alignment**, which tests whether Assistants alignment with both direction and intensity of Seeker preferences improves Seeker satisfaction; and **MLLM as recommender**, which evaluates whether current models can move beyond surface-level dialogue cues toward robust and generalizable preference modeling.

To this end, we propose three Research Questions (RQs):

RQ1: **Visual grounding & dynamics.** How does visual grounding alter the dynamics and stages of dialogue, compared to text-only settings? In an inherently multimodal domain like fashion, what structures and processes emerge, and how should CRS adapt to these?

RQ2: **Impact of Human Assistant alignments on Seeker experience.** To what extent do the item-level preference ratings predicted by Assistants after a conversation match the ground-truth ratings from Seekers? And does this alignment on preference actually correlate with the subjective satisfaction of Seekers?

RQ3: **MLLM versus Human Assistant Alignment.** How well do MLLMs perform when they are asked to recover a Seeker's ratings given the conversation? What strengths and weaknesses do they show compared to human Assistants?

Table 3: Assistant intent, building on Cai and Chen [3], Lyu et al. [13]. We present them hierarchically, with higher-fidelity tags grouped under broader categories. 3-letter tag codes as used in the data release and analysis are also specified.

Category (Code)	Description (Assistant...)	Example
Request		
Request Task Initiation (RTI)	...initiates dialogue with an open-ended prompt	"...how can I help you today?"
Request Preferences (RP)	...requests the Seeker's situational or background information	"Are you looking for...something with sort of a longer..."
Request Context (RC)	...requests situational or background information	"Ah, okay, so where is this farm going to be?"
Clarify Question (CQ)	...asks for clarification on a previous requirement	"...you're looking for something to wear in a more formal setting?"
Ask Opinion (A)	...requests the Seeker's opinion on a choice question	"So would you like to purchase this one?"
Ensure Fulfillment (EF)	...confirms task fulfillment during the conversation	"Sure and yeah, any items you want?"
Inform Progress (IP)	...discloses the current item being processed	"...so Item29 is a...give me one second..."
Acknowledgement (ACK)	...shows understanding towards a previous Seeker utterance	"Casual, stylish, and fit for indoors."
Answer (ANS)	...answers the question issued by the Seeker	"I would say waterproof is one level above water-resistant."
Recommend		
Recommend - Show (RS)	...provides recommendations by showing items directly	"Take a look at Item31, Item32, and Item28."
Recommend - Combine (RC)	...provides a recommendation compatible with another item	"You can wear Item22 or Item13 underneath Item17."
Explain		
Explain - Preference (EP)	...explains recommendations based on the Seeker's preferences	"Item03 is your best option because it's waterproof."
Explain - Additional Info (EAI)	...adds details not previously discussed	"It's machine washable, so easy to clean."
Personal Opinion		
Comparison (PCM)	...compares two or more mentioned items	"Between those two, only Item11 is machine washable."
Persuasion (PER)	...provides positive comments towards the recommended item	"Both the materials are pretty good."
Prior Experience (PEX)	...refers to past experience with the recommended item	"I have one of these myself, it's my go-to winter boot."
Context (PCN)	...provides opinion considering the given context	"If you're ever in deeper snow, that might be a consideration."
Others (OTH)	Greetings, gratitude expression, chit-chat utterance	"Hi."

Table 4: Vogue dataset statistics.

Statistic	Count
Conversations	60
Participants	20
Total turns	899
Total utterances	2100
Total tokens	22,717
Average turns per conversation	14.8
Average tokens per turn	36.1

By framing our analysis around these questions, we highlight how Vogue both enables deeper evaluation of existing systems and suggests design principles needed for user-centric multimodal CRS.

4.1 RQ1: Visual grounding & dynamics.

To address **RQ1**, we ask how visual grounding alters the dynamics and stages of dialogue compared to text-only settings, and what structural and behavioral changes result from this grounding. In this section, we examine the distribution of conversational intent, item mentions, and intent tag transitions across the course of the interaction and compare with related datasets. For the sake of clarity, our analysis will omit *Acknowledgment* and *Others* tags, since we find that they do not meaningfully contribute to the goal of the conversation.

As shown in Fig. 3, Assistants consistently introduce new items in two distinct waves: the first following preference and context elicitation, and the second after refinement based on earlier feedback. We also observe that Assistants frequently group recommendations, presenting up to four items simultaneously for discussion. This contrasts with conversational flows in text-only datasets, where item presentation tends to be more linear and sequential. Chiefly, we observe that the conversations flow naturally through a common implicit procedure shaped by visual grounding. The distributions

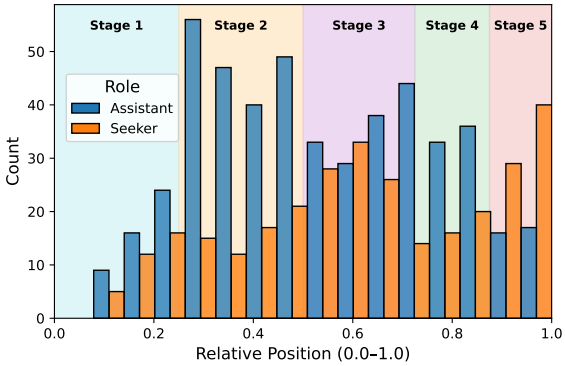


Figure 3: Seeker & Assistant Item Mention Distribution by Conversation Proportion. We observe a bimodal distribution in the Assistant, corresponding to *First Recommendation* and *Refinement* stages respectively.

in Fig. 3 further suggest a clear stage-like structure: activity is clustered into distinct peaks rather than being evenly spread across the dialogue, with a characteristic ebb and flow as initiative flips between participants. We therefore segment the conversations into five stages, each defined by the dominant (i.e, most prevalent) intents and role dynamics at that point. In Fig. 4, we provide an illustrative example of each stage based on conversations in Vogue.

Stage 1: **Preference Elicitation.** Early exchanges are dominated by preference elicitation (Assistant *Request*; Seeker *Provide*, *Ask for Recommendation*, and *Answer*).

Stage 2: **Recommendation.** This is followed by the first recommendation wave (Assistant *Recommend*), with *Explain* rising soon after. Interestingly, we observed that participants often discussed the items in groups according to abstract unifying properties of those item groups.

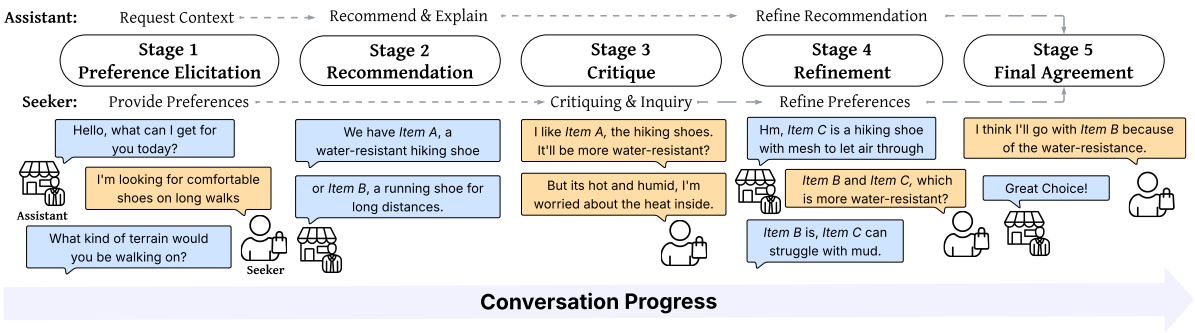


Figure 4: Our proposed stage splits, illustrating Recommender–User and Seeker-Assistant initiative and major intent tags. Examples of dialogue are included, along with each role's motivation at each stage.

Stage 3: Critique. The conversation then transitions to a bridge characterized by Seeker *Critique*, *Inquire*, and evaluative ratings, while Assistant *Explain* and *Personal Opinion* remain high.

Stage 4: Refinement. Stage 4 is strongly characterized by Assistants introducing a new, reduced set of items (usually one or two) to refine the recommendation based on Seeker critiques. Concurrently, Seekers refine their preferences based on available attributes across recommended items, continuing to utilize property grouping as observed in *Stage 2*.

Stage 5: Final Agreement. Finally, Stage 5 is marked by Seeker critique and *Recommendation Rating* peaking, particularly *Accept* tags, signaling choice and resolution.

Significance to CRS. Our analysis indicates that visual grounding and multimodality reshape conversational dynamics compared to text-only CRS settings. Dialogues in Vogue tend to follow a staged structure, supported by shared images that keep exchanges relatively compact. Conversations average 14.8 turns and 36 tokens per turn, yet still converge on final agreements with strong preference alignment (see Sec. 4.2).

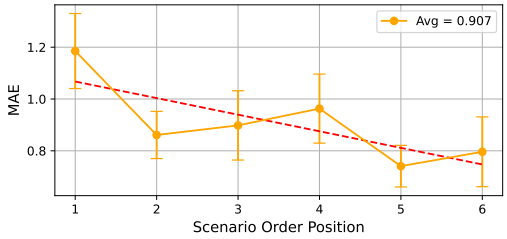
Anecdotal inspection of the transcripts indicates frequent usage of positional and unique identifying stylistic references in the shared catalogue that reduces the need for verbose description or minute back-and-forth; this is in contrast to similar studies without a shared visual catalogue [13], and allows participants to anchor their discussion and reasoning in concrete visual referents.

Taken together, these observations point to design implications for CRS: multimodal grounding can enable parallel exploration of grouped items, comparative reasoning, and collaborative refinement—interaction patterns that are far less accessible in text-only dialogue flows. In practice, these mechanisms could be directly incorporated into CRS design, providing a way to make interactions more efficient and user-friendly. Exploring how such patterns translate into deployed systems offers a promising direction for future work.

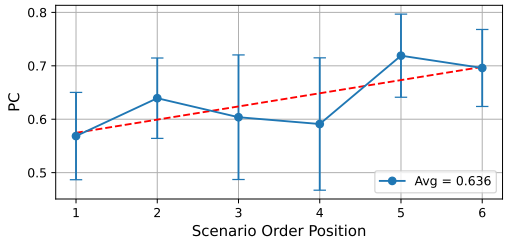
4.2 RQ2: Human Assistant Alignment and Impact on Seeker experience.

We now evaluate the quality of human Assistant interaction in Vogue using two complementary metrics: (1) quantitative alignment between human Assistant predictions and Seeker ground-truth ratings, and (2) post-task satisfaction surveys from Seekers.

Following Jannach et al. [7] and McNee and Riedl [14], we choose alignment as recovery of both the structure and intensity of preferences over accuracy, reporting it via Pearson Correlation (PC) and Mean Absolute Error (MAE) between item-level ratings. Satisfaction is included as an orthogonal but essential user-centered measure and reported directly on a 1-5 Likert scale.



(a) MAE↓ by scenario order position.



(b) PC by scenario order position.

Figure 5: Alignment metrics as a function of scenario order position. Scenario Order is the order in which individual trials were conducted for each Assistant-Seeker pair.

Alignment between Assistants and Seekers. We observed that Seekers strongly favored low scores (44% 1s, only 8.6% 5s), resulting in a highly imbalanced distribution. Assistants nonetheless

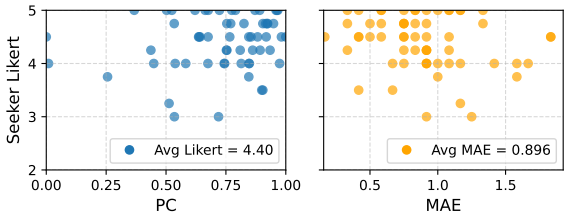


Figure 6: Relationship between Seeker satisfaction and alignment metrics. PC (left) shows strong clustering with higher satisfaction, while MAE (right) reveals broader variability.

achieved consistently good average PC (~0.64) as shown in Fig. 5, showing they recovered preference directions well. MAE, however, varied widely across scenarios, from as low as 0.25 to nearly 2.0, underscoring persistent challenges in modeling preference strength.

Since each Seeker engaged with the same Assistant in six independent scenarios, Fig. 5 also analyzes the performance across repeated interactions. PC improved slightly and MAE declined over successive scenarios, suggesting that Assistants became more effective with practice, despite neither party ever seeing the other's item ratings. While improvement trends were modest, the result underscores the value of shared history in conversational recommendation settings and indicates potential for adaptive learning without explicit item-level feedback.

Correlation between Alignment and Satisfaction. Unlike prior CRS resources, our design captures not only objective alignment through item ratings but also subjective satisfaction from Seekers, enabling a more holistic evaluation of recommendation quality. In Fig. 6, we observe that Seeker satisfaction (y-axis) was consistently high. Likert surveys captured perceived decision quality, justification, intent to purchase, and Assistant helpfulness. Most scores exceeded 4.5 (out of 5), with helpfulness and justification especially positive.

We further analyze the correlation between the Assistant's alignment and the Seeker's satisfaction (cf. Fig. 6). Encouragingly, higher satisfaction generally coincides with stronger alignment (measured by PC and MAE), indicating that better alignment contributes to better user experiences. This strong correlation also supports the use of alignment metrics to evaluate overall CRS performance in cases where user satisfaction is unknown, such as for the MLLM-based Assistant evaluated in the following sections.

Nonetheless, satisfaction varied across perceived decision impact and purchase intent, underscoring that it is not monolithic. This suggests that measuring only alignment, as is common in existing datasets, is insufficient. By contrast, the comprehensive design of Vogue highlights the opportunity for CRS to maximize impact by coupling robust preference modeling with user-centered design. In accordance with prior research [14], a system that feels supportive and transparent, while also aligning with user preferences, is likely to foster sustained user satisfaction and trust.

4.3 RQ3: MLLM versus Human Assistant Alignment.

The recent surge of MLLMs into conversational recommendation [20] leads to questions about their ability to operate on human-human dialogue in recommender and preference modelling roles. Having established the impact that alignment can have on user experience, we now examine how many MLLMs perform in this domain. To answer **RQ3**, we provide an evaluation of several prevalent MLLMs on the proposed dataset. We provide the prompt template as follows:

```
You are an expert sales associate and fashion expert,
 with deep experience in apparel, styling, sales, and
 customer relations, reviewing a recorded conversation
 between an assistant and a seeker. The assistant
 provides shopping advice and help to the seeker, who is
 attempting to make a clothes purchase. You must rate
 each of the 12 items on a 1-5 scale (inclusive).

Criteria: Rate the items based on how likely you think
 the Seeker is to purchase them, given the conversation
 and scenario that just concluded.

Provide only your rating in the format:

ItemXX, ItemYY, ItemZZ...
A, B, C...

Where A, B, C... is the 1-5 rating. There are always 12
 items per conversation, and each must receive a rating,
 even if it was not mentioned.
---
<metadata and conversation history appended here>
```

We note that the goal of this experiment is to evaluate the internal reasoning ability of MLLMs as recommenders. We do not compare specific fine-tuned or purpose-built MLLM-based CRS for fashion recommendation, as it is beyond the scope of this work. Accordingly, all models are provided with the same prompt as above, the full item catalogue & metadata, and the complete tagged transcript at the conversation level. We then instruct them to complete the same item rating survey task as the Assistants. We also include two baselines: (i) *Baseline: Random*, which randomly sample ratings from the Seeker's empirical groudn thruth distribution, and (ii) *Baseline: Mode Rating*, which returns the mode (i.e. most common) rating of 1. Our analysis first examines rating distributions, followed by item-level alignment and preference generalization.

Performance Overview. Similarly as before, we report PC and MAE; however, due to class imbalance (which most strongly affects the MLLMs) we also report per class (MAE[GT = k] for Seeker ground truth (GT) rating k) and Macro MAE (M-MAE). We report accuracy for all models, though we discuss later that this statistic can be misleading due to class imbalance. Due to the retroactive nature of these evaluations, no model is actively participating in the recommendation process so satisfaction metrics are not relevant here. The results are presented in Tab. 5. We note that the Assistants score best or 2nd best in nearly every measure. *Critically, no MLLM alternative tested was competitive with real human recommenders.*

Calibration and Distribution. Examining a distribution of ratings, we note that both families of models experience systemic failure in replicating Seeker ratings; Gemini models heavily overestimate the proportion of 1 ratings, while heavily collapsing on 3s and 4s; GPT-4o-mini adheres to a strongly normal distribution, peaking at 3, while GPT-5-mini produces a nearly linear decline across ratings (Fig. 7). All models maintain a relatively fair proportion of 5 ratings. This deficit is directly reflected in the PC

Table 5: Alignment metrics and accuracy for human Assistants and MLLMs. Best results from models and Assistants are bolded, second-best are underlined. Arrows indicate whether higher or lower is better. *PC undefined (uniform rating, zero variance)

Model	Accuracy ↑	PC ↑	MAE ↓	M-MAE ↓	MAE[GT=1] ↓	MAE[GT=2] ↓	MAE[GT=3] ↓	MAE[GT=4] ↓	MAE[GT=5] ↓
Human Assistants	0.397	**0.636**	**0.896**	**0.849**	**0.746**	0.891	1.171	1.284	**0.597**
GPT-5-mini	0.344	0.083	1.296	1.543	1.041	**0.845**	1.209	2.222	2.516
GPT-4o-mini	0.156	0.000	1.619	1.491	2.160	1.186	**0.953**	**1.160**	1.726
Gemini-2.5-Flash	**0.406**	0.057	1.356	1.681	0.853	0.930	1.628	2.444	2.839
Gemini-2.0-Flash	0.402	0.041	1.356	1.772	0.693	1.031	1.736	2.568	3.065
Baseline: Random	0.240	-0.005	1.713	1.730	1.997	1.403	1.147	1.469	2.387
Baseline: Mode Rating	0.443	NaN*	1.219	1.932	0.000	1.000	2.000	3.000	4.000

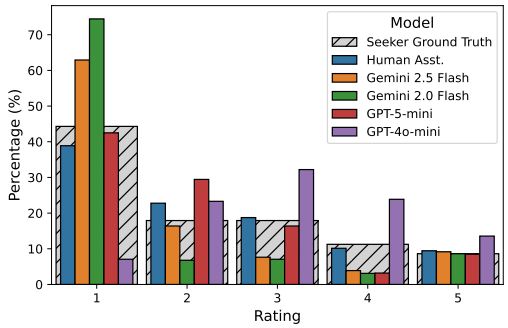


Figure 7: Distribution of Predicted Ratings vs. Seeker Ground Truth. The ground truth human rating distribution is presented in grey as the background.

performance of each model, where they perform very poorly, especially in comparison to human Assistants. Poor performance at capturing natural human rating distributions directly undermines the ability of these models to effectively predict preferences.

Weak Preference Generalization. Next we specifically examine how well different recommenders are able to infer and generalize preference knowledge within the conversation. Since it has been shown that most state-of-the-art MLLMs are able to parse explicit dialogue signals well [15], we foreground the ability to extend and generalize explicit preferences onto the full catalogue. Since many of these items *may not have been explicitly discussed*, the ability to generalize from the discussion to accurately predict all item ratings in the full catalogue is paramount for effective recommendation.

From the per-conversation M-MAE for the MLLM models and two baselines, we note a distinct and consistent gap from the Assistants' performance (cf. Tab. 5 and Fig. 1). Both Gemini models perform worse than the Random baseline. This underscores that extending preference inference beyond explicitly discussed items remains a challenging open problem, even for state-of-the-art MLLMs.

Implications for MLLMs in CRS. Viewed from the naive lens of accuracy, many MLLMs *appear* to perform comparably to human Assistants. However, this apparent parity is largely an artifact of class imbalance and the relative ease of recovering extreme ratings from conversational cues. When results are disaggregated by rating (i.e., MAE[GT = k] in Tab. 5), none clearly surpasses the Assistants. In fact, a trivial and uninformed strategy of predicting all 1s (Baseline: Mode Rating) beats all MLLMs in raw accuracy.

This dissonance stems in part from systematic distortions in rating distributions, which prevent MLLM-based recommenders from faithfully capturing Seeker ambivalence or partial fit. In turn, this weakens their ability to generalize preference reasoning beyond explicit conversational cues and actively discussed items. We do not observe substantial evidence that any of these models form a coherent preference model of the Seeker.

In direct response to **RQ3**, these results demonstrate that extending preference inference beyond explicit conversational cues remains an open challenge for state-of-the-art MLLMs. The dataset introduced here directly foregrounds this gap by enabling evaluation across fully featured candidate catalogues and comprehensive and detailed item-level feedback, providing a benchmark that moves beyond surface-level metrics like accuracy and supports rigorous comparison of MLLM-based CRS systems in terms of their capacity for nuanced and generalizable preference modeling.

5 Conclusion

We introduced Vogue, a multimodal conversational recommendation dataset that addresses current gaps by combining (i) **multimodality**, with fashion catalogs containing both product images and metadata to ground dialogue in visual and textual context; (ii) **human–human conversations**, with 60 natural Seeker–Assistant dialogues in contrast to synthetic or model-generated interactions; (iii) **domain specificity**, focusing on fashion, where subjective, context-sensitive choices make CRS evaluation especially challenging; and (iv) **explicit user profiles**, derived from pre-task surveys capturing style preferences, values, and behaviors.

Our stage analysis shows a consistent five-step rhythm: Assistant-led elicitation, Seeker critique, and convergence on choices. Visual grounding and grouped recommendations support comparative reasoning and faster convergence, while transparent justification remains crucial when alignment falters. At the item level, human Assistants still outperform large models: while MLLMs can follow explicit cues, they fail to capture relative preferences or generalize across all items. This underscores preference modeling and rating calibration as central open challenges.

Taken together, Vogue foregrounds item-level alignment, conversational stage dynamics, and satisfaction in a realistic, visually grounded setting. Most importantly for the development of future multimodal CRS, our findings establish Vogue as a challenge dataset beyond the current recommendation capabilities of existing top-tier MLLMs such as GPT-4o-mini, GPT-5-mini, and Gemini-2.5-Flash.

References

[1] Marco Alessio, Simone Merlo, Tommaso Di Noia, Guglielmo Faggioli, Marco Ferrante, Nicola Ferro, Cristina Ioana Muntean, Franco Maria Nardini, Fedelucio Narducci, Raffaele Perego, Giuseppe Santucci, and Nicola Viterbo. 2025. CoSRec: A Joint Conversational Search and Recommendation Dataset. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 3466–3477. doi:10.1145/3726302.3730319

[2] Nolwenn Bernard and Krisztian Balog. 2023. MG-ShopDial: A Multi-Goal Conversational Dataset for e-Commerce. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. ACM, 2775–2785. doi:10.1145/3539618.3591883

[3] Wanling Cai and Li Chen. 2020. Predicting User Intents and Satisfaction with Dialogue-based Conversational Recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 33–42. doi:10.1145/3340631.3394856

[4] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 815–824. doi:10.1145/2939672.2939746

[5] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2 (2021), 100–126. doi:10.1016/j.aiopen.2021.06.002

[6] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. arXiv:2009.14306 [cs.CL] https://arxiv.org/abs/2009.14306

[7] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *Comput. Surveys* 54, 5 (May 2021), 1–36. doi:10.1145/3453154

[8] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 4903–4912. doi:10.18653/v1/2021.emnlp-main.401

[9] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 9748–9758.

[10] 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 *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 675–684. doi:10.1145/3404835.3462970

[11] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 4335–4347. doi:10.18653/v1/2021.emnlp-main.356

[12] 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. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 3515–3533. doi:10.18653/v1/2023.findings-acl.217

[13] Shengnan Lyu, Arpit Rana, Scott Sanner, and Mohamed Reda Bouadjenek. 2021. A Workflow Analysis of Context-driven Conversational Recommendation. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 866–877. doi:10.1145/3442381.3450123

[14] Sean McNee and John Riedl. 2006. Being accurate is not enough: How accuracy metrics have hurt recommender systems. *CHI '06 Extended Abstracts*, 1097–1101. doi:10.1145/1125451.1125659

[15] Lalchand Pandia and Allyson Ettinger. 2021. Sorting through the noise: Testing robustness of information processing in pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1583–1596. doi:10.18653/v1/2021.emnlp-main.119

[16] Kaicheng Pang, Xingxing Zou, and Waikeung Wong. 2025. FashionM3: Multimodal, Multitask, and Multiround Fashion Assistant based on Unified Vision-Language Model. arXiv:2504.17826 [cs.CV] https://arxiv.org/abs/2504.17826

[17] Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2018. Towards Building Large Scale Multimodal Domain-Aware Conversation Systems. arXiv:1704.00200 [cs.CL] https://arxiv.org/abs/1704.00200

[18] Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval*. 235–244.

[19] Zihan Wang, Xiaocui Yang, Yongkang Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2025. Muse: A Multimodal Conversational Recommendation Dataset with Scenario-Grounded User Profiles. arXiv:2412.18416 [cs.MM] https://arxiv.org/abs/2412.18416

[20] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2024. Large Language Models for Recommendation: Progresses and Future Directions. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1268–1271. doi:10.1145/3589335.3641247

[21] 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*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 4128–4139. doi:10.18653/v1/2020.coling-main.365

A Protocol Supplement

A.1 Scenarios

We designed the study with well-rounded scenarios to cover a variety of subjective fashion needs (e.g. "for campus", "business casual"), along with requirements based needs (e.g. rain, temperature). We present a few scenarios as an example:

```
(1) It's a weekend evening, and you've been invited to a
     small dinner at a friend's apartment. It's not
     formal (no dress code) but you know people will look
      put together. You'll be indoors, maybe helping in
     the kitchen, lounging on the couch, or playing a
     group game. You want to look good without looking
     like you tried too hard, casual, clean, and
     comfortable. Which layering piece would you wear in
     this setting?

(2) You're heading out for a weekend trip to a coastal
     town in early summer. The forecast promises a bright
     , chilly morning, with temperatures rising quickly,
     especially around midday. What type of jacket would
     you want to have for this occasion?
```

A.2 Platform UI

We also present samples of the UI used during the trials in Fig. 8. Both views are presented. In the trials, each participant only saw their respective UI. While the Seeker did not need to scroll to see every item, due to more product info on the Assistant side, they needed to scroll. Between conversations and participants, item locations were randomized to mitigate locational bias.

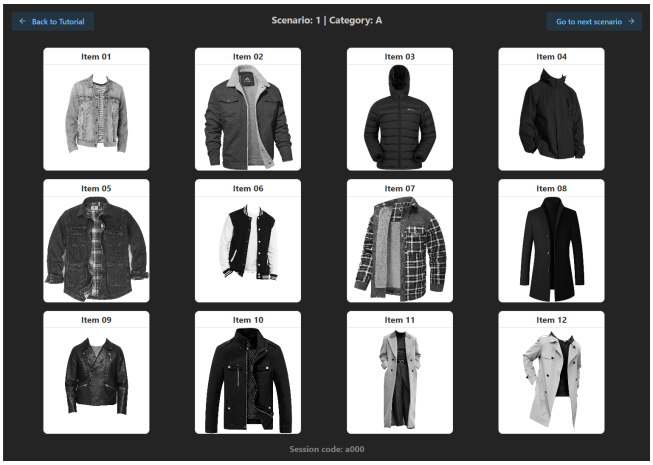
A.3 Satisfaction Surveys

We present Tab. 6 which lists the questions asked to assess Seeker satisfaction.

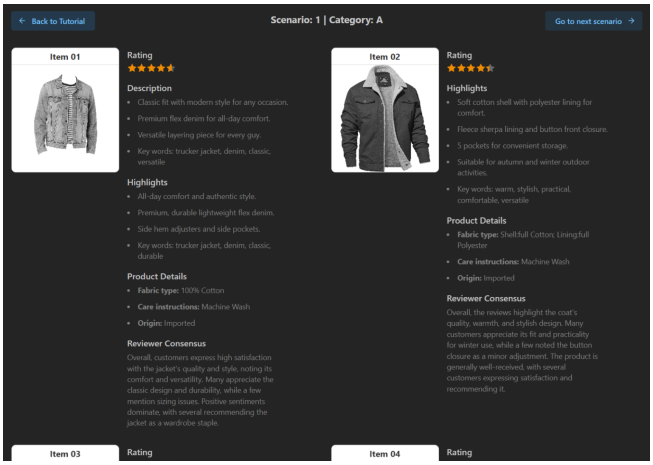
A.4 Sample Product and Metadata

We present here an example of the bundled metadata, with accompanying picture in Fig. 9

```
Categories:
  - Clothing, Shoes & Accessories
```



(a) Seeker User Interface



(b) Assistant User Interface

Figure 8: Study User Interfaces used in our experiment. The Seeker UI is a 3 x 4 grid of product photos. The Assistant UI is 2 columns, with product photos and metadata.

Table 6: Seeker satisfaction items.

Label	Question
Likert-1	The conversation allowed me to make a good selection on what to buy.
Likert-2	I made a well justified final selection.
Likert-3	I would actually purchase the item selected.
Likert-4	During the conversation, I felt like I was in control of the process.
Likert-5	I found the recommendations given by the Assistant helpful.

```
  - Men
  - Shoes
  - Fashion Sneakers

Product Name: Under Armour Mens Charged Assert 9 Running
     Shoes
Product Brand: Visit the Under Armour Store
```

Figure 9: Example product image from the dataset metadata.

```
Product Rating: 4.5

Product Description:
  - Built for speed and comfort
  - Responsive Charged Cushioning midsole
  - Solid rubber outsole for durability
  - Key words: running, performance, support, lightweight

About Product:
  - Lightweight mesh upper for breathability
  - Durable leather overlays for stability
  - EVA sockliner for comfort
  - Charged Cushioning for responsiveness
  - Key words: running shoes, breathable, comfortable,
      durable

Product Detail:
  - Care instructions: Machine Wash
  - Sole material: Rubber
  - Shaft height: Low-top
  - Outer material: Mesh

Reviews:
  Overall, customers express high satisfaction with the
       shoes, highlighting their comfort, support, and
       durability for various activities. Many appreciate
        the lightweight design and breathability, making
       them suitable for both running and everyday wear.
       Some users noted sizing issues but still found the
        shoes to be a great choice for their needs.
```

In JSON format:

```
{
  "ID": "S002",
  "ASIN": "B08CFPVBTW",
  "ProductName": "Under Armour...",
  "Brand": "Under Armour",
  "Category": "Shoes",
  "Rating": 4.5,
  "Description": "Lightweight...",
  "Detail": {
    "Care": "Machine Wash",
    ...
  },
  ...
}
```

Due to the large number of reviews, which Assistants could not feasibly parse, we instead include a GPT-4o-mini summary of the reviews. The prompt for this summary is presented here:

```
    Exclude all non-English reviews. Write a single-
        paragraph summary capturing the overall
        sentiment and recurring themes. Do not mention
        anything related to colors, sizing, fit, or
        prices.
```

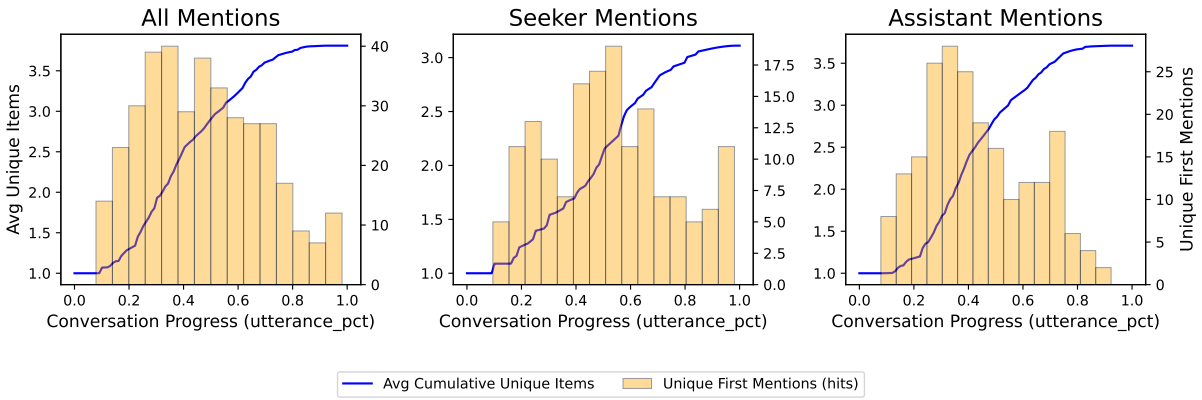


Figure 10: Cumulative distribution of first item mentions over the course of the conversation by role.

A.5 Minimum Item Mentions

In 3.1, we instruct Assistants to discuss at least 3 items during each conversation. We observed that this does not meaningfully affect the structure of dialogue, or number of items mentioned. We present Fig. 10, where first item mention histograms are collated by participant role. Cumulative totals are also presented. We note the distinct lack of plateau around 2 mentioned items, and that the final total mentions do not cluster around 3. This reinforces observations that the first recommendation cluster features 2-4 items, while the second is supplementary as needed. The mode of Assistant first-mentioned items is 4 items.

B Dataset

B.1 Example Conversation and Annotations

We provide a representative dialogue from Vogue in the following section, including utterances, intent tags, and item mentions. This excerpt demonstrates typical conversational flow: initial context elicitation, grouped recommendations, refinement, and final acceptance.

```
Conversation Id: 1 | Scenario: 1 | Final Choice: 2 |
     Mentioned: 2, 7, 9

Turn 1 (Assistant201, 00:06:34)
  U1: "Hi, how can I help you today?"
      [Request-Task-Initiation (RTI)]

Turn 2 (Seeker201, 00:06:36)
  U1: "Yeah, so I want to choose a right outerwear for a
       fall visit to a farm."
      [Provide-Context (PCT)]
  U2: "So do you have any recommendations for me?"
      [Initial-Query (IQ)]

[...]

Turn 13 (Assistant201, 00:08:58)
  U1: "Okay."
      [Acknowledgement (ACK)]
  U2: "Yeah, in that case, I would recommend... probably
       Item02 will be your best bet."
      [Recommend-Show (RS)]
  U3: "Because it is waterproof and breathable, perfect
       for being outdoors."
      [Explain-Preference (EP)]
  U4: "So what do you think about that?"
      [Ask-Opinion (A)]

Turn 14 (Seeker201, 00:09:24)
  U1: "Yeah, yeah, I think I will go with Item02."
      [Accept (ACT)]

Turn 15 (Assistant201, 00:09:29)
  U1: "Okay, cool."
      [Acknowledgement (ACK)]
  U2: "Sounds good."
      [Acknowledgement (ACK)]
```

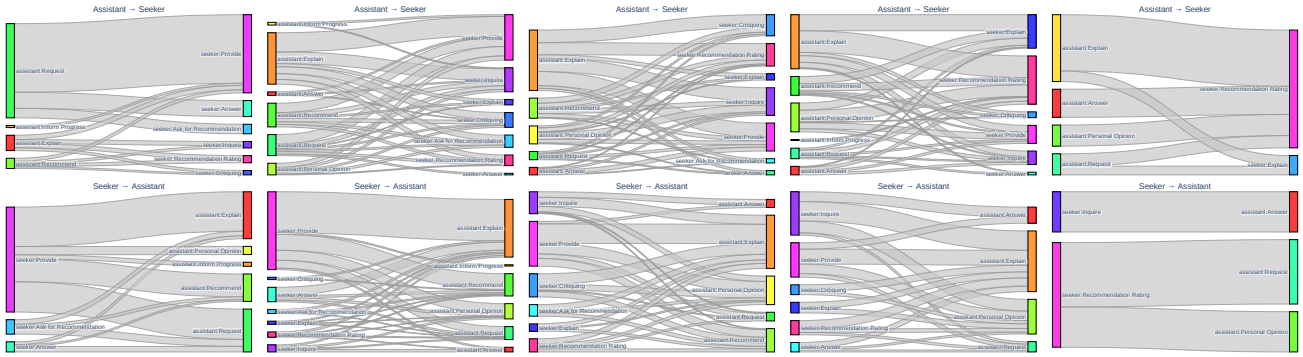
C In-depth Stage Analysis

With the stage analysis in place, we provide a closer account of how conversations narrow from broad exploration to final choice. Stage-wise histograms are shown in Appendix Fig. 12, with Sankey diagrams of tag transitions in Appendix Fig. 11. Although Seekers could reject all options, this never occurred in practice.

Stage 1: Preference Elicitation. Conversations begin asymmetrically, with Assistants using *Request* tags to probe context, and Seekers responding with *Provide* to supply details and constraints. This establishes the initial search space, positioning the Assistant as interviewer and the Seeker as explainer.

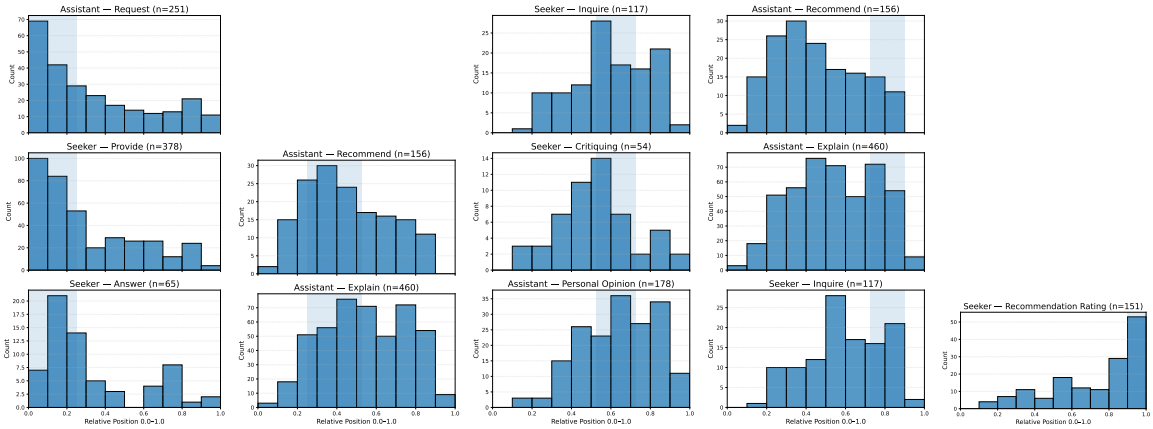
Stage 2: Recommendation. Assistants expand the space by introducing small groups of items with *Recommend–Show*, often supported by *Explain*. Both parties also leverage the shared catalogue to prompt grouped, comparative reasoning: Seekers weigh abstract properties against one another rather than evaluating items in isolation. The dialogue thus broadens quickly while setting up efficient downstream pruning.

Stage 3: Critique. Initiative shifts as Seekers employ *Critique* and *Inquire* moves, while Assistants counter with *Explain* and *Answer*. The search space narrows through this back-and-forth: Seekers articulate trade-offs, and Assistants refine understanding by clarifying features. Dialogue becomes more balanced and collaborative, with preferences expressed at an abstract, attribute-level rather than tied to single items.



(a) Stage 1 (b) Stage 2 (c) Stage 3 (d) Stage 4 (e) Stage 5

Figure 11: Stage-wise Sankey diagrams of dialogue act transitions.



(a) Stage 1 (b) Stage 2 (c) Stage 3 (d) Stage 4 (e) Stage 5

Figure 12: Stage-wise histograms of dialogue act distributions.

Stage 4: Refinement. Building on this feedback, Assistants reintroduce a reduced set of candidates with *Recommend*, often paired with *Explain* or *Personal Opinion*. This reduced set may include new introductions, or simply re-affirm a previous recommendation. Item group discussion may continue, as Seekers contribute short *Inquire* or *Rating* turns that either reinforce or challenge convergence. This iterative narrowing prunes the options towards a final candidate.

Stage 5: Final Agreement. The dialogue closes as *Recommendation Rating–Accept* peaks: Seekers confirm a choice, often adding a brief *Explain* for justification. Assistants reciprocate with *Explain*, *Personal Opinion*, or *Acknowledge*, reinforcing closure. At this stage, both roles have actively narrowed the path from broad exploration to final commitment, completing the collaborative arc.

Received 07 October 2025