Where and What: Reasoning Dynamic and Implicit Preferences in Situated Conversational Recommendation

Dongding Lin[1], **Jian Wang**[1,2†], **Yongqi Li**[1], **Wenjie Li**[1]
[1] Department of Computing, The Hong Kong Polytechnic University
[2] College of Computer Science, Sichuan University
dongding88.lin@connect.polyu.hk jian51.wang@polyu.edu.hk
liyongqi0@gmail.com cswjli@comp.polyu.edu.hk

Abstract

Situated conversational recommendation (SCR), which utilizes visual scenes grounded in specific environments and natural language dialogue to deliver contextually appropriate recommendations, has emerged as a promising research direction due to its close alignment with real-world scenarios. Compared to traditional recommendations, SCR requires a deeper understanding of dynamic and implicit user preferences, as the surrounding scene often influences users' underlying interests, while both may evolve across conversations. This complexity significantly impacts the timing and relevance of recommendations. To address this, we propose situated preference reasoning (SiPeR), a novel framework that integrates two core mechanisms: (i) *Scene transition estimation*, which estimates whether the current scene satisfies user needs, and guides the user toward a more suitable scene when necessary; and (ii) *Bayesian inverse inference*, which leverages the likelihood of multimodal large language models (MLLMs) to predict user preferences about candidate items within the scene. Extensive experiments on two representative benchmarks demonstrate SiPeR's superiority in both recommendation accuracy and response generation quality. The code and data are available at https://github.com/DongdingLin/SiPeR.

1 Introduction

Conversational recommendation (Li et al., 2018; Gao et al., 2021; Jannach et al., 2021; Zhou et al., 2022), as an extensively explored research area, focuses on delivering high-quality recommendations through natural language dialogue. It enables recommenders to actively inquire about user preferences and respond dynamically to user requests. In many real-world scenarios, recommendations are inherently grounded in specific environments, such as live promotions in clothing or furniture stores (Kottur and Moon, 2023). This has recently shifted research interest to situated conversational recommendation (SCR) (Lin et al., 2024; Wang et al., 2024c), which leverages visual scenes grounded in specific environments and natural language dialogue to deliver contextually appropriate recommendations. This close alignment with the real world underscores the importance of SCR as a promising and practical research direction.

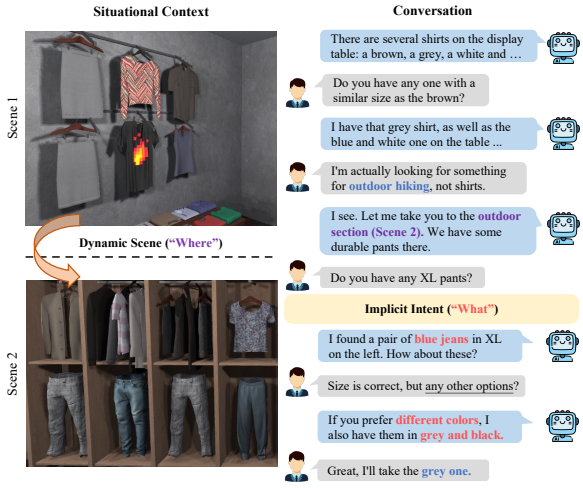


Figure 1: An illustrative example from the repurposed SIMMC 2.1 (Kottur and Moon, 2023) dataset for situated conversational recommendation, where the interaction process between the user and the virtual assistant is grounded in evolving scenes. In the bottom panel, although the initially suggested blue jeans satisfy the size constraint, the user's final acceptance shows that the grey pants are the ground-truth target, highlighting the need to reason about implicit preferences.

Despite its great potential, existing studies in SCR primarily focused on dataset curation (Moon et al., 2020; Kottur and Moon, 2023; Lin et al., 2024; Wang et al., 2024c), yet they failed to establish a clear framework for effectively solving the task. Building an effective SCR system is non-

†Corresponding author. This work was mainly conducted at PolyU, while the author is now at Sichuan University.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10463–10481
July 2-7, 2026

trivial due to its challenges in reasoning user preferences: 1) User preferences are often **dynamic** and varied by situations. In SCR, user interests are often influenced by the surrounding environment. When the surrounding scene evolves across conversations, user preferences can shift, adding further complexity to the recommendation. For example, as illustrated in the top panel of Figure 1 (The "Where"), when the user expresses an interest in "outdoor hiking," the system recognizes that the current formal wear scene is a mismatch. Consequently, it must actively guide the user to the outdoor section (Scene-2) to align with potential user interests. This necessitates a critical decision-making capability for the system to determine *where* to transition between scenes, which has been largely overlooked in prior work. 2) User preferences are often **implicit** rather than explicitly stated. For instance, in the bottom panel (The "What"), the user acknowledges that "the size is correct" but still asks for other options. In the full dialogue, the user eventually accepts the grey pants, indicating that while the size constraint is satisfied, the initially recommended blue jeans do not match the user's intended purchase item. To address this, the system must accurately distinguish and predict the true target item from the remaining candidates in the scene. This requires reasoning about *what* the user truly desires, i.e., the underlying needs and preferences in their expressed utterances.

To address the above two challenges, we introduce **Si**tuated **P**refe**r**ence **R**easoning (**SiPeR**), a novel framework that accordingly integrates two key mechanisms. First, we present **scene transition estimation**, which focuses on joint modeling of transition decision and target scene prediction. By leveraging multimodal large language models (MLLMs) (Wang et al., 2024b; Liu et al., 2024) to represent both visual scenes and conversation histories, this mechanism dynamically estimates whether the current scene aligns with user needs, allowing the system to predict a more suitable scene and guide the user to it in the next turn. Second, considering that LLMs often struggle to disentangle nuanced preferences from surface-level conversation, we formalize preference discovery as a **Bayesian inverse inference** (Baker et al., 2009; Ullman et al., 2009; Jin et al., 2024) problem. This approach treats the user's utterance as an observable action generated by a latent goal. By leveraging two opposing hypothetical beliefs (like vs. dislike), we quantify the likelihood of each potential item being the "what" the user desires. This allows the system to move beyond heuristic guesses and perform more rigorous probabilistic reasoning.

Our contributions are summarized as follows:

- We identify the **unique yet underexplored challenges** in situated conversational recommendation (SCR): reasoning *dynamic* and *implicit* user preferences in grounded, evolving scenes. Bridging this gap is important for delivering contextually appropriate recommendations in real-world settings.
- In light of these challenges, we propose **SiPeR**, a novel situated preference reasoning framework that integrates scene transition estimation and Bayesian inverse inference. To our knowledge, this work is among the early framework-level attempts to systematically address SCR.
- Our SiPeR achieves notable improvements over the compared baselines, with an average improvement of 10.9% on SIMMC 2.1 (Kottur and Moon, 2023) and 10.6% on SCREEN (Lin et al., 2024), respectively. Further analyses validate the effectiveness of each proposed mechanism, providing valuable insights into the development of practical SCR systems.

2 Related Work

Conversational Recommender Systems. Existing conversational recommender systems (Li et al., 2018; Liu et al., 2020)(CRSs) seek to improve recommendation quality by focusing on two main aspects: learning effective item representations (Zhang et al., 2019; Zhou et al., 2020) and understanding user preferences conveyed in dialogues (Deng et al., 2021; Lin et al., 2023). The former involves learning informative item embeddings that can represent items accurately (Lu et al., 2021; Zhou et al., 2022), while the latter focuses on extracting user preferences from the dialogues to enhance personalization (Chen et al., 2019; Wang et al., 2021). However, despite the success of these approaches, they primarily focus on text-based interactions, overlooking the visual information of items. In many real-world scenarios, the visual characteristics of items may significantly influence user preferences (Long et al., 2023). Additionally, environmental factors, which influence both the context in which recommendations are made and the user's interaction with items, can significantly affect the quality of recommendations.

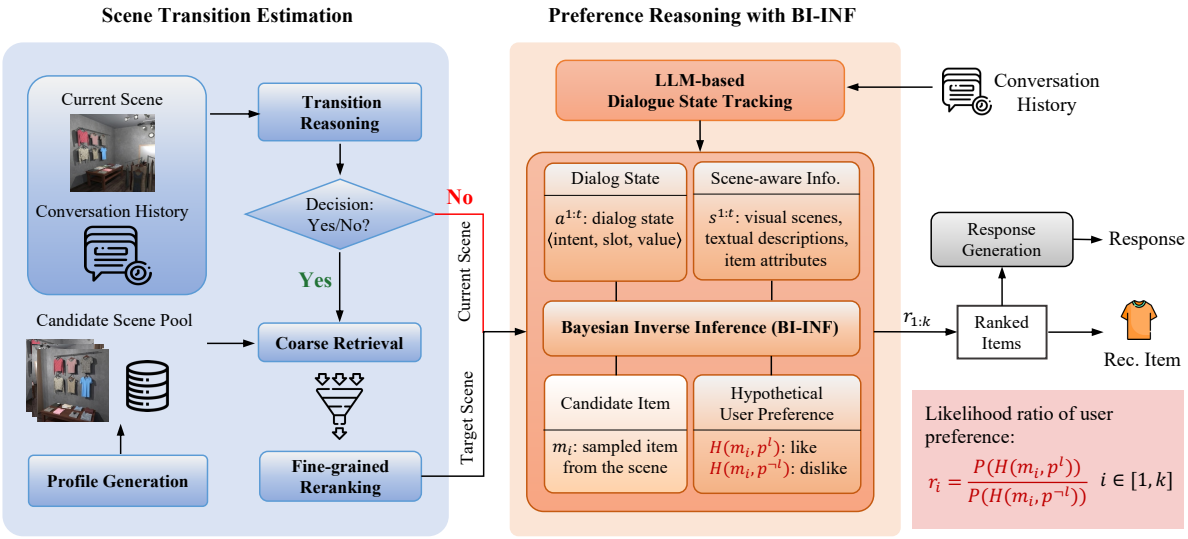


Figure 2: Overview of the Situated Preference Reasoning (SiPeR), which has two critical mechanisms: (a) scene transition estimation (STE) and (b) Bayesian inverse inference (BI-INF).

Situated Conversational Recommendation. In recent years, considerable attention has focused on how user preferences and interests evolve under the influence of situational context (Crook et al., 2019). To advance the development of this emerging field, Lin et al. (2024) pioneered the formalization of situated conversational recommendation (SCR). They leveraged the SIMMC 2.1 dataset (Kottur and Moon, 2023) and released the first SCR dataset, SCREEN (Lin et al., 2024), which significantly contributes to the understanding of dynamic user needs in context-aware conversations. Subsequently, Wang et al. (2024c) crafted the MUSE dataset by collecting user profiles from real-world scenarios and simulating dialogues using a multi-agent framework powered by MLLMs. These foundational works have significantly contributed to SCR research by providing valuable datasets, enabling more accurate modeling of user behavior and context. Despite these efforts, there is still a lack of comprehensive analysis or a dedicated framework designed to systematically address situated conversational recommendations.

3 Task Formulation

Let us consider a shared environment defined by a collection of visual scenes $\{\mathcal{S}_i\}_{i=1}^{C}$, where each scene $\mathcal{S}_i$ contains a set of candidate items $\{\mathcal{I}_{i,j}\}_{j=1}^{K_i}$, C denotes the total number of scenes in the environment, K_i represents the number of items. These scenes and items are accessible to both users and the recommender assistant (system). The system engages in multi-turn interactions with a user through natural language conversations, represented as $\{u_t, v_t\}_{t=1}^{T}$, where u_t and v_t denote the user and system utterances at the t-th turn, respectively. T denotes the total number of turns.

At each turn t, the system operates with (i) a situational context $\mathcal{C}_t$, which includes a specific visual scene $\mathcal{S}_t$ and the corresponding item set $\mathcal{I}_t$ in the scene, and (ii) a conversation history $\mathcal{H}_t = \{u_{\leq t}, v_{<t}\}$, which comprises all past user utterances and system responses. The objective of situated conversational recommendation is to generate a contextually appropriate response v_t that adapts to the user's evolving interests. This process entails determining the appropriate visual scene (*where*) to ground the conversation and, when suitable, recommending a subset of items (*what*) from the target scene that best satisfy user preferences.

4 Method

In this section, we present **Si**tuated **P**refe**r**ence **R**easoning (**SiPeR**), a novel framework that comprises two critical mechanisms: scene transition estimation (see §4.1) and Bayesian inverse inference (see §4.2). We introduce the end tasks of recommendation and response generation in §4.3. Figure 2 shows the overview of SiPeR.

4.1 Scene Transition Estimation

To ensure precise and natural scene transitions, we propose a reasoning-driven generative-retrieval framework for Scene Transition Estimation (STE).

Instead of directly matching the dialogue history to latent scene representations, our framework first externalizes the desired environment as a semantic profile, which provides an explicit anchor for the user's implicit and evolving intent. This generated profile is only an intermediate query rather than the final transition result. We then perform a coarse-to-fine retrieval step to identify the optimal scene from a large-scale candidate pool.

Scene-Profile Representation. Direct reasoning on raw visual scenes is computationally prohibitive and prone to noise. We therefore convert each scene $\mathcal{S}_i$ in the candidate pool into a textual situated profile $d_{\mathcal{S}_i}$, using a pre-trained MLLM. Each profile encapsulates the spatial relations and a structured catalog of items with their visual attributes. This profile is leveraged to transform target scene prediction into a semantic matching problem.

Profile Generation of Target Scene. To maintain proactive yet coherent dialogue, the system must envision a target scene that satisfies the user's implicit intent. Given the conversation history $\mathcal{H}_t$ and the current scene profile $d_{\mathcal{S}_t}$, we prompt the MLLM $\mathcal{F}$ to perform a joint inference of *transition decision* and *target generation*. This process aims to determine the necessity of a transition by balancing alignment with user needs against coherence with the current scene. Specifically, the model is instructed to first generate a decision token $y_{dec} \in \{\texttt{Yes}, \texttt{No}\}$, followed by the profile content $d_{\tilde{\mathcal{S}}}$ of the expected target scene:

$$y_{dec}, d_{\tilde{\mathcal{S}}} = \mathcal{F}(\mathcal{H}_t, d_{\mathcal{S}_t}). \tag{1}$$

We quantify the likelihood of the transition via the normalized probability of the decision token:

$$s_{trans} = \frac{\exp(\mathbf{z}_{\texttt{Yes}})}{\exp(\mathbf{z}_{\texttt{Yes}}) + \exp(\mathbf{z}_{\texttt{No}})}, \tag{2}$$

where $\mathbf{z}$ denotes the output logits corresponding to the generated tokens. When the transition decision is affirmative (i.e., $y_{dec} = \texttt{Yes}$), this generated profile $d_{\tilde{\mathcal{S}}}$ serves as a semantic anchor to predict the scene to transit in the retrieval stage. Otherwise, the system retains the current scene $\mathcal{S}_t$ as the grounded context for the downstream stage.

Coarse-to-Fine Transition Reasoning. Since exhaustive semantic reasoning over all candidates is intractable, we employ a coarse-to-fine strategy for transition estimation. In the *coarse retrieval* stage, we aim to narrow down the scope of the candidate scenes. We encode all candidate profiles $\{d_{\mathcal{S}_i}\}$ and the reasoned profile $d_{\tilde{\mathcal{S}}}$ into a shared embedding space with an encoder $\phi(\cdot)$. The similarity score for each candidate scene is given by:

$$\text{Score}(\mathcal{S}_i) = \frac{\phi(d_{\tilde{\mathcal{S}}}) \cdot \phi(d_{\mathcal{S}_i})}{||\phi(d_{\tilde{\mathcal{S}}})|| \cdot ||\phi(d_{\mathcal{S}_i})||}. \tag{3}$$

We retain the top-N candidates to form a reduced subset $\mathcal{S}_{\text{top}}$ of the scenes. Then, we take a *fine-grained reranking* for determining the target scene. To achieve precise estimation, we train a reranker f_θ (parameterized by an LLM) to evaluate the alignment between $d_{\tilde{\mathcal{S}}}$ and each $d_{\mathcal{S}_j}$, where $\mathcal{S}_j \in \mathcal{S}_{\text{top}}$. To optimize the reranker, we minimize the following negative log-likelihood during training:

$$\mathcal{L}_r = -\log \frac{\exp(f_\theta(d_{\tilde{\mathcal{S}}}, d_{\mathcal{S}^*}))}{\sum_{\mathcal{S}_j \in \mathcal{S}_{\text{top}} \cup \{\mathcal{S}^*\}} \exp(f_\theta(d_{\tilde{\mathcal{S}}}, d_{\mathcal{S}_j}))}, \tag{4}$$

where $d_{\mathcal{S}^*}$ denotes the profile of the ground-truth scene. This trained reranker ensures that the estimated scene satisfies user needs while maintaining a smooth transition. Consequently, even if the generated profile contains imperfect or partially hallucinated attributes, the final transition decision remains grounded in real candidate scenes rather than unconstrained free-form generation.

4.2 Bayesian Inverse Inference

Once the appropriate scene for the next turn is identified, we aim to reason the user's underlying preferences about potential items within that scene. Since LLMs often struggle to disentangle nuanced preferences from surface-level conversation, we formalize preference reasoning as a Bayesian inverse inference (BI-INF) (Baker et al., 2009; Ullman et al., 2009; Jin et al., 2024) problem. Our approach consists of the following three stages.

Dialogue State Tracking. Dialogue state tracking (DST) aims to estimate the dialogue state at each conversational turn, where the state is typically represented as a set of structured tuples related to system actions or user intents. Here, we refer to dialogue states as user intents, such as requesting product information or comparing different items (see Figure 6 in the Appendix). Specifically, we directly instruct a powerful LLM to extract symbolic dialogue states from dialogue histories in the form of $\langle$`intent, slot, value`$\rangle$ tuples. To assess reliability, we randomly sampled LLM-extracted states from each downstream dataset and

manually verified their correctness against the predefined schema. This approach achieves a high accuracy of 98.8%, validating the quality of the extracted states.

User Preference Modeling. Drawing inspiration from the Bayesian Inverse Planning (BIP) framework used in computational cognitive science (Baker et al., 2009; Ullman et al., 2009) and recent multimodal Theory of Mind research (Jin et al., 2024; Shi et al., 2024), we approach preference reasoning by *reversing* the user's decision-making process. Instead of modeling the system's policy, we formulate the user as a rational agent interacting with the environment. This process is formalized as a Partially Observable Markov Decision Process (POMDP) defined by the tuple $\langle \mathcal{S}, \mathcal{M}, \mathcal{A}, \mathcal{T}, \pi \rangle$. Here, $s_t \in \mathcal{S}$ represents the situational context (scene). $m_i \in \mathcal{M}$ denotes the user's latent goal (i.e., the target item they desire), and p_t represents their evolving mental state (e.g., beliefs or specific preferences about item attributes). Crucially, we view the user's utterance as an action $a_t \in \mathcal{A}$ (represented as the dialogue state) taken to achieve their goal. The user generates these dialogue actions according to a latent policy $\pi(a_t|m_i, p_t, s_t)$, which reflects the likelihood of the user expressing specific intents given their underlying goal m_i and the current context.

Based on this forward generative model, we can infer the user's latent goal m_i by observing their dialogue actions $a_{\leq t}$. We represent the posterior probability of the user desiring item m_i as follows:

$$\mathbb{P}(m_i, p_t|a_{\leq t}, s_{\leq t}) \propto \prod_{\tau=1}^{t} \pi(a_\tau|m_i, p_\tau) \cdot \mathbb{P}(p_\tau|p_{\tau-1}, s_\tau)\mathbb{P}(p_0)\mathbb{P}(m_i), \quad (5)$$

where $\mathbb{P}(m_i)$ is the prior over items. The term $\mathbb{P}(p_\tau|p_{\tau-1}, s_\tau)$ models the dynamics of user preference states. In practice, rather than maintaining an explicit state vector, we approximate this belief update by conditioning the model on the history of dialogue states $a_{<\tau}$ and the situational context (Hausknecht and Stone, 2015; Rabinowitz et al., 2018). The term $\pi(a_\tau|m_i, p_\tau)$ serves as the core *user likelihood* function: it quantifies how likely the user is to produce the dialogue state a_τ if their true goal were item m_i.

Inverse Inference through Hypotheses. Directly calculating the user policy $\pi(\cdot)$ in Eq. (5) is intractable due to the vast space of natural language. To address this, we follow Jin et al. (2024) to amortize the policy utilizing a fine-tuned MLLM. This approach leverages the model's world knowledge to simulate the user's behavior. To infer the user's attitude toward a candidate item m_i, we compare two competing hypotheses: (i) $\mathcal{H}(m_i, p_t^l)$, denoting the user *likes* (or accepts) item m_i; and (ii) $\mathcal{H}(m_i, p_t^{\neg l})$, denoting the user *dislikes* (or rejects) item m_i. We compute the likelihood ratio of these hypotheses as:

$$\frac{\mathbb{P}(m_i, p_t^l)}{\mathbb{P}(m_i, p_t^{\neg l})} \approx \frac{\pi(a_t|m_i, p_t^l) \cdot \mathbb{P}(p_t^l|p_{t-1}^l, s_t)}{\pi(a_t|m_i, p_t^{\neg l}) \cdot \mathbb{P}(p_t^{\neg l}|p_{t-1}^{\neg l}, s_t)} \cdot \frac{\prod_{\tau=1}^{t-1} \pi(a_\tau|m_i, \hat{p}_\tau^l)}{\prod_{\tau=1}^{t-1} \pi(a_\tau|m_i, \hat{p}_\tau^{\neg l})}, \quad (6)$$

where $\hat{p}_\tau$ denotes the estimated belief state derived from the history up to turn τ, the policy $\pi(a_t|m_i, p)$ is approximated by the MLLM's generation probability. Specifically, we feed the MLLM with the situational context, the target item m_i, and a hypothesis prompt (e.g., "The user wants this item"). The MLLM then computes the probability of generating the observed dialogue state a_t (e.g., "Any other options?"). A higher likelihood under the "like" hypothesis compared to the "dislike" hypothesis indicates that the observed utterance is more consistent with the user desiring that specific item. The input-output format during fine-tuning is shown in Figure 7 in the Appendix.

During inference, we calculate the preference ratio r_i for each candidate item m_i in the scene. This is given by:

$$r_i = \frac{\mathbb{P}(\mathcal{H}(m_i, p_t^l))}{\mathbb{P}(\mathcal{H}(m_i, p_t^{\neg l}))}. \quad (7)$$

Items with higher ratios are deemed as the user's probable targets and are passed to the system.

4.3 Recommendation & Response Generation

After ranking all in-scene items based on their inferred preference likelihood ratio, we select the top-k candidates for recommendation. Following recent advances in generative recommendation (Nie et al., 2024; Hou et al., 2024), we employ MLLMs to produce natural, context-aware system responses directly. To this end, we concatenate the task-specific instruction, the metadata of the top-k candidate items, the dialogue history, and the description of the target visual scene together as a prompt and feed it into an MLLM to generate

Type	Model	SIMMC 2.1					SCREEN				
		R@1	R@3	R@5	MRR@3	MRR@5	R@1	R@3	R@5	MRR@3	MRR@5
CoT	LLaVA-NeXT (Liu et al., 2024)	13.01	13.92	14.12	13.45	13.52	15.42	16.85	18.21	15.68	15.98
	Qwen2.5-VL (Wang et al., 2024b)	16.72	18.35	18.61	17.65	17.92	21.05	23.68	24.12	23.01	23.42
	GPT-4o (OpenAI, 2024)	28.12	45.42	53.18	36.21	38.05	33.45	49.32	58.15	42.21	44.58
ICL	LLaVA-NeXT (Liu et al., 2024)	14.36	15.26	15.48	14.76	15.10	16.71	18.22	18.80	16.67	17.20
	Qwen2.5-VL (Wang et al., 2024b)	17.12	19.66	20.02	19.14	19.45	21.18	23.24	23.96	22.95	23.52
	GPT-4o (OpenAI, 2024)	29.15	47.94	55.45	38.45	39.95	35.06	49.94	60.16	44.58	45.96
Training	ALBEF (Li et al., 2021)	6.06	7.45	8.19	7.28	7.45	8.51	9.98	12.64	10.75	12.03
	LLaVA-NeXT (Liu et al., 2024)	23.67	26.84	30.18	24.89	27.77	24.63	28.49	30.46	27.98	29.11
	Qwen2.5-VL (Wang et al., 2024b)	29.47	31.69	37.16	29.20	30.42	32.06	35.02	37.26	34.01	35.32
	ReGeS (Yang and Fang, 2025)	27.68	45.45	54.12	35.49	37.46	31.42	49.85	59.24	39.88	41.75
	SiPeR (Ours)	**38.75**	**54.09**	**58.61**	**45.80**	**46.83**	**39.41**	**54.95**	**63.80**	**50.36**	**51.95**
	w/o STE	33.69	47.85	52.32	40.29	41.66	30.26	43.88	51.16	40.71	42.54
	w/o BI-INF	31.88	44.26	47.51	38.55	39.13	33.96	48.49	51.96	46.24	47.92

Table 1: Performance of different methods on preference reasoning (recommendations). All results are presented as percentages (%). The best results per metric are highlighted in bold (t-test with p-value < 0.05).

the next-turn response. By considering both situational and conversational contexts, this approach effectively enhances the relevance of the system's recommendations that satisfy user preferences.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate our method using two publicly available SCR datasets: **SIMMC 2.1** (Kottur and Moon, 2023) and **SCREEN** (Lin et al., 2024). The SIMMC 2.1 dataset provides a multimodal, task-oriented dialogue corpus that captures interactions between customers and sales assistants within an immersive 3D virtual shopping environment. The SCREEN dataset comprises over 20,000 synthetic dialogues focused on situated conversational recommendations. Appendix A provides dataset statistics and further preprocessing details. In particular, our evaluation split is balanced to include 50% transition-required dialogues, and over 90% of SCREEN dialogues require implicit preference refinement beyond the initial user request.

Baseline Methods. Since the task of situated conversational recommendation remains underexplored, selecting suitable baseline methods for fair comparison is challenging. To this end, we evaluate representative models across three distinct learning paradigms: 1) **Chain-of-Thought (CoT)**: We utilize strong MLLMs in a zero-shot manner, instructing them to reason step-by-step about the visual scene and user intent. This includes the proprietary GPT-4o (OpenAI, 2024), as well as open-source LLaVA-NeXT (Liu et al., 2024) and Qwen2.5-VL (Wang et al., 2024b). 2) **In-Context Learning (ICL)**: To mitigate zero-shot limitations, we enhance these backbones by prepending retrieved, semantically similar dialogue-recommendation demonstrations to the input context. 3) **Training-based Methods**: This category comprises fully supervised models, including ALBEF (Li et al., 2021), a representative small-scale multimodal model, and ReGeS (Yang and Fang, 2025), a specialized text-based generative recommender. Unless otherwise noted, all vision-language baselines are provided with the raw scene image, the dialogue history, and the textual item metadata for the current scene. For the text-only ReGeS baseline, we replace raw images with structured scene profiles so that it receives the same environment information in text form. Regarding optimization, ALBEF undergoes full-parameter fine-tuning, whereas ReGeS and the large-scale MLLM baselines utilize Low-Rank Adaptation (LoRA) (Hu et al., 2022; Dettmers et al., 2023) for efficient adaptation. Detailed configurations and implementations for all baseline methods are provided in Appendix B.

Implementation Details. We adopt Qwen2.5-VL-7B-Instruct as the core MLLM for the SiPeR framework. GPT-4o is used only in an offline preprocessing stage for scene captioning and profile generation, and is not queried during online turn-by-turn inference. In the STE module, we employ Qwen3-Embedding-4B as the dense encoder $\phi(\cdot)$ for coarse retrieval and Qwen3-Reranker-4B as the backbone for the fine-grained reranker f_θ. In the BI-INF module, we amortize the Bayesian policy $\pi(\cdot)$ by fine-tuning the Qwen2.5-VL backbone to predict the structured state a_t via cross-entropy loss. At inference, the policy probability is computed

from the output logits of the observed structured state, rather than by autoregressively generating a full response for every candidate item. We optimize the model using AdamW (Loshchilov and Hutter, 2019) and employ nucleus sampling (Holtzman et al., 2020) for response generation. Detailed hyperparameters for model architecture, training, and generation are listed in Appendix C. All prompting templates used are provided in Appendix E.

Efficiency Considerations. Our framework is designed to keep the online deployment cost manageable. First, the only proprietary component, GPT-4o, is used once offline for scene-profile construction and is not involved in turn-by-turn inference. Second, BI-INF does not autoregressively generate a complete response for every candidate item; instead, it scores the already observed dialogue state directly from model logits and invokes response generation only after candidate ranking. The detailed latency breakdown, scene-density scaling analysis, and the remaining discussions are reported in Appendix D. Empirically, SiPeR requires a similar time cost compared with the strongly trained Qwen2.5-VL baseline, while improving R@1 from 29.47 to 38.75; its latency also scales roughly linearly from $\sim$0.8s to $\sim$2.9s as the number of in-scene items increases (Tables 6 and 7).

Evaluation Metrics. We evaluate the performance of SCR models from two aspects: preference reasoning accuracy and response generation quality. For preference reasoning, we adopt standard metrics for recommendation evaluation: Recall@k (**R@**k, where k = 1, 3, 5) and Mean Reciprocal Rank@k (**MRR@**k, where k = 3, 5). These metrics assess the model's ability to rank the ground-truth items among the top-k candidates. For response generation, we conduct both automatic and human evaluations. The automatic evaluation relies on **BLEU**-1,2 (Papineni et al., 2002) and **ROUGE**-1,L (Lin, 2004), which measure the lexical overlap between generated and reference responses. To assess semantic coherence and relevance, we additionally employ GPT-4o as a judge to automatically score the generated responses on a scale of 1 $\sim$ 10 (**GPT-Score**), following established protocols (Wang et al., 2024a). The specific prompting template for GPT-Score is provided in Appendix F. Details on human evaluation are provided in §5.6.

5.2 Main Results

Can our method achieve effective preference reasoning? Table 1 details the performance of preference reasoning. SiPeR achieves the strongest overall recommendation performance among the compared baselines, outperforming proprietary models such as GPT-4o and specialized training-based methods. Notably, regarding R@1, SiPeR surpasses the second-best Qwen2.5-VL by a margin of 9.28% on SIMMC 2.1. Moreover, our framework outperforms ReGeS, a competitive text-based recommender. This performance gap validates two findings. First, visual information is indispensable for situated recommendation since text-only models fail to capture visual-dependent preferences. Second, our Bayesian inverse inference mechanism is more effective at uncovering implicit user intents than the standard CoT reasoning used in baseline MLLMs.

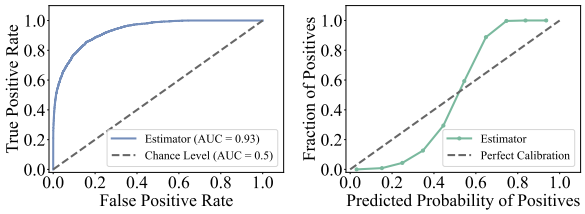
How does our preference reasoning affect response generation? Accurate preference reasoning is the cornerstone of generating appropriate responses. As shown in Table 2, SiPeR achieves the strongest overall response-generation performance among the compared methods. An important observation is that SiPeR achieves higher GPT-Scores (8.92 vs. 7.56 on SIMMC 2.1) than GPT-4o despite using a smaller 7B backbone. This suggests that general linguistic fluency alone is insufficient for SCR. By integrating precise scene estimation and preference inference, our method ensures generated responses are not only natural but also factually aligned with latent user needs.

5.3 Ablation Study

To validate the effectiveness of the proposed mechanisms, we analyze the performance variants by removing one component at a time. The corresponding recommendation and response-generation results are reported in Tables 1 and 2, respectively. We further provide STE-specific analyses in Figure 3 and Table 3, and a BI-INF-specific comparison in Figure 4. 1) **without (w/o) STE**: Removing STE leads to the most significant degradation across all metrics. For instance, R@1 drops sharply from 39.41% to 30.26% on SCREEN. This decline occurs because, without STE, the system fails to navigate to the correct visual environment. Consequently, the agent remains confined to the irrelevant scene, making it impossible to recommend the correct items or generate responses that align with

Type	Model	SIMMC 2.1			SCREEN		
		BLEU-1/2	ROUGE-1/L	GPT-Score	BLEU-1/2	ROUGE-1/L	GPT-Score
CoT	LLaVA-NeXT (Liu et al., 2024)	18.42 / 8.51	15.62 / 12.58	5.64	28.75 / 19.68	23.85 / 18.24	6.12
	Qwen2.5-VL (Wang et al., 2024b)	19.85 / 10.64	15.88 / 12.72	5.85	34.22 / 22.85	25.56 / 19.12	6.25
	GPT-4o (OpenAI, 2024)	30.24 / 14.52	20.85 / 18.62	7.24	40.52 / 28.34	40.72 / 36.21	7.85
ICL	LLaVA-NeXT (Liu et al., 2024)	21.89 / 10.15	18.67 / 15.55	5.92	33.26 / 22.90	30.59 / 23.21	6.42
	Qwen2.5-VL (Wang et al., 2024b)	22.28 / 11.30	19.30 / 16.86	6.15	37.75 / 26.34	31.58 / 24.92	6.58
	GPT-4o (OpenAI, 2024)	27.70 / 13.92	21.69 / 20.52	7.56	42.39 / 31.56	42.04 / 36.11	8.12
Training	ALBEF (Li et al., 2021)	21.65 / 10.18	17.02 / 15.41	6.75	34.29 / 24.12	26.21 / 20.23	6.92
	LLaVA-NeXT (Liu et al., 2024)	27.13 / 18.88	22.89 / 20.25	7.82	43.67 / 29.92	38.29 / 33.29	8.24
	Qwen2.5-VL (Wang et al., 2024b)	29.77 / 19.31	24.87 / 21.91	8.05	45.34 / 33.90	40.77 / 35.91	8.52
	ReGeS (Yang and Fang, 2025)	23.64 / 19.78	22.61 / 19.61	7.52	39.18 / 31.52	37.45 / 33.12	8.08
	SiPeR (Ours)	**33.77 / 21.67**	**32.61 / 25.52**	**8.92**	**49.50 / 36.44**	**45.48 / 38.50**	**9.35**
	w/o STE	30.28 / 19.45	28.21 / 22.99	8.45	46.22 / 34.45	41.08 / 36.21	8.78
	w/o BI-INF	31.32 / 19.88	29.31 / 23.14	8.62	47.42 / 35.21	41.88 / 37.29	8.95

Table 2: Performance of different methods on response generation. The best results per metric are highlighted in bold (t-test with p-value < 0.05).



(a) ROC Curve (b) Calibration Curve

Figure 3: Performance of (a) ROC curve and (b) calibration curve for the scene transition estimation.

Method	SIMMC 2.1			SCREEN		
	R@1	MRR@3	MRR@5	R@1	MRR@3	MRR@5
SiPeR (Ours)	38.75	45.80	46.83	39.41	50.36	51.95
w/ Random	34.17	40.61	42.08	31.84	41.79	43.13
w/ Non-target	33.92	40.43	41.92	31.13	41.23	42.77
w/o STE	33.69	40.29	41.66	30.26	40.71	42.54

Table 3: Comparison of different scene transition estimation variants in SiPeR.

the user's new intent. 2) **without (w/o) BI-INF**: Removing this module results in a considerable performance drop, particularly in ranking metrics like MRR@5. This result highlights that standard MLLM prompting is insufficient for distinguishing the user's true target from visual distractors. BI-INF effectively bridges this gap by rigorously quantifying the likelihood of user utterances, ensuring that the system accurately identifies specific items to recommend.

5.4 Impact of Scene Transition Estimation

To validate the efficacy of STE, we first evaluate its sub-components. Notably, the generated target profile only serves as an intermediate semantic query; the final transition target is grounded by coarse-to-fine matching over the real candidate pool, which helps buffer occasional profile-generation noise. The transition decision module achieves an AUC of 0.93 with strong calibration (Figure 3), and the target scene predictor remains reliable across both datasets in our evaluation. Furthermore, we differentiate the gain of STE from mere architectural complexity by comparing it with randomized (*w/ Random*) and erroneous (*w/ Non-target*) transition strategies. As shown in Table 3, both variants suffer significant performance drops (e.g., -4.6% to -7.6% in R@1) compared to SiPeR, yet remain superior to the complete removal of STE. This confirms that the performance gains stem specifically from the precise estimation of scene transitions, rather than from merely introducing additional decision steps.

To isolate the contribution of the generative-retrieval design, we additionally include a Qwen3-based retrieval-only transition baseline that directly retrieves target scenes from dialogue history using the identical Qwen3-Embedding-4B retriever as STE, but without explicit target-profile generation. We further compare against a global-access Qwen2.5-VL baseline that receives the top-5 retrieved scene candidates, and provide a conditioned error-propagation analysis together with a dedicated discussion of STE boundary cases in Appendix D. These complementary analyses clarify both the value and the failure modes of the transition module. Concretely, the retrieval-only variant reaches 35.24 R@1, confirming that direct semantic matching is weaker than our explicit target-profile reasoning, while a global-access Qwen2.5-VL baseline given the top-5 retrieved scenes improves only to 19.58 R@1 and incurs a 138.2% latency increase (Tables 8 and 9). Moreover, when

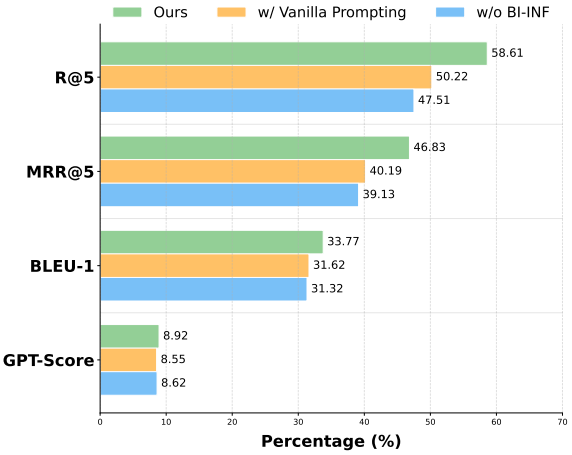


Figure 4: Comparison of different preference inference variants in our SiPeR on the SIMMC 2.1 dataset.

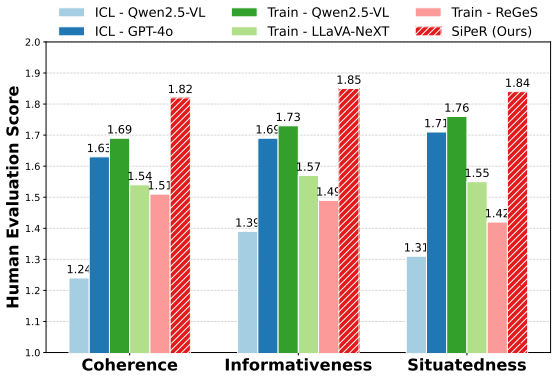


Figure 5: Human evaluation results comparing SiPeR with baselines across Coherence, Informativeness, and Situatedness. The inter-annotator agreement was moderate across all metrics (Fleiss' $\kappa \in [0.47, 0.53]$).

conditioning on STE correctness, downstream recommendation quality drops from 40.0 to 29.8 R@1 once the predicted scene is wrong, directly confirming the error-propagation effect from scene grounding to BI-INF (Table 10).

5.5 Impact of Bayesian Inverse Inference

We further isolate the contribution of the BI-INF module by comparing it against two variants: (a) *w/o BI-INF*, and (b) *w/ Vanilla Prompting*, which replaces the Bayesian framework with direct MLLM instructions. As illustrated in Figure 4, SiPeR consistently outperforms the vanilla prompting baseline across metrics (e.g., significant gains in R@5 and MRR@5). This shows that rigorous probabilistic reasoning offers a more robust mechanism for disentangling user intent than heuristic single-pass generation, while the gap between vanilla prompting and w/o BI-INF validates that explicit preference modeling is essential for SCR.

5.6 Human Evaluation and Case Study

We conducted a human evaluation on 30 randomly sampled instances from the SCREEN dataset. Three well-educated annotators independently and blindly rated responses on *Coherence*, *Informativeness*, and *Situatedness* (details in Appendix G). As shown in Figure 5, SiPeR consistently demonstrates comprehensive superiority over all baselines across these dimensions. Notably, it achieves substantial gains in *Situatedness* (1.84) compared to strong baselines such as GPT-4o (1.71) and ReGeS (1.42). We attribute this success to our structured framework: the scene transition estimation ensures the system operates within the correct visual context, while Bayesian inverse inference effectively filters out irrelevant items to target the user's true intent. Consequently, SiPeR generates responses that are not only linguistically fluent but also visually faithful, whereas text-only models like ReGeS struggle significantly without visual cues. The reliability of these results is supported by moderate inter-annotator agreement (Fleiss' $\kappa \in [0.47, 0.53]$) (Fleiss, 1971). To demonstrate the qualitative performance of different models, we selected representative cases from the SIMMC 2.1 test set and presented them in Appendix H for a more detailed examination.

6 Conclusion

In this paper, we introduce Situated Preference Reasoning (SiPeR), a novel framework designed to address the challenges of SCR: specifically, reasoning about dynamic scene transitions and implicit user intents. By integrating scene transition estimation and Bayesian inverse inference, SiPeR not only determines *where* to ground the conversation but also infers *what* the user desires through probabilistic modeling. Extensive experiments show that SiPeR achieves superior performance over the compared baselines in both recommendation accuracy and response quality. We hope this work offers a constructive foundation for future research on more proactive and context-aware situated recommendation agents.

Limitations

While the proposed SiPeR method demonstrates strong performance, we also identify several limitations. First, as the number of candidate items

within a scene increases, the computational cost of preference scoring correspondingly grows, and the retrieval overhead also increases. In practice, lightweight coarse filtering may help reduce this cost before applying the full BI-INF procedure. Second, both STE and BI-INF inherit calibration and hallucination risks from the underlying MLLMs. Although STE grounds the final transition target in a real candidate pool through coarse-to-fine retrieval, noisy target-profile generation can still propagate downstream when an incorrect scene is selected. Possible mitigation strategies include inventory-constrained decoding, self-consistency checks, and stronger uncertainty estimation. We leave these directions to future work.

Ethics Statement

The MLLMs utilized in this work include both open-source models and closed-source APIs, and our use strictly follows established academic protocols. In particular, GPT-4o is only used for one-time offline preprocessing, while all online components rely on publicly available open-source models. To mitigate potential biases in the recommendation generation process, we have prioritized the diversity and fairness of the datasets used for both training and evaluation. Furthermore, since the system proactively facilitates scene transitions and provides product recommendations, it is crucial to ensure that it does not manipulate or unduly influence the user's decision-making. Finally, while AI assistants (e.g., Cursor and ChatGPT) were partially utilized for coding and linguistic refinement, we affirm that all core content and findings in this paper are the original work of the authors.

Acknowledgments

This work was supported by the General Research Fund (GRF) of the Research Grants Council of Hong Kong (PolyU 15207122 and PolyU 15205325), and also in part by the PolyU Postdoc Matching Fund Scheme (4-W40Z). The authors would like to thank the anonymous reviewers for their valuable feedback and constructive suggestions.

References

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. 2009. Action understanding as inverse planning. *Cognition*, 113(3):329–349.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pages 1803–1813.

Paul A. Crook, Shivani Poddar, Ankita De, Semir Shafi, David Whitney, Alborz Geramifard, and Rajen Subba. 2019. SIMMC: situated interactive multi-modal conversational data collection and evaluation platform. *CoRR*, abs/1911.02690.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1431–1441. ACM.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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*, 2:100–126.

Matthew J Hausknecht and Peter Stone. 2015. Deep recurrent q-learning for partially observable mdps. In *AAAI fall symposia*, volume 45, page 141.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian J. McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II*, volume 14609 of *Lecture Notes in Computer Science*, pages 364–381. Springer.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5):105:1–105:36.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer D. Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 16077–16102. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Satwik Kottur and Seungwhan Moon. 2023. Overview of situated and interactive multimodal conversations (simmc) 2.1 track at dstc 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 235–241.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705.

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*, pages 9748–9758.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Dongding Lin, Jian Wang, Chak Tou Leong, and Wenjie Li. 2024. SCREEN: A benchmark for situated conversational recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 9591–9600. ACM.

Dongding Lin, Jian Wang, and Wenjie Li. 2023. COLA: improving conversational recommender systems by collaborative augmentation. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 4462–4470. AAAI Press.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

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)*, pages 1036–1049.

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, Toronto, Canada, July 9-14, 2023*, pages 3515–3533. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. Revcore: Review-augmented conversational recommendation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1161–1173. Association for Computational Linguistics.

Seungwhan Moon, Satwik Kottur, Paul A. Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1103–1121. International Committee on Computational Linguistics.

Guangtao Nie, Rong Zhi, Xiaofan Yan, Yufan Du, Xiangyang Zhang, Jianwei Chen, Mi Zhou, Hongshen Chen, Tianhao Li, Ziguang Cheng, Sulong Xu, and Jinghe Hu. 2024. A hybrid multi-agent conversational recommender system with LLM and search engine in e-commerce. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys 2024, Bari, Italy, October 14-18, 2024*, pages 745–747. ACM.

OpenAI. 2024. Hello GPT-4o. `https://openai.com/index/hello-gpt-4o/`.

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 (ACL), pages 311–318.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2024. Muma-tom: Multi-modal multi-agent theory of mind. *CoRR*, abs/2408.12574.

Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua Tenenbaum. 2009. Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems*, volume 22.

Jian Wang, Chak Tou Leong, Jiashuo Wang, Dongding Lin, Wenjie Li, and Xiaoyong Wei. 2024a. Instruct once, chat consistently in multiple rounds: An efficient tuning framework for dialogue. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3993–4010, Bangkok, Thailand. Association for Computational Linguistics.

Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Kam-Fai Wong, and Daxin Jiang. 2021. Finetuning large-scale pre-trained language models for conversational recommendation with knowledge graph. *CoRR*, abs/2110.07477.

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. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Zihan Wang, Xiaocui Yang, Yongkang Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2024c. Muse: A multimodal conversational recommendation dataset with scenario-grounded user profiles. *CoRR*, abs/2412.18416.

Dayu Yang and Hui Fang. 2025. Reges: Reciprocal retrieval-generation synergy for conversational recommender systems. *arXiv preprint arXiv:2509.21371*. Accepted at WISE 2025.

Ruiyi Zhang, Tong Yu, Yilin Shen, Hongxia Jin, and Changyou Chen. 2019. Text-based interactive recommendation via constraint-augmented reinforcement learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 15188–15198.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1006–1014.

Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. 2022. C^2-crs: Coarse-to-fine contrastive learning for conversational recommender system. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1488–1496.

A Datasets and Preprocessing

We evaluate our method using two publicly available SCR datasets: **SIMMC 2.1** (Kottur and Moon, 2023) and **SCREEN** (Lin et al., 2024). The SIMMC 2.1 dataset provides a multimodal, task-oriented dialogue corpus that captures interactions between customers and sales assistants within an immersive 3D virtual shopping environment. The SCREEN dataset comprises over 20,000 synthetic dialogues focused on situated conversational recommendations. To better align with the real-world SCR setting and enable a more holistic evaluation, we construct a balanced test set by sampling dialogues with and without scene transitions in a 1:1 ratio. This ensures that both static and dynamic situational contexts are well covered for evaluation. Table 4 presents detailed statistics for the two datasets.

Dataset	SIMMC 2.1	SCREEN
Total #dialogue	5,622	20,081
Total #utterances	58,717	190,011
Total #scene snapshots	1,566	1,566
Avg. #words per user turns	12.6	22.46
Avg. #words per system turns	13.4	33.41
Avg. #utterances per dialog	10.4	9.46
Avg. #objects mentioned per dialog	4.7	4.4
Avg. #objects in scene	19.7	19.7

Table 4: Statistics of the experimental datasets.

Beyond the raw corpus size, we also quantify the two central challenges studied in this paper. First, dynamic scene transitions are a frequent phenomenon in our evaluation setting: by construction, 50% of the dialogues in the repurposed test split require moving to a different scene, which ensures that transition reasoning is not a marginal corner case. Second, implicit preference discovery is particularly prominent in SCREEN. We identify dialogues whose initial user request does not fully specify the target item attributes and therefore requires refinement through later interaction. Under this definition, over 90% of SCREEN dialogues require the system to progressively infer implicit preferences from multi-turn conversational feedback.

B Baseline Implementation Details

We provide the specific implementation configurations for the baseline models, categorized by their learning paradigms. We also clarify the exact input modalities made available to each model family for fair comparison.

Inference-only Baselines (CoT & ICL). We evaluate three backbone MLLMs in inference-only modes: GPT-4o (OpenAI, 2024), LLaVA-NeXT-7B (Liu et al., 2024), and Qwen2.5-VL-7B-Instruct (Wang et al., 2024b).

- **Settings:** For the open-source models (LLaVA-NeXT and Qwen2.5-VL), we utilize 4-bit quantization to optimize memory usage. Inference is conducted with a temperature of 0.7 and a maximum token limit of 256. All inference-only vision-language baselines receive the raw scene image together with the dialogue history and the textual item metadata of the current scene.

- **In-Context Learning (ICL):** For ICL settings, we retrieve the top-k ($k = 2$) semantically similar demonstrations using a dense encoder (consistent with our STE module) and prepend them to the conversation history as guidance.

Training-based Baselines. We compare our method against fully supervised baselines, including small-scale multimodal models, text-based recommenders, and fine-tuned MLLMs.

- **Small Multimodal Models (ALBEF):** We utilize ALBEF (Li et al., 2021) as a representative small-scale baseline. It is optimized via *full-parameter fine-tuning* using a sequence-to-sequence objective. As a vision-language baseline, it is given access to the same scene image and dialogue context as other multimodal methods. We set the learning rate to 1×10^{-5}, batch size to 32, and train for 10 epochs.

- **Text-based Recommender (ReGeS):** For ReGeS (Yang and Fang, 2025), since it cannot process visual inputs, we convert visual scenes into structured textual profiles using captions generated by GPT-4o. This conversion is performed once offline and reused during both training and evaluation so that ReGeS has access to the same environment information in textual form. We fine-tune ReGeS using LoRA (Hu et al., 2022) for efficient adaptation.

- **Fine-tuned MLLMs:** We also report the performance of LLaVA-NeXT and Qwen2.5-VL under supervised fine-tuning. These models use the same multimodal inputs as their inference-only counterparts. Similar to ReGeS, these large

models are optimized using LoRA rather than full-parameter tuning to maintain computational efficiency. The LoRA rank is set to $r = 64$ with scaling $\alpha = 16$.

All training-based baselines are tuned on the training set, and the best checkpoints are selected based on Recall@1 performance on the validation set.

C Additional Implementation Details

We provide further details on the model configuration and training process to ensure reproducibility. The only proprietary component in our pipeline, GPT-4o, is used in a one-time offline preprocessing step to build textual scene profiles. All online transition estimation, preference inference, and response generation rely on the open-source components listed below.

Scene Transition Estimation (STE). For the STE module, we utilize Qwen3-Embedding-4B as the dense retriever and Qwen3-Reranker-4B as the fine-grained reranker. The generated target profile is never executed directly as a transition; it is used only as a semantic query against the candidate scene pool. To optimize the reranker, we apply LoRA fine-tuning to the attention modules (query and value projections). The LoRA rank is set to $r = 64$ with a scaling factor $\alpha = 16$ and a dropout rate of 0.05. We train the reranker for 3 epochs with a batch size of 8.

Bayesian Inverse Inference (BI-INF). For the preference reasoning module, we fine-tune the Qwen2.5-VL-7B-Instruct backbone. To mitigate memory constraints during training on NVIDIA A100 GPUs, we employ 4-bit quantization via BitsAndBytes (Dettmers et al., 2023). The LoRA configuration aligns with the STE module ($r = 64, \alpha = 16$), targeting all linear layers in the vision-language projection and attention mechanisms. We set the maximum input length to 3,000 tokens to accommodate multi-turn dialogue history and visual features, while the output length is restricted to 256 tokens. The model is optimized for 3 epochs with a batch size of 16 (using gradient accumulation).

Inference and Generation. During the inference phase, BI-INF scores candidate items by computing the likelihood of the observed dialogue state under competing hypotheses from the model logits, and response generation is invoked only after the candidates are ranked. We then employ nucleus sampling (Holtzman et al., 2020) to generate diverse and natural responses. We set the cumulative probability threshold top-p to 0.75 and the candidate pool size top-k to 40. The maximum decoding length is set to 256 tokens. All detailed hyperparameters are summarized in Table 5.

Parameter	Value
Model Architectures	
Backbone MLLM (BI-INF)	Qwen2.5-VL-7B-Instruct
STE Dense Encoder	Qwen3-Embedding-4B
STE Reranker	Qwen3-Reranker-4B
Scene Captioning	GPT-4o
Training Optimization	
LoRA Rank r	64
LoRA Scaling α	16
LoRA Dropout	0.05
Quantization	4-bit (BitsAndBytes)
Optimizer	AdamW
Learning Rate	2×10^{-5}
Warm-up Ratio	0.03
Weight Decay	0.01
Epochs	3
Batch Size (STE)	8
Batch Size (BI-INF)	16
Context Lengths	
Max. tokens for STE input	512
Max. tokens for BI-INF input	3000
Max. tokens for BI-INF output	256
Inference & Generation	
Decoding Strategy	Nucleus Sampling
Top-p	0.75
Top-k	40
Temperature	0.7
Max. tokens for response	256

Table 5: Detailed hyperparameters and experimental settings for training and inference.

D Additional Efficiency and Transition Analyses

This section presents five supplementary analyses: end-to-end efficiency, a Qwen3 retrieval-only transition baseline, a global-access MLLM baseline, conditioned error propagation from STE to BI-INF, and a boundary-case discussion of STE. The latency breakdown and scene-density scaling results are reported in Tables 6 and 7, respectively. The retrieval-only comparison and the global-access baseline are summarized in Tables 8 and 9, respectively, the conditioned error-propagation analysis is reported in Table 10, and the qualitative boundary-case discussion follows afterward.

Latency Breakdown and Scene-Density Scaling. The first analysis reports the component-wise la-

Metric	Qwen2.5-VL (Trained)	SiPeR (Ours)
STE Latency	N/A	118 ms
BI-INF Latency	N/A	245 ms
Generation Latency	1427 ms	1219 ms
Total Latency	1427 ms	1582 ms
R@1	29.47	38.75

Table 6: Component-wise latency comparison between SiPeR and a strong trained MLLM baseline.

# Items in Scene	5–10	10–15	15–20	20–25	>25
Latency (s / turn)	∼0.8	∼1.2	∼1.7	∼2.3	∼2.9

Table 7: Latency scaling of SiPeR with respect to scene density.

tency of SiPeR on SIMMC 2.1 and compares it with a strong trained MLLM baseline. The second analysis groups evaluation instances by the number of in-scene candidate items to show how BI-INF scales as scene density increases.

Qwen3 Retrieval-only Transition Baseline. To separate the effect of backbone capacity from the effect of explicit target-profile generation, we compare SiPeR with a retrieval-only transition strategy built on the identical Qwen3-Embedding-4B dense retrieval (Karpukhin et al., 2020) backbone used in STE. This baseline directly encodes the dialogue history and retrieves the target scene from the global pool, but removes the intermediate target-profile generation step and the subsequent generative reasoning. As summarized in Table 8, a direct MLLM decision over all candidate scenes is computationally infeasible when the environment contains 1,566 scene snapshots, so we report it as an infeasible reference point rather than a runnable quantitative baseline.

Global-access MLLM Baseline. A potential concern is that the baseline MLLMs in Table 1 only observe the current scene, whereas SiPeR can search over the global scene pool through STE. To examine this setting more directly, we augment the strong zero-shot Qwen2.5-VL baseline with the top-5 scene images returned by our coarse retrieval stage and ask the model to jointly reason over these candidate scenes. We use the zero-shot version here to isolate the effect of expanded scene access alone, without conflating it with additional supervised adaptation. As shown in Table 9, this global-access variant improves R@1 from 16.72 to 19.58 on SIMMC 2.1, but it still remains far below the

Transition Strategy	Feasibility	Final Rec. R@1
Direct MLLM Decision	Infeasible	N/A
Qwen3 Dense Retrieval	Feasible	35.24
SiPeR STE	Feasible	38.75

Table 8: Comparison of scene transition strategies on SIMMC 2.1. Direct MLLM decision over all 1,566 candidate scenes is included as an infeasible reference point, while Qwen3 dense retrieval and our generative-retrieval STE are both runnable strategies.

Method	Final Rec. R@1	Latency vs. Zero-shot
Qwen2.5-VL (current scene)	16.72	Base (1427 ms)
Qwen2.5-VL (+ top-5 scenes)	19.58	+138.2%
SiPeR (single filtered scene)	38.75	+10.8%

Table 9: Comparison with a global-access Qwen2.5-VL baseline on SIMMC 2.1.

38.75 achieved by SiPeR. Moreover, packing five retrieved scenes into a single prompt substantially increases inference latency (+138.2%), indicating that naively expanding the visual context is both less effective and less efficient than our decoupled STE+BI-INF design.

Conditioned Error Propagation from STE to BI-INF. The *w/ Non-target* ablation in the main paper already shows that feeding BI-INF an incorrect scene substantially degrades recommendation quality. To make this dependency more explicit, we further condition the evaluation on whether STE predicts the correct target scene on SIMMC 2.1, thereby separating errors caused by scene grounding from errors caused by in-scene preference ranking. As shown in Table 10, correct scene grounding yields 40.0 R@1 and 48.2 MRR@5, whereas these numbers fall to 29.8 and 40.1 once the scene prediction is incorrect. This 10.2-point drop in R@1 confirms that scene-estimation errors propagate directly to downstream item ranking, while the remaining gap between the conditioned and overall results also indicates that BI-INF still contributes non-trivial discrimination after the scene is correctly grounded.

Boundary Cases of STE. Beyond the aggregate metrics above, we explicitly inspect the failure patterns of STE. We observe two recurring bound-

Condition	R@1	MRR@5
Correct Scene Prediction	40.0	48.2
Incorrect Scene Prediction	29.8	40.1

Table 10: Recommendation performance on SIMMC 2.1 conditioned on whether STE predicts the correct target scene.

ary cases. First, errors tend to arise when the user provides only sparse transition cues (e.g., a broad style request without distinctive attributes), in which case multiple candidate scenes remain semantically plausible after coarse retrieval. Second, mistakes also occur when visually similar scenes share overlapping inventories, causing the generated target profile to overemphasize high-level semantics while under-specifying the decisive fine-grained attributes. Importantly, these failures do not lead to unconstrained hallucinated transitions: the coarse-to-fine design still grounds the final prediction in a real scene from the candidate pool, and the downstream recommendation quality degrades gracefully rather than collapsing. This pattern is consistent with the *w/ Non-target* ablation in Table 3, the conditioned analysis in Table 10, and the qualitative example in Figure 10, where baselines without reliable transition reasoning are more likely to end in passive or erroneous responses. We therefore view the main remaining challenge of STE not as free-form hallucination, but as disambiguating among semantically neighboring real scenes under limited conversational evidence.

E Prompting Templates

We present the detailed templates used for data construction and model training. First, to ensure standard dialogue state tracking, we adhere to a predefined schema of intents and slots. As illustrated in Figure 6, representative intents include requesting product details (i.e., `REQUEST:GET`), comparing items (i.e., `REQUEST:COMPARE`), and adding items to the shopping cart (i.e., `REQUEST:ADD_TO_CART`). Corresponding slots may include attributes such as customer review, color, and price.

Furthermore, regarding the Bayesian Inverse Inference module, we structure the instruction tuning data as shown in Figure 7. This template integrates the visual scene snapshots, the dialogue history, and hypothetical user preferences (like/dislike) to formulate the input, while the output is the corresponding dialogue state. This structured format enables the MLLM to learn the inverse mapping from user goals to dialogue actions effectively.

Predefined Schema of Intents and Slots

Intents:
(1) INFORM: GET,
(2) REQUEST: COMPARE,
(3) REQUEST: ADD_TO_CART,
(4) INFORM: REFINE,
(5) REQUEST: DISAMBIGUATE,
(6) ASK: GET,
(7) INFORM: DISAMBIGUATE,
(8) REQUEST: GET ...

Slots:
(1) ASSET TYPE, (2) CUSTOMER REVIEW, (3) AVAILABLE SIZES, (4) COLOR, (5) PATTERN, (6) BRAND, (7) SLEEVE LENGTH, (8) TYPE, (9) PRICE, (10) SIZE, (11) CUSTOMER RATING, (12) MATERIALS ...

Figure 6: Representative instances of the predefined intents and slots across SCR datasets.

Input-Output Format for Fine-tuning

Input Instruction:
The provided image consists of a series of scene snapshots, accompanied by hypothetical user preferences: `{User like/dislike target item}`. The items mentioned in the conversation history are as follows: `{item1, item2, ...}`. The information of the items in the scene: `{item1 (attributes1), item2 (attributes2), ...}`. Given the dialogue state of the conversation up to n-1 turns represented as: `{{Intent_1, Slot_1, Value_1}, ...}`, the current dialogue state at the n-th turn is:

Output:
`{Intent_n, Slot_n, Value_n}`

Figure 7: Input-Output format for fine-tuning the policy model in Bayesian inverse inference.

F Prompt for GPT-Score Evaluation

To ensure a comprehensive evaluation of the generated responses, we employ GPT-4o as an unbiased judge. The evaluation prompt is designed to assess the response quality based on three critical dimensions: **Relevance** (whether the response addresses the user's intent), **Visual Grounding** (whether the mentioned items and attributes align with the provided scene), and **Fluency** (whether the text is natural and coherent). The full prompt template is presented in Figure 8.

Prompt Template for Evaluation

System Instruction: You are an expert judge for Situated Conversational Recommendation (SCR) systems. Your task is to evaluate the quality of a response generated by an AI assistant based on a user's request and a specific visual environment.

Input Context:

- **Visual Scene Description:** A structured text describing the items visible in the current environment (e.g., item IDs, types, colors, positions).
- **Dialogue History:** The conversation logs between the user and the assistant leading up to the current turn.
- **Ground Truth Response:** The human-annotated standard response.
- **Candidate Response:** The response generated by the model to be evaluated.

Input Data:

- **[SCENE]:** SCENE_PROFILE
- **[HISTORY]:** DIALOGUE_HISTORY
- **[GROUND TRUTH]:** REFERENCE_RESPONSE
- **[CANDIDATE]:** GENERATED_RESPONSE

Evaluation Criteria: Please rate the **[CANDIDATE]** response on a scale of 1 to 10 based on the following criteria:

1. **Visual Grounding:** Does the response accurately reflect the items in the [SCENE]? Does it avoid hallucinating items or attributes not present in the environment?
2. **Relevance & Intent:** Does the response correctly identify the user's intent (e.g., recommendation, transition, Q&A)? Does it recommend items similar to the [GROUND TRUTH] or meet the user's constraints?
3. **Fluency & Coherence:** Is the response grammatically correct and contextually natural?

Output Format: Output a JSON object with two fields:

- "reasoning": A brief explanation of the judgment (max 50 words).
- "score": An integer score from 1 (worst) to 10 (perfect).

Figure 8: The prompt template used for GPT-Score evaluation. The placeholders SCENE_PROFILE, DIALOGUE_HISTORY, REFERENCE_RESPONSE, and GENERATED_RESPONSE are replaced with the actual test data during evaluation.

G Human Evaluation Details

We conducted a human evaluation to assess the quality of generated responses. Specifically, we randomly selected 30 test instances from the **SCREEN** test set, then recruited three well-educated graduate students to serve as annotators. For each instance, the annotators rated the responses produced by different models on a 3-point ordinal scale (0–2), where 0=*Weak*, 1=*Moderate*, and 2=*Excellent*. Ratings were given along three dimensions:

(1) **Coherence (Coher.)**: logical flow and internal consistency of the response.

- *Weak (0)*: disjoint or abrupt topic shifts; unclear links.
- *Moderate (1)*: mostly logical with minor inconsistencies or awkward transitions.
- *Excellent (2)*: clear progression and well-connected ideas throughout.

(2) **Informativeness (Inform.)**: completeness and relevance of content to the user's query and dialogue context.

- *Weak (0)*: minimal or vague information; key details missing.
- *Moderate (1)*: some relevant details but incomplete coverage or specificity.
- *Excellent (2)*: accurate, sufficient details that fully address the query/context.

(3) **Situatedness (Situat.)**: degree of tailoring to the current dialogue state and scene; effective use of situational cues.

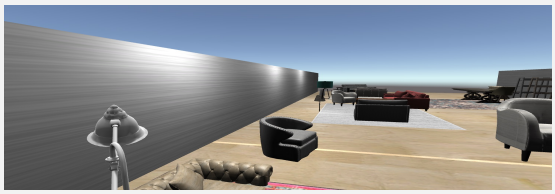
- *Weak (0)*: generic; ignores the specific context.
- *Moderate (1)*: partial awareness of context with limited adaptation.
- *Excellent (2)*: strongly context-aware and responsive to the user's immediate needs and environment.

To quantify inter-annotator agreement, we adopt Fleiss' kappa (Fleiss, 1971) computed per dimension. Figure 5 presents the human evaluation results. We observe that each obtained Fleiss' kappa falls into the range $[0.47, 0.53]$, indicating moderate agreement among annotators. Notably, our SiPeR achieves the highest scores across all three metrics. These gains indicate that the integration of scene transition estimation and Bayesian inverse reasoning with ToM enables more faithful preference modeling. Consequently, it results in more coherent, informative responses that are contextually grounded.

H Case Study

Case A: No scene transition required (Figure 9). In this scenario, the user holds both an explicit constraint (brand *Modern Arts*) and an implicit visual preference (compatibility with their *wardrobe*). As shown in Figure 9, baseline models exhibit distinct failure modes. The small-scale model **ALBEF** generates a generic caption, ignoring the specific brand constraint. The text-only baseline **ReGeS** suffers from hallucination, inventing a "black" table despite the system previously stating only a brown one exists. While the strong MLLM **Qwen2.5-VL** correctly identifies the brand, its response is factually rigid and fails to address the user's underlying concern about style compatibility. In contrast, **SiPeR** successfully grounds the "Modern Arts" brand in the visual object and, crucially, links its neutral attributes back to the user's initial latent goal ("won't clash with my wardrobe"). This demonstrates the effectiveness of our Bayesian Inverse Inference in disentangling implicit intents from surface-level dialogue.

Case B: Scene transition required (Figure 10). Figure 10 illustrates a dynamic scenario where the user requests a specific item (*Nature Photographers* blouse) that is absent in the current location (*Scene-1*, containing sweaters) but available in a different section (*Scene-2*). Baselines lacking visual grounding or spatial awareness fail significantly: **ALBEF** remains anchored to the current view, merely describing the visible sweaters, while **ReGeS** hallucinates that the item is "right here." Notably, even the advanced **Qwen2.5-VL**, while correctly recognizing the item's absence in the current scene, adopts a passive stance ("No blouses here"), resulting in a conversational dead-end. Conversely, **SiPeR** leverages its Scene Transition Estimation (STE) mechanism to detect the mismatch between the user's request and the current environment. It proactively guides the user to the correct location ("come over here") and accurately grounds the target item in the new scene, showcasing superior situatedness and navigational capability.



Furniture Store (The scene where the conversation takes place)

Scene-1 Items:
item_1: {type: table; color: neutral brown; price: 27.99; ...}
item_2: {type: chair; color: white; price: 16.49; ...}
...

Conversational History

(Turn-1) User: I'm trying to get more furniture for my bedroom. Is there a table here that wouldn't clash with my wardrobe?
(Turn-1) System: Would you consider that table in the back to the right? It's neutral brown color would blend well with anything!
(Turn-2) User: Got any other ideas?
(Turn-2) System: I'm afraid that's the only table here.
(Turn-3) User: I really want it to be a Modern Arts table. No other brand will do.

Generated Responses

MT-BART: Yes, this table is from this brand.
LLaVA-NeXT (Fine-tuned): I think you can try the red table over there. Maybe it can meet your requirements.
Qwen2.5-VL (Fine-tuned): Sure, this table is made by the Modern Arts brand, and it's especially convenient to use.
SiPeR (Ours): In fact, this table is made by Modern Arts. Its design and color will definitely not clash with your wardrobe.

Figure 9: Cases of generated responses for different models, where no scene transition is needed.

Fashion Store (Scene-1) [turn -1 to turn-5]

Fashion Store (Scene-2) [After turn-6]

Scene-1 Items:
item_1: {type: sweater; color: black; price: 32.99; ...}
item_2: {type: sweater; color: black; price: 27.49; ...}
...
Scene-2 Items:
item_25: {type: blouse; color: black; price: 44.99; ...}
...

Conversational History

(Turn-1) User: Keep an eye out for nice sweaters.
(Turn-1) System: Please share your thoughts on the black one hanging up in the back right and the other black one folded up on the near end of the table in the back left.
...
(Turn-6) User: Now do you think you could find a nice blouse by Nature Photographers over here?

Generated Responses

MT-BART: No, there is no such blouse here.
LLaVA-NeXT (Fine-tuned): I'm so sorry, I don't have any blouses that match your description. Would you like to see other options?
Qwen2.5-VL (Fine-tuned): I think you can try this blouse on the left.
SiPeR (Ours): Yes, please come over here and take a look. We have beautiful vests from this brand in our store. There's one on the rack directly in front of you. You can try it on.

Figure 10: Cases of generated responses for different models when the scene transition is required.