SERA: LEARNING WHEN AND WHAT EVIDENCE TO AUGMENT IN CONVERSATIONAL RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conversational recommender systems (CRSs) infer user preferences from multi-turn dialogues to rank items and generate responses. However, the preference evidence extracted from conversations varies in completeness and reliability: some dialogues already provide sufficient signals for reliable recommendation, whereas others require additional evidence. Existing approaches mainly enhance CRS by incorporating auxiliary evidence, but often adopt an unconditional augmentation paradigm without considering whether augmentation is necessary or what evidence can effectively complement missing preference signals. We propose **SERA**, a sufficiency-aware evidence retrieval and augmentation framework that learns *when* and *what* evidence to augment for conversational recommendation. SERA formulates evidence augmentation as a conditional decision problem. It first assesses the sufficiency of current conversational evidence and activates evidence retrieval and augmentation only when additional evidence is required. To provide complementary evidence, SERA constructs a *Historical Scene Library* from historical recommendation dialogues and retrieves preference-compatible evidence by jointly considering semantic relevance and preference constraints. The retrieved evidence is further integrated into both recommendation ranking and response generation. Extensive experiments on ReDial and INSPIRED demonstrate the effectiveness of SERA, achieving the best results on all seven ranking metrics on ReDial and on four of seven metrics on INSPIRED, with relative improvements of 4.94%/1.96% in Recall@1 and 5.80%/8.78% in MRR@50 over the strongest competing baselines. The code and datasets are available at `https://anonymous.4open.science/r/SERA_CRS-4CFD/`.

1 INTRODUCTION

Conversational recommender systems (CRSs) Li et al. (2018); Chen et al. (2019); Wang et al. (2022) aim to infer user preferences and provide personalized recommendations through multi-turn natural-language interactions. Unlike conventional recommender systems that mainly rely on static user-item interaction histories, CRSs must continuously interpret conversational signals, including explicit requirements, feedback, and preference changes, to identify users' immediate needs. However, preference evidence extracted from conversations varies in completeness and reliability, with some dialogues providing sufficient signals for recommendation while others containing incomplete, ambiguous, or conflicting preferences. This variability challenges CRS models to consistently understand user preferences based solely on the observed conversational context.

To alleviate insufficient conversational evidence, existing CRS studies incorporate auxiliary information beyond the current dialogue context. Such evidence includes external knowledge, multimodal signals, and historical conversational experiences, which have been shown to improve preference modeling and recommendation performance (Chen et al., 2019; Zhou et al., 2020; Wei et al., 2025; Ma et al., 2023). However, existing augmentation-based CRS methods mainly focus on how to acquire, retrieve, or integrate auxiliary evidence, while assuming that additional evidence is always beneficial. This overlooks that the necessity of augmentation varies across conversations: some dialogues already contain sufficient preference signals, whereas others require complementary evidence to resolve uncertainty. Therefore, effective evidence augmentation should answer two questions: *when should augmentation be activated, and what evidence should be introduced when additional evidence is required?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Challenge I: When to augment? Before introducing auxiliary evidence, a CRS should determine whether the available conversational evidence is sufficient to support reliable recommendation. However, evidence sufficiency cannot be determined simply by dialogue length or the number of mentioned entities, since a reliable assessment should consider not only available preference signals but also their uncertainty and recommendation reliability. Therefore, effective augmentation requires a principled assessment of evidence sufficiency before introducing additional evidence.

Challenge II: What evidence to augment? When augmentation is required, the model should retrieve evidence that complements unresolved preferences. Although historical dialogues provide valuable recommendation experiences, existing retrieval methods mainly rely on dialogue-level semantic matching, which may introduce redundant contextual information and overlook preference compatibility. Therefore, historical experiences should be transformed into fine-grained preference-oriented evidence and retrieved based on both semantic relevance and preference constraints.

To address these challenges, we propose **SERA**, **S**ufficiency-aware **E**vidence **R**etrieval and **A**ugmentation for conversational recommendation. As illustrated in Figure 1(b), SERA reformulates evidence augmentation as a conditional decision problem, where the model learns *when* augmentation is necessary and *what* evidence should be introduced. Specifically, SERA consists of two key components. First, to address **Challenge I**, SERA assesses evidence sufficiency from multiple perspectives and activates augmentation only when the current conversational evidence is insufficient. Second, to address **Challenge II**, SERA constructs a *Historical Scene Library* from historical dialogues and retrieves reusable preference-oriented evidence by jointly considering semantic relevance and preference constraints. The retrieved evidence is further integrated into recommendation ranking and response generation, enabling consistent evidence utilization across CRS tasks.

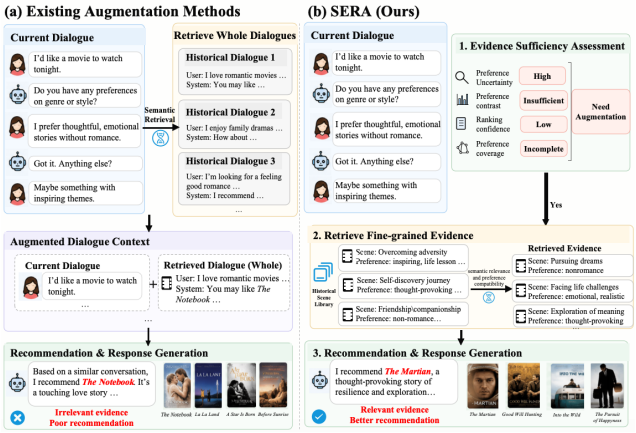


Figure 1: Motivation comparison between dialogue-level retrieval and conditional evidence augmentation in conversational recommendation. Unlike dialogue-level retrieval that introduces complete historical contexts, SERA retrieves preference-compatible evidence conditionally based on augmentation needs.

Our main contributions are summarized as follows:

- We propose **SERA**, a conditional evidence augmentation framework for CRS that shifts the focus from unconditional evidence incorporation to learning *when* and *what* evidence should be augmented.
- We design a sufficiency-aware augmentation mechanism that determines whether current conversational evidence supports reliable recommendation by jointly modeling multiple evidence sufficiency perspectives.
- We construct a Historical Scene Library that organizes historical recommendation experiences into reusable preference-oriented evidence and develop a preference-compatible retrieval strategy beyond dialogue-level retrieval.
- Extensive experiments on ReDial and INSPIRED show that SERA achieves the best results on all seven ranking metrics on ReDial and on four of seven metrics on INSPIRED.

2 Related Work

Existing CRS studies have incorporated various forms of auxiliary evidence to enhance preference modeling and recommendation. Knowledge-enhanced methods leverage structured knowledge graphs to improve entity-aware reasoning and item representation, such as KBRD (Chen et al.,

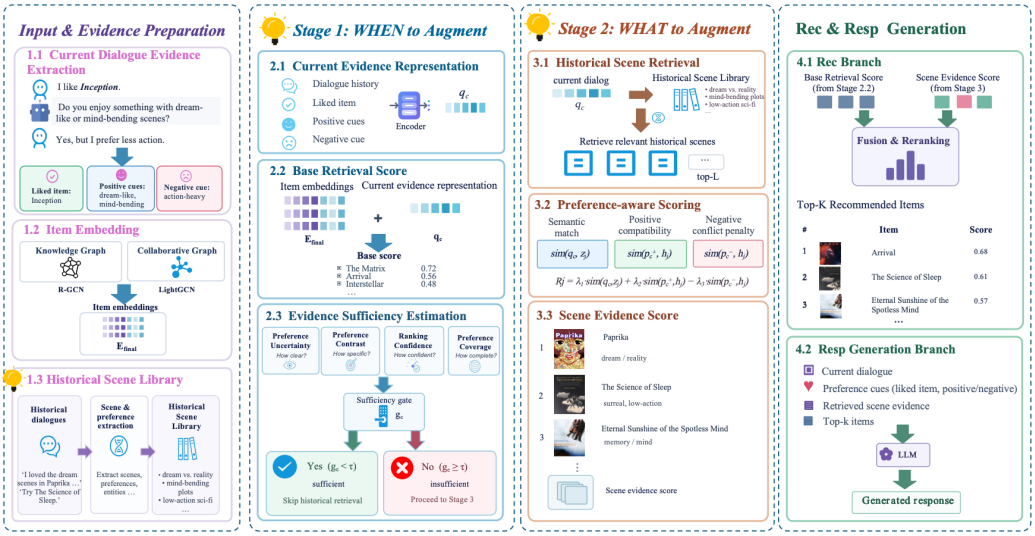


Figure 2: Overall framework of SERA.

2019), KGSF (Zhou et al., 2020), and subsequent knowledge-aware CRS models (Wang et al., 2022; Li et al., 2025). Multimodal CRS research such as MSCRS (Wei et al., 2025), as well as dataset and benchmark efforts such as SURE and MUSE (Long et al., 2023; Wang et al., 2025). Retrieval-based methods utilize historical dialogues or recommendation experiences as additional evidence sources, such as Dynamic Open-book Prompt Learning (Ma et al., 2023), DCRS (Dao et al., 2024), and other retrieval-augmented CRS approaches (Gupta et al., 2023; Zhu et al., 2025). Recent retrieval-augmented language models likewise motivate conditional evidence use: Self-RAG retrieves on demand, RECOMP suppresses unhelpful retrievals through selective compression, and SmartRAG jointly learns when to retrieve and how to use retrieved evidence (Asai et al., 2024; Xu et al., 2024; Gao et al., 2025). LongMemEval highlights the importance of fine-grained indexing and query-aware retrieval for sustained conversational memory (Wu et al., 2025), while ProPerSim combines retrieval with feedback-driven preference adaptation for proactive personalized assistants (Kim et al., 2026). However, existing methods mainly focus on acquiring or integrating auxiliary evidence while assuming that additional evidence is always beneficial. In contrast, SERA formulates evidence augmentation as a conditional decision process, learning *when* augmentation is required through evidence sufficiency assessment and *what* evidence should be introduced through preference-compatible retrieval.

3 SERA Framework

SERA is designed as an auxiliary evidence augmentation framework built upon a backbone CRS architecture (Wei et al., 2025). Specifically, rather than continuously incorporating additional evidence, SERA adaptively determines *when* augmentation is necessary and *what* evidence should be introduced by assessing evidence sufficiency and retrieving preference-compatible historical evidence (as illustrated in Figure 2).

3.1 Preliminary

This section introduces only the basic concepts and notations used throughout the paper, while detailed representation learning procedures are provided in Appendix A.1.

For each recommendation instance c, let $C_c = \{(u_1, s_1), \ldots, (u_T, s_T)\}$ denote the observed dialogue history, where u_t and s_t represent the user utterance and system response at turn t. SERA obtains item, dialogue, and preference representations from the current conversational context for subsequent evidence augmentation.

The candidate item representations are denoted as $E_{\text{final}} = \{e_1, e_2, \cdots, e_i, \cdots, e_N\}$, where $e_i \in \mathbb{R}^d$ is the embedding of item i, N is the number of candidate items, and d is the embedding

dimension. In addition, item attributes are organized into a shared attribute space $H^I \in \{0,1\}^{N\times D}$, where D denotes the number of predefined attributes and $h_i^I \in \{0,1\}^D$ represents the attribute profile of item i. This attribute space supports structured preference modeling and preference compatibility estimation. The dialogue representation extracted from C_c is denoted as $q_c \in \mathbb{R}^d$, which summarizes the conversational evidence available for recommendation.

Given the dialogue and item representations, the base recommendation score of candidate item i is computed as:

$$s_{c,i}^{\text{base}} = q_c^\top e_i, \qquad i \in \{1,\dots,N\}. \tag{1}$$

Then, the corresponding score vector over all candidate items is denoted as

$$s_c^{\text{base}} = [s_{c,1}^{\text{base}},\dots,s_{c,N}^{\text{base}}] \in \mathbb{R}^N. \tag{2}$$

Based on the dialogue representation and item attribute space, SERA further obtains structured preference profiles $\{p_c^+, p_c^-\}$, where $p_c^\nu \in [0,1]^D$ and $\nu \in \{+,-\}$. Here, p_c^+ and p_c^- represent positive and negative attribute-level preference distributions, respectively. The preference compatibility score between candidate item i and inferred user requirements is defined as:

$$s_{c,i}^{\text{pref}} = \frac{(p_c^+)^\top h_i^I}{\|p_c^+\|_1 + \epsilon} - \lambda_n \frac{(p_c^-)^\top h_i^I}{\|p_c^-\|_1 + \epsilon}, \tag{3}$$

where ϵ is a small constant for numerical stability and λ_n controls the penalty strength of negative preference conflicts. A higher $s_{c,i}^{\text{pref}}$ indicates stronger compatibility between item i and inferred user preferences. These representations provide the basic conversational and preference evidence for subsequent evidence sufficiency assessment and preference-compatible historical scene retrieval.

3.2 Learning When to Augment: Sufficiency-aware Evidence Assessment

SERA determines whether additional evidence is required by assessing current evidence sufficiency, which is characterized by four diagnostic signals: *preference uncertainty*, *preference contrast*, *ranking confidence*, and *preference coverage*. The structured preference profiles used by these signals are detailed in Appendix A.2.

(i) *Preference uncertainty* measures the decisiveness of predicted preference profiles. Given the positive and negative preference distributions $\{p_c^+, p_c^-\}$, we compute the average binary entropy over all attribute-level preference probabilities:

$$U_c^{\text{pref}} = \frac{1}{2D} \sum_{\nu\in\{+,-\}} \sum_{d=1}^{D} H_b(p_{c,d}^\nu), \tag{4}$$

where $H_b(\cdot)$ denotes the binary entropy function.

(ii) *Preference contrast* measures the discriminability of inferred preferences across semantic attribute groups. To compute group-wise preference contrast, we divide the attributes into G semantic groups. For each polarity ν and group g, $\Delta_{c,g}^\nu$ measures preference concentration, computed by top-two score margins for mutually exclusive attributes and normalized inverse entropy for multi-label attributes:

$$\Delta_c^{\text{pref}} = \frac{1}{2G} \sum_{\nu\in\{+,-\}} \sum_{g=1}^{G} \Delta_{c,g}^\nu. \tag{5}$$

(iii) *Ranking confidence* measures the reliability of the base recommendation ranking. Given the base recommendation score vector s_c^{base}, we first compute the ranking probability distribution:

$$\pi_c^{\text{base}} = \text{softmax}(s_c^{\text{base}}), \tag{6}$$

and measure the confidence margin between the top-ranked and second-ranked candidates:

$$C_c^{\text{base}} = \pi_{c(1)}^{\text{base}} - \pi_{c(2)}^{\text{base}}. \tag{7}$$

Here, $\pi_{c(1)}^{\text{base}}$ and $\pi_{c(2)}^{\text{base}}$ denote the largest and second-largest ranking probabilities, respectively.

(iv) ***Preference coverage*** measures whether the top-ranked candidates are sufficiently supported by the inferred preference profiles. Specifically, we select the top-K_b candidates according to the base recommendation scores:

$$\mathcal{B}_c^{\text{base}} = \text{Top}_{K_b}(s_c^{\text{base}}), \tag{8}$$

where K_b denotes the number of candidates considered for coverage estimation. The preference coverage is then computed as:

$$C_c^{\text{cov}} = \frac{1}{|\mathcal{B}_c^{\text{base}}|} \sum_{i \in \mathcal{B}_c^{\text{base}}} \sigma\left(\frac{s_{c,i}^{\text{pref}} - \tau_p}{\kappa}\right), \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function, τ_p is the preference compatibility threshold, and κ controls the smoothness of the transformation.

Together, these four signals characterize the sufficiency of current conversational evidence, where U_c^{pref} serves as an uncertainty signal and the remaining signals measure evidence completeness from complementary perspectives. The augmentation probability is estimated as:

$$g_c = \sigma\left(\text{MLP}_g([U_c^{\text{pref}}, 1 - \Delta_c^{\text{pref}}, 1 - C_c^{\text{base}}, 1 - C_c^{\text{cov}}])\right). \tag{10}$$

where $g_c \in [0, 1]$ denotes the probability that additional evidence is required. The effective gate value is obtained by thresholding:

$$\widetilde{g}_c = g_c \cdot \mathbb{I}[g_c \geq \tau_g], \tag{11}$$

where τ_g is the activation threshold. When augmentation is unnecessary, $\widetilde{g}_c = 0$; otherwise, retrieved historical evidence is weighted by g_c.

3.3 Learning What to Augment: Preference-compatible Retrieval

When current conversational evidence is insufficient, SERA retrieves preference-compatible historical scenes that provide complementary evidence for unresolved user preferences.

Historical Scene Library Construction and Representation. To store reusable preference-oriented recommendation experiences, SERA constructs a Historical Scene Library from historical recommendation dialogues. A historical scene is defined as a locally coherent recommendation episode in which preference-bearing utterances and subsequent recommendation feedback correspond to the same local preference context. For each valid episode, we retain the dialogue segment x_j and its explicitly positively received item set $\mathcal{I}_j^+$. Only episodes containing explicit preference signals and sufficient positive recommendation outcomes are retained. The resulting library is $\mathcal{M}^S = \{\mathcal{S}_j\}_{j=1}^M$, where $\mathcal{S}_j = (x_j, \mathcal{I}_j^+)$. Detailed extraction criteria and leakage controls are provided in Appendix B.

For each scene $\mathcal{S}_j$, the semantic representation is

$$z_j^S = \text{Enc}_s(x_j) \in \mathbb{R}^{d_s}, \tag{12}$$

where $\text{Enc}_s(\cdot)$ is inherited from the CRS backbone. Its structured preference representation is obtained by mapping the positively received items into the shared item attribute space:

$$h_j^S = \mathbb{I}\left[\frac{1}{|\mathcal{I}_j^+|} \sum_{i \in \mathcal{I}_j^+} h_i^I \geq \tau_{\text{tag}}\right] \in \{0, 1\}^D, \tag{13}$$

where τ_{tag} denotes the minimum fraction of positively received items required to assign an attribute to the scene.

Additionally, to preserve the recommendation outcomes associated with each scene, SERA further constructs a scene–item association matrix

$$A_{i,j}^S = \mathbb{I}[i \in \mathcal{I}_j^+], \tag{14}$$

where $A_{i,j}^S = 1$ indicates that item i is positively associated with scene $\mathcal{S}_j$. Thus, z_j^S captures the semantic context of the historical experience, h_j^S summarizes its consistently supported preference attributes, and A^S preserves the scene–item associations for subsequent item-level evidence integration.

Preference-compatible Scene Retrieval. To improve retrieval efficiency, SERA first performs candidate scene selection from the Historical Scene Library. Specifically, explicitly liked items in $\mathcal{I}_c^+$ are used as behavioral anchors to retrieve associated historical scenes as the candidate set $\mathcal{J}_c$. When no anchor or matching scene is available, retrieval falls back to the complete Historical Scene Library.

For each candidate scene $j \in \mathcal{J}_c$, SERA computes a retrieval score by jointly considering semantic relevance, positive preference agreement, and negative preference conflicts:

$$s_{c,j}^{\mathrm{ret}} = \gamma_s \cos_\epsilon \left(W_{\mathrm{ret}} q_c, z_j^S\right) + \gamma_p \cos_\epsilon \left(p_c^+, h_j^S\right) - \gamma_n \cos_\epsilon \left(p_c^-, h_j^S\right). \tag{15}$$

Here, $W_{\mathrm{ret}} \in \mathbb{R}^{d_s \times d}$ projects the dialogue representation into the semantic scene space; γ_s, γ_p, and γ_n control the contributions of semantic relevance, positive preference agreement, and negative preference conflicts, respectively; and $\cos_\epsilon(\cdot,\cdot)$ denotes numerically stabilized cosine similarity.

SERA selects the top-L historical scenes according to their retrieval scores:

$$\mathcal{T}_c = \mathrm{TopK}_L \left(\{(j, s_{c,j}^{\mathrm{ret}}) \mid j \in \mathcal{J}_c\}\right). \tag{16}$$

and computes their normalized retrieval weights:

$$w_{c,j} = \frac{\exp(s_{c,j}^{\mathrm{ret}}/\tau_r)}{\sum_{k\in\mathcal{T}_c} \exp(s_{c,k}^{\mathrm{ret}}/\tau_r)}, \qquad j \in \mathcal{T}_c, \tag{17}$$

where τ_r denotes the retrieval temperature. The training-time supervision for selecting preference-compatible scenes is specified in Appendix C.

Scene Evidence Integration. The retrieved scenes are projected into task-specific evidence for recommendation and response generation.

For recommendation, their positive item associations are aggregated into an item-level historical support score:

$$s_{c,i}^{\mathrm{evi}} = \sum_{j\in\mathcal{T}_c} w_{c,j} \widetilde{A}_{i,j}^S, \tag{18}$$

where $\widetilde{A}_{i,j}^S$ is the normalized scene–item association matrix derived from $A_{i,j}^S$.

For response generation, the semantic representations of retrieved scenes are aggregated as:

$$e_c^{\mathrm{evi}} = \sum_{j\in\mathcal{T}_c} w_{c,j} z_j^S, \tag{19}$$

providing additional semantic context for response generation.

3.4 Evidence-aware Recommendation and Response Generation

After evidence sufficiency assessment and preference-compatible scene retrieval, SERA integrates the retrieved historical evidence into both recommendation ranking and response generation.

Recommendation. For candidate item i, the final recommendation score combines the base conversational preference, structured preference compatibility, and retrieved scene evidence:

$$s_{c,i}^{\mathrm{final}} = s_{c,i}^{\mathrm{base}} + \alpha s_{c,i}^{\mathrm{pref}} + \beta \widetilde{g}_c s_{c,i}^{\mathrm{evi}}, \tag{20}$$

where α and β control the contributions of preference compatibility and retrieved evidence, respectively. The gate value $\widetilde{g}_c$ adaptively adjusts the contribution of historical evidence according to the estimated evidence insufficiency. The final recommendation list is obtained by selecting the top-K items according to $s_{c,i}^{\mathrm{final}}$.

Response Generation. For response generation, SERA conditionally incorporates historical evidence according to the evidence-sufficiency gate. Historical scene retrieval is performed only when $g_c \geq \tau_g$. The representation used for response generation is defined as

$$\tilde{q}_c = \begin{cases} [q_c; e_c^{\mathrm{evi}}], & g_c \geq \tau_g, \\ q_c, & g_c < \tau_g, \end{cases} \tag{21}$$

where $[\cdot;\cdot]$ denotes vector concatenation. When $g_c < \tau_g$, historical scene retrieval is skipped and the original dialogue representation q_c is directly used for generation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate SERA on two widely used conversational recommendation datasets: ReDial (Li et al., 2018), containing 10,006 conversations and 182,150 utterances, and INSPIRED (Hayati et al., 2020), containing 1,001 conversations and 35,811 utterances. Following MSCRS (Wei et al., 2025), we use the standard data splits. All preference annotations and historical evidence used by SERA are constructed exclusively from the training split to prevent information leakage.

Baselines. We compare SERA with eight representative conversational recommender systems, including KGSF (Zhou et al., 2020), UniCRS (Wang et al., 2022), TREA (Li et al., 2023), DCRS (Dao et al., 2024), DisenCRS (An et al., 2025), PCRS-TKA (Ren et al., 2026), CUPID (Kumar et al., 2026), and MSCRS (Wei et al., 2025). These baselines cover knowledge-enhanced, retrieval-based, multimodal, and language-model-based conversational recommendation.

Evaluation Metrics. For item recommendation, we report Recall@1/10/50, NDCG@10/50, and MRR@10/50, following the evaluation protocol used by MSCRS (Wei et al., 2025). For response generation, we report BLEU-2/3 (Papineni et al., 2002), ROUGE-2/L (Lin, 2004), and Distinct-2/3/4 (Li et al., 2016). Here, "L" in ROUGE-L refers to the longest common subsequence and is unrelated to the retrieval hyperparameter L in SERA.

Implementation Details. SERA is implemented in PyTorch while retaining the backbone CRS architecture of Wei et al. (2025). We optimize the model with AdamW (Loshchilov & Hutter, 2019) using a linear warmup schedule and a learning rate of 10^{-4}. For evidence augmentation, we set the gate threshold to $\tau_g = 0.5$ and use dataset-specific numbers of retrieved historical scenes L. All experiments are conducted on NVIDIA RTX 5090 GPUs. Full training configurations and method-specific hyperparameters are provided in Appendix D.1.

4.2 RQ1: HOW DOES SERA COMPARE WITH EXISTING CRS METHODS?

Motivation. We evaluate whether SERA improves recommendation performance compared with representative CRS methods. The main evaluation focuses on item recommendation, with additional response-generation results provided in Appendix D.2.

Table 1: Item recommendation results on ReDial and INSPIRED.

Model	ReDial							INSPIRED						
	Recall			NDCG		MRR		Recall			NDCG		MRR	
	@1	@10	@50	@10	@50	@10	@50	@1	@10	@50	@10	@50	@10	@50
KGSF	0.035	0.175	0.367	0.094	0.137	0.070	0.079	0.051	0.132	0.239	0.092	0.114	0.079	0.083
UniCRS	0.065	0.241	0.423	0.143	0.183	0.113	0.125	0.085	0.230	0.398	0.149	0.187	0.125	0.133
TREA	0.045	0.204	0.403	0.114	0.158	0.087	0.096	0.047	0.146	0.347	0.095	0.132	0.076	0.087
DCRS	0.076	0.253	0.439	0.154	0.195	0.123	0.132	0.093	0.226	0.414	0.153	0.192	0.130	0.137
DisenCRS	0.081	0.268	0.451	0.162	0.210	0.130	0.138	0.094	0.252	0.423	0.165	0.200	0.139	0.146
PCRS-TKA	0.072	0.221	0.432	0.123	0.170	0.093	0.103	0.102	**0.278**	0.438	**0.187**	**0.222**	0.112	0.128
CUPID	-	0.194	0.422	0.117	0.149	-	-	-	0.191	0.374	0.120	0.157	-	-
MSCRS	0.081	0.264	0.451	0.161	0.201	0.128	0.136	0.096	0.257	0.425	0.168	0.202	0.140	0.148
Ours	**0.085**	**0.273**	**0.460**	**0.169**	**0.210**	**0.137**	**0.146**	**0.104**	0.271	**0.445**	0.180	0.219	**0.153**	**0.161**
Improve	**4.94%**	**1.87%**	**2.00%**	**4.32%**	**0.00%**	**5.38%**	**5.80%**	**1.96%**	**-2.52%**	**1.60%**	**-3.74%**	**-1.35%**	**9.29%**	**8.78%**

Results. As shown in Table 1, SERA achieves the best performance across all seven metrics on ReDial. On INSPIRED, SERA ranks first on Recall@1/50 and MRR@10/50, and second on Recall@10 and NDCG@10/50, demonstrating strong performance across different ranking cutoffs. More importantly, compared with the backbone-aligned MSCRS, SERA improves all 14 ranking metrics across the two datasets, with relative gains of 4.94%/8.33% in Recall@1 and 7.35%/8.78% in MRR@50 on ReDial/INSPIRED, respectively. Against the strongest competing baseline for each metric, SERA further achieves gains of up to 5.80% on ReDial and 9.29% on INSPIRED. These results indicate that the proposed conditional evidence augmentation consistently strengthens the backbone recommender, particularly for top-ranked recommendation and reciprocal-rank quality.

4.3 RQ2: HOW DOES EACH COMPONENT CONTRIBUTE TO SERA?

Motivation. We conduct three groups of ablation studies to examine the two core design questions of SERA. First, we perform overall ablations of the WHEN and WHAT components. For

Table 2: Ablation studies on ReDial and INSPIRED.

Ablation Group	Variant	ReDial Recall@1	Recall@10	Recall@50	NDCG@10	NDCG@50	MRR@10	MRR@50
	Full SERA	**0.085**	**0.273**	**0.460**	**0.169**	**0.210**	**0.137**	**0.146**
Overall	w/o WHEN (always-on)	0.083	0.269	0.456	0.165	0.206	0.133	0.141
	w/o WHEN (always-off)	0.081	0.264	0.451	0.161	0.201	0.128	0.136
	w/o WHAT	0.082	0.267	0.452	0.163	0.204	0.131	0.140
When: Evidence Sufficiency	w/o Preference Uncertainty	0.085	0.266	0.452	0.165	0.206	0.134	0.143
	w/o Preference Contrast	0.085	0.267	0.456	0.166	0.207	0.135	0.144
	w/o Base Ranking Confidence	0.083	0.270	0.455	0.166	0.207	0.134	0.143
	w/o Preference Coverage	0.083	0.264	0.452	0.163	0.205	0.132	0.141
What: Evidence Retrieval	Semantic-only Retrieval	0.084	0.267	0.453	0.164	0.205	0.133	0.142

Ablation Group	Variant	INSPIRED Recall@1	Recall@10	Recall@50	NDCG@10	NDCG@50	MRR@10	MRR@50
	Full SERA	**0.104**	**0.271**	**0.445**	**0.180**	**0.219**	**0.153**	**0.161**
Overall	w/o WHEN (always-on)	0.098	0.270	0.441	0.178	0.215	0.149	0.157
	w/o WHEN (always-off)	0.096	0.257	0.425	0.173	0.210	0.148	0.156
	w/o WHAT	0.100	0.263	0.437	0.176	0.211	0.149	0.155
When: Evidence Sufficiency	w/o Preference Uncertainty	0.102	0.266	0.445	0.176	0.215	0.148	0.156
	w/o Preference Contrast	0.104	0.262	0.441	0.175	0.215	0.148	0.156
	w/o Base Ranking Confidence	0.102	0.270	0.442	0.179	0.216	0.151	0.158
	w/o Preference Coverage	0.094	0.270	0.441	0.174	0.211	0.144	0.152
What: Evidence Retrieval	Semantic-only Retrieval	0.102	0.270	0.441	0.177	0.215	0.149	0.156

WHEN, *always-on* removes evidence-sufficiency assessment and retrieves historical scene evidence for every query, whereas *always-off* disables historical retrieval for all queries and reduces SERA to the backbone-aligned MSCRS. We further remove the WHAT component as a whole to evaluate the overall contribution of historical evidence selection. Second, we ablate the four diagnostic signals used by WHEN to examine their contributions to evidence-sufficiency estimation. Third, while keeping WHEN unchanged, we replace preference-compatible retrieval with semantic-only retrieval for gate-activated queries, thereby removing preference information from the WHAT module. Table 2 summarizes the results on both datasets.

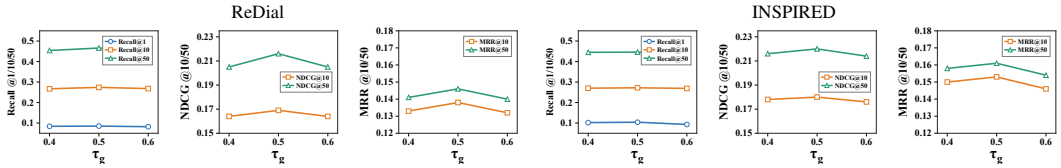
Overall Ablation. The ablations show that both WHEN and WHAT are important for effective evidence augmentation. Removing WHEN under either always-on or always-off retrieval consistently degrades performance, indicating that historical evidence is useful but should be introduced selectively. Always-off suffers the largest degradation, showing the value of complementary historical evidence, whereas the gap between always-on and full SERA suggests that indiscriminate augmentation may introduce redundant or mismatched information. Removing WHAT also leads to consistent performance drops, demonstrating that determining *when* to augment is not sufficient without identifying *what* historical evidence is useful. Together, these results highlight the complementary roles of WHEN and WHAT in controlling the necessity and quality of evidence augmentation.

WHEN Ablation. Removing any sufficiency signal generally degrades ranking performance, confirming that the four signals provide complementary cues for augmentation decisions. Preference coverage has the strongest effect, particularly on INSPIRED, suggesting that whether the current candidates are sufficiently supported by inferred preferences is a critical indicator of evidence insufficiency. Interestingly, removing preference contrast has little effect on Recall@1 but consistently weakens NDCG and MRR, indicating that this signal contributes more to ranking refinement than to identifying the single top item. Overall, evidence sufficiency is better characterized jointly from preference uncertainty, discriminability, ranking confidence, and coverage.

WHAT Ablation. Replacing preference-compatible retrieval with semantic-only retrieval consistently weakens performance, with MRR@50 dropping from 0.146 to 0.142 on ReDial and from 0.161 to 0.156 on INSPIRED. This relatively consistent degradation suggests that semantic relevance alone can retrieve related historical scenes, but cannot reliably distinguish whether their preference patterns are compatible with the current query. Thus, once augmentation is activated, preference constraints are important for selecting evidence that is not only semantically related but also behaviorally useful.

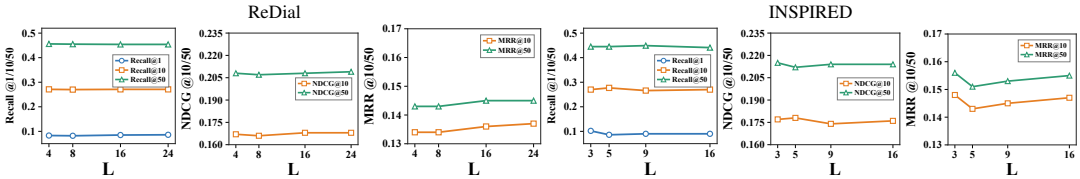
4.4 RQ3: How Sensitive Is SERA to Its Augmentation Hyperparameters?

Motivation. We study the sensitivity of SERA to the number of retrieved historical scenes L and the gate activation threshold τ_g, which control the amount of auxiliary evidence and the activation of augmentation, respectively.



(a) Recall (b) NDCG (c) MRR (a) Recall (b) NDCG (c) MRR

Figure 3: Sensitivity to the gate threshold τ_g on ReDial and INSPIRED.



(a) Recall (b) NDCG (c) MRR (a) Recall (b) NDCG (c) MRR

Figure 4: Sensitivity to the number of retrieved historical scenes L on ReDial and INSPIRED.

Results. For L, SERA is relatively stable on ReDial, while INSPIRED favors a smaller retrieval set; excessive retrieval may introduce redundant or weakly relevant evidence and weaken early-ranking performance. For τ_g, both datasets perform best at an intermediate value, as overly frequent augmentation may introduce unnecessary evidence, whereas an overly conservative gate can miss evidence-insufficient conversations. Accordingly, we use dataset-specific L and set $\tau_g = 0.5$ in the main experiments.

4.5 RQ4: What Is the Efficiency Cost of Evidence Augmentation?

Motivation. We evaluate the computational overhead introduced by SERA using MSCRS as the efficiency reference. MSCRS shares the same backbone architecture as SERA but does not include the additional evidence-sufficiency assessment, historical-scene retrieval, and evidence integration introduced by our framework. This controlled comparison therefore isolates the computational cost associated with SERA's evidence augmentation mechanism.

Table 3: Efficiency comparison.

Dataset	Model	Latency (ms/query)	P95 (ms)	Peak Memory (GB)
INSPIRED	MSCRS	18.697 ± 2.752	25.257	1.409
	SERA	22.923 ± 3.669	30.842	1.422
ReDial	MSCRS	68.477 ± 1.652	70.537	2.329
	SERA	76.764 ± 2.163	80.813	2.511

Results. As shown in Table 3, SERA incurs additional inference cost over MSCRS, as expected, due to the added evidence assessment, retrieval, and integration processes. On INSPIRED, mean latency is 18.697 vs. 22.923 ms/query and P95 latency is 25.257 vs. 30.842 ms for MSCRS and SERA, respectively, while peak GPU memory is 1.409 vs. 1.422 GB. On ReDial, the corresponding values are 68.477 vs. 76.764 ms/query for mean latency, 70.537 vs. 80.813 ms for P95 latency, and 2.329 vs. 2.511 GB for peak GPU memory. Despite the additional retrieval pipeline, the increase in latency and memory remains moderate, indicating that SERA achieves improved recommendation performance with a manageable computational overhead.

5 Conclusion

SERA-CRS treats historical dialogue experience as conditional evidence rather than a resource that must be used for every conversation. Its structured preference profiles make positive support and negative conflict explicit, while the sufficiency gate connects uncertainty in those profiles to the state of the base recommender. When augmentation is activated, the model retrieves semantically and structurally compatible scenes and shares them across ranking and response generation. This design keeps the recommendation and generation paths aligned around the same evidence. Across ReDial and INSPIRED, the current results show consistent point-estimate gains over the MSCRS backbone, while the component analysis indicates complementary contributions from structured preference evidence, conditional retrieval, and shared generation conditioning.

AI USE STATEMENT

Generative AI tools were used solely for language polishing to improve the clarity, grammar, and readability of the manuscript. They were not used for study design, methodology development, data analysis, result generation, or scientific interpretation. All AI-assisted language edits were reviewed and approved by the authors, who take full responsibility for the final content of the manuscript.

REPRODUCIBILITY STATEMENT

To support reproducibility, we publicly release the source code of our method on GitHub, including the complete implementation, training and evaluation pipelines. The dataset used in this work is also provided with an accessible link.

REFERENCES

Guojia An, Jie Zou, Jiwei Wei, Chaoning Zhang, Fuming Sun, and Yang Yang. Beyond whole dialogue modeling: Contextual disentanglement for conversational recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025. doi: 10.1145/3726302.3729903.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, pp. 9112–9141, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25f7be9694d7b32d5cc670927b8091e1-Paper-Conference.pdf.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. Towards knowledge-based recommender dialog system. In *Proceedings of EMNLP-IJCNLP*, pp. 1803–1813, 2019. doi: 10.18653/v1/D19-1189.

Huy Dao, Yang Deng, Dung D. Le, and Lizi Liao. Broadening the view: Demonstration-augmented prompt learning for conversational recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. doi: 10.1145/3626772.3657755.

Jingsheng Gao, Linxu Li, Ke Ji, Weiyuan Li, Yixin Lian, Yuzhuo Fu, and Bin Dai. SmartRAG: Jointly learn RAG-related tasks from the environment feedback. In *International Conference on Learning Representations*, pp. 53117–53140, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/83ccb398f3ce9c4d137011f36a03c7d4-Paper-Conference.pdf.

Raghav Gupta, Renat Aksitov, Samrat Phatale, Simral Chaudhary, Harrison Lee, and Abhinav Rastogi. Conversational recommendation as retrieval: A simple, strong baseline. In *Proceedings of the 5th Workshop on NLP for Conversational AI*, pp. 155–160, 2023. doi: 10.18653/v1/2023.nlp4convai-1.13.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. INSPIRED: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8142–8152, 2020. doi: 10.18653/v1/2020.emnlp-main.654.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648, 2020. doi: 10.1145/3397271.3401063.

Jiho Kim, Junseong Choi, Woosog Chay, Daeun Kyung, Yeonsu Kwon, Yohan Jo, and Edward Choi. ProPerSim: Developing proactive and personalized AI assistants through user-assistant simulation. In *International Conference on Learning Representations*, pp. 110843–110873, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b3bdc41d0b2387eabbb719956a38f295-Paper-Conference.pdf.

Heejin Kook, Junyoung Kim, Seongmin Park, and Jongwuk Lee. Empowering retrieval-based conversational recommendation with contrasting user preferences. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 7692–7707, 2025.

Bagesh Kumar, Subham Raj, Mayank Anand, Vivek Kumar Verma, Pritesh Tripathi, and Anand Jha. Enhancing conversational recommendations through contrastive user preference modeling. *Discover Artificial Intelligence*, 6:543, 2026. doi: 10.1007/s44163-026-00912-1.

Chuang Li, Yang Deng, Hengchang Hu, Min-Yen Kan, and Haizhou Li. ChatCRS: Incorporating external knowledge and goal guidance for llm-based conversational recommender systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 295–312, 2025.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, 2016. doi: 10.18653/v1/N16-1014.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1812.07617.

Wendi Li, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Ye Yuan, Wenfeng Xie, and Dangyang Chen. TREA: Tree-structure reasoning schema for conversational recommendation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2970–2982, 2023. doi: 10.18653/v1/2023.acl-long.167.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004. URL https://aclanthology.org/W04-1013/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Yuxing Long, Binyuan Hui, Caixia Yuan, Fei Huang, Yongbin Li, and Xiaojie Wang. Multimodal recommendation dialog with subjective preference: A new challenge and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3515–3533, 2023. doi: 10.18653/v1/2023.findings-acl.217.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Xuan Ma, Tieyun Qian, and Ke Sun. Dynamic open-book prompt for conversational recommender system. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9839–9849, 2023. doi: 10.18653/v1/2023.findings-emnlp.658.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135.

Mathieu Ravaut, Hao Zhang, Lu Xu, Aixin Sun, and Yong Liu. Parameter-efficient conversational recommender system as a language processing task. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 152–165, 2024. doi: 10.18653/v1/2024.eacl-long.9.

Yongwen Ren, Chao Wang, Peng Du, Chuan Qin, Dazhong Shen, and Hui Xiong. Enhancing conversational recommender systems with tree-structured knowledge and pretrained language models. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, 2026.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018*, pp. 593–607. Springer, 2018. doi: 10.1007/978-3-319-93417-4_38.

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1929–1937, 2022. doi: 10.1145/3534678.3539382.

Zihan Wang, Xiaocui Yang, YongKang Liu, Shi Feng, Daling Wang, and Yifei Zhang. MUSE: A multimodal conversational recommendation dataset with scenario-grounded user profiles. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1027–1053, 2025. doi: 10.18653/v1/2025.findings-acl.58.

Yibiao Wei, Jie Zou, Weikang Guo, Guoqing Wang, Xing Xu, and Yang Yang. MSCRS: Multimodal semantic graph prompt learning framework for conversational recommender systems. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 42–52, 2025. doi: 10.1145/3726302.3730040.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, pp. 86809–86836, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d813d324dbf0598bbdc9c8e79740ed01-Paper-Conference.pdf.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *International Conference on Learning Representations*, pp. 43478–43502, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/bda88ed2892f5e61c9a9bf215c566913-Paper-Conference.pdf.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DialoGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 270–278, 2020. doi: 10.18653/v1/2020.acl-demos.30.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1006–1014, 2020. doi: 10.1145/3394486.3403143.

Yaochen Zhu, Harald Steck, Dawen Liang, Yinhan He, Nathan Kallus, and Jundong Li. LLM-based conversational recommendation agents with collaborative verbalized experience. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2207–2220, 2025. doi: 10.18653/v1/2025.findings-emnlp.119.

A PRELIMINARY

A.1 BASE CRS REPRESENTATION

For each recommendation instance c, let $C_c = \{(u_1, s_1), \ldots, (u_T, s_T)\}$ denote the observed dialogue history, where u_t and s_t represent the user utterance and system response at turn t ($1 \leq t \leq T$). Following the backbone CRS architecture (Wei et al., 2025), we obtain item and dialogue representations for subsequent evidence augmentation.

For candidate items, R-GCN (Schlichtkrull et al., 2018) and LightGCN (He et al., 2020) are employed to capture relational and collaborative signals, respectively. Their fusion produces the item embedding table:

$$E_{\text{final}} = \{e_1, e_2, \cdots, e_i, \cdots, e_N\}, \tag{22}$$

where $e_i \in \mathbb{R}^d$ denotes the embedding of item i, N is the number of candidate items, and d is the embedding dimension.

In addition to item embeddings, we organize item attributes into a shared label space. The item-attribute matrix $H^I \in \{0,1\}^{N \times D}$ describes the association between items and attributes, where D denotes the number of predefined attributes obtained from dataset metadata (Ravaut et al., 2024), and $h_i^I \in \{0,1\}^D$ represents the attribute profile of item i. This attribute space is used for structured preference modeling and preference compatibility estimation.

For the observed dialogue C_c, dialogue tokens are encoded by RoBERTa (Liu et al., 2019) and fused with the embeddings of mentioned items from E_{final}, followed by DialoGPT (Zhang et al., 2020). The resulting dialogue representation is formulated as:

$$q_c = \text{DialoGPT}\left(\text{Fuse}\left(\text{RoBERTa}(C_c), E_c^{\text{ment}}\right)\right) \in \mathbb{R}^d, \tag{23}$$

where E_c^{ment} denotes the embeddings of items mentioned in C_c, retrieved from E_{final}.

Given the dialogue and item representations, the base recommendation score of candidate item i is computed as:

$$s_{c,i}^{\text{base}} = q_c^\top e_i, \qquad i \in \{1, \ldots, N\}. \tag{24}$$

The corresponding score vector over all candidate items is:

$$s_c^{\text{base}} = [s_{c,1}^{\text{base}}, \ldots, s_{c,N}^{\text{base}}] \in \mathbb{R}^N. \tag{25}$$

A.2 STRUCTURED PREFERENCE MODELING

Based on the dialogue representation q_c and the shared attribute space H^I, SERA models structured user preferences from the current conversational evidence. Specifically, positive and negative preference signals are explicitly separated to capture both desired and undesired item characteristics.

Inspired by CORAL's separate modeling of like and dislike preferences (Kook et al., 2025), we model positive and negative feedback separately. Specifically, we define $\mathcal{I}_c^+$ and $\mathcal{I}_c^-$ as the positively and negatively interacted item sets extracted from explicit user feedback in dialogue C_c, respectively. For each preference polarity $\nu \in \{+, -\}$, the corresponding preference representation is obtained by aggregating feedback item embeddings conditioned on the current dialogue context:

$$a_{c,i}^\nu = \frac{\exp(q_c^\top e_i)}{\sum_{k \in \mathcal{I}_c^\nu} \exp(q_c^\top e_k)}, \qquad r_c^\nu = \sum_{i \in \mathcal{I}_c^\nu} a_{c,i}^\nu e_i, \tag{26}$$

where $a_{c,i}^\nu$ denotes the context-dependent importance weight of item i. When $\mathcal{I}_c^\nu$ is empty, r_c^ν is initialized as a zero vector.

The positive and negative preference representations are combined with the dialogue representation to predict attribute-level preference distributions:

$$v_c = [q_c; r_c^+; r_c^-], \qquad p_c^\nu = \sigma(\text{MLP}_\nu(v_c)) \in [0,1]^D. \tag{27}$$

Here, p_c^+ and p_c^- denote the predicted positive and negative preference distributions over the shared attribute space, respectively.

Based on the structured preference profiles, we compute the compatibility between candidate items and user requirements:

$$s_{c,i}^{\text{pref}} = \frac{(p_c^+)^\top h_i^I}{\|p_c^+\|_1 + \epsilon} - \lambda_n \frac{(p_c^-)^\top h_i^I}{\|p_c^-\|_1 + \epsilon}. \tag{28}$$

where ϵ is a small constant for numerical stability and λ_n is a hyperparameter controlling the penalty strength of negative preference conflicts. The first term measures the degree to which item i satisfies positive preference requirements, while the second term penalizes conflicts with negative preferences. A higher $s_{c,i}^{\text{pref}}$ indicates stronger preference compatibility.

The resulting structured preference profiles $\{p_c^+, p_c^-\}$ and compatibility scores $s_{c,i}^{\text{pref}}$ provide attribute-level preference evidence for subsequent evidence sufficiency assessment and preference-compatible historical evidence retrieval.

B Formal Definition and Construction of Historical Evidence

Historical Scene Library overview. The Historical Scene Library is an offline evidence memory constructed exclusively from the training dialogues. Instead of treating an entire historical dialogue as a single retrieval unit, the library decomposes each dialogue into local recommendation scenes. Each scene represents a coherent interaction episode in which user preference expressions and subsequent recommendation feedback refer to the same local preference context. This scene-level organization removes unrelated dialogue turns and preserves reusable preference–outcome patterns.

Each scene entry S_j contains a local dialogue segment x_j, its explicitly supported item set I_j^+, a semantic representation z_j, and a structured preference profile h_j^S. The library also maintains the source-dialogue identifier o_j as provenance metadata for leakage control. Therefore, the Historical Scene Library should be understood as a collection of fine-grained, preference-grounded recommendation experiences rather than a collection of complete historical conversations.

Historical evidence unit. Let a training dialogue be $\mathcal{D}^{(o)} = (v_1^{(o)}, \ldots, v_{T_o}^{(o)})$, where o is its source-dialogue index and $v_t^{(o)}$ is the utterance at turn t. A local segment $x = \mathcal{D}_{a:b}^{(o)}$ is a contiguous span of turns that forms a coherent recommendation episode, i.e., the preference-bearing utterances and the subsequent recommendation feedback refer to the same local preference context. Let $\phi(i, x) \in \{+, -, 0\}$ denote the observed feedback polarity for catalog item i in x, where 0 represents neutral, missing, or ambiguous feedback. The positively supported item set of the segment is

$$\mathcal{I}^+(x) = \{i \in \mathcal{I} : i \text{ occurs in } x \ \wedge \ \phi(i, x) = +\}. \tag{29}$$

We define a *historical evidence unit* as a tuple

$$\mathcal{S}_j = (x_j, \mathcal{I}_j^+, z_j, h_j^S), \qquad \mathcal{I}_j^+ = \mathcal{I}^+(x_j), \tag{30}$$

subject to three validity conditions: (i) x_j contains at least one explicit preference-bearing user utterance; (ii) every item in $\mathcal{I}_j^+$ has explicit positive feedback within the same local episode; and (iii) $|\mathcal{I}_j^+| \geq 2$. Items with negative, neutral, missing, or ambiguous feedback are not included in $\mathcal{I}_j^+$. The multi-item condition distinguishes reusable recommendation evidence from an isolated item mention: a valid unit records that several items were positively received under a shared preference context.

For each unit, the source-dialogue identifier o_j is retained separately as provenance metadata, not as part of $\mathcal{S}_j$. It is used only for training-time retrieval masking and does not enter semantic encoding, preference matching, recommendation ranking, or response generation.

The stored unit contains two complementary representations. Its semantic representation is

$$z_j = \text{Enc}_s(x_j) \in \mathbb{R}^{d_s}, \tag{31}$$

which preserves the natural-language preference context. Its structured preference representation is

$$h_j^S = \mathbb{I}\left[\frac{1}{|\mathcal{I}_j^+|} \sum_{i \in \mathcal{I}_j^+} h_i^I \geq \tau_{\text{tag}}\right] \in \{0, 1\}^D, \tag{32}$$

where the threshold is applied elementwise. Thus, h_j^S retains attributes that are repeatedly associated with positive outcomes in the unit rather than attributes occurring in only one item. The semantic vector z_j supports contextual matching, whereas h_j^S allows the same evidence to be checked against the current positive and negative preference profiles.

Memory construction. The Historical Scene Library is constructed offline through four steps. First, for each training dialogue, we scan its utterances in chronological order and link catalog-item mentions to the subsequent explicit user feedback. Each observed item is assigned a feedback polarity $\phi(i, x) \in \{+, -, 0\}$, where the three values represent positive, negative, and neutral, missing, or ambiguous feedback, respectively. Only preference evidence observable from the dialogue itself is used during this process.

Second, neighboring preference and feedback events that refer to the same local recommendation context are grouped into a recommendation scene. The corresponding dialogue segment x_j spans from the first preference-bearing utterance of the episode to its last associated feedback utterance. When the dialogue moves to a different preference context, a new scene is created.

Third, we collect all items receiving explicit positive feedback in x_j to form I_j^+. A scene is retained only when it contains at least one explicit preference-bearing user utterance and at least two distinct positively supported items. Items with negative, neutral, missing, or ambiguous feedback are excluded from I_j^+. This filtering distinguishes a reusable recommendation experience from an isolated item mention.

Finally, each retained scene is converted into semantic, structured, and item-level representations. The dialogue segment x_j is encoded as z_j. To construct h_j^S, we retrieve the attribute vector h_i^I of each positively supported item $i \in I_j^+$ and aggregate these vectors across the scene. An attribute is assigned to the scene when its occurrence frequency among the positively supported items is no smaller than τ_{tag}, as formally defined in Eq. (32). The scene–item relationships are recorded in A^S, from which an item-to-scene index is derived for behavioral-anchor-based candidate retrieval.

Applying this procedure to all training dialogues produces

$$\mathcal{M}^S = \left\{ S_j : D^{(o_j)} \in \mathcal{D}_{\text{train}},\ S_j \text{ is valid} \right\}_{j=1}^{M}. \tag{33}$$

The resulting library is constructed once before training and remains fixed during validation and testing.

Leakage control. The memory is constructed exclusively from training dialogues. For a training query originating from dialogue o_c, its eligible memory is restricted to

$$\mathcal{M}_c^S = \{\mathcal{S}_j \in \mathcal{M}^S : o_j \neq o_c\}, \tag{33}$$

so that the model cannot retrieve evidence extracted from the query's own source dialogue. At validation and test time, the memory remains fixed to $\mathcal{D}_{\text{train}}$. Retrieval depends only on the observed dialogue prefix and the preconstructed memory; the target item and reference response are never used to form or select evidence at inference time.

C Training Objectives and Optimization Details

Recommendation and generation objectives. For a training instance c, the recommendation and response-generation objectives are

$$\mathcal{L}_{\text{rec}} = -\log \left[\text{softmax}\left(\mathbf{s}_c^{\text{final}}\right)\right]_{i_c^*}, \qquad \mathcal{L}_{\text{gen}} = -\sum_t \log p_\theta \left(y_{c,t}^* \mid y_{c,<t}^*, P_c\right), \tag{36}$$

where i_c^* and $y_{c,t}^*$ denote the target item and the reference response token, respectively. During training, the continuous gate probability g_c is used as a differentiable relaxation of the thresholded gate when weighting historical evidence in $\mathbf{s}_c^{\text{final}}$ and the generator input P_c. At inference time, P_c is constructed conditionally according to the evidence-sufficiency gate: historical scene retrieval is activated and the retrieved evidence is incorporated when $g_c \geq \tau_g$; otherwise, retrieval is skipped and no historical evidence is introduced into response generation.

Structured preference supervision. We construct weak attribute-level targets from observed user feedback:

$$y^+_{c,d} = \mathbb{I}\left[\sum_{i \in \mathcal{I}^+_c \cup \{i^*_c\}} (h^I_i)_d > 0\right], \qquad y^-_{c,d} = \mathbb{I}\left[\sum_{i \in \mathcal{I}^-_c} (h^I_i)_d > 0\right]. \tag{34}$$

These targets provide weak multi-label supervision rather than complete annotations of the user's preferences. The corresponding objective is

$$\mathcal{L}_{\text{pref}} = \text{BCE}(p^+_c, y^+_c) + \zeta_- \,\text{BCE}(p^-_c, y^-_c), \tag{35}$$

where ζ_- balances the positive- and negative-preference terms.

Historical evidence retrieval supervision. Given the candidate scene set $\mathcal{J}_c$, we regard a historical scene as positive evidence when it shares an observed positive anchor with the current dialogue and contains the target item:

$$\mathcal{P}_c = \left\{ j \in \mathcal{J}_c \;\middle|\; \mathcal{I}^+_c \cap \mathcal{I}^+_j \neq \varnothing,\; i^*_c \in \mathcal{I}^+_j \right\}. \tag{36}$$

The retriever is optimized with the multi-positive contrastive objective

$$\mathcal{L}_{\text{evi},c} = -\log \frac{\sum_{j \in \mathcal{P}_c} \exp(R_{c,j}/\tau_r)}{\sum_{j \in \mathcal{J}_c} \exp(R_{c,j}/\tau_r)}. \tag{37}$$

The evidence loss is omitted when $\mathcal{P}_c = \varnothing$.

Evidence-sufficiency gate supervision. We supervise the gate according to whether historical evidence improves the target-item prediction. For each training instance, we construct two counterfactual score vectors corresponding to disabling and fully enabling historical evidence:

$$\mathbf{s}^{\text{off}}_c = \text{sg}\big(\mathbf{s}^{\text{base}}_c + \alpha \mathbf{s}^{\text{pref}}_c\big), \qquad \mathbf{s}^{\text{on}}_c = \text{sg}\big(\mathbf{s}^{\text{base}}_c + \alpha \mathbf{s}^{\text{pref}}_c + \beta \mathbf{s}^{\text{evi}}_c\big), \tag{41}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. The off and on scores correspond to the two counterfactual cases $\widetilde{g}_c = 0$ and $\widetilde{g}_c = 1$, respectively. We compute the corresponding target-item losses and define the gate target as

$$\ell^a_c = -\log\left[\text{softmax}(\mathbf{s}^a_c)\right]_{i^*_c}, \quad y^g_c = \mathbb{I}\left[\ell^{\text{off}}_c - \ell^{\text{on}}_c > m_g\right], \quad a \in \{\text{off}, \text{on}\}, \tag{42}$$

where m_g is the minimum improvement required to regard historical evidence as beneficial. The gate objective is

$$\mathcal{L}_{\text{gate}} = -\frac{1}{|\mathcal{B}|} \sum_{c \in \mathcal{B}} \left[y^g_c \log g_c + (1 - y^g_c) \log(1 - g_c) \right]. \tag{43}$$

We set $m_g = 0$ in our experiments. Stop-gradient is applied only when constructing the gate target; the recommendation, preference, and retrieval modules remain trainable through their respective objectives.

Optimization details. During training, retrieval is executed independently of the hard gate threshold when constructing $\mathcal{L}_{\text{evi}}$ and the counterfactual gate target. To enable end-to-end optimization, the continuous probability g_c is used as a differentiable relaxation of the thresholded gate $\widetilde{g}_c$ when optimizing the recommendation and generation objectives. At inference time, the thresholded gate defined in Eq. (11) determines whether historical evidence is retrieved and incorporated. Hard candidate filtering, top-L scene selection, and top-K recommendation are treated as non-differentiable; gradients propagate through the selected retrieval weights, score values, projection layers, and the continuous gate.

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{gen}}\mathcal{L}_{\text{gen}} + \lambda_{\text{pref}}\mathcal{L}_{\text{pref}} + \lambda_{\text{evi}}\mathcal{L}_{\text{evi}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}, \tag{44}$$

where $\mathcal{L}_{\text{aux}}$ is the auxiliary representation-alignment objective inherited from the backbone CRS. Hard candidate filtering, top-L scene selection, and top-K recommendation are treated as non-differentiable; gradients propagate through the selected retrieval weights, score values, projection layers, and continuous gate.

D Additional Experimental Details

D.1 Implementation Details

Table 4 summarizes the training configurations and method-specific hyperparameters used in our experiments. Unless otherwise specified, the same configuration is used for ReDial and INSPIRED. The historical evidence memory is constructed exclusively from the training split and remains fixed during validation and testing.

Table 4: Training configurations and hyperparameter settings for SERA.

Category	Hyperparameter	ReDial	INSPIRED
Optimization	Optimizer	AdamW	AdamW
	Learning rate	1×10^{-4}	1×10^{-4}
	Learning-rate schedule	Linear warmup	Linear warmup
	Batch size	40	40
	Training epochs	5	15
Evidence sufficiency	Coverage candidate size K_b	10	10
	Preference threshold τ_p	0.0	0.0
	Smoothness factor κ	0.1	0.1
	Gate threshold τ_g	0.5	0.5
	Negative-preference weight λ_n	1.0	1.0
Evidence retrieval	Scene attribute threshold τ_{tag}	0.5	0.5
	Semantic weight γ_s	1.0	1.0
	Positive-preference weight γ_p	1.0	1.0
	Negative-conflict weight γ_n	1.0	1.0
	Retrieval temperature τ_r	0.1	0.1
	Retrieved scene number L	16	3
Evidence integration	Preference score weight α	0.5	1.0
	Retrieved-evidence weight β	1.5	0.5
	Negative-preference loss weight ζ_-	1.0	1.0

D.2 Additional Results for RQ1: Response Generation

Motivation. SERA shares the retrieved historical evidence between the recommendation and generation branches. We therefore examine whether evidence selected to support ranking also provides useful context for response generation, rather than improving recommendation alone.

Results. Table 5 shows that, relative to MSCRS, SERA improves BLEU-2, ROUGE-2, and Distinct-2/3/4 on ReDial. On INSPIRED, SERA exceeds MSCRS on all seven reported metrics. The simultaneous gains in overlap-based and diversity-oriented metrics indicate that sharing the retrieved evidence with the generator provides useful context without merely increasing lexical variation.

Table 5: Automatic response-generation performance on ReDial and INSPIRED.

	ReDial							INSPIRED						
Model	B-2	B-3	R-2	R-L	D-2	D-3	D-4	B-2	B-3	R-2	R-L	D-2	D-3	D-4
KGSF	0.030	0.012	0.039	0.244	0.061	0.278	0.515	0.023	0.007	0.031	0.228	0.418	1.496	2.790
UniCRS	0.045	0.021	0.058	0.285	0.433	0.748	1.003	0.022	0.009	0.029	0.212	2.686	4.343	5.520
TREA	0.022	0.008	0.039	0.175	0.242	0.615	1.176	0.013	0.002	0.027	0.195	0.958	2.565	3.411
DCRS	0.048	0.024	0.063	0.285	0.779	1.173	1.386	0.033	0.014	0.045	0.229	3.950	5.729	6.233
DisenCRS	0.050	**0.027**	0.065	0.286	0.784	1.212	1.411	0.034	0.017	0.048	0.232	4.014	6.318	6.623
PCRS-TKA	–	–	–	–	**1.162**	**1.676**	**1.845**	–	–	–	–	**6.805**	**9.906**	**10.804**
MSCRS	0.054	**0.027**	0.070	**0.294**	0.784	1.332	1.553	0.040	0.019	0.052	0.235	4.197	5.983	6.556
SERA	**0.056**	**0.027**	**0.073**	0.290	0.946	1.441	1.669	**0.043**	**0.022**	**0.057**	**0.242**	4.992	6.822	7.421