

LightCRS: Lightweight Multi-Domain Conversational Recommendation System

He Ma
University of Sydney
Sydney, NSW, Australia

Qiyang Wu
University of Sydney
Sydney, NSW, Australia

Lan You
University of Sydney
Sydney, NSW, Australia

Chen Liu*
liu.chen@nankai.edu.cn
Nankai University
Tianjin, China

Abstract

Conversational recommender systems have recently adopted vision-language models to match user dialogue intents with multimodal product information. However, extending these methods to many product domains faces a dilemma: deploying a separately adapted VLM per domain scales cost linearly with the number of domains and requires routing each dialogue to the right model, while naively merging all domains into a single VLM degrades accuracy because product characteristics vary widely across categories. We propose LightCRS, a lightweight, query-conditioned framework for unified multi-domain multimodal conversational recommendation. LightCRS precomputes multimodal product representations once offline with frozen encoders. Online, a compact 5.3M-parameter module fuses this evidence for each dialogue via query-conditioned cross-attention, whose parameters are shared across all domains, and compares candidates jointly with an inter-candidate Transformer. To enable rigorous evaluation, we construct MM-R2A-25, a 25-domain benchmark pairing natural Reddit recommendation-seeking conversations with Amazon product listings. LightCRS achieves state-of-the-art average HR@1, NDCG@5, and MRR, improving HR@1 by 35.4% over the strongest baseline while reducing per-query latency by 266 $\times$. Joint multi-domain training itself improves HR@1 by 15.2% over separately trained single-domain models.

CCS Concepts

• Information systems $\rightarrow$ Recommender systems.

Keywords

Conversational Recommender Systems, Multi-Modal Recommendation, Multi-Domain Recommendation

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

ACM Reference Format:

He Ma, Lan You, Qiyang Wu, and Chen Liu. 2018. LightCRS: Lightweight Multi-Domain Conversational Recommendation System. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 15 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Conversational recommender systems (CRSs) predict items that satisfy user intents expressed through natural-language dialogue [11, 14, 45, 48, 50, 51]. Recent progress in CRSs has been driven by vision-language models (VLMs), which integrate visual and textual product evidence to improve product understanding [8, 14, 51]. Meanwhile, e-commerce platforms host a large and growing number of product categories, and maintaining a separately adapted VLM for each category is untenable in practice. Therefore, a unified multimodal CRS that serves multiple domains is desirable.

Existing multimodal CRSs, however, face a dilemma when extended to many domains. Deploying per-domain models [14, 45, 50] scales training and inference costs approximately linearly with the number of domains, and introduces a routing problem: the system must decide which model should handle each dialogue, which is difficult when the relevant domain is ambiguous. Conversely, naively merging all domains into a single VLM can degrade recommendation accuracy, because product attributes, visual characteristics, and recommendation criteria vary widely across categories [14]. Nor can task-specific training be skipped: general-purpose proprietary models and VLMs applied through zero-shot prompting perform poorly on conversational recommendation [13, 14, 22, 43, 46].

To address these limitations, we propose LightCRS, a lightweight, query-conditioned, multi-domain framework that resolves this dilemma. Following the candidate-based CRS setting of LaViC [14], LightCRS ranks a set of candidates for each dialogue through two stages. Offline, since candidate information is static, each product is represented by five views: title, features, description, image captions, and visual tokens. These representations are precomputed once with frozen encoders, which removes VLM inference from the online path. Online, a compact 5.3M-parameter trainable module scores candidates with two components: a query-conditioned cross-attention that fuses the five views for each dialogue across 25 domains, and an inter-candidate Transformer that lets candidates be compared jointly rather than scored in isolation.

The key observation behind our design is that which product evidence matters is determined by the dialogue itself: a user describing a product’s appearance needs visual signals, whereas one asking about CPU specifications needs textual attributes. Query-conditioned cross-attention therefore allows a single set of shared parameters to behave differently for every dialogue. This addresses all three obstacles at once. A single unified model eliminates per-domain replication and the routing mechanism it requires; offline precomputation with a compact online module keeps training and inference costs low; and sharing the attention parameters across categories turns joint multi-domain training from a liability into an asset, as our multi-domain model outperforms 25 separately trained single-domain counterparts by 15.2% in HR@1. Overall, LightCRS improves average HR@1 by 35.4% over the strongest baseline while reducing per-query latency by 266×. Although designed for supervised multi-domain recommendation, LightCRS also provides modest generalization gains on unseen product categories.

To enable rigorous evaluation, we introduce *MM-R2A-25*, a 25-domain dataset linking conversational queries from Reddit recommendation communities to Amazon product listings. To the best of our knowledge, this is the first CRS benchmark designed to probe multi-domain visual and textual product understanding at scale.

Our main contributions are as follows:

- We propose LightCRS, a lightweight multi-domain multimodal CRS that decouples offline product representation from online scoring: five textual and visual product views are precomputed with frozen encoders, and a compact trainable module combines query-conditioned cross-attention, which adaptively fuses five product views for each query while sharing parameters across all 25 domains, with an inter-candidate Transformer that captures relative relationships among candidates.
- We release *MM-R2A-25*, a 25-domain multimodal CRS benchmark that links natural Reddit recommendation dialogues to Amazon product listings with verified images.
- Extensive experiments show that LightCRS achieves state-of-the-art average HR@1, NDCG@5, and MRR, improving HR@1 by 35.4% over the strongest baseline while reducing training and inference costs. Moreover, LightCRS benefits from joint multi-domain training, achieving a 15.2% HR@1 improvement over single-domain training, and exhibits modest generalization to unseen categories.

2 MM-R2A-25 Dataset

To evaluate multi-domain multimodal conversational recommendation at scale, we introduce *MM-R2A-25*, a large-scale benchmark that follows the Reddit-Amazon construction protocol [14] while extending its domain coverage to 25 diverse product categories. Unlike existing CRSs datasets, which are often limited in scale, domain coverage, or multimodal richness, *MM-R2A-25* provides aligned Reddit conversations, comprehensive Amazon product metadata, and verified product images spanning all 25 categories.

Each conversation is drawn from Reddit communities¹ where users seek product recommendations, and is paired with candidate product listings from the corresponding Amazon category². Ground-truth items are identified through explicit product mentions or links within the final utterance of each retained thread. Product images are verified and linked to the corresponding Amazon listings, enabling multimodal modeling with real visual signals. To construct this corpus, we apply GPT-3.5-turbo to each Reddit thread using the filtering template shown in Table 8 in Appendix D. We retain only conversations labeled as “yes” by the model, corresponding to threads whose final utterance contains an explicit product recommendation; these retained instances form *MM-R2A-25*. We further use LLaVA-v1.6 with the captioning prompt in Table 9 in Appendix D to generate visually grounded product descriptions. The dataset spans a wide range of domain sizes ensuring that models are evaluated across both resource-rich and resource-sparse categories. The full per-category breakdown is provided in Table 4 in Appendix A. Table 5 in Appendix A further positions *MM-R2A-25* relative to existing conversational recommendation benchmarks.

3 Method

3.1 Problem Statement

We formulate multi-domain multimodal conversational recommendation as a re-ranking problem. Following LaViC [14], we define the dialogue query at recommender turn k as the prefix $q = C_{<k} = \{u_t\}_{t=1}^{k-1}$ containing every utterance before the recommender’s k -th turn. The recommender’s k -th utterance u_k , which mentions the recommended product, is treated as the ground-truth label only and is excluded from q , ensuring that the target product does not appear in the model input. Given q , an SBERT retriever [10, 37] returns a candidate set $C = \{c_1, c_2, \dots, c_K\}$ of $K = 10$ candidates following the Reddit-Amazon setting [14], where each candidate c_i denotes a product from one of multiple e-commerce domains. Each candidate set contains exactly one ground-truth item c^+ and $K - 1$ negative candidates. Conversations associated with multiple ground-truth items are split into multiple instances. In addition, if c^+ is initially absent in the candidates, we swap out one negative candidate for c^+ to ensure precisely one ground-truth item. After candidate-set construction, we randomly shuffle all K candidate items to remove the original retrieval order while prevent positional bias. Each candidate is associated with heterogeneous multimodal product evidence, including its title $\mathbf{t}_i \in \mathbb{R}^{d_{\text{txt}}}$, features $\mathbf{f}_i \in \mathbb{R}^{d_{\text{txt}}}$, description $\mathbf{d}_i \in \mathbb{R}^{d_{\text{txt}}}$, image captions $\mathbf{I}_i \in \mathbb{R}^{N \times d_{\text{txt}}}$, and visual tokens $\mathbf{V}_i \in \mathbb{R}^{5 \times d_v}$, where $d_{\text{txt}} = 768$ and $d_v = 4096$. The goal is to learn a single multi-domain scoring model $s(c_i | q, C) \in \mathbb{R}$ that efficiently integrates multimodal product evidence across diverse domains and ranks candidates according to their alignment with the user’s conversational intent, assigning the ground-truth item $c^+ \in C$ a higher rank than all remaining candidates. Furthermore, we focus on improving multi-domain recommendation accuracy, rather than generating fluent dialogue [11, 14, 48].

¹<https://pushshift.io/>

²<https://amazon-reviews-2023.github.io/>

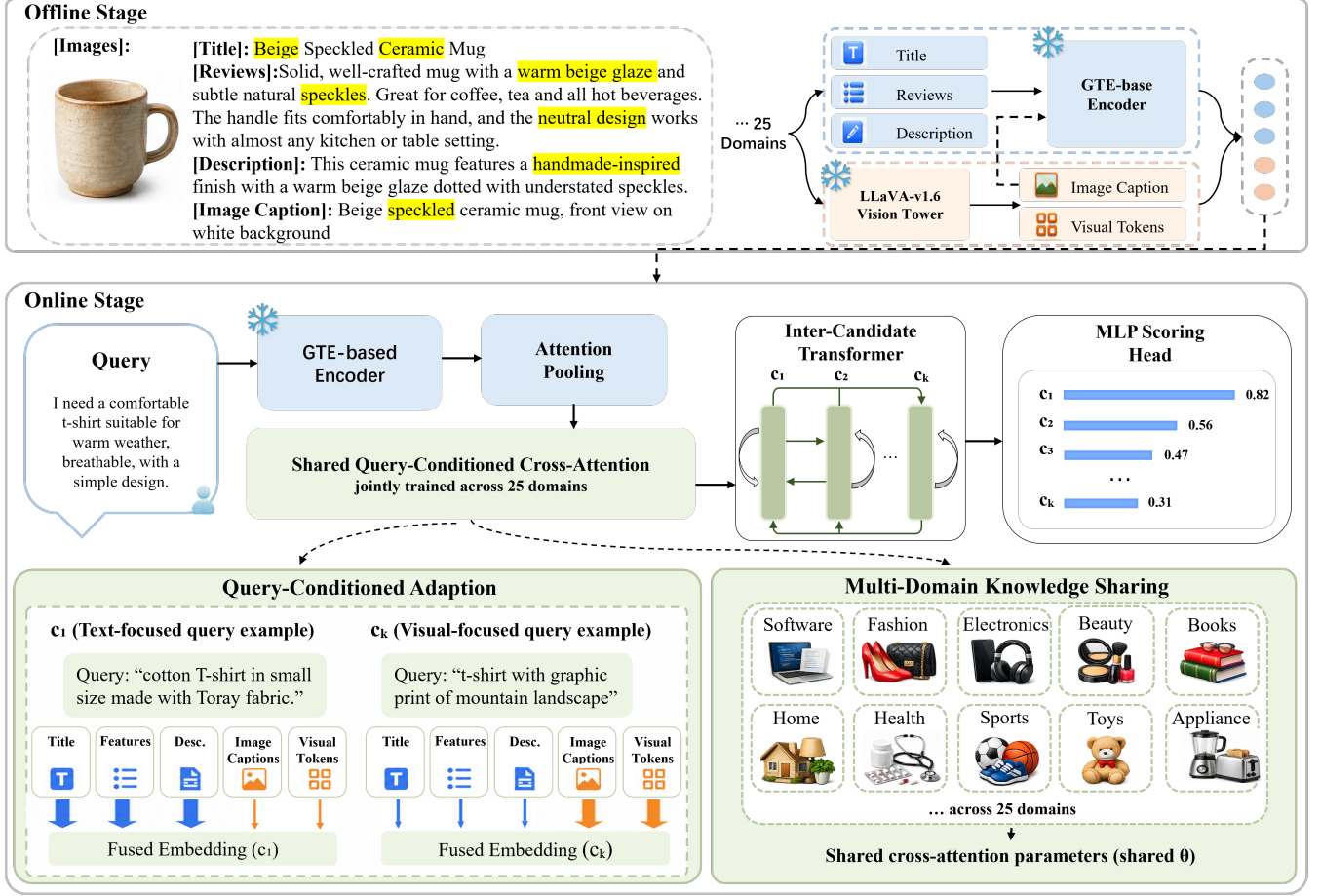


Figure 1: Architecture of LightCRS. The offline stage extracts and caches multi-view product representations using frozen GTE-base and LLaVA-v1.6. The online stage encodes the query, performs query-conditioned cross-attention over candidate representations, models candidate interactions with an inter-candidate Transformer, and outputs relevance scores through an MLP scoring head.

3.2 Proposed Framework

In this section, we propose LightCRS, a lightweight, query-adaptive framework that aligns user dialogue with product representations across diverse domains. As illustrated in Figure 1, LightCRS operates in two stages.

3.2.1 Offline Stage. To improve computational efficiency, we pre-compute static candidates’ multi-view representations offline. These cached representations are shared across training and evaluation. We next describe the encoding process for each view representation.

Text encoding. All three textual views, including the title, features, and description, are encoded by GTE-base [20]. Each view is mean-pooled over its token sequence into a single d_{txt} -dimensional embedding: title $\mathbf{t}_i \in \mathbb{R}^{d_{\text{txt}}}$, features $\mathbf{f}_i \in \mathbb{R}^{d_{\text{txt}}}$, description $\mathbf{d}_i \in \mathbb{R}^{d_{\text{txt}}}$.

Image caption encoding. Product images are captioned offline using LLaVA-v1.6. The resulting captions are then encoded by the same GTE-base model, mean-pooled per caption, and stacked to

form $\mathbf{I}_i \in \mathbb{R}^{N \times d_{\text{txt}}}$, where N denotes the number of captions per product.

Visual-token extraction. Visual features are extracted with the vision tower of LLaVA-v1.6. For one representative image, we divide it into five spatial crops and encode each crop independently. The resulting crop-level visual features are projected into the language-model hidden space, yielding one d_v -dimensional visual token per crop. This yields five spatially grounded visual tokens, each summarizing a distinct region of the product image. The resulting visual representation is $\mathbf{V}_i \in \mathbb{R}^{5 \times d_v}$.

The actual cost of the offline stage is reported in Appendix F.

3.2.2 Online Stage. At this stage, LightCRS uses a shared cross-attention module to align the on-the-fly encoded query with each candidate’s precomputed multi-view representations, enabling query-adaptive fusion and knowledge sharing across all 25 domains. An inter-candidate Transformer then compares candidates each other, and a scoring MLP outputs the final logits.

Query encoding. We tokenize the dialogue query with the same frozen GTE-base encoder into T_q subword tokens, embed them as $\mathbf{Q} \in \mathbb{R}^{T_q \times d_{\text{txt}}}$, project to $\mathbb{R}^{T_q \times D}$, and compress the resulting token sequence into a single vector $\mathbf{q} \in \mathbb{R}^D$ via learned attention pooling:

$$\mathbf{q} = \text{softmax}\left(\frac{\mathbf{w}_q(\mathbf{Q}\mathbf{W}_p)^\top}{\sqrt{D}}\right)(\mathbf{Q}\mathbf{W}_p), \quad (1)$$

with $\mathbf{W}_p \in \mathbb{R}^{d_{\text{txt}} \times D}$ and $\mathbf{w}_q \in \mathbb{R}^{1 \times D}$ learnable, yielding $\mathbf{q} \in \mathbb{R}^D$.

Query-conditioned cross-attention. For each candidate c_i , we project all five feature views into a shared D -dimensional space via learnable linear projections $\hat{\mathbf{t}}_i = \mathbf{W}_t \mathbf{t}_i$, $\hat{\mathbf{f}}_i = \mathbf{W}_f \mathbf{f}_i$, $\hat{\mathbf{d}}_i = \mathbf{W}_d \mathbf{d}_i$, $\hat{\mathbf{I}}_i = \mathbf{I}_i \mathbf{W}_I$, and $\hat{\mathbf{V}}_i = \mathbf{V}_i \mathbf{W}_V$, and concatenate them:

$$\mathbf{C}_i = [\hat{\mathbf{t}}_i; \hat{\mathbf{f}}_i; \hat{\mathbf{d}}_i; \hat{\mathbf{I}}_i; \hat{\mathbf{V}}_i] \in \mathbb{R}^{T_c \times D}, \quad (2)$$

where $T_c = 3 + N + 5$. A binary attention mask zeroes out padded positions for image-caption, handling the variable number of images per product.

The query vector is replicated across all K candidates and processed through L cross-attention layers, where each layer applies multi-head cross-attention followed by a feed-forward network, both with post-norm residuals:

$$\mathbf{q}_i^{(l)} = \text{LN}(\mathbf{q}_i^{(l-1)} + \text{MHA}(\mathbf{q}_i^{(l-1)}, \mathbf{C}_i, \mathbf{C}_i)), \quad (3)$$

$$\mathbf{q}_i^{(l)} = \text{LN}(\mathbf{q}_i^{(l)} + \text{FFN}(\mathbf{q}_i^{(l)})). \quad (4)$$

Rather than committing a priori to a fixed weighting of modalities, the cross-attention splits its weights across all five views in proportion to how relevant each view is to the current dialogue. This lets the model adapt to each user’s intent within a domain: it attends more to visual tokens when a user describes the appearance of a product, and more to textual descriptions when the user asks about CPU or GPU type of a computer. The same product can therefore be scored differently depending on the information highlighted in the dialogue.

Beyond query-specific view weighting, cross-attention also serves as the shared mechanism for multi-domain knowledge fusion. All 25 domains jointly update the same attention parameters, allowing domain-specific modality preferences to be consolidated into a unified query-view alignment function. For example, queries in text-oriented domains such as Software tend to emphasize titles, features, and descriptions, whereas queries in visually oriented domains such as Handmade place greater weight on image captions and visual tokens. Because all domains share the same cross-attention parameters, query-view alignment patterns learned from one domain can transfer to others, helping the model interpret query intent and identify the most relevant product views. This design enables a single cross-attention module to support all 25 domains.

Inter-candidate comparison. The fusion above scores each candidate independently against the query. To decide which candidate is the best, the per-candidate fused representations $\mathbf{q}_i^{(L)}$, which now carry the query-conditioned multi-view evidence of candidate c_i , are exposed to each other through a Transformer encoder. Stacking these across the K candidates gives $\mathbf{F} = [\mathbf{q}_1^{(L)}; \mathbf{q}_2^{(L)}; \dots; \mathbf{q}_K^{(L)}] \in \mathbb{R}^{K \times D}$, which is fed to the encoder as

$$\mathbf{H} = \text{TransformerEncoder}(\mathbf{F}) \in \mathbb{R}^{K \times D}, \quad (5)$$

whose i -th row $\mathbf{h}_i \in \mathbb{R}^D$ is the comparison-aware representation of candidate c_i : it preserves the query-conditioned multi-view evidence from $\mathbf{q}_i^{(L)}$ and is refined through inter-candidate self-attention.

Scoring head. The representation of each candidate is then mapped to a scalar logit by a two-layer MLP with GELU activation:

$$s_i = \mathbf{w}_2^\top \text{GELU}(\mathbf{W}_1 \mathbf{h}_i + \mathbf{b}_1) + b_2, \quad (6)$$

with $\mathbf{W}_1 \in \mathbb{R}^{D/2 \times D}$, $\mathbf{b}_1 \in \mathbb{R}^{D/2}$, $\mathbf{w}_2 \in \mathbb{R}^{D/2}$, and $b_2 \in \mathbb{R}$ learnable. Stacking the K candidates yields the logit vector $\mathbf{s} = [s_1, \dots, s_K]^\top \in \mathbb{R}^K$, which is the model’s output $s(\cdot | q, C)$ for the given query and candidate set.

3.3 Training Objective

We train the full model end-to-end with a cross-entropy loss over the K -way candidate set. The scoring head produces logits $\mathbf{s} \in \mathbb{R}^K$, and we treat reranking as K -way classification in which the target is the index y^+ of the ground-truth candidate c^+ :

$$\mathcal{L}_{\text{ce}} = -\log \frac{\exp(s_{y^+})}{\sum_{j=1}^K \exp(s_j)}. \quad (7)$$

4 Evaluations

4.1 Experimental Setup

Dataset. We evaluate all methods on the newly constructed *MM-R2A-25* dataset, which spans 25 product domains. Section 2 details the dataset construction, while Table 4 summarizes its statistics. For each domain, we partition the conversations into disjoint training, validation, and test sets using an 8:1:1 ratio.

Baselines. We compare our proposed method with three baseline categories: task-specific CRS methods, generative methods, and retrieval methods. Although earlier knowledge graph-based CRSs methods have shown effectiveness [3, 42, 54], they rely heavily on domain-specific knowledge graphs and have mainly been evaluated in movie recommendation. Generalizing such approaches consistently to our 25 diverse product domains is non-trivial. Moreover, prior studies have shown that zero-shot large language models often outperform knowledge graph-based conversational recommender systems on Reddit dialogues [10]. Therefore, we focus on these three baseline categories because they are relevant to candidate-based conversational recommendation and applicable across multiple product domains.

(1) Task-specific CRS methods. These methods are specifically designed or adapted for conversational recommendation. Among the multimodal methods, LaViC [14] distills visual knowledge from LLaVA and finetunes the resulting model with LoRA. MSCRS [45] integrates collaborative, textual, and visual semantic graphs through GNN-based prompt learning. PRIME [50] adopts a retrieval-reranking pipeline in which a Qwen2-VL-2B multimodal ranker reranks candidates using online preference feedback. Among methods without direct visual signals, RTA [11] reindexes item titles as individual tokens and calibrates item-level generation probabilities, while NBCRS [48] models collaborative preferences through item neighborhoods.

(2) Generative methods. These methods take the conversation and candidate-item information as input and use shared prompt templates to select a single most relevant item from the candidate list. MiniMax-M2.7 and MiniMax-M3 are proprietary models accessed through APIs. Vicuna-v1.5 [4] is a text-only instruction-following LLM. LLaVA-v1.5 [25] extends Vicuna through visual instruction tuning, while LLaVA-v1.6 [26] further improves multi-modal reasoning and multi-image processing.

(3) Retrieval methods. These methods rank candidate items according to the similarity between the conversation and item titles. BM25 [39] performs lexical matching based on token overlap. SBERT [37] generates sentence embeddings using a siamese BERT architecture. RoBERTa_{large} [27] uses representations from a large pretrained language encoder, while SimCSE_{large} [7] learns sentence embeddings through contrastive training.

Metrics. We use *HitRatio@1* (*HR@1*), *NormalizedDiscountedCumulativeGain@5* (*NDCG@5*), and *MeanReciprocalRank* (*MRR*) as recommendation accuracy metrics. *HR@1* directly measures how often the model identifies the correct item as the best candidate. *NDCG@5* evaluates whether the ground-truth item is placed within a top-ranked shortlist. *MRR* further considers the exact position of the ground-truth item across the entire candidate list. Beyond recommendation accuracy, we report both training and inference costs to evaluate whether performance improvements are achieved. Trainable parameters measure the optimization burden during task adaptation, whereas total parameters and on-disk model size reflect the complete deployment footprint. For inference, peak GPU memory and per-query MFLOPs characterize memory and computational demand, while wall-clock latency (ms/query) and throughput (queries/second) measure end-to-end serving efficiency.

Implementation Details. All locally executed training and inference experiments are conducted on a single NVIDIA DGX Spark workstation equipped with an NVIDIA GB10 Grace Blackwell Superchip and 128 GB of unified system memory. Using a single workstation provides a consistent hardware environment for comparing the computational costs of LightCRS and baselines. In addition, we obtain open-source pretrained models (SBERT, RoBERTa_{large}, SimCSE_{large}, Vicuna, and LLaVA) from the official HuggingFace³ repositories, while MiniMax-M3 and MiniMax-M2.7 are accessed through their official APIs and therefore are excluded from hardware-dependent efficiency comparisons. Moreover, for all methods, we report the mean and standard deviation of *HR@1*, *NDCG@5*, and *MRR* over three independent runs. Retrieval methods (BM25, SBERT, RoBERTa, and SimCSE) are deterministic given fixed inputs, so we omit their zero standard deviations. Hyperparameter details and model-specific configurations are provided in Appendix E to support reproducibility.

4.2 Overall Performance

We compare LightCRS with three categories of baselines: task-specific CRS methods, generative methods, and retrieval methods. For task-specific CRS methods, generative methods and our method, we rely on candidate-based recommendation setup with 10 items, using SBERT which shows the highest performance among retrievals

methods [14]. The generative baselines are evaluated under two input configurations: using item titles only (Vicuna-v1.5, LLaVA-v1.5, and LLaVA-v1.6) and using both item titles and product images (LLaVA-v1.6, MiniMax-M2.7, and MiniMax-M3). Table 1 compares LightCRS with all baselines in terms of average recommendation performance, training cost, and inference efficiency. Since the multi-modal task-specific CRS methods are the only baselines that remain competitive in recommendation accuracy, we restrict the detailed training- and inference-cost comparison to this group. The complete per-domain comparison is divided between two tables. Table 6 in Appendix B reports LightCRS model with the task-specific CRSs baselines, while Table 7 in Appendix C provides the comparison with the retrieval and generative baselines. Considered jointly, the two tables compare LightCRS model against all baseline groups. Table 7 additionally reports the performance of LightCRS on unseen domains, which is further analyzed in Section 5.

4.2.1 Comparison with task-specific CRSs baselines. Table 6 compares LightCRS with the non-visual task-specific methods RTA and NBCRS, as well as the multimodal methods LaViC, MSCRS, and PRIME. Across the 25 domains, LightCRS achieves the best performance in 24 domains across *HR@1*, *NDCG@5*, and *MRR*. The results show a gap between the non-visual and multimodal task-specific baselines. RTA achieves an average *HR@1* of 0.0812, exceeding the generative and retrieval baselines but remaining well below the multimodal subgroup. Its performance also varies considerably across domains, reaching an *HR@1* of 0.2183 on Software and 0.2091 on Handmade, both above its overall average and competitive with several multimodal methods in these domains.

Among the multimodal baselines, PRIME achieves the strongest average results, with an *HR@1* of 0.1934, an *NDCG@5* of 0.3798, and an *MRR* of 0.3680. The results also show that the approximately 7.6B-parameter LaViC underperforms the considerably smaller MSCRS (0.3B) and PRIME (2.2B), suggesting that increasing model scale alone does not guarantee higher recommendation performance. In contrast, LightCRS achieves the best average results across all three metrics, with an *HR@1* of 0.2618, an *NDCG@5* of 0.4540, and an *MRR* of 0.4367. Compared with PRIME, LightCRS improves *HR@1* by 35.4%, *NDCG@5* by 19.5%, and *MRR* by 18.7%. The larger gain in *HR@1* indicates that LightCRS is particularly effective at identifying the single best item from the candidate set, while its consistent improvements in *NDCG@5* and *MRR* show that it also produces a stronger overall candidate ranking. LightCRS achieves these performance gains with lower training and inference costs. As reported in Table 1, it requires only 5.3M trainable parameters, 0.1B total parameters, a 0.3 GB model size, 0.3 GB peak GPU memory, 97K MFLOPs per query, and 9.8 ms per-query latency, while supporting a throughput of 101.6 queries/s. Compared with PRIME, LightCRS uses 1.7× fewer trainable parameters and 22× fewer total parameters, while reducing model size by 15×, peak GPU memory by 30×, per-query FLOPs by 153×, latency by 266×, and increasing throughput by 254×. These results demonstrate that LightCRS provides a stronger balance between recommendation accuracy and computational efficiency than existing task-specific CRSs baselines.

4.2.2 Comparison with retrieval and generative baselines. Combining LightCRS results reported in Table 6 with the baseline results

³<https://huggingface.co/>

reported in Table 7, we compare LightCRS with the retrieval methods BM25, SBERT, RoBERTa, and SimCSE; the generative models Vicuna-v1.5, LLaVA-v1.5, and LLaVA-v1.6; and the proprietary models MiniMax-M2.7 and MiniMax-M3. The retrieval methods achieve average HR@1 values ranging from 0.0142 for BM25 to 0.0603 for SBERT. The open-source generative models achieve only 0.0629–0.0753 average HR@1 despite relying on larger 6.7–7.1B-parameter backbones, yielding only limited improvements over SBERT. The results of both groups vary considerably across domains. For example, SBERT reaches an HR@1 of 0.2166, an NDCG@5 of 0.4017, and an MRR of 0.3700 on Home. Its HR@1 in this domain surpasses every open-source generative baseline, as well as RTA and MSCRS. LLaVA-v1.6 with image inputs achieves an HR@1 of 0.2227 on Software, exceeding its overall average performance. This indicates that incorporating product images into LLaVA-v1.6 has inconsistent effects across domains, improving recommendation performance in some domains while degrading it in others. This observation aligns with previous findings that directly incorporating candidate images does not necessarily improve recommendation accuracy [14, 28]. These domain-specific spikes show that textual retrieval and general-purpose large model prompting can be effective in particular categories, but their advantages do not transfer consistently across the full set of domains.

The proprietary models perform better on average than the open-source generative baselines, with MiniMax-M2.7 and MiniMax-M3 achieving average HR@1 values of 0.1025 and 0.1085, respectively. However, their results are also strongly domain-dependent. MiniMax-M3 reaches an HR@1 of 0.3325 on Home, giving it the second-highest baseline result in that domain, but achieves only 0.1025 on Baby Products, where it falls behind the multimodal task-specific methods and LightCRS. In contrast, LightCRS achieves an average HR@1 of 0.2618, outperforming both retrieval and generative models. At the per-domain level, LightCRS achieves the highest HR@1 in 24 of the 25 domains against open-source baselines. It also achieves the highest NDCG@5 and MRR in all 25 domains against the retrieval methods; these two metrics are not reported for the generative baselines because they output only a single recommendation rather than a candidate ranking. Taken together, these results show that neither textual-similarity retrieval nor prompt-based scoring with general-purpose large models can achieve consistent recommendation performance across diverse product domains. In contrast, LightCRS leverages joint multi-domain training and multimodal product evidence, leading to robust performance across domains.

4.3 Ablation Study

4.3.1 Component Ablation. To understand the contribution of each component in our framework, we conduct ablation experiments by removing one component at a time and evaluating the resulting recommendation performance, as shown in Table 2.

Inter-candidate comparison drives top-1 accuracy. Although removing inter-candidate comparison (ICC) retains the strongest performance among all ablated variants (HR@1 of 0.1432, NDCG@5 of 0.3505, MRR of 0.3389), it still reduces HR@1 by 45.3% relative to the full model, confirming that direct candidate-to-candidate comparison is essential for accurate top-1 selection.

Cross-attention and visual tokens contribute most. Removing cross-attention (CA) or visual tokens (VT) has the greatest impact on top-1 accuracy, with HR@1 dropping to 0.0998 and 0.0987, respectively, underscoring the importance of query-conditioned multimodal interaction and visually grounded representations. Image descriptions (ID) further supply useful visual semantics, though their contribution is smaller than that of the raw visual tokens.

Title is the most informative textual view. Among the textual feature views, removing the title yields the lowest NDCG@5 (0.2832) and MRR (0.2871) among all ablated variants, while product descriptions and features contribute complementary signals.

4.3.2 Single-domain vs Multi-domain. Table 3 compares our full multi-domain model with a variant trained separately on each domain. In the single-domain setting, a separate model is trained for each product category, and HR@1, NDCG@5, and MRR are averaged across the 25 domain-specific models. Compared with this setting, our full multi-domain model improves HR@1, NDCG@5, and MRR by 15.2%, 5.5%, and 6.2%, respectively, demonstrating that multi-domain knowledge sharing improves recommendation performance.

5 Further Analysis

CRSs on unseen domains remains challenging, as even strong proprietary models achieve limited performance without task-specific adaptation [13, 14, 22, 43, 46]. A key challenge in unseen-domain CRSs is the domain-dependent semantics. For example, “light” may denote low weight in hardware, light wearing comfort, a thin texture in cosmetics, or low computational cost in software. Although LightCRS is not explicitly designed for recommendation on unseen domains, we evaluate its transferability through 5-fold domain-split testing across 25 domains. In each fold, the model is trained on 20 domains and evaluated on 5 unseen domains. Table 7 in Appendix C shows the comparison of LightCRS on unseen domains with retrieval and generative baselines, for which the evaluation domains are also unseen. LightCRS slightly outperforms the strongest MiniMax-M3 baseline. However, the absolute performance remains relatively low, indicating that recommendation on unseen domains remains challenging.

6 Related Work

Recommendation systems have long studied how to infer user preferences and rank relevant items from behavioral signals. Conventional sequential recommendation focuses on predicting users’ next items from historical interaction sequences. Early methods model sequential transitions with Markov chains and recurrent networks [12, 38], followed by convolutional and self-attention architectures such as Caser [41], SASRec [17], and BERT4Rec [40]. More recent self-supervised and contrastive learning methods further improve sequence representations by exploiting auxiliary sequence-level signals [47, 53]. Along another line, candidate reranking methods refine an initially retrieved item list by modeling listwise context and candidate interactions. Representative methods include listwise context models [1], groupwise scoring functions [2], SetRank [33],

Table 1: Average recommendation performance and resource usage across methods. Gray -shaded rows denote multimodal task-specific CRS methods, training and inference costs for our method are compared only with those of methods in this group. “–” indicates that the value cannot be calculated. Bold and underlined values denote the best and second-best results. Abbreviations: Rec. Perf.=recommendation performance, Train. P.=trainable parameters, Total P.=total parameters, GPU Mem.=peak GPU memory, Lat.=latency (ms/query), and Thpt.=throughput (queries/s).

Method	Rec. Perf.			Training Cost			Inference Efficiency			
	HR@1 ↑	NDCG@5 ↑	MRR ↑	Train. P. ↓	Total P. ↓	Size ↓	GPU Mem. ↓	MFLOPs ↓	Lat. ↓	Thpt. ↑
<i>Retrieval</i>										
BM25 [39]	0.0142	0.0434	0.1724	–	–	8.4MB	0.0GB	0.0	65.7	15.21
SBERT [37]	0.0603	0.1224	0.2282	–	0.1B	0.4GB	0.5GB	70,675.1	17.5	58.14
RoBERTa [27]	0.0483	0.1065	0.2164	–	0.4B	1.4GB	1.5GB	154,634.3	24.5	40.94
SimCSE [7]	0.0195	0.0567	0.1837	–	0.4B	1.4GB	1.5GB	295,071.7	29.1	34.72
<i>Generative</i>										
Vicuna-v1.5 [4]	0.0633 $\pm$ 0.0083	–	–	–	6.7B	13.5GB	14.6GB	28,435,892.3	1825.4	0.55
LLaVA-v1.5 [25]	0.0629 $\pm$ 0.0112	–	–	–	7.1B	14.1GB	15.3GB	32,148,273.5	2087.6	0.48
LLaVA-v1.6 [26]	0.0683 $\pm$ 0.0108	–	–	–	7.1B	14.1GB	15.8GB	35,832,914.7	2341.2	0.43
LLaVA-v1.6 + title/image [26]	0.0753 $\pm$ 0.0098	–	–	–	7.1B	14.1GB	16.1GB	39,527,168.8	2589.5	0.39
MiniMax-M2.7	0.1025 $\pm$ 0.0140	–	–	–	–	–	–	–	–	–
MiniMax-M3	0.1085 $\pm$ 0.0143	–	–	–	–	–	–	–	–	–
<i>Task-specific CRSs</i>										
NBCRS [48]	0.0490 $\pm$ 0.0118	0.0941 $\pm$ 0.0120	0.1473 $\pm$ 0.0097	22.9M	0.02B	0.1GB	0.2GB	7,344.7	3.2	319.4
RTA [11]	0.0812 $\pm$ 0.0074	–	–	40.3M	6.8B	13.6GB	14.1GB	6,352,337.1	116.8	8.57
LaViC [14]	0.1338 $\pm$ 0.0117	–	–	24.6M	7.6B	15.2GB	16.2GB	37,765,450.8	1940.3	0.5
MSCRS [45]	0.1582 $\pm$ 0.0077	0.2998 $\pm$ 0.0078	0.2856 $\pm$ 0.0069	22.2M	<u>0.3B</u>	<u>1.2GB</u>	<u>2.0GB</u>	<u>130,642.0</u>	<u>112.2</u>	<u>8.9</u>
PRIME [50]	0.1934 $\pm$ 0.0143	0.3798 $\pm$ 0.0117	0.3680 $\pm$ 0.0111	<u>9.2M</u>	2.2B	4.4GB	8.9GB	14,849,125.6	2609.1	0.4
LightCRS	0.2618$\pm$0.0104	0.4540$\pm$0.0093	0.4367$\pm$0.0083	5.3M	0.1B	0.3GB	0.3GB	97,078.6	9.8	101.6

Table 2: Per-component ablation. Each row removes a single component from the full model, and we report the resulting average HR@1, NDCG@5, and MRR across all 25 product domains. Abbreviations: ICC – Inter-Candidate Comparison; ID – Image Descriptions; Feat – Features; Desc – Description; CA – Cross-Attention; VT – Visual Tokens.

Model Variant	HR@1 ↑	NDCG@5 ↑	MRR ↑
w/o ICC	0.1432 $\pm$ 0.0079	0.3505 $\pm$ 0.0107	0.3389 $\pm$ 0.0089
w/o ID	0.1109 $\pm$ 0.0069	0.3004 $\pm$ 0.0034	0.3015 $\pm$ 0.0039
w/o Feat	0.1054 $\pm$ 0.0064	0.2994 $\pm$ 0.0054	0.2969 $\pm$ 0.0047
w/o Desc	0.1021 $\pm$ 0.0002	0.3011 $\pm$ 0.0017	0.2972 $\pm$ 0.0006
w/o Title	0.1002 $\pm$ 0.0012	0.2832 $\pm$ 0.0023	0.2871 $\pm$ 0.0015
w/o CA	0.0998 $\pm$ 0.0036	0.3052 $\pm$ 0.0041	0.2971 $\pm$ 0.0001
w/o VT	0.0987 $\pm$ 0.0024	0.2938 $\pm$ 0.0033	0.2907 $\pm$ 0.0002
LightCRS (Ours)	0.2618$\pm$0.0104	0.4540$\pm$0.0093	0.4367$\pm$0.0083

Table 3: Single-domain vs. Multi-domain. The single-domain regime trains a separate model per product category and we average metrics across the 25 domains; the full model trains one shared model jointly on all 25 domains. Abbreviations: Sing.=Single-domain, Multi.=Multi-domain.

Training Regime	HR@1 ↑	NDCG@5 ↑	MRR ↑
Sing.	0.2272 $\pm$ 0.0090	0.4305 $\pm$ 0.0030	0.4113 $\pm$ 0.0040
Multi. (Ours)	0.2618$\pm$0.0104	0.4540$\pm$0.0093	0.4367$\pm$0.0083

and personalized reranking modules [34]. Recent industrial generative recommender systems, such as OneMall [52], further reformulate large-scale e-commerce recommendation as autoregressive semantic-ID generation. Together, these studies provide important foundations for sequence modeling, candidate comparison, and industrial-scale recommendation, but they mainly rely on behavioral histories, retrieved item lists, or platform-specific interaction signals rather than natural-language conversational interactions.

CRSs extend recommendation by eliciting user preferences through multi-turn dialogue and refining recommendations according to the evolving conversational context. Early CRSs datasets and systems, such as ReDial [19], GoRecDial [16], and TG-ReDial [55], established conversational recommendation as the task of predicting relevant items from natural-language interactions, mostly in narrow domains such as movies. With the development of neural approaches, CRSs began to jointly model dialogue context, user preference, external knowledge, and item recommendation [3, 42, 54]. More recently, LLM-based CRSs methods have enhanced conversational recommendation by leveraging the instruction-following ability and world knowledge of large language models [10, 11, 13, 57]. Despite these advances, most existing CRSs methods remain primarily text-based and are not designed to exploit rich multimodal product evidence in e-commerce scenarios.

Recent studies have further extended CRSs from text-only interactions to multimodal product understanding by leveraging VLMs to jointly encode visual and textual product evidence [8, 14, 51]. These methods show that visual signals can complement textual dialogue, especially in visually grounded e-commerce scenarios where product appearance, style, color, and packaging are important for

user decisions. However, VLMs-based CRSs still face three major limitations when deployed across diverse e-commerce categories: (1) training and inference costs that scale approximately linearly with the number of product domains, (2) performance degradation when combining heterogeneous domains into a single VLM, and (3) an additional routing challenge when multiple domain-specific models must be selected per dialogue. These limitations motivate LightCRS, which precomputes multimodal product views offline and performs lightweight dialogue-conditioned reranking online, using query-conditioned cross-attention and inter-candidate comparison to improve multi-domain multimodal CRSs recommendation accuracy while maintaining low training and inference costs.

7 Conclusion

We presented LightCRS, a lightweight framework for unified multi-domain multimodal conversational recommendation. By moving product representation computation offline, LightCRS avoids repeatedly encoding static candidate information during training and inference. Its compact online design enables a single model to adapt to different dialogue intents, share knowledge across product domains, and compare candidates jointly. In addition, we introduce MM-R2A-25, a benchmark linking Reddit recommendation conversations to Amazon product listings across 25 categories. Experiments show that LightCRS provides state-of-the-art average recommendation performance compared to the evaluated baselines while using fewer trainable parameters and lower inference costs.

Limitations and Future Work. Our method follows a candidate-based reranking pipeline, so its overall performance remains partly constrained by the quality of the first-stage retriever. Improving candidate retrieval could therefore yield further gains in recommendation performance. On unseen domains, LightCRS still underperforms its in-domain supervised counterpart, while even strong proprietary models achieve limited performance. This indicates that achieving effective recommendation on unseen domains remains an important direction for future work. Moreover, our model focuses on improving recommendation accuracy, while extending it to support dialogue generation remains another important direction for future work.

References

- [1] Qingyao Ai, Keping Bi, Jiafeng Guo, and W Bruce Croft. 2018. Learning a deep listwise context model for ranking refinement. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 135–144.
- [2] Qingyao Ai, Xuanhui Wang, Sebastian Bruch, Nadav Golbandi, Michael Bender-sky, and Marc Najork. 2019. Learning groupwise multivariate scoring functions using deep neural networks. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*. 85–92.
- [3] 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)*. 1803–1813.
- [4] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) 2, 3 (2023), 6.
- [5] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931* (2015).
- [6] Zuohui Fu, Yikun Xian, Yaxin Zhu, Shuyuan Xu, Zelong Li, Gerard De Melo, and Yongfeng Zhang. 2021. Hoops: Human-in-the-loop graph reasoning for conversational recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2415–2421.
- [7] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*. 6894–6910.
- [8] Ramin Giah, Kehui Yao, Sriram Korpeoglu, Kai Zhao, Vahid Mirjalili, Jianpeng Xu, Topojoy Biswas, Evren Korpeoglu, and Kannan Achan. 2025. VI-clip: Enhancing multimodal recommendations via visual grounding and llm-augmented clip embeddings. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 482–491.
- [9] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyang Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. 8142–8152.
- [10] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 720–730.
- [11] Zhankui He, Zhouhang Xie, Harald Steck, Dawen Liang, Rahul Jha, Nathan Kallus, and Julian McAuley. 2024. Reindex-then-adapt: Improving large language models for conversational recommendation. *arXiv preprint arXiv:2405.12119* (2024).
- [12] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).
- [13] Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2025. Recommender ai agent: Integrating large language models for interactive recommendations. *ACM Transactions on Information Systems* 43, 4 (2025), 1–33.
- [14] Hyunsik Jeon, Satoshi Koide, Yu Wang, Zhankui He, and Julian McAuley. 2025. Adapting Large Vision-Language Models to Visually-Aware Conversational Recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 1037–1048.
- [15] Meihuizi Jia, Ruixue Liu, Peiying Wang, Yang Song, Zexi Xi, Haobin Li, Xin Shen, Meng Chen, Jinhui Pang, and Xiaodong He. 2022. E-ConvRec: a large-scale conversational recommendation dataset for E-commerce customer service. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 5787–5796.
- [16] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul A Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. 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)*. 1951–1961.
- [17] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.
- [18] Minjin Kim, Minju Kim, Hana Kim, Beong-woo Kwak, Seongku Kang, Youngjae Yu, Jinyoung Yeo, and Dongha Lee. 2024. Pearl: A review-driven persona-knowledge grounded conversational recommendation dataset. *Findings of the Association for Computational Linguistics: ACL 2024* (2024), 1105–1120.
- [19] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. *Advances in neural information processing systems* 31 (2018).
- [20] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281* (2023).
- [21] Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. 2024. Llm-redial: A large-scale dataset for conversational recommender systems created from user behaviors with llms. In *Findings of the Association for Computational Linguistics: ACL 2024*. 8926–8939.
- [22] Hao Liao, Jiwei Zhang, Jianxun Lian, Wensheng Lu, Mingqi Wu, Yong Zhang, Yitian Huang, Mingyang Zhou, Rui Mao, et al. 2026. Eliminating Out-of-Domain Recommendations in LLM-based Recommender Systems: A Unified View. In *Findings of the Association for Computational Linguistics: ACL 2026*. 6251–6271.
- [23] Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. MMConv: an environment for multimodal conversational search across multiple domains. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 675–684.
- [24] 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*. 9591–9600.
- [25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 26296–26306.
- [26] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA: Improved reasoning, ocr, and world knowledge.

- [27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [28] Yuqing Liu, Yu Wang, Lichao Sun, and Philip S Yu. 2024. Rec-gpt4v: Multimodal recommendation with large vision-language models. *arXiv preprint arXiv:2402.08670* (2024).
- [29] Yuanxing Liu, Weinan Zhang, Baohua Dong, Yan Fan, Hang Wang, Fan Feng, Yifan Chen, Ziyu Zhuang, Hengbin Cui, Yongbin Li, et al. 2023. U-need: A fine-grained dataset for user needs-centric e-commerce conversational recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*. 2723–2732.
- [30] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 4335–4347.
- [31] Srijata Maji, Moghis Fereidouni, Vinaik Chhetri, Umar Farooq, and AB Siddique. 2024. Mobileconvrec: A conversational dataset for mobile apps recommendations. *arXiv preprint arXiv:2405.17740* (2024).
- [32] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Open-dialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 845–854.
- [33] Liang Pang, Jun Xu, Qingyao Ai, Yanyan Lan, Xueqi Cheng, and Jirong Wen. 2020. Setrank: Learning a permutation-invariant ranking model for information retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 499–508.
- [34] Changhua Pei, Yi Zhang, Yongfeng Zhang, Fei Sun, Xiao Lin, Hanxiao Sun, Jian Wu, Peng Jiang, Junfeng Ge, Wenwu Ou, et al. 2019. Personalized re-ranking for recommendation. In *Proceedings of the 13th ACM conference on recommender systems*. 3–11.
- [35] Ahmad Bin Rabiah, Nafis Sadeq, and Julian McAuley. 2024. Bridging conversational and collaborative signals for conversational recommendation. *arXiv preprint arXiv:2412.06949* (2024).
- [36] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the 20th Annual SIGDial Meeting on Discourse and Dialogue*. 353–360.
- [37] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*. 3982–3992.
- [38] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.
- [39] Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Vol. 4. Now Publishers Inc.
- [40] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.
- [41] Jiayi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 565–573.
- [42] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 1929–1937.
- [43] Yancheng Wang, Ziyang Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Yanbin Lu, Xiaojiang Huang, and Yingzhen Yang. 2024. Recmind: Large language model powered agent for recommendation. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 4351–4364.
- [44] Zihan Wang, Xiaocui Yang, Yongkang Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2025. Muse: A multimodal conversational recommendation dataset with scenario-grounded user profiles. In *Findings of the Association for Computational Linguistics: ACL 2025*. 1027–1053.
- [45] Yibiao Wei, Jie Zou, Weikang Guo, Guoqing Wang, Xing Xu, and Yang Yang. 2025. MSCRS: Multi-modal 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*. 42–52.
- [46] Yunjia Xi, Weiwen Liu, Jianghao Lin, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. Memocrs: Memory-enhanced sequential conversational recommender systems with large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2585–2595.
- [47] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 1259–1273.
- [48] Zhouhang Xie, Junda Wu, Hyunsik Jeon, Zhankui He, Harald Steck, Rahul Jha, Dawen Liang, Nathan Kallus, and Julian McAuley. 2024. Neighborhood-based collaborative filtering for conversational recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1045–1050.
- [49] Hu Xu, Seungwhan Moon, Honglei Liu, Bing Liu, Pararth Shah, and Philip S Yu. 2020. User memory reasoning for conversational recommendation. In *Proceedings of the 28th international conference on computational linguistics*. 5288–5308.
- [50] Zhenrui Yue, Huimin Zeng, Yueqi Wang, Julian McAuley, and Dong Wang. 2025. Preference-Optimized Retrieval and Ranking for Efficient Multimodal Recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 3692–3703.
- [51] Daoze Zhang, Chenghan Fu, Zhanheng Nie, Jianyu Liu, Wanxian Guan, Yuan Gao, Jun Song, Pengjie Wang, Jian Xu, and Bo Zheng. 2026. MOON: Generative MLLM-based multimodal representation learning for e-commerce product understanding. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*. 924–933.
- [52] Kun Zhang, Jingming Zhang, Wei Cheng, Yansong Cheng, Jiaqi Zhang, Hao Lu, Xu Zhang, Haixiang Gan, Jiangxia Cao, Tenglong Wang, et al. 2026. OneMall: One Architecture, More Scenarios—End-to-End Generative Recommender Family at Kuaishou E-Commerce. *arXiv preprint arXiv:2601.21770* (2026).
- [53] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1893–1902.
- [54] 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 *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1006–1014.
- [55] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. In *Proceedings of the 28th international conference on computational linguistics*. 4128–4139.
- [56] Tianyu Zhu, Myong Chol Jung, and Jesse Clark. 2025. Generalized contrastive learning for multi-modal retrieval and ranking. In *Companion Proceedings of the ACM on Web Conference 2025*. 661–670.
- [57] Yaochen Zhu, Harald Steck, Dawen Liang, Yinhan He, Vito Ostuni, Jundong Li, and Nathan Kallus. 2025. Rank-GRPO: Training LLM-based Conversational Recommender Systems with Reinforcement Learning. *arXiv preprint arXiv:2510.20150* (2025).

A Dataset Details

Table 4 presents the full per-category breakdown of the *MM-R2A-25* dataset. Table 5 compares it with existing CRSs benchmarks. Early efforts (e.g., ReDial [19], GoRecDial [16], and TG-ReDial [55]) primarily focus on movies, often with crowd-sourced or synthetic dialogues. Recent datasets have expanded into broader domains (e.g., beauty, home and fashion) or used natural Reddit conversations [15, 29, 35], but still lack multimodal product information. In contrast, *MM-R2A-25* covers 25 diverse product categories, enabling comprehensive evaluations of multi-domain multimodal conversational recommender systems. Following LaViC [14], we use SBERT to retrieve candidate sets for each conversation, with a Recall@10 of 0.6479.

B Per-Domain Results

Table 6 reports the full per-domain breakdown of HR@1, NDCG@5, and MRR across all 25 domains in *MM-R2A-25*, comparing LightCRS with three baseline groups: task-specific CRS, generative, retrieval methods. These results complement the aggregate analysis in Section 4.2.

C Generalization to Held-Out Unseen Domains

To provide a supplementary assessment of unseen-domain generalization, we conduct a five-fold held-out-domain evaluation. The 25 product domains are partitioned into 5 non-overlapping groups,

Table 4: Summary of the MM-R2A-25 dataset. #Conv: number of conversations. #Turns: number of conversational turns. #Items: number of unique products. #Images: total number of verified product images.

Category	#Conv	#Turns	#Items	#Images
Appliances	2,291	4,582	739	4,062
Arts, Crafts & Sewing	1,267	2,315	823	3,473
Automotive	1,512	2,869	811	3,657
Baby Products	3,336	6,672	1,975	11,561
Beauty	7,672	22,966	5,433	28,082
Books	4,707	13,250	1,978	1,864
CDs & Vinyl	3,818	7,345	1,023	1,562
Cell Phones & Accessories	3,576	7,612	1,926	11,580
Digital Music	1,350	2,700	613	988
Electronics	3,996	7,992	2,913	12,323
Fashion	8,039	21,831	6,716	31,162
Grocery & Gourmet Food	2,864	5,222	1,940	8,186
Handmade Products	2,856	5,712	781	4,672
Health	1,590	2,974	1,134	6,050
Industrial & Scientific	3,978	7,411	2,268	10,520
Home	3,701	6,675	3,077	18,505
Movies & TV	2,284	4,455	972	1,315
Musical Instruments	2,230	4,460	1,327	4,693
Office Products	2,905	5,513	1,684	8,375
Patio, Lawn & Garden	1,148	2,044	874	3,989
Pet Supplies	5,219	9,628	2,708	13,293
Software	2,277	4,554	283	1,625
Sports & Outdoors	3,180	7,405	1,156	6,459
Toys & Games	1,909	3,714	1,113	5,462
Video Games	19,142	38,284	5,854	15,810

and in each fold, LightCRS is trained on 20 domains and evaluated on the remaining 5 domains, whose data are entirely excluded from training. Table 7 compares LightCRS under this unseen-domain setting with retrieval and generative baselines, for which the evaluation domains are also unseen.

D Prompt Templates

This appendix consolidates the four prompt templates that govern our data curation and model training pipeline. Table 8 isolates Reddit threads that end in an explicit product recommendation, Table 9 elicits visually grounded captions, and Tables 10–11 drive the recommendation step in the text-only and the text-plus-image configurations, respectively.

E Hyperparameter Details

No domain-specific hyperparameter tuning was conducted. We used one fixed configuration across all 25 product domains and trained the model (Section 3) under this unified protocol.

The common projection dimension is $D = 256$. The query–candidate cross-attention stack has $L = 2$ layers with 4 attention heads and a feed-forward dimension of 1024 ($4 \times D$), and the inter-candidate comparison Transformer mirrors this configuration (2 layers, 4 heads, 1024 FFN). The scoring head is a 2-layer MLP ($D \rightarrow D/2 \rightarrow 1$), and dropout is fixed at 0.1 across all trainable modules. Frozen encoder outputs feed in at their native hidden

sizes: $d_q = 768$ for the GTE-base text encoder and $d_v = 4096$ for the LLaVA-v1.6 visual tokens, with a maximum of $N = 15$ image descriptions and 5 visual tokens per candidate item. The training objective is a cross-entropy classification loss over the K -way candidate set. Models are trained for up to 20 epochs with early stopping (patience = 2) on validation loss, using a batch size of 16 and the AdamW optimizer ($\text{lr} = 1 \times 10^{-4}$, weight decay 1×10^{-5}) under a constant learning-rate schedule with no warmup or decay. Gradient clipping (max norm 1.0) and BF16 mixed precision are applied throughout.

F Offline Preprocessing Cost and Storage Footprint

LightCRS moves product-side multimodal encoding from online scoring to offline preprocessing (Section 3.2.1). This appendix reports the computational cost and storage overhead of this offline process.

Measurement methodology. All measurements are conducted on a single NVIDIA DGX Spark workstation. We evaluate four offline processing steps: text encoding(title/features/descriptions), image caption generation, caption encoding, and visual-token extraction. Each step is measured independently. We sample 50 products from

Table 5: Comparison of existing conversational recommendation datasets. MM-R2A-25 consists of natural Reddit conversations aligned with Amazon Reviews 2023 across 25 diverse product categories.

Dataset	#Conv	#Turns	#Items	Domain	Source
Facebook Rec [5]	1M	6M	–	Movies	Synthetic
ReDial [19]	10K	182K	6.2K	Movies	Crowd-sourced
GoRecDial [16]	9K	170K	–	Movies	Crowd-sourced
OpenDialKG [32]	15K	91K	–	Movies, Music, etc.	Crowd-sourced
TG-ReDial [55]	10K	129K	–	Movies	Synthetic
DuRecDial 2.0 [30]	16.5K	255K	–	Movies, Music, etc.	Crowd-sourced
CCPE-M [36]	502	11K	–	Movies	Crowd-sourced
INSPIRED [9]	1K	35K	1.9K	Movies	Crowd-sourced
Reddit-Movie Base [35]	85K	133K	24.3K	Movies	Natural
Reddit-Movie Large [35]	634K	1.6M	51.2K	Movies	Natural
U-Need [29]	7K	53K	–	E-commerce	Natural
E-ConvRec [15]	25K	775K	–	E-commerce	Natural
Hoops [6]	–	11.6M	–	E-commerce	Synthetic
MGConvRec [49]	7K	73K	–	Restaurant	Crowd-sourced
MMConv [23]	5K	39K	–	Travel	Crowd-sourced
MobileConvRec [31]	12.2K	156K	1.7K	Mobile Apps	Synthetic
Pearl [18]	57.2K	482K	9.4K	Movies	Synthetic
LLM-REDIAL [21]	47.6K	482.6K	–	Movie, Book, Sports	Synthetic
MUSE [44]	7K	83K	13.7K	Clothing, Shoes, Jewelry	Synthetic
SCREEN [24]	20K	–	–	Movies	Synthetic
Reddit-Amazon [14]	19K	51K	15K	Beauty, Fashion, Home	Natural
Marqo-GS-10M [56]	98.2K	–	5.50M	Fashion, Home	Natural
MM-R2A-25 (Ours)	97K	208K	50K	Multi-domain	Natural

the Books category, measure the runtime and peak GPU memory after one warm-up pass, and linearly extrapolate the results to the full MM-R2A-25 catalog since each product is processed independently.

Offline preprocessing cost. Table 12 reports the runtime, throughput, peak GPU memory, and extrapolated cost on the full catalog. Caption generation dominates the preprocessing cost, requiring 593.46 GPU-hours for generating captions of all product images,

while the remaining encoding stages require less than 3 GPU-hours in total. This cost is incurred only once per catalog. After preprocessing, the online module operates only on cached representations without additional product-side encoding.

Storage footprint. Table 13 summarizes the storage cost of cached product representations. We notice that visual tokens accounts for the majority of storage usage.

Table 6: Per-domain comparison of the task-specific CRS baselines and LightCRS across 25 Amazon product categories. Panel A reports HR@1 for all methods, whereas Panel B reports NDCG@5 and MRR only for ranking-based methods. Bold and underline indicate the best and second-best results, respectively.

Panel A: HR@1.							
Domain	HR@1 ↑					Ours	
	Task-specific CRSs						
	NBCRS [48]	RTA [11]	LaViC [14]	MSCRS [45]	PRIME [50]		
Appliances	0.0957±0.0101	0.2130±0.0090	0.2709±0.0014	0.2391±0.0063	0.2261±0.0223	0.2873±0.0152	
Arts & Crafts	0.0078±0.0091	0.0469±0.0048	0.1013±0.0031	0.0859±0.0059	0.1016±0.0036	0.1349±0.0270	
Automotive	0.0197±0.0090	0.0395±0.0051	0.0560±0.0253	0.0526±0.0053	0.0852±0.0128	0.1184±0.0139	
Baby Products	0.0567±0.0190	0.0090±0.0058	0.3373±0.0113	0.5254±0.0168	0.5791±0.0136	0.7194±0.0093	
Beauty	0.0326±0.0137	0.0618±0.0069	0.1187±0.0224	0.1135±0.0050	0.2317±0.0122	0.4009±0.0106	
Books	0.0275±0.0065	0.0678±0.0044	0.0758±0.0032	0.0742±0.0093	0.0994±0.0025	0.1081±0.0061	
CDs & Vinyl	0.0470±0.0091	0.0940±0.0096	0.0887±0.0031	0.1017±0.0025	0.0877±0.0045	0.1048±0.0129	
Cell Phones	0.0279±0.0144	0.0641±0.0060	0.1002±0.0077	0.0752±0.0048	0.1170±0.0051	0.1225±0.0036	
Digital Music	0.1149±0.0101	0.1897±0.0078	0.2643±0.0054	0.1782±0.0169	0.1842±0.0025	0.2674±0.0129	
Electronics	0.0449±0.0083	0.1870±0.0011	0.0075±0.0080	0.2843±0.0193	0.2868±0.0174	0.5885±0.0051	
Fashion	0.0130±0.0043	0.0281±0.0011	0.1232±0.0056	0.0982±0.0104	0.3276±0.0207	0.3517±0.0025	
Grocery	0.0070±0.0102	0.0314±0.0075	0.0940±0.0165	0.1185±0.0073	0.1209±0.0114	0.1376±0.0051	
Handmade	0.1429±0.0219	0.2091±0.0112	0.2264±0.0028	0.2976±0.0003	0.2509±0.0302	0.3175±0.0097	
Health	0.0278±0.0146	0.0389±0.0074	0.0777±0.0128	0.0722±0.0163	0.1160±0.0155	0.1389±0.0060	
Home	0.0252±0.0183	0.0579±0.0050	0.3197±0.0025	0.0605±0.0066	0.3575±0.0034	0.2002±0.0071	
Industrial	0.0150±0.0123	0.0551±0.0099	0.0902±0.0195	0.0501±0.0043	0.0982±0.0287	0.1441±0.0146	
Movies & TV	0.0480±0.0037	0.0655±0.0112	0.0786±0.0044	0.0655±0.0072	0.1059±0.0133	0.1376±0.0158	
Musical Instruments	0.1045±0.0118	0.0261±0.0064	0.0475±0.0229	0.3848±0.0096	0.4821±0.0259	0.6627±0.0076	
Office Prod.	0.0206±0.0065	0.0619±0.0053	0.0846±0.0042	0.0845±0.0048	0.0790±0.0067	0.0922±0.0011	
Patio & Garden	0.0086±0.0090	0.0345±0.0051	0.1206±0.0022	0.0431±0.0032	0.0862±0.0266	0.1609±0.0289	
Pet Supplies	0.0287±0.0231	0.0593±0.0096	0.0792±0.0246	0.0899±0.0079	0.0955±0.0117	0.1128±0.0058	
Software	0.1659±0.0084	0.2183±0.0078	0.3249±0.0320	0.3319±0.0031	0.3057±0.0098	0.4607±0.0244	
Sports	0.0280±0.0115	0.0592±0.0111	0.0903±0.0134	0.0841±0.0056	0.0709±0.0322	0.1106±0.0057	
Toys & Games	0.0104±0.0170	0.0417±0.0124	0.0833±0.0138	0.0625±0.0052	0.1301±0.0068	0.1422±0.0051	
Video Games	0.1055±0.0124	0.0700±0.0128	0.0851±0.0239	0.3817±0.0093	0.2089±0.0192	0.5240±0.0045	
Avg.	0.0490±0.0118	0.0812±0.0074	0.1338±0.0117	0.1582±0.0077	0.1934±0.0143	0.2618±0.0104	

Panel B: NDCG@5 and MRR.								
Domain	NDCG@5 ↑				MRR ↑			
	Task-specific CRSs			Ours	Task-specific CRSs			Ours
	NBCRS [48]	MSCRS [45]	PRIME [50]	LightCRS	NBCRS [48]	MSCRS [45]	PRIME [50]	LightCRS
Appliances	0.1614±0.0112	0.4365±0.0068	0.4437±0.0048	0.5160±0.0180	0.1861±0.0109	0.3925±0.0065	0.4202±0.0127	0.4781±0.0140
Arts & Crafts	0.0296±0.0051	0.2332±0.0045	0.2810±0.0054	0.3123±0.0103	0.1070±0.0070	0.2170±0.0043	0.2691±0.0269	0.3130±0.0161
Automotive	0.0544±0.0071	0.1812±0.0037	0.2842±0.0101	0.3070±0.0048	0.1187±0.0062	0.1832±0.0032	0.2726±0.0027	0.3068±0.0068
Baby Products	0.1065±0.0172	0.5908±0.0124	0.6910±0.0098	0.8339±0.0069	0.1515±0.0196	0.5887±0.0143	0.6713±0.0088	0.8099±0.0103
Beauty	0.0581±0.0104	0.2132±0.0046	0.3814±0.0135	0.6133±0.0067	0.1295±0.0118	0.2008±0.0047	0.3732±0.0141	0.5767±0.0083
Books	0.0682±0.0073	0.2381±0.0101	0.2817±0.0155	0.2984±0.0080	0.1250±0.0059	0.2221±0.0088	0.2817±0.0163	0.2990±0.0059
CDs & Vinyl	0.1398±0.0103	0.2730±0.0040	0.2700±0.0031	0.2892±0.0150	0.1443±0.0093	0.2584±0.0041	0.2691±0.0020	0.2861±0.0125
Cell Phones	0.0792±0.0141	0.2295±0.0038	0.3189±0.0043	0.3203±0.0044	0.1256±0.0077	0.2148±0.0023	0.3133±0.0078	0.3170±0.0032
Digital Music	0.1696±0.0099	0.3693±0.0139	0.4979±0.0137	0.5104±0.0093	0.2045±0.0077	0.3310±0.0116	0.4704±0.0013	0.4726±0.0091
Electronics	0.0718±0.0051	0.3894±0.0136	0.4346±0.0315	0.7475±0.0026	0.1424±0.0066	0.3781±0.0157	0.4145±0.0260	0.7129±0.0041
Fashion	0.0266±0.0020	0.1558±0.0076	0.4500±0.0045	0.5745±0.0021	0.1127±0.0023	0.1564±0.0053	0.4278±0.0073	0.5354±0.0018
Grocery	0.0290±0.0267	0.2482±0.0102	0.2926±0.0116	0.3237±0.0102	0.1067±0.0134	0.2395±0.0057	0.2958±0.0133	0.3206±0.0031
Handmade	0.2208±0.0147	0.4762±0.0158	0.5001±0.0131	0.5217±0.0010	0.2286±0.0158	0.4505±0.0180	0.4591±0.0101	0.4902±0.0063
Health	0.0455±0.0132	0.2061±0.0166	0.3392±0.0207	0.3653±0.0057	0.1255±0.0113	0.2010±0.0136	0.3194±0.0040	0.3302±0.0052
Home	0.0453±0.0136	0.1278±0.0040	0.5073±0.0076	0.4408±0.0044	0.1227±0.0137	0.1557±0.0044	0.4860±0.0086	0.4068±0.0058
Industrial	0.0456±0.0127	0.1704±0.0042	0.3087±0.0123	0.3358±0.0144	0.1145±0.0103	0.1666±0.0030	0.2935±0.0249	0.3311±0.0108
Movies & TV	0.0823±0.0030	0.2111±0.0055	0.3013±0.0020	0.3272±0.0120	0.1433±0.0029	0.2001±0.0047	0.3121±0.0028	0.3239±0.0098
Musical Instruments	0.1761±0.0122	0.5504±0.0080	0.6189±0.0038	0.7932±0.0081	0.1942±0.0092	0.5182±0.0071	0.6018±0.0059	0.7663±0.0061
Office Prod.	0.0498±0.0063	0.2369±0.0046	0.2648±0.0031	0.2760±0.0059	0.1187±0.0053	0.2284±0.0049	0.2623±0.0109	0.2792±0.0050
Patio & Garden	0.0172±0.0109	0.1736±0.0031	0.2875±0.0146	0.3429±0.0389	0.1079±0.0089	0.1702±0.0038	0.2933±0.0055	0.3448±0.0354
Pet Supplies	0.0582±0.0183	0.2232±0.0045	0.2813±0.0213	0.3234±0.0024	0.1259±0.0166	0.2159±0.0052	0.2903±0.0134	0.3145±0.0027
Software	0.3367±0.0075	0.5535±0.0021	0.5477±0.0160	0.6819±0.0147	0.2994±0.0059	0.5149±0.0029	0.5047±0.0121	0.6359±0.0131
Sports	0.0877±0.0128	0.2283±0.0053	0.2794±0.0148	0.3005±0.0064	0.1279±0.0072	0.2152±0.0035	0.2766±0.0171	0.2998±0.0023
Toys & Games	0.0311±0.0265	0.2179±0.0084	0.2978±0.0092	0.3228±0.0157	0.1252±0.0168	0.1971±0.0054	0.3013±0.0079	0.3195±0.0078
Video Games	0.1617±0.0210	0.5617±0.0188	0.3338±0.0264	0.6724±0.0051	0.1955±0.0092	0.5233±0.0084	0.3296±0.0151	0.6483±0.0031
Avg.	0.0941±0.0120	0.2998±0.0078	0.3798±0.0117	0.4540±0.0093	0.1473±0.0097	0.2856±0.0069	0.3680±0.0111	0.4367±0.0083

Table 7: Comparison of retrieval, generative baselines, and LightCRS on unseen domains across 25 Amazon product categories. Panel A reports HR@1; Panel B reports NDCG@5 and MRR for ranking-based methods. Bold and underline indicate the best and second-best results within each evaluation setting. Abbreviations: RoB. denotes RoBERTa_{large}; Vic. denotes Vicuna-v1.5; 1.5, 1.6, and 1.6+I denote LLaVA-v1.5, LLaVA-v1.6, and LLaVA-v1.6+image; M3 denotes MiniMax-M3; and M2.7 denotes MiniMax-M2.7.

Panel A: HR@1.											
Domain	HR@1 ↑										
	Retrieval				Generative					Ours	
	BM25 [39]	SBERT [37]	RoB. [27]	SimCSE [7]	Vic. [4]	1.5 [25]	1.6 [26]	1.6+I [26]	M2.7	M3	LightCRS
Appliances	0.0076	0.0684	0.0456	0.0000	0.0783±0.0024	0.0783±0.0059	0.0739±0.0133	0.1174±0.0105	0.1957±0.0033	0.2174 ±0.0014	0.0994±0.0208
Arts & Crafts	0.0000	0.0153	0.0230	0.0154	0.0547±0.0061	0.0546±0.0076	0.0551±0.0154	0.0560±0.0110	0.0703±0.0240	0.0880±0.0035	0.1198 ±0.0252
Automotive	0.0000	0.0128	0.0132	0.0134	0.0197±0.0008	0.0399±0.0126	0.0197±0.0168	0.0526±0.0100	0.0612±0.0151	0.0329±0.0211	0.1096 ±0.0312
Baby Products	0.0627	0.0657	0.0478	0.0209	0.1164±0.0128	0.1194±0.0162	0.1463±0.0156	0.0657±0.0042	0.1007±0.0220	0.1025±0.0174	0.1363 ±0.0125
Beauty	0.0169	0.0551	0.0640	0.0326	0.0533±0.0131	0.0476±0.0165	0.0770±0.0168	0.0584±0.0106	0.1148±0.0219	0.1237±0.0132	0.1511 ±0.0133
Books	0.0085	0.0486	0.0275	0.0042	0.0593±0.0002	0.0508±0.0195	0.0297±0.0006	0.0593±0.0043	0.0720±0.0174	0.0663±0.0302	0.1017 ±0.0041
CDs & Vinyl	0.0233	0.0129	0.0078	0.0000	0.0627±0.0153	0.0601±0.0113	0.0339±0.0057	0.0548±0.0115	0.0679±0.0074	0.0762±0.0113	0.1036 ±0.0123
Cell Phones	0.0055	0.0302	0.0330	0.0110	0.0362±0.0074	0.0306±0.0151	0.0306±0.0124	0.0362±0.0105	0.0864±0.0138	0.0931±0.0066	0.1105 ±0.0088
Digital Music	0.0185	0.1343	0.0417	0.0231	0.1264±0.0102	0.1149±0.0009	0.1322±0.0124	0.1437±0.0100	0.2299 ±0.0061	0.2241±0.0119	0.1021±0.0226
Electronics	0.0100	0.1371	0.1122	0.0175	0.0973±0.0028	0.1272±0.0111	0.1297±0.0166	0.1546±0.0039	0.0050±0.0081	0.0050±0.0044	0.1398 ±0.0100
Fashion	0.0140	0.0681	0.0631	0.0301	0.0481±0.0063	0.0441±0.0015	0.0827±0.0196	0.0459±0.0101	0.0979±0.0195	0.0952±0.0229	0.1421 ±0.0215
Grocery	0.0034	0.0137	0.0103	0.0103	0.0000±0.0000	0.0070±0.0106	0.0070±0.0105	0.0139±0.0186	0.0801±0.0054	0.0836±0.0158	0.1370 ±0.0092
Handmade	0.0167	0.0639	0.0500	0.0389	0.0732±0.0109	0.0557±0.0019	0.0592±0.0074	0.1359±0.0186	0.1295±0.0138	0.1117±0.0290	0.1686 ±0.0234
Health	0.0110	0.0495	0.0549	0.0165	0.0278±0.0039	0.0389±0.0125	0.0444±0.0167	0.0434±0.0007	0.0907±0.0187	0.0980±0.0103	0.1222 ±0.0118
Home	0.0479	0.2166	0.1814	0.0957	0.1184±0.0068	0.0932±0.0109	0.2030±0.0075	0.1089±0.0071	0.3224±0.0023	0.3325±0.0054	0.1371±0.0103
Industrial	0.0099	0.0471	0.0521	0.0074	0.0627±0.0173	0.0526±0.0147	0.0351±0.0049	0.0576±0.0191	0.0795±0.0157	0.0923±0.0036	0.1178 ±0.0199
Movies & TV	0.0306	0.0393	0.0087	0.0175	0.0349±0.0133	0.0306±0.0072	0.0437±0.0150	0.0439±0.0081	0.0905±0.0176	0.0956±0.0133	0.1339 ±0.0222
Musical Instruments	0.0166	0.1140	0.1116	0.0214	0.1140±0.0083	0.1283±0.0171	0.1140±0.0157	0.1568±0.0042	0.0214±0.0197	0.0285±0.0158	0.1473 ±0.0328
Office Prod.	0.0034	0.0172	0.0103	0.0172	0.0103±0.0128	0.0105±0.0181	0.0034±0.0058	0.0206±0.0127	0.0653±0.0162	0.0609±0.0237	0.0951 ±0.0107
Patio & Garden	0.0000	0.0259	0.0172	0.0086	0.0431±0.0168	0.0433±0.0027	0.0437±0.0043	0.0517±0.0134	0.0690±0.0347	0.1034±0.0291	0.1609 ±0.0284
Pet Supplies	0.0057	0.0209	0.0152	0.0076	0.0554±0.0038	0.0561±0.0011	0.0402±0.0042	0.0535±0.0032	0.0937±0.0133	0.0889±0.0176	0.1128 ±0.0052
Software	0.0288	0.1439	0.1187	0.0397	0.1572±0.0115	0.1703±0.0208	0.1965±0.0057	0.2227±0.0186	0.2358±0.0020	0.2882±0.0063	0.0946±0.0186
Sports	0.0000	0.0280	0.0312	0.0218	0.0187±0.0039	0.0156±0.0196	0.0187±0.0097	0.0218±0.0140	0.0561±0.0145	0.0717±0.0264	0.1194 ±0.0186
Toys & Games	0.0000	0.0104	0.0000	0.0052	0.0208±0.0148	0.0209±0.0069	0.0052±0.0053	0.0211±0.0037	0.0840±0.0043	0.0869±0.0024	0.1042 ±0.0146
Video Games	0.0151	0.0695	0.0658	0.0110	0.0945±0.0066	0.0815±0.0182	0.0825±0.0127	0.0850±0.0052	0.0430±0.0129	0.0470±0.0158	0.1222 ±0.0050
Avg.	0.0142	0.0603	0.0483	0.0195	0.0633±0.0083	0.0629±0.0112	0.0683±0.0108	0.0753±0.0098	0.1025±0.0140	0.1085±0.0143	0.1236 ±0.0165

Panel B: NDCG@5 and MRR.										
Domain	NDCG@5 ↑					MRR ↑				
	Retrieval				Ours	Retrieval				Ours
	BM25 [39]	SBERT [37]	RoB. [27]	SimCSE [7]	LightCRS	BM25 [39]	SBERT [37]	RoB. [27]	SimCSE [7]	LightCRS
Appliances	0.0145	<u>0.1351</u>	0.0863	0.0053	0.3020 ±0.0091	0.1799	0.2248	<u>0.2454</u>	0.1720	0.2924 ±0.0110
Arts & Crafts	0.0093	<u>0.0604</u>	0.0437	0.0328	0.2994 ±0.0267	0.1681	<u>0.2245</u>	0.2065	0.1969	0.3016 ±0.0226
Automotive	0.0066	<u>0.0416</u>	<u>0.0514</u>	0.0314	0.3295 ±0.0148	0.1733	0.2061	<u>0.2159</u>	0.1958	0.3086 ±0.0160
Baby Products	0.0995	<u>0.1302</u>	0.0978	0.0733	0.3720 ±0.0149	0.1564	<u>0.1591</u>	0.1430	0.1188	0.3472 ±0.0151
Beauty	0.1326	<u>0.1678</u>	<u>0.1696</u>	0.1588	0.3617 ±0.0145	0.1389	<u>0.1632</u>	0.1607	0.1584	0.3472 ±0.0152
Books	0.0184	<u>0.0917</u>	0.0553	0.0211	0.3007 ±0.0114	0.1724	0.2224	0.2099	0.1871	0.2984 ±0.0081
CDs & Vinyl	0.0359	<u>0.0368</u>	0.0286	0.0026	0.2971 ±0.0279	0.1987	<u>0.2013</u>	0.1940	0.1693	0.2954 ±0.0138
Cell Phones	0.0141	<u>0.0708</u>	0.0673	0.0307	0.2939 ±0.0075	0.1799	<u>0.2324</u>	0.2172	0.1812	0.2980 ±0.0076
Digital Music	0.0348	<u>0.1919</u>	0.0966	0.0339	0.2993 ±0.0202	0.1984	0.3362	0.2563	0.1967	0.2981 ±0.0197
Electronics	0.0225	<u>0.2367</u>	0.2025	0.0415	0.3584 ±0.0235	0.1090	<u>0.2234</u>	0.2010	0.1158	0.3328 ±0.0163
Fashion	0.1239	<u>0.1952</u>	0.1774	0.1573	0.3682 ±0.0235	0.1291	<u>0.1845</u>	0.1745	0.1535	0.3473 ±0.0207
Grocery	0.0143	<u>0.0342</u>	0.0187	0.0219	0.3194 ±0.0054	0.1648	0.1307	0.1302	0.1444	0.3184 ±0.0058
Handmade	0.0386	<u>0.1176</u>	<u>0.1323</u>	0.0728	0.3044 ±0.0207	0.2025	0.2736	0.2906	0.2330	0.2969 ±0.0157
Health	0.0357	<u>0.1196</u>	<u>0.1334</u>	0.0536	0.3095 ±0.0106	0.2005	0.2371	0.2186	0.2175	0.3061 ±0.0117
Home	0.2417	0.4017	<u>0.3864</u>	0.3147	0.3584±0.0168	0.2505	0.3700	<u>0.3696</u>	0.3137	0.3384±0.0094
Industrial	0.0320	0.1017	<u>0.1036</u>	0.0319	0.3088 ±0.0095	0.1831	<u>0.2260</u>	0.2110	0.1949	0.3100 ±0.0113
Movies & TV	0.0462	<u>0.0745</u>	0.0323	0.0312	0.3235 ±0.0419	0.1886	<u>0.2285</u>	0.1975	0.1950	0.3183 ±0.0253
Musical Instruments	0.0220	<u>0.1754</u>	<u>0.1841</u>	0.0401	0.3716 ±0.0268	0.1149	<u>0.2026</u>	0.2004	0.1193	0.3514 ±0.0268
Office Prod.	0.0183	<u>0.0472</u>	0.0430	0.0322	0.2955 ±0.0050	0.1824	<u>0.2085</u>	0.2006	0.1960	0.2948 ±0.0087
Patio & Garden	0.0100	<u>0.0647</u>	<u>0.0648</u>	0.0317	0.3249 ±0.0177	0.1767	<u>0.2271</u>	<u>0.2286</u>	0.1969	0.3305 ±0.0141
Pet Supplies	0.0177	<u>0.0668</u>	0.0570	0.0279	0.2953 ±0.0069	0.1758	<u>0.2300</u>	0.2159	0.1933	0.3000 ±0.0028
Software	0.0474	<u>0.2670</u>	0.2358	0.0966	0.2693 ±0.0073	0.2093	0.3993	<u>0.3574</u>	0.2567	0.2738±0.0124
Sports	0.0142	<u>0.0651</u>	0.0565	0.0330	0.3130 ±0.0199	0.1809	<u>0.2188</u>	0.2180	0.1960	0.3116 ±0.0153
Toys & Games	0.0104	<u>0.0462</u>	0.0225	0.0212	0.3087 ±0.0032	0.1629	<u>0.2111</u>	0.1892	0.1794	0.2986 ±0.0046
Video Games	0.0234	<u>0.1210</u>	0.1149	0.1099	0.3451 ±0.0060	0.1136	<u>0.1626</u>	0.1592	0.1099	0.3248 ±0.0021
Avg.	0.0434	<u>0.1224</u>	0.1065	0.0567	0.3212 ±0.0157	0.1724	<u>0.2282</u>	0.2164	0.1837	0.3136 ±0.0133

Table 8: Prompt used to identify product-recommendation conversations. Only conversations labeled yes by GPT-3.5-turbo are retained in MM-R2A-25.

You are given a Reddit conversation containing multiple numbered messages. Each message follows the format <message>(i) [content]</message>.

Determine whether the final message explicitly recommends one or more specific items in response to a recommendation request in an earlier message.

Answer yes if both conditions are satisfied, no otherwise, or uncertain if the conversation does not provide enough information. Return only yes, no, or uncertain, without any explanation.

Conversation: {conversation}

Table 9: Prompt used to generate visually grounded product descriptions.

You are given an Amazon product title and image. Generate a detailed, visually grounded description that helps users evaluate the product. Focus on its appearance, visible features, and other visually informative details. Do not repeat or mention the product title.

Product title: {title}

Product image: {image}

Table 10: Prompt for text-only product recommendation, used by generative baselines.

You are given a conversation between a user requesting product recommendations, marked by <submission>, and other users providing responses, marked by <comment>. You are also given candidate products formatted as ID: title.

Select the candidate that best matches the user’s needs. Return only its ID without any explanation.

Conversation: {conversation}

Candidates: {list of candidates}

Table 11: Prompt for image-aware product recommendation, used by generative baselines.

You are given a conversation between a user requesting product recommendations, marked by <submission>, and other users providing responses, marked by <comment>. You are also given candidate products, each formatted as ID: title followed by its image.

Select the candidate that best matches the user’s needs based on the conversation, title, and image. Return only its ID without any explanation.

Conversation: {conversation}

Candidates: {list of candidates}

Table 12: Per-stage offline preprocessing cost measurement. “Wall-clock” denotes the measured runtime on the 50-item sample after one warm-up pass; “Throughput” denotes the processing rate in items or encodings per second; “Peak GPU mem” denotes the maximum allocated GPU memory during the timed run; “Full MM-R2A-25” reports the extrapolated cost for 50,121 products and 219,268 product images.

#	Stage	View	Encoder	Wall-clock (50 items)	Throughput (items/s)	Peak GPU mem (GB)	Full MM-R2A-25 (GPU-hours)
1	Text encoding	title / features / description	GTE-base	0.47 s	319.3	0.29	0.13
2	Caption generation	image captions	LLaVA-v1.6 (full VLM)	487.18 s	0.10	15.03	593.46
3	Caption encoding	caption embeddings	GTE-base	0.18 s	277.3	0.28	0.22
4	Visual extraction	visual tokens	LLaVA-v1.6 vision tower	9.32 s	5.36	14.43	2.60

Table 13: Cached product-representation footprint. Each view is stored as a tensor with the indicated shape and FP32 precision. Per-product KB uses the catalog-wide average of 4.38 images per product for the caption row; full-catalog totals apply this same average to all products.

View	Tensor shape	Source	Per-product (KB, FP32)	Full MM-R2A-25 (GB, FP32)
Title embedding	(768,)	GTE-base	3.1	0.15
Features embedding	(768,)	GTE-base	3.1	0.15
Description embedding	(768,)	GTE-base	3.1	0.15
Caption embeddings	(N, 768)	GTE-base over LLaVA-v1.6 captions	13.5	0.68
Visual tokens	(5, 4096)	LLaVA-v1.6 vision tower	81.9	4.11