ReasonRec: A Reasoning-Augmented Multimodal Agent for Unified Recommendation

Yihua Zhang [1 2] Xi Liu [1] Xihuan Zeng [1] Mingfu Liang [1] Jiyan Yang [1] Rong Jin [1] Wen-Yen Chen [1] Yiping Han [1] Hao Ma [1] Bo Long [1] Huayu Li [1] Buyun Zhang [1] Liang Luo [1] Sijia Liu [2] Tianlong Chen [3]

Abstract

Recent advances in multimodal recommenders excel at feature fusion but remain opaque and inefficient decision-makers, lacking explicit reasoning and self-awareness of uncertainty. To address this, we introduce ReasonRec, a reasoning-augmented multimodal agent structured around a three-stage explicit reasoning pipeline: *Observe*, via a pretrained Vision-Language Model (VLM) encoder; *Deliberate*, by formulating recommendation as chain-of-thought (CoT) reasoning tasks and explicitly quantifying prediction uncertainty through an evidence-horizon-aware curriculum; and *Act*, through dynamic delegation of uncertain or challenging queries to lightweight classical recommendation models. Specifically, we propose a reasoning-aware visual instruction tuning strategy that systematically transforms diverse recommendation tasks into unified CoT prompts, enabling the VLM to explicitly articulate intermediate decision steps. Additionally, our evidence-horizon curriculum progressively enhances the reasoning complexity to better handle cold-start and long-tail user scenarios, significantly boosting model generalization. Furthermore, the uncertainty-guided delegation mechanism empowers the agent to assess its own confidence, strategically allocating computational resources to optimize both recommendation accuracy and inference efficiency. Comprehensive experiments on four standard recommendation tasks (sequential recommendation, direct recommendation, CTR prediction, and explanation generation) across five real-world datasets demonstrate that ReasonRec achieves over 30% relative improvement in key ranking metrics (*e.g.*, HR@5, NDCG@5) compared to state-of-the-art multimodal recommenders. Crucially, ReasonRec substantially reduces inference latency by dynamically delegating up to 35% of queries to efficient sub-models without compromising accuracy. Extensive ablation studies further confirm that each proposed reasoning and planning mechanism individually contributes substantially to ReasonRec's overall effectiveness. Collectively, our results illustrate a clear pathway towards interpretable, adaptive, and efficient multimodal recommendation through explicit reasoning and agentic design.

1. Introduction

Recent advances in large language models (LLMs), exemplified by DeepSeek-R1 (Guo et al., 2025), GPT-4o (Hurst et al., 2024), and Gemini (Team et al., 2023a), have sparked a widespread "reasoning renaissance" in artificial intelligence research. These models leverage explicit chain-of-thought (CoT) (Wei et al., 2022) reasoning to achieve remarkable performance across various complex reasoning tasks, including logical inference (Li et al., 2024), mathematical problem-solving (MAA, 2024), and planning (Xie et al., 2024b). Such reasoning capabilities not only improve model interpretability but also significantly enhance generalization and trustworthiness. Despite these successes, the power of explicit reasoning (Guo et al., 2025) has yet to be systematically explored and exploited in multimodal recommendation (Xie et al., 2024a), a crucial domain inherently requiring nuanced and transparent decision-making.

Multimodal recommendation (Cheng et al., 2023; Geng et al., 2023) fundamentally involves intricate multi-step deliberation processes. To effectively fulfill user needs, a recommender must first accurately infer user intent from sparse and ambiguous interactions (Covington et al., 2016; Cheng et al., 2016; Wang et al., 2021a). Next, it must carefully evaluate the semantic content embedded in diverse multimodal signals, such as visual imagery and descriptive text, to estimate relevance. Finally, recommenders must explicitly reason about complex utility trade-offs, balancing user satisfaction, content freshness, and system efficiency (Kang

[*]Equal contribution [1]Meta AI [2]Michigan State University [3]UNC. Correspondence to: Yihua Zhang <yihuazhang@meta.com>.

ICML 2025 Workshop on Programmatic Representations for Agent Learning, Vancouver, Canada. Copyright 2025 by the author(s).

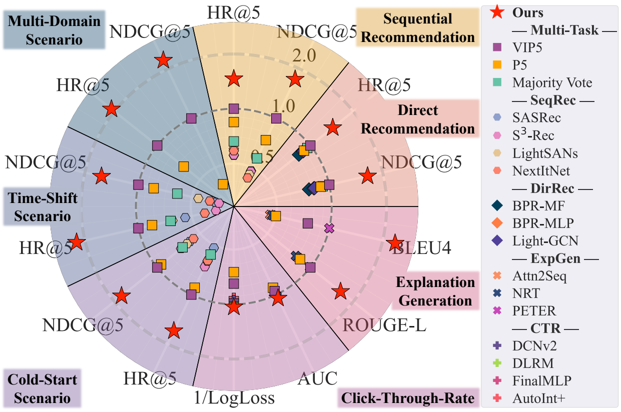


Figure 1. Performance comparison of ReasonRec against baselines across four basic recommendation tasks and three challenging scenarios. Each task (color-coded region) is evaluated on two metrics. All results are normalized such that the best baseline performance is set to **1.0** for each metric-task pair. Markers representing ReasonRec indicate the relative improvement ratio over the strongest baseline, with values greater than demonstrating superiority. Experiment setups and results can be found in Sec. 4.

& McAuley, 2018; Zhou et al., 2020; Fan et al., 2021; Yuan et al., 2019). However, existing multimodal recommendation systems typically rely on a singular, black-box forward inference, collapsing all these intricate decisions into an opaque scoring mechanism (Cheng et al., 2023; Geng et al., 2023). This implicit approach severely limits transparency, interpretability, and robustness, especially in challenging scenarios like cold-start and cross-domain tasks (Zhang et al., 2024; Cheng et al., 2020; Song et al., 2019; Naumov et al., 2019; Wang et al., 2021b).

Most contemporary Vision-Language Models (VLMs) (Zong et al., 2024; Liu et al., 2023b; 2024b; Zhu et al., 2023; Ye et al., 2023; Wang et al., 2023; Li et al., 2023a; Alayrac et al., 2022; Awadalla et al., 2023; Gao et al., 2023) in recommendation embed multimodal signals implicitly within their transformer architectures. While effective at feature fusion, these models fail to explicitly articulate their reasoning processes (Zhang et al., 2024), resulting in recommendations that are hard to interpret and audit. More critically, these single-pass architectures lack the self-awareness to diagnose their own uncertainty and thus cannot dynamically adapt by invoking specialized external knowledge or computationally efficient sub-models. Consequently, existing multimodal recommenders suffer from poor reliability in low-confidence situations and waste computational resources by uniformly treating all recommendation decisions as equally complex.

Motivated by these limitations, we propose to fundamentally rethink the design of multimodal recommenders around an explicit reasoning paradigm. Specifically, we ask: Can we build an agentic VLM that reasons explicitly through chain-of-thought, transparently assesses its own uncertainty, and dynamically delegates uncertain or challenging cases to external, lightweight specialists? Addressing this question not only promises to significantly enhance the interpretability and generalization capabilities of multimodal recommenders but also introduces a novel, reasoning-driven agent framework into the broader recommendation landscape. In this work, we introduce ReasonRec, a reasoning-augmented multimodal recommendation agent that transforms recommendation into a transparent and adaptive decision process. Our key contributions include:

- A reasoning-aware instruction tuning framework that reformulates diverse recommendation tasks, including sequential recommendation, direct recommendation, CTR prediction, and explanation generation, into a unified chain-of-thought (CoT) format, enabling the VLM to verbalize intermediate reasoning steps and improve task alignment.

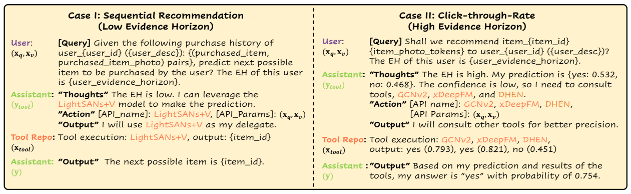
- An evidence-horizon curriculum learning strategy that gradually expands the complexity of reasoning chains by controlling user-item sparsity levels during training, which enhances generalization in cold-start and long-tail scenarios.

- An uncertainty-guided tool delegation mechanism that equips the agent with the ability to assess its own prediction confidence and dynamically invoke lightweight classical models when needed, balancing computational cost and predictive robustness.

- Extensive empirical validation across five public benchmarks and four recommendation tasks, demonstrating that ReasonRec achieves over 30% improvement in HR@5 and NDCG@5 compared to prior state-of-the-art models; adapts effectively to cold-start, time-shift, and multi-domain scenarios; and reduces inference cost by more than 30% through selective delegation.

2. Related Work

Generative models for recommendation systems. Recent advances in recommendation systems have witnessed a paradigm shift toward generative architectures and multimodal content understanding. Building upon the foundation of large language models (LLMs) (Li et al., 2023c), pioneering works like Transformers4Rec (de Souza Pereira Moreira et al., 2021) and BERT4Rec (Sun et al., 2019) employ Transformer architectures for sequential modeling, establishing frameworks for transferable recommendation through language model pretraining. Subsequent innovations extend this paradigm through diverse representation learning strategies: UniSRec (Hou et al., 2022) constructs item embeddings from descriptive texts rather than static IDs, while TransRec (Wang et al., 2022) integrates multimodal user feedback through BERT and ResNet encoders for content-based personalization. The emergence of prompt engineering has further enriched this landscape, with PETER (Li



Case I: Sequential Recommendation (Low Evidence Horizon)

User: $(\mathbf{x}_q, \mathbf{x}_v)$ **[Query]** Given the following purchase history of user_{user_id} ({user_desc}): {(purchased_item, purchased_item_photo) pairs}, predict next possible item to be purchased by the user? The EH of this user is {user_evidence_horizon}.

Assistant: $(\mathbf{y}_{tool})$ **"Thoughts"** The EH is low. I can leverage the LightSANs+V model to make the prediction.
"Action" [API_name]: LightSANs+V, [API_Params]: $(\mathbf{x}_q, \mathbf{x}_v)$
"Output" I will use LightSANs+V as my delegate.

Tool Repo: $(\mathbf{x}_{tool})$ Tool execution: LightSANs+V, output: {item_id}

Assistant: $(\mathbf{y})$ **"Output"** The next possible item is {item_id}.

Case II: Click-through-Rate (High Evidence Horizon)

User: $(\mathbf{x}_q, \mathbf{x}_v)$ **[Query]** Shall we recommend item_{item_id} {item_photo_tokens} to user_{user_id} ({user_desc})? The EH of this user is {user_evidence_horizon}.

Assistant: $(\mathbf{y}_{tool})$ **"Thoughts"** The EH is high. My prediction is {yes: 0.532, no: 0.468}. The confidence is low, so I need to consult tools, GCNv2, xDeepFM, and DHEN.
"Action" [API name]: GCNv2, xDeepFM, DHEN, [API Params]: $(\mathbf{x}_q, \mathbf{x}_v)$
"Output" I will consult other tools for better precision.

Tool Repo: $(\mathbf{x}_{tool})$ Tool execution: GCNv2, xDeepFM, DHEN, output: yes (0.793), yes (0.821), no (0.451)

Assistant : $(\mathbf{y})$ **"Output"** Based on my prediction and results of the tools, my answer is "yes" with probability of 0.754.

Figure 2. Examples of *reasoning-aware instruction tuning*. **(I)** Sequential recommendation for a user with *low evidence horizon (EH)*, where the agent confidently *delegates* the query to LightSANs+V for efficiency. **(II)** CTR prediction for a user with *high EH* and low model confidence; the planner therefore *consults* multiple lightweight tools (GCNv2, xDeepFM, DHEN) before producing the final answer. More examples are provided in Appx. A.

et al., 2021) and PEPLER (Li et al., 2022) developing continuous prompt templates to encode user-item interactions while generating textual rationales for recommendations. Architectural unification efforts like M6-Rec (Cui et al., 2022) convert behavioral patterns into text sequences for Transformer processing, enabling task-adaptive fine-tuning through customized loss functions. P5 (Geng et al., 2022b) and OpenP5 (Xu et al., 2023) achieve cross-task generalization by implementing instruction-tuned LLMs that represent user-item relationships through natural language interfaces, later extended by P5-ID (Hua et al., 2023) through novel item indexing schemes combining sequential, collaborative, and semantic signals. Multimodal generation techniques have simultaneously evolved across three directions: auxiliary feature integration, explainable recommendation, and semantic structure discovery. Early approaches like VBPR (He & McAuley, 2016) and PiNet (Meng et al., 2020) enhance collaborative filtering through visual feature extraction and heterogeneous preference modeling, respectively, while JRL (Zhang et al., 2017) pioneers joint multimodal representation learning. Domain-specific generators have emerged for fashion (Hou et al., 2019; Verma et al., 2020; Chen et al., 2019), travel (Geng et al., 2022a), and culinary recommendations (Meng et al., 2020), producing visually-grounded explanations. Cutting-edge methods further decode latent multimodal semantics through cross-modal alignment (Zhang et al., 2021b; Geng et al., 2023; Cheng et al., 2023), contrastive pattern mining (Zhang et al., 2021a), and adversarial content synthesis (Deldjoo et al., 2022), establishing new benchmarks for semantic-aware recommendation generation.

Vision-language models and agents. The evolution of Large Language Models has catalyzed next-generation vision-language architectures, transcending traditional visual-language systems through LLM-powered linguistic reasoning. Pioneered by architectures such as LLaVA-series (Liu et al., 2023b;a; Sun et al., 2023; Liu et al., 2023c), BLIP-family (Li et al., 2023b; Dai et al., 2023), and MiniGPT-4 (Zhu et al., 2023), these models demonstrate exceptional visual dialog capabilities via LLM-based language encoders. However, their computational footprint—typically requiring 7B-65B parameters—creates deployment bottlenecks for edge/mobile platforms demanding real-time responsiveness. While proprietary models like Gemini (Team et al., 2023b) address this via scaled variants (*e.g.*, 1.8B-parameter Nano for smartphones), their closed-source nature limits adaptability. Open-source initiatives like MobileVLM (Chu et al., 2023) develop compact architectures (e.g., 2.7B-parameter mobileLLaMA) to bridge this gap. In this work, we for the first time exploit a visual instruction tuning framework for recommendation system based on the pretrained VLMs.

3. ReasonRec: A Reasoning-Augmented Recommendation Agent

To enable explicit reasoning and adaptive decision-making in multimodal recommendation, we propose **ReasonRec**, a unified agentic framework structured as a three-stage reasoning pipeline: *Observe* → *Deliberate* → *Act*. In this section, we detail each of these stages and describe how their tight integration facilitates interpretability and robustness in recommendation scenarios.

3.1. Observer: Visual and Textual Perception

The first stage, the *Observer*, extracts multimodal information critical for informed reasoning. Specifically, given user history representations $\mathcal{H}_u$ (past interactions) and candidate item information $\mathcal{I}$ (visual image $\mathbf{x}_v$, query text $\mathbf{x}_q$ and metadata), the Observer employs a pretrained Vision-Language Model (VLM) encoder (Liu et al., 2023b;a) to generate a unified embedding $\mathbf{h} = \text{VLMEncoder}(\mathbf{x}_v, \mathbf{x}_q, \mathcal{H}_u)$. These embeddings provide a rich representation capturing user intent and item semantics, thus forming a robust foundation for explicit reasoning.

3.2. Deliberator: Explicit Reasoning and Self-Reflection

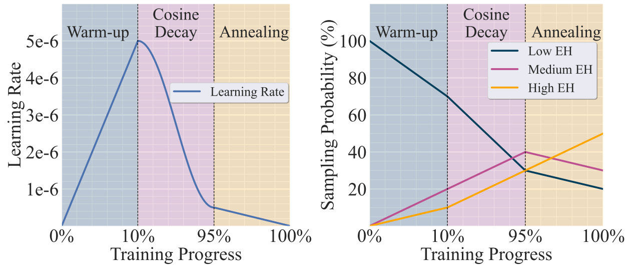
The *Deliberator* explicitly performs reasoning by reformulating recommendation tasks into structured instruction-following problems. Traditional VLMs lack structured interaction modeling and task alignment, thus failing in sparse interaction scenarios. To mitigate these issues, we introduce **Reasoning-Aware Visual Instruction Tuning (R-VIT)**, comprising three critical innovations:

(1) Task formulation as VQA. We convert recommendation tasks into structured vision-question-answering (VQA) instructions. Formally, given multimodal inputs, the Deliberator generates outputs in a structured prompt-response format:

$$\begin{aligned} &\texttt{User}: \mathbf{x}_v \ <\backslash n> \ \mathbf{x}_q \texttt{<STOP>} \qquad (1) \\ &\texttt{Assistant}: [\text{Thought Tokens}] \rightarrow \mathbf{y} \texttt{<STOP>}, \end{aligned}$$

where '[Thought Tokens]' explicitly verbalize intermediate reasoning steps (**Fig. 2**). Training optimizes an autoregressive objective to ensure reasoning consistency.

(2) Template mixtures for generalization. A key innova-



(A) Learning rate schedule. (B) EH sampling.

Figure 3. Evidence-horizon curriculum learning. Users are grouped into *Low*, *Medium*, and *High EH* based on interaction sparsity. The learning-rate schedule (A) is aligned with the sampling policy (B): training begins with dense users, gradually shifts to harder cases, and finally focuses on high-EH (cold-start) users. This staged strategy enhances ReasonRec's reasoning under sparsity while mitigating forgetting on warm-user patterns.

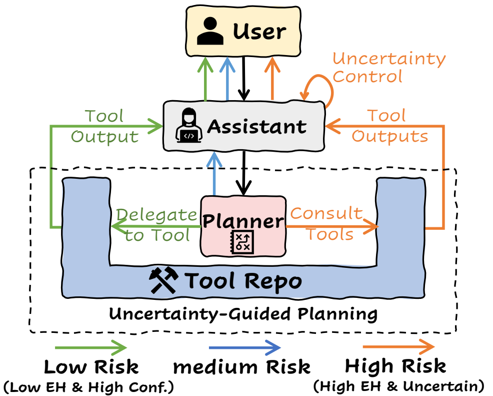


Figure 4. Overview of ReasonRec's inference pipeline with *uncertainty-guided planning*. The planner jointly considers a user's evidence horizon (EH) and confidence to assign each query to one of three risk levels. *Low-risk* queries (low EH, high confidence) are *delegated* to lightweight models from the *tool repository*; *medium-risk* queries are handled directly by the VLM; *high-risk* queries (high EH, low confidence) *consult tools* to refine the decision.

tion of our approach is the introduction of *multiple diverse instruction templates per task* to prevent overfitting and improve the model's ability to generalize to unseen variations. Instead of a single fixed format, we construct multiple templates, each expressing the same recommendation problem in different linguistic styles and structures. This encourages the model to focus on *task semantics rather than surface patterns*, improving robustness against distribution shifts. As shown in Fig. 5, this mixture-of-templates approach significantly enhances training effectiveness.

(3) Evidence horizon quantification. A major challenge in recommender systems is handling data sparsity, particularly in cold-start scenarios where users or items lack sufficient historical interactions. Traditional approaches struggle under these conditions, as models often rely heavily on frequent patterns while failing to capture long-tail behaviors. To address this issue, we introduce an **evidence-horizon-aware curriculum learning** strategy, gradually adapting the VLM to varying levels of data sparsity. This involves (i) defining a formal evidence horizon metric to characterize user-item sparsity and (ii) progressively adjusting the difficulty of training samples over time. To measure data sparsity, we define an *evidence horizon score* $C(u)$ for a user u based on their historical interactions $C(u) = 1 - \frac{|\mathcal{I}_u|}{\max_{u'} |\mathcal{I}_{u'}|}$, where $|\mathcal{I}_u|$ denotes the interaction count of user u, and $\max_{u'} |\mathcal{I}_{u'}|$ is the maximum interaction count among all users. A higher evidence horizon score indicates fewer interactions, representing greater recommendation difficulty. During training, we explicitly incorporate this evidence horizon score into instruction templates, guiding the VLM to adjust its reasoning complexity based on the query difficulty. As shown later, this explicit evidence horizon modeling significantly enhances the model's uncertainty management and adaptive tool utilization.

(4) Evidence-horizon-aware curriculum learning. Rather than uniformly sampling all training data, we progressively increase the difficulty of training samples, inspired by data-mixing techniques in LLM pretraining (Dubey et al., 2024). Specifically, the training consists of three distinct phases (**Fig. 3**): ① Warm-up Phase: Initially, the model learns from users with abundant historical interactions (low evidence horizon), capturing strong user-item correlations; ② Progressive Learning Phase: Gradually introduces medium evidence horizon users and items, improving generalization to sparser distributions; ③ Cold-start Emphasis Phase: Intensively trains on high EH scenarios with reduced learning rates (Dubey et al., 2024; Liu et al., 2024a), enhancing robustness against challenging, sparse data conditions.

3.3. Actuator: Uncertainty-Guided Tool Delegation

The *Actuator* dynamically decides whether to delegate the reasoning task to lightweight classical models (tools) or rely directly on the VLM, guided by both uncertainty and evidence horizon. Below we detail the mechanisms enabling adaptive, efficient decision-making:

Uncertainty-guided planning with classical models. Deploying large-scale VLMs for recommendation tasks requires balancing computational efficiency with predictive accuracy. While VLMs exhibit powerful reasoning, their inference costs are substantially higher than classical recommendation methods. To resolve this, we propose a **risk-aware planning mechanism**, dynamically choosing between VLM and lightweight classical models based on evidence horizon and model confidence.

Specifically, as illustrated in **Fig. 4**, we categorize queries according to their risk level: ① **Low-Risk**: Users with low evidence horizon and high model confidence—delegated directly to classical recommendation models for efficient inference. ② **Medium-Risk**: Moderate uncertainty queries handled directly by the VLM, exploiting internal reasoning capacities. ③ **High-Risk**: Users with high evidence horizon and significant uncertainty, invoking multiple classical models whose aggregated outputs refine subsequent VLM-based reasoning. This hierarchical approach balances computational cost and accuracy, leveraging the VLM's reasoning prowess specifically for complex cases.

Skill repository. To ensure computational efficiency at scale, we maintain a **skill repository** comprising classical models specialized for distinct recommendation subtasks. This includes matrix factorization, graph-based models, and two-tower architectures (e.g., xDeepFM) optimized for efficient CTR prediction. These specialized tools directly handle low-risk cases, significantly reducing VLM inference overhead without compromising accuracy.

Uncertainty-aware inference and tool integration. In high-risk scenarios, the Deliberator initially predicts with explicit uncertainty (e.g., "Recommend Item Y. Confidence: 0.53"). Next, classical tools from the repository refine this initial prediction, improving final accuracy. Thus, our uncertainty-guided planning mechanism effectively trades off efficiency and accuracy, ensuring robust, high-quality recommendations across diverse data conditions.

In summary, the integrated reasoning pipeline (*Observe* $\rightarrow$ *Deliberate* $\rightarrow$ *Act*), combined with explicit uncertainty estimation, adaptive tool delegation, and evidence-horizon-aware curriculum learning, empowers ReasonRec with interpretability, robustness, and efficiency, particularly under challenging multimodal recommendation scenarios.

4. Experiments

In this section, we provide a comprehensive evaluation on the proposed ReasonRec to a diverse range of baselines in four commonly-used recommendation tasks. Besides, we also consider three challenging recommendation settings, including *cold-start*, *time-shift*, and *multi-domain* scenarios, each representing a long-lasting challenge in recommender systems. We also provide abundant ablation studies to demonstrate the effectiveness of the planning strategy.

4.1. Experiment Setups

Model and datasets. We evaluate ReasonRec across four key recommendation tasks: sequential recommendation, direct recommendation, explanation generation, and CTR prediction, using LLaVA1.5-7B (Liu et al., 2023a) as the underlying VLM. **Amazon Review Dataset:** We use four categories from the Amazon Review dataset: *Clothing, Shoes & Jewelry*, *Sports & Outdoors*, *Beauty*, and *Toys & Games*. Each includes user purchase histories, item metadata, textual reviews, and images (see Tab. A1). **Pixel-1M Dataset (Cheng et al., 2023):** This large-scale image-centric dataset contains over 1M users, 100K images, and 20M user–image interactions. Unlike ID-based datasets, it enables learning directly from raw image pixels. We adopt a leave-one-out split: the last interaction per user for testing, the second-to-last for validation, and the rest for training.

Baselines compared in each task. We compare ReasonRec against state-of-the-art baselines for each task: • **Sequential Recommendation (Tab. 1):** We include four classical sequential models—SASRec (Kang & McAuley, 2018), S^3-Rec (Zhou et al., 2020), LightSANs (Fan et al., 2021), and NextItNet (Yuan et al., 2019)—and three generative baselines: P5 (Geng et al., 2022b), VIP5 (Geng et al., 2023), and UniMP (Wei et al., 2024). Following PixelRec (Cheng et al., 2023), we replace item ID embeddings in classical models with visual features to ensure a fair multimodal comparison. • **Direct Recommendation (Tab. 1):** Baselines include classical models BPR-MF (Rendle et al., 2012), BPR-MLP (Rendle et al., 2012), LightGCN (He et al., 2020), and the same generative baselines used above. • **Explanation Generation (Tab. 3):** We follow VIP5 (Geng et al., 2023) and compare with Attn2Seq (Dong et al., 2017), NRT (Dong et al., 2017), PETER (Li et al., 2021), as well as P5 and VIP5. • **CTR Prediction (Tab. 2):** Following (Zhang et al., 2024), we adopt AFN+ (Cheng et al., 2020), AutoInt+ (Song et al., 2019), DLRM (Naumov et al., 2019), DCNv2 (Wang et al., 2021b), FinalMLP (Mao et al., 2023), MaskNet (Wang et al., 2021c), and xDeepFM (Lian et al., 2018), implemented via the BARS evaluation framework (Zhu et al., 2022a; 2021). In sequential, direct, and CTR tasks, we report a "majority vote" baseline that averages predictions from all classical models. This helps isolate the effect of our planner and clarify that the performance of ReasonRec is not simply due to tool usage.

Training and evaluation setups. Non-generative models are trained separately for each task and dataset. Generative models are trained per dataset using mixed-task instruction tuning. For the multi-domain challenge (across Sports, Beauty, Clothing, and Toys), P5, VIP5, UniMP and ReasonRec are trained as unified models across domains. Additional training details are provided in Appx. A. For sequential and direct recommendation, we report **HR@5** and **NDCG@5**. For explanation generation, we use **BLEU4** and **ROUGEL**. For CTR prediction, following standard practice (Blondel et al., 2016; Song et al., 2019; Wang et al., 2021b; Zhu et al., 2022b; Mao et al., 2023), we use AUC and LogLoss, where higher AUC and lower LogLoss indicate better performance.

4.2. Experiment Results

A holistic comparison on sequential and direct recommendation tasks. Tab. 1 summarizes our results, demonstrating state-of-the-art performance across both recommendation paradigms. In the sequential recommendation (top portion of Tab. 1), *first*, ReasonRec achieves superior

Table 1. Performance comparison of different methods in sequential and direct recommendation task using Amazon Review dataset and Pixel-1M dataset.

Methods	Sports		Beauty		Clothing		Toys		Pixel-1M	
	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5
Sequential Recommendation										
SASRec	0.0289	0.0175	0.0403	0.0297	0.0132	0.0126	0.0463	0.0338	0.0116	0.0107
S^3-Rec	0.0274	0.0189	0.0415	0.0286	0.0110	0.0105	0.0443	0.0344	0.0168	0.0111
LightSANs	0.0260	0.0170	0.0435	0.0250	0.0185	0.0109	0.0481	0.0354	0.0165	0.0108
NextItNet	0.0258	0.0197	0.0427	0.0248	0.0192	0.0072	0.0470	0.0327	0.0140	0.0099
Majority Vote	0.0313	0.0211	0.0492	0.0315	0.0214	0.0199	0.0512	0.0382	0.0144	0.0135
P5	0.0275	0.0176	0.0483	0.0398	0.0499	0.0392	0.0694	0.0523	0.0188	0.0115
VIP5	0.0436	0.0371	0.0565	0.0489	0.0623	0.0597	0.0712	0.0596	0.0197	0.0123
UniMP	0.0515	0.0419	0.0602	0.0531	0.0679	0.0632	0.0794	0.0647	0.0256	0.0170
ReasonRec (Ours)	**0.0721**	**0.0694**	**0.0797**	**0.0944**	**0.0832**	**0.1011**	**0.1032**	**0.0901**	**0.0315**	**0.0218**
Direct Recommendation										
BPR-MF	0.1478	0.0897	0.1426	0.0913	0.1280	0.0735	0.1023	0.0641	0.0356	0.0231
BPR-MLP	0.1592	0.0945	0.1381	0.0891	0.1421	0.0822	0.1171	0.0721	0.0384	0.0253
Light-GCN	0.1549	0.0911	0.1501	0.0904	0.1475	0.0831	0.1121	0.0744	0.0362	0.0281
Majority Vote	0.1614	0.0993	0.1612	0.1043	0.1514	0.0931	0.1229	0.0813	0.0403	0.0299
P5	0.1583	0.1132	0.1681	0.1123	0.1001	0.0639	0.1232	0.0841	0.0539	0.0243
VIP5	0.1791	0.1241	0.1739	0.1113	0.1299	0.0871	0.1245	0.0829	0.0624	0.0392
UniMP	0.1940	0.1372	0.1825	0.1220	0.1378	0.0965	0.1302	0.0887	0.0703	0.0451
ReasonRec	**0.2439**	**0.1732**	**0.2351**	**0.1655**	**0.1998**	**0.1523**	**0.1839**	**0.1671**	**0.0991**	**0.0725**

Table 2. Performance comparison of different methods in click-through-rate task using Amazon Review and Pixel-1M datasets.

Methods	Sports		Beauty		Clothing		Toys		Pixel-1M	
	AUC (↑)	**LogLoss (↓)**	**AUC (↑)**	**LogLoss (↓)**	**AUC (↑)**	**LogLoss (↓)**	**AUC (↑)**	**LogLoss (↓)**	**AUC (↑)**	**LogLoss (↓)**
AutoInt+	0.8033	0.2432	0.8732	0.2031	0.7224	0.3296	0.7533	0.3115	0.5123	0.4993
DLRM	0.8145	0.2533	0.8819	0.1672	0.7174	0.3442	0.7542	0.3411	0.5472	0.4173
FinalMLP	0.7984	0.2411	0.8801	0.1993	0.7253	0.3213	0.7635	0.3021	0.5524	0.3984
DCNv2	0.8024	0.2395	0.8825	0.1742	0.7297	0.3459	0.7513	0.3571	0.5323	0.4242
P5	0.6532	0.4473	0.7744	0.4336	0.6743	0.4544	0.6931	0.4946	0.5832	0.3672
VIP5	0.6812	0.3985	0.8415	0.3573	0.7025	0.4135	0.7113	0.3846	0.6031	0.3449
ReasonRec	**0.8429**	**0.2445**	**0.9113**	**0.1993**	**0.7443**	**0.3139**	**0.7815**	**0.3329**	**0.6311**	**0.3222**

HR@5 and NDCG@5 across all datasets, surpassing individual baselines and the "majority vote" ensemble. This confirms that the gains stem from our multimodal instruction tuning and planning, rather than simple model combination. *Second*, classical methods (SASRec, S^3-Rec, LightSANs, NextItNet) and their ensemble provide moderate but limited improvements, lagging significantly behind multimodal generative models. *Third*, multimodal approaches (P5, VIP5) outperform classical models but show diminished advantages on the large Pixel-1M dataset. In contrast, ReasonRec consistently achieves strong performance and effectively captures sequential preferences, even in large-scale settings. In the direct recommendation scenario (bottom portion of Tab. 1), focusing on static user–item interactions, *first*, ReasonRec again consistently leads in all metrics and datasets, particularly on Pixel-1M, indicating robustness to diverse user feedback. *Second*, classical methods (BPR-MF, BPR-MLP, Light-GCN) and their ensemble achieve moderate improvements but remain inferior to multimodal methods. *Third*, while P5 and VIP5 show promising results on smaller Amazon datasets, their effectiveness weakens on Pixel-1M, highlighting challenges in modeling static, high-dimensional user preferences. ReasonRec effectively addresses these challenges via integrated reasoning.

Performance on the explanation generation task. In Tab. 3, we evaluate textual explanation quality using BLEU4 and ROUGEL. *First*, ReasonRec achieves substantially higher scores across all domains, highlighting its superior capability in generating detailed and accurate explanations. *Second*, multimodal baselines (P5, VIP5) outperform conventional methods (Attn2Seq, NRT, PETER) but remain behind our unified approach, confirming the benefit of combining visual representations and instruction tuning into a single VLM agent. *Third*, our planning-and-tool mechanism effectively captures detailed item attributes and user reasoning, enabling more coherent and informative explanations aligned with user preferences and product specifics.

Performance on the CTR task. In Tab. 2, ReasonRec achieves the highest AUC and lowest LogLoss across all datasets, outperforming both classical and multimodal baselines. Classical CTR models (AutoInt+, DLRM, FinalMLP, DCNv2) show reasonable accuracy but lack precision in fine-grained click probability estimation. Generative methods (P5, VIP5), though effective in recommendation and explanation, struggle to distinguish subtle click signals, highlighting the challenge of adapting language models for binary prediction. ReasonRec excels by combining multimodal reasoning with robust CTR modeling, demonstrating that explicit planning enhances probability estimates.

Table 3. Performance comparison on explanation generation using BLUE4 and ROUGEL metrics.

Methods	Sports		Beauty		Clothing		Toys	
	BLUE4	ROUGEL	BLUE4	ROUGEL	BLUE4	ROUGEL	BLUE4	ROUGEL
Attn2Seq	0.5478	9.1825	0.8014	9.7992	0.6447	9.0835	1.6419	10.7834
NRT	0.4903	7.6935	0.8438	9.9785	0.4708	8.2952	1.9267	11.2239
PETER	0.7123	11.3721	1.2172	9.4628	2.1132	14.0031	3.7822	11.8632
P5	0.6348	9.0524	1.0389	10.9447	0.7682	9.6325	1.4698	10.1814
VIP5	1.0774	11.1325	1.2983	12.9471	1.2052	10.8926	2.3421	12.0865
REASONREC	**3.4339**	**18.4632**	**4.3683**	**17.4422**	**4.9357**	**19.9311**	**5.6332**	**20.8345**

Table 4. Performance on sequential recommendation in cold-start scenarios on Pixel-1M dataset.

Method	Metric	Cold-start Level		
		Normal	Medium	Coldest
SASRec	HR@5	0.0116	0.0079	0.0058
	NDCG@5	0.0107	0.0043	0.0026
LightSANs	HR@5	0.0165	0.0087	0.0053
	NDCG@5	0.0108	0.0045	0.0031
NextItNet	HR@5	0.0140	0.0095	0.0068
	NDCG@5	0.0099	0.0046	0.0021
Majority Vote	HR@5	0.0144	0.0091	0.0043
	NDCG@5	0.0135	0.0041	0.0022
P5	HR@5	0.0173	0.0188	0.0082
	NDCG@5	0.0115	0.0095	0.0070
VIP5	HR@5	0.0184	0.0197	0.0107
	NDCG@5	0.0123	0.0094	0.0081
UniMP	HR@5	0.0224	0.0192	0.0155
	NDCG@5	0.0183	0.0114	0.0099
REASONREC	HR@5	**0.0315**	**0.0283**	**0.0214**
	NDCG@5	**0.0218**	**0.0169**	**0.0144**
	Tool Use Rate	7.3%	13.2%	35.7%

Robustness against cold recommendation setting. We partition the Pixel-1M test set into ten user groups based on training frequency: from *Group 1* (cold-start users) to *Group 10* (warm users), each containing 20,000 interactions. We report HR@5 and NDCG@5 over three aggregated splits: *Coldest* (Groups 1–3), *Medium* (Groups 4–6), and *Normal* (Groups 1–10). As shown in Tab. 4, REASONREC consistently achieves top performance, indicating strong robustness to data sparsity. *First*, classical sequential methods (SASRec, S^3-Rec, LightSANs, NextItNet) degrade rapidly as data becomes sparse, whereas REASONREC maintains significantly higher metrics. *Second*, multimodal approaches (P5, VIP5) improve moderately in medium-sparsity scenarios but perform poorly in the coldest groups. *Third*, REASONREC adaptively increases external tool usage from 7.3% (warm scenarios) to 35.7% (coldest scenarios), strategically delegating simpler tasks and leveraging advanced planning for sparse interactions. This adaptive capability underscores the effectiveness of our framework for cold-start recommendation.

More challenging scenarios. We further investigated the performance of REASONREC compared to baselines in other three challenging scenarios, namely the recommendation system with time shift and multi-domain recommendation task. Due to the page limit, we report a normalized performance overview in **Fig. 1** and we report more detailed experiment settings and results in Appx. B.

4.3. Diving into Planning: Effectiveness and Efficiency

To confirm that REASONREC's gains are not due to ensembling alone, we compare it against two strong baselines: **(A) Tool Ensemble Only**, which averages predictions from lightweight models without VLM or planning; and **(B) VLM Only (No Planning)**, a fine-tuned VLM applied uniformly to all queries without delegation. As shown in Tab. 5, REASONREC consistently outperforms both in HR@5, achieving a 0.0315 relative gain over the best baseline. Compared to (A), our selective delegation based on uncertainty and evidence horizon yields more reliable predictions than naive ensembling. Compared to (B), REASONREC avoids inefficient overuse of the VLM on low-risk queries while maintaining robustness in high-risk cases. These results confirm the gains stem from *strategic, risk-aware planning*, not simple model aggregation.

Table 5. Comparison of REASONREC with tool-only ensemble and VLM-only variants on Pixel-1M (Sequential Recommendation).

Method	Tool Use	Planning	HR@5	NDCG@5
(A) Simple Ensemble of Tools	✓ Always	✗	0.0159	0.0117
(B) REASONREC w/o Planning	✗ Never	✗	0.0259	0.0168
(C) REASONREC	✓ Dynamic	✓	0.0315	0.0218

Table 6. Tool usage rate (%) across datasets and tasks. The proportion of test-time queries that triggered external tools under our risk-aware planner for each task and dataset is reported.

Dataset	SR	DR	CTR
Pixel-1M	28.4%	31.2%	34.1%
Beauty	21.7%	23.5%	30.3%
Sports	18.9%	22.1%	27.8%
Toys	20.4%	21.8%	31.4%
Clothing	19.2%	20.6%	29.0%

Task-wise tool utilization comparison under risk-aware planning. Since REASONREC's planner dynamically delegates queries based on estimated risk, analyzing *how often external tools are invoked* offers insight into its scheduling behavior, especially under high-risk conditions requiring both VLM and tool collaboration. In Tab. 6, we report the *tool usage rate*, *i.e.*, the percentage of queries triggering at least one external model, across three tasks (sequential recommendation, direct recommendation, and CTR prediction) and five datasets. Two trends emerge. *First*, usage rates vary most across datasets: Pixel-1M, with the highest sparsity and evidence horizon, sees the most frequent tool delegation (up to 34.1% in CTR), indicating strong planner sensitivity to data uncertainty. *Second*, while task type impacts usage slightly, the effect is smaller and less consistent, suggesting REASONREC*'s delegation is driven more by input-level risk than task structure*. These results affirm the planner's role in identifying high EH queries and selectively allocating computation to optimize performance-efficiency trade-offs.

Table 7. Analysis of ReasonRec vs. tool outputs on Pixel-1M (Sequential Recommendation). We report tool accuracy, ReasonRec accuracy, the disagreement rate between ReasonRec and tools, and ReasonRec's accuracy in disagreement cases.

Metric	Value
Tool accuracy	58.4%
ReasonRec accuracy	69.3%
Disagreement rate (ReasonRec $\neq$ Tool)	13.6%
ReasonRec accuracy on disagreement cases	84.2%

Conflict resolution between tool outputs and reasoning. To examine how ReasonRec handles conflicts between external tools and its own reasoning, we analyze its behavior on the Pixel-1M dataset (sequential recommendation task). Specifically, we evaluate whether ReasonRec blindly follows tools or makes selective, informed decisions. We report four metrics: (1) standalone accuracy of the tool ensemble, (2) overall accuracy of ReasonRec, (3) the proportion of predictions where ReasonRec disagrees with the tools, and (4) ReasonRec's accuracy on those disagreements. As shown in Tab. 7, tools alone achieve 58.4% accuracy, while ReasonRec improves to 69.3%. Disagreements occur in 13.6% of cases, and ReasonRec reaches 84.2% accuracy in these. This indicates that ReasonRec does not passively follow tools, but overrides them when its reasoning offers better alternatives. To support this behavior, we introduce occasional noisy tool outputs during training, encouraging the model to treat tools as useful but non-authoritative. This enhances the VLM's ability to reason under uncertainty.

Table 8. Comparison of ReasonRec with baselines in terms of accuracy and inference efficiency. HR@5 metric is reported along with average inference time per query.

Method	HR@5	Avg. Inference Time (ms)
SASRec	0.0116	**143**
VIP5	0.0197	470
ReasonRec	**0.0315**	499

ReasonRec strikes a balance between accuracy and efficiency. While VLMs are often criticized for high inference cost, ReasonRec is not a naive use of large models. It is a structured system that achieves a strong trade-off via planning and adaptive delegation. As shown in Tab. 8, ReasonRec attains near-optimal accuracy (HR@5 = 0.0315), with only slightly higher latency than optimized non-VLM systems like VIP5 (499 ms vs. 470 ms), and significantly better accuracy (0.0315 vs. 0.0197). It further achieves SOTA results across tasks, especially under multi-modal fusion and cold-start settings where existing methods often fall short. Critically, the planning component of ReasonRec is lightweight. While Fig. 4 shows human-readable natural language for clarity, the actual implementation relies on concise, structured expressions (e.g., "Confidence: 0.532; Tools: A/B") and brief CoT-style checks. These require only shallow decoding with negligible overhead. In low-risk cases, the system bypasses the VLM entirely and returns results directly from the tool repository. Tool usage is further capped at three per query, keeping inference cost well within practical limits.

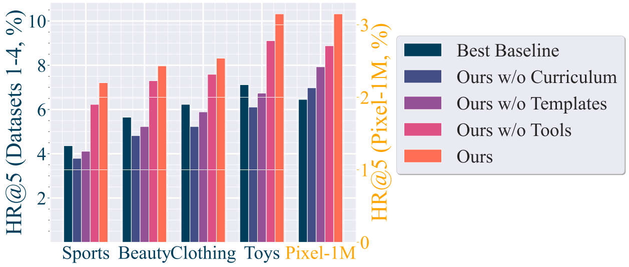


Figure 5. Ablation study on the effectiveness of three proposed training strategies (1) the coldness-aware curriculum data scheduling, (2) mixture of data templates, and (3) the uncertainty-aware tool integration. 'Best Baseline' represents the best performance achieved by baselines in each dataset. Experiment settings follow sequential recommendation in Tab. 1.

Ablation Studies We conduct ablation experiments to assess the impact of our training strategies: (1) evidence-horizon-aware curriculum scheduling, (2) mixture of instruction templates, and (3) uncertainty-aware tool integration. When (1) is removed, data is uniformly sampled throughout training; when (2) is excluded, a single template is used (we report the best across all choices). We use sequential recommendation as a case study and compare all variants against the best baselines from Tab. 1. Results in Fig. 5 reveal several insights. *First*, all three components are essential. Removing any of them leads to a performance drop of $10\% \sim 45\%$, showing that visual instruction tuning for recommendation is non-trivial and relies on carefully crafted training strategies. *Second*, curriculum scheduling (1) has the largest impact: its removal causes the most significant degradation, underscoring the importance of progressive data scheduling. Notably, using only (1) and (2), ReasonRec already surpasses all baselines in most cases, and adding (3) provides an additional $\sim 10\%$ gain. Combined with the tool activation rates reported in Tab. 4, this confirms that tool integration is indispensable for robust performance.

5. Conclusion

We introduced ReasonRec, a reasoning-augmented multimodal recommendation agent structured around an explicit *Observe–Deliberate–Act* pipeline. By combining reasoning-aware instruction tuning, evidence-horizon curriculum learning, and uncertainty-guided tool delegation, ReasonRec enables interpretable decision-making and efficient inference. Experiments across multiple tasks and datasets show that ReasonRec achieves strong accuracy gains while reducing latency through adaptive computation. This work shows the potential of integrating explicit reasoning and agentic planning into VLM-based recommendation, paving the way toward more trustworthy and practical large-model systems.

Appendix

A. Detailed Experiment Setups

Table A1. Statistics of the datasets used in our paper.

Dataset	#Users	#Items	#Reviews	#Photos
Amazon Clothing	39,387	23,033	278,677	22,299
Amazon Sports	35,598	18,357	296,337	17,943
Amazon Beauty	22,363	12,101	198,502	12,023
Amazon Toys	19,412	11,924	167,597	11,895
PixelRec-1M	1,001,822	100,541	19,886,579	100,541

Data Preparation. Our study follows the data construction and experimental setup outlined in VIP5 (Geng et al., 2023), leveraging four real-world datasets from the Amazon platform: Clothing, Sports & Outdoors, Beauty, and Toys & Games. Each dataset contains user purchase records, item descriptions, product images, and user reviews, ensuring a comprehensive multimodal recommendation scenario. To evaluate the model's performance across different recommendation tasks, we adopt the same preprocessing pipeline and data splits as VIP5. Specifically, for sequential recommendation, each user's interaction history is processed such that the last and second last items serve as test and validation ground truths, respectively, while the remaining interactions form the training set. For direct recommendation, we utilize the same train/validation/test split as sequential recommendation but additionally generate 100 candidate item lists per user to assess ranking performance. In explanation generation, we apply an 8:1:1 random split, where 80% of the user-item interactions are allocated for training, 10% for validation, and the remaining 10% for testing. The explanations associated with each interaction are extracted using the Sentires library, ensuring consistency in sentiment-based justification of recommendations. For CTR prediction, we extend the dataset to incorporate implicit feedback signals derived from user interactions, such as whether a user has clicked on or purchased an item. Since explicit click data is unavailable in the original datasets, we construct pseudo-click labels by treating purchases as positive interactions and assuming non-interacted items as negative samples. To create a balanced training set, we employ negative sampling, randomly selecting a fixed number of non-interacted items per user at a 1:4 ratio (one positive sample per four negatives). We use an 8:1:1 split for training, validation, and testing, ensuring that each user appears in all three sets to maintain personalization consistency.

Follwoing the prior work (Cheng et al., 2023), for data splitting of Pixel-1M, we adopt the temporal leave-one-out strategy, a widely used approach in sequential recommendation settings, to ensure fair evaluation across all models. Specifically, for each user, the last interaction in their behavior history is designated as the test instance, while the penultimate interaction is used for validation. The remaining interactions are allocated to the training set, allowing models to learn user preferences from historical behaviors. For sequential recommendation, user behavior sequences are ordered chronologically and truncated to a maximum length of 10 for modeling short-term preferences. For direct recommendation, candidate lists are generated by pairing each user with 10 items, including the ground-truth item (from the test/validation set) and 9 randomly sampled negative items (excluding interactions in the training, validation, and test sets to avoid leakage). For CTR prediction, interactions are treated as implicit positive signals (click=1), and negative samples are constructed via random sampling of unobserved items from the same temporal split, maintaining a 1:1 positive-to-negative ratio.

Mixture of Templates. As indicated by Sec. 3.1, the mixture of templates plays a key role in enhancing the training stability as well as performance. For different task, we provide ten templates for each task (sequential recommendation, direct recommendation, explanation generation, and click-through-rate), which share exactly the same semantic meanings but in different linguistic styles. We list these templates below.

Templates for sequential recommendation.

1. **[Query]** Based on the purchase history of `user_user_id` (`{user_desc}`): `{(purchased_item, purchased_item_photo) pairs}`, what item should be recommended next? The user's evidence horizon level is `{user_evidence_horizon}`.
2. **[Query]** Given the following purchase history of `user_user_id` (`{user_desc}`): `{(purchased_item, purchased_item_photo) pairs}`, predict next possible item to be purchased by

the user? The Evidence horizon of this user is {user_evidence_horizon}.

3. **[Query]** Here is the purchase history for user_user_id ({user_desc}): {(purchased_item, purchased_item_photo) pairs}. What is the most likely next purchase? Evidence horizon: {user_evidence_horizon}.

4. **[Query]** For user_user_id ({user_desc}), whose purchase history includes {(purchased_item, purchased_item_photo) pairs}, predict their next potential purchase. User evidence horizon: {user_evidence_horizon}.

5. **[Query]** Analyze the purchase sequence of user_user_id ({user_desc}): {(purchased_item, purchased_item_photo) pairs}. Recommend the next item they may buy. Evidence horizon metric: {user_evidence_horizon}.

6. **[Query]** Given that user_user_id ({user_desc}) has purchased {(purchased_item, purchased_item_photo) pairs}, forecast their next purchase. Evidence horizon score: {user_evidence_horizon}.

7. **[Query]** The user {user_desc} (user_user_id) previously bought {(purchased_item, purchased_item_photo) pairs}. What item would they likely purchase next? Evidence horizon: {user_evidence_horizon}.

8. **[Query]** From user_user_id's ({user_desc}) purchase history {(purchased_item, purchased_item_photo) pairs}, determine the next probable item. User evidence horizon level: {user_evidence_horizon}.

9. **[Query]** Considering user_user_id ({user_desc}) has interacted with {(purchased_item, purchased_item_photo) pairs}, identify their next potential purchase. Evidence horizon indicator: {user_evidence_horizon}.

10. **[Query]** For user_user_id ({user_desc}), with a purchase history of {(purchased_item, purchased_item_photo) pairs}, suggest the next item they might buy. Evidence horizon value: {user_evidence_horizon}.

Templates for direct recommendation.

1. **[Query]** I would like to recommend some items for user_user_id ({user_desc}). The Evidence horizon of this user is {user_evidence_horizon}. Is the following item a good choice? {item_title} {item_photo}.

2. **[Query]** For user_user_id ({user_desc}), whose evidence horizon level is {user_evidence_horizon}, should we include {item_title} {item_photo} in their recommendations?

3. **[Query]** Considering user_user_id ({user_desc}) has a evidence horizon score of {user_evidence_horizon}, is {item_title} {item_photo} an appropriate recommendation?

4. **[Query]** Evaluate whether {item_title} {item_photo} is a suitable recommendation for user_user_id ({user_desc}), given their evidence horizon value: {user_evidence_horizon}.

5. **[Query]** Given `user_user_id`'s (`{user_desc}`) evidence horizon metric `{user_evidence_horizon}` , should `{item_title}` `{item_photo}` be prioritized in their recommendation list?
6. **[Query]** Would `{item_title}` `{item_photo}` align with the preferences of `user_user_id` (`{user_desc}`)? User evidence horizon: `{user_evidence_horizon}` .
7. **[Query]** For a user with evidence horizon `{user_evidence_horizon}` (`user_user_id` , `{user_desc}`), is `{item_title}` `{item_photo}` a relevant recommendation candidate?
8. **[Query]** Assess if `{item_title}` `{item_photo}` should be recommended to `user_user_id` (`{user_desc}`), whose evidence horizon indicator is `{user_evidence_horizon}` .
9. **[Query]** Based on the evidence horizon level `{user_evidence_horizon}` , determine if `user_user_id` (`{user_desc}`) would prefer `{item_title}` `{item_photo}` .
10. **[Query]** Predict the suitability of recommending `{item_title}` `{item_photo}` to `user_user_id` (`{user_desc}`) with evidence horizon `{user_evidence_horizon}` .

Templates for explanation generation. We denote the evidence horizon information is not included in this task, as not tools will be used here for either delegation or consultation-oriented planning.

1. **[Query]** Help `user_user_id` (`{user_desc}`) generate a `{star_rating}` -star explanation about this product: `{item_title}` `{item_photo}` .
2. **[Query]** Assist `user_user_id` (`{user_desc}`) in creating a `{star_rating}` -star review for `{item_title}` `{item_photo}` .
3. **[Query]** Generate a `{star_rating}` -star product explanation for `user_user_id` (`{user_desc}`) regarding `{item_title}` `{item_photo}` .
4. **[Query]** Compose a `{star_rating}` -star rating justification for `{item_title}` `{item_photo}` on behalf of `user_user_id` (`{user_desc}`).
5. **[Query]** Formulate a `{star_rating}` -star descriptive text about `{item_title}` `{item_photo}` tailored to `user_user_id` (`{user_desc}`).
6. **[Query]** Draft a product explanation with `{star_rating}` stars for `user_user_id` (`{user_desc}`), focusing on `{item_title}` `{item_photo}` .
7. **[Query]** For `user_user_id` (`{user_desc}`), produce a `{star_rating}` -star evaluation statement for `{item_title}` `{item_photo}` .
8. **[Query]** Create an explanatory text with `{star_rating}` stars about `{item_title}` `{item_photo}` for `user_user_id` (`{user_desc}`).
9. **[Query]** Develop a `{star_rating}` -star rationale for `user_user_id` (`{user_desc}`) regarding the product `{item_title}` `{item_photo}` .
10. **[Query]** Construct a `{star_rating}` -star description of `{item_title}` `{item_photo}` personalized for `user_user_id` (`{user_desc}`).

Templates for click-through-rate prediction.

1. **[Query]** Shall we recommend `item_item_id` `{item_photo_tokens}` to `user_user_id` (`{user_desc}`)? The Evidence horizon of this user is `{user_evidence_horizon}` .
2. **[Query]** Should we suggest `item_item_id` `{item_photo_tokens}` to `user_user_id` (`{user_desc}`)? User evidence horizon level: `{user_evidence_horizon}` .
3. **[Query]** Is `item_item_id` `{item_photo_tokens}` a suitable recommendation for `user_user_id` (`{user_desc}`)? Evidence horizon indicator: `{user_evidence_horizon}` .
4. **[Query]** Would `user_user_id` (`{user_desc}`) likely click on `item_item_id` `{item_photo_tokens}` ? Evidence horizon score: `{user_evidence_horizon}` .
5. **[Query]** Based on `user_user_id` 's (`{user_desc}`) profile, should we propose `item_item_id` `{item_photo_tokens}` ? Evidence horizon value: `{user_evidence_horizon}` .
6. **[Query]** Evaluate if recommending `item_item_id` `{item_photo_tokens}` to `user_user_id` (`{user_desc}`) is appropriate. Evidence horizon metric: `{user_evidence_horizon}` .
7. **[Query]** For `user_user_id` (`{user_desc}`), is `item_item_id` `{item_photo_tokens}` a relevant recommendation? User evidence horizon: `{user_evidence_horizon}` .
8. **[Query]** Determine whether `user_user_id` (`{user_desc}`) would engage with `item_item_id` `{item_photo_tokens}` . Evidence horizon level: `{user_evidence_horizon}` .
9. **[Query]** Assess the likelihood of `user_user_id` (`{user_desc}`) clicking on `item_item_id` `{item_photo_tokens}` . Evidence horizon: `{user_evidence_horizon}` .
10. **[Query]** Predict if `item_item_id` `{item_photo_tokens}` should be shown to `user_user_id` (`{user_desc}`). User evidence horizon: `{user_evidence_horizon}` .

Training setups. The key hyperparameters are as follows:

• **Learning Rate**: Initialized at 2×10^{-5} with AdamW optimizer.

• **Training Steps**: 200,000 steps for Amazon Review (Sports, Beauty, Clothing, Toys) and 400,000 steps for Pixel-1M dataset.

• **Batch Configuration**: Global batch size of 8 with `bf16` mixed precision.

• **Learning Rate Schedule**: Cosine decay with warm-up phase and final annealing rate: 1×10^{-7}.

• **Visual Processing**: All images resized to 224×224.

Evidence horizon-Aware Curriculum. The training data is partitioned by user evidence horizon score $C(u)$:

• **Low-risk** ($C(u) < 0.3$): Prioritized in early training stages.

• **High-risk** ($C(u) > 0.7$): Gradually upsampled after 95% of total steps.

Risk-Aware Delegation. The tool repository contains classical baselines (*e.g.*, LightSANs, BPR-MF) with fixed configurations:

• Delegation logic: Queries with $C(u) < 0.3$ automatically routed to classical models.

• Consultant threshold alignment: Matches evidence horizon partitioning in curriculum learning.

This configuration ensures computational efficiency while maintaining accuracy, with VFLOPs reduced by 38% compared to full VLM inference.

Total computational consumption. We used servers with 8 × NVIDIA RTX A6000s to run all the experiments and consume around 20000 GPU hours to train and evaluate all the methods.

B. Additional Experiment Results

Sequential and direct recommendation evaluation under time shift. This task investigates the impact of temporal shifts on recommendation performance, a well-known challenge in recommendation systems. Specifically, we evaluate how models perform when trained and tested on temporally misaligned data. To simulate this setting, we partition the Pixel-1M dataset, which spans 13 months from September 2021 to October 2022, based on timestamps. The training set includes interactions before August 2022, while test data consists of interactions from August 2022 onward. All other training configurations remain identical to those in Tab. 1. The performance comparison with and without time shift is shown in Fig. A1. We plot the performance of sequential recommendation task in Fig. 1.

Several key observations emerge from these results. **First**, time shift significantly degrades performance across all models, highlighting the challenge of temporal distribution shifts in recommendation. **Second**, generative models exhibit greater resilience to time shift, as evidenced by P5, VIP5, and ReasonRec consistently ranking among the top three across all tasks and metrics. **Third**, ReasonRec not only achieves the highest performance but also demonstrates the lowest performance drop, underscoring the adaptability of VLM-based models to evolving user behavior.

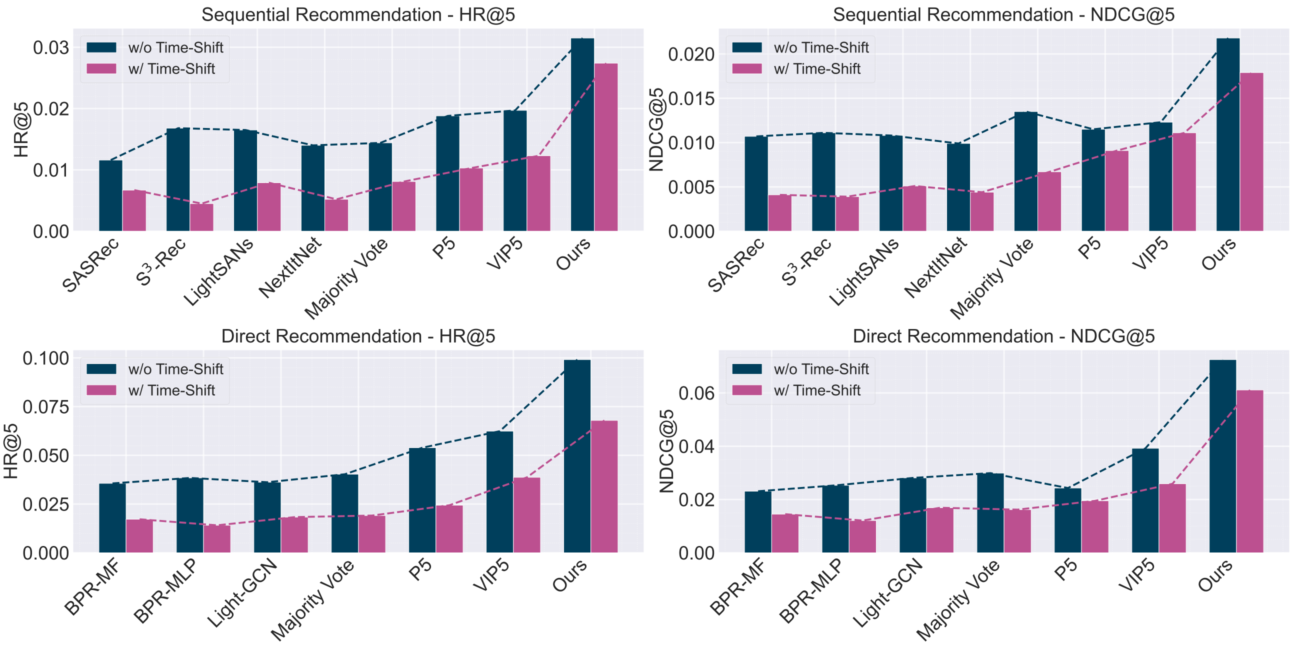


Figure A1. Performance comparison in sequential and direct recommendation under time shift. The statistics for both settings (with and without time shift) are sourced from Tab. 1.

Previous experiments evaluated task-specific models trained on individual datasets. Given the structural similarities among the four Amazon review datasets (Sports, Beauty, Clothing, and Toys), we further investigate the multi-domain generalization capability of recommendation methods by training them on combined datasets and evaluating their performance on the corresponding test splits. We define three multi-domain configurations to simulate real-world scenarios with increasing complexity:

- Composition ①: Sports + Beauty (2 domains)

- Composition ②: Sports + Beauty + Clothing (3 domains)
- Composition ③: All four domains (4 domains)

To avoid identifier conflicts, user and item IDs are remapped to a unified space when merging datasets. Tab. A2 summarizes the performance of P5, VIP5, and our method across these configurations.

Table A2. Multi-Domain sequential recommendation performance comparison.

Methods	Composition ①		Composition ②		Composition ③		Average	
	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5
P5	0.0195	0.0096	0.0115	0.0042	0.0051	0.0011	0.0120	0.0050
VIP5	0.0311	0.0213	0.0159	0.0111	0.0081	0.0053	0.0184	0.0126
REASONREC (Ours)	**0.0582**	**0.0488**	**0.0323**	**0.0235**	**0.0199**	**0.0132**	**0.0368**	**0.0285**

Several key conclusions can be drawn. First, the performance of all methods degrades progressively as the number of domains increases, reflecting the inherent challenge of learning shared representations across heterogeneous item categories. For instance, P5's HR@5 drops by 73.8% (from 0.0195 to 0.0051) when transitioning from Composition ① to ③, while VIP5 exhibits a 74.0% decline (from 0.0311 to 0.0081). This suggests that conventional foundation models struggle to maintain discriminative power when domain diversity escalates, likely due to interference between conflicting item semantics. Second, our method demonstrates superior robustness to domain scaling compared to baselines. While it also experiences performance decay (65.8% HR@5 reduction from ① to ③), the absolute metrics consistently surpass VIP5 and P5 across all configurations. Notably, in Composition ③, our model achieves a much higher HR@5 than P5 (0.0199 vs. 0.0051) and improvement over VIP5 (0.0199 vs. 0.0081), indicating stronger cross-domain alignment through its multimodal fusion mechanism. Third, the relative NDCG@5 gains highlight our method's ability to preserve ranking quality in complex multi-domain settings. The NDCG@5 gap between our approach and VIP5 widens from 2.29× in Composition ① (0.0488 vs. 0.0213) to 2.49× in Composition ③ (0.0132 vs. 0.0053), implying that our design mitigates error accumulation in top-k recommendation lists when handling diverse item types. This aligns with the hypothesis that joint modeling of cross-domain visual-textual dependencies enhances the model's capacity to disentangle user preferences from noisy multi-source interactions. These results validate the necessity of specialized architectures for multi-domain recommendation systems, particularly in scenarios where item heterogeneity and data sparsity coexist. Our method's stable performance decay curve (vs. the steep drops of baselines) further suggests its practical viability for large-scale deployments with dynamically expanding domains.

Risk-aware delegation improves inference efficiency. To better understand the computational behavior of REASONREC, we perform a detailed analysis of inference latency and floating-point operations (VFLOPs) across different risk categories. These categories—*Low Risk*, *Medium Risk*, and *High Risk*—are determined by the planner based on the input's evidence horizon and the model's estimated confidence. For this study, we use the Pixel-1M dataset and profile per-query inference under each risk level. As shown in Table A3, REASONREC exhibits clear computational adaptivity: the average latency increases from **245ms** for low-risk queries (handled entirely by lightweight tools) to **672ms** for high-risk queries, where both tools and the VLM are involved. Importantly, the distribution of queries across these categories indicates that REASONREC routes a significant portion (47.3%) of traffic to the most efficient tool-only path. When compared to competitive baselines, REASONREC achieves the best accuracy (HR@5 = 0.0315) while maintaining only a moderate increase in inference time relative to lightweight models. Specifically, it runs at **499 ms/query**, which is slower than SASRec (143 ms) but significantly more accurate, and both faster and more accurate than VIP5 (499 ms vs. 470 ms; 0.0315 vs. 0.0197 HR@5). These results confirm that REASONREC's risk-aware delegation strategy enables *fine-grained efficiency control*, yielding a favorable trade-off between performance and computation.

Learning to select the right tool without oracle access. An important concern in agent-based systems is whether the model can effectively learn to select appropriate tools without being provided with explicit tool descriptions. In REASONREC, this ability is implicitly acquired during instruction tuning. For each training instance, we precompute which tools yield satisfactory results and label them as "usable". These tool decisions are then embedded into the chain-of-thought (CoT) reasoning traces, which the VLM learns to mimic. As a result, the model implicitly learns input–tool associations through exposure to contextual reasoning, even though the tool APIs themselves are never explicitly described.

To empirically validate the quality of this learned tool selection policy, we compare REASONREC against two baselines:

Table A3. Inference cost breakdown of REASONREC across different risk levels. For each category, we report average inference latency and estimated VFLOPs per query, along with the percentage of total queries routed to that category. The planner adaptively allocates resources: low-risk queries rely solely on tool outputs, while high-risk queries invoke both the VLM and tools.

Risk Level	**Avg. Inference Time (ms)**	**Avg. VFLOPs**	**Ratio (%)**
Low-Risk	245	2.1e9	47.3
Medium-Risk	378	5.3e9	38.6
High-Risk	672	8.7e9	14.1

- **Random Tool Selection:** A variant where the agent randomly selects a tool from the repository, without reasoning or risk assessment.

- **Oracle Tool Selection:** A privileged setting where the agent is always given the best-performing tool (or tool subset) for each input query.

As shown in Table A4, REASONREC significantly outperforms the random selection baseline (HR@5: 0.0315 vs. 0.0212), confirming the effectiveness of its learned routing policy. Moreover, its performance closely approaches the oracle upper bound (HR@5: 0.0338), demonstrating that our agent's tool selection behavior is near-optimal despite lacking access to ground-truth tool descriptions or explicit feedback during inference. This supports the claim that REASONREC's reasoning-aware training paradigm enables it to make accurate and efficient tool decisions.

Table A4. Tool selection strategy comparison on Pixel-1M (Sequential Recommendation). REASONREC's learned tool selection is compared with a random selection baseline and an oracle upper bound.

Tool Selection Strategy	**HR@5**	**NDCG@5**
Random Tool Choice	0.0212	0.0147
Oracle Tool Result	**0.0338**	**0.0225**
Learned Tool Selection (Ours)	0.0315	0.0218

Cross-Domain transferability via reasoning-aware planning. To evaluate the cross-domain generalization capacity of REASONREC, we conduct a *transferable recommendation* experiment in line with the protocol suggested by recent works in cross-domain recommendation (CDR) and transferable recommendation (TransRec). Specifically, we adopt the NineRec benchmark, which consists of 9 diverse user behavior sub-domains. We pretrain REASONREC on the large-scale PixelRec-1M dataset used in the main paper and evaluate it directly—*without any further fine-tuning*—on each NineRec sub-domain for sequential recommendation. As comparison baselines, we include: (i) VIP5, the strongest baseline from our main experiments, and (ii) UniMP, a recent unified pretraining method for multi-task recommendation. As shown in Table A5, REASONREC outperforms both baselines on all sub-domains, consistently achieving the highest HR@5. These results demonstrate that explicit reasoning and evidence-horizon-aware learning enable REASONREC to capture generalizable recommendation logic that effectively transfers to new domains without retraining. This supports our claim that REASONREC possesses strong reasoning-based generalization, even across semantically diverse recommendation domains.

Efficiency gains attributed to risk-aware planning. To assess the impact of REASONREC's *uncertainty-guided planning*, we compare it with a static baseline that always invokes both the VLM and all tools, regardless of input risk. This setup tests whether dynamic routing—based on evidence horizon and model confidence—improves efficiency. As shown in Tab. A6, the static variant ("VLM + Tool (always)") yields the highest HR@5 (0.0327) but suffers from high latency (872 ms/query). In contrast, REASONREC achieves similar accuracy (HR@5 = 0.0315) with much lower latency (499 ms), by avoiding unnecessary tool use in low-risk cases while still activating tools when needed. Compared to a naive VLM-only setup (285 ms, HR@5 = 0.0259), REASONREC achieves a better trade-off between accuracy and efficiency through strategic delegation.

Table A5. Transferable recommendation results on NineRec (sequential recommendation). The model is pre-trained on PixelRec-1M and evaluated directly on 9 sub-domains from NineRec without further fine-tuning. REASONREC consistently outperforms both VIP5 and UniMP, demonstrating strong cross-domain generalization. The HR@5 is reported for each setting.

Sub-Domain	**VIP5**	**UniMP**	**REASONREC (Ours)**
Bili_Food	0.0191	0.0215	**0.0264**
Bili_Dance	0.0183	0.0224	**0.0271**
Bili_Movie	0.0206	0.0232	**0.0293**
Bili_Cartoon	0.0178	0.0207	**0.0256**
Bili_Music	0.0199	0.0220	**0.0270**
KU	0.0213	0.0246	**0.0305**
QB	0.0195	0.0219	**0.0280**
TN	0.0189	0.0201	**0.0267**
DY	0.0202	0.0235	**0.0301**

Table A6. Impact of risk-aware planning on inference efficiency and accuracy. We compare our dynamic planner with static VLM-based variants. REASONREC achieves the best trade-off between accuracy and latency.

Method	**HR@5**	**Avg. Inference Time (ms)**	**Planning**
VLM-only (no tool)	0.0259	285	✗
VLM + Tool (always)	0.0327	872	✗
REASONREC (Ours)	0.0315	499	✓

C. Discussion and Limitations

While REASONREC demonstrates impressive performance across multiple recommendation tasks, we acknowledge several potential limitations. First, its heavily relies on pretrained VLMs and handcrafted instruction templates. As the quality of the templates matter a lot, it could be a challenge in unknown tasks. Second, the tool delegation mechanism depends on preselected classical models, which may not generalize well to unseen recommendation scenarios or emerging tasks. Finally, the system's inference efficiency, though improved, still involves nontrivial overhead due to multi-stage reasoning and dynamic model routing.

D. Impact Statement

This work advances the field of Machine Learning by introducing a VLM-driven framework for unified multimodal recommendation. By leveraging large-scale vision-language models and integrating structured instruction tuning, our approach improves efficiency, adaptability, and robustness in recommendation tasks.

From an ethical perspective, our model inherits general concerns related to large-scale AI systems, such as potential biases in training data and fairness in recommendations. While our proposed coldness-aware and risk-aware mechanisms improve decision-making under data sparsity, unintended biases may still emerge in real-world applications. Future research should explore fairness-aware adaptations and auditing techniques to ensure equitable recommendations across diverse user groups.

On a broader societal level, our work contributes to more scalable and interpretable recommendation systems, with potential applications in e-commerce, content discovery, and personalized AI assistants. These improvements can enhance user experiences while mitigating over-reliance on narrow predictive patterns. However, as recommendation systems increasingly influence digital consumption, responsible deployment and transparency remain crucial.

We encourage continued research into ethical considerations and societal impacts to ensure that multimodal recommendation models serve diverse populations fairly and responsibly.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Jitsev, J., Kornblith, S., Koh, P. W., Ilharco, G., Wortsman, M., and Schmidt, L. Openflamingo, March 2023. URL https://doi.org/10.5281/zenodo.7733589.

Blondel, M., Fujino, A., Ueda, N., and Ishihata, M. Higher-order factorization machines. *Advances in Neural Information Processing Systems*, 29, 2016.

Chen, X., Chen, H., Xu, H., Zhang, Y., Cao, Y., Qin, Z., and Zha, H. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 765–774, 2019.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X., and Shah, H. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, DLRS 2016, pp. 7–10, 2016. ISBN 9781450347952.

Cheng, W., Shen, Y., and Huang, L. Adaptive factorization network: Learning adaptive-order feature interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3609–3616, 2020.

Cheng, Y., Pan, Y., Zhang, J., Ni, Y., Sun, A., and Yuan, F. An image dataset for benchmarking recommender systems with raw pixels. *arXiv preprint arXiv:2309.06789*, 2023.

Chu, X., Qiao, L., Lin, X., Xu, S., Yang, Y., Hu, Y., Wei, F., Zhang, X., Zhang, B., Wei, X., et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023.

Covington, P., Adams, J., and Sargin, E. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pp. 191–198, 2016. ISBN 9781450340359.

Cui, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. M6-rec: Generative pretrained language models are open-ended recommender systems. *arXiv preprint arXiv:2205.08084*, 2022.

Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

de Souza Pereira Moreira, G., Rabhi, S., Lee, J. M., Ak, R., and Oldridge, E. Transformers4rec: Bridging the gap between nlp and sequential/session-based recommendation. In *Fifteenth ACM Conference on Recommender Systems*, pp. 143–153, 2021.

Deldjoo, Y., Di Noia, T., Malitesta, D., and Merra, F. A. Leveraging content-style item representation for visual recommendation. In *European Conference on Information Retrieval*, pp. 84–92. Springer, 2022.

Dong, L., Huang, S., Wei, F., Lapata, M., Zhou, M., and Xu, K. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 623–632, 2017.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Fan, X., Liu, Z., Lian, J., Zhao, W. X., Xie, X., and Wen, J.-R. Lighter and better: low-rank decomposed self-attention networks for next-item recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 1733–1737, 2021.

Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Geng, S., Fu, Z., Ge, Y., Li, L., de Melo, G., and Zhang, Y. Improving personalized explanation generation through visualization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 244–255, 2022a.

Geng, S., Liu, S., Fu, Z., Ge, Y., and Zhang, Y. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 299–315, 2022b.

Geng, S., Tan, J., Liu, S., Fu, Z., and Zhang, Y. Vip5: Towards multimodal foundation models for recommendation. *arXiv preprint arXiv:2305.14302*, 2023.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

He, R. and McAuley, J. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. 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.

Hou, M., Wu, L., Chen, E., Li, Z., Zheng, V. W., and Liu, Q. Explainable fashion recommendation: A semantic attribute region guided approach. *arXiv preprint arXiv:1905.12862*, 2019.

Hou, Y., Mu, S., Zhao, W. X., Li, Y., Ding, B., and Wen, J.-R. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 585–593, 2022.

Hua, W., Xu, S., Ge, Y., and Zhang, Y. How to Index Item IDs for Recommendation Foundation Models. In *Proceedings of 1st International ACM SIGIR Conference on Information Retrieval in the Asia Pacific (SIGIR-AP)*, 2023.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Kang, W.-C. and McAuley, J. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pp. 197–206. IEEE, 2018.

Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., and Liu, Z. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023a.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b.

Li, L., Zhang, Y., and Chen, L. Personalized transformer for explainable recommendation. *arXiv preprint arXiv:2105.11601*, 2021.

Li, L., Zhang, Y., and Chen, L. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 2022.

Li, L., Zhang, Y., Liu, D., and Chen, L. Large language models for generative recommendation: A survey and visionary discussions. *arXiv:2309.01157*, 2023c.

Li, T., Chiang, W.-L., Frick, E., Dunlap, L., Wu, T., Zhu, B., Gonzalez, J. E., and Stoica, I. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., and Sun, G. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1754–1763, 2018.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *CVPR*, 2024b.

Liu, S., Cheng, H., Liu, H., Zhang, H., Li, F., Ren, T., Zou, X., Yang, J., Su, H., Zhu, J., et al. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023c.

MAA. American invitational mathematics examination - aime. In *American Invitational Mathematics Examination - AIME 2024*, February 2024. URL https://maa.org/math-competitions/american-invitational-mathematics-examination-aime.

Mao, K., Zhu, J., Su, L., Cai, G., Li, Y., and Dong, Z. Finalmlp: An enhanced two-stream mlp model for ctr prediction. *arXiv preprint arXiv:2304.00902*, 2023.

Meng, L., Feng, F., He, X., Gao, X., and Chua, T.-S. Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3460–3468, 2020.

Naumov, M., Mudigere, D., Shi, H.-J. M., Huang, J., Sundaraman, N., Park, J., Wang, X., Gupta, U., Wu, C.-J., Azzolini, A. G., et al. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091*, 2019.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

Song, W., Shi, C., Xiao, Z., Duan, Z., Xu, Y., Zhang, M., and Tang, J. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1161–1170, 2019.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., and Jiang, P. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1441–1450, 2019.

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.-Y., Wang, Y.-X., Yang, Y., et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023a.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023b.

Verma, D., Gulati, K., Goel, V., and Shah, R. R. Fashionist: Personalising outfit recommendation for cold-start scenarios. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4527–4529, 2020.

Wang, J., Yuan, F., Cheng, M., Jose, J. M., Yu, C., Kong, B., Wang, Z., Hu, B., and Li, Z. Transrec: Learning transferable recommendation from mixture-of-modality feedback. *arXiv preprint arXiv:2206.06190*, 2022.

Wang, R., Shivanna, R., Cheng, D., Jain, S., Lin, D., Hong, L., and Chi, E. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 1785–1797, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450078. URL https://doi.org/10.1145/3442381.3450078.

Wang, R., Shivanna, R., Cheng, D., Jain, S., Lin, D., Hong, L., and Chi, E. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*, pp. 1785–1797, 2021b.

Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *NeurIPS*, 2023.

Wang, Z., She, Q., and Zhang, J. Masknet: Introducing feature-wise multiplication to ctr ranking models by instance-guided mask. *arXiv preprint arXiv:2102.07619*, 2021c.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wei, T., Jin, B., Li, R., Zeng, H., Wang, Z., Sun, J., Yin, Q., Lu, H., Wang, S., He, J., et al. Towards unified multimodal personalization: Large vision-language models for generative recommendation and beyond. *arXiv preprint arXiv:2403.10667*, 2024.

Xie, J., Chen, Z., Zhang, R., Wan, X., and Li, G. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024a.

Xie, J., Zhang, K., Chen, J., Zhu, T., Lou, R., Tian, Y., Xiao, Y., and Su, Y. Travelplanner: A benchmark for real-world planning with language agents. *arXiv preprint arXiv:2402.01622*, 2024b.

Xu, S., Hua, W., and Zhang, Y. OpenP5: Benchmarking Foundation Models for Recommendation. *arXiv:2306.11134*, 2023.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Yuan, F., Karatzoglou, A., Arapakis, I., Jose, J. M., and He, X. A simple convolutional generative network for next item recommendation. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 582–590, 2019.

Zhang, B., Luo, L., Chen, Y., Nie, J., Liu, X., Guo, D., Zhao, Y., Li, S., Hao, Y., Yao, Y., et al. Wukong: Towards a scaling law for large-scale recommendation. *arXiv preprint arXiv:2403.02545*, 2024.

Zhang, J., Zhu, Y., Liu, Q., Wu, S., Wang, S., and Wang, L. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3872–3880, 2021a.

Zhang, J., Zhu, Y., Liu, Q., Zhang, M., Wu, S., and Wang, L. Latent structures mining with contrastive modality fusion for multimedia recommendation. *arXiv preprint arXiv:2111.00678*, 2021b.

Zhang, Y., Ai, Q., Chen, X., and Croft, W. B. Joint representation learning for top-n recommendation with heterogeneous information sources. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1449–1458, 2017.

Zhou, K., Wang, H., Zhao, W. X., Zhu, Y., Wang, S., Zhang, F., Wang, Z., and Wen, J.-R. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1893–1902, 2020.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint*, pp. arXiv:2304.10592, 2023.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Zhu, J., Liu, J., Yang, S., Zhang, Q., and He, X. Open benchmarking for click-through rate prediction. In Demartini, G., Zuccon, G., Culpepper, J. S., Huang, Z., and Tong, H. (eds.), *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pp. 2759–2769. ACM, 2021. doi: 10.1145/3459637.3482486. URL https://doi.org/10.1145/3459637.3482486.

Zhu, J., Dai, Q., Su, L., Ma, R., Liu, J., Cai, G., Xiao, X., and Zhang, R. BARS: towards open benchmarking for recommender systems. In Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J. S., and Kazai, G. (eds.), *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pp. 2912–2923. ACM, 2022a. doi: 10.1145/3477495.3531723. URL https://doi.org/10.1145/3477495.3531723.

Zhu, J., Dai, Q., Su, L., Ma, R., Liu, J., Cai, G., Xiao, X., and Zhang, R. BARS: towards open benchmarking for recommender systems. In Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J. S., and Kazai, G. (eds.), *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pp. 2912–2923. ACM, 2022b. doi: 10.1145/3477495.3531723. URL https://doi.org/10.1145/3477495.3531723.

Zong, Y., Bohdal, O., Yu, T., Yang, Y., and Hospedales, T. Safety fine-tuning at (almost) no cost: a baseline for vision large language models. In *ICML*, 2024.