

# OneRec-THINK: In-Text Reasoning for Generative Recommendation

Zhanyu Liu\*, Shiyao Wang\*, Xingmei Wang, Rongzhou Zhang, Jiaxin Deng,  
 Honghui Bao, Jinghao Zhang, Wuchao Li, Pengfei Zheng, Xiangyu Wu,  
 Yifei Hu, Qigen Hu, Xincheng Luo, Lejian Ren, Zixing Zhang,  
 Qianqian Wang, Kuo Cai, Yunfan Wu, Hongtao Cheng, Zexuan Cheng,  
 Lu Ren, Huanjie Wang, Yi Su, Ruiming Tang, Kun Gai, Guorui Zhou<sup>†</sup>

Kuaishou Inc., Beijing, China

{liuzhanyu, wangshiyao08, zhouguorui}@kuaishou.com

## Abstract

The powerful generative capacity of Large Language Models (LLMs) has instigated a paradigm shift in recommendation. However, existing generative models (e.g., OneRec) operate as implicit predictors, critically lacking the capacity for explicit and controllable reasoning—a key advantage of LLMs. To bridge this gap, we propose OneRec-Think, a unified framework that seamlessly integrates dialogue, reasoning, and personalized recommendation. OneRec-Think incorporates: (1) **Itemic Alignment**: cross-modal Item-Textual Alignment for semantic grounding; (2) **Reasoning Activation**: Reasoning Scaffolding to activate LLM reasoning within the recommendation context; and (3) **Reasoning Enhancement**, where we design a recommendation-specific reward function that accounts for the multi-validity nature of user preferences. Experiments across public benchmarks show state-of-the-art performance. Moreover, our proposed "Think-Ahead" architecture enables effective industrial deployment on Kuaishou, achieving a 0.159% gain in APP Stay Time and validating the practical efficacy of the model's explicit reasoning capability.

## 1 Introduction

The rapid advancement of Large Language Models (LLMs) has fundamentally reshaped recommender systems, ushering in the generative retrieval paradigm(GR) (Peng et al., 2025; Zhang et al., 2025b; Deldjoo et al., 2024; Li et al., 2023; Wang et al., 2023). This approach represents a profound shift from traditional query-candidate matching, utilizing Transformer-based sequence-to-sequence models to autoregressively decode the identifiers of target candidates. Capitalizing on this, a major research frontier is the development

\*: Equal contributions.

<sup>†</sup>: Corresponding Author.



Figure 1: Examples of OneRec-Think’s Unified Dialogue, Reasoning and Recommendation Framework.

of end-to-end generative frameworks, including OneRec, OneLoc, OneSug, and OneSearch (Deng et al., 2025; Zhou et al., 2025a,b; Wei et al., 2025; Guo et al., 2025b; Chen et al., 2025). These unified models replace the traditional multi-stage recommendation funnel (involving separate retrieval and ranking stages), enabling holistic optimization towards the final objective and concentrating computational resources for better industrial scaling and performance.

While these models successfully harness the LLMs’ capacity for output generation, they fundamentally lack the explicit, verifiable reasoning pathways that define modern LLM breakthroughs,

such as text-based Chain-of-Thought (CoT) (Rajput et al., 2023; Zheng et al., 2024; Wang et al., 2024b). To bridge this critical gap, we propose **OneRec-Think, a novel framework that integrates dialogue, reasoning, and personalized generative recommendations within a single, unified model**. It is capable of generating high-quality, interpretable textual reasoning paths, significantly enhancing both recommendation accuracy and user trustworthiness. The model’s inherent dialogic nature further enables dynamic tailoring of suggestions to specific user constraints (as shown in Fig. 1). Our approach is realized through a three-stage framework: (1) **Itemic Alignment**, which maps item semantics into the LLM’s textual embedding space, establishing a unified representational continuum that unlocks the model’s capacity for reasoning. (2) **Reasoning Activation**, which aims to induce the LLM’s inherent reasoning ability directly within the context of recommender systems; (3) **Reasoning Enhancement**, which utilizes a recommendation-specific reward function that captures the multi-validity (i.e., multiple valid choices) nature of user preferences. Furthermore, we introduce the OneRec-Think Inference Architecture to ensure efficient deployment and real-time responsiveness in industrial-scale serving scenarios.

Our contributions are summarized as follows:

- We introduce a unified framework that bridges the semantic gap between discrete recommendation items and continuous reasoning spaces, enabling seamless integration of personalized recommendation within LLMs’ natural language understanding.
- We design a novel reasoning paradigm that orchestrates multi-step deliberation with recommendation optimization, achieving interpretable and accuracy-aware personalized recommendation through synergistic training.
- The proposed approach achieves state-of-the-art results on multiple public benchmarks, while our deployment-friendly "Think-Ahead" architecture enables significant industrial impact with a 0.159% gain in APP Stay Time.

## 2 Related Work

### 2.1 Reasoning in Large Language Models

Large language models achieve complex reasoning through various prompting techniques, with

CoT prompting (Wei et al., 2022) being the foundational approach that decomposes problems into intermediate reasoning steps. This has inspired numerous extensions including zero-shot CoT (Kojima et al., 2022), self-consistency decoding (Wang et al., 2022), and tree-of-thoughts (Yao et al., 2023). These techniques enable test-time scaling where additional computational budget during inference improves performance (Snell et al., 2024). Recent work has shifted focus from prompting to post-training enhancement of reasoning capabilities by using techniques such as Reinforcement Learning. Models including DeepSeek-R1 (Guo et al., 2025a) and Seed-1.5 (Seed et al., 2025) optimize reasoning behaviors via techniques such as GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), and VAPO (Yue et al., 2025), which demonstrates promising advancements in this direction.

### 2.2 Reasoning-Based Recommendation

Although generative recommendation models such as TIGER (Rajput et al., 2023), HSTU (Zhai et al., 2024), and OneRec (Zhou et al., 2025a) have demonstrated effectiveness, they inherently lack reasoning capabilities. Recently, the Reasoning-based recommendation systems aim to perform multi-step deduction for more accurate and interpretable recommendations. Existing approaches fall into two categories: **explicit reasoning** methods generate human-readable rationales but are confined to discriminative tasks (Tsai et al., 2024; Bisimay et al., 2024; Fang et al., 2025; Kim et al.), while **implicit reasoning** methods (Zhang et al., 2025a; Tang et al., 2025) perform latent reasoning without textual interpretability. Our work introduces explicit reasoning into generative recommendation, bridging this gap to enable both interpretable rationales and scalable item generation.

## 3 Preliminary

**Itemic Token** An itemic token is a discrete, semantic-rich representation unit for an item, analogous to a word token in natural language. Following OneRec (Zhou et al., 2025a; Rajput et al., 2023), we map each item  $v$  to a sequence of such tokens  $s_v = (s_v^1, \dots, s_v^L)$ , which are generated from the item’s multi-modal and collaborative content.

**Problem Definition** Let  $\mathcal{U}$  and  $\mathcal{V}$  denote the sets of users and items, respectively. Each user  $u \in \mathcal{U}$  has a chronological interaction history  $V_u = (v_1^u, v_2^u, \dots, v_{n_u}^u)$  of length  $n_u$ . For brevity,

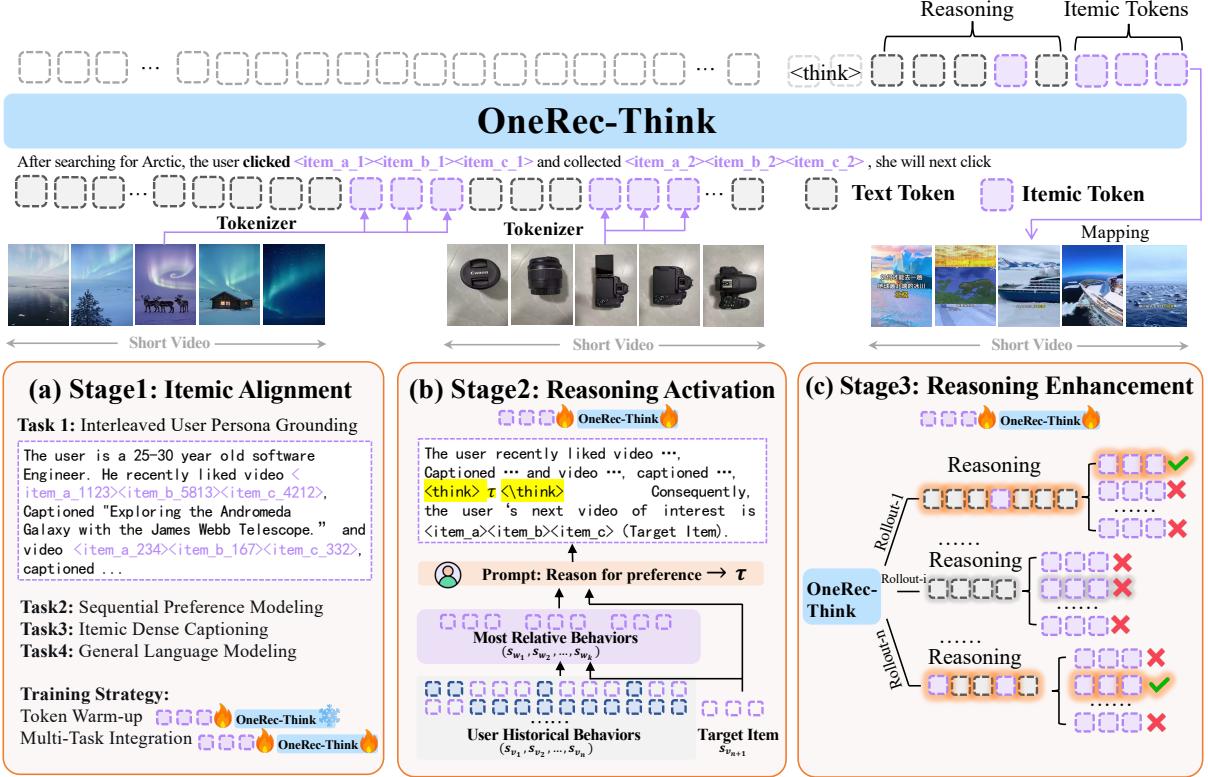


Figure 2: The framework of the OneRec-Think. In the first stage, we achieve item-level semantic alignment through multi-task pre-training. In the second stage, we activate explicit reasoning by prompting the model to generate preference rationales. In the third stage, we refine the reasoning paths through RL based on a reward tailored for recommendations.

we omit the user index  $u$ . By using the itemic tokens, the user’s interaction history is thus represented by the sequence  $S_u = (\mathbf{s}_{v_1}, \dots, \mathbf{s}_{v_n})$ .

Conventional generative recommenders (Rajput et al., 2023; Zhou et al., 2025a,b; Zheng et al., 2024) define their task with a generation target of the next itemic tokens as:

$$\mathbf{s}_{v_{n+1}} \sim P(\cdot | \mathbf{s}_{v_1}, \dots, \mathbf{s}_{v_n}; \theta) \quad (1)$$

In contrast, we reformulate this task to unify reasoning and recommendation in a single autoregressive pass. Conditioned on a prompted user history, we generate tokens sequentially, beginning with a reasoning sequence  $\tau = (r_1, \dots, r_M)$  and concluding with the next itemic tokens  $\mathbf{s}_{v_{n+1}}$ . This end-to-end process is captured by:

$$\begin{aligned} \tau &\sim P(\cdot | \mathcal{P}(\mathbf{s}_{v_1}, \dots, \mathbf{s}_{v_n}); \theta) \\ \mathbf{s}_{v_{n+1}} &\sim P(\cdot | \mathcal{P}(\mathbf{s}_{v_1}, \dots, \mathbf{s}_{v_n}), \tau; \theta) \end{aligned} \quad (2)$$

where  $\mathcal{P}(\cdot)$  means a valid prompt constructed for the recommendation.

## 4 Methodology

We now present OneRec-Think, a scalable framework for end-to-end generative reasoning recommendation. Our approach comprises three core components: an Itemic Alignment stage, a Reasoning Activation stage, and a "Think Ahead" architecture for industrial deployment. The illustration of OneRec-Think is shown in Fig. 2.

### 4.1 Itemic Alignment through Multi-Task Pre-training

To align recommender knowledge with the LLM’s linguistic space, we design a Multi-Task Pre-training strategy enabling seamless processing of natural language and itemic tokens via four complementary tasks under Next Token Prediction.

**Interleaved User Persona Grounding** Unlike prior work that uses either pure textual data or isolated item sequences, this task interleaves the itemic tokens and text tokens of User Persona. It includes serialized static attributes, active search behaviors, interactive sequences, and summarized user interests. This composition creates rich, dual-

modality training instances where itemic tokens are grounded in their semantic context.

**Sequential Preference Modeling** As the core recommendation task, this task constructs data that teaches the model to predict subsequent item interactions from chronological user histories.

**Itemic Dense Captioning** This task requires the model to decode an item’s descriptive content from its itemic tokens. By learning to generate detailed textual descriptions, the model establishes a fundamental understanding of the semantic characteristics represented by item combinations.

**General Language Modeling** This task continues pre-training the model on general text corpora, preserving the model’s fundamental language capabilities during applying the model to recommendation scenarios.

To enable effective knowledge integration while preserving the linguistic capabilities of the model, we implement a two-substage training strategy to ensure stable alignment. The **Token Warm-up** substage exclusively trains itemic token embeddings on the Interleaved User Persona Grounding task while keeping the base LLM frozen. The subsequent **Multi-Task Integration** substage jointly optimizes all parameters on the combined task using a designed ratio (see Appendix A.3 for details).

## 4.2 Reasoning Activation

Despite robust itemic alignment, direct application to industrial recommendation scenarios often fails to yield effective CoT reasoning due to the noisy and lengthy nature of real-world user behavior sequences. To address this, we propose a supervised fine-tuning framework that first extracts coherent reasoning trajectories from pruned user contexts, then leverages these trajectories to guide rationale generation over raw behavioral data, enabling effective contextual distillation for noisy industrial settings (as shown in Fig. 2(b)).

**Bootstrapping with Pruned Contexts:** To bootstrap reasoning capabilities, we first construct easy-to-learn instances where logical relationships are preserved despite sequence pruning. For each user, we select the target item  $s_{v_{n+1}}$  and form a context-target pair  $\langle (s_{v_1}, \dots, s_{v_n}), s_{v_{n+1}} \rangle$ . We then retrieve the top- $k$  most relevant historical items using a similarity function  $g(\cdot, \cdot)$ :

$$g((s_{v_1}, \dots, s_{v_n}), s_{v_{n+1}}) = (s_{w_1}, \dots, s_{w_k}). \quad (3)$$

Using these relevant items, we query our pre-aligned model to generate a rationale  $\tau$  explaining the target interaction:

$$\tau \sim P(\cdot | \mathcal{P}_r((s_{w_1}, \dots, s_{w_k}), s_{v_{n+1}}); \theta) \quad (4)$$

where  $\mathcal{P}_r(a, b)$  means constructs a prompt to query the rationale why a user who interacts with item sequence  $a$  would interact with item  $b$ . This process yields high-quality rationales that are both logically sound and target-aligned, providing ideal training signals for reasoning induction.

**Learning to Reason from Noisy Sequences:** The distilled rationales serve as supervision for learning to reason from raw sequences. The training objective minimizes the negative log-likelihood of generating both the rationale and target item:

$$\begin{aligned} \mathcal{L}_{RA} = & - \left( \sum_{i=1}^M \log P(r_i | \mathcal{P}(s_{v_1}, \dots, s_{v_n}), r_{<i}; \theta) \right. \\ & \left. + \sum_{j=1}^L \log P(s_{v_{n+1}}^j | \mathcal{P}(s_{v_1}, \dots, s_{v_n}), \tau, s_{v_{n+1}}^{<j}; \theta) \right), \end{aligned} \quad (5)$$

where  $\tau = \{r_1, \dots, r_M\}$  represents the rationale tokens and  $s_{v_{n+1}} = \{s_{v_{n+1}}^1, \dots, s_{v_{n+1}}^L\}$  denotes the target item tokens. By optimizing  $\mathcal{L}_{RA}$ , the model learns to internally distill relevant context from noisy sequences and generate coherent rationales that bridge user history to target interactions, significantly enhancing its CoT capabilities in challenging recommendation scenarios.

## 4.3 Reasoning Enhancement

Building upon the CoT capabilities from Reasoning Activation, we address the challenge of ensuring consistently high-quality reasoning through Reinforcement Learning. This stage refines the recommendation accuracy using a novel reward mechanism tailored for generative recommendation.

**Beam Candidate Reward Maximization:** Standard verifiable pass rewards face significant sparsity challenges in recommendation scenarios, as most reasoning rollouts fail to hit the target item and consequently yield identical zero rewards—thereby neutralizing group advantages in algorithms such as GRPO (Shao et al., 2024). To overcome this, we introduce the **Rollout-Beam** reward that evaluates reasoning capability by the model’s best achievable performance within a constrained beam. Our approach employs beam search with width  $K$  to

explore multiple generation candidates after reasoning trajectory generation:

$$\mathcal{R}_{\text{Rollout-Beam}} = \max_{\hat{s}_{v_{n+1}} \in \mathcal{B}} \sum_{l=1}^L \mathbb{I}(\hat{s}_{v_{n+1}}^l = s_{v_{n+1}}^l), \quad (6)$$

where the beam search result set is defined as :

$$\begin{aligned} \mathcal{B} &= \{(\hat{s}_{v_{n+1}}^{1,(j)}, \dots, \hat{s}_{v_{n+1}}^{L,(j)})_{j=1}^K\}, \\ &= \text{BeamSearch}\left(P(s_{v_{n+1}} | \mathbf{H}, \tau; \theta), K\right) \end{aligned} \quad (7)$$

which contains the items with the top- $K$  probability in the beam search within the distribution  $P(s_{v_{n+1}} | \mathbf{H}, \tau; \theta)$ .  $\mathbf{H} = \mathcal{P}(s_{v_1}, \dots, s_{v_n})$  is a valid prompt of history sequence. BeamSearch( $P, K$ ) means the top- $K$  result of beam search within distribution  $P$ . Subsequently, we optimize the model using GRPO (Shao et al., 2024) based on  $\mathcal{R}_{\text{Rollout-Beam}}$ , which effectively leverages the enriched reward signals from the multi-validity nature of user preferences.

Overall, this design establishes training-inference consistency by aligning reward computation with beam search-based inference, providing denser learning signals through multi-path evaluation.

#### 4.4 Industrial Deployment: A "Think-Ahead" Architecture

The deployment of OneRec-Think in industrial recommendation systems presents a fundamental challenge: reconciling the computational demands of multi-step reasoning with the stringent latency requirements of real-time user interactions.

To address this critical bottleneck, we introduce a novel *Think-Ahead* Inference Architecture. Our solution strategically decouples the model's inference into two stages: **In the first stage**, the computationally intensive reasoning path and the initial item-tokens (e.g., the first two itemic tokens) are generated offline by the full OneRec-Think model. These initial tokens are designed to capture the user's broad intent or general preference context. Subsequently, **the second stage** then employs a real-time updated OneRec model following (Zhou et al., 2025a) for online finalization. It utilizes the pre-generated item-tokens as a constrained prefix to rapidly produce the final itemic token. This design ensures real-time responsiveness and achieves production-grade performance by leveraging current contextual data. The details of this architecture are in Appendix A.3.4.

## 5 Experiments

### 5.1 Experimental Settings

#### Datasets and Baselines.

We use three real-world recommendation datasets from the popular Amazon review benchmark<sup>1</sup>: Beauty, Toys, and Sports. We compare OneRec-Think against two groups of competitive baselines: (1) Classic sequential methods like BERT4Rec (Sun et al., 2019), HGN (Ma et al., 2019), GRU4Rec (Hidasi et al., 2016), and SAS-Rec (Kang and McAuley, 2018); and (2) Generative Recommender Models, such as TIGER (Rajput et al., 2023), HSTU (Zhai et al., 2024), and ReaRec (Tang et al., 2025). Top-K Recall (R@K) and NDCG (N@K) with K=5 and 10 are used as metrics, following (Rajput et al., 2023). Implementation<sup>2</sup> details are in Appendix A.1.

### 5.2 Overall Performance

The results are shown in Table 1. We could observe that models leveraging powerful reasoning-based architectures (ReaRec and our OneRec-Think) consistently outperform both traditional sequential recommenders and generative recommenders. This robust trend confirms that effective sequential prediction necessitates robust inference and contextual reasoning capabilities. Furthermore, OneRec-Think further gets the best performance across all benchmarks. This significant superiority is directly attributed to explicit, text-based reasoning ability for item generation, in contrast to the more implicit, purely learned generation mechanisms in prior works.

### 5.3 Ablation Study

We conduct an ablation study on the **Beauty** dataset, comparing three configurations: the *Base* model tuned by the raw itemic token sequence, the *Base+IA* model enhanced with Itemic Alignment, and the full *Base+IA+R* model incorporating our enhanced reasoning mechanism, as shown in Table 2. It demonstrates that each component is indispensable: Itemic Alignment provides a foundational boost by creating coherent semantic representations of itemic tokens, while the reasoning mechanism yields a further significant gain, confirming that both components synergistically address core challenges in sequential recommendation.

<sup>1</sup><https://jmcauley.ucsd.edu/data/amazon/>.

<sup>2</sup>code and data are in <https://github.com/wangshy31/OneRec-Think>

Table 1: Overall performance comparison between the baselines and OneRec-Think on three datasets. The bold results highlight the best results, while the second-best ones are underlined.

Dataset	Method	BERT4Rec	HGN	GRU4Rec	SASRec	TIGER	HSTU	ReaRec	OneRec-Think
Beauty	R@5	0.0232	0.0319	0.0395	0.0402	0.0405	0.0424	<u>0.0450</u>	<b>0.0563</b>
	R@10	0.0396	0.0536	0.0584	0.0607	0.0623	0.0652	<u>0.0704</u>	<b>0.0791</b>
	N@5	0.0146	0.0196	0.0265	0.0254	0.0267	<u>0.0280</u>	0.0262	<b>0.0398</b>
	N@10	0.0199	0.0266	0.0326	0.0320	0.0337	<u>0.0353</u>	0.0344	<b>0.0471</b>
Sports	R@5	0.0102	0.0183	0.0190	0.0199	0.0215	<u>0.0268</u>	0.0214	<b>0.0288</b>
	R@10	0.0175	0.0313	0.0312	0.0301	<u>0.0347</u>	0.0343	0.0332	<b>0.0412</b>
	N@5	0.0065	0.0109	0.0122	0.0106	0.0137	<u>0.0173</u>	0.0116	<b>0.0199</b>
	N@10	0.0088	0.0150	0.0161	0.0141	0.0179	<u>0.0226</u>	0.0154	<b>0.0239</b>
Toys	R@5	0.0215	0.0326	0.0330	0.0448	0.0337	0.0366	<u>0.0523</u>	<b>0.0579</b>
	R@10	0.0332	0.0517	0.0490	0.0626	0.0547	0.0566	<u>0.0764</u>	<b>0.0797</b>
	N@5	0.0131	0.0192	0.0228	<u>0.0300</u>	0.0209	0.0245	0.0298	<b>0.0412</b>
	N@10	0.0168	0.0254	0.0279	0.0358	0.0276	0.0309	<u>0.0376</u>	<b>0.0482</b>



Figure 3: Demonstration of context-aware recommendation adaptation: our model dynamically shifts recommendations to relaxing content based on the user’s command.

Table 2: Ablation Study of different variants of OneRec-Think on Beauty dataset.

Training Method	R@5	R@10	N@5	N@10
Base	0.0460	0.0654	0.0314	0.0377
Base+IA	0.0532	0.0735	0.0342	0.0402
Base+IA+R	<b>0.0563</b>	<b>0.0791</b>	<b>0.0398</b>	<b>0.0471</b>

## 5.4 Industrial Experiments

### 5.4.1 Training Settings

We adopt Qwen-8B (Yang et al., 2025) as our backbone model, initializing its parameters from the publicly available pre-trained weights. The model’s vocabulary is extended with 24,576 new tokens representing the three-level hierarchical itemic tokens (8,192 tokens per level), plus two special bound-

ary tokens, <|item\_begin|> and <|item\_end|>. For our production environment, we implement a daily incremental training pipeline. The model is updated each day on a cluster of 80 flagship GPUs, processing approximately 20B tokens per day to stay current with newly generated user interaction data. The details are shown in Appendix A.3.

### 5.4.2 Results

**Online A/B Result.** We deploy OneRec-Think on Kuashou, a short-video platform with hundreds of millions of daily active users. Using a 1.29% traffic experimental group, we compare OneRec-Think with our online model for one week and report the result in Table 3, where the primary metric is APP Stay Time (reflecting total user engagement time). The primary metric, App Stay Time, shows signif-



Figure 4: Demonstration of fine-grained interest reasoning, which shows the end-to-end process from user behavior analysis to interpretable recommendations.

Table 3: The relative improvement of our online A/B testing on a short-video recommendation scenario.

Online Metrics	OneRec-Think
App Stay Time	+0.159%
Watch Time	+0.169%
Video View	+0.150%
Follow	+0.431%
Forward	+0.758%
Like	+0.019%
Collect	+0.098%

icant gains that increase by 0.159%. Note that in industrial recommendation systems, 0.1% improvements are considered substantial. Furthermore, interactive metrics such as Video View and Forward exhibit positive trends, indicating enhanced user engagement. We conducted multiple experiments at different times and consistently observed significant improvements in stay time and related interaction metrics.

Table 4: The BertScore for User Understanding and Short Video Understanding Benchmark.

Benchmark	Qwen3	Qwen3 + TW	Qwen3 + TW + MI
User	0.6588	0.6492	<b>0.7053</b>
Short Video	0.6031	0.6443	<b>0.7300</b>

**Ablation on Itemic Alignment on industrial benchmark** We evaluate Token Warm-up (TW) and Multi-Task Integration (MI) of the Itemic Alignment stage on our industrial User and Short Video Understanding benchmarks using BertScore (Zhang et al., 2019) (details in Appendix A.3.2). Results in Table 4 reveal distinct roles for each component. On the text-heavy User Understanding task, TW provides limited gain over the strong Base model since the LLM can effectively process the abundant textual information directly, while MI delivers a substantial boost by translating aligned representations into actionable insights. In contrast, in the pure itemic token Short Video Understanding task, it shows progressive gains from both TW and MI, confirming their ne-

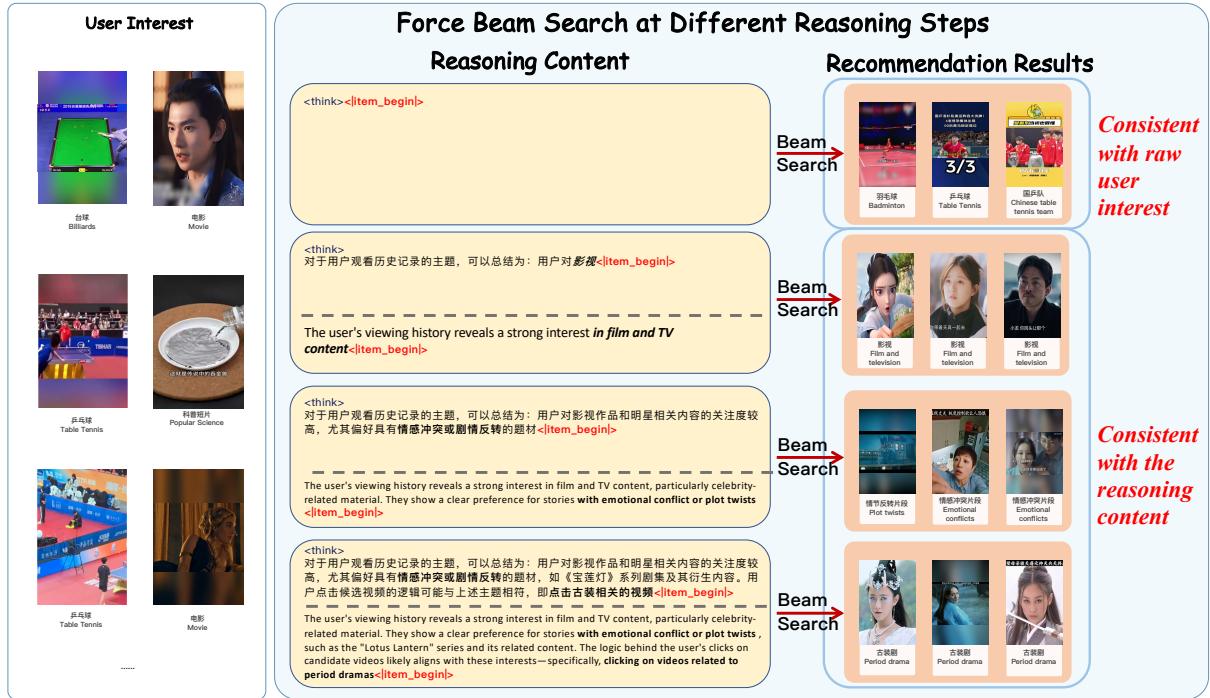


Figure 5: The model’s reasoning process evolves from broad interest matching (left) to fine-grained theme specification (middle), with recommendations (right) showing semantic consistency with each reasoning step.

cessity for interpreting non-textual item information. These results validate that these two substages both contribute to the final performance of Itemic Alignment.

## 5.5 Case Study

Our case studies demonstrate the model’s sophisticated reasoning capabilities across different scenarios. In conversational settings (Fig. 3), when the user expresses negative emotions, the model detects this affective signal and strategically shifts recommendations from general interests toward relaxing and positive content, demonstrating its ability to actively optimize the viewing experience through the interaction with the user. In reasoning-based short-video recommendation (Fig. 4), the model generates diverse reasoning paths that capture fine-grained user preferences, such as specific gameplay mechanics and narrative patterns, enabling more precise recommendations beyond coarse topic matching. Furthermore, our consistency analysis (Fig. 5) reveals strong alignment between reasoning textx and recommended items when applying beam search at intermediate reasoning steps, confirming that the reasoning process genuinely guides recommendation generation rather than serving as post-hoc justification. These results collectively validate our model’s capacity

for authentic, multi-faceted reasoning by demonstrating its ability to adapt to real-time interactions, capture fine-grained preferences, and maintain semantic consistency across diverse recommendation scenarios.

## 6 Conclusion

We present OneRec-Think, a novel framework that bridges reasoning capabilities with generative recommendation through three key innovations: hierarchical itemic token alignment, reasoning activation via CoT supervised fine-tuning, and reinforcement-based reasoning refinement. Our method fundamentally transforms recommendation systems from mere item predictors into reasoning-aware models that generate interpretable rationales alongside high-quality recommendations. Extensive experiments demonstrate that OneRec-Think not only achieves state-of-the-art performance across multiple benchmarks, but also translates to concrete industrial impact with a 0.15% gain in primary metrics like APP Stay Time. Future work will focus on exploring user long-sequence modeling and dense RL reward for finer-grained preference modeling, further bridging LLM-based reasoning with industrial recommendation systems.

## Limitations

Despite promising empirical results, current public datasets exhibit quality constraints through their limited behavior sequence lengths and restricted item spaces. These limitations hinder our Reasoning Activation and Reasoning Enhancement modules from acquiring high-quality reasoning capabilities comparable to those learned from industrial-scale data. Consequently, we simplify and adapt our approach to achieve a stable yet simplified reasoning capacity, which remains robust within the public datasets. To address these issues, we are actively constructing a large-scale benchmark with extended behavioral trajectories and diversified item catalogs that will enable more comprehensive evaluation of reasoning capabilities for reasoning-based recommendation models.

## Ethics Statement

In this work, we have conducted experiments for two settings: one for open-source benchmark datasets and one for the industrial scenario. For the experiments for open-source benchmark datasets, all datasets are publicly available from previous works or public APIs while maintaining anonymity. For the industrial scenario, we utilize user interaction data collected from our platform to train the recommendation model. All data collection and usage strictly comply with our platform’s privacy policy and terms of service, to which users have provided explicit consent. Importantly, our training process operates solely on aggregated behavioral sequences, textual content, and user base information without accessing or processing any personally identifiable information.

## References

- Millennium Bismay, Xiangjue Dong, and James Caverlee. 2024. Reasoningrec: Bridging personalized recommendations and human-interpretable explanations through llm reasoning. *arXiv preprint arXiv:2410.23180*.
- Ben Chen, Xian Guo, Siyuan Wang, Zihan Liang, Yue Lv, Yufei Ma, Xinlong Xiao, Bowen Xue, Xuxin Zhang, Ying Yang, and 1 others. 2025. Onesearch: A preliminary exploration of the unified end-to-end generative framework for e-commerce search. *arXiv preprint arXiv:2509.03236*.
- Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnaud Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 2024. A review of modern recommender systems using generative models (genrecsys). In *Proceedings of the 30th ACM SIGKDD conference on Knowledge Discovery and Data Mining*, pages 6448–6458.
- Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. 2025. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965*.
- Yi Fang, Wenjie Wang, Yang Zhang, Fengbin Zhu, Qifan Wang, Fuli Feng, and Xiangnan He. 2025. Reason4rec: Large language models for recommendation with deliberative user preference alignment. *CoRR*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Xian Guo, Ben Chen, Siyuan Wang, Ying Yang, Chenyi Lei, Yuqing Ding, and Han Li. 2025b. Onesug: The unified end-to-end generative framework for e-commerce query suggestion. *arXiv preprint arXiv:2506.06913*.
- Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *ICLR*. ICLR.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *ICDM*, pages 197–206. IEEE.
- Jieyong Kim, Hyunseo Kim, Hyunjin Cho, SeongKu Kang, Buru Chang, Jinyoung Yeo, and Dongha Lee. 2024. Review-driven personalized preference reasoning with large language models for recommendation. *corr, abs/2408.06276*, 2024. doi: 10.48550. *arXiv preprint ARXIV.2408.06276*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. 2023. Large language models for generative recommendation: A survey and visionary discussions. *arXiv preprint arXiv:2309.01157*.
- Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *KDD*. ACM.

- Qiyao Peng, Hongtao Liu, Hua Huang, Qing Yang, and Minglai Shao. 2025. A survey on llm-powered agents for recommender systems. *arXiv preprint arXiv:2502.10050*.
- Shashank Rajput, Nikhil Mehta, Anima Singh, Raghu-nandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, and 1 others. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315.
- ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, and 1 others. 2025. Seed1. 5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.
- 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 *CIKM*, pages 1441–1450. ACM.
- Jiakai Tang, Sunhao Dai, Teng Shi, Jun Xu, Xu Chen, Wen Chen, Jian Wu, and Yuning Jiang. 2025. Think before recommend: Unleashing the latent reasoning power for sequential recommendation. *arXiv preprint arXiv:2503.22675*.
- Alicia Tsai, Adam Kraft, Long Jin, Chenwei Cai, Anahita Hosseini, Taibai Xu, Zemin Zhang, Lichan Hong, Ed H. Chi, and Xinyang Yi. 2024. Leveraging LLM reasoning enhances personalized recommender systems. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13176–13188, Bangkok, Thailand. Association for Computational Linguistics.
- Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024a. Learnable item tokenization for generative recommendation. In *CIKM*. ACM.
- Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2023. Generative recommendation: Towards next-generation recommender paradigm. *arXiv preprint arXiv:2304.03516*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Yidan Wang, Zhaochun Ren, Weiwei Sun, Jiyuan Yang, Zhixiang Liang, Xin Chen, Ruobing Xie, Su Yan, Xu Zhang, Pengjie Ren, and 1 others. 2024b. Content-based collaborative generation for recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2420–2430.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Zipeng Wei, Kuo Cai, Junda She, Jie Chen, Ming-hao Chen, Yang Zeng, Qiang Luo, Wencong Zeng, Ruiming Tang, Kun Gai, and 1 others. 2025. One-loc: Geo-aware generative recommender systems for local life service. *arXiv preprint arXiv:2508.14646*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Bin Yuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL <https://arxiv.org/abs/2305.10601>, 3:1.
- Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.
- Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, and 1 others. 2025. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*.
- Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, and 1 others. 2024. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152*.
- Junjie Zhang, Beichen Zhang, Wenqi Sun, Hongyu Lu, Wayne Xin Zhao, Yu Chen, and Ji-Rong Wen. 2025a. Slow thinking for sequential recommendation. *arXiv preprint arXiv:2504.09627*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yu Zhang, Shutong Qiao, Jiaqi Zhang, Tzu-Heng Lin, Chen Gao, and Yong Li. 2025b. A survey of large language model empowered agents for recommendation and search: Towards next-generation information retrieval. *arXiv preprint arXiv:2503.05659*.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *ICDE*. IEEE.

Guorui Zhou, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Shiyao Wang, and 1 others. 2025a. Onerec technical report. *arXiv preprint arXiv:2506.13695*.

Guorui Zhou, Hengrui Hu, Hongtao Cheng, Huanjie Wang, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Lu Ren, Liao Yu, and 1 others. 2025b. Onerec-v2 technical report. *arXiv preprint arXiv:2508.20900*.

## A Appendix

### A.1 Experiment settings

**Details of Baselines** We compare OneRec-Think with competitive baselines within two groups of work, traditional recommender models and generative recommender models: 1) **BERT4Rec** (Sun et al., 2019) leverages BERT’s pre-trained language representations to capture semantic user-item relationships. 2) **HGN** (Ma et al., 2019) utilizes graph neural networks to learn user and item representations for predicting user-item interactions. 3) **GRU4Rec** (Hidasi et al., 2016) is a lightweight graph convolutional network model focusing on high-order connections between users and items. 4) **SASRec** (Kang and McAuley, 2018) employs self-attention mechanisms to capture long-term dependencies in user interaction history. 5) **TIGER** (Rajput et al., 2023) introduces codebook-based identifiers via RQ-VAE, which quantizes semantic information into code sequences for LLM-based generative recommendation. 6) **HSTU** (Zhai et al., 2024) reformulates recommendation problems as sequential transduction tasks within a generative modeling framework and proposes a new architecture for streaming data. 7) **ReaRec** (Tang et al., 2025) enhances user representations through implicit multi-step reasoning within an inference-time computing framework for recommendation. **Evaluation Metrics.** We use two metrics: top- $K$  Recall

(R@ $K$ ) and NDCG (N@ $K$ ) with  $K = 5$  and 10, following (Rajput et al., 2023).

### Details of experiments on open-source datasets

We adopt Qwen3-1.7B (Yang et al., 2025) as our backbone model. The model’s vocabulary is extended with 1,024 new tokens representing the four-level hierarchical semantic IDs (256 tokens per level), in addition to two special boundary tokens, <|item\_begin|> and <|item\_end|>. All models were trained on a server equipped with flagship GPUs. To generate the top-K recommendations during evaluation, we employ a beam search strategy with a beam width of 10. Given the inherent challenge of deriving a robust reasoning path  $\tau$  from the short and sparse item sequences typical of public benchmarks, we strategically employ manually constructed category-based CoT as the pruned content for Reasoning Activation to ensure stable semantic guidance. For the training data, we adopt the pre-processing techniques from previous work (Rajput et al., 2023; Wang et al., 2024a), discarding sparse users and items with interactions less than 5. We consider the sequential recommendation setting and use *leave-one-out* strategy (Rajput et al., 2023; Zheng et al., 2024) to split datasets. For training, we follow (Kang and McAuley, 2018) to restrict the number of items in a user’s history to 50.

### A.2 Model Demonstration

**Reasoning Cases for Short-video Recommendation** In this part, we show some cases of short video recommendations of another user. Figure 6 presents two reasoning paths generated by our model, demonstrating its capacity for multi-step interest inference and underlying need identification. In the first case, the model connects the user’s gaming preferences with hardware comparison behaviors to deduce an unstated need for performance optimization, ultimately recommending monitor analysis videos. The second case reveals deeper psychological needs by associating sports/military viewing history with adolescent rebellion searches, identifying parenting challenges as the core concern. Both examples showcase our model’s ability to transcend superficial topic matching and perform causal reasoning about user motivations, enabling recommendations that address both expressed interests and latent needs.

**Semantic Comprehension Validation** After semantic alignment, the model acquires the capability

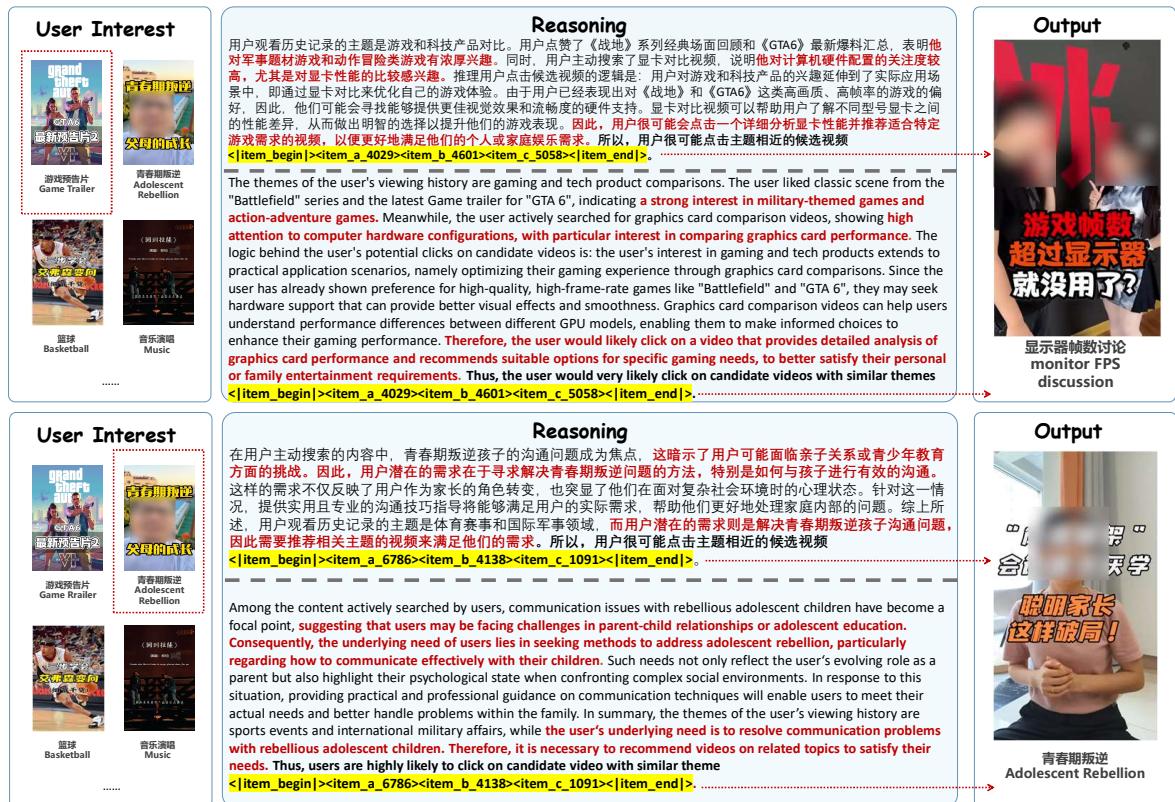


Figure 6: A reasoning example for a short-video recommendation scenario.

to comprehend and articulate the semantic meaning of item tokens through natural language. To validate this emergent ability, we evaluate the capacity of the model to generate descriptive captions for itemic tokens without explicit training on this task. When prompted to explain what an itemic token represents, the model leverages the learned semantic correspondences to produce coherent textual descriptions that accurately capture the characteristics of the items. This demonstrates that the alignment process successfully establishes genuine semantic understanding rather than superficial pattern matching, as the model can now bidirectionally translate between discrete itemic tokens and their rich natural language semantics. The following cases in Beauty dataset showcase the model’s caption generation capability for item semantic tokens:

#### Case: Itemic Token Explaining 1

##### Question + Answer

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
Can you explain the meaning of <|
item_begin|><item_a_219><item_b_189><
item_c_49><item_d_227><|item_end|>?<|
im_end|>
```

```
<|im_start|>assistant
Its title is Manuka Honey and Manuka Oil Skin Cream, its category is Beauty > Skin Care > Face > Cleansers.
```

#### Ground Truth

```
"title": "Manuka Honey and Manuka Oil Skin Cream"
"categories": "Beauty > Skin Care > Face > Creams & Moisturizers > Fluids & Lotions"
"sid": "<|item_begin|><item_a_219><item_b_189><item_c_49><item_d_227><|item_end|>"
```

#### Case: Itemic Token Explaining 2

##### Question + Answer

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
Can you explain the meaning of <|
item_begin|><item_a_8><item_b_18><
item_c_86><item_d_131><|item_end|>?<|
im_end|>
<|im_start|>assistant
Its title is 25 Color Professional Camouflage Makeup Palette, its category is Beauty > Makeup > Face > Concealer & Base.
```

## Ground Truth

```
"title": "Professional 15 Color Concealer  
Camouflage Makeup Palette",  
"categories": "Beauty > Makeup > Face >  
Concealers & Neutralizers",  
"sid": "<|item_begin|><item_a_8><  
item_b_18><item_c_86><item_d_131|><|  
item_end|>"
```

critics specializing in sci-fi.

### # User Summary

**Primary Interests:** This user enjoys science, especially astronomy. This user also engages with Honor of Kings content, indicating a casual gaming interest.

**Secondary Interests:** Diverse explorations include pet (cat) videos, traditional culture, and local food content.

## A.3 Implementation Details

### A.3.1 Itemic Alignment

**Task details** Here, we first introduce the aforementioned four types of tasks.

**1. Interleaved User Persona Grounding.** This part contains the interleaved itemic tokens with rich, natural-language text extracted from static user profiles. This process forces the model to create a robust mapping between the itemic tokens and their real-world meanings, grounded in user attributes, stated interests, and historical behaviors. The sample data is shown below.

#### Interleaved User Persona Grounding data

```
# User Profile Narrative  
The user is a 25-30 year-old male based  
in Beijing.  
  
# User Profile Narrative  
The user's recent searches on the  
platform include: "best space opera  
novels," and "latest NASA discoveries."  
  
# Live Stream Behavior  
He recently commented 5 times on live  
streams in the "Science & Tech" category.  
  
# Like Behavior  
He recently liked video <|item_begin|><  
item_a_1123><item_b_5813><item_c_4212|><  
item_end|>, captioned "Exploring the  
Andromeda Galaxy with the James Webb  
Telescope."; and video <|item_begin|><  
item_a_3421><item_b_8812><item_c_1234|><  
item_end|>, captioned "Top 10 Paradoxes  
of Time Travel."  
...  
  
# Comment Behavior  
He recently commented "Incredible footage  
!" on Video: <|item_begin|><item_a_5813|><  
item_b_1123><item_c_9876|><|item_end|>, a  
documentary about black holes; and  
commented "Mind-blowing concept!" on  
Video: <|item_begin|><item_a_8812|><  
item_b_3421><item_c_5432|><|item_end|>,  
explaining the Fermi Paradox.  
...  
  
# Followed Creators  
He follows creators on the platform  
across various fields, including science  
popularizers, book reviewers, and film
```

**2. Sequential Preference Modeling.** The model is given a sequence of a user's recent interactions (up to 256 items) and is tasked with predicting the next item. The loss is computed only on the tokens of the target itemic token. The sample data is shown below.

#### Sequential Preference Modeling data

```
<|im_start|>system  
You are a sequential recommendation  
engine. Your task is to analyze the  
provided sequence of user-item  
interactions and predict the single next  
item the user is most likely to engage  
with.  
<|im_end|>  
<|im_start|>user  
User interaction history: <|item_begin|><  
item_a_1024><item_b_2048><item_c_4096|><  
item_end|>...<|item_begin|><item_a_1234><  
item_b_5678><item_c_5876|><|item_end|>.  
Predict the next item.  
<|im_end|>  
<|im_start|>assistant  
The next recommended item is <|item_begin|><  
item_a_5555><item_b_6666><item_c_7777|><  
item_end|>.  
<|im_end|>
```

**3. Itemic Dense Captioning.** This task asks the model to generate a textural description of a given itemic token, ensuring a robust mapping between itemic tokens and natural language. The loss is computed on the tokens of the target textual description. The sample data is shown below.

#### Itemic Dense Captioning data

```
<|im_start|>system  
You are an expert content analyst. Given  
an itemic token, generate a concise and  
accurate textual description of its  
content.  
<|im_end|>  
<|im_start|>user  
Provide a description for the itemic  
token: <|item_begin|><item_a_1357|><  
item_b_2468><item_c_7753|><|item_end|>.  
<|im_end|>  
<|im_start|>assistant  
This video showcases a local food stall,  
highlighting traditional cooking methods  
and signature dishes like grilled skewers
```

Table 5: Data details for the Recommendation Enhancing.

Task Type	Data Percentage
Interleaved User Persona Grounding	24.30%
Sequential Preference Modeling	65.73%
Item Dense Captioning	4.94%
General Language Modeling	5.03%

to attract customers by celebrating local food culture.  
<|im\_end|>

**4. General Language Modeling.** This part of the data contains the pretraining and instruction fine-tuning data of the general corpus to maintain the base ability of the language model.

To prepare our model for in-text reasoning recommendations, we design a two-stage training curriculum to first build a foundational semantic understanding and then develop sophisticated recommendation capabilities. The initial substage, **Token Warm-up**, focuses on aligning itemic tokens with their rich textual meanings. The second substage, **Multi-Task Integration**, develops the model’s core recommendation abilities while preserving the initial semantic alignment.

**Token Warm-up** The initial and foundational step is to imbue the model with a deep understanding of the semantic content represented by each itemic tokens ( $s_v$ ). In the open-source benchmark datasets, the dataset is consist of the user profile of the full training data. In the industrial setting, the entire alignment process leverages a corpus of 6 billion tokens of the Interleaved User Persona Grounding task, ensuring a comprehensive semantic grounding.

During this stage, we freeze the parameters of the LLM backbone and exclusively train the embeddings of the newly introduced itemic tokens. This focused approach allows the randomly initialized embeddings to efficiently converge to meaningful positions within the model’s existing semantic space without disrupting its pre-trained knowledge. We use a higher learning rate of  $5 \times 10^{-4}$  to train only the new token embeddings, which allows the randomly initialized embeddings to converge quickly and efficiently.

**Multi-Task Integration** With the semantic item vocabulary successfully grounded, the second substage of our curriculum aims to develop the model’s

core recommendation capabilities. Simply training on a single objective, such as next-item prediction, could cause the model to gradually treat the itemic tokens as conventional, non-semantic identifiers, thereby losing the benefits of the initial grounding substage. To prevent this and foster a more holistic understanding, we employ the Multi-Task Integration. The mixture of different tasks, detailed in Table 5, is designed to balance the learning of collaborative patterns with the reinforcement of semantic understanding and contextual reasoning.

For the open-source benchmark dataset, we train the full parameters of the model. For the industrial scenario, we unfreeze the model backbone and utilize LoRA (Hu et al., 2022) for parameter-efficient fine-tuning. We employ a more conservative learning rate of  $3 \times 10^{-4}$  and fine-tune the model, which ensures a stable adaptation of the model’s internal mechanisms to the recommendation tasks without disrupting its foundational knowledge.

### A.3.2 Reasoning Activation

The Reasoning Activation stage employs a carefully orchestrated two-substage training pipeline designed to progressively develop the model’s reasoning capabilities from simplified contexts to complex, noisy industrial sequences.

**Bootstrapping with Pruned Contexts:** In this substage, we construct a high-quality reasoning dataset by strategically pruning user behavior sequences to retain only the most relevant interactions. For each training instance, we employ a semantic similarity function  $g(\cdot, \cdot)$  to extract the top- $k$  items (we set  $k = 10$  in our experiments) most relevant to the target interaction. The similarity function leverages pre-trained item embeddings to compute cosine similarity scores between historical items and the target. Using these pruned sequences, we prompt the semantically aligned model to generate step-by-step reasoning trajectories that explain why the target item aligns with the user’s demonstrated preferences. The example is shown below.

#### Rationale Generation from Pruned Context

<|im\_start|>system  
You are an expert recommendation analyst.  
Given a user’s relevant interaction history, provide a step-by-step reasoning process explaining why a specific target item aligns with their preferences.  
<|im\_end|>  
<|im\_start|>user

User's relevant interaction history:

```
<|item_begin|><item_a_1111><item_b_2222><item_c_3333><|item_end|> - Beginner's guide to sourdough bread making  
<|item_begin|><item_a_2947><item_b_8156><item_c_3402><|item_end|> - Professional chef knife skills tutorial  
<|item_begin|><item_a_5618><item_b_3927><item_c_7084><|item_end|> - Italian pasta recipes from scratch  
<|item_begin|><item_a_7293><item_b_5471><item_c_1836><|item_end|> - Kitchen organization and meal prep tips  
<|item_begin|><item_a_9105><item_b_6248><item_c_4519><|item_end|> - Comparison of stand mixers for home baking
```

Target item: <|item\_begin|><item\_a\_3672><item\_b\_9814><item\_c\_2657><|item\_end|> - Advanced French pastry techniques and equipment guide

Explain why this target item matches the user's interests.

<|im\_end|>

<|im\_start|>assistant

Reasoning: The user's viewing history shows a strong interest in advanced cooking techniques and kitchen equipment, with engagement in sourdough baking, handmade pasta, and stand mixer comparisons. The user has progressed from basic skills to time-intensive culinary projects, demonstrating commitment to mastering traditional cooking methods. French pastry represents a natural next step that combines their baking experience with technical precision. Therefore, the user would very likely click on candidate videos with similar themes <|item\_begin|><item\_a\_3672><item\_b\_9814><item\_c\_2657><|item\_end|>.

## Learning to Reason from Noisy Sequences

Building upon the distilled reasoning patterns, this substage trains the model to transfer its reasoning capabilities to raw, unfiltered user behavior sequences. The training data consists of complete user histories paired with the high-quality rationales generated in Substage 1. We employ the joint optimization objective defined in Equation 5, which simultaneously trains the model to: (1) generate coherent reasoning trajectories by predicting rationale tokens conditioned on noisy context, and (2) produce accurate target item predictions informed by both the context and generated rationale. The loss function applies equal weighting to both components, ensuring the model learns to perform contextual distillation while maintaining recommendation accuracy. The example is shown below.

### Reasoning from Noisy Sequences

<|im\_start|>system

You are an intelligent recommendation system. Given a user's complete interaction history, provide reasoning for your recommendation and then suggest the most appropriate next item.

<|im\_end|>

<|im\_start|>user

User's complete interaction history:

```
<|item_begin|><item_a_1847><item_b_6392><item_c_5104><|item_end|>, <|item_begin|><item_a_9263><item_b_4571><item_c_8026><|item_end|>, <|item_begin|><item_a_1111><item_b_2222><item_c_3333><|item_end|>, <|item_begin|><item_a_5419><item_b_8736><item_c_2948><|item_end|>, <|item_begin|><item_a_3082><item_b_7145><item_c_6379><|item_end|>, <|item_begin|><item_a_2947><item_b_8156><item_c_3402><|item_end|>, <|item_begin|><item_a_7651><item_b_2493><item_c_1087><|item_end|>, <|item_begin|><item_a_5618><item_b_3927><item_c_7084><|item_end|>, <|item_begin|><item_a_8304><item_b_9672><item_c_4215><|item_end|>, <|item_begin|><item_a_7293><item_b_5471><item_c_1836><|item_end|>, <|item_begin|><item_a_4926><item_b_1358><item_c_8740><|item_end|>, <|item_begin|><item_a_9105><item_b_6248><item_c_4519><|item_end|>, <|item_begin|><item_a_1537><item_b_8904><item_c_3261><|item_end|>
```

What item should be recommended next?

Explain your reasoning first.

<|im\_end|>

<|im\_start|>assistant

<think>

The user's viewing history shows a strong interest in advanced cooking techniques and kitchen equipment, with engagement in sourdough baking, handmade pasta, and stand mixer comparisons. The user has progressed from basic skills to time-intensive culinary projects, demonstrating commitment to mastering traditional cooking methods. French pastry represents a natural next step that combines their baking experience with technical precision. Therefore, the user would very likely click on candidate videos with similar themes <|item\_begin|><item\_a\_3672><item\_b\_9814><item\_c\_2657><|item\_end|>.

</think>

Recommendation: <|item\_begin|><item\_a\_3672><item\_b\_9814><item\_c\_2657><|item\_end|>

<|im\_end|>

Similarly, for these tasks, we train the full parameters of the model. For the industrial scenario, we unfreeze the model backbone and utilize LoRA for parameter-efficient fine-tuning. We employ a learning rate of  $3 \times 10^{-4}$  and fine-tune the model.

To quantify the impact of Semantic Grounding (SG) and Recommendation Enhancement (RE)

in Itemic Alignment stage, we calculate the BertScore (Zhang et al., 2019) between the ground truth and prediction of OneRec-Think on the User Understanding and Short Video Understanding benchmarks. The cases of User and Short Video Understanding of Table 4 are shown below, and the ground truth is the accurate answer from our online multi-modal large language model.

#### Case for Short Video Understanding Benchmark

```
<|im_start|>system
You are an expert content analyst. Given
an itemic token, generate a concise and
accurate textual description of its
content.
<|im_end|>
<|im_start|>user
Provide a description for the itemic
token: <|item_begin|><item_a_1357><
item_b_2468><item_c_7753><|item_end|>.
<|im_end|>
<|im_start|>assistant
# Answers
```

#### Case for User Understanding Benchmark

```
<|im_start|>system
You are an expert content analyst. Your
task is to analyze the provided user
interaction history and generate a
structured, actionable user profile
summary..
<|im_end|>
<|im_start|>user
<|item_begin|><item_a_1111><item_b_2222><
item_c_3333><|item_end|> - Beginner's
guide to sourdough bread making
<|item_begin|><item_a_2947><item_b_8156><
item_c_3402><|item_end|> - Professional
chef knife skills tutorial
<|item_begin|><item_a_5618><item_b_3927><
item_c_7084><|item_end|> - Italian pasta
recipes from scratch
<|item_begin|><item_a_7293><item_b_5471><
item_c_1836><|item_end|> - Kitchen
organization and meal prep tips
<|item_begin|><item_a_9105><item_b_6248><
item_c_4519><|item_end|> - Comparison of
stand mixers for home baking
...
<|im_end|>
<|im_start|>assistant
# Answers
```

### A.3.3 Reasoning Enhancement

Building upon the reasoning capabilities established in the previous stage, we employ Reinforcement Learning to further refine both reasoning coherence and recommendation accuracy. We leverage the VERL framework (Sheng et al., 2024) for efficient optimization, which provides robust infrastructure for distributed training and scalable reward computation. Our training utilizes the GRPO

algorithm and implements the Rollout-Beam reward within VERL. During training, we sample  $|G| = 16$  CoT paths and do beam search with width  $K = 32$  for each path to compute rewards, balancing exploration breadth with computational efficiency. We train for 2 epochs with a learning rate of 1e-5, KL divergence coefficient  $\beta = 0.001$ , clip ratio  $\epsilon = 0.2$  to ensure stable policy updates.

### A.3.4 System Deployment: A "Think-Ahead" Architecture

Here, we show the details of the "Think-Ahead" Architecture. Our "Think-Ahead" architecture decouples inference into an offline stage that generates reasoning paths and initial item tokens to capture broad user intent, followed by an online stage that utilizes these tokens as constrained prefixes for real-time finalization, ensuring production-grade latency.

#### Stage 1: Reasoning-Guided Prefix Generation

**Reasoning Path Synthesis:** For user  $u$  with interaction history  $H_u = \mathcal{P}(\mathbf{s}_{v_1}, \dots, \mathbf{s}_{v_n})$ , we sample  $T$  diverse reasoning paths:

$$\tau^{(i)} \sim P(\cdot | H_u; \theta), \quad (8)$$

where each path  $\tau^{(i)}$  for  $i \in \{1, \dots, T\}$  encapsulates a distinct logical pathway connecting the user's behavioral patterns to potential interests.

**Constrained Prefix Generation:** Given our hierarchical itemic tokens with  $L$  identifiers, we perform beam search to generate candidate prefixes. Here, in our industrial setting, we use  $L = 3$ . For each reasoning path  $\tau^{(i)}$ , we decode *only the first two itemic tokens*, yielding:

$$\mathcal{A}_u^{(i)} = \text{BeamSearch}\left(P(\hat{\mathbf{s}}_{v_{n+1}}^1, \hat{\mathbf{s}}_{v_{n+1}}^2 | H_u, \tau^{(i)}; \theta), m\right) \quad (9)$$

where  $\mathcal{A}_u^{(i)}$  denotes the set of  $m$  candidate item prefixes derived from path  $\tau^{(i)}$  for user  $u$ .

**Semantic Space Materialization:** The union of all item prefix sets forms the user's personalized candidate space:

$$\mathcal{C}_u = \bigcup_{i=1}^T \mathcal{A}_u^{(i)} \quad (10)$$

This set, containing  $T \times m$  high-potential item prefixes, is cached in an industrial distributed storage system, effectively materializing the model's deliberative reasoning into actionable semantic hypotheses.

**Stage 2: Prefix-Constrained Finalization** The online stage executes rapid inference while leveraging the pre-computed semantic priors from Stage 1 upon user request arrival.

**Constrained Decoding:** Upon receiving a request for user  $u$ , we retrieve  $\mathcal{C}_u$  and infer the last token using a real-time updated OneRec following (Zhou et al., 2025a)  $h_{\text{online}}$ . The search space is restricted so that the prefix of the candidate target item  $(\hat{s}_{v_{n+1}}^1, \hat{s}_{v_{n+1}}^2)$  should be in the candidate prefix set, which means  $(\hat{s}_{v_{n+1}}^1, \hat{s}_{v_{n+1}}^2) \in \mathcal{C}_u$ . Consequently, we could construct the decoding target:

$$\begin{aligned} \hat{s}_{v_{n+1}} = \arg \max_{\mathbf{s}_{v_{n+1}}} P_{h_{\text{online}}}(\mathbf{s}_{v_{n+1}} \mid \mathbf{s}_{v_1}, \dots, \mathbf{s}_{v_n}) \\ \text{s.t. } (\hat{s}_{v_{n+1}}^1, \hat{s}_{v_{n+1}}^2) \in \mathcal{C}_u \end{aligned} \quad (11)$$

where the items with top-K probability are then served to the users.

In summary, our *Think-Ahead* architecture presents a novel paradigm for industrial recommendation, effectively reconciling sophisticated reasoning with strict latency requirements.