Multi-granularity Intent Modeling with Adversarial Robustness for Sequential Recommendation

Yangyi Fang[1], **Haolin Shi**[1*]
[1]Tsinghua University
yangyi.fang06@gmail.com, shihaolin0720@gmail.com

Abstract

User purchase decisions are driven by complex, multi-faceted intentions that evolve across different temporal horizons (e.g., immediate needs, transitional interests, and long-term preferences). However, existing sequential methods often treat user sequences as unified blocks, overlooking the dynamic evolution of user intents at different granularities, while also lacking robustness against prevalent noise in real-world interaction data. This paper proposes Multi-granularity Intent Modeling with Adversarial Robustness for Sequential Recommendation (MIMAR-SRec), a framework that models latent user intentions at multiple granularities. Specifically, MIMAR-SRec integrates multi-granularity intent representation across different contextual windows to capture evolving user interests, dual-perspective contrastive learning that aligns user representations with both intent prototypes and cross-user sequences, and intent-similarity adversarial robustness that systematically enhances model stability against interaction, temporal, and preference noise through controlled perturbations. By integrating multi-granularity intent modeling with adversarial training, MIMAR-SRec enables simultaneous fine-grained underlying intent modeling and noise-resistant recommendations. Extensive experiments on four widely used benchmark datasets demonstrate that MIMAR-SRec outperforms baselines, particularly in long-tail item recommendation and noisy interaction scenarios.

Introduction

Recommendation systems have emerged as essential tools for navigating the overwhelming volume of information on digital platforms, with their effectiveness hinging on accurately capturing users' evolving preferences. Sequential Recommendation (SR), which predicts future interactions by modeling temporal behavior patterns, has gained significant traction in addressing this challenge (Fan et al. 2022; Lin et al. 2024a,b). The field has evolved from early Markov chain approaches that captured basic sequential dependencies (Cheng et al. 2013) to sophisticated deep learning architectures like CNNs (Tang and Wang 2018), RNNs (Hidasi et al. 2015; Wu et al. 2017), and attention mechanisms (Fan et al. 2022) that extract features, including localized patterns, temporal dynamics, and relative item importance. Despite these advances, a fundamental limitation persists: These methods often fail to develop sufficiently intent-aware user representations, particularly when training data is sparse (Li et al. 2023). This has led to a shift toward self-supervised learning paradigms, where contrast learning (CL) has proven particularly effective by generating diverse views through data augmentation and reinforcing semantic consistency through contrastive objectives (Chen et al. 2022).

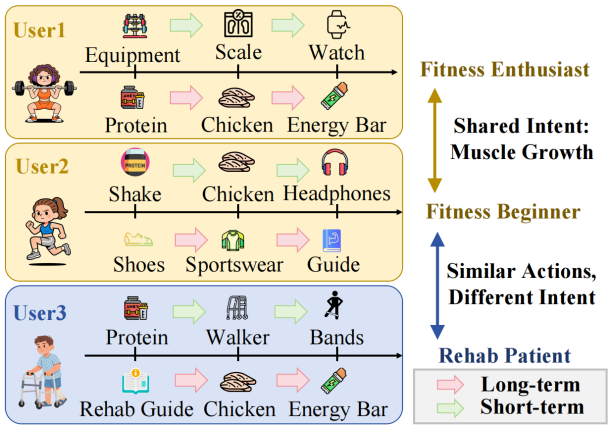

Figure 1: Consumer behavior, as shown, is driven by latent intentions. Despite both User1 and User2's apparent interest in healthy food, User1's purchases suggest a goal of muscle gain (high protein), while User2's indicate a need for blood sugar management (low sugar).

Figure 1 illustrates this problem: when two users interact with similar health-related products, conventional systems might simplistically classify both as "health enthusiasts" and recommend generic items. This surface-level analysis misses crucial intent differences—User 1's behavior indicates muscle-building goals (seeking high-protein options), while User 2's suggests blood sugar management (focusing on low-glycemic foods).

To effectively capture users' multi-granularity intents and enhance robustness against behavioral noise, we propose a novel framework named MIMAR-SRec (Multi-granularity Intent Modeling with Adversarial Robustness for Sequential Recommendation). We first leverage Large Language

*Corresponding author.

Models (LLMs) to enrich item representations with semantic features from textual descriptions, then dynamically segment user historical interaction sequences into multiple subsequences across different temporal granularities, each representing a coarse-grained intent at different interaction contexts. These multi-granularity intent representations serve as units for dual-perspective contrastive learning, where we design intent alignment modules that connect user representations with both intent prototypes and cross-user behavioral patterns. Specifically, the multi-granularity intent alignment module brings closer the similar intent representations across different granularity levels, while the cross-user target-aware contrast module leverages behavioral outcome similarities to identify latent intent correlations across different users. To enhance model robustness, we propose an adversarial learning framework motivated by the Actor-Critic paradigm in reinforcement learning. Our design employs an Adversarial Augmentor to generate semantically consistent perturbations and an Intent Critic to evaluate perturbation quality, achieving adaptive augmentation through a two-phase training strategy.

To address these challenges, we propose MIMAR-SRec, a novel framework with three key contributions:

- An adaptive contextual segmentation mechanism that captures user intents across multiple temporal granularities through hierarchical sequence partitioning and clustering-based prototype discovery.
- A contrastive learning framework that aligns user representations with both intent prototypes and cross-user behavioral patterns, enabling comprehensive intent modeling from complementary perspectives.
- An adversarial learning approach with an Adversarial Augmentor and Intent Critic that enhances model robustness against interaction, temporal, and preference noise while preserving semantic intent consistency.

Related Work

Sequential Recommendation

Sequential recommendation focuses on capturing users' dynamic interests by modeling historical behavior sequences (Zheng et al. 2022; Li et al. 2021; Chen et al. 2021; Kang and McAuley 2018). Early methods relied on Markov chains (Cheng et al. 2013) to model sequential transitions between items based on probabilistic state transitions. FPMC (Rendle, Freudenthaler, and Schmidt-Thieme 2010) combined Markov chains with matrix factorization to integrate both sequential patterns and individual user preferences through factorized personalized transition matrices.

Deep learning approaches have since dominated this field, including RNNs (Wu et al. 2017; Hidasi et al. 2015) and CNNs (Tang and Wang 2018). Transformer-based architectures (Vaswani et al. 2017) have inspired models like SASRec (Kang and McAuley 2018), which employs unidirectional self-attention, and BERT4Rec (Sun et al. 2019), which uses bidirectional transformers with masked item prediction. Recent innovations include LSAN (Li et al. 2021) with time-aware contextual embeddings and STOSA (Fan et al. 2022) with Wasserstein self-attention. However, these methods often struggle with noisy user behavior data, lacking explicit mechanisms to distinguish between meaningful interactions and random noise.

Contrastive Self-Supervised Learning

Self-supervised learning has achieved significant success in computer vision (Chen et al. 2020; He et al. 2020) and natural language processing (Gao, Yao, and Chen 2021; Lan et al. 2019; Mnih and Kavukcuoglu 2013) through contrastive learning paradigms. These successful paradigms are increasingly being adopted in recommendation systems (Chuang et al. 2020; Qiu et al. 2022).

In sequential recommendations, various contrastive learning frameworks have been developed with diverse augmentation and optimization strategies. SGL (Wu et al. 2021), S^3-Rec (Zhou et al. 2020), and CL4SRec (Xie et al. 2022) employ contrastive strategies that range from explicit data augmentation techniques—including sequence cropping, masking, and reordering—to mutual information maximization between differently augmented sequence views. These methods construct positive pairs through stochastic data transformations while maintaining semantic consistency. In contrast, DuoRec (Qiu et al. 2022) and REDA (Bian et al. 2022) adopt an alternative paradigm by constructing contrastive views at the model representation level through stochastic encoding and recency-based sampling, bypassing explicit data manipulation. Recent advances have begun integrating adversarial training mechanisms with contrastive learning objectives. AdvCL (Fan et al. 2021) and ACVAE (Xie et al. 2021) demonstrate the effectiveness of applying adversarial perturbations and variational inference to generate robust and discriminative contrastive representations across different application domains.

User Intent for Recommendation

In sequential recommendation, user intent modeling has evolved from simple to sophisticated approaches. DSS-Rec (Ma et al. 2020) introduced a seq2seq-based training paradigm that captures the mutual information relationship between users' historical and future behavior sequences by incorporating intent variables. ICLRec (Chen et al. 2022) further integrated sequence clustering techniques to construct intent prototypes, leveraging random data augmentation to generate positive views for contrastive learning. To enhance the quality of intent representations, IOCRec (Li et al. 2023) improved upon ICLRec by introducing global and local modules to model intents while mitigating noise issues in contrastive learning tasks. Similarly, ICSRec (Qin et al. 2024) designed a supervision mechanism based on user interaction signals to achieve more precise intent representation learning.

Our approach differs by introducing multi-granularity intent contrastive learning across different temporal scales and adversarial robustness mechanisms guided by intent-similarity constraints, addressing both the hierarchical structure of user intents and model stability against behavioral noise.

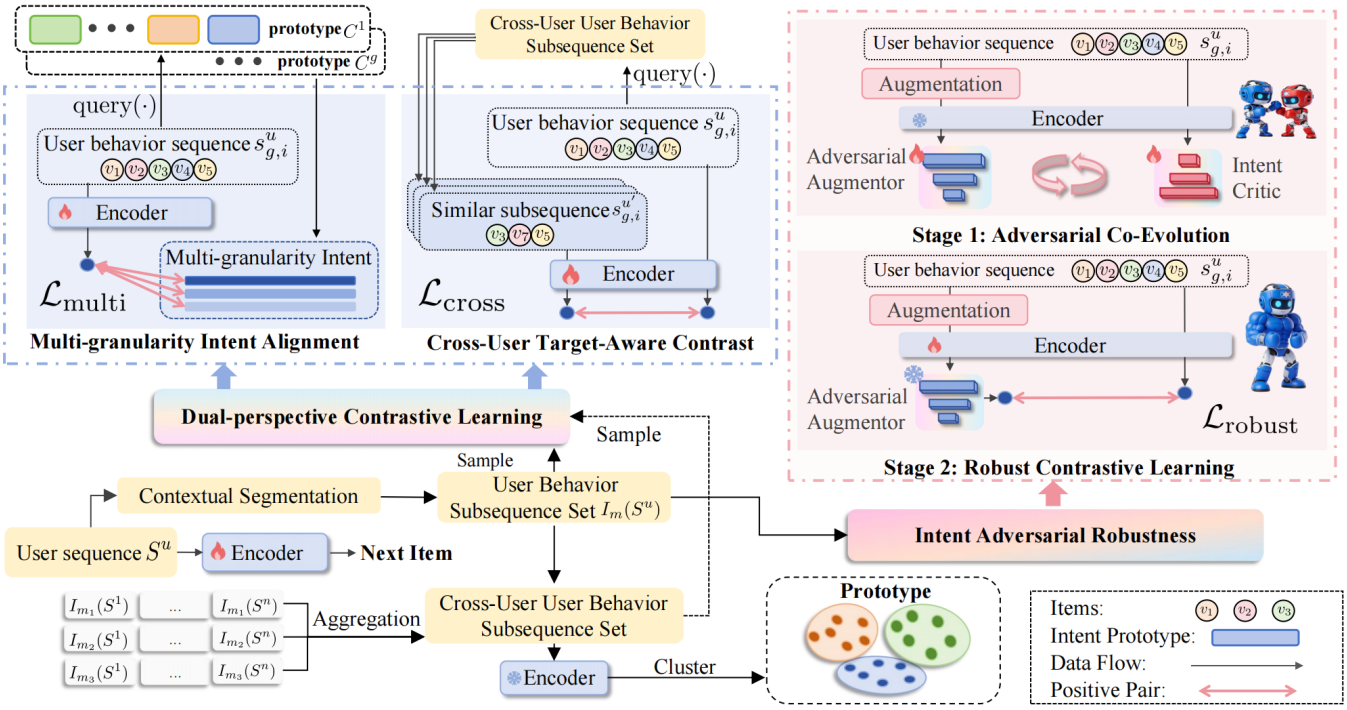

Figure 2: The architecture of MIMAR-SRec. User behavior sequences S^u are adaptively partitioned into multiple subsequences, which serve as multi-granularity intent representations for robust modeling.

Method

In this section, we first formalize the sequential recommendation problem. Let the sets of users and items be denoted by U and V, respectively. For any user $u \in U$, their behavior can be described by an ordered sequence of items, denoted as $S^u = \{v_1^u, v_2^u, \ldots, v_{|S^u|}^u\}$, where $|S^u|$ represents the length of the sequence, and $v_t^u \in V$ $(1 \le t \le |S^u|)$ indicates the item with which user u interacted at position t. The core task is to predict the next item $v_{|S^u|+1}^u$ that the user is likely to interact with, formalized as: $\arg\max_{v_i \in V} P(v_{|S^u|+1} = v_i \mid S^u)$.

We present MIMAR-SRec, a novel framework for sequential recommendation that effectively captures user intents at multiple granularities while maintaining robustness against noise. Figure 2 illustrates the overall architecture of MIMAR-SRec, which consists of three primary components: (1) Intent Representation through multi-granularity contextual segmentation, (2) Intent Alignment via dual-perspective contrastive learning for capturing intent patterns from different perspectives, and (3) Intent Adversarial Robustness for enhancing model stability. These components work in concert to effectively learn comprehensive user intent representations that are both fine-grained and resilient to perturbations.

Multi-granularity Contextual Segmentation

User behavior sequences inherently exhibit multi-level temporal patterns, ranging from short-term interactions to long-term preference evolution. To effectively model this multi-granularity structure, we segment user interaction sequences into contextually coherent units across multiple temporal scales. Given a user interaction sequence $S^u = \{v_1^u, v_2^u, \ldots, v_{|S^u|}^u\}$, where v_t^u represents the item with which user u interacted at time step t, we construct a multi-granularity intent representation set $I(S^u)$ through an adaptive contextual segmentation approach:

$$I_m(S^u) = \begin{cases} \{\{v_1^u, v_2^u\}, \{v_1^u, v_2^u, v_3^u\}, \\ \ldots, \{v_1^u, v_2^u, \ldots, v_{|S^u|}^u\}\} & |S^u| \le m \\ \\ I_m(S^u_{1:m}) \cup \{\{v_2^u, v_3^u, \ldots, v_{m+1}^u\}, \\ \ldots, \{v_{|S^u|-m+1}^u, \ldots, v_{|S^u|}^u\}\} & |S^u| > m, \end{cases} \tag{1}$$

where $S^u_{1:m} = \{v_1^u, v_2^u, \ldots, v_m^u\}$ represents the first m items in sequence S^u, and $m \in \mathcal{M}$ represents different context window sizes capturing varying intent scopes from immediate interests to long-term preferences.

The specific values of $\mathcal{M}$ are determined through scenario-adaptive optimization maintaining the hierarchical constraint $m_1 < m_2 < m_3$.

To effectively capture sequential dependencies within each segment in $I_m(S^u)$, we employ SASRec or GRU as the sequence encoder $f(\cdot)$. To enrich item representations with semantic understanding, we leverage Large Language Models (LLMs) (Lin et al. 2025) to extract semantic features from item textual descriptions, which are aligned with collaborative embeddings learned from interaction patterns through a unified representation space. Our approach incor-

porates adaptive weighting mechanisms through learnable parameters β_g and γ_g that automatically adjust the influence of different granularity levels during training.

For notation clarity, we denote $s^u_{g,i}$ as the i-th segment at granularity level g for user u, where $s^u_{g,i} \in I_g(S^u)$ represents a specific subsequence extracted from the multi-granularity segmentation process. The set $\mathcal{I}$ represents the complete item vocabulary used for substitution and insertion operations in adversarial training.

To discover shared intent patterns across users, we implement K-means clustering within each granularity level for computational efficiency and interpretability. This clustering process: (1) groups segment representations according to their parameter m, establishing $|\mathcal{M}|$ distinct granularity levels, and (2) executes K-means clustering within each granularity level g to identify level-specific intent prototypes. This yields a multi-granularity intent prototype set $C = \{c^g | g \in \mathcal{M}\}\}$, where each C^g contains the intent prototypes $\{c^g_1, ..., c^g_{K_g}\}$ at that granularity level.

Dual-perspective Contrastive Learning

To comprehensively capture user intents from different perspectives, we propose a dual-alignment contrastive learning framework that connects user representations with both intent prototypes and cross-user sequences sharing similar behavioral outcomes. This approach enables our model to learn both prototype-based and outcome-based intent patterns simultaneously.

Multi-granularity Intent Alignment To effectively model the multi-granularity nature of user intents, we establish connections between user sequence representations and intent prototypes at different granularity levels. This alignment process leverages the multi-granularity organized intent prototype set $C = \{C^g | g \in \{1, 2, ..., |\mathcal{M}|\}\}$ derived from the clustering process in Section 3.1.

For each user sequence S^u, we first sample from it and then compute its representation using the sequence encoder $f(\cdot)$, denoted as h^u. We then align this representation with the most relevant intent prototypes at each granularity level through a multi-granularity contrastive loss:

$$\mathcal{L}_{\text{multi}} = \sum_{g=1}^{|\mathcal{M}|} \beta_g \mathcal{L}_{\text{ICL}}(h^u, c^g_{k^*}), \tag{2}$$

where β_g is a learnable parameter for granularity level g, $c^g_{k^*}$ is the intent prototype in C^g that is closest to h^u, and k^* represents the index of this closest prototype. The $\mathcal{L}_{\text{ICL}}$ is defined as:

$$\begin{aligned} \mathcal{L}_{\text{ICL}}(x_1, x_2) = &- \log \frac{\exp(\text{sim}(x_1, x_2))}{\sum_{x_i \notin F} \exp(\text{sim}(x_1, x_i))} \\ &- \log \frac{\exp(\text{sim}(x_2, x_1))}{\sum_{x_i \notin F} \exp(\text{sim}(x_2, x_i))}, \end{aligned} \tag{3}$$

where $\text{sim}(\cdot)$ denotes the dot product, and F is constructed based on the False Negative Mitigation strategy (Liu et al. 2021a; Qin et al. 2024). Specifically, F contains sequence representations from the current mini-batch that potentially represent similar user intents: when x_2 is a sequence representation, F includes representations whose corresponding sequences terminate with the same item as x_1; when x_2 is an intent prototype, F comprises representations that belong to the same intent cluster as x_2.

This loss encourages the user sequence representation to align with its most semantically relevant intent prototype at each granularity level while maintaining distinctiveness from other prototypes at the same level.

Cross-User Target-Aware Contrast To enhance intent representation learning through cross-user knowledge transfer, we propose a cross-user target-aware contrastive learning framework that leverages behavioral outcome similarities to identify latent intent correlations across different users.

In sequential recommendation, users with distinct interaction histories may exhibit convergent behavioral patterns, indicating shared underlying motivations. We leverage this through a cross-user contrastive learning mechanism that uses behavioral outcome alignment as supervision signals. This approach facilitates the identification of transferable intent patterns while preserving robustness against individual preference variations and contextual noise.

Formally, given a user sequence S^u, we construct positive pairs by sampling from the $|\mathcal{M}|$ granularity-specific clusters to obtain cross-user sequences $\{S^{u'}_g\}^{|\mathcal{M}|}_{g=1}$ where each $S^{u'}_g$ exhibits similar behavioral outcomes as S^u at granularity level g. Each sampled sequence is processed through the encoder to derive its representation $h^{u'}_g$.

The cross-user contrastive learning objective is formulated as:

$$\mathcal{L}_{\text{cross}} = \sum_{g=1}^{|\mathcal{M}|} \gamma_g \mathcal{L}_{\text{ICL}}(h^u, h^{u'}_g), \tag{4}$$

where γ_g represents a learnable parameter that determines the contribution of contrastive learning at granularity level g. This cross-user contrastive learning mechanism enables knowledge transfer between users with similar interaction outcomes, allowing the model to identify common intent patterns across multiple granularity levels while providing a complementary perspective to the multi-granularity intent alignment.

Intent Adversarial Robustness

Motivated by the Actor-Critic paradigm in reinforcement learning (Schulman et al. 2017; Li et al. 2025), we design an Adversarial Augmentor-Intent Critic architecture that mirrors the actor-critic interaction: the Adversarial Augmentor functions analogously to an actor by learning to generate semantically consistent perturbations, while the Intent Critic acts as an evaluator providing adversarial feedback signals to guide the augmentor's learning process.

Phase 1: Adversarial Co-Evolution The Adversarial Augmentor and Intent Critic engage in adversarial co-evolution. The Adversarial Augmentor $\mathcal{A}(\cdot)$ is formulated as:

$$h^u_{\text{adv}} = \mathcal{A}(s^u_{g,i}, \omega) = f_a(\text{Encoder}(\text{Perturb}(s^u_{g,i}, \omega))), \quad (5)$$

where $s^u_{g,i} \in I_g(S^u)$ represents a segment from the multi-granularity intent representation, ω denotes the perturbation strategy, Perturb$(\cdot)$ applies noise operations, Encoder$(\cdot)$ obtains the initial representation, and $f_a(\cdot)$ refines this representation. The perturbation operations are:

$$\text{Perturb}(s^u_{g,i}, \omega) \in \{\text{Mask}, \text{Substitute}, \text{Swap}, \text{Insert}\}. \quad (6)$$

These operations introduce controlled noise while preserving sequence semantics.

The Intent Critic $\mathcal{C}(\cdot)$ distinguishes real from perturbed representations:

$$\mathcal{C}(h^u) = \sigma(f_c(h^u)), \quad (7)$$

where $f_c(\cdot)$ is a neural network and $\sigma(\cdot)$ is the sigmoid function.

The Intent Critic loss is:

$$\mathcal{L}_{\text{critic}} = -\mathbb{E}[\log \mathcal{C}(h^u)] - \mathbb{E}[\log(1 - \mathcal{C}(h^u_{\text{adv}}))]. \quad (8)$$

To ensure semantic consistency, inspired by the KL penalty in PPO (Schulman et al. 2017) that prevents policy deviation in RLHF (Ouyang et al. 2022), we introduce a soft intent alignment constraint:

$$\mathcal{L}_{\text{sem}} = \sum_{g=1}^{|\mathcal{M}|} \text{KL}\big(\text{softmax}(\text{sim}(h^u, C^g)/\tau) \,\|\, \text{softmax}(\text{sim}(h^u_{\text{adv}}, C^g)/\tau)\big), \quad (9)$$

where $\text{sim}(h, C^g) = [\text{sim}(h, c^g_1), ..., \text{sim}(h, c^g_{K_g})]$ represents the similarity vector to all prototypes at granularity g, and τ is the temperature parameter.

The Adversarial Augmentor objective is:

$$\mathcal{L}_{\text{aug}} = \mathbb{E}[\log \mathcal{C}(h^u)] + \mathbb{E}[\log(1 - \mathcal{C}(h^u_{\text{adv}}))] + \lambda \mathcal{L}_{\text{sem}}, \quad (10)$$

where λ controls the semantic consistency constraint weight.

Phase 2: Robust Contrastive Learning with Frozen Augmentor After Phase 1 convergence, we freeze the Adversarial Augmentor to serve as a stable perturbation generator. This prevents the augmentor from overfitting to the downstream task while ensuring consistent data augmentation quality. For each sequence segment $s^u_{g,i} \in I_g(S^u)$, we obtain adversarial representation h^u_{adv} and treat it as positive samples:

$$\mathcal{L}_{\text{robust}} = \mathcal{L}_{\text{ICL}}(h^u, h^u_{\text{adv}}), \quad (11)$$

where $\mathcal{L}_{\text{ICL}}$ refers to the Intent Contrastive Learning loss.

The complete ISAR training procedure is summarized in Algorithm 1, which integrates both Augmentor-Critic training and robust contrastive learning phases.

Algorithm 1: Intent-Similarity Adversarial Robustness

Input: User sequences $\{S^u\}$, Intent prototypes $\{C^g\}_{g=1}^{|\mathcal{M}|}$
Output: Robust user representations
Phase 1: Adversarial Co-Evolution
Initialize Adversarial Augmentor $\mathcal{A}$ and Intent Critic $\mathcal{C}$
for epoch = 1 to E_1 **do**
  for each user sequence segment $s^u_{g,i} \in I_g(S^u)$ **do**
    Generate perturbed sequence: $h^u_{\text{adv}} = \mathcal{A}(s^u_{g,i}, \omega)$
    Compute semantic constraint:
      $\mathcal{L}_{\text{sem}} = \sum_{g=1}^{|\mathcal{M}|} \text{KL}(p^u_g \,\|\, p^{\text{adv}}_g)$
      where $p^u_g = \text{softmax}(\text{sim}(h^u, C^g)/\tau)$
      and $p^{\text{adv}}_g = \text{softmax}(\text{sim}(h^u_{\text{adv}}, C^g)/\tau)$
  end for
  Update Intent Critic: $\mathcal{C} \leftarrow \arg\min_{\mathcal{C}} \mathcal{L}_{\text{critic}}$
  Update Adversarial Augmentor: $\mathcal{A} \leftarrow \arg\min_{\mathcal{A}} \mathcal{L}_{\text{aug}}$
end for
Phase 2: Robust Contrastive Learning
Freeze Adversarial Augmentor: $\mathcal{A}_{\text{frozen}}$
for epoch = 1 to E_2 **do**
  for each user sequence segment $s^u_{g,i} \in I_g(S^u)$ **do**
    Generate adversarial sample:
      $h^u_{\text{adv}} = \mathcal{A}_{\text{frozen}}(s^u_{g,i}, \omega)$
    Compute robust contrastive loss:
      $\mathcal{L}_{\text{robust}} = \mathcal{L}_{\text{ICL}}(h^u, h^u_{\text{adv}})$
  end for
  Update model parameters with $\mathcal{L}_{\text{robust}}$
end for

Multi-Task Learning Objective

The complete MIMAR-SRec framework integrates multi-granularity intent modeling with adversarial robustness through a multi-task learning objective. The final loss function combines the standard recommendation loss with intent modeling and robustness components:

$$\mathcal{L}_{\text{MIMAR}} = \mathcal{L}_{\text{Rec}} + \mu \mathcal{L}_{\text{multi}} + \nu \mathcal{L}_{\text{cross}} + \eta \mathcal{L}_{\text{robust}}, \quad (12)$$

where $\mathcal{L}_{\text{Rec}}$ is the sequential recommendation loss, while μ, ν, and η are learnable parameters that are automatically optimized during training to control the contribution of each component. $\mathcal{L}_{\text{multi}}$ represents the multi-granularity intent modeling loss, $\mathcal{L}_{\text{cross}}$ captures the cross-granularity intent relationships, and $\mathcal{L}_{\text{robust}}$ enhances model robustness against noise in user behavior sequences.

This unified multi-task learning objective enables the model to simultaneously learn fine-grained intent representations at multiple granularities while developing robustness against various types of noise in user behavior sequences.

Experiments

Experimental Setting

Datasets We evaluate our method on four datasets collected from four real-world platforms , including three Amazon product categories (Sports, Beauty, Toys) and the MovieLens-1M (ML-1M) movie recommendation dataset. Amazon subsets contain post-2019 user-product interactions and metadata (McAuley et al. 2015), while ML-1M comprises long-sequence user-movie ratings (Liu et al. 2021b).

DataSet	Metric	BPR	Caser	SASRec	BERT4Rec	CoSeRec	DuoRec	ICLRec	IOCRec	ICSRec	Ours	Improve
Sports	HR@10	0.0210	0.0254	0.0327	0.0351	0.0432	0.0459	0.0429	0.0445	0.0549	**0.0698***	27.14%
	HR@20	0.0362	0.0391	0.0493	0.0596	0.0628	0.0687	0.0638	0.0676	0.0782	**0.0966***	23.53%
	NDCG@10	0.0100	0.0129	0.0171	0.0183	0.0237	0.0236	0.0231	0.0213	0.0321	**0.0415***	29.28%
	NDCG@20	0.0138	0.0171	0.0211	0.0244	0.0286	0.0294	0.0284	0.0272	0.0381	**0.0482***	26.51%
Beauty	HR@10	0.0289	0.0335	0.0616	0.0593	0.0717	0.0843	0.0736	0.0766	0.0935	**0.1134***	21.28%
	HR@20	0.0467	0.0635	0.0886	0.0976	0.1026	0.1219	0.1050	0.1138	0.1286	**0.1534***	19.28%
	NDCG@10	0.0141	0.0219	0.0335	0.0293	0.0402	0.0433	0.0396	0.0388	0.0564	**0.0698***	23.76%
	NDCG@20	0.0185	0.0291	0.0379	0.0384	0.0479	0.0528	0.0476	0.0482	0.0652	**0.0799***	22.55%
Toys	HR@10	0.0191	0.0326	0.0644	0.0517	0.0747	0.0951	0.0826	0.0796	0.1032	**0.1212***	17.44%
	HR@20	0.0320	0.0535	0.0949	0.0752	0.1029	0.1284	0.1131	0.1124	0.1356	**0.1596***	17.70%
	NDCG@10	0.0094	0.0162	0.0313	0.0302	0.0434	0.0482	0.0472	0.0374	0.0641	**0.0756***	17.94%
	NDCG@20	0.0126	0.0214	0.0390	0.0361	0.0505	0.0566	0.0549	0.0457	0.0723	**0.0852***	17.84%
ML-1M	HR@10	0.0406	0.1434	0.1802	0.2211	0.1853	0.3069	0.2265	0.2681	0.3222	**0.3392***	5.28%
	HR@20	0.0743	0.2220	0.2737	0.3346	0.2942	0.4089	0.3360	0.3823	0.4336	**0.4498***	3.74%
	NDCG@10	0.0206	0.0727	0.0941	0.1089	0.0907	0.1741	0.1167	0.1479	0.1928	**0.2032***	5.39%
	NDCG@20	0.0291	0.0924	0.1149	0.1376	0.1239	0.1999	0.1442	0.1767	0.2202	**0.2313***	5.04%

Table 1: Performance comparisons of different methods. The results of the best baseline are underlined in each row. The last column is the relative improvements compared with the best baseline results. An asterisk (*) denotes statistically significant improvements over the best baseline, as assessed by a two-tailed t-test with $p < 0.05$.

Baseline Models MIMAR-SRec is benchmarked against the following sequential recommendation models:

- **Non-sequential**: BPR (Rendle et al. 2009)
- **General Sequential**: Caser (Tang and Wang 2018), SASRec (Kang and McAuley 2018)
- **SSL-based**: BERT4Rec (Sun et al. 2019), CoSeRec (Li et al. 2021), DuoRec (Qiu et al. 2022)
- **Intent-Guided**: ICLRec (Chen et al. 2022), IOCRec (Li et al. 2023), ICSRec (Qin et al. 2024)

Evaluation Methodology Following (Wang et al. 2019), we perform full-ranking over the entire item corpus without negative sampling and quantify model performance using Hit Ratio@k (HR@k) and Normalized Discounted Cumulative Gain@k (NDCG@k) with $k \in \{10, 20\}$.

Implementation Details We use batch size 256, Adam optimizer with learning rate 10^{-3}, and temperature $\tau = 1.0$. Dropout rates and regularization parameters $\{\lambda, \beta\}$ are selected from {0.1, 0.2, 0.3, 0.4, 0.5}, cluster numbers from {64, 128, 256, 512, 1024}, sampling ratios $\{\eta, \gamma, \delta\}$ from [0.1, 0.9], and $\{\epsilon, \omega\}$ from [0.5, 1.0]. Experiments run on NVIDIA A100 GPU with early stopping.

Performance Comparison

Table 1 demonstrates MIMAR-SRec's superior performance across all datasets and metrics, consistently outperforming baseline methods in both sparse (Sports, Beauty, Toys) and dense (ML-1M) scenarios. Notably, MIMAR-SRec surpasses the strong baseline IOCRec[1] (Wang et al. 2025) by 13.41%, 11.75%, 9.23%, and 2.80% on NDCG@20, respectively. Across baseline categories, non-sequential methods (BPR) show limited temporal modeling, general sequential methods (Caser, SASRec) lack explicit intent modeling, SSL-based methods (BERT4Rec, CoSeRec, DuoRec) enhance representation learning, while intent-guided methods (ICLRec, IOCRec, ICSRec) represent the most competitive baselines. MIMAR-SRec's advantages are particularly pronounced in sparse datasets, where multi-granularity intent modeling effectively extracts meaningful patterns from limited interactions. The three key components work synergistically to capture temporal patterns, facilitate collaborative learning, and enhance stability against noise.

Ablation Study

To validate component contributions and interaction effects, we conduct experiments with the following variants: (A) Full Model with all components: Multi-granularity Intent Alignment (MIA), Cross-User Target-Aware Contrast (CUTAC), and Intent-Similarity Adversarial Robustness (ISAR). (B-D) Single component removal: w/o MIA, w/o CUTAC, w/o ISAR. (E-G) Pairwise component removal: w/o MIA+CUTAC, w/o MIA+ISAR, w/o CUTAC+ISAR.

Table 2 presents the results. Key findings include:

Component Significance: CUTAC exhibits the most substantial impact when removed independently, demonstrating its critical role in capturing collaborative signals. MIA

[1]We implement IOCRec following the original paper as the code is not publicly available.

Model Variant	Sports		Beauty		Toys		ML-1M	
	HR@20	NDCG@20	HR@20	NDCG@20	HR@20	NDCG@20	HR@20	NDCG@20
Full Model	**0.0966**	**0.0482**	**0.1534**	**0.0799**	**0.1596**	**0.0852**	**0.4498**	**0.2313**
w/o MIA	0.0912	0.0454	0.1456	0.0758	0.1531	0.0817	0.4421	0.2291
w/o CUTAC	0.0847	0.0423	0.1389	0.0719	0.1467	0.0779	0.4356	0.2264
w/o ISAR	0.0889	0.0437	0.1423	0.0741	0.1498	0.0798	0.4387	0.2276
w/o MIA+CUTAC	0.0798	0.0401	0.1312	0.0681	0.1398	0.0741	0.4289	0.2238
w/o MIA+ISAR	0.0834	0.0418	0.1356	0.0703	0.1442	0.0763	0.4312	0.2251
w/o CUTAC+ISAR	0.0801	0.0395	0.1298	0.0672	0.1381	0.0728	0.4267	0.2225

Table 2: Comprehensive ablation study of MIMAR-SRec components across all datasets. MIA: Multi-granularity Intent Alignment, CUTAC: Cross-User Target-Aware Contrast, ISAR: Intent-Similarity Adversarial Robustness.

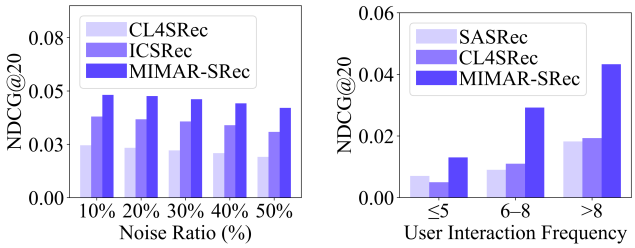

Figure 3: Robustness analysis (NDCG@20, Sports).

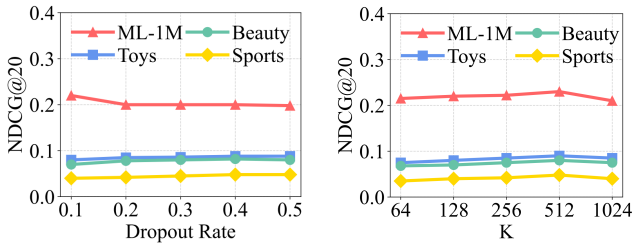

Figure 4: Hyperparameter sensitivity analysis on NDCG@20.

shows consistent degradation across datasets, while ISAR's importance varies with dataset sparsity.

Interaction Effects: Combined removals reveal non-additive patterns. The MIA+CUTAC removal produces severe degradation in sparse datasets, indicating complementary mechanisms, while CUTAC+ISAR interaction proves critical for handling data sparsity.

Dataset Sensitivity: Sparse datasets show heightened sensitivity to component interactions, confirming the architecture's effectiveness for challenging sparse scenarios.

Robustness Analysis

Figure 3 presents the robustness analysis results. MIMAR-SRec exhibits smaller performance degradation compared to CL4SRec and ICSRec under increasing noise levels. To evaluate robustness during inference, we trained models on clean data and systematically injected varying proportions (10% to 50%) of noisy interactions into each test sequence. These noises comprised random negative items (selected from popular but non-interacted items among similar user cohorts) and more complex perturbations including temporally disordered interactions and contextually irrelevant behaviors, integrated within a unified experimental framework. Experimental results demonstrate that MIMAR-SRec maintains superior performance stability, indicating enhanced robustness against diverse interaction noise patterns.

Additionally, MIMAR-SRec outperforms baselines across different interaction lengths, especially for users with short sequences (≤ 5). This highlights its effectiveness in cold-start scenarios, where the multi-granularity intent modeling and adversarial training help produce reliable recommendations even with limited user history.

Parameter Sensitivity Analysis

We analyze the impact of key hyperparameters on model performance, as shown in Figure 4.

Dropout Rate The results show that moderate dropout rates generally yield the best performance across all datasets. When the dropout rate is too low, the model is prone to overfitting, especially on smaller datasets. Conversely, excessive dropout can impair the model's representation capacity, leading to underfitting. The optimal dropout rate balances regularization and representation learning, enabling effective multi-granularity intent modeling while preventing overfitting.

Intent Prototype Number (K) Performance initially improves with increasing K, reaching optimal results at moderate values before degrading with larger K. Excessive prototype numbers lead to over-fragmentation and computational inefficiency. The optimal K balances representational capacity and model complexity.

Conclusion

This paper proposes MIMAR-SRec, a novel sequential recommendation framework addressing the challenges of capturing user intents and resisting interaction data noise. The framework employs multi-granularity intent modeling and intent-similarity adversarial robustness to effectively model the complexity of user preferences while enhancing robustness against data perturbations.

References

Bian, S.; Zhao, W. X.; Wang, J.; and Wen, J.-R. 2022. A relevant and diverse retrieval-enhanced data augmentation framework for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2923–2932.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Chen, Y.; Li, J.; Liu, C.; Li, C.; Anderle, M.; McAuley, J.; and Xiong, C. 2021. Modeling dynamic attributes for next basket recommendation. *arXiv preprint arXiv:2109.11654*.

Chen, Y.; Liu, Z.; Li, J.; McAuley, J.; and Xiong, C. 2022. Intent contrastive learning for sequential recommendation. In *Proceedings of the ACM web conference 2022*, 2172–2182.

Cheng, C.; Yang, H.; Lyu, M. R.; and King, I. 2013. Where you like to go next: Successive point-of-interest recommendation. In *IJCAI*, volume 13, 2605–2611.

Chuang, C.-Y.; Robinson, J.; Lin, Y.-C.; Torralba, A.; and Jegelka, S. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33: 8765–8775.

Fan, L.; Liu, S.; Chen, P.-Y.; Zhang, G.; and Gan, C. 2021. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? *Advances in neural information processing systems*, 34: 21480–21492.

Fan, Z.; Liu, Z.; Wang, Y.; Wang, A.; Nazari, Z.; Zheng, L.; Peng, H.; and Yu, P. S. 2022. Sequential recommendation via stochastic self-attention. In *Proceedings of the ACM web conference 2022*, 2036–2047.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

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

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Li, M.; Lin, J.; Zhao, X.; Lu, W.; Zhao, P.; Wermter, S.; and Wang, D. 2025. Curriculum-RLAIF: Curriculum Alignment with Reinforcement Learning from AI Feedback. *arXiv preprint arXiv:2505.20075*.

Li, X.; Sun, A.; Zhao, M.; Yu, J.; Zhu, K.; Jin, D.; Yu, M.; and Yu, R. 2023. Multi-intention oriented contrastive learning for sequential recommendation. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, 411–419.

Li, Y.; Chen, T.; Zhang, P.-F.; and Yin, H. 2021. Lightweight self-attentive sequential recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 967–977.

Lin, J.; Guo, Y.; Han, Y.; Hu, S.; Ni, Z.; Wang, L.; Chen, M.; Liu, H.; Chen, R.; He, Y.; et al. 2025. Se-agent: Self-evolution trajectory optimization in multi-step reasoning with llm-based agents. *arXiv preprint arXiv:2508.02085*.

Lin, J.; Li, Q.; Xie, G.; Guan, Z.; Jiang, Y.; Xu, T.; Zhang, Z.; and Zhao, P. 2024a. Mitigating Sample Selection Bias with Robust Domain Adaption in Multimedia Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7581–7590.

Lin, J.; Peng, S.; Zhang, Z.; and Zhao, P. 2024b. TLRec: A Transfer Learning Framework to Enhance Large Language Models for Sequential Recommendation Tasks. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 1119–1124.

Liu, Z.; Chen, Y.; Li, J.; Yu, P. S.; McAuley, J.; and Xiong, C. 2021a. Contrastive self-supervised sequential recommendation with robust augmentation. *arXiv preprint arXiv:2108.06479*.

Liu, Z.; Fan, Z.; Wang, Y.; and Yu, P. S. 2021b. Augmenting sequential recommendation with pseudo-prior items via reversely pre-training transformer. In *Proceedings of the 44th international ACM SIGIR conference on Research and development in information retrieval*, 1608–1612.

Ma, J.; Zhou, C.; Yang, H.; Cui, P.; Wang, X.; and Zhu, W. 2020. Disentangled self-supervision in sequential recommenders. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 483–491.

McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 43–52.

Mnih, A.; and Kavukcuoglu, K. 2013. Learning word embeddings efficiently with noise-contrastive estimation. *Advances in neural information processing systems*, 26.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Qin, X.; Yuan, H.; Zhao, P.; Liu, G.; Zhuang, F.; and Sheng, V. S. 2024. Intent contrastive learning with cross subsequences for sequential recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*, 548–556.

Qiu, R.; Huang, Z.; Yin, H.; and Wang, Z. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, 813–823.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, 452–461. Arlington, Virginia, USA: AUAI Press. ISBN 9780974903958.

Rendle, S.; Freudenthaler, C.; and Schmidt-Thieme, L. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, 811–820.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 1441–1450.

Tang, J.; and Wang, K. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 565–573.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W.; Ma, J.; Zhang, Y.; Zhang, K.; Jiang, J.; Yang, Y.; Zhou, Y.; and Zhang, Z. 2025. Intent Oriented Contrastive Learning for Sequential Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12748–12756.

Wang, X.; He, X.; Wang, M.; Feng, F.; and Chua, T.-S. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, 165–174.

Wu, C.-Y.; Ahmed, A.; Beutel, A.; Smola, A. J.; and Jing, H. 2017. Recurrent recommender networks. In *Proceedings of the tenth ACM international conference on web search and data mining*, 495–503.

Wu, J.; Wang, X.; Feng, F.; He, X.; Chen, L.; Lian, J.; and Xie, X. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 726–735.

Xie, X.; Sun, F.; Liu, Z.; Wu, S.; Gao, J.; Zhang, J.; Ding, B.; and Cui, B. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*, 1259–1273. IEEE.

Xie, Z.; Liu, C.; Zhang, Y.; Lu, H.; Wang, D.; and Ding, Y. 2021. Adversarial and contrastive variational autoencoder for sequential recommendation. In *Proceedings of the web conference 2021*, 449–459.

Zheng, Y.; Gao, C.; Chang, J.; Niu, Y.; Song, Y.; Jin, D.; and Li, Y. 2022. Disentangling long and short-term interests for recommendation. In *Proceedings of the ACM web conference 2022*, 2256–2267.

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