Neurocomputing 698 (2026) 134241

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Similarity-aware generalization framework for zero-shot cross-domain sequential recommendation

Zihao Pan [a], Yuzhuo Dang [a], Xin Zhang [a], Zhiqiang Pan [a], Taihua Shao [b], Fei Cai [a,*]

[a] National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha, 410073, China
[b] Institute of Data and Target Engineering, Information Engineering University, Zhengzhou, 450001, China

ARTICLE INFO

Communicated by H. Peng

Keywords:
Zero-shot recommendation
Cross-domain recommendation
Sequential recommendation
Large language models
Recommender systems

ABSTRACT

Zero-shot cross-domain sequential recommendation aims to transfer sequential knowledge from an interaction-rich source domain to interaction-unseen target domains without any target-domain interaction data during training. Recent large language model (LLM)-based methods have improved this setting by leveraging item-side semantic information. However, existing alignment strategies typically enforce uniform cross-domain attraction over all item pairs. Such non-selective alignment may distort the latent semantic structures and weaken the domain-specific discrimination, especially when semantically weakly related items are forced to align. To address this issue, we propose **SAGERec**, a similarity-aware generalization framework for zero-shot cross-domain sequential recommendation. SAGERec mainly comprises a *Similarity-Aware Inter-domain Compactness* (SIC) module that selectively aligns the semantically compatible cross-domain item pairs, and a *Domain-Adaptive Generalization* (DAG) module that dynamically adjusts the strength of semantic regularization according to the discrepancy between the source and target domains. In this way, the proposed framework facilitates reliable transfer while reducing noisy cross-domain attraction. Extensive experiments on multiple real-world datasets and representative sequential recommendation backbones demonstrate that SAGERec consistently improves recommendation performance within backbone families and achieves the best overall results across zero-shot transfer settings. Further analyses indicate that, compared with existing semantic-based and generalization-based baselines, SAGERec yields more structured cross-domain alignment and a more favorable efficiency–effectiveness trade-off.

1. Introduction

Sequential recommendation aims to model users' dynamic preferences from historical interactions and has become a fundamental component of modern recommender systems [1,2]. However, most existing sequential recommenders are developed under the assumption that sufficient interaction data are available within a specific domain for model training and adaptation [3–6]. In practice, this assumption is frequently violated. For example, when models trained on interaction-rich domains are deployed to newly emerging services or privacy-constrained platforms, the target-domain interactions may be sparse or entirely unavailable. Therefore, a primary challenge is how to effectively transfer knowledge from a high-resource source domain to support recommendations in a low-resource target domain [7].

Cross-Domain Sequential Recommendation (CDSR) addresses the aforementioned issue by transferring preference signals across domains. Nevertheless, most existing CDSR methods still rely on overlapping users or items, or require target-domain interactions for model adaptation [7]. These requirements restrict their applicability in realistic deployment scenarios, which motivates a practical and challenging setting, namely *Zero-shot Cross-Domain Sequential Recommendation* (ZCDSR). In this setting, a model trained on an interaction-rich source domain is directly transferred to a previously unseen target domain without using any target-domain interaction data for optimization, while only the item-side content is available for transfer [8]. This objective is particularly challenging for sequential recommendation as the model must generalize across domains with substantial semantic and behavioral discrepancies while preserving sequential dependency patterns learned from the source domain.

Naturally, item-side content becomes a key source of transferable signals for ZCDSR. Although recent content-enhanced recommendation

* Corresponding author.
Email addresses: panzihao@nudt.edu.cn (Z. Pan), dangyuzhuo@nudt.edu.cn (Y. Dang), zhangxin16@nudt.edu.cn (X. Zhang), panzhiqiang@nudt.edu.cn (Z. Pan), shaotaihua13@nudt.edu.cn (T. Shao), caifei08@nudt.edu.cn (F. Cai).

https://doi.org/10.1016/j.neucom.2026.134241
Received 8 April 2026; Received in revised form 21 May 2026; Accepted 5 June 2026
Available online 11 June 2026

studies have exploited auxiliary item signals to better model user preferences [9,10], these signals still need to be transformed into transferable representations for bridging heterogeneous domains. Advances in large language models (LLMs) provide a more direct semantic pathway for this purpose, owing to their strong semantic understanding and generalization ability acquired through large-scale pretraining [11,12]. By encoding item-side metadata, such as titles, descriptions, and attributes, into dense semantic representations, LLMs make it possible to construct a shared semantic space across domains even without user or item overlap. This capability is especially appealing in zero-shot settings, where target-domain interactions are unavailable for training while item content is accessible at deployment time. Motivated by this advantage, recent studies have begun to explore LLM-driven zero-shot recommendation and have shown that pretrained semantic representations can substantially improve transferability to unseen domains [8,13,14].

However, existing semantic generalization methods typically approach the cross-domain transfer in a rigid and uniform manner, mainly manifested at two levels. At the item level, they impose a global alignment between the source and target semantic spaces without distinguishing the varying degrees of semantic compatibility among specific item pairs. Such non-selective alignment overlooks the fine-grained structure of cross-domain relations. Specifically, the highly compatible pairs provide reliable transfer signals while the weakly related pairs may introduce semantic noise and lead to negative transfer. At the domain level, existing approaches consistently apply a fixed regularization strategy. This assumes that the trade-off between recommendation accuracy and generalization strength remains constant across different target domains. Such a one-size-fits-all strategy fails to account for the substantial semantic gaps caused by differences in catalog structure, content modality and linguistic style. These limitations highlight the critical need for a new ZCDSR framework. Rather than relying on indiscriminate alignment, an effective solution can dynamically adapt its transfer strength based on specific discrepancies at both the item and domain levels.

To address the above challenges, we propose **SAGERec**, a **S**imilarity-**A**ware **GE**neralization framework for zero-shot cross-domain sequential recommendation, which systematically resolves the limitations of existing methods at both the item and domain levels. First, to address the item-level problem of non-selective alignment, we design a similarity-aware semantic generalization objective. Within this objective, a *Similarity-Aware Inter-domain Compactness* (SIC) term selectively aligns semantically compatible cross-domain item pairs according to thresholded semantic affinity in the LLM embedding space, while an intra-domain diversity preservation term counteracts over-compression and retains the fine-grained item discrimination within each domain. Second, to address the domain-level problem of static regularization across heterogeneous domain pairs, we introduce a *Domain-Adaptive Generalization* (DAG) mechanism. DAG dynamically modulates the strength of semantic regularization according to the estimated discrepancy between the source and target domains, allowing stronger transfer across semantically close domains while encouraging more conservative alignment for distant ones. Ultimately, SAGERec offers a plug-and-play solution. It can be seamlessly incorporated into existing encoder-based semantic-sequential recommenders without modifying the backbone architecture. Extensive experiments on three real-world datasets show that SAGERec consistently outperforms the semantic-based and generalization-based baselines in zero-shot transfer scenarios.

The main contributions of this work are summarized as follows:

- We propose a **S**imilarity-**A**ware **GE**neralization framework for ZCDSR, named **SAGERec**, which enhances the LLM-based semantic-sequential transfer pipeline by incorporating a similarity-aware semantic regularization into the transfer process.
- We design a *Similarity-Aware Inter-domain Compactness* mechanism to address the issue of non-selective cross-domain alignment, thereby reducing the noisy attraction by preventing weakly related items from being overly aligned.
- We introduce a *Domain-Adaptive Generalization* mechanism to address the limitation of static regularization across heterogeneous source-target domain pairs, thereby enabling the transfer strength to adapt to varying degrees of semantic discrepancy across domains.
- Extensive experiments on three real-world domains and representative sequential backbones verify the effectiveness, robustness, and efficiency of SAGERec, with additional analyses on sensitivity, representation structure, encoder backbones, and qualitative item-pair cases.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 formulates the problem definition. Section 4 details the proposed SAGERec framework. Subsequently, Section 5 reports the experimental results and analysis. Finally, Section 6 concludes the paper and discusses future directions.

2. Related work

2.1. Cross-domain sequential recommendation

Cross-domain sequential recommendation (CDSR) aims to improve recommendation quality by transferring sequential preference signals across domains, and has been widely studied as an effective solution to data sparsity and cold-start issues. Early methods such as π-Net [15] and PSJNet [16] mainly rely on explicit information sharing across domains, while later approaches such as C^{2}DSR [17] further incorporate graph modeling and contrastive learning to capture richer cross-domain dependencies. More recent studies have sought to relax the assumptions of conventional CDSR. For example, AMID [18] points out that many existing CDSR methods still depend on restrictive conditions such as overlapping users or stable train–test distributions, which may not hold in realistic environments. ADR-Rec [19] further improves CDSR by disentangling domain-specific and cross-domain preferences and dynamically controlling cross-sequence information with a gating mechanism. A recent survey [7] also shows that CDSR is gradually evolving toward more open and realistic transfer settings. Nevertheless, most existing CDSR methods still require some form of cross-domain overlap or target-domain interaction signal for adaptation, and are therefore not well suited to the more challenging setting where the target domain is previously unseen and no interaction data are available during training.

2.2. Zero-shot recommendation

Zero-shot recommendation studies how to generalize recommendation models to unseen users, unseen items, or entirely new domains without conventional supervised adaptation. Early work such as ZESRec [20] explicitly formulates zero-shot recommendation without overlapping users or items, while MAIL [21] addresses zero-shot cold-start recommendation through aligned auxiliary representations. Another related line focuses on transferable or pre-trained recommenders for zero-shot generalization. For example, PrepRec [22] improves transferability by exploiting universal interaction statistics and pre-trained sequential representations. UniSRec [23] learns universal sequence representations from item description text with a lightweight item encoding architecture and an MoE-enhanced adapter, making it a strong text-enhanced transferable recommendation baseline. In addition, prompt-oriented paradigms such as P5 [24] and SCRIPT [25] further extend this direction by reformulating recommendation tasks into prompt-based prediction or generation problems. More recently, TCLF [26] highlights that, in the absence of target-domain interaction supervision, item content becomes a crucial bridge for zero-shot transfer, and richer alignment signals beyond pure content similarity can better support preference transfer. Although these studies demonstrate the importance of transferable semantic representations, they are not specifically designed for zero-shot cross-domain sequential recommendation, nor do they explicitly address how to organize and align cross-domain semantics under sequential transfer.

2.3. LLM-based recommendation

The rapid development of large language models (LLMs) has stimulated extensive research on LLM-based recommendation [11]. Existing approaches broadly follow two paradigms: generative recommendation, which uses LLMs as recommenders or ranking agents, such as TALLRec [27], GenRec [28], LlamaRec [29], and zero-shot conversational recommendation [30]; and encoder-based recommendation, where LLMs provide semantic embeddings for recommenders. Generative methods rely on instruction tuning, prompting, or autoregressive decoding, whereas our work follows the encoder-based paradigm and uses LLM-derived item representations as transferable semantic priors for downstream sequential recommendation. Harte et al. [31] show that LLM-derived embeddings improve sequential recommendation, while LLM-ESR [32] demonstrates cached semantic features enhance recommendation without extra inference overhead. Safar et al. [33] further show that LLM-generated semantic embeddings with FAISS-based similarity search improve recommendation quality and retrieval efficiency across domains, highlighting the value of high-quality semantic representations for downstream recommendation and retrieval. Closely related to our work, LLM-RecG [8] studies zero-shot cross-domain sequential recommendation and identifies domain semantic bias as a key transfer obstacle. These studies confirm that LLM-derived semantic representations offer a foundation for zero-shot cross-domain recommendation. However, existing methods still largely rely on uniform semantic generalization and pay limited attention to transfer heterogeneity across semantically different item pairs and domain pairs, motivating our similarity-aware and domain-adaptive framework for robust zero-shot cross-domain sequential transfer.

3. Problem definition

We study zero-shot cross-domain sequential recommendation (ZCDSR), where a model is trained on a source domain and transferred to a previously unseen target domain without using any target-domain interaction data for parameter optimization. The core challenge is to learn transferable sequential knowledge from an interaction-rich source domain and generalize it to a target domain where only source-domain interaction supervision is available during training.

We denote the source domain by

$$D^s = (U^s, V^s, X^s, S^s), \tag{1}$$

where U^s and V^s denote the user set and item set, respectively, X^s denotes the textual metadata of source-domain items, and S^s denotes the set of source-domain user interaction sequences. For each user $u^s \in U^s$, the interaction sequence is written as

$$\mathbf{s}_{u^s} = [v^s_{u^s,1}, v^s_{u^s,2}, \dots, v^s_{u^s,T_{u^s}}], \tag{2}$$

where $v^s_{u^s,t} \in V^s$ is the item interacted with at position i, and T_{u^s} denotes the sequence length of user u^s.

Similarly, the target domain is denoted by

$$D^t = \left(U^t, V^t, X^t, S^t\right), \tag{3}$$

where U^t, V^t, X^t, and S^t are defined analogously for the target domain. In addition to the current target domain, we consider a set of auxiliary domains, denoted by

$$\mathcal{D}^a = \{D^{a_1}, D^{a_2}, \dots, D^{a_M}\}, \tag{4}$$

where each auxiliary domain is a non-source domain that is not selected as the evaluation target in the current zero-shot task. For notational convenience, we define the set of all non-source domains as

$$\mathcal{D}^{ns} = \{D^t\} \cup \mathcal{D}^a. \tag{5}$$

Following the zero-shot setting, all non-source domains are disjoint from the source domain in both users and items:

$$\forall D^q \in \mathcal{D}^{ns}, \quad U^s \cap U^q = \emptyset, \qquad V^s \cap V^q = \emptyset. \tag{6}$$

During training, only source-domain interaction sequences S^s are available for supervised recommendation learning, while for each non-source domain $D^q \in \mathcal{D}^{ns}$, only item-side metadata X^q is assumed to be available. Accordingly, the model cannot rely on interaction-based adaptation in the target domain, but must instead learn transferable sequential preference patterns from the source domain and generalizable item semantics from item-side information across domains.

Formally, let $f_\Theta(\mathbf{s}, v)$ denote a parameterized scoring function that maps an interaction sequence $\mathbf{s}$ and a candidate item v to a real-valued relevance score. For notational simplicity, let $X = \{X^\delta\}_{D^\delta \in \{D^s\} \cup \mathcal{D}^{ns}}$ denote the collection of item-side metadata from all domains. The goal of ZCDSR is to learn the optimal parameters Θ^* by jointly optimizing a source-domain recommendation objective and a cross-domain generalization objective:

$$\Theta^* = \arg\min_\Theta \ \mathcal{L}_{\text{rec}}(\Theta; S^s, X) + \Omega(\Theta; X)\ \mathcal{L}_{\text{gen}}(\Theta; X), \tag{7}$$

where $\mathcal{L}_{\text{rec}}$ denotes the source-domain recommendation objective defined over training instances derived from S^s, $\mathcal{L}_{\text{gen}}$ denotes a cross-domain generalization objective induced by item-side information from all domains, and $\Omega(\Theta; X)$ denotes a balancing term that controls the contribution of cross-domain generalization. Here, $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{gen}}$ are task-level abstractions, whose concrete formulations will be specified in the method section.

At inference time, the historical interaction sequence of a target-domain user is assumed to be observable. The learned scoring function is then applied to target-domain next-item ranking:

$$f_{\Theta^*} : S^t \times V^t \to \mathbb{R}, \tag{8}$$

which assigns a relevance score $f_{\Theta^*}(\mathbf{s}_{u^t}, v^t)$ to each candidate item $v^t \in V^t$ conditioned on the observed target-domain interaction history $\mathbf{s}_{u^t} \in S^t$, and ranks candidate items accordingly.

4. Method

Under the zero-shot protocol defined in Section 3, the goal is to transfer sequential knowledge learned from an interaction-rich source domain to a target domain without using any target-domain interaction data for optimization. In our setting, supervised learning is performed only on source-domain interaction sequences, while for non-source domains, including the target domain and auxiliary domains, only item-side semantic metadata are accessible and used for interaction-free semantic regularization. To address this problem, we propose **SAGERec**, a **S**imilarity-**A**ware **GE**neralization framework for zero-shot cross-domain sequential recommendation.

As illustrated in Fig. 1, SAGERec builds on a standard LLM-enhanced semantic-sequential transfer pipeline and improves its robustness through two key mechanisms: *Similarity-Aware Inter-domain Compactness* (SIC) and *Domain-Adaptive Generalization* (DAG). The framework consists of four components: (1) an LLM-based item semantic encoding module with lightweight dual projection heads (Section 4.1); (2) a sequential modeling module and final recommendation (Section 4.2); (3) a similarity-aware semantic generalization module for selective inter-domain alignment (Section 4.3); and (4) a domain-adaptive weighting mechanism for discrepancy-aware regularization (Section 4.4).

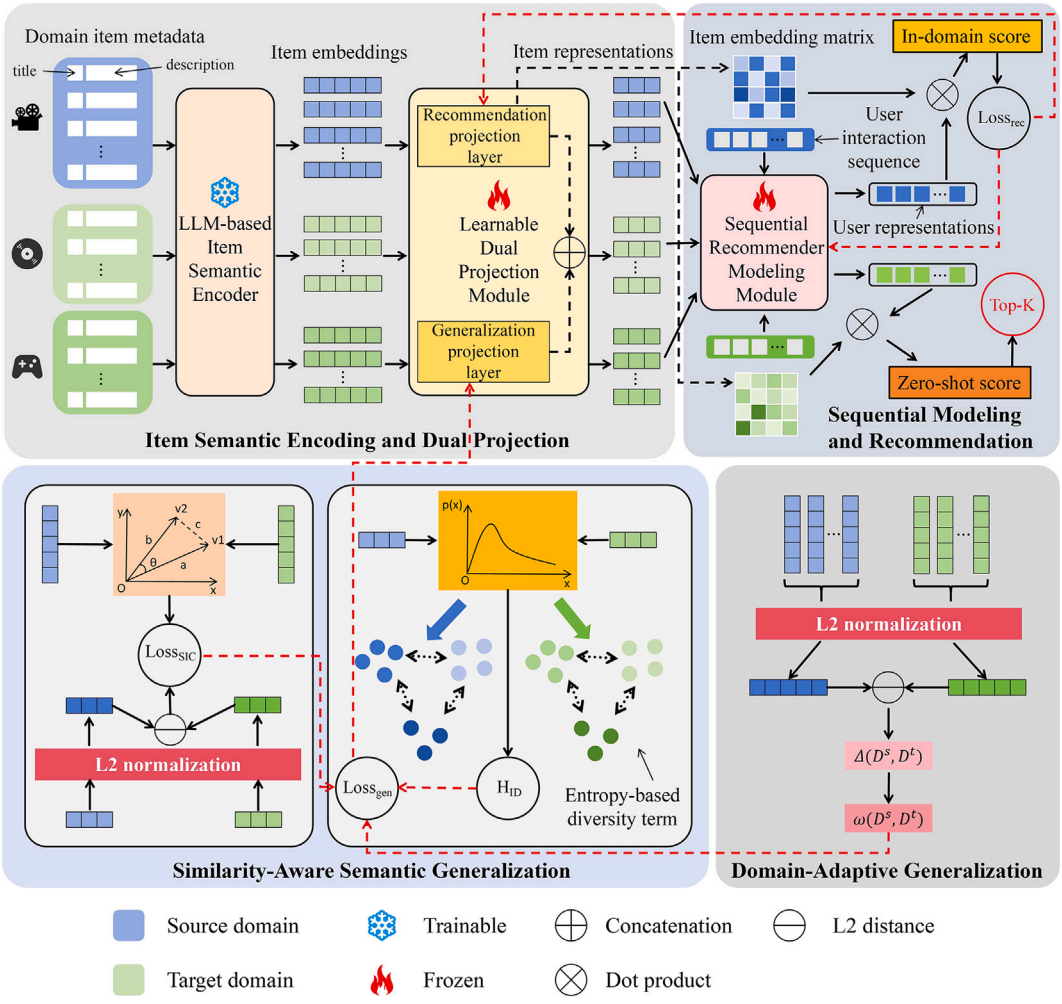


Fig. 1. Overall framework of **SAGERec**.

4.1. Item semantic encoding and dual projection

Let x_i denote the textual metadata of item v_i, such as its title and description text. We first employ a pretrained large language model as the encoder $E(\cdot)$ to extract the semantic representation of item v_i:

$$\mathbf{e}_i^{\text{sem}} = E(x_i), \tag{9}$$

where $\mathbf{e}_i^{\text{sem}} \in \mathbb{R}^{d_h}$, and d_h is the hidden dimension of the semantic encoder. This step provides transferable semantic priors and helps construct a shared semantic space across domains.

Since the raw LLM embedding space is not directly compatible with the downstream sequential recommender, we introduce two lightweight projection heads to map semantic embeddings into task-specific latent spaces:

$$\mathbf{z}_i^{\text{rec}} = \mathbf{W}_r \mathbf{e}_i^{\text{sem}} + \mathbf{b}_r, \tag{10}$$

$$\mathbf{z}_i^{\text{gen}} = \mathbf{W}_g \mathbf{e}_i^{\text{sem}} + \mathbf{b}_g, \tag{11}$$

where $\mathbf{W}_r, \mathbf{W}_g \in \mathbb{R}^{d \times d_h}$ and $\mathbf{b}_r, \mathbf{b}_g \in \mathbb{R}^d$ are trainable parameters. Here, $\mathbf{z}_i^{\text{rec}} \in \mathbb{R}^d$ is the recommendation-oriented item representation, while $\mathbf{z}_i^{\text{gen}} \in \mathbb{R}^d$ is the generalization-oriented item representation used for cross-domain semantic regularization.

To provide a unified item input for the subsequent sequential recommender, we further fuse the two projected representations as

$$\tilde{\mathbf{z}}_i = \mathbf{W}_f \left[\mathbf{z}_i^{\text{rec}} \,\|\, \mathbf{z}_i^{\text{gen}}\right] + \mathbf{b}_f, \tag{12}$$

where $[\cdot \,\|\, \cdot]$ denotes vector concatenation, $\mathbf{W}_f \in \mathbb{R}^{d \times 2d}$ and $\mathbf{b}_f \in \mathbb{R}^d$ are trainable parameters, and $\tilde{\mathbf{z}}_i \in \mathbb{R}^d$ is the fused item representation fed into the sequential modeling module.

Since the source and non-source domains have disjoint item ID spaces, we use distinct indices to avoid confusion. Specifically, source-domain items are indexed by i, while non-source-domain items are indexed by j. Accordingly, their fused representations are denoted by $\tilde{\mathbf{z}}_{i^s}$ and $\tilde{\mathbf{z}}_{j^q}$, respectively, where D^q may refer to the target domain or an auxiliary domain. In this way, the semantic encoder serves as a transferable item initializer [34], while the dual projection heads enable joint optimization for recommendation learning and cross-domain generalization.

4.2. Sequential modeling and recommendation

Given a source-domain user $u^s \in U^s$ with interaction sequence $\mathbf{s}_{u^s} = [v^s_{u^s,1}, v^s_{u^s,2}, \ldots, v^s_{u^s,T_{u^s}}]$, we first replace each interacted item with its fused semantic representation and obtain the source-side input sequence:

$$\tilde{\mathbf{Z}}_{u^s} = [\tilde{\mathbf{z}}_{u^s,1}, \tilde{\mathbf{z}}_{u^s,2}, \ldots, \tilde{\mathbf{z}}_{u^s,T_{u^s}}], \tag{13}$$

where $\tilde{\mathbf{z}}_{u^s,i}$ denotes the fused representation of item $v^s_{u^s,i}$. The sequence is then encoded by a sequential recommender $\text{SeqRec}(\cdot)$ to capture the user's dynamic preference state:

$$\mathbf{h}_{u^s} = \text{SeqRec}(\tilde{\mathbf{Z}}_{u^s}), \tag{14}$$

where $\mathbf{h}_{u^s} \in \mathbb{R}^d$ is the sequence-level user representation. Here, $\mathrm{SeqRec}(\cdot)$ can be implemented using standard sequential recommenders, such as GRU4Rec, SASRec, or BERT4Rec.

Based on the learned sequence representation, the source-domain relevance score for a candidate item $v_i^s \in V^s$ is computed as

$$\mathrm{score}^s(u^s, i) = \mathbf{h}_{u^s}^\top \mathbf{z}_{i^s}^{\mathrm{rec}}, \tag{15}$$

where $\mathbf{z}_{i^s}^{\mathrm{rec}}$ denotes the recommendation-oriented projected representation of item v_i^s. This scoring process constitutes the semantic-sequential recommendation backbone in the source domain.

At inference time, for a target-domain user $u^t \in U^t$ with an observed interaction sequence $\mathbf{s}_{u^t} = [v_{u^t,1}^t, v_{u^t,2}^t, \dots, v_{u^t,T_{u^t}}^t]$, we analogously construct the target-side input sequence

$$\tilde{\mathbf{Z}}_{u^t} = [\tilde{\mathbf{z}}_{u^t,1}, \tilde{\mathbf{z}}_{u^t,2}, \dots, \tilde{\mathbf{z}}_{u^t,T_{u^t}}], \tag{16}$$

and obtain its sequence representation by

$$\mathbf{h}_{u^t} = \mathrm{SeqRec}(\tilde{\mathbf{Z}}_{u^t}). \tag{17}$$

The learned scoring function is then directly transferred to the target domain for zero-shot ranking:

$$\mathrm{score}^t(u^t, j) = \mathbf{h}_{u^t}^\top \mathbf{z}_{j^t}^{\mathrm{rec}}, \tag{18}$$

where $\mathbf{z}_{j^t}^{\mathrm{rec}}$ denotes the recommendation-oriented projected representation of the candidate item $v_j^t \in V^t$.

The above module defines the semantic-sequential transfer pipeline of SAGERec, where source-domain interaction sequences are used to learn sequential preference patterns, and the learned scoring function is directly generalized to target-domain next-item recommendation in a zero-shot manner.

4.3. Similarity-aware semantic generalization

A major challenge in zero-shot transfer is that indiscriminate global alignment may over-compress or distort the latent space, potentially triggering negative transfer when items from the source domain and non-source domains are only weakly related. Instead of enforcing coarse global attraction, we introduce a similarity-aware semantic generalization objective, which selectively aligns semantically compatible inter-domain items while preserving discrimination within each domain.

During training, we sample non-source items from D^{ns} for cross-domain semantic regularization. Importantly, only item-side semantic metadata from non-source domains are used, while no interaction data from these domains is involved in optimization. Therefore, the proposed regularization remains compatible with the zero-shot setting.

4.3.1. Similarity-aware inter-domain compactness

Let $v_i^s \in V^s$ and $v_j^q \in V^q$ denote a source-domain item and a sampled non-source item, respectively, where $D^q \in D^{\mathrm{ns}}$. For each source item and sampled non-source item, their semantic embeddings and projected representations are obtained following Section 4.1. We first measure their semantic affinity based on the original LLM embeddings:

$$a_{ij} = \cos(\mathbf{e}_i^{\mathrm{sem}}, \mathbf{e}_j^{\mathrm{sem}}) = \frac{(\mathbf{e}_i^{\mathrm{sem}})^\top \mathbf{e}_j^{\mathrm{sem}}}{\|\mathbf{e}_i^{\mathrm{sem}}\|_2 \, \|\mathbf{e}_j^{\mathrm{sem}}\|_2}. \tag{19}$$

It should be noted that the LLM-based cosine affinity is used as a soft semantic prior rather than an absolute ground-truth compatibility label. Therefore, SIC does not force all item pairs to be aligned according to raw LLM similarity, but uses the affinity only to modulate selective compactness.

Since not all positive semantic affinities are equally reliable for cross-domain transfer, we further introduce a similarity threshold θ to control the selectivity of SIC. Specifically, the semantic affinity is converted into a non-negative thresholded weight:

$$w_{ij}^{(\theta)} = \max(0, a_{ij} - \theta), \tag{20}$$

where θ controls the activation boundary of cross-domain item pairs. When $a_{ij} > \theta$, the pair is regarded as sufficiently compatible and contributes to the compactness objective with weight $a_{ij} - \theta$; when $a_{ij} \le \theta$, its weight becomes zero and the pair is excluded from direct SIC-based attraction. The original non-negative affinity gate can be recovered by setting $\theta = 0$.

We further apply L2 normalization to the generalization-oriented projected representations before measuring pairwise distances:

$$\bar{\mathbf{z}}_{i^s}^{\mathrm{gen}} = \frac{\mathbf{z}_{i^s}^{\mathrm{gen}}}{\|\mathbf{z}_{i^s}^{\mathrm{gen}}\|_2}, \qquad \bar{\mathbf{z}}_{j^q}^{\mathrm{gen}} = \frac{\mathbf{z}_{j^q}^{\mathrm{gen}}}{\|\mathbf{z}_{j^q}^{\mathrm{gen}}\|_2}. \tag{21}$$

Let $\mathcal{B}^s$ and $\mathcal{B}^q$ denote the source-domain and non-source item sets appearing in the current mini-batch, respectively, and let $\mathcal{P} \subseteq \mathcal{B}^s \times \mathcal{B}^q$ denote the set of sampled cross-domain item pairs. The similarity-aware inter-domain compactness loss is defined as:

$$\mathcal{L}_{\mathrm{SIC}} = \frac{\sum_{(i,j)\in\mathcal{P}} w_{ij}^{(\theta)} \, \|\bar{\mathbf{z}}_{i^s}^{\mathrm{gen}} - \bar{\mathbf{z}}_{j^q}^{\mathrm{gen}}\|_2^2}{\sum_{(i,j)\in\mathcal{P}} w_{ij}^{(\theta)} + \varepsilon}, \tag{22}$$

where $\varepsilon > 0$ is a small constant for numerical stability.

This formulation has two desirable properties. First, activated source–non-source item pairs with larger $w_{ij}^{(\theta)}$ are pulled closer, encouraging consistent and transferable semantics across domains. Second, weakly related or semantically irrelevant pairs receive negligible weights, thereby suppressing noisy attraction caused by excessive uniformity in alignment. Therefore, $\mathcal{L}_{\mathrm{SIC}}$ can be viewed as a selective compactness objective that explicitly considers item-level transfer heterogeneity.

4.3.2. Intra-domain diversity preservation

Although inter-domain compactness is beneficial for semantic transfer, overly strong attraction may compress the representation space and weaken item discrimination within each domain. To alleviate such representation collapse, we introduce an intra-domain diversity preservation term.

For each domain $\kappa \in \{s, q\}$, let $\mathcal{B}^{(\kappa)}$ denote the set of indices of items from domain κ. We then define a similarity-based probability distribution $p_{mn}^{(\kappa)}$ over items in the mini-batch $\mathcal{B}^{(\kappa)}$ as:

$$p_{mn}^{(\kappa)} = \frac{\exp(\cos(\mathbf{z}_m^{(\kappa),\mathrm{gen}}, \mathbf{z}_n^{(\kappa),\mathrm{gen}})/\tau_\kappa)}{\sum_{b\in\mathcal{B}^{(\kappa)}} \exp(\cos(\mathbf{z}_m^{(\kappa),\mathrm{gen}}, \mathbf{z}_b^{(\kappa),\mathrm{gen}})/\tau_\kappa)}, \tag{23}$$

where $\tau_\kappa > 0$ is a temperature parameter controlling the smoothness of the distribution, and $\mathbf{z}_m^{(\kappa),\mathrm{gen}}, \mathbf{z}_n^{(\kappa),\mathrm{gen}}$ denote the generalization-oriented projected representations of the items indexed by $m, n \in \mathcal{B}^{(\kappa)}$, respectively.

In practice, this distribution is estimated within the current mini-batch for computational efficiency, serving as a stochastic surrogate of full-domain intra-domain diversity. Across training iterations, randomly sampled mini-batches expose the model to different local subsets of each domain, allowing the diversity regularization to accumulate over multiple stochastic observations.

Based on this distribution, we define the entropy-based diversity term as

$$\mathcal{H}_{\mathrm{ID}} = -\sum_{\kappa\in\{s,q\}} \frac{1}{|\mathcal{B}^{(\kappa)}|} \sum_{m\in\mathcal{B}^{(\kappa)}} \sum_{n\in\mathcal{B}^{(\kappa)}} p_{mn}^{(\kappa)} \log p_{mn}^{(\kappa)}. \tag{24}$$

A larger $\mathcal{H}_{\mathrm{ID}}$ corresponds to a higher-entropy intra-domain similarity distribution and hence better preservation of intra-domain diversity. This term counterbalances overly aggressive inter-domain attraction and helps maintain fine-grained semantic discrimination.

4.4. Domain-adaptive generalization

Besides item-level heterogeneity, zero-shot transfer also exhibits domain-level heterogeneity: different non-source domains may vary significantly in their semantic distance from the source domain. Using a fixed regularization strength for all transfer scenarios is therefore suboptimal. To address this issue, we introduce a domain-adaptive weighting strategy to enhance semantic generalization.

Specifically, for the current mini-batch, we compute the mean semantic embeddings of the source domain and the sampled non-source query domain:

$$\bar{\mathbf{e}}_s^{\text{sem}} = \frac{\sum_{i\in\mathcal{B}^s} \mathbf{e}_i^{\text{sem}}}{\|\sum_{i\in\mathcal{B}^s} \mathbf{e}_i^{\text{sem}}\|_2}, \qquad \bar{\mathbf{e}}_q^{\text{sem}} = \frac{\sum_{j\in\mathcal{B}^q} \mathbf{e}_j^{\text{sem}}}{\|\sum_{j\in\mathcal{B}^q} \mathbf{e}_j^{\text{sem}}\|_2}. \tag{25}$$

Moreover, the semantic discrepancy between the two domains is measured by

$$\Delta(D^s, D^q) = \|\bar{\mathbf{e}}_s^{\text{sem}} - \bar{\mathbf{e}}_q^{\text{sem}}\|_2. \tag{26}$$

Based on the discrepancy obtained by Eq. (26), we define a domain-adaptive generalization coefficient as follows:

$$\omega(D^s, D^q) = \lambda \exp\big(-\gamma \Delta(D^s, D^q)\big), \tag{27}$$

where $\lambda > 0$ controls the base strength of semantic regularization, and $\gamma > 0$ controls the sensitivity of the decay with respect to the domain discrepancy.

This weighting mechanism yields a monotonic relationship between semantic proximity and regularization strength. Specifically, when the source domain and the query domain are semantically close, $\Delta(D^s, D^q)$ is small and $\omega(D^s, D^q)$ remains relatively large, imposing more rigorous semantic constraints on domains with high proximity. Conversely, when the semantic gap is large, $\omega(D^s, D^q)$ is reduced, thus encouraging the model to behave more conservatively and avoid over-aligning distant domains. Therefore, by sampling different query domains $D^q \in \mathcal{D}^{\text{ns}}$ during training, SAGERec approximates heterogeneous transfer difficulty and dynamically adjusts the regularization strength accordingly.

4.5. Training objective

To optimize the generalization-oriented projection space for cross-domain semantic alignment and intra-domain diversity preservation, the semantic generalization term is defined as:

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{SIC}} - \beta \mathcal{H}_{\text{ID}}, \quad (\beta > 0) \tag{28}$$

where β is a hyper-parameter to balance inter-domain selective compactness and intra-domain diversity preservation.

To optimize the model on the source domain, we adopt a standard pairwise ranking loss:

$$\mathcal{L}_{\text{rec}} = -\sum_{(u^s, v^{s,+}, v^{s,-})} \log \sigma\big(\text{score}^s(u^s, v^{s,+}) - \text{score}^s(u^s, v^{s,-})\big), \tag{29}$$

where $(u^s, v^{s,+}, v^{s,-})$ denotes a training triplet consisting of user u^s, a positive item $v^{s,+}$, and a sampled negative item $v^{s,-}$, and $\sigma(\cdot)$ is the sigmoid function.

Overall, the training objective is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \omega(D^s, D^q)\, \mathcal{L}_{\text{gen}}. \tag{30}$$

Overall, SAGERec is a lightweight enhancement to the standard semantic-sequential transfer pipeline, not a replacement for the backbone itself. It combines similarity-aware selective compactness at the item-pair level with domain-adaptive weighting at the domain level, enabling more robust zero-shot cross-domain transfer across diverse semantic gaps while maintaining full compatibility with existing sequential recommenders.

5. Experiments

In this section, we conduct comprehensive experiments to evaluate the proposed **SAGERec** framework. Our experiments are designed to answer the following research questions:

- **RQ1**: How effective is SAGERec for zero-shot cross-domain sequential recommendation compared with strong semantic and generalization-based baselines?
- **RQ2**: What are the individual contributions of similarity-aware inter-domain compactness and domain-adaptive generalization?
- **RQ3**: How sensitive is SAGERec to its key hyperparameters and training settings, including the SIC similarity threshold and the mini-batch size?
- **RQ4**: How efficient is SAGERec compared with representative baselines in terms of computational cost and effectiveness–efficiency trade-off?
- **RQ5**: How does SAGERec affect cross-domain semantic alignment and representation quality?
- **RQ6**: How does the choice of semantic encoder backbone affect SAGERec?
- **RQ7**: What do representative item-pair cases reveal about the behavior and limitations of semantic alignment in SAGERec?

5.1. Experimental setup

5.1.1. Datasets

We evaluate **SAGERec** on three public real-world datasets from different domains: **Amazon Movies & TV (AMT)**, **Amazon CDs & Vinyl (ACV)**, and **Steam**. Following prior work [35], AMT and ACV are selected from the Amazon Review corpus, while Steam is adopted as a widely used benchmark dataset for sequential recommendation. All three datasets provide sequential user–item interactions together with item-side textual metadata, which makes them suitable for evaluating recommendation models that jointly exploit behavioral signals and semantic item information.

Domain Split. The dataset split follows the zero-shot cross-domain setting considered in this paper. AMT is the largest dataset in terms of both users and items and is therefore used as the source domain with relatively rich interaction signals. ACV and Steam are treated as non-source domains. In each transfer scenario, one of them is used as the evaluation target domain, while the other serves as an auxiliary domain that provides item-side semantic information for interaction-free regularization. Consequently, the two zero-shot tasks considered in this paper are AMT → ACV (with Steam as the auxiliary domain) and AMT → Steam (with ACV as the auxiliary domain). This split also covers different transfer gaps: ACV shares the Amazon platform and media-related metadata style with AMT, whereas Steam comes from a different platform and contains game-oriented item metadata, making AMT → Steam a relatively more distant cross-platform transfer scenario.

Data Preprocessing. Following prior studies [5,36–38], we apply 10-core filtering to remove users and items with fewer than 10 interactions, and discard items without valid textual metadata required for semantic encoding. The remaining interactions of each user are then sorted chronologically to form behavior sequences. We adopt the leave-one-out protocol, where the last interaction of each user is used for testing, the penultimate one for validation, and the remaining interactions for training. During evaluation, each ground-truth target item is ranked against 100 randomly sampled negative items. Table 1 reports the statistics of the processed datasets.

5.1.2. Baseline methods

To comprehensively evaluate the effectiveness of SAGERec, we compare it with seven semantic-transfer baseline models under a unified experimental protocol. Following the baseline design of LLM-RecG [8], we select models that are directly comparable within the same semantic-sequential transfer pipeline, ensuring that performance gains can be

Table 1
Statistics of the processed datasets after preprocessing.

Dataset	Items	Users	Avg. Seq. Len.
Amazon Movies & TV (AMT)	84,909	179,784	18.6
Amazon CDs & Vinyl (ACV)	20,089	22,170	19.8
Steam	8156	26,541	19.7

clearly attributed to our generalization strategy rather than architectural differences. Specifically, we build our comparisons around three representative sequential backbones and their semantic variants. In addition, we include a recent universal sequence representation method as a strong text-enhanced transferable recommendation baseline.

Sequential Backbones. We adopt three representative sequential recommenders as backbone architectures for constructing semantic transfer baselines. Specifically, **GRU4Rec** [3] represents recurrent sequential modeling, **SASRec** [5] represents unidirectional self-attentive sequential modeling, and **BERT4Rec** [6] represents bidirectional Transformer-based sequence encoding. These backbones cover the major modeling paradigms in sequential recommendation and provide diverse foundations for evaluating the robustness of the proposed framework.

Semantic Variants (-Sem). For the *-Sem* baselines, we construct **GRU4Rec-Sem**, **SASRec-Sem**, and **BERT4Rec-Sem**. These variants incorporate item-side textual metadata through the same LLM-based semantic encoding and projection pipeline, but do not introduce any additional generalization objective. They are used to evaluate the contribution of semantic enhancement alone on different sequential backbones.

Generalization Variants (-RecG). For the *-RecG* baselines, we construct **GRU4Rec-RecG**, **SASRec-RecG**, and **BERT4Rec-RecG**. These variants further introduce the recommendation generalization mechanism on top of the corresponding *-Sem* models, so as to enhance the transferability of learned item representations in the zero-shot target domain. The comparison between each *-Sem* variant and its corresponding *-RecG* variant reveals the contribution of additional generalization-oriented optimization.

Universal Sequence Representation. We also compare with **UniSRec** [23], a strong universal sequence representation learning method that leverages textual semantics to improve recommendation transferability across scenarios. As a competitive text-enhanced recommendation baseline beyond the six backbone-derived variants, UniSRec provides an important external reference in our experiments. We use UniSRec in its original transferable representation-learning form and do not construct additional *-Sem*, *-RecG*, or *-SAGE* variants for it. This is because UniSRec is not a plain sequential backbone to which the same semantic projection and generalization modules can be attached; it already integrates text-derived item representations with an MoE adapter for universal sequence representation.

5.1.3. Evaluation metrics

We adopt Recall@K and NDCG@K as evaluation metrics, with $K \in \{10, 20\}$. Recall@K measures whether the ground-truth item appears in the top-K recommendation list, while NDCG@K further evaluates the ranking quality by assigning higher scores to correct items ranked closer to the top positions. Since each test instance contains one positive item and multiple sampled negatives, Recall@K is equivalent to Hit Ratio@K in this setting. Higher values of both metrics indicate better recommendation performance.

5.1.4. Implementation details

All models are implemented in Python 3.10.19 with PyTorch 2.6.0 and optimized using Adam on a server equipped with an NVIDIA RTX 5880 Ada GPU. For semantic encoding, we employ LLM2Vec [39] based on Llama-3-8B-Instruct [40] to transform item metadata into 4096-dimensional dense vectors, which are then projected into the recommendation latent space. Before semantic encoding, each item's available metadata is converted into a prompt-formatted text string. For Amazon items, the prompt follows the template "Title: {title}. Features: {feature}. Description: {description}."; for Steam items, the same field-ordering rule is applied to the available product-level textual fields. Missing fields are filled with a predefined placeholder "N/A", and the same prompt construction rule is applied to all semantic-aware methods. The semantic encoder is used only for offline item encoding and is kept frozen during recommendation training; the resulting item embeddings are pre-computed once and reused by all semantic-aware methods. Unless otherwise specified, we set the hidden size to 256, the maximum sequence length to 50, and the dropout rate to 0.2. For Transformer-based backbones, we use 2 layers and 2 attention heads. Models are trained with a batch size of 128 and a learning rate of 1×10^{-4} for at most 100 epochs.

For the loss-balancing hyperparameters (λ, γ, β), we perform grid search following the validation protocol described below. In the main comparison and ablation studies, SIC uses the default threshold $\theta = 0$, which recovers the non-negative semantic affinity gate in the original formulation. The candidate ranges for (λ, γ, β), together with diagnostic sensitivity analyses of θ and the mini-batch size, are reported in the sensitivity analysis.

In each transfer scenario, model selection and early stopping are performed according to the best validation performance on the auxiliary domain, while the final reported results are evaluated on the corresponding target domain. This protocol is designed to avoid using target-domain interaction signals for model selection and thus remains consistent with the zero-shot setting. For the main comparison experiments, all methods are evaluated with fixed random seeds, and the final comparison results are reported as the average over three independent runs.

5.1.5. Fair comparison and reproducibility

To ensure a fair comparison, all methods use the same train/validation/test split, the same negative sampling strategy, and the same evaluation metrics. More importantly, all semantic-aware methods, including SAGERec and the baselines, are given access to exactly the same item-side semantic information from all domains, while no interaction data from the target domain is used for optimization. For baselines that require semantic representations, we further use the same textual prompts, the same pre-computed semantic embeddings and the same preprocessing pipeline whenever applicable. Therefore, performance differences can be attributed to the proposed similarity-aware compactness and domain-adaptive generalization mechanisms rather than to discrepancies in accessible information, data preprocessing, or training protocol. The implementation, preprocessing scripts, and experimental configurations are publicly available at https://github.com/Zihao-Pan-666/SAGERec.

5.2. Overall comparison (RQ1)

Table 2 reports the overall zero-shot recommendation performance when AMT is used as the source domain and ACV and Steam are treated as unseen target domains. To more clearly assess the contribution of the proposed framework itself, the results are organized by sequential backbone, so that each *-SAGE* variant can be directly compared with its corresponding *-Sem* and *-RecG* counterparts under the same architecture. Overall, SAGERec consistently improves performance within all three backbone families, and BERT4Rec-SAGE achieves the best average result among all compared models.

(1) *Consistent family-wise gains.* Across all three sequential backbones, the *-SAGE* variant achieves the best performance within its corresponding model family. For the GRU4Rec backbone, the average score improves from 16.79 (*-Sem*) and 17.55 (*-RecG*) to 18.74 (*-SAGE*). For SASRec, the average score further increases

Table 2
Overall zero-shot recommendation performance on two zero-shot transfer tasks. Results are reported in percentage (%). To evaluate the model-agnostic effectiveness of the proposed framework, variants are grouped by their sequential backbones. Underlined values denote the best-performing variant within each backbone family, while **bold** values denote the best overall results in each column. "Avg." represents the arithmetic mean across all metrics and domains.

Models	ACV				Steam				Avg.
	R@10	N@10	R@20	N@20	R@10	N@10	R@20	N@20	
UniSRec	40.16	21.14	58.14	25.67	20.42	10.15	34.65	13.71	28.01
GRU4Rec-Sem	17.16	8.48	30.32	11.78	16.76	8.65	29.36	11.80	16.79
GRU4Rec-RecG	17.82	8.90	31.21	12.25	18.06	9.12	30.71	12.29	17.55
GRU4Rec-SAGE	19.65	9.85	33.13	13.23	20.46	10.01	30.95	12.64	18.74
SASRec-Sem	38.40	20.65	55.71	25.02	21.54	10.79	35.53	14.30	27.74
SASRec-RecG	42.25	22.61	59.58	26.99	22.75	11.44	37.91	15.23	29.85
SASRec-SAGE	45.50	25.20	62.70	29.54	24.30	**12.31**	38.10	**15.78**	31.68
BERT4Rec-Sem	39.74	21.49	56.62	25.74	22.39	11.29	35.94	14.69	28.49
BERT4Rec-RecG	40.65	22.18	57.89	26.52	22.51	11.00	36.34	14.47	28.95
BERT4Rec-SAGE	**45.99**	**25.43**	**62.93**	**29.71**	**24.45**	12.15	**38.36**	15.65	**31.83**

from 27.74 and 29.85 to 31.68. A similar trend is observed for BERT4Rec, where *-SAGE* reaches the highest average score of 31.83, compared with 28.49 for *-Sem* and 28.95 for *-RecG*. These results indicate that the gains of SAGERec are not merely due to stronger backbones, but can also be consistently observed within the same backbone architecture.

(2) *Advantages over uniform generalization.* Comparing each *-SAGE* variant with its corresponding *-RecG* counterpart further suggests that the proposed similarity-aware and discrepancy-adaptive design is more effective than a uniform cross-domain generalization strategy. The advantage is particularly clear on the Transformer-based backbones. For SASRec, the *-SAGE* variant outperforms *-RecG* on both ACV and Steam across the evaluated metrics. The same trend is observed for BERT4Rec, where the gains are especially notable on ACV and remain positive on Steam. On GRU4Rec, the improvement is also consistent, but the absolute margins are relatively smaller than those on SASRec and BERT4Rec. This indicates that selective cross-domain alignment is generally more beneficial than uniform semantic attraction, although the magnitude of the gain still depends on backbone expressiveness.

(3) *Stronger benefits on expressive backbones.* The benefit of SAGERec becomes more substantial on stronger sequential backbones. In terms of average performance, SAGERec improves over the corresponding *-RecG* baseline by 1.19 points on GRU4Rec, 1.83 points on SASRec, and 2.88 points on BERT4Rec. This pattern suggests that more expressive sequential encoders are better able to exploit the transferable semantic structure introduced by SAGERec. At the same time, the table also shows that the advantage is not equally large in every setting. In particular, the margins within the GRU4Rec family are visibly smaller, indicating that a weaker sequential encoder may limit the extent to which semantic regularization can be translated into recommendation gains.

(4) *Overall competitiveness against a strong external baseline.* Beyond the within-family comparisons, UniSRec serves as a strong external transferable baseline. The results show that both SASRec-SAGE and BERT4Rec-SAGE outperform UniSRec in terms of average performance, while BERT4Rec-SAGE achieves the best overall average score of 31.83 among all compared models. More specifically, BERT4Rec-SAGE obtains the best results on all four metrics in the AMT→ACV task, as well as the best Recall@10 and Recall@20 in the AMT→Steam task. Meanwhile, SASRec-SAGE achieves the best NDCG@10 and NDCG@20 on Steam, indicating slightly better ranking quality at the top positions in that scenario.

Table 3
Ablation study of SAGERec on different sequential backbones. AMT is used as the source domain, while ACV and Steam are treated as unseen target domains for zero-shot transfer. "Avg." denotes the average of ACV/Steam results on R@10 and N@10. All results are reported in percentage (%).

Backbone	Variant	ACV		Steam		Avg.
		R@10	N@10	R@10	N@10	
GRU4Rec	**SAGE**	**19.65**	**9.85**	**20.46**	**10.01**	**14.99**
	w/o SIC	17.79	8.95	17.02	8.92	13.17
	w/o DAG	16.91	8.42	18.05	9.17	13.14
	RecG	17.82	8.90	18.06	9.12	13.47
SASRec	**SAGE**	**45.50**	**25.20**	**24.30**	**12.31**	**26.83**
	w/o SIC	43.26	23.67	22.20	11.19	25.08
	w/o DAG	44.67	25.02	23.61	11.67	26.24
	RecG	42.25	22.61	22.75	11.44	24.76
BERT4Rec	**SAGE**	**45.99**	**25.43**	**24.45**	12.15	**27.01**
	w/o SIC	41.37	22.17	24.23	**12.16**	24.98
	w/o DAG	43.74	24.03	23.39	**12.16**	25.83
	RecG	40.65	22.18	22.51	11.00	24.09

Overall, these results support the effectiveness and robustness of SAGERec in zero-shot cross-domain sequential recommendation.

5.3. Ablation study (RQ2)

To evaluate the contribution of the proposed components, we conduct ablation studies on three representative sequential backbones. Following prior work [8], the intra-domain diversity regularization is retained in all variants to ensure a fair comparison. We compare the full SAGERec model with two degraded variants (*w/o SIC* and *w/o DAG*) as well as the corresponding *-RecG* baseline. The results are summarized in Table 3.

Overall, the full SAGERec model achieves the best average (Avg.) performance across all backbones, indicating that the proposed components contribute positively and work in a complementary manner. Two observations can be drawn.

(1) *Effect of SIC.* Removing the SIC module leads to the most noticeable performance degradation in most cases, especially on the Transformer-based backbones. On BERT4Rec, the average performance drops from 27.01 to 24.98 when SIC is removed, corresponding to a relative decrease of about 7.5%. Similar trends can be observed on SASRec and GRU4Rec, where the Avg. score

decreases from 26.83 to 25.08 and from 14.99 to 13.17, respectively. These results suggest that item-level selective alignment plays a central role in effective zero-shot transfer, as it directly filters out weakly related cross-domain correspondences that might otherwise introduce semantic noise.

(2) *Effect of DAG.* Removing DAG also results in consistent performance drops across different backbones, although the magnitude is generally smaller than that of removing SIC. For example, the Avg. score decreases from 27.01 to 25.83 on BERT4Rec and from 26.83 to 26.24 on SASRec when DAG is removed. On GRU4Rec, the corresponding drop is from 14.99 to 13.14. This indicates that discrepancy-aware weighting further improves transfer robustness by adjusting the strength of semantic regularization according to domain differences, even though its contribution is typically more moderate than that of SIC.

It is also observed that the full SAGERec model does not strictly outperform all degraded variants on every individual metric. For instance, on BERT4Rec, the N@10 result on Steam (12.15) is slightly lower than that of both *w/o SIC* and *w/o DAG* (12.16). However, such minor fluctuations are negligible compared with the overall average gains across backbones and domains. Overall, the ablation results show that SIC and DAG provide complementary benefits, with SIC serving as the primary source of improvement and DAG acting as a stabilizing factor for heterogeneous cross-domain transfer.

5.4. Sensitivity analysis (RQ3)

We further investigate the robustness of SAGERec with respect to key hyperparameters and training settings. Specifically, we analyze three aspects: the main loss-balancing hyperparameters (λ, β, γ), the SIC similarity threshold θ, and the mini-batch size. Unless otherwise specified, all sensitivity analyses are conducted on BERT4Rec-SAGE, and the results are averaged over the two zero-shot transfer scenarios, i.e., AMT→ACV and AMT→Steam. These analyses are intended to diagnose the robustness of SAGERec under controlled settings, rather than to select the final models reported in the main comparison and ablation studies.

5.4.1. Sensitivity to main hyperparameters

We examine the sensitivity of SAGERec to three key hyperparameters: the base regularization strength λ, the intra-domain diversity weight β, and the semantic decay rate γ. To isolate the effect of each parameter while controlling computational cost, we adopt a coordinate-wise evaluation strategy on the BERT4Rec-based variant, varying one hyperparameter at a time while fixing the others at their default values. Specifically, we evaluate $\lambda \in \{0.005, 0.010, 0.050\}$, $\beta \in \{0.1, 0.3, 0.7\}$, and $\gamma \in \{0.05, 0.10, 0.20\}$. The results are shown in Fig. 2.

Overall, SAGERec shows relatively stable behavior within the tested ranges, while still exhibiting meaningful sensitivity to the three hyperparameters. Several observations can be drawn.

- **Base regularization strength (λ).** Performance improves steadily as λ increases within the tested range, and the best results are obtained at $\lambda = 0.05$ for both Recall@10 and NDCG@10. This trend suggests that stronger semantic regularization within the considered range is beneficial for reducing cross-domain discrepancy and improving zero-shot transfer.
- **Intra-domain diversity weight (β).** The best performance is obtained at the smallest tested value, $\beta = 0.1$, while larger values lead to a consistent decline in both metrics. This result indicates that a small diversity-preservation weight is sufficient in this setting, whereas assigning it an excessively large weight may over-constrain the representation space and reduce the flexibility required for effective cross-domain alignment.
- **Semantic decay rate (γ).** The performance follows a non-monotonic pattern, peaking at $\gamma = 0.10$. When γ is too large, the weighting scheme becomes overly selective and may under-utilize transferable semantic relations from moderately distant domains; conversely, when γ is too small, the selective effect is weakened, and the weighting becomes nearly uniform, which may introduce noisy correspondences from semantically distant domains. This pattern is consistent with the intended role of γ in balancing selectivity and coverage in cross-domain alignment.

5.4.2. Sensitivity to the SIC similarity threshold

We further analyze the sensitivity of SAGERec to the similarity threshold θ in the SIC module. This controlled diagnostic analysis is conducted on the BERT4Rec-SAGE variant by varying θ in $\{-0.05, 0.00, 0.05, 0.10, 0.20\}$ while keeping the remaining settings fixed. The results are averaged over the two target domains, i.e., AMT→ACV and AMT→Steam.

As shown in Fig. 3, the performance curve generally peaks at a mildly positive threshold, $\theta = 0.05$, indicating that SIC can benefit from filtering out weakly related cross-domain item pairs. The default setting $\theta = 0$ also remains competitive, which suggests that the original non-negative affinity gate is already a reasonable and robust choice.

The results reveal a trade-off between semantic coverage and alignment selectivity. According to the thresholded weighting in Eq. (20), a smaller threshold activates more cross-domain item pairs and thus provides broader semantic coverage, but it may also admit weakly related pairs caused by superficial textual overlap, generic category words, or platform-specific metadata patterns. In contrast, a mildly positive threshold such as $\theta = 0.05$ filters out near-zero weak affinities and increases the relative emphasis on more reliable semantic correspondences, which explains its slight advantage in this controlled diagnostic analysis. However, when θ is further increased to 0.10 or 0.20, the activated pair set becomes too sparse, and SIC may discard moderately related but still transferable pairs. Therefore, the threshold should balance coverage and selectivity. Although $\theta = 0.05$ achieves the best result in this diagnostic sweep, $\theta = 0$ remains a competitive setting and

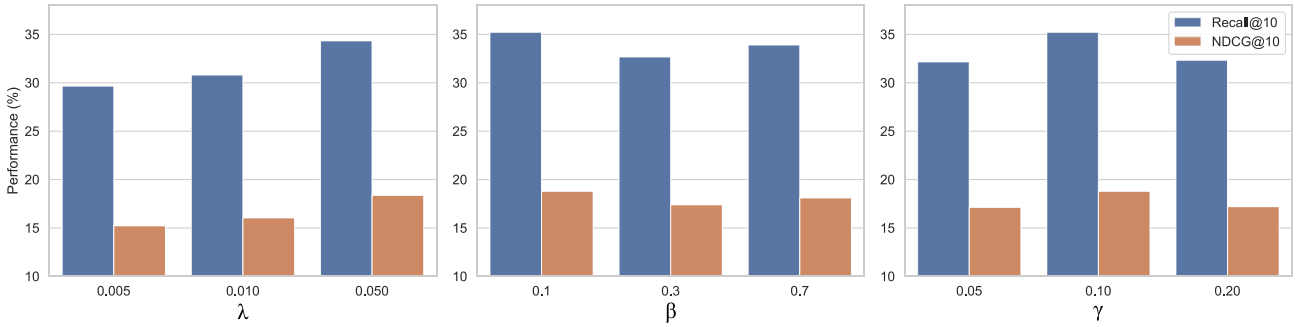


Fig. 2. Sensitivity analysis of the key hyperparameters (λ, β, and γ) for BERT4Rec-SAGE. Results are reported in terms of average Recall@10 and average NDCG@10 over the two target domains.

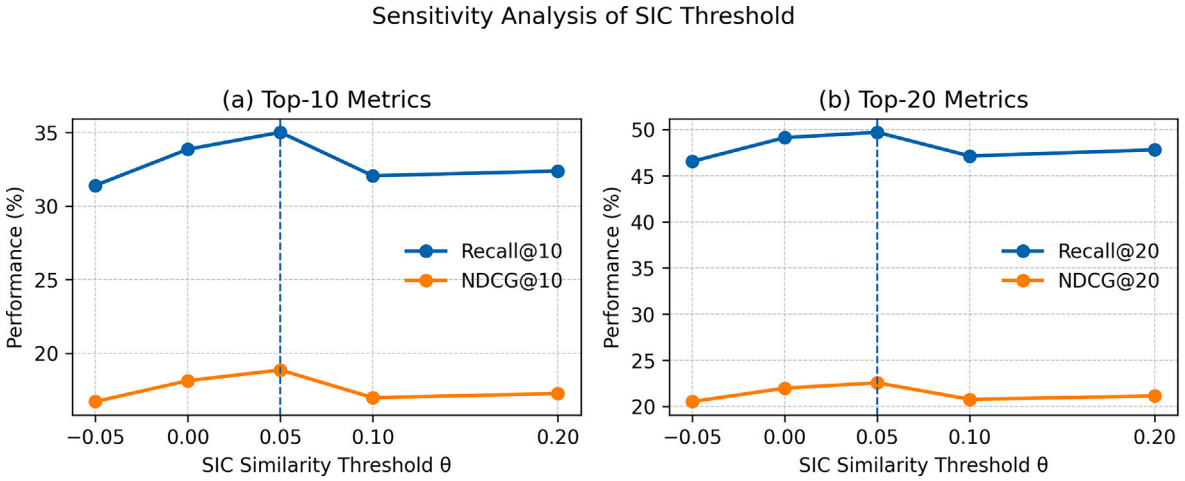


Fig. 3. Sensitivity analysis of the SIC similarity threshold θ on BERT4Rec-SAGE. Metrics are averaged over the two target domains, i.e., AMT→ACV and AMT→Steam.

corresponds exactly to the non-negative semantic affinity gate used in the original formulation. Therefore, the main comparison and ablation studies report results under $\theta = 0$, while the sensitivity curve further shows that mildly positive thresholds can be explored when stronger alignment selectivity is desired.

5.4.3. Sensitivity to the mini-batch size

Since the intra-domain diversity term is estimated from items in the current mini-batch, we further examine whether SAGERec is sensitive to the mini-batch size. Specifically, we vary the batch size in $\{32, 64, 128, 256, 512\}$ on the BERT4Rec-SAGE variant while keeping the other hyperparameters fixed. To remain consistent with the main experimental setting, the SIC threshold is fixed to $\theta = 0$ in this analysis. The reported metrics are averaged over the two target domains, i.e., AMT→ACV and AMT→Steam.

As shown in Table 4, the best overall performance is obtained at batch size 128. When the batch size is too small, such as 32, the performance drops noticeably, supporting the concern that small mini-batches may provide a less representative estimate of the intra-domain similarity structure. Increasing the batch size from 32 to 64 and 128 consistently improves the averaged performance, suggesting that a moderately larger mini-batch can provide a more reliable local diversity estimate.

However, further increasing the batch size does not lead to continuous improvement. The performance slightly decreases at a batch size of 256 and drops more clearly at a batch size of 512. This suggests that simply using a very large mini-batch is unnecessary and may even be less favorable, possibly because very large batches reduce the number of parameter updates per epoch and weaken the stochasticity of mini-batch training. Overall, SAGERec remains reasonably stable within the practical batch-size range from 64 to 256, and the default batch size of 128 provides a favorable balance between recommendation accuracy and diversity-estimation stability.

Table 4
Sensitivity analysis with respect to the mini-batch size on BERT4Rec-SAGE. Results are averaged over the two target domains, i.e., AMT→ACV and AMT→Steam. "Avg." denotes the arithmetic mean over the four reported metrics. All values are reported in percentage (%).

Batch Size	R@10	N@10	R@20	N@20	Avg.
32	31.65	16.90	46.99	20.77	29.08
64	33.50	18.07	48.59	21.86	30.51
128	**35.22**	**18.79**	**50.65**	**22.68**	**31.83**
256	34.13	18.09	49.38	21.94	30.89
512	31.43	16.57	47.28	20.56	28.96

5.5. Efficiency analysis (RQ4)

Efficiency is an important practical factor in zero-shot cross-domain sequential recommendation, because a useful transfer method should improve transferability without introducing prohibitive computational overhead. In this subsection, we analyze the efficiency of SAGERec from three complementary perspectives: offline LLM-based item encoding cost, training-stage cost and convergence behavior, and the effectiveness–efficiency trade-off. For training-stage measurements and recommendation performance, we focus on the BERT4Rec backbone and average the results over the two transfer scenarios where AMT is used as the source domain and ACV and Steam are used as target domains. The LLM encoding cost is reported separately at the dataset level, since semantic encoding is performed once as offline preprocessing and the resulting embeddings are reused by all semantic-aware methods.

5.5.1. Offline LLM-based item encoding cost

We first report the offline cost of obtaining LLM-based item representations using the same LLM2Vec encoder described in the implementation details. This cost is separated from the per-epoch training cost because item metadata are encoded only once as offline preprocessing. The resulting semantic embeddings are cached and reused by all semantic-aware variants during training and inference, so the recommendation model does not repeatedly call the LLM during optimization.

As shown in Table 5, encoding the three datasets takes 262.15min in total, excluding the shared model loading time. The per-item encoding time is around 139 ms on average, and the cached embeddings require about 2.54 GB of storage. Although this preprocessing stage is more expensive than ordinary ID-based embedding lookup, it is a one-off and reusable cost shared by *-Sem*, *-RecG*, and *-SAGE*. During recommendation training, *-SAGE* operates on cached semantic embeddings and mini-batch item representations, so the LLM encoding cost should be interpreted separately from the training-stage SIC/DAG regularization overhead.

Table 5
Offline LLM-based item encoding cost. The one-time model loading time of 22.56 s is excluded. Storage denotes the on-disk size of cached semantic embedding files.

Dataset	#Items	Total (min)	Enc./item (ms)	Cached emb. storage (MB)
AMT	84,909	185.77	131.27	1844.58
ACV	20,089	57.90	172.93	492.96
Steam	8156	18.48	135.95	198.54
Total	113,154	262.15	139.01	2536.08

Table 6
Training-stage cost on the BERT4Rec backbone. Results are averaged over AMT→ACV and AMT→Steam.

Method	Avg./epoch (s)	Avg. epochs	Total (min)
Sem	35.12	64	46.08
RecG	113.12	50	109.65
SAGE	50.97	48	52.69

5.5.2. Training cost and convergence behavior

We next examine the training-stage cost of different semantic-aware variants on the BERT4Rec backbone. Table 6 reports the average training time per epoch, the average number of training epochs, and the total wall-clock training time measured from logs under the same experimental protocol. These metrics distinguish the optimization cost per epoch from the end-to-end wall-clock cost required to complete training, including validation and early-stopping overhead.

As shown in Table 6, *-RecG* incurs the largest training cost, requiring 113.12 s per epoch and 109.65min in total. In contrast, *-SAGE* requires 50.97 s per epoch and 52.69min in total, reducing the per-epoch time by 54.94% and the total training time by 51.94% compared with *-RecG*. Since *-RecG* and *-SAGE* require a comparable number of training epochs, the efficiency advantage of *-SAGE* mainly comes from its lower per-epoch computational overhead.

Compared with the lightweight *-Sem* variant, *-SAGE* introduces additional per-epoch computation because it performs SIC- and DAG-based semantic regularization. However, this extra cost is partly offset by faster convergence: *-SAGE* uses fewer training epochs on average than *-Sem*, reducing the gap in total wall-clock training time. Specifically, although the average epoch time increases from 35.12 s to 50.97 s, the average number of epochs decreases from 64 to 48, and the total training time only increases from 46.08min to 52.69min. This suggests that *-SAGE* achieves a practical training cost by using selective similarity-aware regularization over cached semantic embeddings, instead of performing dense cross-domain generalization as in *-RecG*.

5.5.3. Effectiveness–efficiency trade-off

Finally, we analyze whether the additional computational cost of *-SAGE* is justified by its recommendation improvements. Fig. 4 visualizes the trade-off between total wall-clock training time and average recommendation performance. The average recommendation score is consistent with the "Avg." metric in Table 2, i.e., the arithmetic mean of R@10, N@10, R@20, and N@20 after averaging over AMT→ACV and AMT→Steam. In addition to the BERT4Rec-based *-Sem*, *-RecG*, and *-SAGE* variants, UniSRec is included as an external universal sequence representation baseline for reference.

As shown in Fig. 4, BERT4Rec-SAGE provides a favorable balance between effectiveness and efficiency. Compared with BERT4Rec-Sem, BERT4Rec-SAGE improves the average recommendation score from 28.49 to 31.83, while the total wall-clock training time increases only moderately from 46.08min to 52.69min. This indicates that although *-SAGE* introduces additional semantic regularization, the extra cost is accompanied by a substantial improvement.

Compared with BERT4Rec-RecG, BERT4Rec-SAGE is both more effective and more efficient. It improves the average recommendation score from 28.95 to 31.83, while reducing the total training time from 109.65min to 52.69min, corresponding to a 51.94% reduction. This suggests that dense cross-domain generalization can be computationally expensive, whereas the proposed similarity-aware compactness and domain-adaptive generalization strategy retains useful semantic transfer signals with much lower training cost.

UniSRec provides another useful reference point. Although UniSRec is designed as a universal sequence representation baseline, it requires 291.76min of training time in our setting, while its average recommendation score is 28.01, lower than that of BERT4Rec-SAGE. This comparison further shows that *-SAGE* achieves a more practical effectiveness–efficiency trade-off: it obtains the best average recommendation performance among the compared methods while avoiding the substantially higher training cost of UniSRec.

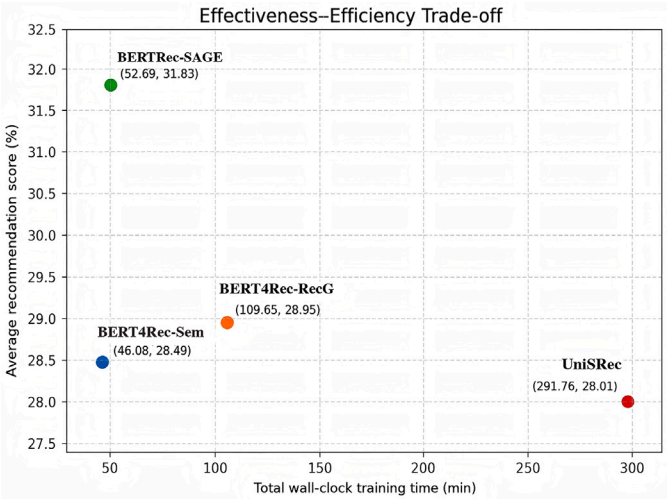


Fig. 4. Effectiveness–efficiency trade-off. For BERT4Rec-based variants, results are averaged over AMT→ACV and AMT→Steam. Each point shows the total wall-clock training time and the average recommendation score, with UniSRec included as an external reference.

5.6. Representation analysis (RQ5)

To further understand how SAGERec influences zero-shot transfer, we visualize both the learned item representations and the resulting cross-domain alignment patterns. Fig. 5 provides two complementary views of the learned representation space.

(1) *Item-level cross-domain weighting patterns.* The top part of Fig. 5 visualizes the row-normalized cross-domain weight matrices between 50 sampled source items and 50 sampled target items in the AMT→ACV setting. Under a shared color scale, the *-RecG* variant exhibits a comparatively diffuse and less differentiated response pattern, suggesting limited selectivity when aligning cross-domain item pairs. By contrast, SAGERec shows a more structured and contrastive distribution, with clearer bands and localized high-response regions over cross-domain correspondences. This observation is consistent with the intended role of the SIC module, namely, to prioritize reliable transfer signals while suppressing noisier alignments.

(2) *Domain-level alignment in the latent space.* In the bottom part of Fig. 5, we project the learned item representations of the BERT4Rec-based variants into a two-dimensional space using t-SNE. The *-Sem* variant shows relatively separated clusters across the source and non-source domains, suggesting that semantic enhancement alone is insufficient to establish effective cross-domain alignment. The *-RecG* variant produces substantially stronger overlap among the three domains, but the projected structure also appears more mixed and less domain-distinguishable. In contrast, SAGERec yields a more structured arrangement: some cross-domain overlap is retained to support transfer, while clearer domain-specific discrimination is still preserved. This pattern is consistent with the design goal of achieving selective rather than uniform alignment.

Overall, the visualization results are qualitatively consistent with the quantitative improvements reported in Table 2. They suggest that SAGERec better balances cross-domain alignment and intra-domain discrimination than semantic enhancement alone or uniform generalization.

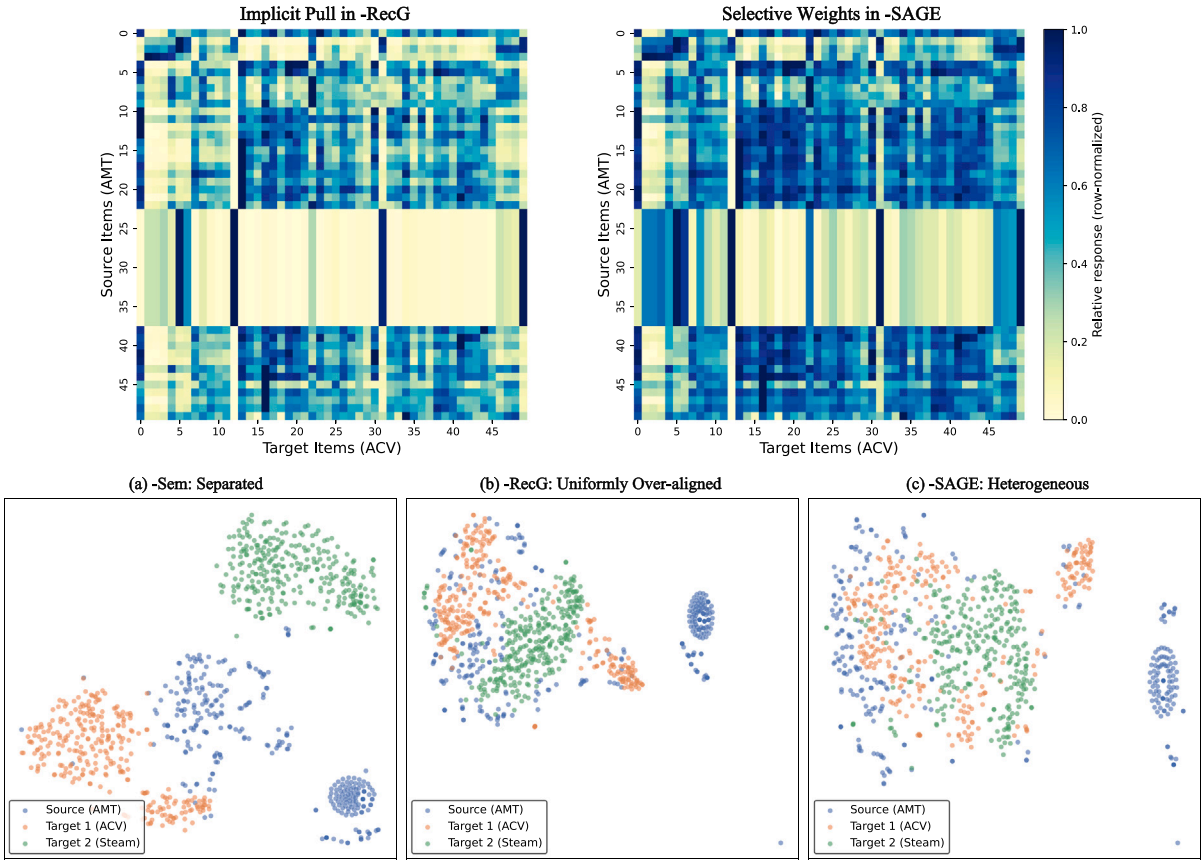


Fig. 5. Representation analysis of SAGERec. Top: row-normalized cross-domain alignment weight matrices between 50 sampled source items and 50 target items in the AMT→ACV setting, displayed under a shared color scale for structural comparison. Bottom: t-SNE visualization of projected item representations from the BERT4Rec-based variants across the source domain (AMT) and two non-source domains (ACV and Steam).

5.7. Semantic encoder backbone analysis (RQ6)

To examine whether SAGERec is tied to a specific item semantic encoder, we compare a BERT-based text encoder with the Llama-based encoder used in the main experiments. We fix the sequential recommendation backbone to BERT4Rec and keep the same zero-shot transfer protocol, source domain, target domains, and evaluation settings. The only changed component is the offline item semantic encoder used to produce cached item embeddings.

Fig. 6 provides an intuitive comparison between BERT-based and Llama-based item semantic encoders. The Llama-based encoder consistently yields stronger overall performance than the BERT-based encoder across all three variants, indicating that higher-quality semantic representations provide more informative cross-domain item affinities.

More importantly, the benefit of SAGERec depends on the quality of semantic embeddings. Under BERT-based embeddings, BERT4Rec-SAGE does not outperform the semantic-only or *-RecG* variants, whereas under Llama-based embeddings it achieves the best average score. This contrast indicates that SAGERec is not architecturally restricted to a particular item encoder, but its selective alignment mechanism requires sufficiently reliable cross-domain affinity estimates; otherwise, noisy or less discriminative embeddings may limit the benefit of SIC-based weighting.

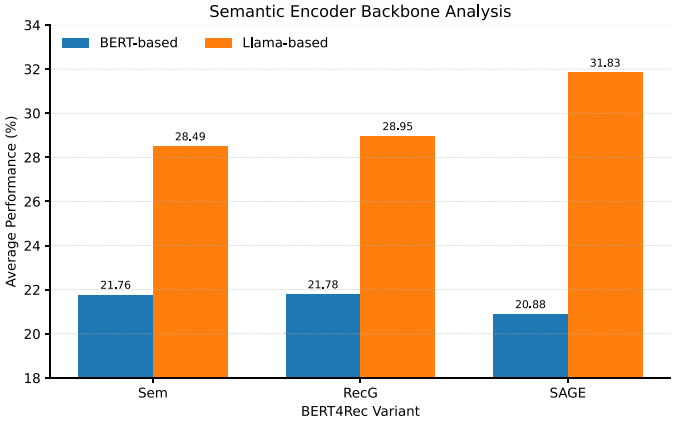


Fig. 6. Average performance of BERT4Rec variants under BERT-based and Llama-based item semantic encoders. Results are averaged over the two target domains. "Average Performance" denotes the arithmetic mean of R@10, N@10, R@20, and N@20.

5.8. Case study (RQ7)

To complement the quantitative results, we conduct a qualitative case study to examine when semantic alignment provides reliable transfer signals and when it becomes uncertain. Table 7 presents representative AMT–ACV item-pair cases under the main setting with $\theta = 0$. For traceability, we report raw Amazon ASINs together with semantic affinity and behavioral-neighborhood overlap. These cases cover different alignment patterns, including behavior-supported semantic matches, partial-title matches with weak behavioral support, and behavior-unconfirmed semantic matches.

Table 7 shows that SIC can identify useful cross-domain alignment when semantic affinity is supported by behavioral evidence. The two

Table 7
Representative AMT–ACV item-pair cases. all numeric values are rounded to three decimals. Item IDs are raw Amazon ASINs from the AMT source domain and the ACV target domain. "Beh. overlap" denotes behavioral-neighborhood overlap, and only α_{ij} is reported because the main setting uses $\theta = 0$.

Case	Source item (AMT)	Target item (ACV)	α_{ij}	Beh. overlap	Signal
Behavior-supported aligned	Celtic Thunder: It's Entertainment! ID: B003019LWY	It's Entertainment ID: B003019LWO	0.999	0.714	Semantic and behavioral agreement
Behavior-supported aligned	Celtic Thunder: Storm ID: B005DETYW8	Storm ID: B005DEVH1E	0.997	0.571	Semantic and behavioral agreement
Partial-title weak support	Celtic Thunder: Take Me Home [DVD] ID: B002AK9UUO	Take Me Home ID: B00942S4OY	0.998	0.167	Shared phrase, limited behavior
Exact-title semantic edge	Cargo ID: B003ZXNCHC	Cargo ID: B001U9BS4W	1.000	0.000	Title-identical, behavior-unconfirmed
Lexical-only alignment	Abandoned ID: B003MI2GHM	Abandon ID: B000007N8O	0.993	0.000	Lexical similarity without observed behavior

"Celtic Thunder" cases obtain near-unit semantic affinity scores and relatively high behavioral-overlap values of 0.714 and 0.571, respectively, indicating that these activated pairs are not merely matched by surface titles but also share similar behavioral neighborhoods. The partial-title case further shows a weaker alignment pattern: although "Celtic Thunder: Take Me Home [DVD]" and "Take Me Home" have high semantic affinity, their behavioral overlap is much lower. This suggests that shared title phrases may still provide transferable semantic clues, but their reliability can vary across domains; therefore, semantic alignment should be weighted rather than treated as hard item matching.

The exact-title and lexical-only cases reveal the boundary of item-side semantic alignment. Although "Cargo"–"Cargo" and "Abandoned"–"Abandon" receive high semantic affinity scores, their behavioral-neighborhood overlap is zero in the observed interaction data, suggesting that similar titles or word forms do not necessarily imply consistent user intent across domains. Such cases may introduce noisy attraction if item representations are forced to be overly close. These observations clarify why SAGERec treats semantic affinity as a soft prior: SIC provides selective weighted alignment, while DAG adjusts the regularization strength according to domain-level discrepancy. Nevertheless, DAG cannot fully resolve item-level behavior–semantic mismatch, which remains an important limitation and motivates future behavior-aware or user-intent-aware alignment strategies.

5.9. Discussion and limitations

The results from RQ1–RQ7 collectively show that our SAGERec improves zero-shot cross-domain sequential recommendation by balancing semantic transfer and domain-specific discrimination. The main comparison and ablation studies demonstrate that the gains are consistent across sequential backbones and mainly come from the complementary effects of SIC and DAG: SIC selects more reliable item-level semantic correspondences, while DAG adjusts cross-domain regularization according to domain discrepancy. This interpretation is supported by the representation analysis, where the alignment heatmaps become more structured and localized, and the t-SNE visualization shows that SAGERec encourages useful cross-domain overlap without fully mixing domain-specific item structures. The sensitivity and efficiency analyses further indicate that SAGERec behaves stably under practical settings and achieves a favorable effectiveness–efficiency trade-off. Meanwhile, the encoder-backbone and case studies show that selective semantic generalization is most effective when semantic similarity is supported by reliable item representations and is consistent with behavioral evidence.

These findings also reveal the boundaries of the current framework. Since SAGERec uses LLM-derived item embeddings to estimate cross-domain semantic affinity, its alignment quality is inevitably affected by the reliability of item-side textual metadata and the semantic encoder. Noisy descriptions, missing fields, platform-specific writing styles, or superficial lexical similarities may lead to inaccurate affinities, even though the similarity scores are used only as soft priors rather than ground-truth compatibility. In addition, SIC and DAG regularize representation learning through thresholded semantic weights and mini-batch-based structure estimates, which may become less reliable when cross-domain semantic overlap is weak or when local batch statistics are not sufficiently representative. Accordingly, SAGERec should be interpreted as an encoder-based selective semantic generalization framework for zero-shot transfer. It does not make direct claims about prompt-based or generative LLM recommenders, which usually rely on different supervision signals, prompting strategies, and inference protocols.

Future research can build on these observations in several directions. A promising direction is to move beyond purely text-based affinity estimation by incorporating behavior-aware or uncertainty-aware signals, so that cross-domain alignment can better distinguish true user-intent consistency from surface-level semantic similarity. Another direction is to improve robustness under incomplete, noisy, or heterogeneous metadata, especially in more distant domain-transfer scenarios. It would also be valuable to develop evaluation protocols that jointly examine accuracy, representation structure, diversity, and coverage, and that fairly connect encoder-based zero-shot transfer with prompt-based or generative LLM recommendation paradigms. These directions would deepen the understanding of when semantic generalization is reliable and how it can be extended to more realistic recommendation environments.

6. Conclusion

In this paper, we study zero-shot cross-domain sequential recommendation from the perspective of transfer heterogeneity. We identify a key limitation of existing semantic transfer methods, namely, their tendency to impose overly uniform cross-domain alignment, which may introduce noisy correspondences and weaken intra-domain structure. To address this issue, we propose SAGERec, a similarity-aware generalization framework that improves zero-shot transfer through two components: the Similarity-Aware Inter-domain Compactness (SIC) module for selective item-level alignment, and the Domain-Adaptive Generalization (DAG) mechanism for discrepancy-aware regularization.

Extensive experiments on real-world datasets and multiple sequential backbones show that SAGERec consistently improves recommendation performance within matched backbone families and achieves strong overall results across transfer settings. The ablation, sensitivity, efficiency, representation, encoder-backbone, and case analyses further indicate that SAGERec provides an effective balance between cross-domain alignment, domain-specific discrimination, and computational cost. Despite these encouraging results, SAGERec still relies on the quality of item-side semantic representations and metadata. Future work may explore behavior-aware or uncertainty-aware alignment strategies to better distinguish true user-intent consistency from surface-level semantic similarity in more heterogeneous recommendation scenarios.

CRediT authorship contribution statement

Zihao Pan: Writing – original draft, Validation, Methodology, Conceptualization. **Yuzhuo Dang:** Validation, Methodology, Formal

analysis. **Xin Zhang:** Formal analysis, Conceptualization. **Zhiqiang Pan:** Project administration, Investigation. **Taihua Shao:** Writing – review & editing, Investigation, Funding acquisition. **Fei Cai:** Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the support provided by the COSTA: Complex System Optimization Team of the College of Systems Engineering at NUDT. This research was supported by Natural Science Foundation of Henan under Grant No. 252300420989.

Data availability

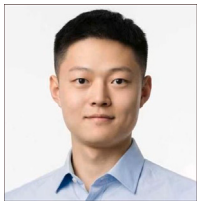
Data will be made available upon request.

References

[1] T.F. Boka, Z. Niu, R.B. Neupane, A survey of sequential recommendation systems: techniques, evaluation, and future directions, Inf. Syst. 125 (2024) 102427.
[2] L. Pan, W. Pan, M. Wei, H. Yin, Z. Ming, A survey on sequential recommendation, Front. Comput. Sci. 20 (3) (2026) 2003606.
[3] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks, in: ICLR (Poster), 2016.
[4] J. Tang, K. Wang, Personalized top-N sequential recommendation via convolutional sequence embedding, in: WSDM, ACM, 2018, pp. 565–573.
[5] W.-C. Kang, J.J. McAuley, Self-attentive sequential recommendation, in: ICDM, IEEE Computer Society, 2018, pp. 197–206.
[6] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer, in: CIKM, ACM, 2019, pp. 1441–1450.
[7] S. Chen, Z. Xu, W. Pan, Q. Yang, Z. Ming, A survey on cross-domain sequential recommendation, in: IJCAI, Ijcai.org, 2024, pp. 7989–7998.
[8] Y. Li, J. Wang, H. Sundaram, Z. Liu, LLM-RecG: a semantic bias-aware framework for zero-shot sequential recommendation, in: RecSys, ACM, 2025, pp. 237–246.
[9] Y. Dang, Z. Pan, X. Zhang, W. Chen, F. Cai, H. Chen, Discrepancy learning guided hierarchical fusion network for multi-modal recommendation, Knowl. Based Syst. 317 (2025) 113496.
[10] Y. Dang, W. Chen, Z. Pan, X. Zhang, Y. Duan, F. Cai, H. Chen, Dual-space feature representation learning network for multimodal recommender systems, Adv. Eng. Inform. 72 (2026) 104462.
[11] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, H. Xiong, E. Chen, A survey on large language models for recommendation, World Wide Web (WWW) 27 (5) (2024) 60.
[12] Z. Zhao, W. Fan, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang, Q. Li, Recommender systems in the era of large language models (LLMs), IEEE Trans. Knowl. Data Eng. 36 (11) (2024) 6889–6907.
[13] L. Wang, E.-P. Lim, Zero-shot next-item recommendation using large pretrained language models, CoRR arXiv:2304.03153, 2023.
[14] Y. Jiang, X. Ren, L. Xia, D. Luo, K. Lin, C. Huang, RecGPT: a foundation model for sequential recommendation, in: EMNLP, Association for Computational Linguistics, 2025, pp. 10129–10143.
[15] M. Ma, P. Ren, Y. Lin, Z. Chen, J. Ma, M. de Rijke, π-net: a parallel information-sharing network for shared-account cross-domain sequential recommendations, in: SIGIR, ACM, 2019, pp. 685–694.
[16] W. Sun, M. Ma, P. Ren, Y. Lin, Z. Chen, Z. Ren, J. Ma, M. de Rijke, Parallel split-join networks for shared account cross-domain sequential recommendations, IEEE Trans. Knowl. Data Eng. 35 (4) (2023) 4106–4123.
[17] J. Cao, X. Cong, J. Sheng, T. Liu, B. Wang, Contrastive cross-domain sequential recommendation, in: CIKM, ACM, 2022, pp. 138–147.
[18] W. Xu, Q. Wu, R. Wang, M. Ha, Q. Ma, L. Chen, B. Han, J. Yan, Rethinking cross-domain sequential recommendation under open-world assumptions, in: WWW, ACM, 2024, pp. 3173–3184.
[19] S. Moon, J. Koo, Y. Lee, ADR-rec: adaptive disentanglement for cross-domain sequential recommendation with cross attention gating mechanisms, Neurocomputing (2025) 132254.
[20] H. Ding, Y. Ma, A. Deoras, Y. Wang, H. Wang, Zero-shot recommender systems, CoRR arXiv:2105.08318, 2021.
[21] P.J. Feng, P. Pan, T. Zhou, H. Chen, C. Luo, Zero shot on the cold-start problem: model-agnostic interest learning for recommender systems, in: CIKM, ACM, 2021, pp. 474–483.
[22] J. Wang, P. Rathi, H. Sundaram, A pre-trained zero-shot sequential recommendation framework via popularity dynamics, in: RecSys, ACM, 2024, pp. 433–443.
[23] Y. Hou, S. Mu, W.X. Zhao, Y. Li, B. Ding, J.-R. Wen, Towards universal sequence representation learning for recommender systems, in: KDD, ACM, 2022, pp. 585–593.
[24] S. Geng, S. Liu, Z. Fu, Y. Ge, Y. Zhang, Recommendation as language processing (RLP): a unified pretrain, personalized prompt & predict paradigm (P5), in: RecSys, ACM, 2022, pp. 299–315.
[25] X. Zhou, J. Jin, L. Ma, X. Gao, J. Yang, X. Yang, L. Xiao, SCRIPT: sequential cross-meta-information recommendation in pretrain and prompt paradigm, in: ICDM, IEEE, 2023, pp. 888–897.
[26] H. Zhou, X. Chen, W. Cha, R. Gu, P. Wang, Triple contrastive learning framework for domain-level zero-shot recommendation, Neurocomputing 654 (2025) 131354.
[27] K. Bao, J. Zhang, Y. Zhang, W. Wang, F. Feng, X. He, TALLRec: an effective and efficient tuning framework to align large language model with recommendation, in: RecSys, ACM, 2023, pp. 1007–1014.
[28] J. Ji, Z. Li, S. Xu, W. Hua, Y. Ge, J. Tan, Y. Zhang, GenRec: large language model for generative recommendation, in: ECIR (3), Lecture Notes in Computer Science, Springer, 2024, pp. 494–502.
[29] Z. Yue, S. Rabhi, G. de Souza Pereira Moreira, D. Wang, E. Oldridge, LlamaRec: two-stage recommendation using large language models for ranking, CoRR arXiv:2311.02089, 2023.
[30] Z. He, Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B.P. Majumder, N. Kallus, J. McAuley, Large language models as zero-shot conversational recommenders, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 720–730, https://doi.org/10.1145/3583780.3614949
[31] J. Harte, W. Zorgdrager, P. Louridas, A. Katsifodimos, D. Jannach, M. Fragkoulis, Leveraging large language models for sequential recommendation, in: RecSys, ACM, 2023, pp. 1096–1102.
[32] Q. Liu, X. Wu, Y. Wang, Z. Zhang, F. Tian, Y. Zheng, X. Zhao, LLM-ESR: large language models enhancement for long-tailed sequential recommendation, in: NeurIPS, 2024.
[33] S. Safar, B.R. Jose, J. Mathew, T. Santhanakrishnan, Recommendation systems with LLM-based semantic embeddings and FAISS similarity search, Neurocomputing 649 (2025) 130753.
[34] Z. Qu, R. Xie, C. Xiao, Z. Kang, X. Sun, The elephant in the room: rethinking the usage of pre-trained language model in sequential recommendation, in: RecSys, ACM, 2024, pp. 53–62.
[35] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, J.J. McAuley, Bridging language and items for retrieval and recommendation, CoRR arXiv:2403.03952, 2024.
[36] Y. Li, Y. Ding, B. Chen, X. Xin, Y. Wang, Y. Shi, R. Tang, D. Wang, Extracting attentive social temporal excitation for sequential recommendation, in: CIKM, ACM, 2021, pp. 998–1007.
[37] R. He, J.J. McAuley, Fusing similarity models with Markov chains for sparse sequential recommendation, in: ICDM, IEEE Computer Society, 2016, pp. 191–200.
[38] S. Rendle, C. Freudenthaler, L. Schmidt-Thieme, Factorizing personalized Markov chains for next-basket recommendation, in: WWW, ACM, 2010, pp. 811–820.
[39] P. BehnamGhader, V. Adlakha, M. Mosbach, D. Bahdanau, N. Chapados, S. Reddy, LLM2VEC: large language models are secretly powerful text encoders, CoRR arXiv:2404.05961, 2024.
[40] L. Team, The LLaMA 3 herd of models, CoRR arXiv:2407.21783, 2024.

Author biography

Zihao Pan: is a MSc student at the National University of Defense Technology. His research interests include zero-shot learning and LLM-based sequential recommender systems. He earned his Bachelor degree at the National University of Defense Technology majoring in Information Systems.



Yuzhuo Dang: is a PhD student at the National University of Defense Technology. His research interests include recommender systems and multimodal learning. He earned his Bachelor and Master degrees at the National University of Defense Technology majoring in Information Systems. He has published several related papers in venues such as SIGIR, INFFUS, AEI, KBS, etc. In addition, he serves as a reviewer for SIGIR, AAAI KBS, etc.

Xin Zhang: is a PhD student at the National University of Defense Technology. His research interests include few-shot learning and graph neural networks. He received his master degree at the National University of Defense Technology majoring in Information Systems in 2022. He has published papers in IP&M, KBS, CIKM and WWW. In addition, he serves as a reviewer for SIGIR, WWW and TNNLS.

Zhiqiang Pan: is a lecturer at the National University of Defense Technology. His research interests include graph analytics and recommender systems. He earned his Bachelor, Master and PhD degrees at the National University of Defense Technology majoring in Information Systems. He has published several related papers in conference proceedings and journals such as SIGIR, CIKM, TOIS, TNNLS, NCAA, IRJ, etc.

Taihua Shao: is a lecturer at the Information Engineering University, Zhengzhou, China. He received his Ph.D. degree in management science and engineering from the National University of Defense Technology, Changsha, China, in 2023. His research interests include person-post matching, data mining and machine learning.

Fei Cai: is an associate professor at the National University of Defense Technology. His research interests include information retrieval and query formulation. He earned his Doctorate degree in Computer Science from the University of Amsterdam under the supervision of Prof. Maarten de Rijke. He has several papers published in WWW, SIGIR, CIKM, FnTIR, TOIS, TKDE, IPM, etc. In addition, he serves as a PC member for KDD, WWW, SIGIR, CIKM, WSDM and CCIR as well as a reviewer for SIGIR, WWW, WSDM, CIKM, TOIS, VLDBJ, TKDE, IPM, JASIST, etc. He joined the Editorial Board of IPM in 2015.