SEMANTIC-GUIDED STRUCTURAL INFORMATION BORROWING FOR ITEM COLD-START RECOMMENDATION ON GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

In graph-based recommender systems, items are ranked by embeddings learned from the structural information of the user-item interaction graph. A cold (new) item has no interaction history and is isolated in this graph, so it lacks such structural information and cannot be ranked effectively like warm items. Existing methods address this issue in two ways. The first is injecting content features, which provide a cold item with a semantic embedding, but not the structural embedding required for ranking. The second is complementing the structural information through simulated or reconstructed edges. Since these edges are generated instead of real, the complemented structural information may be noisy. To overcome these limitations, we propose an end-to-end Semantic-Guided Structural information Borrowing framework (SSB), which allows a cold item to borrow the structural information from its similar warm items to represent itself, thus the cold item obtains a structural embedding for accurate recommendation. Specifically, we use a fine-tuning-free LLM to help a cold item to accurately locate the similar warm items, ensuring that the borrowed structural information is indeed relevant. Then, we propose a soft-structure borrowing mechanism composed of a soft attention mechanism and hierarchical structure aggregation. This mechanism enables a cold item to flexibly determine how much information to borrow from each located similar warm item, thereby making a cold item obtain a continuous combination of warm item structural embeddings rather than embeddings learned from discrete edges. Moreover, we optimize a pseudo-cold mask method to address the lack of supervision signals caused by missing structural information of cold items, letting gradients flow back to the learnable parameters during training. Finally, we conduct extensive experiments on three Amazon datasets: Sports, Toys, and Beauty. The results show that SSB outperforms SOTA cold-start baselines on Sports and Toys, and is comparable to CGRC on Beauty.

1 INTRODUCTION

Recommender systems are trained on historical user-item interactions, which give every warm item an embedding that reflects how users interact with it. However, new items continuously enter the product catalog. Since these new items lack interaction history upon launch, the system cannot learn corresponding embeddings for them, rendering it unable to recommend these cold items. Cold items constitute a significant proportion of the catalog, and recommending them poorly deprives new products of exposure.

In graph-based recommender systems, item embeddings are constructed using structural information derived from neighboring nodes within the user-item interaction graph. A warm item is connected by edges to the users who interacted with it, so it occupies a position in that graph. Collaborative signals propagate along these edges, allowing the embedding of the item to encode both its own identity and the preference characteristics of its neighbors. In contrast, a cold item has no such edges at launch, so it is an isolated node cut off from all propagation paths and therefore has no position in the collaborative embedding space. Consequently, to be ranked alongside warm items, cold items must find a way to establish their own position within that space.

To address this issue, prior work on item cold-start has largely followed two strategies. The first is injecting content features, which provide a cold item with a semantic embedding, while it fails to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

provide the structural embeddings required for ranking. The second is complementing the structural information through simulated or reconstructed edges. However, since these edges are generated instead of real, the resulting structural information may contain noise, thereby hindering reliable ranking for cold-start items.

In interaction graphs, similar items are often associated with the same users (McPherson et al., 2001), so warm items similar to a cold item already carry the structural information required by that cold item. Existing simulation and reconstruction methods use this characteristic but generate the structural information. In fact, this structural information can instead be derived from the warm items, without the need for construction from scratch. This requires accurate semantic localization. The textual descriptions of items must be fine-grained and highly discriminative.

We instantiate this idea in SSB (Semantic-Guided Structural information Borrowing), which proceeds in two steps—locate and borrow. A fine-tuning-free LLM locates the warm items that are semantically close to a cold item, and a soft attention mechanism determines how much information to borrow from each item.

Our contributions are as follows.

- We formulate the item cold-start as a locate-and-borrow problem. Based on this, we propose an end-to-end Semantic-Guided Structural information Borrowing framework (SSB), which utilizes a fine-tuning-free LLM to encode the textual description of a cold item into an embedding that is used to measure and locate similar warm items. And missing structural information of cold items is borrowed from these warm items, better than existing generated methods.
- We propose a soft-structure borrowing mechanism to enable a cold item to flexibly determine how much information to borrow from each located warm semantic neighbor. This mechanism consists of two components. The component of the attention mechanism is adopted to learn the weights of located warm semantic neighbors, and the hierarchical structure aggregation is applied to learn the weights of LightGCN propagation layers. In this way, SSB could represent a cold item as a continuous and learnable combination of the structural embeddings of warm semantic neighbors for ranking.
- We design a pseudo-cold masking approach to ensure the existence of the gradient during the training phase, avoiding the issue of gradient unavailability caused by missing structural information of cold items.
- We conduct extensive experiments on three Amazon datasets, where SSB outperforms SOTA cold-start baselines on Sports and Toys and is comparable to CGRC on Beauty, with the largest gains on the rarest items of the sparsest datasets.

2 Related Work

2.1 Item Cold-Start Recommendation

Work on item cold-start falls into four families. The first simulates or aligns a behavioral embedding for a cold item so that it can be ranked without real interactions. DropoutNet (Volkovs et al., 2017) applies dropout to the collaborative embeddings of warm items during training, forcing the model to rely on content features when collaborative signals are unavailable. Heater (Zhu et al., 2020) learns a mapping from item attributes into the collaborative embedding space. MTPR (Du et al., 2020) drops the collaborative features of warm items to simulate cold items. CLCRec aligns content and collaborative views through contrastive learning, and ALDI aligns content embeddings with behavioral embeddings by distillation (Wei et al., 2021; Huang et al., 2023). Generative, contrastive, and semantic-behavior alignment variants further extend this idea (Chen et al., 2022; Zhou et al., 2023; Wang et al., 2024c; Xu et al., 2024; Yao et al., 2025). These methods give the cold item a usable vector, but one still derived from content, which cannot substitute for the structural information the item lacks.

The second family enriches the cold item with content or graph signals at the representation level, for example by propagating features from neighboring items, completing feature-structure adaptive graphs, or reconstructing the interaction graph (Kim et al., 2024a; Lei et al., 2025; Wang et al., 2024b;

Qian et al., 2022). These approaches move the cold item closer to the collaborative embedding space. However, the edges or features they rely on are usually built by heuristics or by a separate reconstruction stage, outside the ranking objective.

The third family simulates the missing interactions themselves. ColdLLM couples a funnel filter with an LLM to generate interactions for a cold item (Huang et al., 2025), and SetLLM prompts an LLM to produce a set of interacted users that serves as supervision (Liu et al., 2026). Related methods use LLMs to augment the interaction graph or reason over knowledge graphs (Wei et al., 2024; Wang et al., 2024a; Sakurai et al., 2025). These methods place a cold item directly in the collaborative embedding space, but they rely on synthetic, discrete interactions that can carry hallucinated preferences. The LLM-based simulators among them, such as ColdLLM and SetLLM, also require LLM fine-tuning or online inference at deployment.

A fourth, orthogonal family applies meta-learning to adapt quickly to new items, but it targets few-shot adaptation, not the complete absence of structural information (Dong et al., 2020; Lu et al., 2020). SSB differs from all four families. It neither simulates nor aligns embeddings nor reconstructs edges or generates interactions. Instead, the LLM stays offline as an encoder, and the cold item borrows the learned structural information of similar warm items through continuous, differentiable soft weights.

2.2 LLMs for Recommendation

Large language models have entered recommendation in several roles. As a recommender, the LLM is prompted or fine-tuned to rank users and items directly (Liao et al., 2024; Liang et al., 2025; Kieu et al., 2025), which is expressive but the most expensive and least reproducible option. As a feature encoder, it turns item text into a semantic vector for downstream ranking (Kim et al., 2024b; Hu et al., 2024; Ren et al., 2024). This is the cheapest way to benefit from an LLM, though many such methods still fine-tune the encoder or couple it tightly to the ranking model, blurring the boundary between semantics and collaboration. As a simulator or graph augmenter, it produces interactions or augments the interaction graph for cold items, as in the third family above. This is more powerful, yet it demands online inference or fine-tuning and risks hallucinated signals. SSB deliberately occupies the second role in its lightest form.

3 Method

3.1 Problem Formulation

Let $\mathcal{U}$ and $\mathcal{I}$ be the sets of users and items, and let $H(u) \subseteq \mathcal{I}$ denote the history of items that a user u has interacted with. The training interactions define the user-item interaction graph, a bipartite graph over users and items, and every item carries a textual description that is available at launch. We split items into a warm set $\mathcal{I}_w$, whose items appear in the training interactions, and a cold set $\mathcal{I}_c$, whose items have all their interactions withheld for validation and testing only. A cold item is therefore completely absent from the interaction graph. Unlike warm items, a cold item has no embedding derived from the graph and must be represented from its textual description alone.

The task of item cold-start recommendation is to rank the cold items for each user with the same scoring function used for warm items, so that items the user would interact with rank above the rest.

3.2 Overview

SSB assembles the representation of a cold item in two steps, which we call locate and borrow. Locate uses a fine-tuning-free LLM to encode item text and retrieve the top-K warm items that are semantically closest to the cold item. In parallel, a LightGCN encodes the structural information of warm items over the interaction graph. Borrow then lets the cold item read this structural information through a soft-structure borrowing mechanism, which combines a soft attention mechanism with hierarchical structure aggregation. The semantic prior of the cold item and the borrowed structural information are fused into its final representation, and the whole model is trained end to end. Figure 1 gives an overview.

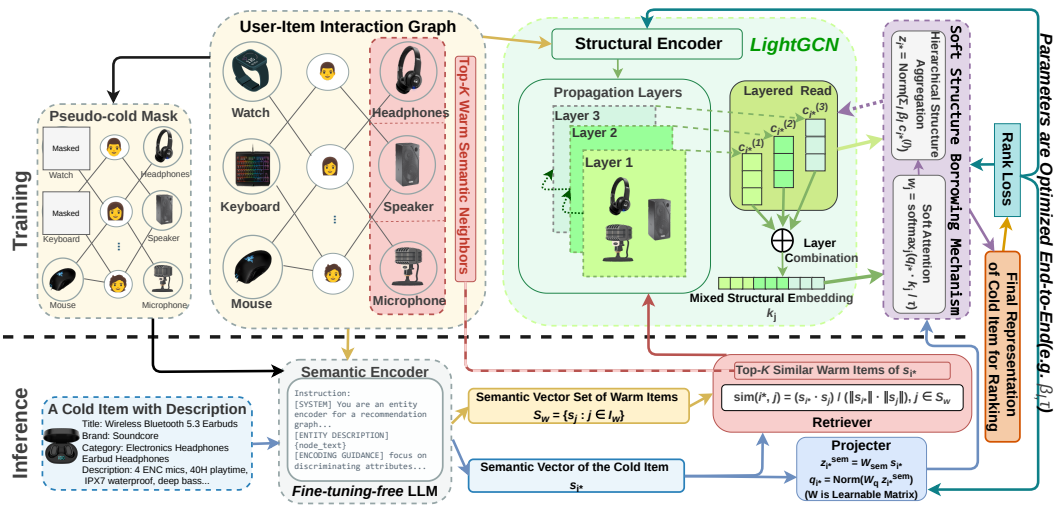

Figure 1: Overview of SSB. A LightGCN encodes the interaction graph, Locate finds similar warm items by semantic similarity, and Borrow obtains their structural information through a soft-structure borrowing mechanism.

3.3 Semantic Encoding and Locate

For each item i, we assemble its description from the available text fields (title, brand, category, and description). Before training, we pass the assembled text through a fine-tuning-free LLM with a fixed, dataset-specific prompt (Appendix A), and mask-aware mean pooling over the final hidden states yields $s_i \in \mathbb{R}^{d_{\text{LLM}}}$. The vector is L_2-normalized, stored offline, and frozen, so the LLM is never invoked during training or evaluation.

SSB then uses the frozen semantic vector of the cold item to locate the warm items whose structural information it can borrow. For a cold item i^*, its top-K warm semantic neighbors $\mathcal{N}^K_{\text{sem}}(i^*)$ are retrieved offline by cosine similarity, keeping the top-K warm items among $3K$ candidates so that the cold item can always borrow from K warm neighbors. The located semantic vector is projected into the common d-dimensional space as the semantic prior $z^{\text{sem}}_{i^*} = W_{\text{sem}} s_{i^*}$, where $W_{\text{sem}} \in \mathbb{R}^{d \times d_{\text{LLM}}}$ is a learned projection, and serves as the query of soft attention (Section 3.5). We freeze the encoder so that the semantic vectors depend only on item text, since fine-tuning would drift them toward items connected in the interaction graph.

3.4 Structural Encoding

We encode structural information with a LightGCN (He et al., 2020) over the interaction graph. Let $E^{(0)}_i, E^{(0)}_u \in \mathbb{R}^d$ be learnable initial embeddings for item i and user u in a common d-dimensional space. After L layers of degree-normalized propagation, the structural representations are the sums over layers,

$$E_i = \sum_{l=0}^{L} E^{(l)}_i, \qquad E_u = \sum_{l=0}^{L} E^{(l)}_u, \tag{1}$$

with both E_i and E_u L_2-normalized. Here $E^{(l)}_i$ is the representation of item i after l layers of propagation, which hierarchical structure aggregation later reads layer by layer (Section 3.5). The user side is a standard LightGCN representation, where a user u is encoded by its structural embedding,

$$z_u = E_u, \tag{2}$$

which lives in the same d-dimensional space as the item embeddings, so the dot product $z_u \cdot z_i$ is a well-defined ranking score. Since the datasets provide no user-side text, the user representation is purely structural.

3.5 Soft-Structure Borrowing Mechanism

Borrow assembles the structural information of the cold item from its located warm semantic neighbors, each providing its structural embedding E_j. To decide how much to borrow from each neighbor, SSB uses a unidirectional and cross-item attention, where the semantic prior acts as the query and the neighbors' structural embeddings act as the keys, both mapped through learned projections,

$$q_{i^*} = \mathrm{Norm}(W_q z_{i^*}^{\mathrm{sem}}), \qquad k_j = \mathrm{Norm}(W_k E_j), \tag{3}$$

with attention weight $\alpha_j = q_{i^*} \cdot k_j / \tau_{\mathrm{attn}}$, where $W_q, W_k \in \mathbb{R}^{d \times d}$ and $\tau_{\mathrm{attn}} > 0$ is a learnable temperature initialized to 0.1, and the soft attention weights are

$$w_j = \mathrm{softmax}_j(\alpha_j), \qquad a_{i^*} = \sum_{j \in \mathcal{N}_{\mathrm{sem}}^K(i^*)} w_j E_j. \tag{4}$$

As the soft read is differentiable, the ranking loss shapes how much each neighbor contributes, without fixing these weights by an external similarity threshold (e.g., borrowing only when the similarity exceeds 0.8). Given that a smaller temperature yields a sharper softmax, the small initial value already concentrates attention on the few most similar neighbors, and τ_{attn} is then adjusted by the ranking objective during training. Finally, the semantic prior and the borrowed structural information are fused,

$$z_{i^*} = \mathrm{Norm}\Big(\mathrm{MLP}\big[z_{i^*}^{\mathrm{sem}} \,\|\, g \cdot a_{i^*}\big]\Big), \tag{5}$$

where g is a learnable scalar gate.

The aggregation above borrows from the mixed structural embeddings E_j, which collapse multi-hop collaborative signals into a single vector. SSB therefore extends the read across the individual layers of the interaction graph, where $E_j^{(l)}$ is the item-side representation of warm item j after l layers of propagation (Section 3.4). The cold item keeps its semantic prior as layer zero and reads each subsequent layer with the same soft weights, computed once from E_j and reused across layers so that how much to borrow is decoupled from which order of signal to read. All layers are then combined with learned layer weights,

$$z_{i^*} = \mathrm{Norm}\Big(\sum_{l=0}^{H} \beta_l \, c_{i^*}^{(l)}\Big), \qquad c_{i^*}^{(0)} = z_{i^*}^{\mathrm{sem}}, \quad c_{i^*}^{(l)} = \sum_{j \in \mathcal{N}_{\mathrm{sem}}^K(i^*)} w_j E_j^{(l)}, \tag{6}$$

with $\beta = \mathrm{softmax}(\theta)$. This one-way read keeps the warm graph untouched and the cold item's representation identical at training and inference. SSB uses this hierarchical structure aggregation by default, combining the semantic prior and the soft readout structure layers by a learned linear convex combination and L_2-normalization. The single-layer fusion above is the variant removed in the ablation.

3.6 End-to-End Training

SSB scores a pair by the dot product of the normalized representations, $\mathrm{score}(u, i) = z_u \cdot z_i$, and trains with a listwise InfoNCE loss. For each observed interaction (u, i), the true user is the positive and a set $\mathcal{N}^-$ of sampled non-interacted users are negatives,

$$\mathcal{L} = -\log \frac{\exp(z_u \cdot z_i / \tau)}{\exp(z_u \cdot z_i / \tau) + \sum_{u' \in \mathcal{N}^-} \exp(z_{u'} \cdot z_i / \tau)}, \tag{7}$$

where τ is a temperature. Item-conditioned negative sampling runs the forward pass once per positive item, rather than per negative, keeping training efficient. Since the locate and borrow steps are all learned inside this objective, borrowed structural information is optimized directly for ranking.

SSB additionally trains with an optimized pseudo-cold mask method. Cold items alone cannot train the forward pass, since their interactions are withheld for testing and provide no gradient, and warm items train through LightGCN. To train this forward pass, SSB masks a fraction ρ of warm items as cold in each epoch, while their edges stay in the LightGCN graph and their own representation is excluded from the semantic-neighbor retrieval, so these masked items reconstruct their representation through structural borrowing. Unlike pseudo-cold or dropout strategies in prior work that remove

edges or interactions to simulate cold items, this masking at the representation level changes only the representation source, so the warm graph and warm-item representations remain intact and no training signal leaks into testing. The masked warm items still carry ranking supervision, so their gradient flows back through the semantic projection and attention, restoring the gradients that the cold items cannot provide. Masked warm items and real cold items go through the same borrowing computation, so masking keeps training and evaluation on the same cold-item reconstruction task.

4 Experiments

We evaluate SSB to answer four research questions. (RQ1) Does SSB outperform strong baselines on cold-item recommendation, and does the cold-start gain come at a cost to warm items? (RQ2) Does SSB improve on the rarest cold items, those with at most two withheld interactions? (RQ3) How much does each component contribute, including hierarchical structure aggregation, the semantic prior, the pseudo-cold mask method, and the LLM as the semantic source? (RQ4) How do the key hyperparameters K and ρ affect performance?

4.1 Experimental Setup

4.1.1 Datasets

We evaluate on three Amazon datasets, Beauty, Sports and Outdoors, and Toys and Games. Table 1 reports their statistics. They are standard cold-start benchmarks widely used in the community. For baselines whose original papers use other datasets, we reproduce them under the same split and evaluation protocol for a fair comparison. Following common practice, ratings no lower than 4 are treated as positive, a 5-core filter is applied to users and warm items (cold items are exempt), and the interaction sequence of each user is capped at the 50 most recent items. For each dataset, 20% of items are randomly designated as cold items, with all their interactions moved out of the warm graph. The withheld interactions of a cold item are split 1:1 by time into a cold validation set (the earlier half) and a cold test set (the later half). For the remaining warm items we adopt leave-one-out. For each user, the most recent interaction is the test sample, the second most recent is the validation sample, and the remaining interactions form the training set.

Table 1: Statistics of the three datasets after the rating threshold ($\geq$4) filter.

Dataset	#Users	#Items	#Cold	#Interactions	Sparsity
Beauty	16,360	12,085	2,417	124.9K	99.94%
Sports	249,403	103,855	20,772	1.94M	99.99%
Toys	161,939	78,040	15,608	1.30M	99.99%

4.1.2 Baselines

We compare SSB with eight baselines in three groups. (1) Embedding simulation or alignment: DropoutNet (Volkovs et al., 2017) and Heater (Zhu et al., 2020) simulate behavioral embeddings from content, CLCRec (Wei et al., 2021) aligns content and collaborative views through contrastive learning, and ALDI (Huang et al., 2023) aligns content embeddings with behavioral embeddings by distillation. (2) Content-graph reconstruction: CGRC (Kim et al., 2024a) reconstructs user-item edges with a masked graph autoencoder. (3) LLM-based interaction simulation: ColdLLM (Huang et al., 2025) and SetLLM (Liu et al., 2026) simulate interactions for cold items with LLMs. We also include LightGCN (He et al., 2020) as the collaborative backbone. All non-LLM baselines are tuned by grid search on the validation set. For the two LLM-based baselines (ColdLLM and SetLLM), only their downstream LightGCN backbones are grid-searched, as full LLM fine-tuning makes grid search computationally prohibitive.

4.1.3 Evaluation Metrics and Protocol

We report Recall@K and NDCG@K with $K \in \{10, 20\}$ on cold items, treating Recall@20 and NDCG@20 as the primary metrics. Evaluation uses full ranking over the candidate cold items, which are all withheld cold items (the union of the cold validation and cold test sets). Users without

any training interaction are excluded (the warm-user protocol), so the reported numbers reflect warm users ranking cold items. This candidate pool of only cold items follows the standard cold-start evaluation protocol and keeps our results comparable with prior work. We additionally report cold-item performance on the rarest interaction frequency bucket ($\leq$2 withheld interactions) in Section 4.3.

4.1.4 Implementation Details

SSB uses Llama-3.1-8B as the fine-tuning-free base LLM for item encoding. The common space dimension is $d = 256$ (semantic input $d_{\text{LLM}} = 4096$), and the LightGCN has $L = 2$ layers. We optimize with Adam, a batch size of 512, and training for up to 200 epochs with early stopping (patience 20) on cold validation NDCG@20. The learning rate is 10^{-4} on Beauty and 5×10^{-5} on Sports and Toys. Weight decay is 10^{-4}, with a separate value of 10^{-3} on the LightGCN embeddings. Hyperparameters are tuned by grid search on the cold validation set for each dataset, with the search ranges and per-dataset selections listed in Appendix B.

4.2 Main Results (RQ1)

Table 2: Cold-start and warm performance (Recall@K and NDCG@K, in %) over all cold items, the rarest cold bucket, and warm items. SSB and CGRC are reported as mean $\pm$ std over five seeds (42, 123, 2024, 7, 999). Best results are in bold. Second-best results are underlined.

Dataset	Method	Cold (all)				Cold ($\leq$2)				Warm			
		R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
Beauty	LightGCN	0.19	0.34	0.09	0.13	0.00	0.52	0.00	0.13	4.88	7.55	2.53	3.21
	CLCRec	5.70	9.03	3.14	4.01	4.95	8.07	2.63	3.42	4.53	7.41	2.25	2.97
	ALDI	1.86	4.13	0.85	1.44	2.08	5.99	0.69	1.70	3.89	5.99	2.05	2.58
	DropoutNet	1.65	3.05	0.89	1.25	2.60	5.73	1.28	2.05	3.06	5.17	1.43	1.96
	Heater	5.28	7.86	2.91	3.60	8.85	11.98	4.84	5.61	3.89	5.99	2.05	2.58
	CGRC	10.93±0.57	**15.67±0.31**	6.30±0.33	7.55±0.27	**10.68±0.37**	**15.31±1.06**	**6.09±0.46**	**7.27±0.44**	**5.83±0.14**	**9.25±0.07**	**3.01±0.06**	**3.86±0.05**
	ColdLLM	2.17	4.26	1.19	1.73	3.65	7.29	1.85	2.77	4.90	7.33	2.53	3.13
	SetLLM	9.10	12.88	5.23	6.23	7.81	11.46	4.97	5.90	4.93	7.56	2.56	3.22
	SSB	**11.28±0.36**	15.37±0.50	**6.88±0.23**	**7.98±0.26**	10.16±1.19	13.44±0.93	5.80±0.53	6.63±0.47	5.71±0.08	8.87±0.06	2.96±0.04	3.75±0.04
Sports	LightGCN	0.02	0.04	0.01	0.01	0.00	0.00	0.00	0.00	3.29	3.76	**2.77**	**2.89**
	CLCRec	0.44	0.96	0.21	0.34	0.21	0.63	0.12	0.22	2.61	3.35	1.98	2.17
	ALDI	0.26	0.51	0.14	0.20	0.32	0.74	0.13	0.24	3.13	3.46	**2.77**	2.85
	DropoutNet	0.26	0.49	0.13	0.18	0.11	0.21	0.04	0.06	2.79	3.14	2.32	2.41
	Heater	0.69	1.34	0.34	0.51	0.95	1.26	0.54	0.62	3.13	3.47	2.76	2.85
	CGRC	4.76±0.16	6.81±0.23	2.83±0.06	3.37±0.08	3.75±0.40	5.66±0.72	1.94±0.18	2.43±0.20	**3.74±0.03**	**4.88±0.04**	2.42±0.06	2.71±0.06
	ColdLLM	0.15	0.26	0.06	0.09	0.00	0.11	0.00	0.03	2.83	3.28	2.42	2.54
	SetLLM	2.42	3.39	1.53	1.78	1.79	2.63	1.07	1.29	3.04	3.58	2.54	2.67
	SSB	**5.75±0.13**	**8.28±0.11**	**3.33±0.08**	**3.99±0.07**	**4.92±0.74**	**7.21±0.52**	**2.70±0.42**	**3.28±0.32**	2.67±0.36	3.80±0.39	1.47±0.13	1.75±0.13
Toys	LightGCN	0.04	0.10	0.02	0.03	0.00	0.00	0.00	0.00	2.95	4.02	1.75	2.02
	CLCRec	2.40	4.10	1.26	1.71	0.96	1.38	0.39	0.49	2.08	3.19	1.09	1.37
	ALDI	0.65	1.20	0.34	0.48	0.28	0.28	0.11	0.11	2.02	2.67	1.26	1.42
	DropoutNet	1.44	2.46	0.75	1.03	1.24	1.38	0.49	0.52	1.66	2.31	0.96	1.12
	Heater	2.62	4.30	1.42	1.88	1.38	2.20	0.49	0.70	2.02	2.67	1.26	1.42
	CGRC	11.01±0.12	14.78±0.10	7.05±0.05	8.09±0.03	5.81±0.70	8.43±0.71	3.21±0.26	3.88±0.25	3.96±0.04	5.90±0.01	2.14±0.03	2.63±0.02
	ColdLLM	1.02	1.74	0.54	0.73	0.83	1.10	0.30	0.36	2.37	3.26	1.37	1.59
	SetLLM	7.40	9.66	4.76	5.39	4.41	5.37	2.68	2.92	2.36	3.33	1.35	1.59
	SSB	**11.67±0.09**	**15.54±0.16**	**7.37±0.06**	**8.42±0.06**	**7.49±0.52**	**10.74±0.19**	**3.99±0.33**	**4.82±0.24**	**4.57±0.05**	**6.51±0.08**	**2.50±0.02**	**2.99±0.03**

We report the cold-start performance of all methods in Table 2. On Beauty, SSB is comparable to the strongest baseline CGRC (its NDCG@10 is significantly higher, p<0.05, while its Recall@20 is slightly lower and not significant, over five seeds). On Sports and Toys, SSB achieves the best Recall@K and NDCG@K at every cutoff on the overall cold segment, improving NDCG@20 over CGRC from 3.37% to 3.99% on Sports and from 8.09% to 8.42% on Toys. This superiority comes from borrowing real structural information from warm items instead of generating embeddings or interactions. The gain is largest on Sports, the largest and sparsest catalog, where SSB improves Recall@20 over the strongest tuned baseline CGRC from 6.81% to 8.28%. The content-based baselines (DropoutNet, Heater, CLCRec, and ALDI) also fall short of CGRC and SSB. In sparse Sports a cold item has almost no structural information of its own, so the borrowed warm structural information is nearly all the signal it has, while the reconstructed edges of CGRC are noisy when the signal is thin. In dense Beauty the warm structural information is already abundant, so the advantage shrinks.

Table 2 also reports warm-item performance, so the cold-start gains can be weighed against their effect on warm items. On Toys SSB is the best warm method as well (Recall@20 6.51 vs. 5.90 for CGRC), and on Beauty its warm metrics are second only to CGRC (NDCG@20 3.75 vs. 3.86). Only on Sports does the cold gain come with a clear warm cost, concentrated in ranking quality.

Warm Recall@20 is 3.80, second only to CGRC's 4.88, but NDCG@20 falls to 1.75, the lowest among all methods. Warm scoring shares the LightGCN embeddings with the semantic locating and structural borrowing computation, where E_i acts as both the warm collaborative embedding and the attention key/value. The pseudo-cold mask method therefore shifts these shared parameters through the borrowing gradient. The severity of this cost follows the interaction of the masking ratio with the strength of the warm signal in the catalog. On Sports, the highest ratio ($\rho = 0.9$) meets the largest and sparsest catalog. Nearly all warm items train through structural borrowing, so the warm signal, already thin, leaves the shared parameters poorly anchored, which shows up as both a larger drop and a much larger variance across seeds ($\pm$0.39). On Beauty, $\rho = 0.7$ is also high, yet the denser warm signal keeps the drop small (-0.38pp). On Toys, the low $\rho = 0.3$ leaves warm accuracy essentially intact. Across the three datasets, SSB remains competitive on warm items while serving its target cold segment.

4.3 Cold-Start Degree Analysis (RQ2)

For RQ2, we examine the rarest cold items and report their Recall and NDCG at K=10 and K=20 in Table 2. On Sports, the sparsest catalog, SSB's advantage is largest on this rarest bucket. It improves Recall@20 over CGRC by 27.3% and Recall@10 by 31.2%, versus 21.6% and 20.9% over all cold items pooled, and the same holds in NDCG. SSB is also the best in this bucket on Toys. On Beauty, the densest catalog, the pattern reverses. CGRC leads on all four metrics in this bucket, significantly so on Recall@20. This is where the borrowed structural information does the most work, and where its limit shows. A Sports item in the rarest bucket has almost no interaction of its own, so warm-item structural information is the largest share of what the model knows about it. In dense Beauty the warm structural information is abundant and CGRC's reconstruction already supplies enough signal.

4.4 Ablation Study (RQ3)

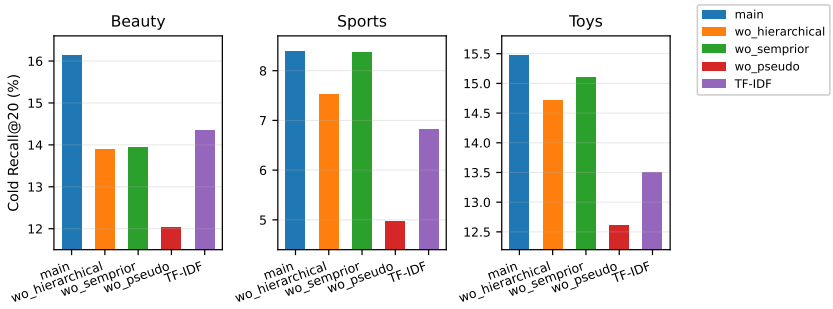

Figure 2: Ablation study: cold Recall@20 of the full SSB and its variants (wo_hierarchical removes hierarchical structure aggregation, wo_semprior removes the semantic prior, wo_pseudo removes the pseudo-cold mask method, and TF-IDF replaces the LLM semantics with TF-IDF vectors).

Figure 2 reports the ablation of SSB's key components and semantic source (all variants use a single fixed seed 42). Removing any component degrades cold Recall@20 on all three datasets. The largest drop comes from removing the pseudo-cold mask method (wo_pseudo), which reduces Recall@20 by 25.4% on Beauty, 40.6% on Sports, and 18.4% on Toys, so training and evaluation should face the same reconstruction of an isolated item. Removing hierarchical structure aggregation (wo_hierarchical) also hurts on all three datasets (13.9% on Beauty, 10.2% on Sports, and 4.9% on Toys), which supports multi-hop reading. Removing the semantic prior (wo_semprior) hurts substantially on Beauty (13.5%) but contributes little on Sports and Toys (0.2% and 2.4%). This ablation drops the semantic vector only from the final readout, keeping the retrieval and attention query, so the gap isolates the contribution of the prior to the final representation. In sparse catalogs, semantically similar items tend to be linked to the same users, so borrowing their structural information already gives a cold item the structural information it is missing. In dense Beauty this overlap is weaker, and the semantic prior still adds information. The learned layer weights confirm this asymmetry. Beauty gives the semantic layer the largest weight among the three (0.357), whereas Sports gives it only

0.118. Replacing the LLM semantic vectors with TF-IDF vectors, with everything else unchanged, also clearly hurts cold Recall@20 on all three datasets (LLM vs. TF-IDF: 16.14% vs. 14.35% on Beauty, 8.39% vs. 6.82% on Sports, and 15.47% vs. 13.51% on Toys). TF-IDF is a classic sparse bag-of-words encoding that captures lexical overlap but no semantic similarity. The richer semantic signal of the LLM substantially improves locating the right warm items to borrow from.

4.5 Parameter Study (RQ4)

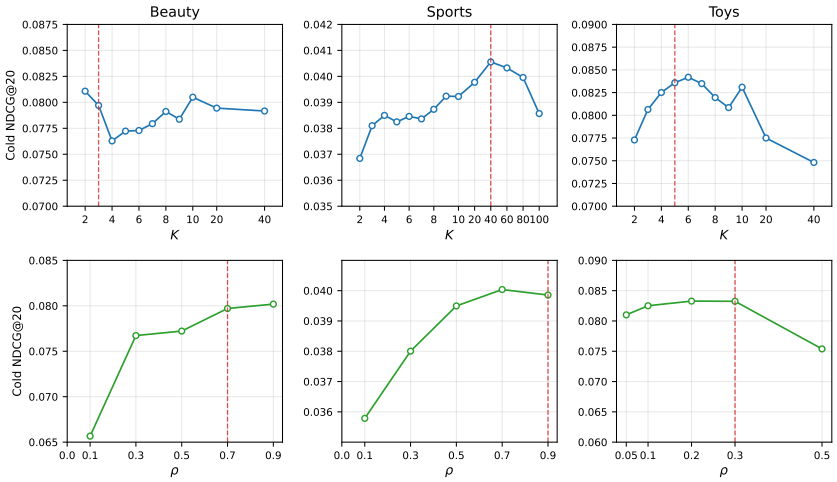

Figure 3: Parameter sensitivity of cold NDCG@20 to the warm semantic-neighbor size K (top row) and the pseudo-cold ratio ρ (bottom row) on Beauty, Sports, and Toys (left to right). Dashed lines mark the selected value.

We study the sensitivity of SSB to its two most influential hyperparameters, the warm semantic-neighbor size K and the pseudo-cold ratio ρ, in Figure 3. The optimal K varies considerably across datasets, shaped by multiple factors. On Sports it peaks at $K = 40$ and declines past $K = 60$, on Toys it peaks at $K \approx 5$–6, and on Beauty it stays flat across $K \in [2, 40]$. An overly large K eventually injects noise from distant and irrelevant neighbors, as both Sports and Toys show. The pseudo-cold ratio ρ matters mainly when it is too small. Performance drops for $\rho < 0.3$ on Beauty and Sports, approaching the fully-removed setting of the ablation (Section 4.4). A large ρ is mostly harmless, except on Toys, where $\rho = 0.5$ already hurts. Performance is otherwise stable within a reasonable range. The remaining hyperparameters (the temperature τ, the number of read layers H, and the negative ratio) are stable within their searched ranges, and their full grids are reported in Appendix B.

5 Conclusion

In this work, we propose an end-to-end framework named SSB (Semantic-guided Structural information Borrowing) to address the item cold-start problem. SSB employs a fine-tuning-free LLM to locate the warm semantic neighbors for a cold item. Subsequently, a soft structural borrowing mechanism represents the cold item as a continuous and learnable combination of the structural embeddings of its warm semantic neighbors. In addition, the proposed pseudo-cold masking method is tailored to avoid missing gradients problem during training. This is because the gradients for cold items would be unavailable due to the lack of structural information. Extensive experiments demonstrate that SSB outperforms SOTA cold-start baselines. Especially, when the cold item appears in the sparse user-item interaction graph and this cold item has little interaction with other users, SSB can still successfully recommend it. Although the user cold-start problem is worth studying, it is beyond the scope of this work. We leave the extension of SSB to cold users by introducing user-side semantics, for future work.

AI Use Statement

We used generative AI tools (Large Language Models) to aid and polish the writing, to retrieve and discover relevant literature, and to draft sections of the paper. We did not use them to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, design the methodology or experiments, implement methods, or interpret results. All research ideas, method design, experiments, data analysis, and final decision-making were conducted by the human authors, who reviewed and edited all AI-assisted text and take full responsibility for the final content of this work.

Ethics Statement

This work uses only publicly available, pre-existing benchmark datasets (the Amazon review subsets Beauty, Sports and Outdoors, and Toys and Games, and MovieLens-1M). It involves no human subjects, no personally identifiable information, and no data collected for this study. The LLM used for offline item encoding (Llama-3.1-8B) is a publicly released model. We see no discrimination, safety, or fairness harms beyond the general considerations of recommender system research.

Reproducibility Statement

We describe the datasets, preprocessing, evaluation protocol, and hyper-parameter search grids in Appendix B, and list the prompt templates in Appendix A. SSB and CGRC are averaged over five random seeds (42, 123, 2024, 7, 999), and the remaining baselines use a single seed. All other experiments use a fixed random seed 42. Code and configuration files will be released upon acceptance.

References

Hao Chen, Zefan Wang, Feiran Huang, Xiao Huang, Yue Xu, Yishi Lin, Peng He, and Zhoujun Li. Generative Adversarial Framework for Cold-Start Item Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2565–2571, Madrid Spain, July 2022. ACM. ISBN 978-1-4503-8732-3. doi: 10.1145/3477495.3531897.

Manqing Dong, Feng Yuan, Lina Yao, Xiwei Xu, and Liming Zhu. MAMO: Memory-Augmented Meta-Optimization for Cold-start Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 688–697, Virtual Event CA USA, August 2020. ACM. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403113.

Xiaoyu Du, Xiang Wang, Xiangnan He, Zechao Li, Jinhui Tang, and Tat-Seng Chua. How to Learn Item Representation for Cold-Start Multimedia Recommendation? In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3469–3477, Seattle WA USA, October 2020. ACM. ISBN 978-1-4503-7988-5. doi: 10.1145/3394171.3413628.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648, Virtual Event China, July 2020. ACM. ISBN 978-1-4503-8016-4. doi: 10.1145/3397271.3401063.

Jun Hu, Wenwen Xia, Xiaolu Zhang, Chilin Fu, Weichang Wu, Zhaoxin Huan, Ang Li, Zuoli Tang, and Jun Zhou. Enhancing Sequential Recommendation via LLM-based Semantic Embedding Learning. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 103–111, Singapore Singapore, May 2024. ACM. ISBN 979-8-4007-0172-6. doi: 10.1145/3589335.3648307.

Feiran Huang, Zefan Wang, Xiao Huang, Yufeng Qian, Zhetao Li, and Hao Chen. Aligning Distillation For Cold-start Item Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1147–1157, Taipei Taiwan, July 2023. ACM. ISBN 978-1-4503-9408-6. doi: 10.1145/3539618.3591732.

Feiran Huang, Yuanchen Bei, Zhenghang Yang, Junyi Jiang, Hao Chen, Qijie Shen, Senzhang Wang, Fakhri Karray, and Philip S. Yu. Large Language Model Simulator for Cold-Start Recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 261–270, Hannover Germany, March 2025. ACM. ISBN 979-8-4007-1329-3. doi: 10.1145/3701551.3703546.

Hai-Dang Kieu, Minh Duc Nguyen, Thanh-Son Nguyen, and Dung D. Le. Keyword-driven Retrieval-Augmented Large Language Models for Cold-start User Recommendations. In *Companion Proceedings of the ACM Web Conference 2025 (WWW '25 Companion)*, pp. 1–5. ACM, April 2025. doi: 10.1145/3701716.3717855.

Jinri Kim, Eungi Kim, Kwangeun Yeo, Yujin Jeon, Chanwoo Kim, Sewon Lee, and Joonseok Lee. Content-based Graph Reconstruction for Cold-start Item Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1263–1273, Washington DC USA, July 2024a. ACM. ISBN 979-8-4007-0431-4. doi: 10.1145/3626772.3657801.

Jooeun Kim, Jinri Kim, Kwangeun Yeo, Eungi Kim, Kyoung-Woon On, Jonghwan Mun, and Joonseok Lee. General Item Representation Learning for Cold-start Content Recommendations. *arXiv preprint arXiv:2404.13808*, April 2024b. doi: 10.48550/arXiv.2404.13808.

Songyuan Lei, Xinglong Chang, Zhizhi Yu, Dongxiao He, Cuiying Huo, Jianrong Wang, and Di Jin. Feature-Structure Adaptive Completion Graph Neural Network for Cold-start Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12022–12030, 2025.

Yueqing Liang, Liangwei Yang, Chen Wang, Xiongxiao Xu, Philip S. Yu, and Kai Shu. Taxonomy-Guided Zero-Shot Recommendations with LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pp. 1520–1530. Association for Computational Linguistics, January 2025. doi: 10.18653/v1/2025.coling-main.102.

Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. LLaRA: Large Language-Recommendation Assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1785–1795, Washington DC USA, July 2024. ACM. ISBN 979-8-4007-0431-4. doi: 10.1145/3626772.3657690.

Ruochen Liu, Hao Chen, Yuanchen Bei, Lijia Chen, Qijie Shen, Feiran Huang, Fakhri Karray, and Senzhang Wang. SetLLM: Set Large Language Model for Cold-Start Item Recommendation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 3282–3293, Jeju Island, Republic of Korea, August 2026. ACM. ISBN 979-8-4007-2259-2. doi: 10.1145/3770855.3817707.

Yuanfu Lu, Yuan Fang, and Chuan Shi. Meta-learning on Heterogeneous Information Networks for Cold-start Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1563–1573, Virtual Event CA USA, August 2020. ACM. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403207.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001. doi: 10.1146/annurev.soc.27.1.415.

Tieyun Qian, Yile Liang, and Qing Li. Solving Cold Start Problem in Recommendation with Attribute Graph Neural Networks. *IEEE Transactions on Knowledge and Data Engineering*, 34(8): 3597–3610, August 2022. ISSN 1041-4347, 1558-2191, 2326-3865. doi: 10.1109/TKDE.2020.3038234.

Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Representation Learning with Large Language Models for Recommendation. In *Proceedings of the ACM Web Conference 2024*, pp. 3464–3475, Singapore Singapore, May 2024. ACM. ISBN 979-8-4007-0171-9. doi: 10.1145/3589334.3645458.

Keigo Sakurai, Ren Togo, Takahiro Ogawa, and Miki Haseyama. LLM is Knowledge Graph Reasoner: LLM's Intuition-aware Knowledge Graph Reasoning for Cold-start Sequential Recommendation. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval (ECIR 2025)*, volume 15573 of *Lecture Notes in Computer Science*, pp. 263–278. Springer, 2025. doi: 10.1007/978-3-031-88711-6_17.

Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. DropoutNet: Addressing Cold Start in Recommender Systems. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4957–4966, 2017. doi: 10.5555/3295222.3295249.

Jianling Wang, Haokai Lu, James Caverlee, Ed H. Chi, and Minmin Chen. Large Language Models as Data Augmenters for Cold-Start Item Recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 726–729, Singapore Singapore, May 2024a. ACM. ISBN 979-8-4007-0172-6. doi: 10.1145/3589335.3651532.

Wenbo Wang, Ben Chen, Bingquan Liu, Xinxin Wang, Luwei Yang, Wen Jiang, Wei Ning, and Jian Guan. Mutual Information Assisted Graph Convolution Network for Cold-Start Recommendation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6785–6789, Seoul, Korea, Republic of, April 2024b. IEEE. ISBN 979-8-3503-4485-1. doi: 10.1109/ICASSP48485.2024.10447656.

Wenbo Wang, Bingquan Liu, Lili Shan, Chengjie Sun, Ben Chen, and Jian Guan. Preference Aware Dual Contrastive Learning for Item Cold-Start Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):9125–9132, March 2024c. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i8.28763.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. LLMRec: Large Language Models with Graph Augmentation for Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 806–815, Merida Mexico, March 2024. ACM. ISBN 979-8-4007-0371-3. doi: 10.1145/3616855.3635853.

Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. Contrastive Learning for Cold-Start Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 5382–5390, Virtual Event China, October 2021. ACM. ISBN 978-1-4503-8651-7. doi: 10.1145/3474085.3475665.

Xiaolong Xu, Hongsheng Dong, Lianyong Qi, Xuyun Zhang, Haolong Xiang, Xiaoyu Xia, Yanwei Xu, and Wanchun Dou. CMCLRec: Cross-modal Contrastive Learning for User Cold-start Sequential Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1589–1598, Washington DC USA, July 2024. ACM. ISBN 979-8-4007-0431-4. doi: 10.1145/3626772.3657839.

Yining Yao, Ziwei Li, Shuwen Xiao, Boya Du, Jialin Zhu, Junjun Zheng, Xiangheng Kong, and Yuning Jiang. SaviorRec: Semantic-Behavior Alignment for Cold-Start Recommendation. *arXiv preprint arXiv:2508.01375*, August 2025. doi: 10.48550/arXiv.2508.01375.

Zhihui Zhou, Lilin Zhang, and Ning Yang. Contrastive Collaborative Filtering for Cold-Start Item Recommendation. In *Proceedings of the ACM Web Conference 2023*, pp. 928–937, Austin TX USA, April 2023. ACM. ISBN 978-1-4503-9416-1. doi: 10.1145/3543507.3583286.

Ziwei Zhu, Shahin Sefati, Parsa Saadatpanah, and James Caverlee. Recommendation for New Users and New Items via Randomized Training and Mixture-of-Experts Transformation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1121–1130, Virtual Event China, July 2020. ACM. ISBN 978-1-4503-8016-4. doi: 10.1145/3397271.3401178.

A Method Details

A.1 Prompt Templates

The LLM encodes the textual description of each item through a fixed prompt. We design a domain-specific prompt for each dataset.

Beauty prompt (node_prompt_Beauty.txt):

```
[SYSTEM]
You are an entity encoder for a beauty and personal-care product
↪ recommendation graph. Your goal is to produce a dense semantic vector
↪ that makes each product as distinguishable as possible from all other
↪ beauty products.

[ENTITY DESCRIPTION]
{node_text}

[ENCODING GUIDANCE]
Focus on what the product does and who it is for: its function and efficacy
↪ claims, skin/hair type suitability, scent or fragrance notes, key
↪ ingredients, form/texture, and intended use occasion. Weight the
↪ attributes that separate this product from other products in the same
↪ category. De-emphasize marketing boilerplate and non-discriminative
↪ wording.
```

Sports prompt (node_prompt_Sports.txt):

```
[SYSTEM]
You are an entity encoder for a sports and outdoor product recommendation
↪ graph. Your goal is to produce a dense semantic vector that makes each
↪ product as distinguishable as possible from all other sports and
↪ outdoor products.

[ENTITY DESCRIPTION]
{node_text}

[ENCODING GUIDANCE]
Focus on what the product is used for and who it suits: the sport or
↪ activity it targets, its function and performance characteristics,
↪ material and construction, usage scenario and environment, and target
↪ user level. Weight the attributes that separate this product from other
↪ products in the same category. De-emphasize marketing boilerplate and
↪ non-discriminative wording.
```

Toys prompt (node_prompt_Toys.txt):

```
[SYSTEM]
You are an entity encoder for a toys and games product recommendation
↪ graph. Your goal is to produce a dense semantic vector that makes each
↪ product as distinguishable as possible from all other toys and games
↪ products.

[ENTITY DESCRIPTION]
{node_text}

[ENCODING GUIDANCE]
```

```
Focus on what the toy is and who it is for: the play type and theme,
↪ recommended age range, developmental or educational value, material and
↪ safety attributes, and intended play context. Weight the attributes
↪ that separate this product from other products in the same category.
↪ De-emphasize marketing boilerplate and non-discriminative wording.
```

ML-1M prompt (node_prompt_ML1M.txt):

```
[SYSTEM]
You are an entity encoder for a movie recommendation graph. Your goal is to
↪ produce a dense semantic vector that makes each movie as
↪ distinguishable as possible from all other movies.

[ENTITY DESCRIPTION]
{node_text}

[ENCODING GUIDANCE]
Focus on the genre and title of the movie: its themes, subject matter, and
↪ intended audience. Weight the attributes that separate this movie from
↪ other movies. De-emphasize boilerplate and non-discriminative wording.
```

B Reproducibility

B.1 Baseline Descriptions

All content-based baselines (DropoutNet, Heater, CLCRec, ALDI, and CGRC) receive the same offline 4096-dimensional Llama-3.1-8B item embeddings as SSB, so encoder capacity is held fixed across methods. Warm metrics for ALDI and Heater are evaluated with their shared teacher (BPRMF) embeddings.

B.2 Hyperparameter Search Space

Table 3 lists the grid-search ranges. A subset of parameters is refined per dataset beyond these ranges. SSB is tuned on the cold validation set.

Table 3: Grid-search ranges. A subset of parameters is refined per dataset beyond these ranges.

Parameter	Candidates
τ	0.03, 0.05, 0.1, 0.2, 0.3
K	2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 80, 100
learning rate	5e-5, 1e-4, 3e-4, 1e-3
gnn_hidden	128, 256
num_layers	1, 2, 3
neg_ratio	10, 30, 60, 90, 120
ρ	0.0, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9
H	1, 2, 3

Table 4: Final per-dataset hyperparameters.

Dataset	K	τ	neg. ratio	ρ	lr	H
Beauty	3	0.1	120	0.7	10^{-4}	2
Sports	40	0.1	90	0.9	5×10^{-5}	2
Toys	5	0.1	60	0.3	5×10^{-5}	2
ML-1M	40	0.05	60	0.95	10^{-3}	1

B.3 Fairness Notes

ColdLLM's low scores, particularly on the large sparse Sports dataset, reflect that its pair-wise online LLM simulation has too little structural information to judge which users would like a cold item in a catalog with hundreds of thousands of users and extreme sparsity.

C Cross-Domain and Semantic Dependence

C.1 Motivation

SSB's locating step retrieves semantically similar warm items, so its behavior depends on how discriminative the item text is. The three Amazon catalogs in the main text carry rich item descriptions (title, brand, category, features, and description), whereas ML-1M items carry only a title and a genre. ML-1M is therefore a useful probe. Its genre signal is a strong *user-item* categorical cue but a weak *item-item* semantic cue. This should favor methods that exploit user-item structure or simulate embeddings from content (CGRC, Heater, DropoutNet) and disadvantage item-item soft structure borrowing (SSB). ML-1M has 6,038 users, 3,382 items (676 cold), and 243,306 interactions.

C.2 Cross-Domain Results on ML-1M

Table 5 reports cold Recall@20 and NDCG@20 on ML-1M under the same protocol as the main text. SSB falls to the middle of the content-based pack. It is behind CGRC, Heater, and DropoutNet, which exploit the strong genre signal, essentially tied with SetLLM, and still clearly ahead of the pure collaborative backbone LightGCN (0.1176 vs. 0.0868 on Recall@20, $+35.5\%$). This matches our hypothesis. When item-item semantics are weak, semantic locating can no longer select informative warm neighbors, and SSB's advantage over methods that do not depend on item-item similarity shrinks accordingly.

Table 5: Cold-start performance on ML-1M (Recall@20 / NDCG@20, in %) over all cold items and the rarest bucket ($\leq$2). Best results are in bold. Second-best results are underlined.

Method	Recall@20	NDCG@20	R@20 ($\leq$2)	N@20 ($\leq$2)
CGRC	**17.59**	**8.78**	**22.64**	**10.83**
Heater	<u>13.44</u>	<u>6.20</u>	14.93	6.30
DropoutNet	12.89	5.81	13.76	6.56
SSB	11.76	5.38	10.45	3.19
SetLLM	11.74	5.78	6.72	1.76
CLCRec	10.82	5.03	7.88	2.31
ColdLLM	8.73	3.82	4.48	1.33
LightGCN	8.68	4.11	4.73	1.56
ALDI	4.30	2.00	3.15	2.15

C.3 Diagnosis: The Rarest-Bucket Reversal

The most informative contrast is on the rarest cold items (at most two interactions). On Amazon Sports, where structural information is scarcest, SSB improves Recall@20 over CGRC by 27.3%. On ML-1M the same bucket reverses, with CGRC more than doubling SSB (22.64% vs. 10.45%). This reversal matches "semantics locate, structure borrows". SSB borrows structural information through semantic neighbors, so it helps most where structural information is scarce *and* semantics are informative, and it fails first where structural information is scarce *but* semantics are uninformative.