

Disentangling Semantics and Popularity: Adversarial Debiasing and Asymmetric Alignment for LLM-Enhanced Long-Tail Sequential Recommendation

Wenzhe Dong*

School of New Media
and Communication
Tianjin University
Tianjin, China
dongwz@tju.edu.cn

Minglai Shao*

School of New Media
and Communication
Tianjin University
Tianjin, China
shaoml@tju.edu.cn

Qiyao Peng†

School of Artificial
Intelligence
Tianjin University
Tianjin, China
qypeng@tju.edu.cn

Abstract

Sequential recommendation models dynamic user preferences but is constrained by the long-tail distribution of real-world interactions. While LLM-derived text embeddings provide semantic priors to alleviate data sparsity, existing adaptation methods are hindered by popularity bias during semantic projection and semantic prior erosion during sequence training. In this paper, we propose **Adversarial Debiasing and Asymmetric Alignment (ADAA)**, a framework designed to alleviate these representation bottlenecks. Specifically, ADAA incorporates an adversarial debiasing mechanism to decouple interaction frequency factors from the projected embeddings, yielding debiased semantic representations. Furthermore, it employs an asymmetric anchoring strategy during the sequence training phase. This strategy restricts the optimization space of sparse items to preserve their injected semantic priors, while permitting head items to flexibly fit sequential transition patterns. Through a parameter-level injection mechanism, ADAA integrates offline-aligned semantics without introducing additional parameters to the online architecture. Extensive experiments on three real-world datasets demonstrate that ADAA improves long-tail recommendation accuracy while preserving the online inference efficiency of the original sequential backbone.

1 Introduction

Sequential Recommendation (SR) models dynamic user preferences by mining chronological interaction histories (Tang and Wang, 2018a; Wang et al., 2019; Peng et al., 2025). The efficacy of these methods relies on capturing sequential transition patterns and behavioral dependencies from discrete item identifiers (IDs) (Kang and McAuley, 2018;

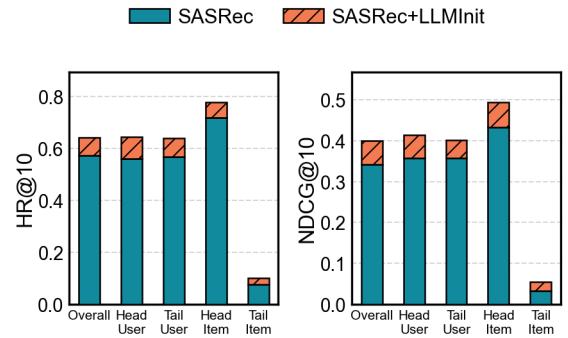


Figure 1: Performance comparison between SASRec and SASRec+LLMInit across different frequency-stratified user and item groups on the Yelp dataset.

Sun et al., 2019). However, real-world feedback data typically follows a long-tail distribution (Yin et al., 2012; Abdollahpouri et al., 2017). While a fraction of popular items yields well-optimized representations through abundant interactions, the majority of items lack sufficient records (Hou et al., 2022b; Kang et al., 2019). Consequently, existing ID-based models fail to establish reliable transition contexts for these sparse items (Jang et al., 2020; Kim et al., 2023), leading to representation degradation and compromised recommendation accuracy.

Recent research leverages Large Language Models (LLMs) to alleviate data sparsity in long-tail recommendation (Wu et al., 2024; Zhao et al., 2021). Derived from extensive text corpora, LLM embeddings encode semantic priors that provide semantic information beyond collaborative interaction signals (Hou et al., 2024; Bao et al., 2023). In this work, LLM embeddings refer to precomputed text embeddings extracted from item metadata by a frozen language model before recommendation training, rather than online generative reasoning during recommendation. A preliminary experiment on the Yelp dataset demonstrates that initializing the SASRec backbone (Kang and McAuley,

*Equal contribution.

†Corresponding author.

2018) with LLM embeddings improves overall recommendation accuracy. However, a frequency-stratified performance analysis reveals an uneven distribution of these gains. As illustrated in Figure 1, the overall improvements are largely concentrated on head items, whereas the enhancements for tail items remain marginal. This observation suggests that standard semantic initialization alone is insufficient to effectively address the long-tail challenge.

This skewed performance improvement indicates that simply injecting semantic priors is insufficient to resolve the long-tail issue. We attribute this to three challenges in existing semantic adaptation methods:

- i) **Inadequate Semantic Alignment:** Textual semantics encode content attributes, whereas collaborative IDs capture behavioral co-occurrences. This discrepancy causes an inherent structural incompatibility. Bridging this gap via standard uniform projections often discards fine-grained semantic features, inducing an information bottleneck.
- ii) **Popularity Bias in Projection:** Standard alignment processes are dominated by the interaction signals of head items, which embeds a popularity bias into the learned mapping. Applying this biased projector to tail items degrades their semantic representations, causing collaborative popularity factors to override their intrinsic textual features (Zhang et al., 2021; Wei et al., 2021; Zheng et al., 2021).
- iii) **Semantic Prior Erosion:** As user sequences consist largely of popular items, the representation updating process is skewed toward high-frequency interactions. For tail items, this biased dynamic erodes their initialized semantic priors, negating the benefits of LLM enhancement.

To address these challenges, we propose the **Adversarial Debiasing and Asymmetric Alignment (ADAA)**, a framework that integrates textual semantic priors into sequential models to mitigate the representation bottlenecks of long-tail items. Specifically, to bridge the structural incompatibility between semantic and collaborative spaces, we develop Collaborative Anchored Semantic Alignment. By recognizing that popular items inherently possess stable collaborative representations, we utilize them as reliable targets to guide the cross-space mapping. This anchors the projection pro-

cess, aligning the feature spaces while preventing the loss of fine-grained semantic properties. Second, to prevent the mapping rules from overfitting head-item frequencies, we introduce an Adversarial Popularity Debiasing mechanism. By formulating the frequency discrepancy as a domain adaptation problem (Ganin et al., 2016; Tzeng et al., 2017), we employ an adversarial minimax game to decouple popularity attributes from the projected embeddings (Ganin and Lempitsky, 2015; Goodfellow et al., 2014), yielding a debiased mapping function. Third, to mitigate semantic prior erosion during sequence training, we propose an Asymmetric Semantic Anchoring strategy. Since sparse items lack sufficient interactions for reliable representation updating, this mechanism applies asymmetric constraints to restrict their optimization space, thereby preserving the injected semantic structures. Concurrently, it permits head items to flexibly fit sequential transition patterns. By systematically mitigating these adaptation bottlenecks, ADAA improves long-tail recommendation accuracy while preserving the architectural integrity and computational efficiency of the underlying sequential backbone.

While ADAA builds on established training objectives such as contrastive alignment, adversarial debiasing, and representation regularization, its key contribution is to identify and implement a head-tail asymmetric adaptation strategy: head items are anchored toward reliable collaborative representations, whereas tail items are anchored toward semantic priors derived from item text. The main contributions of this work are summarized as follows:

- We identify three challenges in applying LLMs to long-tail recommendation, and advocate an asymmetric modeling philosophy tailored to item frequencies.
- We propose ADAA, a two-stage adaptation framework that combines adversarial semantic projection with head-tail asymmetric anchoring, preserving collaborative signals for head items and semantic priors for tail items.
- We introduce a non-intrusive parameter injection strategy. Extensive experiments demonstrate ADAA significantly improves tail-item accuracy while maintaining the online inference cost of vanilla backbones.

2 Preliminary

2.1 Sequential Recommendation

Let $\mathcal{U}$ and $\mathcal{V}$ be the user and item sets. For each user $u \in \mathcal{U}$ with a chronological interaction sequence $S_u = [v_1, \dots, v_{|S_u|}]$, sequential recommendation aims to predict the next item $v^+ \in \mathcal{V}$ by maximizing:

$$\arg \max_{v \in \mathcal{V}} P(v^+ = v \mid S_u). \quad (1)$$

Real-world interactions typically follow a long-tail distribution. Let p_v denote item v 's popularity (total historical interactions). Sorting $\mathcal{V}$ by descending p_v , we follow the empirical Pareto principle (Newman, 2005) to designate the top 20% of items as the head set $\mathcal{V}_H$. The remaining 80% constitutes the sparse tail set $\mathcal{V}_T = \mathcal{V} \setminus \mathcal{V}_H$. This extreme data imbalance ($|\mathcal{V}_T| \gg |\mathcal{V}_H|$) poses the core optimization challenge addressed in this work.

2.2 Generalized Sequential Encoder

Each item $v \in \mathcal{V}$ is parameterized by a trainable collaborative ID embedding $\mathbf{e}_v^{id} \in \mathbb{R}^d$. To capture inherent textual features, we additionally introduce a static semantic embedding $\mathbf{e}_v^{llm} \in \mathbb{R}^{d_{llm}}$, extracted via text-embedding-ada-002 (Neelakantan et al., 2022) from item metadata. The raw textual metadata of items is formatted into natural language descriptions to query the frozen LLM. The exact, structured prompt templates employed for each dataset are presented in Appendix A.1.

A generalized, backbone-agnostic sequential encoder $f_\theta(\cdot)$ processes the stacked input matrix $\mathbf{E}_{S_u} = [\mathbf{e}_{v_1}, \dots, \mathbf{e}_{v_{|S_u|}}]$ to generate a user preference vector:

$$\mathbf{h}_u = f_\theta(\mathbf{E}_{S_u}). \quad (2)$$

The prediction score for a candidate item v is then computed via the inner product $\hat{y}_{u,v} \propto \mathbf{h}_u^\top \mathbf{e}_v$.

3 The Framework of ADAA

Figure 2 illustrates the ADAA framework, which leverages textual side information to enhance long-tail sequential recommendation. First, **Collaborative Anchored Semantic Alignment** projects LLM-encoded features into the collaborative space using head items as targets. Next, **Adversarial Domain Adaptation for Popularity Debiasing** employs adversarial training to decouple interaction frequencies from the projected embeddings, mitigating popularity bias. Finally, **Parameter-Level**

Injection and Asymmetric Alignment initializes the sequential backbone with these debiased representations. During sequence training, asymmetric anchoring constraints preserve tail-item semantic priors while enabling head items to capture sequential transition patterns.

3.1 Collaborative Anchored Semantic Alignment

LLM semantic embeddings and collaborative ID representations inherently reside in heterogeneous latent spaces. To effectively utilize textual priors for recommendation, these semantic representations must be projected into the collaborative space. To achieve this cross-space mapping while preserving the intrinsic semantic topology, we introduce a collaborative-anchored alignment mechanism constrained to the head set $\mathcal{V}_H$.

Collaborative Anchor Extraction. Cross-space alignment requires reliable reference targets. We first pre-train the sequential backbone f_θ using standard sequence modeling objectives. Given the dense interactions within $\mathcal{V}_H$, this pre-training yields stable collaborative embeddings, denoted as $\mathbf{E}_H^{id} \in \mathbb{R}^{|\mathcal{V}_H| \times d}$. We freeze $\mathbf{E}_H^{id}$ to prevent representation drift during the alignment phase. These frozen embeddings serve as structural anchors for the subsequent semantic projection, isolating the alignment process from the under-optimized tail representations.

Frequency-Aware Semantic Projection. We introduce an MLP projector $P(\cdot)$ to map the static semantic embeddings $\mathbf{e}_i^{llm}$ into the collaborative space. Directly applying Mean Squared Error (MSE) enforces absolute coordinate regression, which collapses diverse semantic representations into dense collaborative regions and erases their intrinsic variance.

To prevent mapping collapse, we optimize a contrastive objective (Gutmann and Hyvärinen, 2010; Chen et al., 2020) based on relative similarities. By treating paired semantic and collaborative embeddings of the same item as positive instances and other intra-batch anchors as negatives, the contrastive loss for item i in a sampled batch $\mathcal{B}$ is:

$$\mathcal{L}_{NCE}(i) = -\log \frac{\exp(\text{sim}(\tilde{P}(\mathbf{e}_i^{llm}), \tilde{\mathbf{e}}_i^{id})/\tau)}{\sum_{j \in \mathcal{B}} \exp(\text{sim}(\tilde{P}(\mathbf{e}_i^{llm}), \tilde{\mathbf{e}}_j^{id})/\tau)}, \quad (3)$$

where $\tilde{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$ denotes the ℓ_2 -normalized vector, τ is the temperature, and $\text{sim}(\cdot)$ denotes cosine

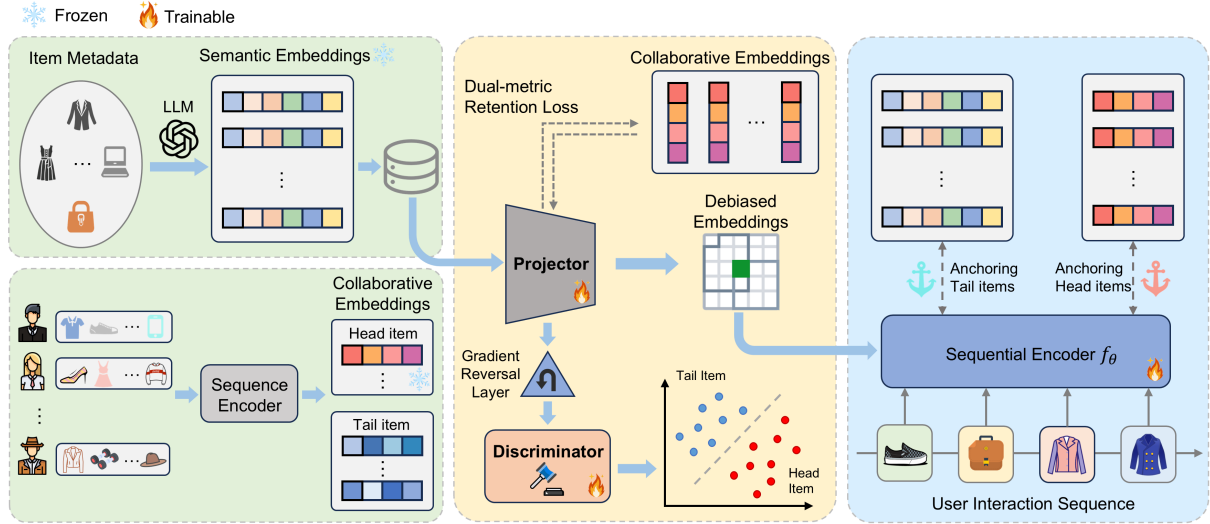


Figure 2: Overview of the proposed ADAA Framework. Frozen semantic embeddings are transformed by a projector trained with collaborative embeddings via dual-metric retention and adversarial head–tail discrimination. During sequential training, tail and head items are anchored to semantic and collaborative embeddings, respectively.

similarity.

Since the stability of collaborative representations correlates with interaction frequencies, sparse items yield unreliable alignment targets. To mitigate interference from these sparse anchors, we scale the contrastive loss with a frequency-based weight c_i :

$$c_i = \omega \times \frac{\log(\text{count}(i) + 1)}{\log(\text{count}_{\max} + 1)} + b, \quad (4)$$

where ω and b are scaling hyperparameters. The overall alignment loss is the frequency-weighted batch average:

$$\mathcal{L}_{\text{align}} = \frac{\sum_{i \in \mathcal{B}} c_i \cdot \mathcal{L}_{\text{NCE}}(i)}{\sum_{i \in \mathcal{B}} c_i}. \quad (5)$$

This formulation removes magnitude biases via ℓ_2 -normalization and ensures the alignment process is guided primarily by the reliable collaborative signals of head items.

3.2 Adversarial Domain Adaptation for Popularity Debiasing

While the projector $P(\cdot)$ establishes a valid cross-space mapping on head items, applying it directly to tail items introduces popularity bias. Because head-item collaborative targets are dominated by high-frequency interactions, $P(\cdot)$ inadvertently fits these popularity-specific features during alignment. Applying this unconstrained projector to tail items shifts their semantic representations into high-frequency regions. To mitigate this, we propose

an adversarial adaptation mechanism to decouple popularity bias from the mapped features.

Adversarial Popularity Decoupling. We formulate popularity debiasing as a domain adaptation problem (Ganin et al., 2016; Tzeng et al., 2017). By treating the high-frequency head set $\mathcal{V}_H$ and low-frequency tail set $\mathcal{V}_T$ as distinct feature domains, we train a domain discriminator $D(\cdot)$ to identify the frequency origin of the projected semantic feature $\mathbf{z}_i = P(\mathbf{e}_i^{\text{llm}})$. The discriminator classifies $\mathbf{z}_i$ into either the head domain ($d_i = 1$) or the tail domain ($d_i = 0$), optimized via binary cross-entropy loss:

$$\begin{aligned} \mathcal{L}_{\text{adv}} = -\mathbb{E}_{v_i \in \mathcal{V}} & \left[d_i \log D(\mathbf{z}_i) \right. \\ & \left. + (1 - d_i) \log(1 - D(\mathbf{z}_i)) \right]. \end{aligned} \quad (6)$$

To decouple the popularity bias, we establish a minimax game between $P(\cdot)$ and $D(\cdot)$ via a Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015). While $D(\cdot)$ minimizes the domain classification error, $P(\cdot)$ is optimized to maximize the discriminator’s uncertainty. This compels $P(\cdot)$ to discard popularity-specific attributes, yielding debiased semantic representations and reducing the influence of item interaction frequencies on the projected representations.

Dual-Metric Mapping Retention. Relying exclusively on the adversarial objective risks mapping collapse. Optimizing $P(\cdot)$ solely to maximize discriminator uncertainty may cause it to map diverse

inputs into a constant vector, which destroys item distinctiveness and discards valid mapping rules.

To prevent this degradation, we apply a dual-metric retention loss $\mathcal{L}_{ret}$ to the head items, integrating Mean Squared Error and cosine similarity:

$$\mathcal{L}_{ret} = \sum_{v_i \in \mathcal{V}_H} \left(\|P(\mathbf{e}_i^{llm}) - \mathbf{e}_i^{id}\|_2^2 + \gamma \cdot (1 - \text{sim}(P(\mathbf{e}_i^{llm}), \mathbf{e}_i^{id})) \right), \quad (7)$$

where γ is a scaling weight. The Mean Squared Error bounds the absolute coordinate distance, while the cosine similarity maintains directional alignment against magnitude perturbations from adversarial updates.

We optimize the objective $\mathcal{L}_{proj} = \mathcal{L}_{ret} + \lambda \mathcal{L}_{adv}$, where λ controls the adversarial penalty. This enables $P(\cdot)$ to decouple popularity features while preserving cross-space topology. The resulting debiased semantic representations are then injected into the sequential encoder.

3.3 Parameter-Level Injection and Asymmetric Alignment

The final phase conducts sequence training of the recommendation backbone via parameter-level initialization and asymmetric anchoring.

Non-intrusive Semantic Initialization. Existing methods often integrate heterogeneous features via auxiliary cross-attention or gating layers (Wu et al., 2019; Hou et al., 2022a; Zhou et al., 2022). Such structural modifications alter the computational graph and increase inference latency.

To avoid structural fusion, we employ a parameter-level injection strategy. Projecting the static LLM embeddings through $P^*(\cdot)$ yields a global, debiased semantic matrix $\mathbf{Z}^* = P^*(\mathbf{E}^{llm}) \in \mathbb{R}^{|\mathcal{V}| \times d}$. We initialize the item embedding matrix of the sequential backbone f_θ by directly overwriting it with $\mathbf{Z}^*$.

Asymmetric Semantic Anchoring. Following initialization, f_θ is optimized via the standard sequential objective $\mathcal{L}_{seq}$. However, sequence training is skewed toward high-frequency collaborative signals. For sparse tail items $\mathcal{V}_T$, this biased optimization causes representation drift, altering their initialized semantic priors.

To address this, we introduce an asymmetric anchoring constraint $\mathcal{L}_{anchor}$. We assign distinct reference targets $\hat{\mathbf{e}}_i$ based on the frequency domain

of each item:

$$\hat{\mathbf{e}}_i = \begin{cases} \mathbf{e}_i^{id}, & \text{if } v_i \in \mathcal{V}_H \\ \mathbf{z}_i^*, & \text{if } v_i \in \mathcal{V}_T \end{cases} \quad (8)$$

Let $\mathbf{h}_i$ denote the updated embedding of item v_i during sequence adaptation. The constraint is:

$$\mathcal{L}_{anchor} = \sum_{v_i \in \mathcal{V}} (1 - \text{sim}(\mathbf{h}_i, \hat{\mathbf{e}}_i)). \quad (9)$$

This asymmetric design serves distinct purposes. For head items, aligning with the pre-trained collaborative ID $\mathbf{e}_i^{id}$ functions as knowledge distillation (Hinton et al., 2015; Tang and Wang, 2018b), transferring collaborative transition patterns into semantic representations. For tail items, aligning with their own debiased semantic feature $\mathbf{z}_i^*$ acts as an elastic regularization, anchoring representations against high-frequency gradients. We apply an exponentially decaying coefficient $\alpha^{(e)} = \alpha_0 \cdot \eta^e$ where $\eta \in (0, 1)$ at epoch e . The overall objective is:

$$\mathcal{L} = \mathcal{L}_{seq} + \alpha^{(e)} \mathcal{L}_{anchor}. \quad (10)$$

This decaying weight prioritizes structural preservation initially and shifts focus toward personalized sequence modeling as training progresses.

3.4 Training and Inference

Training. Optimization follows a decoupled pipeline. First, keeping the static semantic embeddings $\mathbf{E}^{llm}$ and pre-trained collaborative anchors $\mathbf{E}_H^{id}$ frozen, we optimize the projector $P(\cdot)$ using the joint objective:

$$\mathcal{L}_{train} = \mathcal{L}_{align} + \mathcal{L}_{ret} + \lambda \mathcal{L}_{adv}. \quad (11)$$

After convergence, the learned projector $P^*(\cdot)$ generates the global debiased semantic matrix $\mathbf{Z}^* = P^*(\mathbf{E}^{llm})$ to initialize the embedding layer of the sequential backbone f_θ . The backbone f_θ then undergoes sequence training, regularized by the asymmetric anchoring objective:

$$\mathcal{L}_{tune} = \mathcal{L}_{seq} + \alpha^{(e)} \mathcal{L}_{anchor}. \quad (12)$$

Inference. ADAA preserves the architectural integrity of the sequential backbone through parameter-level injection. Item text embeddings are precomputed offline, and auxiliary training components, including the language model, $P(\cdot)$, $D(\cdot)$, and anchoring targets, are discarded post-training. Inference relies solely on the forward propagation

of the optimized backbone f_θ , ensuring long-tail accuracy improvements without increasing online inference time or space complexity. Detailed training and inference procedures are provided in Appendix A.2.

4 Experiments

4.1 Experimental Settings

Dataset. Three public real-world datasets are utilized for performance evaluation: Yelp, Amazon Fashion, and Amazon Beauty (Ni et al., 2019; McAuley et al., 2015). More comprehensive details are provided in Appendix B.1. Following the common practice in sequential recommendation, we conduct standardized preprocessing and data splitting. The detailed statistics and preprocessing steps are summarized in Appendix B.2.

Baselines. To evaluate the effectiveness and flexibility of ADAA, we implement it and all competing baselines upon two representative backbones: SAS-Rec (Kang and McAuley, 2018) and BERT4Rec (Sun et al., 2019). We compare ADAA against two categories of baselines: (1) traditional long-tail enhancement frameworks, including CITIES (Jang et al., 2020) and MELT (Kim et al., 2023); and (2) LLM-based enhancement methods, including RLMRec (Ren et al., 2024), LLMInit (Hu et al., 2024; Harte et al., 2023), SRA-CL (Cui et al., 2026), and LLM-ESR (Liu et al., 2024). Appendix B.3 provides further details on these baselines.

Evaluation Metrics. We evaluate the recommendation performance using the Top-10 ranking list. Specifically, we adopt the widely-used Hit Rate (H@10) and Normalized Discounted Cumulative Gain (N@10). Following common practice in sequential recommendation and the evaluation protocol of LLM-ESR, for each ground-truth item in the test set, we randomly sample 100 non-interacted items as negatives to calculate the main metrics. However, sampled metrics may be inconsistent with full-ranking metrics and may not always preserve the relative ordering of methods (Krichene and Rendle, 2020). Therefore, we additionally conduct full-catalog ranking evaluation in Appendix C.1, where each ground-truth item is ranked against all eligible unseen items. All implementations follow the configurations specified in Appendix B.4.

4.2 Comparison with Baselines

Table 1 summarizes the overall performance of all evaluated models. Overall, our proposed ADAA yields highly competitive results, establishing a new state-of-the-art on the vast majority of evaluation metrics. From the detailed comparison, we draw the following in-depth observations.

Overall Performance Comparison. Compared to traditional ID-based backbones, ADAA consistently improves performance, suggesting that semantic priors effectively mitigate representational limitations caused by interaction sparsity. Furthermore, ADAA outperforms existing LLM-enhanced baselines on most metrics. While standard semantic adaptation often suffers from popularity bias and the erosion of semantic structures during optimization, ADAA addresses these issues systematically. By decoupling interaction frequencies via adversarial learning and asymmetric anchoring, ADAA provides a stable, backbone-agnostic paradigm for sequential recommendation.

Performance on Long-Tail Items. As shown in Table 1, ADAA achieves leading performance on tail items across most evaluations, followed by the competitive LLM-ESR. The general superiority of LLM-enhanced baselines (e.g., LLMInit, SRA-CL) over pure ID-based backbones confirms that semantic priors effectively mitigate collaborative signal scarcity.

More importantly, ADAA achieves a more balanced improvement across tail and head groups in most scenarios. Rather than completely eliminating the common “seesaw effect” in long-tail recommendation, ADAA substantially alleviates it by improving tail-item accuracy while maintaining competitive head-item performance. This balance is attributed to the synergy of adversarial adaptation, which reduces popularity bias, and asymmetric anchoring, which mitigates the semantic-prior erosion of tail items during sequence training. We further verify in Appendix C.1 that ADAA’s tail-item gains persist under full-catalog ranking rather than arising from sampled candidates.

Performance on Long-Tail Users. Regarding tail users, ADAA consistently outperforms all baselines, including MELT and LLMInit. While MELT is specifically devised to enhance tail-user representations, it underperforms ADAA due to its limitations in relying solely on collaborative perspectives.

Table 1: Overall recommendation performance on three datasets. The best and second-best results are indicated in **bold** and underlined, respectively.

Dataset	Model	Overall		Tail Item		Head Item		Tail User		Head User	
		H@10	N@10	H@10	N@10	H@10	N@10	H@10	N@10	H@10	N@10
Yelp	SASRec	0.5724	0.3413	0.0763	0.0318	0.7184	0.4324	0.5782	0.3456	0.5501	0.3247
	- CITIES	0.5828	0.3540	0.1532	0.0700	0.7093	0.4376	0.5785	0.3511	0.5994	0.3649
	- MELT	0.6257	0.3791	0.1015	0.0371	0.7801	0.4799	0.6246	0.3804	0.6299	0.3744
	- RLMRec	0.5990	0.3623	0.0953	0.0412	0.7474	0.4568	0.5966	0.3613	0.6084	0.3658
	- LLMInit	0.6415	0.3997	0.1760	0.0789	0.7785	0.4941	0.6403	0.4010	0.6462	0.3948
	- SRA-CL	0.6534	0.4169	0.1312	0.0564	0.8072	<u>0.5231</u>	0.6518	0.4170	0.6597	<u>0.4166</u>
	- LLM-ESR	0.6673	0.4208	0.1893	0.0845	0.8080	0.5199	0.6685	0.4229	0.6627	0.4128
	- ADAA	0.6770	0.4279	0.1927	0.0840	0.8196	0.5297	0.6769	0.4275	0.6772	0.4296
	BERT4Rec	0.5812	0.3544	0.0173	0.0076	0.7473	0.4565	0.5840	0.3564	0.5707	0.3468
	- CITIES	0.5249	0.3015	0.0041	0.0014	0.6783	0.3899	0.5274	0.3032	0.5155	0.2954
	- MELT	0.6206	0.3770	0.0429	0.0149	0.7907	0.4836	0.6210	0.3780	0.6191	0.3733
	- RLMRec	0.5306	0.3039	0.0104	0.0040	0.6938	0.3922	0.5351	0.3065	0.5137	0.2936
	- LLMInit	0.6199	0.3781	0.0874	0.0330	0.7766	0.4797	0.6204	0.3796	0.6178	0.3723
	- SRA-CL	0.6001	0.3663	<u>0.1330</u>	<u>0.0655</u>	0.7389	0.4528	0.5950	0.3622	0.6217	0.3837
	- LLM-ESR	0.6623	0.4222	0.1227	0.0500	<u>0.8212</u>	<u>0.5318</u>	0.6637	0.4247	<u>0.6571</u>	<u>0.4127</u>
	- ADAA	0.6765	0.4409	0.2178	0.0934	0.8216	0.5433	0.6785	0.4437	0.6689	0.4305
Beauty	SASRec	0.4388	0.3030	0.0870	0.0649	0.5227	0.3598	0.4270	0.2941	0.4926	0.3438
	- CITIES	0.2256	0.1413	0.1363	0.0897	0.2468	0.1536	0.2215	0.1406	0.2441	0.1444
	- MELT	0.4334	0.2775	0.0460	0.0172	0.5258	0.3995	0.4233	0.2673	0.4796	0.3241
	- RLMRec	0.4460	0.3075	0.0924	0.0658	0.5303	0.3652	0.4365	0.3016	0.4892	0.3345
	- LLMInit	0.5455	0.3656	0.1714	0.0965	0.6347	0.4298	0.5359	0.3592	0.5893	0.3948
	- SRA-CL	0.5044	0.3568	0.0827	0.0427	0.6019	0.4294	0.4921	0.3446	0.5587	0.4105
	- LLM-ESR	0.5672	0.3713	0.2257	0.1108	0.6486	0.4334	0.5581	0.3643	0.6087	0.4032
	- ADAA	0.5770	0.4046	0.2698	0.1572	0.6503	0.4637	0.5640	0.3916	0.6364	0.4644
	BERT4Rec	0.4601	0.2935	0.0125	0.0051	0.5656	0.3616	0.4448	0.2812	0.5292	0.3491
	- CITIES	0.3961	0.2339	0.0023	0.0008	0.4900	0.2895	0.3832	0.2250	0.4551	0.2746
	- MELT	0.4716	0.2965	0.0709	0.0291	0.5671	0.3603	0.4596	0.2865	0.5263	0.3423
	- RLMRec	0.3977	0.2365	0.0090	0.0032	0.4903	0.2921	0.3853	0.2277	0.4539	0.2765
	- LLMInit	0.5029	0.3209	0.0927	0.0451	0.6007	0.3867	0.4919	0.3117	0.5530	0.3632
	- SRA-CL	0.4962	0.3196	0.0847	0.0454	0.5823	0.3719	0.4531	0.3015	0.5497	0.3667
	- LLM-ESR	0.5393	0.3590	0.1379	0.0745	0.6350	0.4269	<u>0.5295</u>	<u>0.3507</u>	<u>0.5839</u>	<u>0.3972</u>
	- ADAA	0.5498	0.3821	0.1556	0.0826	0.6428	0.4544	0.5439	0.3696	0.6124	0.4382
Fashion	SASRec	0.4956	0.4429	0.0454	0.0235	0.6748	0.6099	0.3967	0.3390	0.6239	0.5777
	- CITIES	0.4923	0.4423	0.0407	0.0214	0.6721	0.6098	0.3936	0.3392	0.6203	0.5760
	- MELT	0.4875	0.4150	0.0368	0.0144	0.6670	0.5745	0.3792	0.2933	0.6280	0.5729
	- RLMRec	0.4982	0.4457	0.0410	0.0223	0.6803	0.6143	0.3990	0.3415	0.6270	0.5808
	- LLMInit	0.5119	0.4492	0.0596	0.0305	0.6920	0.6159	0.4184	0.3501	0.6332	0.5777
	- SRA-CL	0.5262	0.4649	0.0589	0.0292	0.7097	0.6359	0.4263	0.3606	0.6557	0.6001
	- LLM-ESR	0.5619	0.4743	0.1095	0.0520	0.7420	0.6424	<u>0.4811</u>	0.3769	0.6668	0.6005
	- ADAA	0.6096	0.5240	0.2254	0.1342	0.7605	0.6771	0.5410	0.4444	0.6987	0.6272
	BERT4Rec	0.5139	0.4041	0.0347	0.0131	0.7020	0.5575	0.4251	0.2995	0.6289	0.5396
	- CITIES	0.4926	0.4090	0.0223	0.0099	0.6799	0.5679	0.3952	0.2975	0.6190	0.5535
	- MELT	0.4897	0.3810	0.0059	0.0019	0.6823	0.5319	0.3842	0.2514	0.6266	0.5491
	- RLMRec	0.4744	0.3567	0.0044	0.0015	0.6615	0.4981	0.3626	0.2268	0.6194	0.5251
	- LLMInit	0.4854	0.4035	0.0328	0.0161	0.6655	0.5577	0.3773	0.2846	0.6255	0.5578
	- SRA-CL	0.5260	0.4129	0.0374	0.0178	0.7222	0.5681	0.4355	0.2981	0.6495	0.5618
	- LLM-ESR	0.5487	0.4529	0.0525	0.0225	<u>0.7462</u>	<u>0.6243</u>	0.4629	0.3460	<u>0.6599</u>	<u>0.5916</u>
	- ADAA	0.5688	0.4918	0.1041	0.0468	0.7513	0.6665	0.4933	0.4020	0.6668	0.6083

Table 2: Ablation study of ADAA on the Yelp dataset. We report the detailed HR@10 and NDCG@10 metrics across different item and user groups. The best results are indicated in **bold**, and the second-best are underlined.

Architecture	Overall		Tail Item		Head Item		Tail User		Head User	
	H@10	N@10	H@10	N@10	H@10	N@10	H@10	N@10	H@10	N@10
w/o Adv	0.6246	0.3795	0.1317	0.0519	0.7698	0.4759	0.6268	0.3825	0.6164	0.3680
w/o FT	0.5070	0.3152	0.2992	0.1433	0.5387	0.3364	0.5129	0.3191	0.4845	0.3004
w/o Anc	0.6485	0.4003	0.1163	0.0473	0.8053	0.5042	0.6516	0.4024	0.6370	0.3922
Sym-Col	0.5892	0.3573	0.1096	0.0465	0.7305	0.4488	0.5847	0.3561	0.6066	0.3617
Sym-Sem	0.5924	0.3561	0.1499	0.0655	0.7127	0.4317	0.5944	0.3570	0.5845	0.3528
ADAA	0.6770	0.4279	<u>0.1927</u>	<u>0.0840</u>	0.8196	0.5297	0.6769	0.4275	0.6772	0.4296

4.3 Ablation Study

Table 2 details the ablation study of five ADAA variants. *w/o Adv* removes the adversarial mechanism, and its performance drop on tail items confirms that adversarial training is important for re-

ducing popularity bias. Beyond downstream performance, Appendix C.2 directly probes the learned representations and shows that adversarial debiasing substantially reduces recoverable popularity information.

w/o FT bypasses sequence training entirely. This variant achieves the highest tail-item $H@10/N@10$, indicating that directly preserving LLM-derived semantic priors can strongly benefit sparse items. However, it substantially sacrifices overall and head-item performance, showing that it represents a semantics-oriented extreme rather than a complete sequential recommender. This result reveals an overall–tail trade-off: maximizing semantic preservation alone can improve tail items, but sequence training remains necessary for capturing behavioral transition patterns and maintaining overall recommendation quality.

To analyze the anchoring mechanism, we evaluate three variants. *w/o Anc* removes the anchoring loss entirely, resulting in reduced tail-item accuracy and validating the need for structural regularization. Under the same sequence-training setting, full ADAA improves tail-item $H@10/N@10$ from 0.1163/0.0473 to 0.1927/0.0840 compared with *w/o Anc*, while also improving overall $H@10/N@10$ from 0.6485/0.4003 to 0.6770/0.4279. This shows that asymmetric anchoring mitigates semantic-prior erosion during necessary sequence adaptation, but does not claim to inherit the maximal tail accuracy of the untrained *w/o FT* variant.

The symmetric designs also prove suboptimal. *Sym-Col* forces tail items to align with unreliable sparse targets, while *Sym-Sem* prevents head items from absorbing dense collaborative signals. Their inferior performance indicates that the gain does not come from regularization alone, but from assigning frequency-specific preservation targets to head and tail items. These results show that asymmetric anchoring provides a better balance between preserving sparse-item semantics and learning collaborative transition patterns. Beyond removing individual components, we also test whether ADAA depends on the specific 20/80 head-tail cutoff. Appendix C.3 shows that ADAA consistently improves over SASRec across training-time head ratios from 10% to 40%, indicating that the asymmetric anchoring design is not tied to a single hard split.

We further investigate the robustness of ADAA, with the comprehensive hyper-parameter sensitivity analysis presented in Appendix C.4. The efficiency evaluation is detailed in Appendix C.5.

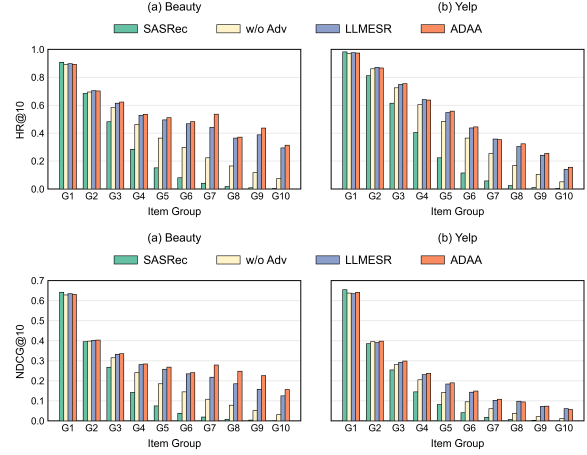


Figure 3: Performance across ten item popularity groups on Yelp and Beauty using SASRec. G1 and G10 contain the most and least popular items, respectively.

4.4 Group Analysis.

Figure 3 presents a fine-grained evaluation across ten item frequency groups on the Yelp and Beauty datasets. While the ID-based SASRec experiences performance degradation on sparse tail items, ADAA achieves competitive performance across item popularity groups and improves over LLMESR in many sparse-item groups. By decoupling interaction frequency via adversarial learning, ADAA generates debiased semantic representations that provide robust semantic initialization for tail items.

Furthermore, ADAA substantially alleviates the seesaw dilemma common in long-tail methods. It improves sparse-item groups while maintaining competitive performance on highly-interacted head items. This suggests that our asymmetric anchoring strategy helps preserve tail-item semantics from erosion while enabling head items to capture collaborative patterns.

5 Related Work

5.1 Sequential Recommendation and the Long-Tail Challenge

Sequential recommendation captures dynamic user preferences from interaction histories, typically parameterizing items as trainable ID embeddings (Kang and McAuley, 2018; Sun et al., 2019). However, real-world interactions follow a long-tail distribution. For tail items, the scarcity of collaborative signals restricts embedding optimization, leading to representation degradation (Xie et al., 2022; Qiu et al., 2022).

Conventional ID-based approaches address this

by optimizing the utilization of existing data through structural signal transfer or dynamic regularization. For example, CITIES (Jang et al., 2020) infers tail representations using co-occurrences with head items, while MELT (Kim et al., 2023) employs curriculum learning to refine tail users and items. Although effective, these intra-domain strategies operate within a closed collaborative loop bounded by available behavioral data. This limitation necessitates external prior knowledge, prompting research into semantic modalities to establish debiased semantic representations.

5.2 LLMs for Recommendation

To alleviate data sparsity, recent methods leverage Large Language Models to encode textual information (Peng et al., 2024; Shao et al., 2024). These approaches generally follow two paradigms: architectural integration and representation alignment. Architectural integration explicitly fuses LLM embeddings into the backbone. For instance, LLMInit (Hu et al., 2024; Harte et al., 2023) explores selective initialization, and LLM-ESR (Liu et al., 2024) combines semantic and collaborative features via a dual-branch network. While improving tail-item accuracy, these structural modifications introduce computational overhead during online serving.

Alternatively, representation alignment bridges semantic and collaborative spaces via auxiliary objectives. RLMRec (Ren et al., 2024) and SRA-CL (Cui et al., 2026) employ contrastive learning, while P5 (Geng et al., 2022) utilizes instruction tuning. However, standard alignment struggles with long-tail data. Because batch-sampled interactions are skewed toward frequent items, the learned mapping overfits dense collaborative signals. This popularity bias shifts tail-item representations toward popular clusters, degrading their initialized semantic features. To address this, our framework utilizes adversarial learning to decouple popularity factors from the semantic mapping, reducing popularity bias in knowledge injection while preserving the inference efficiency of standard sequential backbones.

6 Conclusion

In this paper, we present ADAA to address popularity-induced semantic distortion and semantic prior erosion in long-tail sequential recommendation. Our adversarial mechanism decouples interaction frequencies during cross-space align-

ment, yielding debiased semantic representations for both head and tail items. During sequence adaptation, the asymmetric semantic anchoring mechanism protects sparse tail items from being overwritten by head-dominated gradients while allowing head items to capture collaborative patterns. Experiments on three datasets demonstrate that ADAA achieves competitive performance across most metrics and substantially alleviates the seesaw dilemma. Finally, its architecturally non-intrusive design preserves the online inference efficiency of vanilla sequential backbones.

Limitations

We acknowledge three primary limitations of ADAA. First, our offline alignment pipeline ensures inference efficiency but requires frequent projector retraining in streaming scenarios with continuous new item arrivals. Second, the framework relies on raw textual metadata. For items lacking meaningful descriptions, the limited capability of language-model text embeddings to provide distinguishable semantic representations restricts the effectiveness of the adversarial debiasing mechanism. Third, our experiments are conducted on highly sparse but moderate-sized public datasets; thus, the efficiency results should be interpreted as evidence of unchanged online inference cost rather than large-scale scalability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62272338 and 62602463).

References

- Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*, pages 42–46.
- Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM conference on recommender systems*, pages 1007–1014.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

- Ziqiang Cui, Yunpeng Weng, Xing Tang, Xiaokun Zhang, Shiwei Li, Peiyang Liu, Bowei He, Dugang Liu, Weihong Luo, Xiuqiang He, and Ma Chen. 2026. Semantic retrieval augmented contrastive learning for sequential recommendation. *Advances in Neural Information Processing Systems*, 38:162621–162645.
- Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.
- Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM conference on recommender systems*, pages 299–315.
- Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.
- Jesse Harte, Wouter Zorgdrager, Panos Louridas, Asterios Katsifodimos, Dietmar Jannach, and Marios Fragkoulis. 2023. Leveraging large language models for sequential recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1096–1102.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Yupeng Hou, Binbin Hu, Zhiqiang Zhang, and Wayne Xin Zhao. 2022a. Core: simple and effective session-based recommendation within consistent representation space. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1796–1801.
- Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022b. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 585–593.
- Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European conference on information retrieval*, pages 364–381. Springer.
- Jun Hu, Wenwen Xia, Xiaolu Zhang, Chilin Fu, Weichang Wu, Zhaoxin Huan, Ang Li, Zuoli Tang, and Jun Zhou. 2024. Enhancing sequential recommendation via llm-based semantic embedding learning. In *Companion Proceedings of the ACM Web Conference 2024*, pages 103–111.
- Seongwon Jang, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2020. Cities: Contextual inference of tail-item embeddings for sequential recommendation. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 202–211. IEEE.
- Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.
- Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE.
- Kibum Kim, Dongmin Hyun, Sukwon Yun, and Chanyoung Park. 2023. Melt: Mutual enhancement of long-tailed user and item for sequential recommendation. In *Proceedings of the 46th international ACM SIGIR conference on Research and development in information retrieval*, pages 68–77.
- Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1748–1757.
- Qidong Liu, Xian Wu, Yejing Wang, Zijian Zhang, Feng Tian, Yefeng Zheng, and Xiangyu Zhao. 2024. Llm-esr: Large language models enhancement for long-tailed sequential recommendation. *Advances in Neural Information Processing Systems*, 37:26701–26727.
- Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.
- Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, and 6 others. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

- Mark EJ Newman. 2005. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351.
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.
- Qiyao Peng, Hongtao Liu, Hua Huang, Jian Yang, Qing Yang, and Minglai Shao. 2025. A survey on llm-powered agents for recommender systems. In *EMNLP (Findings)*, pages 11574–11583.
- Qiyao Peng, Hongyan Xu, Yinghui Wang, Hongtao Liu, Cuiying Huo, and Wenjun Wang. 2024. Pept: Expert finding meets personalized pre-training. *ACM Transactions on Information Systems*, 43(1):1–26.
- Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 813–823.
- Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM web conference 2024*, pages 3464–3475.
- Minglai Shao, Hua Huang, Qiyao Peng, and Hongtao Liu. 2024. Ulmrec: User-centric large language model for sequential recommendation. *arXiv preprint arXiv:2412.05543*.
- Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450.
- Jiayi Tang and Ke Wang. 2018a. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 565–573.
- Jiayi Tang and Ke Wang. 2018b. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2289–2298.
- Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176.
- Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z Sheng, and Mehmet Orgun. 2019. Sequential recommender systems: challenges, progress and prospects. *arXiv preprint arXiv:2001.04830*.
- Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1791–1800.
- Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.
- Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 346–353.
- Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 1259–1273. IEEE.
- Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the long tail recommendation. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, volume 5, pages 896–907. Association for Computing Machinery.
- Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 11–20.
- Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2021. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 4653–4664.
- Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the web conference 2021*, pages 2980–2991.
- Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Filter-enhanced mlp is all you need for sequential recommendation. In *Proceedings of the ACM web conference 2022*, pages 2388–2399.

A Appendix

A.1 LLM Prompt Templates

To extract universal semantic representations from textual metadata, we construct structured prompts to query the Large Language Models (LLMs). To ensure a strictly fair comparison and completely isolate the effectiveness of our proposed ADAA framework, we adopt the exact prompt templates established by the baseline LLM-ESR. Specifically, we organize the available item attributes (e.g., brand, category, title) into natural language sentences. Due to the diverse metadata fields recorded across different domains, the prompt templates are tailored for each dataset accordingly.

1. Prompt Template for Yelp:

ITEM PROMPT (YELP)

The point of interest has the following attributes:
name is <NAME>; category is <CATEGORY>; type is <TYPE>; open status is <OPEN>; review count is <COUNT>; city is <CITY>; average score is <STARS>.

2. Prompt Template for Amazon Beauty:

ITEM PROMPT (AMAZON BEAUTY)

The beauty item has the following attributes:
name is <TITLE>; brand is <BRAND>; price is <PRICE>.
The item has the following features: <CATEGORIES>.
The item has the following descriptions: <DESCRIPTION>.

3. Prompt Template for Amazon Fashion:

ITEM PROMPT (AMAZON FASHION)

The fashion item has the following attributes:
name is <TITLE>; brand is <BRAND>; date is <DATE>; price is <PRICE>.
The item has the following features: <FEATURE>.
The item has the following descriptions: <DESCRIPTION>.

A.2 Train and Inference Process

In this section, we provide the comprehensive execution pipeline of the proposed ADAA framework. Algorithm 1 delineates the step-by-step operations during both the decoupled training phase and the highly efficient inference phase.

B Experiment Setting

B.1 Datasets

We conduct experiments on three public real-world datasets: Yelp, Amazon Beauty, and Amazon Fashion. These datasets are selected to cover diverse application scenarios and evaluate the model’s performance under data sparsity.

- **Yelp¹**: A dataset widely used for business recommendation, containing user check-in records and reviews of local businesses. We utilize the interaction sequences along with the textual metadata (e.g., categories and business names) of the point-of-interests to extract semantic features.
- **Amazon²**: A collection of product review datasets sourced from the Amazon e-commerce platform. We adopt two sub-categories for our evaluation: **Beauty** and **Fashion**. Similar to Yelp, product titles, brand names, and category descriptions are used as textual metadata for LLM encoding.

B.2 Preprocessing

Following standard procedures in sequential recommendation, we convert all explicit ratings and reviews into implicit feedback, treating them as chronological interaction sequences based on timestamps. For data splitting, we adopt the leave-one-out evaluation protocol. Specifically, for each user’s interaction sequence, the most recent interaction is held out for testing, the second most recent interaction is used for validation, and all remaining historical interactions are utilized for model training. The detailed statistics of the preprocessed datasets are summarized in Table 3.

Table 3: Statistics of the preprocessed datasets.

Dataset	# Users	# Items	Avg. Length	Sparsity
Yelp	15720	11383	12.23	99.89%
Beauty	52204	57289	7.57	99.99%
Fashion	9049	4722	3.82	99.92%

B.3 Backbones and Baseline Methods

To evaluate the effectiveness and compatibility of ADAA, we instantiate our framework across two

¹<https://www.yelp.com/dataset>

²<http://jmcauley.ucsd.edu/data/amazon/>

Algorithm 1 Training and Inference of ADAA

Require: LLM embeddings $\mathbf{E}^{llm}$,
collaborative anchors $\mathbf{E}_H^{id}$, interactions $\mathcal{S}$,
sequential backbone f_θ .

Train Process

- 1: Initialize projector $P(\cdot)$ and discriminator $D(\cdot)$.
- 2: **for** epoch = 1 **to** E_{proj} **do**
- 3: Compute joint projection loss by Eq. (11).
- 4: Update $P(\cdot)$, $D(\cdot)$ via adversarial process.
- 5: **end for**
- 6: Extract debiased semantics: $\mathbf{Z}^* = P^*(\mathbf{E}^{llm})$.
- 7: Initialize item embeddings of f_θ with $\mathbf{Z}^*$.
- 8: **for** epoch = 1 **to** E_{tune} **do**
- 9: Compute training loss by Eq. (12).
- 10: Update backbone f_θ by minimizing $\mathcal{L}_{tune}$.
- 11: **end for**

Inference Process

- 12: **for** user sequence S_u **in** test set **do**
- 13: Encode user preference by Eq. (2).
- 14: Compute scores $\hat{y}_{u,v} \propto \mathbf{h}_u^\top \mathbf{e}_v$ for candidate items.
- 15: **end for**
- 16: **return** Top- N recommended items for all users.

representative sequential recommendation backbones and compare it against six competitive baselines. These methods are categorized into three distinct paradigms based on their structural designs:

Sequential Recommendation Backbones:

- **SASRec** (Kang and McAuley, 2018): A self-attention-based sequential recommendation model that employs a left-to-right Transformer architecture to capture directional user preference transitions from historical interaction histories.
- **BERT4Rec** (Sun et al., 2019): A bidirectional sequential framework that utilizes a masked language modeling objective to learn cloze-style contextual dependencies within user interaction sequences.

Traditional Long-Tail Enhancement Frameworks:

- **CITIES** (Jang et al., 2020): A collaborative-based framework designed for sparse recommendations, which infers tail-item representations by modeling contextual co-occurrences with frequent head items.

- **MELT** (Kim et al., 2023): A progressive curriculum learning framework developed for long-tail sequential recommendation, which mutually refines and regularizes the representations of tail users and items.

LLM-Enhanced Recommendation Frameworks:

- **RLMRec** (Ren et al., 2024): A framework that utilizes Large Language Models to generate textual user and item profiles, incorporating the derived semantic embeddings to regularize collaborative representation learning.
- **LLMInit** (Hu et al., 2024; Harte et al., 2023): A semantic initialization strategy that leverages pre-trained LLM embedding vectors to directly overwrite the initial embedding parameters of the underlying sequential backbone.
- **LLM-ESR** (Liu et al., 2024): A dual-view modeling architecture that integrates collaborative signals and text semantics through a retrieval-augmented self-distillation mechanism to preserve tail-item properties.
- **SRA-CL** (Cui et al., 2026): A semantic retrieval-augmented contrastive learning paradigm that uses LLM textual features to construct semantic indices, optimizing sequence representations via refined contrastive sample pairs.

B.4 Implementation Details

All experiments are implemented in PyTorch and conducted on an NVIDIA GPU platform. To ensure a fair comparison, the core experimental configurations are aligned with the state-of-the-art baseline LLM-ESR. Specifically, the item embedding dimension is set to 64, and the training batch size is configured to 128. All models are optimized using the Adam optimizer with a learning rate of 0.001 and a dropout rate of 0.5. For the framework design of ADAA, the semantic projector is instantiated as a two-layer MLP incorporating a GELU activation function. For hyper-parameter selection, the adversarial coefficient λ and the initial anchoring weight α_0 are grid-searched within $\{1e-3, 1e-2, 5e-2, 0.1, 0.5\}$ and $\{5, 10, 15, 20, 25\}$, respectively.

C Supplementary Experiments

C.1 Full-Catalog Ranking Evaluation

The main experiments follow the commonly used sampled evaluation protocol, where each ground-truth item is ranked against 100 randomly sampled non-interacted items. To address the concern that sampled evaluation may be sensitive to the sampled candidates, we further conduct a full-catalog ranking evaluation. For each test instance, we rank the ground-truth item against all eligible unseen items. The candidate set is kept identical when computing Overall, Head-item, and Tail-item metrics. The group-specific metrics are obtained only by dividing test instances according to whether the ground-truth item belongs to the head or tail group; the ranking candidates themselves are not restricted to either group. Thus, each tail target competes directly with the full set of eligible head and tail items.

Table 4 reports the full-ranking results over five matched random seeds. We report mean $\pm$ standard deviation and conduct paired significance tests over matched runs. As expected, the absolute metric values under full-catalog ranking are lower than those under sampled evaluation, because the ranking task becomes substantially harder when the candidate set includes all eligible unseen items. Nevertheless, ADAA consistently achieves the best performance across the reported item and user groups. In particular, on the Yelp-SASRec setting, where the sampled-protocol tail results of ADAA and LLM-ESR are closest, ADAA improves the tail-item H/N from .0050/.0021 achieved by the strongest baseline LLM-ESR to .0062/.0029, and the improvement is statistically significant. ADAA also improves the overall H/N from .0482/.0241 to .0491/.0248.

These results show that ADAA’s tail-item improvement is preserved under the stricter full-catalog protocol and is not an artifact of sampled negative candidates. The gains also remain balanced across head and tail groups, supporting our claim that adversarial debiasing and asymmetric anchoring improve tail recommendation while substantially alleviating the seesaw effect of sacrificing head-item or user-group performance.

C.2 Popularity Probing Analysis

The recommendation ablation in Section 4.3 shows that removing the adversarial objective hurts downstream tail-item performance, but it does not di-

rectly establish whether the projected embeddings contain less popularity information. To provide direct evidence, we conduct an independent post-hoc probing analysis.

We freeze each representation and train external probes to predict item popularity. Specifically, we train a linear classifier and a two-layer MLP to predict head/tail membership on balanced and disjoint item splits. These probes do not share parameters with the discriminator used to train ADAA. We also train a regression probe to predict log-transformed interaction frequency, reporting its R^2 and the Spearman correlation between predicted and true popularity. All variants use identical item splits and probe settings, and results are reported as mean $\pm$ standard deviation over five random splits.

Table 5 reports the amount of popularity information recoverable from different item representations. The results confirm that popularity is readily recoverable from collaborative embeddings, which achieve a linear AUC of .815 and an MLP AUC of .864. The variant without adversarial debiasing also retains substantial popularity information from collaborative anchors, with a linear AUC of .718, an MLP AUC of .763, a popularity-regression R^2 of .275, and a Spearman correlation of .492.

After applying adversarial debiasing, recoverable popularity information is substantially reduced. ADAA decreases linear AUC from .718 to .535 and MLP AUC from .763 to .571. The continuous popularity probes show the same pattern: R^2 decreases from .275 to .039, and Spearman’s ρ decreases from .492 to .098. These reductions are significant under paired tests over matched random splits. ADAA therefore brings the projected representations close to the low-popularity-information semantic baseline.

We note that the nonlinear MLP probe still recovers a small amount of popularity information, indicating that the representations are not completely popularity-free. These results indicate that ADAA substantially reduces recoverable popularity information, while not implying complete semantic-popularity disentanglement. This direct probing evidence complements the downstream ablation, where removing the adversarial objective reduces tail-item H/N from .1927/.0840 to .1317/.0519.

C.3 Sensitivity to the Head/Tail Training Cutoff

ADAA uses the head/tail partition to decide whether an item is anchored toward its collabo-

Table 4: Full-ranking results on Yelp, Beauty, and Fashion. We report mean $\pm$ standard deviation and conducted paired significance tests over matched runs. H/N denotes H@10/N@10. $\dagger$ indicates $p < 0.05$ versus the best baseline under a paired test.

Dataset	Model	Overall	Tail Item	Head Item	Tail User	Head User
		H/N	H/N	H/N	H/N	H/N
Yelp	SASRec	.0254 $\pm$.0018/.0118 $\pm$.0009	.0003 $\pm$.0001/.0001 $\pm$.0001	.0496 $\pm$.0027/.0233 $\pm$.0014	.0252 $\pm$.0015/.0121 $\pm$.0007	.0260 $\pm$.0024/.0105 $\pm$.0012
	- CITIES	.0268 $\pm$.0022/.0125 $\pm$.0011	.0005 $\pm$.0002/.0002 $\pm$.0001	.0485 $\pm$.0025/.0227 $\pm$.0012	.0266 $\pm$.0018/.0126 $\pm$.0009	.0274 $\pm$.0022/.0123 $\pm$.0013
	- MELT	.0281 $\pm$.0017/.0134 $\pm$.0008	.0008 $\pm$.0003/.0004 $\pm$.0002	.0502 $\pm$.0029/.0238 $\pm$.0015	.0279 $\pm$.0016/.0132 $\pm$.0008	.0289 $\pm$.0026/.0141 $\pm$.0015
	- RLMRec	.0412 $\pm$.0015/.0205 $\pm$.0007	.0027 $\pm$.0006/.0013 $\pm$.0003	.0605 $\pm$.0021/.0297 $\pm$.0011	.0411 $\pm$.0017/.0207 $\pm$.0008	.0417 $\pm$.0021/.0196 $\pm$.0012
	- LLMInit	.0387 $\pm$.0021/.0189 $\pm$.0010	.0021 $\pm$.0005/.0010 $\pm$.0003	.0583 $\pm$.0024/.0284 $\pm$.0013	.0386 $\pm$.0019/.0191 $\pm$.0009	.0392 $\pm$.0023/.0182 $\pm$.0014
	- SRA-CL	.0435 $\pm$.0019/.0218 $\pm$.0012	.0034 $\pm$.0004/.0016 $\pm$.0002	.0631 $\pm$.0026/.0308 $\pm$.0015	.0434 $\pm$.0020/.0221 $\pm$.0011	.0438 $\pm$.0025/.0208 $\pm$.0014
	- LLM-ESR	.0482 $\pm$.0016/.0241 $\pm$.0009	.0050 $\pm$.0005/.0021 $\pm$.0003	.0643 $\pm$.0021/.0319 $\pm$.0012	.0481 $\pm$.0021/.0243 $\pm$.0010	.0485 $\pm$.0028/.0233 $\pm$.0015
	- ADAA	.0491 $\pm$.0014/.0248 $\pm$.0008†	.0062 $\pm$.0005/.0029 $\pm$.0002†	.0694 $\pm$.0022/.0348 $\pm$.0010†	.0489 $\pm$.0017/.0249 $\pm$.0009†	.0499 $\pm$.0022/.0242 $\pm$.0014†
	BERT4Rec	.0278 $\pm$.0019/.0129 $\pm$.0010	.0004 $\pm$.0001/.0002 $\pm$.0001	.0525 $\pm$.0028/.0252 $\pm$.0015	.0289 $\pm$.0017/.0133 $\pm$.0009	.0235 $\pm$.0025/.0115 $\pm$.0013
	- CITIES	.0290 $\pm$.0023/.0136 $\pm$.0012	.0006 $\pm$.0002/.0003 $\pm$.0001	.0512 $\pm$.0026/.0245 $\pm$.0013	.0298 $\pm$.0019/.0137 $\pm$.0010	.0260 $\pm$.0023/.0132 $\pm$.0014
	- MELT	.0305 $\pm$.0018/.0145 $\pm$.0009	.0009 $\pm$.0003/.0005 $\pm$.0002	.0531 $\pm$.0030/.0256 $\pm$.0016	.0312 $\pm$.0016/.0144 $\pm$.0008	.0278 $\pm$.0027/.0151 $\pm$.0016
	- LLMInit	.0415 $\pm$.0022/.0204 $\pm$.0011	.0024 $\pm$.0005/.0012 $\pm$.0003	.0615 $\pm$.0025/.0302 $\pm$.0014	.0418 $\pm$.0021/.0206 $\pm$.0010	.0402 $\pm$.0024/.0198 $\pm$.0015
	- RLMRec	.0438 $\pm$.0016/.0218 $\pm$.0008	.0030 $\pm$.0006/.0015 $\pm$.0003	.0632 $\pm$.0022/.0312 $\pm$.0012	.0440 $\pm$.0015/.0220 $\pm$.0007	.0430 $\pm$.0022/.0210 $\pm$.0013
	- SRA-CL	.0452 $\pm$.0020/.0229 $\pm$.0013	.0035 $\pm$.0004/.0017 $\pm$.0002	.0648 $\pm$.0027/.0320 $\pm$.0016	.0455 $\pm$.0019/.0232 $\pm$.0012	.0442 $\pm$.0026/.0218 $\pm$.0015
Beauty	- LLM-ESR	.0465 $\pm$.0017/.0240 $\pm$.0010	.0045 $\pm$.0005/.0021 $\pm$.0003	.0662 $\pm$.0022/.0331 $\pm$.0013	.0468 $\pm$.0014/.0242 $\pm$.0009	.0455 $\pm$.0029/.0232 $\pm$.0016
	- ADAA	.0482 $\pm$.0015/.0249 $\pm$.0008†	.0055 $\pm$.0005/.0026 $\pm$.0002†	.0685 $\pm$.0023/.0345 $\pm$.0011†	.0485 $\pm$.0013/.0251 $\pm$.0008†	.0470 $\pm$.0023/.0243 $\pm$.0015†
	BERT4Rec	.0278 $\pm$.0019/.0129 $\pm$.0010	.0004 $\pm$.0001/.0002 $\pm$.0001	.0525 $\pm$.0028/.0252 $\pm$.0015	.0289 $\pm$.0017/.0133 $\pm$.0009	.0235 $\pm$.0025/.0115 $\pm$.0013
	- CITIES	.0290 $\pm$.0023/.0136 $\pm$.0012	.0006 $\pm$.0002/.0003 $\pm$.0001	.0512 $\pm$.0026/.0245 $\pm$.0013	.0298 $\pm$.0019/.0137 $\pm$.0010	.0260 $\pm$.0023/.0132 $\pm$.0014
	- MELT	.0305 $\pm$.0018/.0145 $\pm$.0009	.0009 $\pm$.0003/.0005 $\pm$.0002	.0531 $\pm$.0030/.0256 $\pm$.0016	.0312 $\pm$.0016/.0144 $\pm$.0008	.0278 $\pm$.0027/.0151 $\pm$.0016
	- LLMInit	.0415 $\pm$.0022/.0204 $\pm$.0011	.0024 $\pm$.0005/.0012 $\pm$.0003	.0615 $\pm$.0025/.0302 $\pm$.0014	.0418 $\pm$.0021/.0206 $\pm$.0010	.0402 $\pm$.0024/.0198 $\pm$.0015
	- RLMRec	.0438 $\pm$.0016/.0218 $\pm$.0008	.0030 $\pm$.0006/.0015 $\pm$.0003	.0632 $\pm$.0022/.0312 $\pm$.0012	.0440 $\pm$.0015/.0220 $\pm$.0007	.0430 $\pm$.0022/.0210 $\pm$.0013
	- SRA-CL	.0452 $\pm$.0020/.0229 $\pm$.0013	.0035 $\pm$.0004/.0017 $\pm$.0002	.0648 $\pm$.0027/.0320 $\pm$.0016	.0455 $\pm$.0019/.0232 $\pm$.0012	.0442 $\pm$.0026/.0218 $\pm$.0015
	- LLM-ESR	.0465 $\pm$.0017/.0240 $\pm$.0010	.0045 $\pm$.0005/.0021 $\pm$.0003	.0662 $\pm$.0022/.0331 $\pm$.0013	.0468 $\pm$.0014/.0242 $\pm$.0009	.0455 $\pm$.0029/.0232 $\pm$.0016
	- ADAA	.0482 $\pm$.0015/.0249 $\pm$.0008†	.0055 $\pm$.0005/.0026 $\pm$.0002†	.0685 $\pm$.0023/.0345 $\pm$.0011†	.0485 $\pm$.0013/.0251 $\pm$.0008†	.0470 $\pm$.0023/.0243 $\pm$.0015†
	BERT4Rec	.0278 $\pm$.0019/.0129 $\pm$.0010	.0004 $\pm$.0001/.0002 $\pm$.0001	.0525 $\pm$.0028/.0252 $\pm$.0015	.0289 $\pm$.0017/.0133 $\pm$.0009	.0235 $\pm$.0025/.0115 $\pm$.0013
	- CITIES	.0290 $\pm$.0023/.0136 $\pm$.0012	.0006 $\pm$.0002/.0003 $\pm$.0001	.0512 $\pm$.0026/.0245 $\pm$.0013	.0298 $\pm$.0019/.0137 $\pm$.0010	.0260 $\pm$.0023/.0132 $\pm$.0014
	- MELT	.0305 $\pm$.0018/.0145 $\pm$.0009	.0009 $\pm$.0003/.0005 $\pm$.0002	.0531 $\pm$.0030/.0256 $\pm$.0016	.0312 $\pm$.0016/.0144 $\pm$.0008	.0278 $\pm$.0027/.0151 $\pm$.0016
	- LLMInit	.0415 $\pm$.0022/.0204 $\pm$.0011	.0024 $\pm$.0005/.0012 $\pm$.0003	.0615 $\pm$.0025/.0302 $\pm$.0014	.0418 $\pm$.0021/.0206 $\pm$.0010	.0402 $\pm$.0024/.0198 $\pm$.0015
Fashion	- RLMRec	.0438 $\pm$.0016/.0218 $\pm$.0008	.0030 $\pm$.0006/.0015 $\pm$.0003	.0632 $\pm$.0022/.0312 $\pm$.0012	.0440 $\pm$.0015/.0220 $\pm$.0007	.0430 $\pm$.0022/.0210 $\pm$.0013
	- SRA-CL	.0452 $\pm$.0020/.0229 $\pm$.0013	.0035 $\pm$.0004/.0017 $\pm$.0002	.0648 $\pm$.0027/.0320 $\pm$.0016	.0455 $\pm$.0019/.0232 $\pm$.0012	.0442 $\pm$.0026/.0218 $\pm$.0015
	- LLM-ESR	.0465 $\pm$.0017/.0240 $\pm$.0010	.0045 $\pm$.0005/.0021 $\pm$.0003	.0662 $\pm$.0022/.0331 $\pm$.0013	.0468 $\pm$.0014/.0242 $\pm$.0009	.0455 $\pm$.0029/.0232 $\pm$.0016
	- ADAA	.0482 $\pm$.0015/.0249 $\pm$.0008†	.0055 $\pm$.0005/.0026 $\pm$.0002†	.0685 $\pm$.0023/.0345 $\pm$.0011†	.0485 $\pm$.0013/.0251 $\pm$.0008†	.0470 $\pm$.0023/.0243 $\pm$.0015†
	BERT4Rec	.0278 $\pm$.0019/.0129 $\pm$.0010	.0004 $\pm$.0001/.0002 $\pm$.0001	.0525 $\pm$.0028/.0252 $\pm$.0015	.0289 $\pm$.0017/.0133 $\pm$.0009	.0235 $\pm$.0025/.0115 $\pm$.0013
	- CITIES	.0290 $\pm$.0023/.0136 $\pm$.0012	.0006 $\pm$.0002/.0003 $\pm$.0001	.0512 $\pm$.0026/.0245 $\pm$.0013	.0298 $\pm$.0019/.0137 $\pm$.0010	.0260 $\pm$.0023/.0132 $\pm$.0014
	- MELT	.0305 $\pm$.0018/.0145 $\pm$.0009	.0009 $\pm$.0003/.0005 $\pm$.0002	.0531 $\pm$.0030/.0256 $\pm$.0016	.0312 $\pm$.0016/.0144 $\pm$.0008	.0278 $\pm$.0027/.0151 $\pm$.0016
	- LLMInit	.0415 $\pm$.0022/.0204 $\pm$.0011	.0024 $\pm$.0005/.0012 $\pm$.0003	.0615 $\pm$.0025/.0302 $\pm$.0014	.0418 $\pm$.0021/.0206 $\pm$.0010	.0402 $\pm$.0024/.0198 $\pm$.0015
	- RLMRec	.0438 $\pm$.0016/.0218 $\pm$.0008	.0030 $\pm$.0006/.0015 $\pm$.0003	.0632 $\pm$.0022/.0312 $\pm$.0012	.0440 $\pm$.0015/.0220 $\pm$.0007	.0430 $\pm$.0022/.0210 $\pm$.0013
	- SRA-CL	.0452 $\pm$.0020/.0229 $\pm$.0013	.0035 $\pm$.0004/.0017 $\pm$.0002	.0648 $\pm$.0027/.0320 $\pm$.0016	.0455 $\pm$.0019/.0232 $\pm$.0012	.0442 $\pm$.0026/.0218 $\pm$.0015
	- LLM-ESR	.0465 $\pm$.0017/.0240 $\pm$.0010	.0045 $\pm$.0005/.0021 $\pm$.0003	.0662 $\pm$.0022/.0331 $\pm$.0013	.0468 $\pm$.0014/.0242 $\pm$.0009	.0455 $\pm$.0029/.0232 $\pm$.0016
	- ADAA	.0482 $\pm$.0015/.0249 $\pm$.0008†	.0055 $\pm$.0005/.0026 $\pm$.0002†	.0685 $\pm$.0023/.0345 $\pm$.0011†	.0485 $\pm$.0013/.0251 $\pm$.0008†	.0470 $\pm$.0023/.0243 $\pm$.0015†
	BERT4Rec	.0278 $\pm$.0019/.0129 $\pm$.0010	.0004 $\pm$.0001/.0002 $\pm$.0001	.0525 $\pm$.0028/.0252 $\pm$.0015	.0289 $\pm$.0017/.0133 $\pm$.0009	.0235 $\pm$.0025/.0115 $\pm$.0013
	- CITIES	.0290 $\pm$.0023/.0136 $\pm$.0012	.0006 $\pm$.0002/.0003 $\pm$.0001	.0512 $\pm$.0026/.0245 $\pm$.0013	.0298 $\pm$.0019/.0137 $\pm$.0010	.0260 $\pm$.0023/.0132 $\pm$.0014
	- MELT	.0305 $\pm$.0018/.0145 $\pm$.0009	.0009 $\pm$.0003/.0005 $\pm$.0002	.0531 $\pm$.0030/.0256 $\pm$.0016	.0312 $\pm$.0016/.0144 $\pm$.0008	.0278 $\pm$.0027/.0151 $\pm$.0016

Table 5: Popularity information recoverable from different item representations.

Representation	Linear AUC $\downarrow$	Linear BAcc $\downarrow$	MLP AUC $\downarrow$	MLP BAcc $\downarrow$	Pop. $R^2 \downarrow$	Spearman $\rho \downarrow$
Semantic	.512 $\pm$.011	.508 $\pm$.009	.534 $\pm$.014	.521 $\pm$.012	.018 $\pm$.004	.045 $\pm$.011
Collaborative	.815 $\pm$.014	.742 $\pm$.016	.864 $\pm$.011	.785 $\pm$.013	.442 $\pm$.021	.668 $\pm$.015
w/o Adv	.718 $\pm$.015	.654 $\pm$.018	.763 $\pm$.012	.692 $\pm$.014	.275 $\pm$.019	.492 $\pm$.016
ADAA	.535 $\pm$.012†	.522 $\pm$.011†	.571 $\pm$.015†	.548 $\pm$.013†	.039 $\pm$.006†	.098 $\pm$.014†

Notes. BAcc denotes balanced accuracy; lower values indicate less recoverable popularity information. $\dagger$ indicates a significant reduction from w/o Adv under a paired test ($p < 0.05$).

relative ID representation or its debiased semantic representation during sequence training. To examine whether the method depends on a specific 20/80 cutoff, we vary the training-time head ratio over $p \in \{10, 15, 20, 25, 30, 40\}\%$. For each value of p , the top- $p\%$ most popular items receive collaborative anchoring, while the remaining items receive semantic anchoring. To isolate the effect of the training threshold, we keep the evaluation groups fixed at the original 20/80 split. All settings use the same Yelp data split, SASRec backbone, hyperpa-

rameters, and full-catalog evaluation protocol.

Table 6 shows that ADAA outperforms SASRec on overall, head-item, and tail-item metrics across all tested values of p . Overall and head-item performance remain relatively stable: overall H@10 ranges from .0482 to .0496, and N@10 from .0242 to .0248. Tail-item performance generally decreases as p increases, with identical reported values at $p = 20\%$ and $p = 25\%$.

These results indicate that ADAA’s gains are not confined to the original 20/80 training split. They also show that the cutoff affects the balance between head- and tail-item performance. The overall decline in tail performance is consistent with assigning fewer items to semantic anchoring as p increases. Across the tested range, ADAA improves tail-item recommendation while maintaining head-item performance above the backbone.

Table 6: Sensitivity to the training-time head ratio on Yelp-SASRec under full-catalog ranking.

Model	Overall H/N	Head H/N	Tail H/N
SASRec	.0254/.0118	.0496/.0233	.0003/.0001
ADAA ($p = 10\%$)	.0482/.0242	.0671/.0336	.0068/.0031
ADAA ($p = 15\%$)	.0489/.0246	.0683/.0344	.0065/.0030
ADAA ($p = 20\%$)	.0491/.0248	.0694/.0348	.0062/.0029
ADAA ($p = 25\%$)	.0495/.0247	.0688/.0346	.0062/.0029
ADAA ($p = 30\%$)	.0496/.0248	.0693/.0347	.0051/.0024
ADAA ($p = 40\%$)	.0491/.0244	.0691/.0345	.0042/.0019

Notes. H/N denotes H@10/N@10. The evaluation head/tail groups remain fixed at the original 20/80 split for all rows.

C.4 Hyper-parameter Analysis

Figure 4 presents the sensitivity analyses conducted on two hyper-parameters, the adversarial coefficient λ and the initial anchoring weight α_0 , to evaluate the robustness of ADAA across the evaluated datasets.

The parameter λ controls the adversarial strength during the semantic alignment phase. As observed, the recommendation performance is limited when λ is relatively small. This implies that a weak adversarial penalty is insufficient to decouple the popularity bias, causing tail representations to drift toward dense collaborative regions. Conversely, increasing λ beyond an optimal threshold leads to a noticeable performance drop. A disproportionately high adversarial weight dominates the joint optimization objective, resulting in mapping collapse where the projector fails to establish valid cross-space mapping rules.

The parameter α_0 determines the initial strength of the asymmetric anchoring during the sequence adaptation phase. A small α_0 yields sub-optimal outcomes, indicating that insufficient structural constraints fail to prevent sparse items from representation degradation when updated alongside head items. However, performance also declines when α_0 is set too high. Overly strict anchoring constraints restrict the item embeddings to the static semantic priors, thereby limiting the model’s flexibility to fit dynamic user interactions and personalized sequential patterns during training.

C.5 Efficiency Analysis

We evaluate the inference efficiency of ADAA from both theoretical and empirical perspectives. Let n , d , and L denote the sequence length, hidden dimension, and number of backbone layers, respectively. The theoretical time complexity of vanilla SASRec is $\mathcal{O}(L(n^2d + nd^2))$. Through parameter-

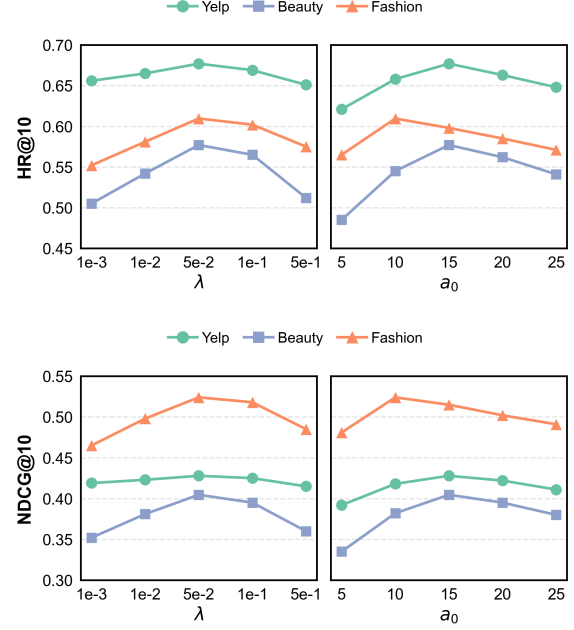


Figure 4: Sensitivity analysis of the adversarial coefficient λ and the initial anchoring weight α_0 on three datasets with the SASRec model.

level injection, ADAA maintains this exact theoretical bound. Similarly, SRA-CL matches this complexity by bypassing its retrieval module during inference. Conversely, LLM-ESR employs a dual-view architecture with auxiliary adapters and cross-attention, increasing its inference complexity to approximately $\mathcal{O}(2L(n^2d + nd^2) + n^2d)$ and introducing significant computational overhead.

Table 7: Efficiency comparison during the online inference phase. Metrics include model parameters (Params), Batch Latency (BL), Throughput (TP), and Peak GPU Memory (Mem).

Method	Params (M)	BL (ms)	TP (QPS)	Mem (GB)
SASRec	0.78	2.39	53568.5	0.08
SRA-CL	0.87	2.39	53298.7	0.08
LLM-ESR	20.31	13.04	9795.7	0.33
ADAA	0.78	2.35	54591.4	0.08

Empirically, the online serving metrics in Table 7, including model parameters (Params), Batch Latency (BL), Throughput (TP), and Peak GPU Memory (Mem), align with our theoretical derivations. ADAA preserves the exact parameter scale of the vanilla backbone, achieving equivalent BL and TP, whereas LLM-ESR incurs significantly higher latency and parameter costs. Although SRA-CL demonstrates comparable efficiency to ADAA, its contrastive learning paradigm alters the backbone’s

original training objective, potentially limiting its compatibility with diverse generative sequence architectures and overall recommendation accuracy. By decoupling semantic alignment from sequence encoding, ADAA provides a flexible integration strategy that preserves both the architectural integrity and computational efficiency of the original backbone.