SOVEREIGN: What is the accuracy impact (Recall@k, NDCG@k) of replacing pretrained multimodal encoders (BERT/ViT) with lig

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Streaming recommender systems (SRSs) are widely deployed in real-world applications, where user interests shift and new items arrive over time. As a result, effectively capturing users' latest preferences is challenging, as interactions reflecting recent interests are limited and new items often lack sufficient feedback. A common solution is to enrich item representations using multimodal encoders (e.g., BERT or ViT) to extract visual and textual features. However, these encoders are pretrained on general-purpose tasks: they are not tailored to user preference modeling, and they overlook the f

1 Introduction

Analysis of: Efficient Multimodal Streaming Recommendation via Expandable Side Mixture-of-Experts. Research goal: What is the accuracy impact (Recall@k, NDCG@k) of replacing pretrained multimodal encoders (BERT/ViT) with lightweight, MoE-adapted encoders in streaming recommendation tasks when new item categories are introduced over time?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

11 papers retrieved. 14 claims extracted, 0 verified. Tribunal: 1.3/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
The XSMoE method uses a side-tuning network with M layers per modality.	×	0.05
The model is warmed up on dataset D0 before processing subsequent data streams.	×	0.02
The model updates on each new dataset Ds for T time steps.	×	0.06
Performance of the model is tested using HR@10 and NDCG@10 metrics.	×	0.08
Utilization scores for experts are computed at each layer to determine which experts to prune.	×	0.05
Each expert contains two projection layers with 2dd' parameters, where d is the input size and d' is the down-projection	×	0.04
The router in each layer has dN + d parameters where d is input size and N is the number of experts.	×	0.05
	×	0.00
At any training stage, only one expert and the router remain trainable per layer.	×	0.03
The per-epoch training time complexity is dominated by forward passes, backward passes, and weight updates.	×	0.02
Per-epoch training time complexity of XSMoE is O(MNdd').	×	0.03
GPU memory usage is dominated by model weights, gradients, optimizer states, and activations.	×	0.03
Existing streaming recommendation systems (SRSs) rely solely on ID-based user-item interactions.	×	0.11
Multimodal recommenders use pretrained encoders such as ViT for visual content and BERT for text.	×	0.12

References

- http://arxiv.org/abs/2407.15411v4
- http://arxiv.org/abs/2002.05063v1
- http://arxiv.org/abs/2508.05993v3