SOVEREIGN: What is the impact of MoE architecture on inference efficiency and accuracy for multimodal reasoning tasks

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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 impact of MoE architecture on inference efficiency and accuracy for multimodal reasoning tasks.

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

3 papers retrieved. 13 claims extracted, 12 verified. Tribunal: 8.2/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

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
Streaming recommender systems are widely deployed in real-world applications.	✓	0.20
In streaming recommender systems, user interests shift and new items arrive over time.	✓	0.28
A common solution to capture users' latest preferences is to enrich item representations using multimodal encoders such	✓	0.25
Multimodal encoders like BERT and ViT are pretrained on general-purpose tasks and are not tailored to user preference mo	✓	0.25
User tastes toward modality-specific features such as visual styles and textual tones can drift over time.	✓	0.29
Fine-tuning large multimodal encoders in streaming scenarios has a high cost.	✓	0.24
Continuous model updates in streaming scenarios risk forgetting long-term user preferences.	✓	0.24
Expandable Side Mixture-of-Experts (XSMoE) is a memory-efficient framework for multimodal streaming recommendation.	✓	0.33
XSMoE attaches lightweight side-tuning modules consisting of expandable expert networks to frozen pretrained encoders.	✓	0.29
XSMoE incrementally expands expert networks in response to evolving user feedback.	✓	0.20
A gating router in XSMoE dynamically combines expert and backbone outputs.	✓	0.18
XSMoE uses a utilization-based pruning strategy to maintain model compactness.	×	0.14
XSMoE learns new patterns through expandable experts without overwriting previously acquired knowledge.	✓	0.19

References

- http://arxiv.org/abs/2603.11114v1
- http://arxiv.org/abs/2508.05993v3
- http://arxiv.org/abs/2504.16021v1