

# SOVEREIGN: Does predictive expert caching in SMOE multimodal models improve robustness to cross-modal distribution shifts

SOVEREIGN Research Kernel

Autonomous draft — Owner review required before publication

May 28, 2026

## Abstract

Graph data are inherently complex and heterogeneous, leading to a high natural diversity of distributional shifts. However, it remains unclear how to build machine learning architectures that generalize to the complex distributional shifts naturally occurring in the real world. Here, we develop GraphMETRO, a Graph Neural Network architecture that models natural diversity and captures complex distributional shifts. GraphMETRO employs a Mixture-of-Experts (MoE) architecture with a gating model and multiple expert models, where each expert model targets a specific distributional shift to produce

## 1 Introduction

Analysis of: GraphMETRO: Mitigating Complex Graph Distribution Shifts via Mixture of Aligned Experts. Research goal: Does predictive expert caching in SMOE multimodal models improve robustness to cross-modal distribution shifts on MMMU, measured by accuracy drop and throughput consistency compared to dense baselines?.

## 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

10 papers retrieved. 6 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
GraphMETRO consistently outperforms baseline models across all datasets.	×	0.06
GraphMETRO achieves improvements of 67.0% and 4.2% relative to EERM on the WebKB and Twitch datasets, respectively.	×	0.11
WebKB is a 5-class prediction task that predicts the classes of university webpages.	×	0.03
Twitch is a binary classification task that predicts whether a user streams mature content.	×	0.03
GraphMETRO uses an individual GNN encoder for each expert in the experiments.	×	0.05
ROC-AUC is used as the evaluation metric on Twitch, and classification accuracy is used on the other datasets.	×	0.03

### References

- <http://arxiv.org/abs/2312.04693v3>
- <http://arxiv.org/abs/2207.05796v1>
- <http://arxiv.org/abs/2603.11114v1>