SOVEREIGN: How does SMoES soft modality-guided routing scale in terms of accuracy-efficiency trade-offs (FLOPs per sample

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Recent advances in vision-language pre-training (VLP) have demonstrated impressive performance in a range of vision-language (VL) tasks. However, there exist several challenges for measuring the community's progress in building general multi-modal intelligence. First, most of the downstream VL datasets are annotated using raw images that are already seen during pre-training, which may result in an overestimation of current VLP models' generalization ability. Second, recent VLP work mainly focuses on absolute performance but overlooks the efficiency-performance trade-off, which is also an impor

1 Introduction

Analysis of: VLUE: A Multi-Task Benchmark for Evaluating Vision-Language Models. Research goal: How does SMoES soft modality-guided routing scale in terms of accuracy-efficiency trade-offs (FLOPs per sample, latency) relative to dense VLMs and standard MoE routing on the ChartQA benchmark when evaluated on unseen chart types, and does this advantage persist under model size scaling (e.g., 1B, 7B, 13B parameters)?.

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

9 papers retrieved. 7 claims extracted, 0 verified. Tribunal: 3.3/10 → 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 VLUE benchmark is publicly available at https://vlue-benchmark.github.io.	×	0.10
The data and codes used for training baseline models are available at https://github.com/MichaelZhouwang/VLUE.	×	0.09
VLUE includes a newly annotated private out-of-distribution (OOD) test set for each representative VL task.	×	0.08
The benchmark covers image-text retrieval, visual question answering, visual reasoning, and visual grounding tasks.	×	0.10
The private OOD test sets are annotated on images from the MaRVL dataset.	×	0.10
Image distribution in the OOD test sets differs from COCO/VG images due to manual collection across cultures.	×	0.06
VLUE maintains a leaderboard tracking the performance of representative studies and new methods on VLP.	×	0.06

References

- http://arxiv.org/abs/2504.13275v4
- http://arxiv.org/abs/2603.11114v1
- http://arxiv.org/abs/2205.15237v1