SOVEREIGN: Can routing signatures learned from few-shot prompts on NLVR2 and SNLI-VE generalize to out-of-distribution co

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Integrating image and text data through multi-modal learning has emerged as a new approach in medical imaging research, following its successful deployment in computer vision. While considerable efforts have been dedicated to establishing medical foundation models and their zero-shot transfer to downstream tasks, the popular few-shot setting remains relatively unexplored. Following on from the currently strong emergence of this setting in computer vision, we introduce the first structured benchmark for adapting medical vision-language models (VLMs) in a strict few-shot regime and investigate v

1 Introduction

Analysis of: Few-shot Adaptation of Medical Vision-Language Models. Research goal: Can routing signatures learned from few-shot prompts on NLVR2 and SNLI-VE generalize to out-of-distribution compositional reasoning tasks (e.g., Winoground, VCR) without retraining, and does this transfer maintain higher accuracy than fixed-ratio MoE 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

12 papers retrieved. 9 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 study evaluates medical vision-language models across three domains: histology, radiology, and ophthalmology.	×	0.09
Quilt-1M was employed for histology tasks with ViT-B/32 vision and GPT2 text encoder.	×	0.03
FLAIR, a foundation model focused on color fundus image understanding, was utilized for ophthalmology tasks.	×	0.06
MedCLIP was used for radiology tasks and was pre-trained on CheXpert and MIMIC-CXR datasets.	×	0.04
FLAIR and MedCLIP both use a dual-encoder architecture with ResNet-50 as vision encoder and BioClinicalBERT as text enco	×	0.03
The models cover a wide range of architectures, including both convolutional and ViT architectures.	×	0.04
Five histology datasets were used for evaluation: NCT-CRC, SICAPv2, SkinCancer, MESSIDOR, and FIVES/ODIR200x3.	×	0.02
The few-shot adaptation protocol uses training subsets with S = {1, 2, 4, 8, 16} images per class for each dataset.	×	0.08
The evaluation metric used is balanced average accuracy (ACA).	×	0.06

References

- http://arxiv.org/abs/2502.07409v5
- http://arxiv.org/abs/2409.03868v1
- http://arxiv.org/abs/2307.03135v3