SOVEREIGN: How do different routing strategies (top-k vs. noisy top-k) in SparseMoE vision-language models influence expe

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Vision-Language Model (VLM) have gained widespread adoption in Open-Vocabulary (OV) object detection and segmentation tasks. Despite they have shown promise on OV-related tasks, their effectiveness in conventional vision tasks has thus far been unevaluated. In this work, we present the systematic review of VLM-based detection and segmentation, view VLM as the foundational model and conduct comprehensive evaluations across multiple downstream tasks for the first time: 1) The evaluation spans eight detection scenarios (closed-set detection, domain adaptation, crowded objects, etc.) and eight seg

1 Introduction

Analysis of: Vision-Language Model for Object Detection and Segmentation: A Review and Evaluation. Research goal: How do different routing strategies (top-k vs. noisy top-k) in SparseMoE vision-language models influence expert utilization balance and multimodal classification accuracy on SEED-Bench and MMBench?.

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. 17 claims extracted, 1 verified. Tribunal: 5.5/10 → REVISE (revision_round=1). 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
Object detection and segmentation are fundamental tasks in computer vision.	✓	0.24
VLM-based detection aligns visual features with text descriptions through contrastive learning approaches.	×	0.11
Current VLMs demonstrate strong performance on open-vocabulary tasks.	×	0.11
VLMs use pre-training on large-scale datasets such as CC12M and YFCC1M.	×	0.02
Some VLMs leverage contrastive learning for feature alignment, while others employ cross-attention for feature fusion.	×	0.02
DA-Pro improves cross-domain detection performance by dynamically generating domain-specific detection heads.	×	0.05
COUNTGD improves instance counting by augmenting text prompts in GroundingDINO with visual exemplars.	×	0.02
DA-Pro uses domain-relevant and domain-agnostic prompt prefixes for each target category.	×	0.02
GroundingDINO achieves an AP of 57.1 on the OV-COCO dataset for the swin-T model configuration.	×	0.03
GroundingDINO achieves an AP of 62.7 on the OV-COCO dataset for the swin-B model configuration.	×	0.03
GLIP-T (A) achieves an AP of 65.9 on the OV-COCO dataset.	×	0.02
RegionCLIP achieves a base AP of 31.4 and a novel AP of 25.2 on the dataset reported.	×	0.00
YOLO-World (Large) achieves an AP of 77.6 on the dataset.	×	0.02
DINO (Swin-L) achieves an AP of 62.8 on the dataset.	×	0.00
Faster R-CNN achieves an AP of 55.2 on the dataset.	×	0.00
YOLO-v8 achieves an AP of 74.1 on the dataset.	×	0.00
GroundingDINO (Swin-T) achieves an AP of 57.1 on the OV-COCO dataset.	×	0.02

References

- http://arxiv.org/abs/2307.16125v2
- http://arxiv.org/abs/2311.17092v1
- http://arxiv.org/abs/2504.09480v1