SOVEREIGN: What is the robustness of SMoES to distribution shifts in multimodal inputs (e.g., adversarial image perturbat

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Pre-trained vision-language (VL) models are highly vulnerable to adversarial attacks. However, existing defense methods primarily focus on image classification, overlooking two key aspects of VL tasks: multimodal attacks, where both image and text can be perturbed, and the one-to-many relationship of images and texts, where a single image can correspond to multiple textual descriptions and vice versa (1:N and N:1). This work is the first to explore defense strategies against multimodal attacks in VL tasks, whereas prior VL defense methods focus on vision robustness. We propose multimodal adver

1 Introduction

Analysis of: Multimodal Adversarial Defense for Vision-Language Models by Leveraging One-To-Many Relationships. Research goal: What is the robustness of SMoES to distribution shifts in multimodal inputs (e.g., adversarial image perturbations or text paraphrasing) compared to standard MoE routing, evaluated by accuracy drop on the RobustVQA or A-OKVQA benchmarks?.

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

11 papers retrieved. 10 claims extracted, 5 verified. Tribunal: 6.9/10 → 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
MAT consistently achieves significantly greater robustness against multimodal attacks than the unimodal AT methods, FARE	×	0.10
The improvements are substantial and consistent for CLIP on Flickr30k and COCO (Tab. 1), as well as ALBEF on both datase	×	0.01
Multimodal adversarial perturbations require multimodal defense.	✓	0.15
Solving the inner-maximization in Eq. 6 is non-trivial, since it requires updating both modalities and involves a high c	×	0.03
Unimodal attacks, such as gradient-based image attacks and BERT-Attack for text, perturb a single modality to mislead th	×	0.07
Multimodal attacks, which perturb both image and text modalities, are significantly more effective than unimodal attacks	✓	0.18
Developing defense strategies against multimodal attacks for VL tasks remains largely unexplored.	✓	0.20
Existing defense strategies for VL models mainly focus on vision robustness, in which adversarial attacks perturb only t	✓	0.24
MAT largely improves multimodal robustness, highlighting the importance of considering multimodal perturbations in VL da	×	0.08
Leveraging one-to-many (1:N) image-text relationships via augmentations enhances robustness, an aspect overlooked in uni	✓	0.17

References

- http://arxiv.org/abs/2602.09258v1
- http://arxiv.org/abs/2212.08044v3
- http://arxiv.org/abs/2405.18770v6