SOVEREIGN: What is the impact of scaling the VLA model size (e.g., from 7B to 13B parameters) on the average reward and t

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Recent advances in the areas of multimodal machine learning and artificial intelligence (AI) have led to the development of challenging tasks at the intersection of Computer Vision, Natural Language Processing, and Embodied AI. Whereas many approaches and previous survey pursuits have characterised one or two of these dimensions, there has not been a holistic analysis at the center of all three. Moreover, even when combinations of these topics are considered, more focus is placed on describing, e.g., current architectural methods, as opposed to also illustrating high-level challenges and oppor

1 Introduction

Analysis of: Core Challenges in Embodied Vision-Language Planning. Research goal: What is the impact of scaling the VLA model size (e.g., from 7B to 13B parameters) on the average reward and task completion rate of LongNav-R1 on the R2R-CE dataset?.

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. 6 claims extracted, 6 verified. Tribunal: 7.5/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

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
Recent advances in multimodal machine learning and AI have led to the development of challenging tasks at the intersecti	✓	0.38
There has not been a holistic analysis at the center of all three dimensions: Computer Vision, Natural Language Processi	✓	0.29
Even when combinations of these topics are considered, more focus is placed on describing current architectural methods,	✓	0.37
Embodied Vision-Language Planning (EVLP) tasks are a family of prominent embodied navigation and manipulation problems t	✓	0.44
The paper proposes a taxonomy to unify EVLP tasks and provides an in-depth analysis and comparison of algorithmic approa	✓	0.31
The paper presents core challenges that new EVLP works should seek to address and advocates for task construction that e	✓	0.33

References

- https://doi.org/10.1613/jair.1.13646
- https://doi.org/10.1007/s42524-025-4136-9
- https://doi.org/10.18653/v1/2021.emnlp-main.328