SOVEREIGN: Does increasing VLA parameter count from 7B to 13B improve long-horizon task completion rate and average rewar

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Incremental decision making in real-world environments is one of the most challenging tasks in embodied artificial intelligence. One particularly demanding scenario is Vision and Language Navigation (VLN) which requires visual and natural language understanding as well as spatial and temporal reasoning capabilities. The embodied agent needs to ground its understanding of navigation instructions in observations of a real-world environment like Street View. Despite the impressive results of LLMs in other research areas, it is an ongoing problem of how to best connect them with an interactive vis

1 Introduction

Analysis of: VELMA: Verbalization Embodiment of LLM Agents for Vision and Language Navigation in Street View. Research goal: Does increasing VLA parameter count from 7B to 13B improve long-horizon task completion rate and average reward on R2R-CE when evaluated with zero-shot cross-dataset generalization?.

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. 8 claims extracted, 8 verified. Tribunal: 8.7/10 $\rightarrow$ 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
Incremental decision making in real-world environments is one of the most challenging tasks in embodied artificial intel	✓	0.30
Vision and Language Navigation (VLN) requires visual and natural language understanding as well as spatial and temporal	✓	0.35
The embodied agent needs to ground its understanding of navigation instructions in observations of a real-world environm	✓	0.41
It is an ongoing problem of how to best connect LLMs with an interactive visual environment.	✓	0.22
VELMA is an embodied LLM agent that uses a verbalization of the trajectory and of visual environment observations as con	✓	0.39
Visual information is verbalized by a pipeline that extracts landmarks from the human written navigation instructions an	✓	0.39
VELMA is able to successfully follow navigation instructions in Street View with only two in-context examples.	✓	0.35
VELMA achieves around 25% relative improvement in task completion over the previous state-of-the-art for two datasets.	✓	0.24

References

- https://doi.org/10.48550/arxiv.2407.06886
- https://doi.org/10.1609/aaai.v38i17.29858

- https://openalex.org/W7120272424