SOVEREIGN: Modality-Native Routing in Agent-to-Agent Networks: A Multimodal A2A Protocol Ex

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Preserving multimodal signals across agent boundaries is necessary for accurate cross-modal reasoning, but it is not sufficient. We show that modality-native routing in Agent-to-Agent (A2A) networks improves task accuracy by 20 percentage points over text-bottleneck baselines, but only when the downstream reasoning agent can exploit the richer context that native routing preserves. An ablation replacing LLM-backed reasoning with keyword matching eliminates the accuracy gap entirely (36% vs. 36%), establishing a two-layer requirement: protocol-level routing must be paired with capable agent-lev

1 Introduction

Analysis of: Modality-Native Routing in Agent-to-Agent Networks: A Multimodal A2A Protocol Extension. Research goal: Can SMoES-trained modality routing generalize to other multimodal benchmarks (e.g., DocVQA, InfographicVQA) under domain shift, and how do accuracy and latency tradeoffs differ from chart-specific distribution shifts?.

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. 5 claims extracted, 5 verified. Tribunal: 8.5/10 → 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
Modality-native routing in Agent-to-Agent (A2A) networks improves task accuracy by 20 percentage points over text-bottle	✓	0.48
An ablation replacing LLM-backed reasoning with keyword matching eliminates the accuracy gap entirely (36% vs. 36%).	✓	0.30
MMA2A achieves 52% task completion accuracy versus 32% for the text-bottleneck baseline on CrossModal-CS (95% bootstrap	✓	0.33
Gains concentrate on vision-dependent tasks: product defect reports improve by +38.5 pp and visual troubleshooting by +1	✓	0.29
This accuracy gain comes at a 1.8× latency cost from native multimodal processing.	✓	0.22

References

- http://arxiv.org/abs/2603.11114v1
- http://arxiv.org/abs/2604.12213v1
- http://arxiv.org/abs/2312.04693v3