Iterative Refinement Trade-offs in BLIP-2 Compositional Accuracy and Inference Latency

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the trade-off between compositional accuracy improvements and inference latency when applying iterative refinement to large multimodal models like BLIP-2 on the GQA dataset. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Looping Back to Move Forward: Recursive Transformers for Efficient and Flexible Large Multimodal Models. Research question: What is the trade-off between compositional accuracy improvements and inference latency when applying iterative refinement to large multimodal models like BLIP-2 on the GQA dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

15 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
RecursiveVLM trained with recursive refinement outperforms the standard non-recursive baseline by 1-2% when using only a	×	0.10
At the second recursion step, RecursiveVLM performance improves by 3% compared with the non-recursive baseline.	×	0.09
This work is the first attempt on a recursive Transformer architecture for Large Multimodal Models (LMMs).	✓	0.21
Existing LMMs universally adopt a single forward-propagate paradigm where each parameter is activated only once per inpu	×	0.05
To date, there has been no systematic exploration of recursive architectures for LMMs prior to this work.	×	0.05
RecursiveVLM achieves an average score of 57.46 on the evaluation benchmarks with 1 recursion step.	×	0.03
RecursiveVLM achieves an average score of 58.86 on the evaluation benchmarks with 2 recursion steps.	×	0.04
RecursiveVLM with 2 recursion steps achieves a score of 74.64 on the AI2D benchmark.	×	0.04
RecursiveVLM with 2 recursion steps achieves a score of 38.69 on the Hallusion benchmark.	×	0.04
The evaluation datasets include AI2D, MM-Star, MM-Vet, MMMU, MMB, MathVista, OCR-Bench, and HallusionBench.	×	0.01

References

- http://arxiv.org/abs/2602.09080v1

- http://arxiv.org/abs/2601.15286v1
- http://arxiv.org/abs/2410.22775v2