SOVEREIGN: How does Flamingo's zero-shot or few-shot generalization ability scale with increasing model size or different

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Large-scale pre-training and instruction tuning have been successful at creating general-purpose language models with broad competence. However, building general-purpose vision-language models is challenging due to the rich input distributions and task diversity resulting from the additional visual input. Although vision-language pretraining has been widely studied, vision-language instruction tuning remains underexplored. In this paper, we conduct a systematic and comprehensive study on vision-language instruction tuning based on the pretrained BLIP-2 models. We gather 26 publicly available

1 Introduction

Analysis of: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. Research goal: How does Flamingo's zero-shot or few-shot generalization ability scale with increasing model size or different pretraining datasets in multimodal tasks?.

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. 9 claims extracted, 9 verified. Tribunal: 9.3/10 $\rightarrow$ 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
The study gathers 26 publicly available datasets covering a wide variety of tasks and capabilities.	✓	0.21
The gathered datasets were transformed into instruction tuning format.	✓	0.15
The study introduces an instruction-aware Query Transformer that extracts informative features tailored to the given ins	✓	0.25
InstructBLIP was trained on 13 held-in datasets.	✓	0.19
InstructBLIP attains state-of-the-art zero-shot performance across all 13 held-out datasets.	✓	0.30
InstructBLIP substantially outperforms BLIP-2 models in zero-shot performance on held-out datasets.	✓	0.18
InstructBLIP substantially outperforms larger Flamingo models in zero-shot performance on held-out datasets.	✓	0.18
InstructBLIP achieves 90.7% accuracy on ScienceQA questions with image contexts when finetuned.	✓	0.18
All InstructBLIP models are open-sourced at https://github.com/salesforce/LAVIS/tree/main/projects/instructblip.	✓	0.29

References

- https://doi.org/10.48550/arxiv.2303.04226
- https://doi.org/10.48550/arxiv.2305.06500
- https://doi.org/10.1093/jamia/ocae045