

SOVEREIGN: How does the number of expert modules in a mixture-of-experts architecture affect cross-domain accuracy degrad

SOVEREIGN Research Kernel

Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

The cost of vision-and-language pre-training has become increasingly prohibitive due to end-to-end training of large-scale models. This paper proposes BLIP-2, a generic and efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models. BLIP-2 bridges the modality gap with a lightweight Querying Transformer, which is pre-trained in two stages. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative

1 Introduction

Analysis of: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. Research goal: How does the number of expert modules in a mixture-of-experts architecture affect cross-domain accuracy degradation when shifting from COCO to Conceptual Captions in multimodal vision-language models?.

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

7 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 9.0/10 \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
BLIP-2 bridges the modality gap with a lightweight Querying Transformer	✓	0.22
BLIP-2 achieves state-of-the-art performance on various vision-language tasks	✓	0.25
BLIP-2 bootstraps vision-language representation learning from a frozen image encoder	✓	0.33
BLIP-2 bootstraps vision-to-language generative learning from a frozen language model	✓	0.32
BLIP-2 achieves state-of-the-art performance on various vision-language tasks	✓	0.25
BLIP-2 has significantly fewer trainable parameters than existing methods	✓	0.24
BLIP-2 outperforms Flamingo80B by 8.7% on zero-shot VQAv2 with 54x fewer trainable parameters	✓	0.31
BLIP-2 demonstrates zero-shot image-to-text generation capabilities that can follow natural language instructions	✓	0.24

References

- <https://doi.org/10.1613/jair.4900>
- <https://doi.org/10.48550/arxiv.2301.12597>
- <https://doi.org/10.1109/tpami.2023.3275156>