SOVEREIGN: BLURR: A Boosted Low-Resource Inference for Vision-Language-Action Models

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Vision-language-action (VLA) models enable impressive zero shot manipulation, but their inference stacks are often too heavy for responsive web demos or high frequency robot control on commodity GPUs. We present BLURR, a lightweight inference wrapper that can be plugged into existing VLA controllers without retraining or changing model checkpoints. Instantiated on the pi-zero VLA controller, BLURR keeps the original observation interfaces and accelerates control by combining an instruction prefix key value cache, mixed precision execution, and a single step rollout schedule that reduces per st

1 Introduction

Analysis of: BLURR: A Boosted Low-Resource Inference for Vision-Language-Action Models. Research goal: How does the inference throughput (tokens per second) of SMoES-based MoE-VLMs with varying expert counts (e.g., 4, 8, 16) compare to dense VLMs on the MMMU benchmark under fixed FLOPs budgets?.

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

14 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 7.5/10 $\rightarrow$ 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
BLURR is a lightweight inference wrapper that can be plugged into existing VLA controllers without retraining or changin	✓	0.33
BLURR accelerates control by combining an instruction prefix key value cache, mixed precision execution, and a single st	✓	0.38
In SimplerEnv based evaluation, BLURR maintains task success rates comparable to the original controller while significa	✓	0.39
BLURR is instantiated on the pi-zero VLA controller.	✓	0.24
BLURR keeps the original observation interfaces.	✓	0.21
An interactive web demo allows users to switch between controllers and toggle inference options in real time while watch	✓	0.36

References

- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2512.11769v1
- http://arxiv.org/abs/2401.04088v1