How does the inference latency of quantized LLaVA-1.5 models vary across different image resolutions in multimodal benchmarks compared to dense inference

Assignee Research

May 29, 2026

Abstract

Visual encoding constitutes the basis of large multimodal models (LMMs) in understanding the visual world. Conventional LMMs process images in fixed sizes and limited resolutions, while recent explorations in this direction are limited in adaptivity, efficiency, and even correctness. In this work, we first take GPT-4V and LLaVA-1.5 as representative examples and expose systematic flaws rooted in their visual encoding strategy. To address the challenges, we present LLaVA-UHD, a large multimodal model that can efficiently perceive images in any aspect ratio and high resolution. LLaVA-UHD

1 Introduction

This paper examines: LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. Research question: How does the inference latency of quantized LLaVA-1.5 models vary across different image resolutions in multimodal benchmarks compared to dense inference?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.7/10.

3 Results

10 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 2.7/10.

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
LLaVA-UHD uses CLIP-ViT-L/14 as the visual encoder with a default resolution of 336 $\times$ 336.	×	0.08
LLaVA-UHD uses Vicuna-13B as the LLM.	×	0.09
LLaVA-UHD uses a shared visual resampler as the projector to connect the visual encoder and LLM.	×	0.08
The number of learnable queries in the resampler is set to 64.	×	0.01
For the image partitioned as N sub-patches, the number of visual tokens fed into LLM is 64 $\times$ (N + 1).	×	0.04
The maximum N is set to be 6, which supports a maximum of 672 $\times$ 1008 resolution images.	×	0.08
Stage 1 pretraining uses the CC-595K dataset for 1 epoch with a learning rate of 1e-3 and a cosine learning rate schedul	×	0.04
Stage 1 pretraining has a global batch size of 256 and takes $\sim$5 hours using 8$\times$A100 GPUs.	×	0.05
Stage 2 instruction-tuning uses a 656K mixture dataset with a learning rate of 2e-5 and a batch size of 128.	×	0.03
Stage 2 instruction-tuning takes $\sim$18 hours using 8$\times$A100 GPUs.	×	0.06
The model is evaluated on 9 popular benchmarks: VQA-V2, GQA, ScienceQA, VizWiz, TextVQA, POPE, MME, MMBench, and MMBench	×	0.03
The computation cost (TFLOPs) in processing an image in the maximum supported resolution is reported.	×	0.05
The accumulated multimodal training data volume is reported, including image-text pairs used during pretraining and inst	×	0.05

References

- http://arxiv.org/abs/2506.17608v1

- http://arxiv.org/abs/2403.11703v1
- http://arxiv.org/abs/2502.00425v2