Memory Efficiency Scaling of LLaVA-UHD Across High-Resolution Visual-LLM Benchmarks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the memory efficiency of LLaVA-UHD scale with image resolution (e.g., 1024x1024 to 8192x8192) compared to dense inference in Visual-LLM benchmarks like LVIS. 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. 18 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. Research question: How does the memory efficiency of LLaVA-UHD scale with image resolution (e.g., 1024x1024 to 8192x8192) compared to dense inference in Visual-LLM benchmarks like LVIS?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

14 papers retrieved. 18 claims extracted; 0 independently verified. Quality review score: 4.7/10.

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.	×	0.07
LLaVA-UHD uses a shared visual resampler as the projector to connect the visual encoder and LLM.	×	0.07
During the encoding of image slices, a minor reshape within half patches (maximum 7-8 pixels) could be performed to fit	×	0.04
The number of learnable queries in the resampler is set to 64.	×	0.02
For the image partitioned as N sub-patches, the number of visual tokens fed into LLM is 64 $\times$ (N + 1), with tokens of the	×	0.10
The maximum N is set to be 6 in experiments, which supports a maximum of 672 $\times$ 1008 resolution images.	×	0.07
Stage 1 of training involves tuning only the perceiver resampler with the CC-595K dataset for 1 epoch.	×	0.03
Stage 1 uses AdamW optimizer with a learning rate of 1e-3 and the cosine learning rate schedule.	×	0.02
The global batch size in Stage 1 is set to 256.	×	0.04
The training cost of Stage 1 is $\sim$5 hours using 8$\times$A100 GPUs.	×	0.08
Stage 2 of training involves fine-tuning the visual resampler and LLM with a 656K mixture dataset.	×	0.03
The 656K mixture dataset includes LLaVA-Instruct, TextVQA, GQA, OCR-VQA, and Visual Genome.	×	0.04
The learning rate in Stage 2 is 2e-5 and the batch size is 128.	×	0.02
The training cost of Stage 2 is $\sim$18 hours using 8$\times$A100 GPUs.	×	0.07
The model is evaluated on 9 popular benchmarks: VQA-V2, GQA, ScienceQA, VizWiz, TextVQA, POPE, MME, MMBench, and MM-Bench	×	0.03
The computation cost (TFLOPs) in processing an image in the maximum supported resolution is reported.	×	0.03
The accumulated multimodal training data volume is reported, including image-text pairs used during pre-training and ins	×	0.13

References

- http://arxiv.org/abs/2501.03895v2
- http://arxiv.org/abs/2412.13871v2
- http://arxiv.org/abs/2403.11703v1