Quantization Trade-offs in LLaVA-UHD: Visual Fidelity vs. Inference Efficiency at INT4 and INT8

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the trade-off between visual fidelity and inference efficiency when quantizing LLaVA-UHD with INT4/INT8 compared to FP16, as measured by SEED-Bench scores and memory footprint reduction. Principal component analysis (PCA) is a popular dimension reduction technique often used to visualize high-dimensional data structures. In genomics, this can involve millions of variables, but only tens to hundreds of observations. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: When and why are principal component scores a good tool for visualizing high-dimensional data?. Research question: What is the trade-off between visual fidelity and inference efficiency when quantizing LLaVA-UHD with INT4/INT8 compared to FP16, as measured by SEED-Bench scores and memory footprint reduction?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

4 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
PCA reduces the dimension of a data matrix by constructing orthogonal linear combinations of the variables	×	0.06
The first principal component is the normalized linear combination of variables with the highest variance	×	0.10
The second principal component will be the linear combination with the highest variance orthogonal to the first componen	×	0.10
The mathematical basis of PCA is the eigendecomposition of the covariance matrix	×	0.01
Let X = [x1, . . . , xn] be a p $\times$ n data matrix, where xi = [xi1, . . . , xip]T are independent and identically distribu	×	0.02
The eigendecomposition of the covariance matrix is given by Σ = VΛVT	×	0.00
The population principal components are defined to be the linear combinations given by the eigenvectors of Σ	×	0.07
The variance of the component scores sTj is given by the jth eigenvalue	×	0.11

References

- http://arxiv.org/abs/2605.08985v1
- http://arxiv.org/abs/2511.21150v1
- http://arxiv.org/abs/1401.2781v4