How does the memory footprint of PowerInfer compare to dense inference methods when running LLaVA-1.5 on NVIDI

Assignee Research

May 29, 2026

Abstract

Recently, ChatGPT, along with DALL-E-2 and Codex,has been gaining significant attention from society. As a result, many individuals have become interested in related resources and are seeking to uncover the background and secrets behind its impressive performance. In fact, ChatGPT and other Generative AI (GAI) techniques belong to the category of Artificial Intelligence Generated Content (AIGC), which involves the creation of digital content, such as images, music, and natural language, through AI models. The goal of AIGC is to make the content creation process more efficient and accessible, a

1 Introduction

This paper examines: A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. Research question: How does the memory footprint of PowerInfer compare to dense inference methods when running LLaVA-1.5 on NVIDIA RTX 3090 GPUs during high-resolution image captioning tasks?.

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 8.3/10.

3 Results

10 papers retrieved. 8 claims extracted; 7 independently verified. Quality review score: 8.3/10.

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
ChatGPT, DALL-E-2, and Codex are examples of Generative AI (GAI) techniques that belong to the category of Artificial In	✓	0.35
AIGC involves the creation of digital content, such as images, music, and natural language, through AI models.	✓	0.31
The goal of AIGC is to make the content creation process more efficient and accessible, allowing for the production of h	✓	0.36
AIGC is achieved by extracting and understanding intent information from instructions provided by humans, and generating	✓	0.30
Large-scale models have become increasingly important in AIGC as they provide better intent extraction and thus, improve	✓	0.31
With the growth of data and the size of the models, the distribution that the model can learn becomes more comprehensive	✓	0.36
This survey provides a comprehensive review on the history of generative models, and basic components, recent advances i	✓	0.36
From the perspective of unimodality, the survey introduces the generation tasks and relative models of text.	×	0.15

References

- https://doi.org/10.3390/app12188972
- https://doi.org/10.1109/access.2024.3365742
- https://doi.org/10.48550/arxiv.2303.04226