How do latency-accuracy trade-offs in automatically generated multimodal pipelines compare to manually optimiz

Assignee Research

June 10, 2026

Abstract

The field of computer vision has experienced significant advancements through scalable vision encoders and multimodal pre-training frameworks. However, existing approaches often treat vision encoders and large language models (LLMs) as independent modules, limiting the integration of hierarchical visual features. In this work, we propose HIVE (Hierarchical Pre-Training of Vision Encoders), a novel framework that enhances vision-language alignment by introducing hierarchical cross-attention between the vision encoder and LLM. Unlike conventional methods that flatten image embeddings, HIVE enabl

1 Introduction

This paper examines: Hierarchical Pre-Training of Vision Encoders with Large Language Models. Research question: How do latency-accuracy trade-offs in automatically generated multimodal pipelines compare to manually optimized models on standard vision-language evaluation suites?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

13 papers retrieved. 14 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
HIVE is evaluated on CIFAR-10, CIFAR-100, ImageNet-1K, Tiny-ImageNet, Food-101, Stanford Cars, Oxford-IIIT Pets, and Cal	×	0.02
HIVE is evaluated on MME, GQA, OK-VQA, and ScienceQA for vision-language model (VLM) evaluation.	✓	0.17
HIVE is compared against two baseline configurations: Base and SA.	×	0.02
Base configuration uses the original foundation models CLIP (clip-vit-large-patch14-336) and SigLIP (siglip-large-patch1	×	0.05
SA configuration is a self-attention-based vision encoder trained using a three-stage pre-training method.	×	0.14
Both SA and HIVE follow identical pre-training strategies for the vision encoder to ensure a fair comparison.	×	0.09
For VLM evaluation, both SA and HIVE are further fine-tuned following the procedure used in LLaVA [23].	×	0.05
MobileLLM-350M [24] is used as the language model backbone for both self-attention and hierarchical cross-attention conf	×	0.11
Optimization is performed using decoupled AdamW [25] with a peak learning rate of 1 $\times$ 10-3 and a cosine decay schedule.	×	0.02
Gradient clipping and a linear warmup phase are applied to maintain stability.	×	0.01
For VLM fine-tuning, LLaVA is adopted using the Llama-3.2-1B-Instruct model following standard VLM training practices an	×	0.05
For classification tasks, a classifier head is appended to the vision encoder and fine-tuned while keeping the vision en	×	0.07
All experiments are conducted on a single RTX 3090 GPU with a maximum batch size of 256 for the early stages.	×	0.02
HIVE consistently outperforms self-attention baselines across classification and vision-language benchmarks while signif	×	0.08

References

- http://arxiv.org/abs/2306.09265v1
- http://arxiv.org/abs/2604.00086v1
- http://arxiv.org/abs/2605.17152v1