Comparative Performance Degradation of CLIP Architectures on WILDS Under Visual Distribution Shifts

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the comparative performance degradation of different CLIP architectures on the WILDS benchmark when subjected to controlled variations in visual factor distributions. 16 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Visual Aesthetic Benchmark: Can Frontier Models Judge Beauty?. Research question: What is the comparative performance degradation of different CLIP architectures on the WILDS benchmark when subjected to controlled variations in visual factor distributions?.

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 5.5/10.

3 Results

4 papers retrieved. 16 claims extracted; 3 independently verified. Quality review score: 5.5/10.

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
The Visual Aesthetic Benchmark (VAB) contains 400 evaluation tasks.	✓	0.22
The Visual Aesthetic Benchmark (VAB) contains 1,195 images.	✓	0.21
VAB spans three visual domains: fine art, photography, and illustration.	×	0.13
VAB covers 24 topics.	×	0.02
Each VAB task presents a set of two to six images sharing a common subject.	×	0.04
The study evaluates 20 MLLMs on the VAB benchmark.	×	0.08
The study evaluates six reward models on the VAB benchmark.	×	0.09
The Human Expert baseline consists of 10 expert judges evaluating each released task.	×	0.09
In the expert consensus example shown, 8 out of 10 experts selected Image D as the best.	×	0.08
In the expert consensus example shown, 7 out of 10 experts selected Image A as the worst.	×	0.08
Comparative ranking yields 42 percentage points higher inter-annotator agreement on best-image selection than score-deri	✓	0.20
The Random Guess baseline is computed analytically for each task size k.	×	0.02
Models are evaluated under three independent random permutations of candidate order.	×	0.13
For tasks with k $\geq$ 3 images, the consensus filter requires both best-image and worst-image votes to meet the threshold.	×	0.09
The evaluated MLLMs include models from the Claude, Gemini, GPT, and o-series families.	×	0.03
One of the evaluated reward models is named ArtiMuse.	×	0.05

References

- http://arxiv.org/abs/1004.3555v1
- http://arxiv.org/abs/2305.08685v5
- http://arxiv.org/abs/2605.12684v1