Lightweight Two-Layer CNN vs. Transformer Models in Adversarial Vision Robustness

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the lightweight two-layer CNN architecture with targeted components for visual robustness in Mouse vs. AI 2025 compare to larger transformer-based vision models in terms of accuracy and. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Navigating Simply, Aligning Deeply: Winning Solutions for Mouse vs. AI 2025. Research question: How does the lightweight two-layer CNN architecture with targeted components for visual robustness in Mouse vs. AI 2025 compare to larger transformer-based vision models in terms of accuracy and inference throughput on adversarial image datasets?.

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 3.1/10.

3 Results

14 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.1/10.

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 final model combining SimpleCNN, GLU, and observation normalization achieves a final score of 95.4% on the Track 1 l	×	0.11
The IMPALA ResNet baseline with 4 layers achieves 87.7% final score on Track 1.	×	0.07
The 24-layer ResNet achieves only 65.98% on Track 1 due to severe overfitting.	×	0.03
The deep 24-layer ResNet exhibits a 30 percentage point gap between ASR and MSR (80.96% vs 51.00%).	×	0.04
The SimpleCNN maintains only a 2.8 percentage point gap between ASR and MSR (96.80% vs 94.00%).	×	0.02
Increasing depth from 4 to 24 layers substantially harms performance on Track 1.	×	0.03
Adding data augmentation to ResNet decreases performance on Track 1.	×	0.03
The visual encoder consists of two convolutional layers with aggressive spatial downsampling.	×	0.05
The first layer of the visual encoder applies a 8$\times$8 kernel with stride 4 to the 86$\times$155$\times$1 input, producing 16 feature cha	×	0.03
The second layer of the visual encoder applies a 4$\times$4 kernel with stride 2, expanding to 32 feature channels.	×	0.02
Both layers of the visual encoder employ LeakyReLU activation with negative slope 0.2.	×	0.04
The resulting feature maps of the visual encoder are flattened and projected to 256 dimensions via a fully-connected lay	×	0.02

References

- http://arxiv.org/abs/2602.00982v1
- http://arxiv.org/abs/2103.15670v3
- http://arxiv.org/abs/2007.11344v2