# Reproducibility study of "Bilinear MLPs enable weight-based mechanistic interpretability"

Hamid Reza Ahmadi • Mihnea Andrei Bârsan
Jan Jelínek • Robert-Ștefan Sofroni

25 September 2026

# Introduction

## Target Paper

Reproducibility study of "Bilinear MLPs for weight-based mechanistic interpretability" by Pearce et al. (2025).

## Core Concept

Replacing standard nonlinearities with bilinear forms enables direct analysis of model weights through eigenvalue decomposition.

## Scope & Extensions

Verifying an interpretable low-rank structure, truncation robustness, eigenvector stability and noise regularization on image classification, alongside extensions to adversarial training and RGB datasets.

# Background

## Mechanistic Interpretability

Aims to reverse-engineer neural network computation by analyzing learned weight parameters directly, rather than relying solely on post-hoc or activation-based probes.

## Bilinear MLP Architecture

Replaces standard element-wise non-linearities with bilinear forms $b(x) = (Wx) \odot (Vx)$, allowing each output coordinate to be expressed as an exact quadratic function $\boldsymbol{x}^\top \boldsymbol{B_i x}$.

## Weight-Space Decomposition

Because each interaction matrix $B_i$ is symmetric, eigendecomposition $B_i = \sum \lambda_j v_j v_j^\top$ extracts low-rank, semantically meaningful feature directions directly from weights without input data.

# Methodology

## Bilinear MLP Architecture

Computes $b(x) = (Wx) \odot (Vx)$ without standard non-linearities. Each output coordinate forms a symmetric matrix $B_i$, analyzed via eigendecomposition $B_i = \sum \lambda_j v_j v_j^\top$

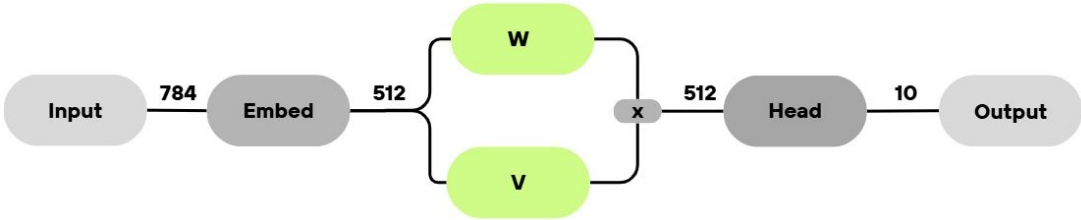
## Datasets & Training

Evaluated on MNIST and Fashion-MNIST grayscale digits/items, and extended to RGB datasets (CIFAR-10, CIFAR-100, Tiny ImageNet) using AdamW optimization.

## Adversarial Extensions

Tested robustness and feature interpretability under Fast Gradient Sign Method (FGSM) and 40-step Projected Gradient Descent (PGD) attacks.

## Architecture Overview




### Key Model Components:

- **Linear Projections:** Parallel weight matrices $W$ and $V$ replace non-linear activation layers.
- **Exact Quadratic Form:** Output features computed directly as $x^T B_i x$ in weight space.
- **Weight Decomposability:** Symmetric interaction tensors allow direct eigendecomposition without data probing.

# Claims

## Claim 1

Eigenvector decomposition of bilinear MLP weights reveals an interpretable low-rank structure across image classification tasks.

## Claim 2

Truncating lower-magnitude eigenvalue components has minimal impact on accuracy, indicating leading eigenvectors capture key task-relevant information.

## Claim 3

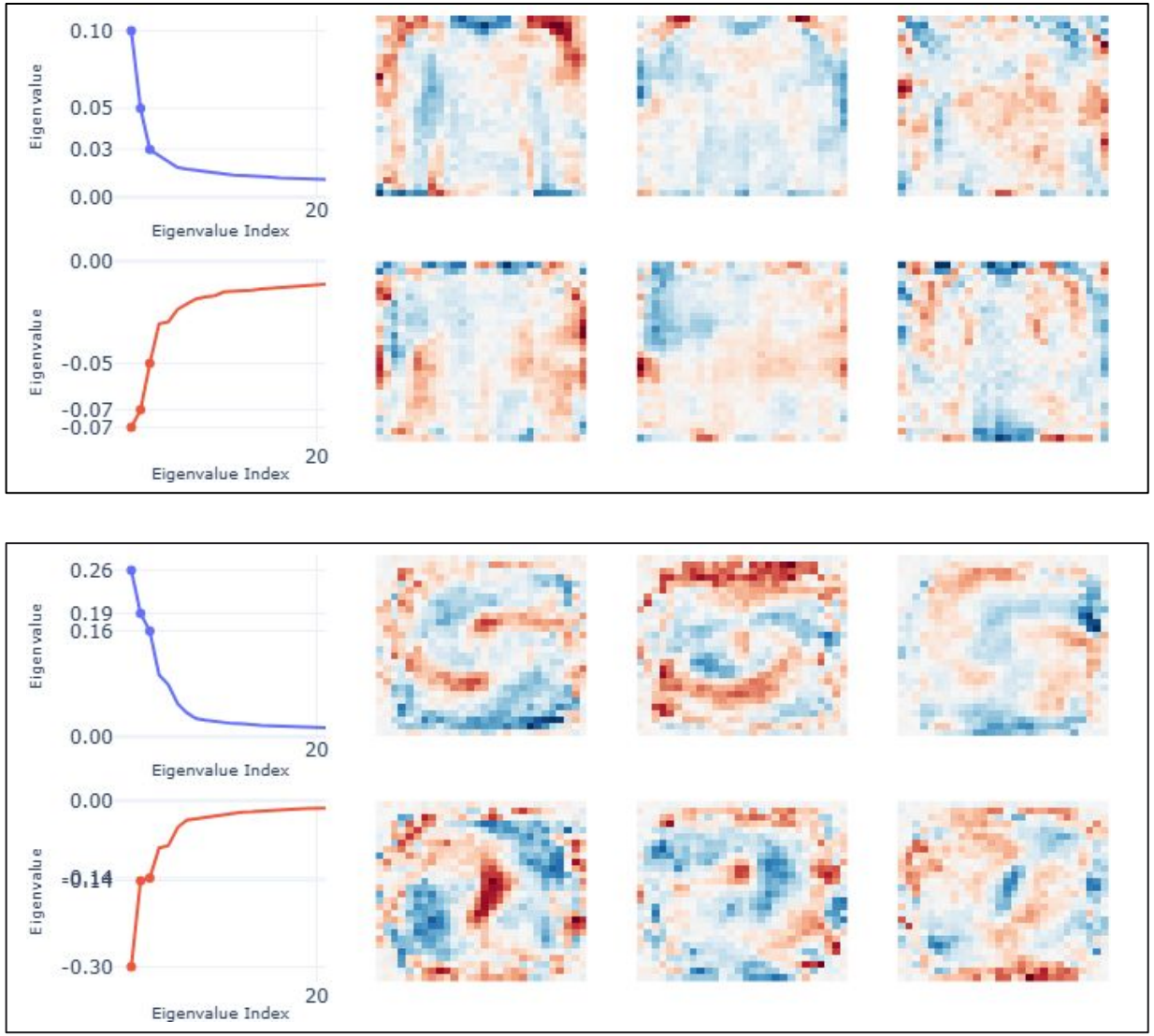
The eigenvectors of bilinear MLPs are stable across random initializations and behave similarly across different model sizes.

## Claim 4

Input noise regularization reduces overfitting and produces bilinear MLPs with eigenvector features that are more interpretable under eigenvalue analysis.

# Results: Claim 1

- Eigenvalue decomposition of bilinear MLP weights reveals an **interpretable low-rank structure** across image classification tasks.
- Spectra exhibit a **sharp, exponential drop-off**, with a small number of leading eigenvalues capturing the majority of spectral mass.
- Remaining eigenvalues **cluster near zero**, indicating that **dominant features govern model behavior**.



# Results: Claim 2

- Truncating lower-magnitude eigenvalue components has **minimal impact on classification performance**, demonstrating that leading eigenvectors capture core task-relevant information.
- The **top 5 eigenvectors account for a significant proportion** of the spectral mass across classes.
- Utilizing t**he top 10 to 15 eigenvectors fully recovers and slightly exceeds full model accuracy**, suggesting a regularizing effect through the removal of noisy components.

| Class | MNIST | Fashion-MNIST |
|---|---|---|
| 0 | 35.96% | 46.97% (T-shirt/top) |
| 1 | 59.65% | 27.24% (Trouser) |
| 2 | 54.76% | 37.40% (Pullover) |
| 3 | 41.04% | 39.64% (Dress) |
| 4 | 35.30% | 21.78% (Coat) |
| 5 | 46.80% | 66.09% (Sandal) |
| 6 | 34.48% | 56.17% (Shirt) |
| 7 | 60.25% | 39.24% (Sneaker) |
| 8 | 16.41% | 31.57% (Bag) |
| 9 | 27.94% | 41.02% (Ankle boot) |

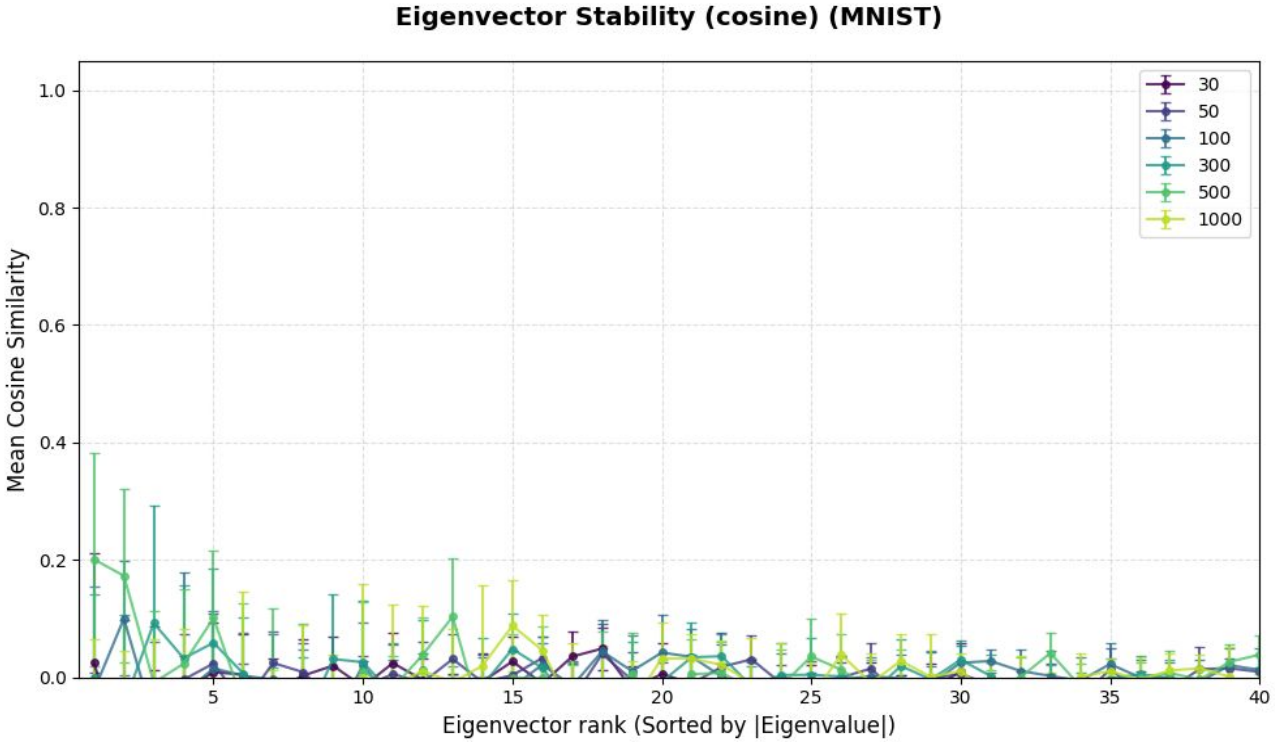
| Dataset | Top $k$ | Accuracy | Drop from Full Model |
|---|---|---|---|
| MNIST | 5 | 93.03% | -1.96% |
| | 10 | 96.20% | **+1.21%** |
| | 15 | 96.43% | **+1.44%** |
| Full Model (MNIST) | All (512) | 94.99% | – |
| Fashion-MNIST | 5 | 83.18% | -0.23% |
| | 10 | 85.25% | **+1.84%** |
| | 15 | 85.37% | **+1.96%** |
| Full Model (Fashion-MNIST) | All (512) | 83.41% | – |

# Results: Claim 3

Claim 3: The eigenvectors of bilinear MLPs are stable across random initializations and behave similarly across different model sizes.




Figure 5 A) from Pearce et al., 2025 (the original paper)

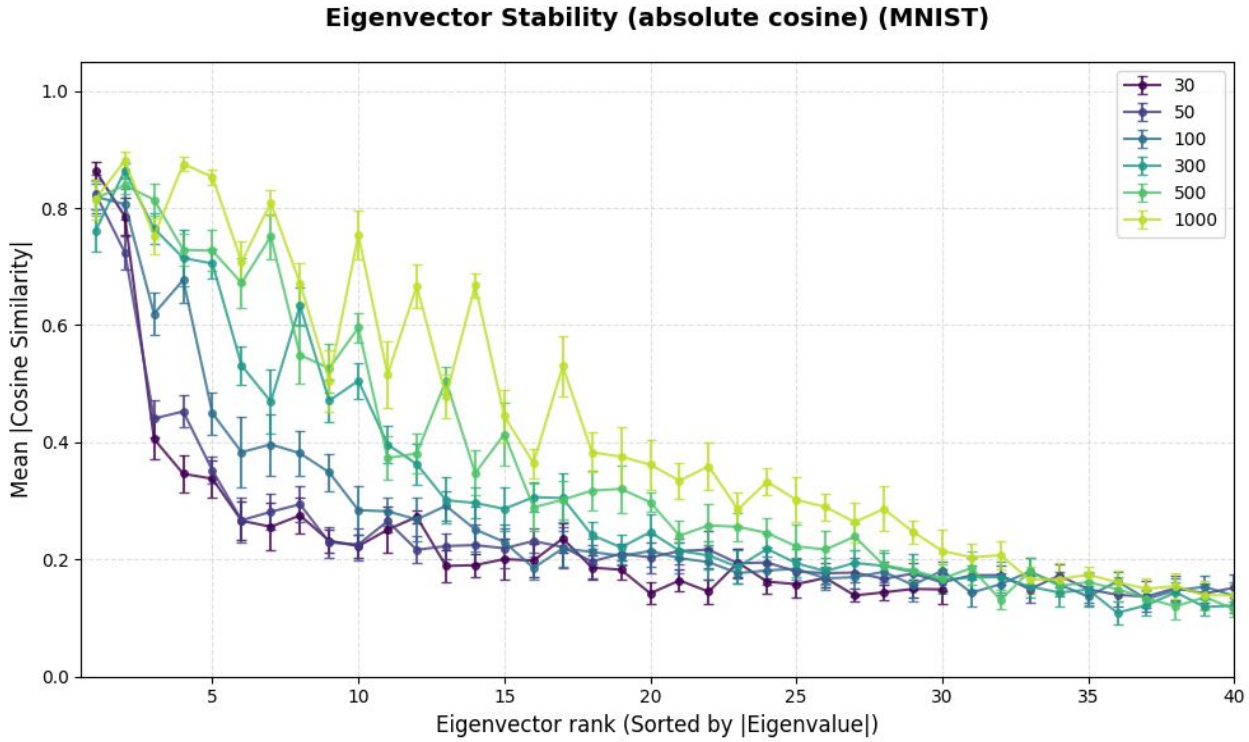



With cosine similarity (as indicated in the reproduced paper)

# Results: Claim 3




Figure 5 A) from Pearce et al., 2025 (the original paper)

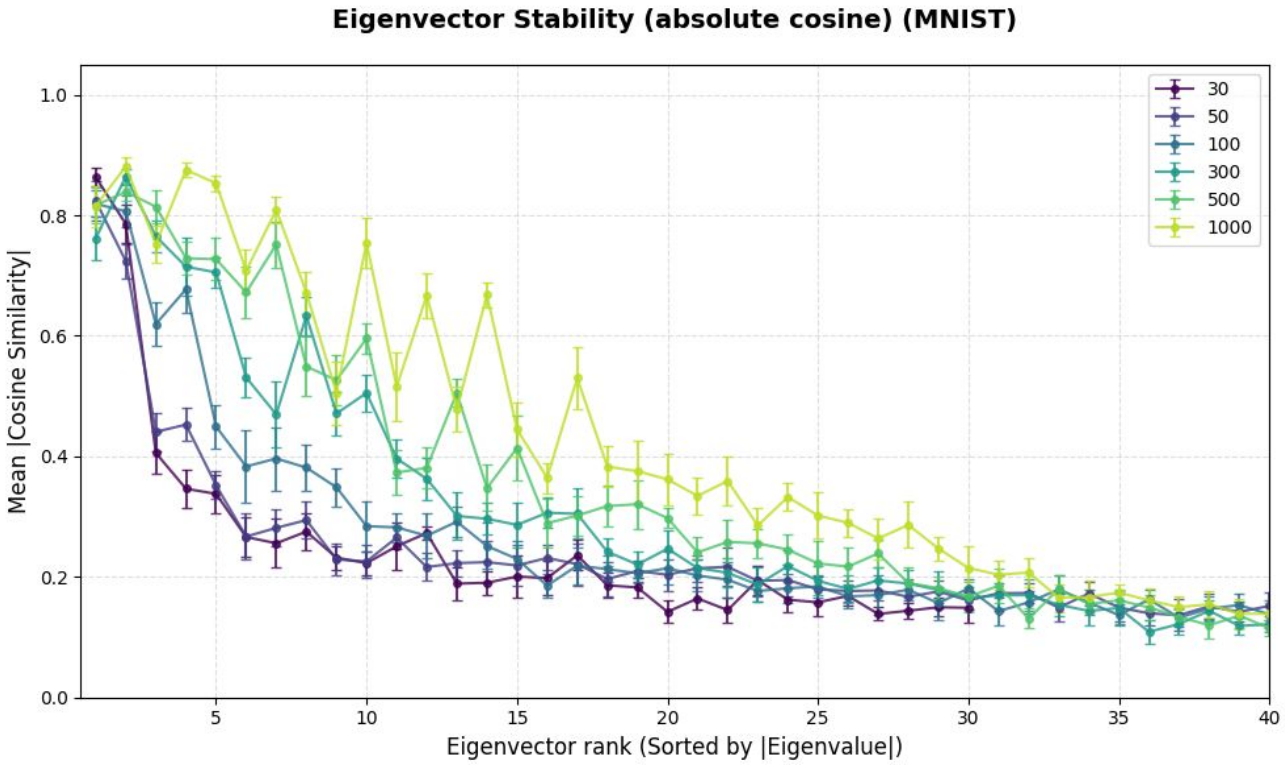



**Why Absolute Cosine Similarity?**

- **The Math:** Bilinear outputs are from quadratic functions.
- **The Problem:** Eigenvector signs are completely arbitrary in quadratic activations. Standard cosine similarity penalizes these arbitrary sign flips, which artificially deflates the similarity score.
- **The Fix:** Applying the absolute value to the cosine similarity prevents this deflation and reveals the true similarity.

# Results: Claim 3

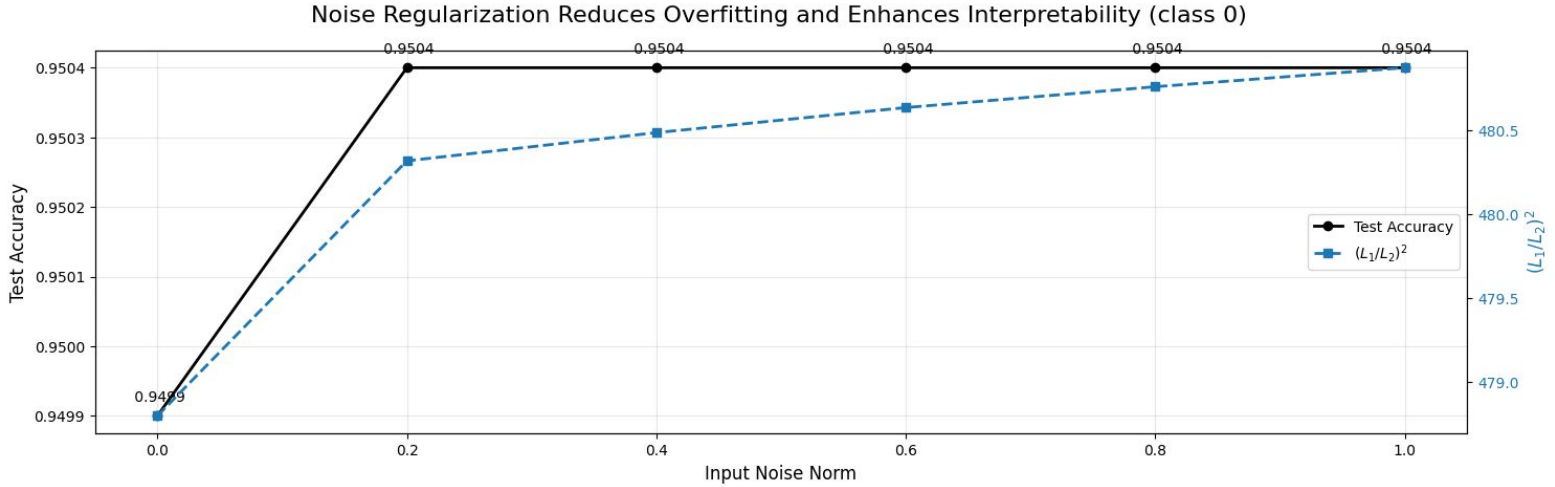
- Claim 3 was confirmed (with the **absolute** cosine similarity).
- The same findings as in the reproduced paper, the main ones being
  - "Both the ordering and contents of top eigenvectors are very consistent across runs. The cosine similarities of the top eigenvector are between 0.8 and 0.9 depending on size (Figure 5). Generally, increasing model sizes results in more similar top eigenvectors." (Pearce et al, 2025).

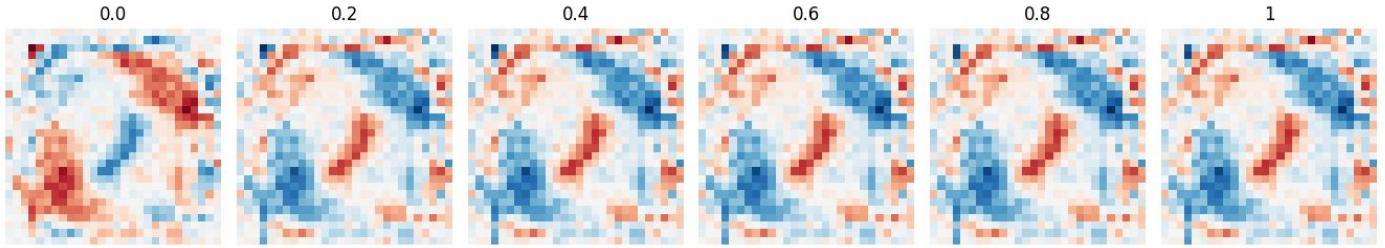


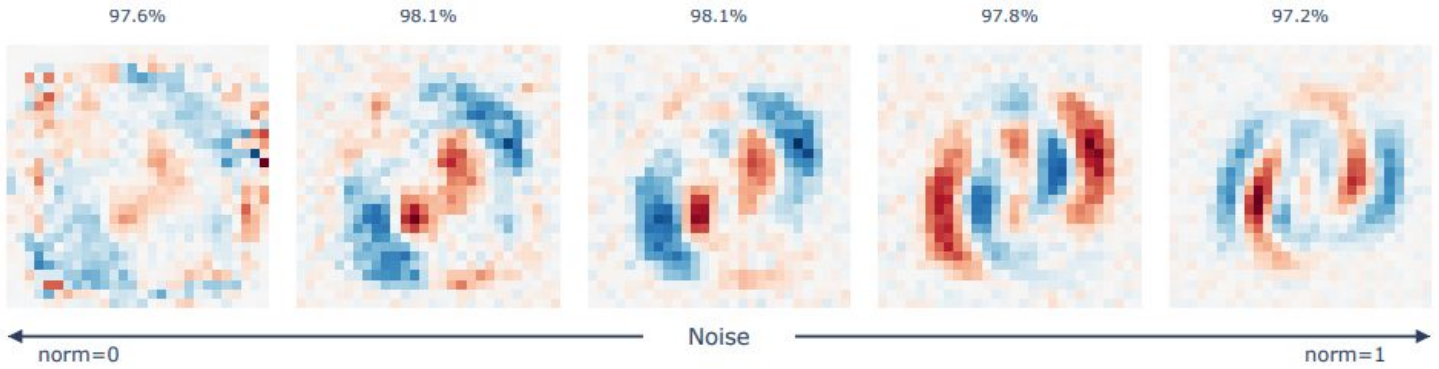

With absolute cosine similarity (the absolute value of the cosine similarity)

# Results: Claim 4

Claim 4: Input noise regularization reduces overfitting and produces bilinear MLPs with eigenvector features that are more interpretable under eigenvalue analysis.



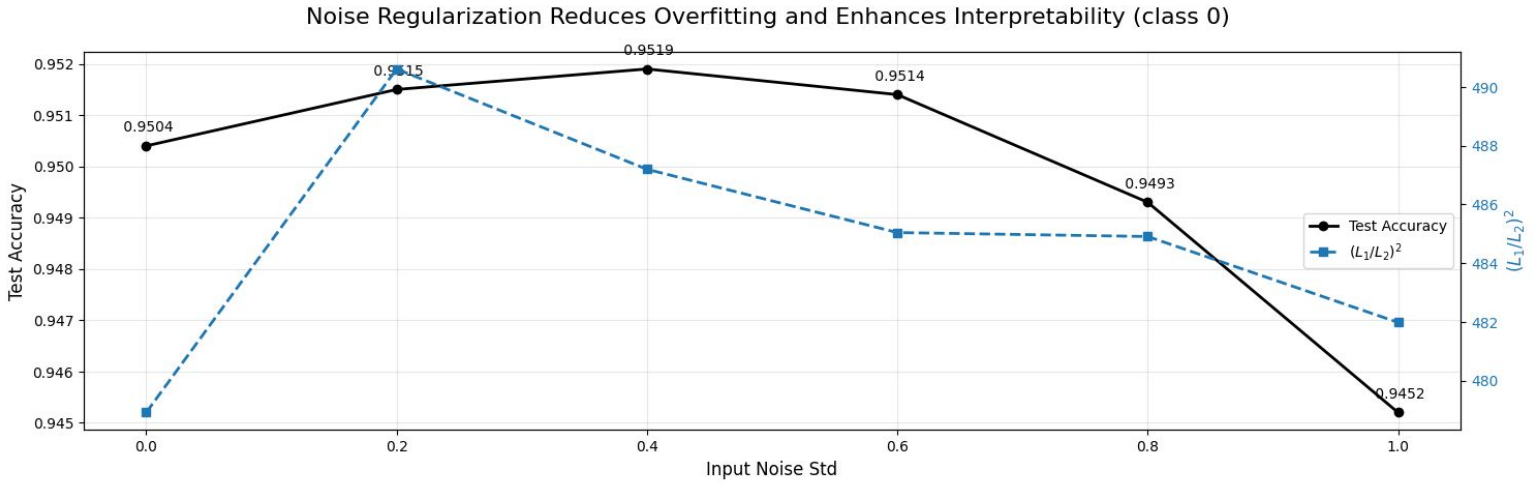



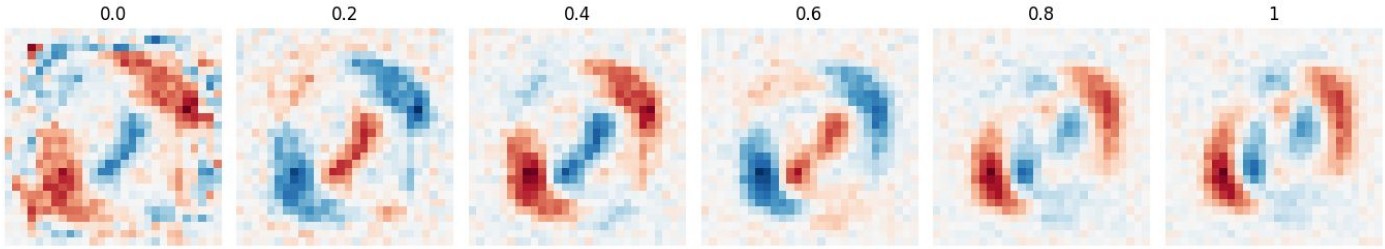

With the added noise scaled to a specific L2 norm (NOT as indicated in the original paper)

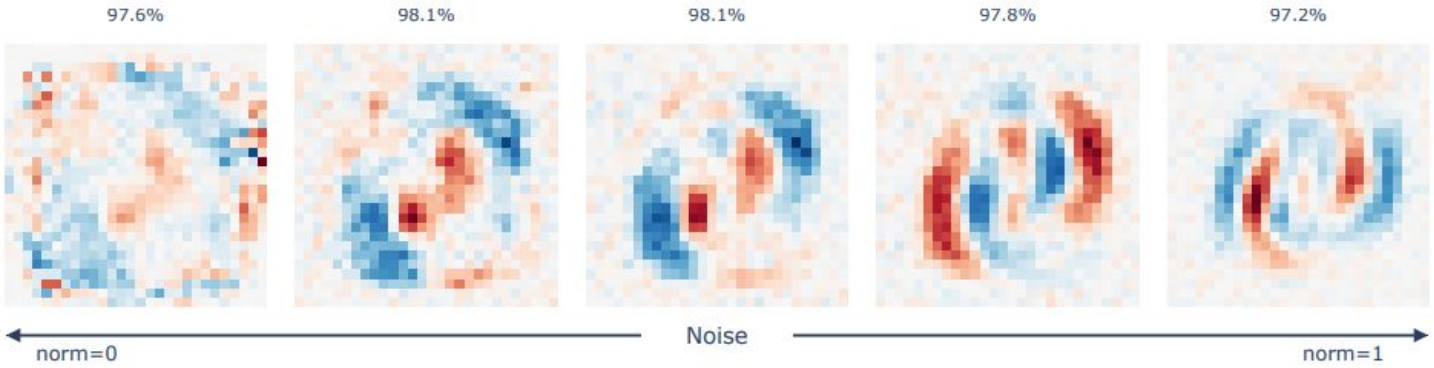
Figure 4 from Pearce et al., 2025 (the original paper)

# Results: Claim 4

Claim 4: Input noise regularization reduces overfitting and produces bilinear MLPs with eigenvector features that are more interpretable under eigenvalue analysis.



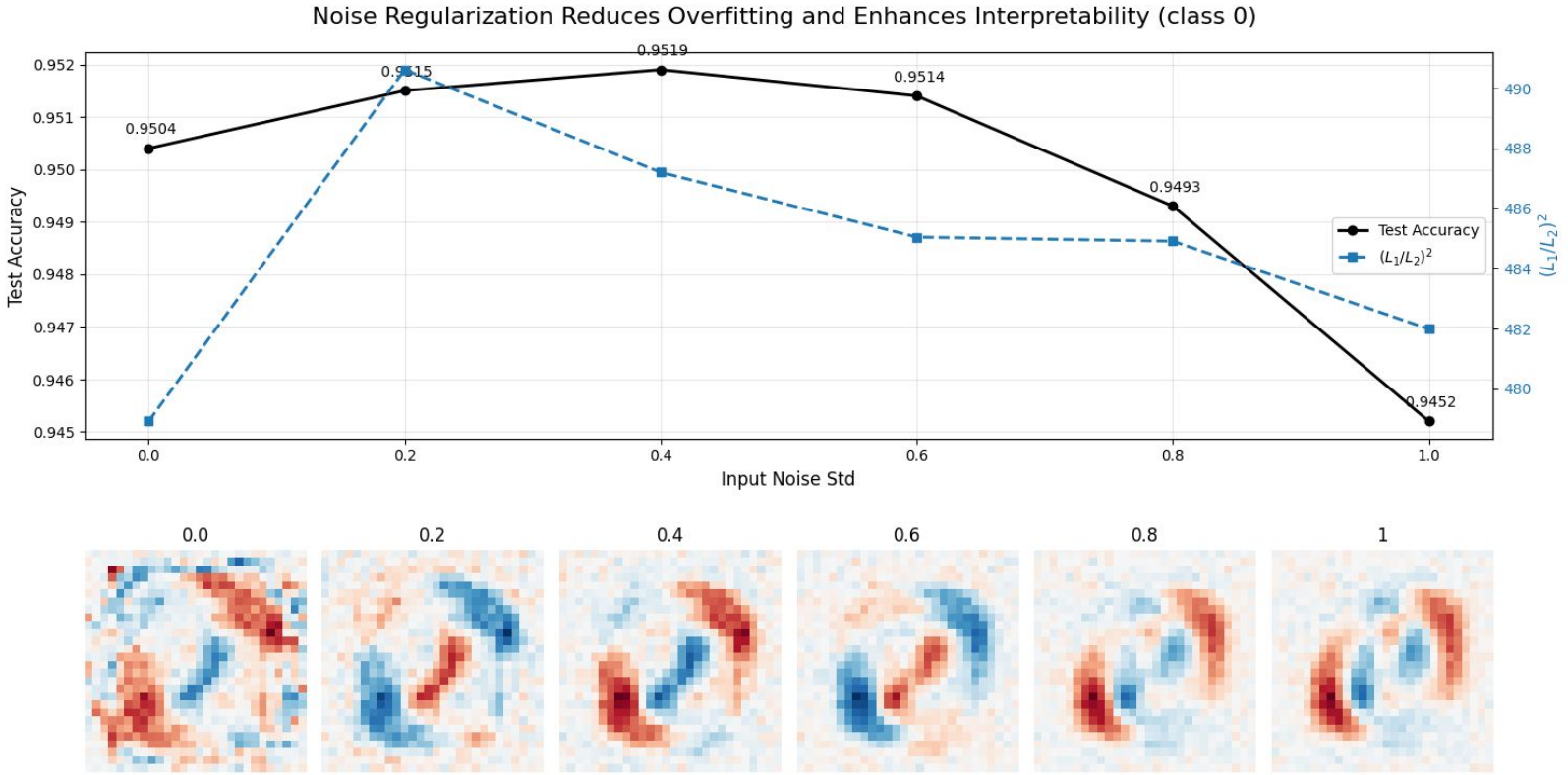




With the added noise scaled with a specific standard deviation

Figure 4 from Pearce et al., 2025 (the reproduced paper)

# Results: Claim 4

- Claim 4 was confirmed (with input noise **standard deviation**)
- The same findings as in the reproduced paper, the main ones of which being
  - "the eigenvectors of unregularized models focus on certain outlying pixels" (Pearce et al., 2025) (signs of overfitting),
  - "Increasing the scale of the added noise results produces more digit-like eigenvectors" (Pearce et al., 2025) (eigenvector features which are more interpretable).




- **Quantitative Nuance**: The $(L_1/L_2)^2$ sparsity surrogate remains highly stable (476-495) even as eigenvectors become visually structured, demonstrating that noise acts as a structural regularizer rather than a mathematical one.

# Extensions

## Extension 1

**Evaluation on adversarial attacks (MNIST):**

- **FGSM:** Single-step linear perturbation method for evaluating local adversarial robustness.
- **PGD:** Multi-step iterative optimization technique for constructing strong adversarial perturbations.

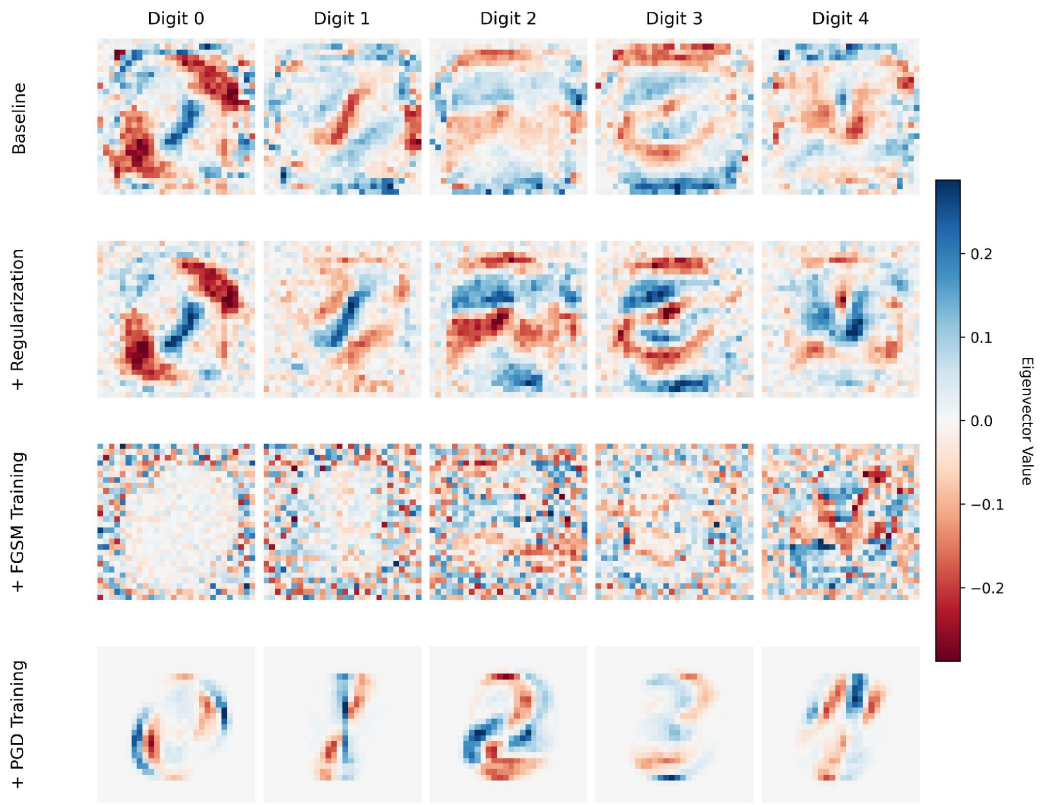
## Extension 2

**Evaluation on other datasets:**

- **CIFAR-10:** 10-class dataset with 32x32 RGB images to test color-channel feature structures.
- **CIFAR-100:** 100-class benchmark evaluating scaling to fine-grained category classification.
- **Tiny ImageNet:** 200 classes at 64x64 resolution to assess generalization on higher-resolution natural images.

# Results: Extension 1

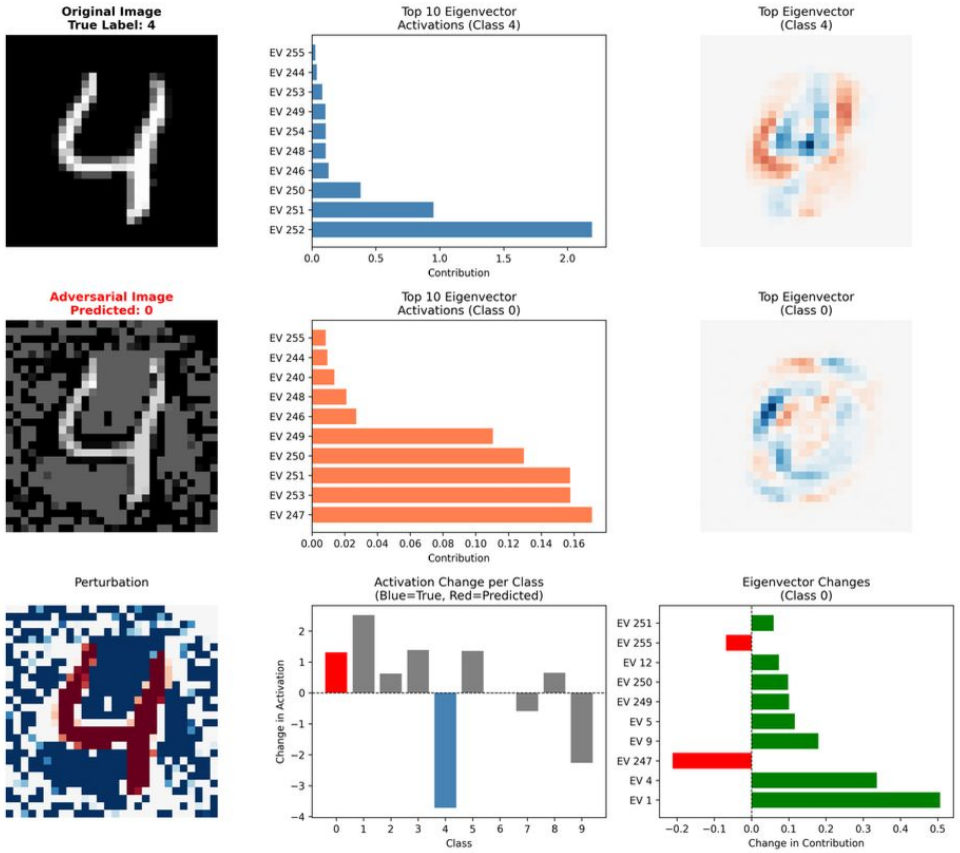
Adversarial training improves interpretability




- Baseline / Regularized → structured but noisy
- FGSM → robustness ↑, interpretability ↓
- PGD → robustness ↑ + clear digit-like features

| Model | Clean Accuracy (%) | FGSM Accuracy (%) | PGD Accuracy (%) |
|---|---|---|---|
| Baseline | 96.5 | 0.02 | 0.13 |
| Baseline + Regularization | 96.6 | 0.20 | 0.01 |
| Baseline + Regularization + FGSM Training | 95.8 | 69.60 | 0.00 |
| Baseline + Regularization + PGD Training | 95.5 | 67.77 | 29.33 |

# Results: Extension 1




- PGD attacks disrupt interpretable eigenvector activations
- Clean “4”
  - activates a dominant 4-like eigenvector
- PGD attack
  - suppresses this feature
  - activates 0-like eigenvectors
  - prediction changes 4 → 0

# Results: Extension 2

Single-layer models consistently outperformed 2- and 3-layer networks

| Dataset | Final Peak Accuracy |
|---|---|
| CIFAR-10 | 0.506 |
| CIFAR-100 | 0.260 |
| Tiny ImageNet | 0.106 |

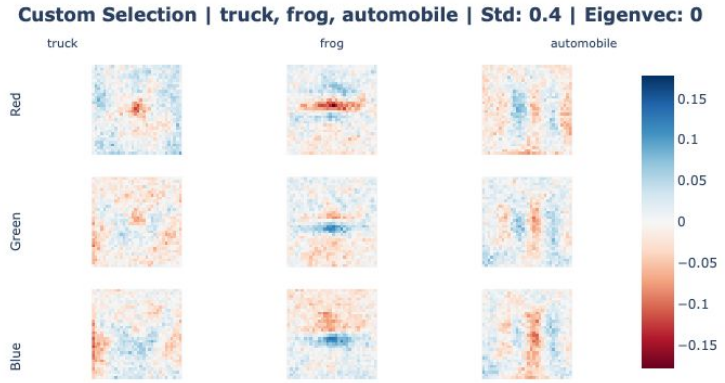
| Standard deviation | Test Accuracy | | Epochs | Layers |
|---|---|---|---|---|
| | CIFAR-100 | Tiny ImageNet | | |
| 0.0 | 0.240 | 0.081 | 20 | 1 |
| 0.0 | 0.215 | 0.031 | 20 | 2 |
| 0.0 | 0.172 | 0.005 | 20 | 3 |
| 0.4 | 0.240 | 0.088 | 20 | 1 |
| 0.4 | 0.223 | 0.014 | 20 | 2 |
| 0.4 | 0.149 | 0.007 | 20 | 3 |
| 0.0 | 0.206 | 0.051 | 100 | 1 |
| 0.0 | 0.201 | 0.077 | 100 | 2 |
| 0.0 | 0.183 | 0.005 | 100 | 3 |
| 0.4 | 0.258 | 0.088 | 100 | 1 |
| 0.4 | 0.234 | 0.064 | 100 | 2 |
| 0.4 | 0.209 | 0.007 | 100 | 3 |

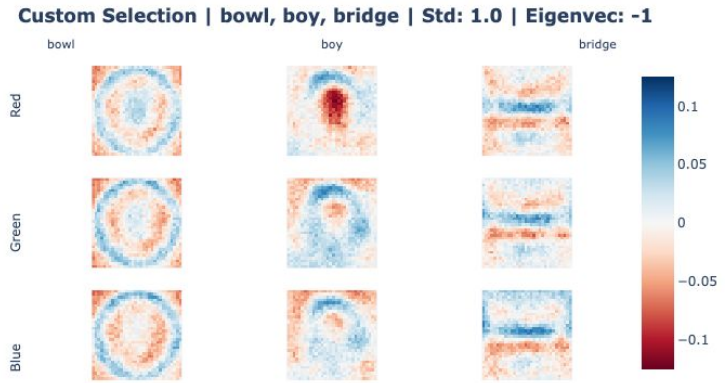
Higher resolution RGB features did not yield human-recognizable semantic shapes



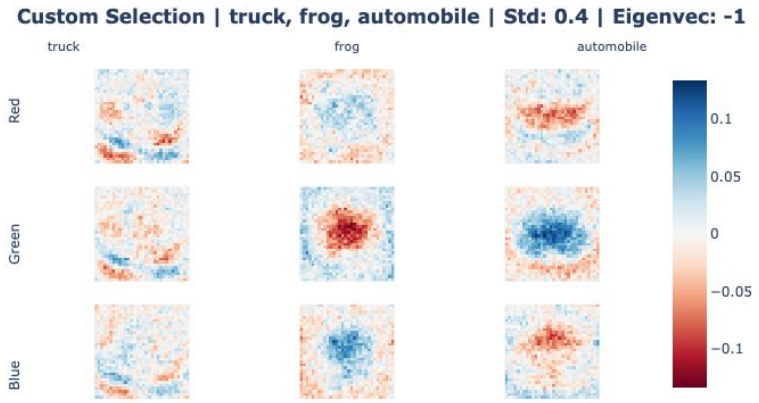
# Results: Extension 2

Increase of noise induces oscillating patterns resembling multi-channel spatial edge detectors

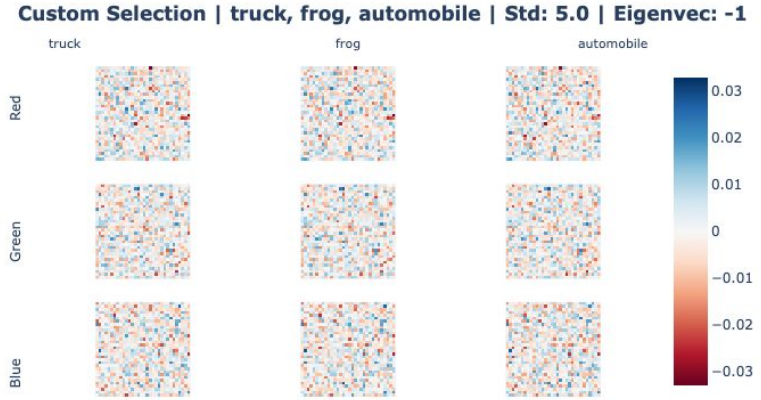



RGB channels complement each other in the activations




Extreme noise degrades feature into noise



# Reproducibility Challenges & Takeaways

- **What was easy**: The original publicly available codebase was modular and default hyperparameters transferred seamlessly for the MNIST and FMNIST datasets.
- **What was difficult**: The original paper had ambiguities regarding the exact type of "norm" and "cosine similarity" used, which required empirical reverse-engineering (e.g., using absolute cosine similarity instead of standard cosine similarity) to successfully reproduce Claims 3 and 4. Furthermore, scaling the method to more complex RGB datasets required extensive hyperparameter tuning for noise standard deviations and model depth.

# Reproducibility Verdict & Conclusion

## 1. Core Claims Verified

**Original claims are reproducible, but required corrections:** Claims regarding low-rank structure and truncation held true. However, verifying similarity and noise regularization required us to resolve mathematical ambiguities present in the original paper.

## 2. Experimental Extensions

- **Sensitivity to Training Signal:** The framework is highly sensitive to the training objective (PGD enhances feature clarity, while FGSM disrupts it).
- **The Scaling Bottleneck:** testing on more complex RGB datasets (CIFAR-100, Tiny ImageNet) reveals that eigenvectors develop spatial edge-detectors rather than human-recognizable shapes.

## 3. Key Takeaways

The Bilinear MLP provides a promising, intrinsically interpretable architecture directly in parameter space, but significant scaling bottlenecks remain for RGB benchmark tasks.

Our open-source, MIT-licensed repository provides a debugged, stress-tested foundation for future researchers to build upon.