Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Large-Model-Assisted Analysis of Adversarial Vulnerability: Separating Evidence-Conflict Detection from Robust Classification

Keji Han

370554@njupt.edu.cn

Nanjing University of Posts and Telecommunications

Research Article

Keywords: Adversarial robustness, Large model assistance, Evidence-conflict detection, Robust classification, Knowledge-based analysis, White-box attacks

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10723486/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Large-Model-Assisted Analysis of Adversarial Vulnerability: Separating Evidence-Conflict Detection from Robust Classification

Keji Han[1*]

[1*]Nanjing University of Posts and Telecommunications, #9 Wenyuan Road, Nanjing, China.

Corresponding author(s). E-mail(s): 370554@njupt.edu.cn;

Abstract

Adversarial examples reveal a persistent mismatch between human visual invariance and the decision boundaries learned by deep neural networks. We study this mismatch as a mechanism problem rather than as a search for a larger backbone or a state-of-the-art robustness number. The proposed workflow uses a large model as an analysis assistant to generate hypotheses about fragile evidence. It then tests these hypotheses with paired clean–adversarial tensors, evidence-switching gates, and white-box attacks. We instantiate the idea on CIFAR-10 with compact ResNet-style backbones, an Aha-ResNet gate, a pairwise ranking variant called AhaV2, and GRPO-style policies that select among base, low-pass, edge, and semantic views. Across tested remote GPU configurations, AhaV2 and GRPO induce larger trigger or semantic-action shifts on adversarial inputs than on clean inputs. The best long-run AhaV2 model reaches 79.85% clean accuracy, 37.45% FGSM accuracy, and 6.65% PGD-50 two-restart accuracy. Human-aligned and joint semantic variants increase PGD-50 Aha shifts to 27.80 and 27.40 percentage points, respectively, but do not improve strong white-box accuracy. A PGD-5 plug-and-play study further shows that a 1,495-parameter AhaV2 trigger improves clean and FGSM accuracy on a frozen adversarially trained backbone, while only slightly improving PGD-50 accuracy. On TinyImageNet, the proposed Aha lens separates PGD-AT, TRADES, MART, and AWP into robust-boundary improvement and unresolved evidence-conflict detection. The main finding is diagnostic: detecting adversarial conflict is trainable and measurable, but it is separable from possessing a robust decision boundary in the alternative evidence path.

Keywords: Adversarial robustness, Large model assistance, Evidence-conflict detection, Robust classification, Knowledge-based analysis, White-box attacks

1 Introduction

Adversarial examples expose a failure mode in which images that appear nearly unchanged to humans induce different neural network predictions [8, 19, 26]. A common diagnosis is that neural networks exploit non-robust or texture-biased features that can be moved by small high-dimensional perturbations [6, 12]. Standard adversarial training improves robustness by optimizing worst-case losses, but it does not directly explain whether the model has detected a conflict in its evidence or changed how it reasons about the image.

We investigate a complementary question: can a compact vision model produce a measurable internal event when adversarial evidence appears, and can that event help explain why the same input remains fragile under stronger attacks? We call this event an *Aha moment*. The term is operational rather than psychological. In this work, an Aha moment occurs when a model increases a trigger gate or shifts its action distribution from a base pixel view toward more stable semantic views under an adversarial perturbation of the same input. This definition supports direct evaluation through clean–adversarial pairs rather than subjective interpretation.

Our central claim is bounded. We do not claim to solve adversarial robustness or to obtain state-of-the-art robust accuracy. Instead, we use large-model-assisted analysis to formulate mechanism-level hypotheses about fragile evidence, then test them with small, controllable vision models. In this workflow, the large model is not a classifier, detector, or defense module. It is a hypothesis generator that helps translate visually plausible explanations, such as preserved object shape or high-frequency texture injection, into quantities that can be checked by tensors and attacks. The experiments indicate that Aha-like evidence switching can be trained, measured, and attached to adversarially trained backbones. They also show that strong white-box attacks expose a gap between detecting adversarial conflict and classifying correctly under attack. This distinction matters for robustness research: a model may notice that its evidence has become suspicious without having a robust decision boundary in the alternative view to which it switches.

We make three contributions. First, we define Aha moments as pairwise evidence-switching events and implement them through Aha-ResNet, AhaV2 ranking gates, and GRPO-style view policies. Second, we convert large-model fragility hypotheses into deterministic tensor audits instead of treating free-form explanations as evidence. Third, we evaluate clean, FGSM, PGD-20, PGD-50, PGD-5 plug-and-play, TinyImageNet, CIFAR-100, and CelebA adversarial-training settings. The results indicate that Aha shift, robust accuracy, and late-training robustness stability are related but separable across datasets.

2 Related Work

Adversarial robustness. Adversarial perturbations were introduced as a surprising failure mode of deep models [26]. Early explanations based on local linearity and robust optimization shaped modern adversarial training [8, 19]. TRADES formalizes the accuracy–robustness trade-off through a regularized objective [39], while MART

and AWP improve robust optimization by emphasizing misclassified adversarial examples and weight-space perturbations [31, 33]. AutoAttack and RobustBench established stricter evaluation protocols that reduce gradient masking artifacts and make robustness comparisons reproducible [3, 4]. Recent work further studies robust overfitting, example selection, instance reweighting, non-zero-sum formulations, and larger-scale robustness evaluation [13, 17, 18, 23, 25, 36]. Our work partially follows this evaluation discipline by reporting PGD-20/50 and by treating weak-attack improvements as insufficient evidence of solved robustness. We do not present the current experiments as a RobustBench-style evaluation.

Recent robust architectures and training. Recent work around 2024–2025 continues to improve adversarial robustness through larger or alternative architectures, better data, robust architecture search, data reweighting, generative purification, and distribution-specific training [2, 9, 15, 16, 20, 32, 38]. Semi-supervised robust learning also remains active: SNORD revisits robust self-training under low-label regimes and reports gains on CIFAR and TinyImageNet [34]. Vision Transformer robustness and adversarially robust foundation-model representations are now studied directly [7, 37]. These works pursue stronger robust classifiers or stronger robust representations. Our focus is different: we use compact ResNet-style models to isolate whether adversarial conflict can be detected and translated into an evidence switch.

Non-robust features and human-aligned vision. Adversarial examples can arise from predictive but non-robust features [12]. A separate line shows that ImageNet-trained CNNs are biased toward texture rather than shape [6]. Frequency-domain analyses also connect adversarial perturbations to high-frequency patterns and feature-trajectory drift [41]. Robust and human-aligned vision has been studied with augmentation, denoising, V1-inspired front ends, and cue-conflict tests [5, 10]. Our Aha policies use these ideas as design pressure: when local evidence becomes unreliable, the model should shift toward semantic or denoised evidence. Unlike work that only reports final accuracy, we explicitly measure the switching event.

Adaptive computation and policy optimization. Conditional computation and mixture-of-experts methods route inputs through different computation paths. Test-time adaptation methods also modify behavior under distribution shift [30, 40]. We study a narrower mechanism: a low-capacity trigger or policy that changes evidence selection for adversarially paired inputs. Our GRPO-style training is used not as a large language model alignment method, but as a compact policy optimization tool for view selection in vision.

Large-model-assisted mechanism analysis. Large models can produce useful visual and semantic hypotheses, but their explanations are not sufficient evidence for adversarial robustness. Automated red-teaming work shows that large models can help search for failures and propose adversarial prompts or behaviors [21, 42]. Broader surveys of LLM-assisted scientific discovery similarly position large models as hypothesis generators rather than experimental proof engines [1]. Recent multimodal attacks show that vision-language systems themselves remain vulnerable to visual adversarial inputs [11, 22]. We therefore use large-model assistance in a restricted role: to propose possible human-aligned evidence factors, such as object shape, edge continuity, texture reliance, and local high-frequency artifacts. These hypotheses are then translated

into measurable view operations and tensor statistics. The large model does not label robustness success; it helps organize mechanism questions that are later accepted or rejected by paired attack experiments. This role is deliberately weaker than an oracle. It suggests what to measure, while strong attacks and paired tensors decide whether the suggestion survives.

Recent multimodal robustness studies further motivate this restrained role. MultiShield uses multimodal large language models to detect and reject adversarial examples by checking visual–textual alignment [28], while Chain of Attack studies transfer-based adversarial attacks against VLMs [35]. Zero-shot adversarial fine-tuning and prompt-based defenses for vision-language models likewise indicate that pretrained multimodal representations can be made more robust, but their robustness depends on the exact threat model and evaluation protocol [14, 29]. Transfer-attack and public-model threat models further show that robustness conclusions change when the attacker is white-box, transfer-based, or constrained to public source models [24, 27]. These works make large or multimodal models relevant to adversarial robustness, but they do not remove the need for white-box tensor-level validation. Our framework therefore treats the large model as a hypothesis generator for fragility mechanisms, not as the final defense.

3 Method

Figure 1 summarizes the proposed mechanism study. The large model proposes candidate fragility factors, but these factors enter the method only after they are translated into paired tensor measurements. The Aha and GRPO modules then test whether a compact model can detect evidence conflict and switch views. Strong white-box evaluation finally determines whether the switched evidence path also has a robust decision boundary.

3.1 Role of the Large Model in the Pipeline

The large model is used as an analysis assistant before model training and evaluation. Its input is a small set of diagnostic artifacts for each clean–adversarial pair: the clean image, the adversarial image, the perturbation map, the predicted labels or margins, and simple summaries of view disagreement. Its output is not a class label or a robustness score. Instead, it produces a structured list of candidate fragility factors. These factors include whether the object identity appears visually preserved, whether the perturbation is likely to affect texture or high-frequency evidence, and whether low-pass, edge, or semantic evidence may be a useful alternative view.

These natural-language hypotheses are then compiled into deterministic operations. A hypothesis about high-frequency evidence becomes a Fourier high-frequency energy test. A hypothesis about preserved visible structure becomes grayscale or edge correlation. A hypothesis about unstable local evidence becomes total-variation gain, view disagreement, or margin collapse. Only this compiled representation enters the rest of the method. The Aha trigger and GRPO policy receive numerical features, view definitions, or reward terms derived from these tensor tests; they do not query the large model during training or inference.

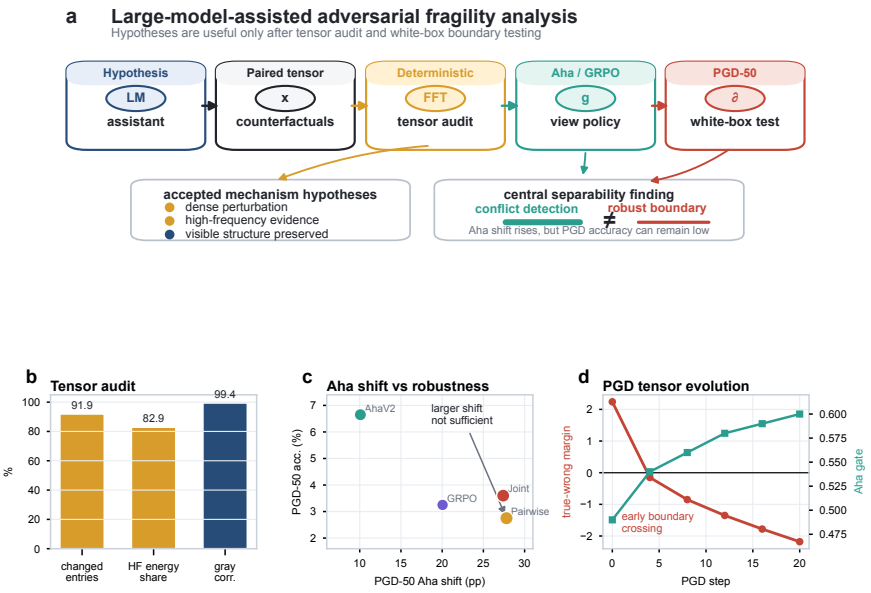


Fig. 1 Overall framework. The method uses a large model only as a hypothesis assistant, converts hypotheses into deterministic tensor audits, trains Aha/GRPO modules to detect evidence conflict and switch views, and then evaluates the resulting models under strong white-box attacks. The central innovation is the explicit separation between detecting adversarial evidence conflict and possessing a robust classification boundary after the switch.

This design gives the large model a narrow and auditable role. It helps decide what to measure, but it does not decide whether the method succeeds. Success is judged by paired clean–adversarial measurements, Aha shift, classification accuracy, and white-box PGD evaluation. The large-model output can therefore be wrong without invalidating the evaluation protocol. Unsupported hypotheses are discarded when they fail the tensor audit, and strong attacks remain the final test of robustness.

3.2 Aha Moment as Pairwise Evidence Switching

Let x be a clean image and x^{adv} be an adversarial counterpart constrained by $\|x^{adv} - x\|_\infty \leq 8/255$. We define an Aha signal as a pairwise increase in either a continuous gate $g(\cdot)$ or a semantic-action probability $p_{\text{sem}}(\cdot)$:

$$\Delta_{\text{Aha}} = g(x^{adv}) - g(x), \tag{1}$$

or, for a policy,

$$\Delta_{\text{sem}} = p_{\text{sem}}(x^{adv}) - p_{\text{sem}}(x). \tag{2}$$

This pairwise definition prevents a trivial solution in which the model always assigns a high trigger value. The desired event is relative: the same image should induce a stronger warning or evidence switch after adversarial perturbation.

Algorithm 1 Large-model-assisted fragility audit

1: **Input:** paired tensors (x_i, x_i^{adv}, y_i), predictions, perturbation maps
2: Ask the large model for candidate fragility factors over each pair.
3: Convert each candidate factor into a measurable tensor statistic or view operation.
4: Keep only factors expressible as tensor tests: density, frequency, edge stability, view disagreement, or margin change.
5: Compute ℓ_∞, changed-entry ratio, Fourier high-frequency share, gray correlation, edge correlation, and total-variation gain.
6: Assign each pair to a fragility bucket using deterministic thresholds from the measured statistics.
7: Use accepted factors to define Aha trigger features, fixed view transformations, or GRPO rewards.
8: **Output:** bucket counts, tensor statistics, and view-selection features.

3.3 Large-Model-Assisted Fragility Hypotheses

The large-model component is used outside the differentiable training loop. Given clean–adversarial image pairs, perturbation maps, and class predictions, it produces a structured hypothesis list. The list specifies which evidence appears stable to a human observer, which evidence is likely texture- or high-frequency-sensitive, and which alternative view should preserve semantic identity. We map these hypotheses to four measurable views: base pixels, low-pass evidence, edge evidence, and their semantic mixture. This design avoids treating free-form large-model explanations as ground truth. A hypothesis is retained only when it predicts an observed tensor-level effect, such as increased gate value, increased semantic-action probability, margin collapse, or failure of the switched-to view under PGD.

The mapping is intentionally simple. The base view tests whether the original classifier remains vulnerable to the adversarial perturbation. The low-pass view tests whether suppressing high-frequency evidence preserves the class signal. The edge view tests whether boundary and shape cues remain stable. The semantic mixture combines low-pass and edge evidence to create a candidate alternative path. The large model motivates these views, but the views themselves are fixed image transformations. This prevents the method from depending on non-reproducible free-form reasoning at evaluation time.

We implement this process as a three-stage audit. First, a large model proposes a fragility schema over the paired images: preserved object identity, dense small-amplitude perturbation, local high-frequency evidence injection, and a predicted need for semantic or denoised evidence. Second, the schema is compiled into deterministic tensor tests. These include ℓ_∞ and changed-entry ratio for perturbation density, Fourier high-frequency energy for local texture injection, grayscale and edge correlation for visible-structure preservation, and total-variation gain for local instability. Third, only hypotheses consistent with these measurements are used to define Aha features or GRPO rewards. Algorithm 1 gives the resulting audit procedure. The audit therefore turns natural-language hypotheses into falsifiable measurements rather than adding an unverified explanation channel.

3.4 Aha-ResNet

Aha-ResNet computes logits from a base image view and a semantic view:

$$v_{\text{sem}} = 0.65\,\text{LowPass}(x) + 0.35\,\text{Sobel}(x). \tag{3}$$

The final prediction interpolates the base and semantic logits:

$$z(x) = (1 - g(x))z_{\text{base}}(x) + g(x)z_{\text{sem}}(x). \tag{4}$$

The gate is trained with classification, robust consistency, and semantic consistency losses. This mechanism is intentionally small: it tests whether explicit evidence switching can emerge without increasing model capacity substantially.

3.5 AhaV2 Pairwise Ranking

AhaV2 replaces the simple gate with a small trigger network over confidence, entropy, margin, view disagreement, high-frequency residuals, total variation, and edge energy. Its objective combines clean and adversarial classification with trigger, consistency, and pairwise ranking terms:

$$\begin{aligned}\mathcal{L}_{\text{AhaV2}} =& \mathcal{L}_{CE}(x, y) + \mathcal{L}_{CE}(x^{adv}, y) \\ &+ \lambda_t \mathcal{L}_{BCE}(g(x), 0) + \lambda_t \mathcal{L}_{BCE}(g(x^{adv}), 1) \\ &+ \lambda_c \text{KL}(f(x^{adv}) \| f(x)) \\ &+ \lambda_r \text{softplus}\left(m - [s_g(x^{adv}) - s_g(x)]\right).\end{aligned} \tag{5}$$

The ranking term directly encodes the Aha event as an increase in trigger score from clean to adversarial input.

3.6 GRPO View Policy

The GRPO-style policy selects among base, low-pass, edge, and semantic views. The policy receives compact features derived from margins, entropy, view disagreement, and high-frequency statistics. Its reward combines label correctness, robust margin, semantic-view preference under perturbation, and penalties for unnecessary clean-image switching. We evaluate both hard action shifts and soft semantic-action probability shifts. We use "GRPO" as a descriptive name for grouped relative preference optimization over image views; the policy is not a language-model alignment module.

3.7 PGD-5 Plug-and-Play Training

To test whether Aha modules are compatible with a stronger backbone, we train a PGD-5 adversarial backbone with

$$\begin{aligned}\mathcal{L}_{bb} =& \text{CE}(x, y) + \text{CE}(x^{adv}, y) \\ &+ \beta \text{KL}(f(x^{adv}) \| f(x)) + \gamma \mathcal{L}_{sem}.\end{aligned} \tag{6}$$

We then freeze the backbone and train only the AhaV2 trigger or the GRPO policy. This isolates plug-and-play behavior from full end-to-end retraining.

4 Experiments

4.1 Setup

All experiments use CIFAR-10 and compact ResNet-style backbones with width 16 and three residual blocks per stage unless otherwise stated. The main subset experiments use 10,000 training examples and evaluate on 2,000 test examples for strong attacks. Attacks use $\epsilon = 8/255$ and, for iterative PGD evaluation, step size 2/255. Strong evaluation reports clean, FGSM, PGD-20 with two restarts, and PGD-50 with two restarts.

We added three remote extensions to test whether the same conclusions survive changes in class granularity and image domain. TinyImageNet experiments use WRN-28-4, a 20k-image training subset, 120 epochs, and PGD-3 evaluation under the same ℓ_∞ budget. CIFAR-100 occurrence experiments use WRN-28-4, a 20k-image training subset, 60 epochs, PGD-5 training/evaluation, and a 4.7k-parameter RobustMemory occurrence student trained on top of the frozen or teacher evidence pathway. CIFAR-100 baseline comparisons use the same WRN-28-4, 20k-subset, 60-epoch, PGD-5 protocol for PGD-AT, TRADES, MART, and AWP. CelebA uses the binary Male attribute, 64×64 images, a 20k-image subset, 40 epochs, batch size 96, and PGD-3 evaluation for the same four baselines. Unless otherwise stated, remote experiments were executed on the NVIDIA GeForce RTX 5090 node with PyTorch 2.8.0–2.11.0+cu128. All table values are reported as percentages.

The experimental protocol follows the mechanism question introduced above. We do not use larger datasets or wider backbones as the main evidence for this paper. Instead, we keep the models small enough that the evidence pathway can be inspected and compare three quantities on the same paired inputs. Classification accuracy answers whether the final decision is correct under clean or adversarial input. Aha shift answers whether the model notices that its evidence has become unreliable, measured as gate shift for Aha-ResNet/AhaV2 or semantic-action shift for GRPO variants. PGD tensor evolution answers when and how the perturbation changes loss, margin, and gate values along attack steps. Large-model assistance is used only before these measurements, to organize qualitative hypotheses about fragile evidence into auditable tensor tests. Acceptance of those hypotheses depends on the quantitative measurements, not on the large model's explanation.

4.2 Evaluation Boundary

Several design choices make the experiments mechanism-oriented rather than benchmark-oriented. CIFAR-10 strong evaluations use 2,000 test examples, PGD-20/50, $\epsilon = 8/255$, step size 2/255, and two restarts. TinyImageNet, CIFAR-100, and CelebA use 20k-image subsets and shorter PGD-3/5 protocols to keep cross-dataset stress tests tractable on a single remote GPU. The reported runs are single-run remote experiments unless a table explicitly states otherwise.

These settings test whether Aha signals and adversarial-training patterns survive stronger attacks than those used during training and survive a change of dataset. They are not sufficient for a final robustness benchmark. We therefore avoid SOTA claims and treat AutoAttack, adaptive attacks against the gate or policy, full-data TinyImageNet/CelebA, and multiple-seed confidence intervals as required follow-up for robustness claims.

4.3 Experimental Questions and Section Organization

The experimental section is organized around mechanism questions rather than around datasets alone. A single robustness number would hide the distinction that the paper aims to test. We first ask whether Aha-style modules behave differently from standard adversarial-training baselines under strong evaluation. This comparison includes CIFAR-10 in Table 2, TinyImageNet in Table 4, and the CIFAR-100/CelebA extension in Table 6. We then ask which component produces the observed behavior: the AhaV2 trigger, the GRPO view policy, the frozen PGD-5 backbone, or the low-capacity RobustMemory occurrence student. Finally, we inspect the mechanism directly through tensor audits, Aha shifts, late-training drift, and PGD step-by-step margin evolution.

This order separates three claims that are easy to conflate. Robust accuracy measures whether the final prediction survives attack. Aha shift measures whether the model detects a change in the evidence it is using. Cross-dataset stability measures whether this diagnostic pattern survives changes in class granularity and image domain. The experiments are therefore designed to test the paper's central claim in pieces: conflict detection and robust classification can help each other, but they are not the same mechanism.

4.4 Analysis Experiment: Large-Model-Assisted Fragility Audit

Before training Aha modules, we test whether the paired clean–adversarial tensors match the large-model-assisted hypothesis schema. Table 1 reports the audit on 50 paired CIFAR-10 tensors. The perturbation is dense, changing 91.92% of tensor entries, yet it remains small under the ℓ_∞ constraint. Its Fourier energy is concentrated in high-frequency components, while the clean–adversarial grayscale correlation remains 0.9939. Thus, the attack changes many local measurements without producing a large visible structural change.

The bucket assignment gives the audit a more concrete interpretation. Forty-four of 50 pairs fall into the dense high-frequency non-robust evidence bucket, four into a weaker local-texture-shift bucket, and two into a structural-check bucket. This pattern supports the mechanism premise used by the Aha models: the visible object is mostly preserved, but a standard classifier can be redirected by small local features. The audit is not treated as evidence of robustness. It only determines which fragility factors the gate and policy experiments should test.

Table 1 Tensor audit of large-model-assisted fragility hypotheses on paired CIFAR-10 clean–adversarial images. The audit is consistent with preserved visible structure and dense high-frequency evidence injection, motivating low-pass, edge, and semantic evidence-switching views.

Audit measurement	Value
Pairs audited	50
Mean ℓ_1 perturbation per entry	0.0200
Mean ℓ_∞ perturbation	0.0289
Changed tensor entries	91.92%
Perturbation high-frequency energy share	0.8288
Median perturbation high-frequency share	0.9034
Clean–adversarial gray correlation	0.9939
Total-variation gain	0.1443
Dense high-frequency bucket	44/50

Table 2 Strong white-box evaluation on CIFAR-10. Aha shift and robust accuracy are not the same objective: Pairwise and Joint variants produce larger PGD-50 shifts, whereas robust_ce120 AhaV2 obtains the highest PGD-50 accuracy.

Method	Clean	FGSM	PGD-20 r2	PGD-50 r2	PGD-50 Aha shift
robust_ce120 Aha-ResNet	63.15	26.40	3.45	3.35	8.83
robust_ce120 AhaV2	**79.85**	37.45	**6.85**	**6.65**	10.08
robust_ce120 GRPO	74.85	35.85	3.30	3.25	20.05
Human-GRPO-Lite	80.65	34.40	4.40	4.15	11.20
Pairwise Human-GRPO	80.35	34.00	3.15	2.75	**27.80**
Joint Robust Semantic Training	**81.55**	**41.45**	4.00	3.60	27.40

4.5 Comparison with Existing Methods on CIFAR-10

Table 2 gives the first separation between conflict detection and robust classification. The strongest PGD-50 accuracy is obtained by robust_ce120 AhaV2 at 6.65%. The largest PGD-50 Aha shift, however, is obtained by Pairwise Human-GRPO at 27.80 percentage points. Joint Robust Semantic Training keeps a similarly high shift of 27.40 points and improves over Pairwise Human-GRPO in PGD-50 accuracy, but it still remains below robust_ce120 AhaV2. The ordering is therefore not monotonic: the model that reacts most strongly to adversarial evidence is not the model that classifies most accurately under the strongest attack.

This mismatch is the key result of the CIFAR-10 comparison. If Aha shift were only a disguised robustness metric, the highest-shift model would also have the highest PGD-50 accuracy. Instead, AhaV2 appears to learn the strongest compact decision boundary among the tested variants, whereas the pairwise and joint semantic models learn a stronger warning or switching response. Their failure under PGD-50 indicates that the alternative evidence path remains attackable. The model can notice that the original evidence has become unreliable and still fail to recover the correct class.

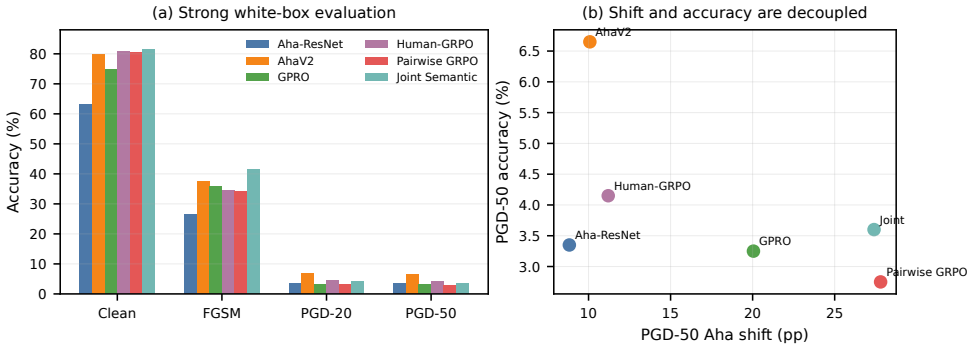


Fig. 2 Strong evaluation and Aha shift. Left: clean and adversarial accuracies under FGSM, PGD-20, and PGD-50. Right: PGD-50 Aha shift versus PGD-50 accuracy. A stronger internal shift does not necessarily imply higher robust accuracy.

The same table also functions as an ablation of the evidence-switching design. Aha-ResNet, AhaV2, and GRPO share the same diagnostic motivation but intervene at different points: a direct gate, a pairwise ranking trigger, and a view-selection policy. AhaV2 improves strong PGD-50 accuracy relative to Aha-ResNet, whereas GRPO-style variants produce larger semantic-action shifts without matching the best robust accuracy. In this compact setting, the pairwise trigger is more useful for the classifier boundary, while the policy objective is more sensitive as a conflict detector. Figure 2 visualizes this non-monotonic relation directly by placing PGD-50 accuracy and Aha shift on the same comparison axis.

4.6 Analysis Experiment: Weak-to-Medium Attacks Can Overestimate Robustness

Weak-to-medium attacks show that the mechanism is learnable, but they overstate its defensive value. During training, AhaV2 and GRPO substantially improve PGD-3 or PGD-10 accuracy. In the main subset run, GRPO raises the best PGD score from 27.55% for the backbone to 41.72%. In long-run training, GRPO reaches 44.63% PGD-10. These numbers indicate that the gate and policy can change model behavior under attack; they are not random alarms.

The strong evaluation changes the interpretation. Under PGD-50 with two restarts, robust_ce120 GRPO drops to 3.25%. Figure 3 visualizes this gap. The result explains why the paper reports weak, medium, and strong attacks together. Weak attacks indicate that evidence switching can be learned, but they can also overestimate robustness when the attacker is not strong enough to optimize through the gate or policy. Detecting a perturbation and switching evidence paths can help against weaker attacks. It becomes insufficient when the switched-to view lacks its own robust class boundary.

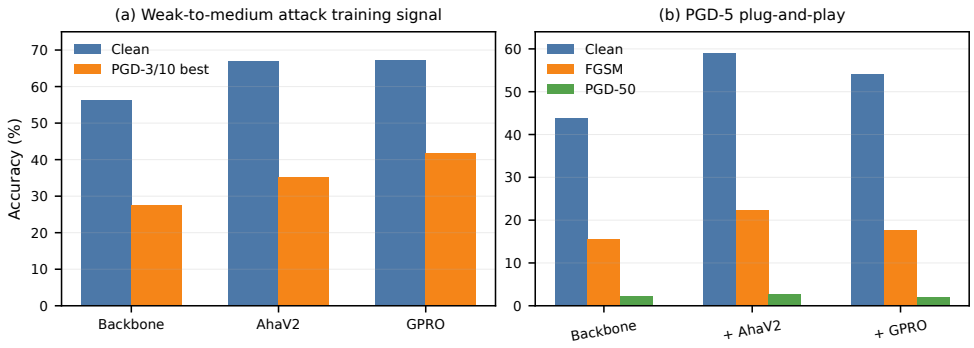


Fig. 3 Training signals and PGD-5 plug-and-play evaluation. Left: weak-to-medium attack training shows large apparent gains. Right: when fitted to a frozen PGD-5 backbone, AhaV2 improves clean and FGSM accuracy but only slightly improves PGD-50 accuracy.

Table 3 Plug-and-play evaluation on a PGD-5 adversarially trained backbone. AhaV2 improves clean and FGSM performance, but strong PGD robustness remains limited.

Method	Clean	FGSM	PGD-20	PGD-50
PGD-5 Backbone	43.70	15.45	2.35	2.30
PGD-5 + AhaV2	**59.05**	**22.25**	**2.65**	**2.60**
PGD-5 + GRPO	54.15	17.65	1.95	1.95

4.7 Ablation: PGD-5 Plug-and-Play Evaluation

Table 3 asks whether the Aha module can be added to an already adversarially trained backbone without retraining the full classifier. The answer is partially positive. AhaV2 improves clean accuracy from 43.70% to 59.05% and FGSM accuracy from 15.45% to 22.25% while training only 1,495 trigger parameters on a frozen backbone. This shows that a small adapter can change the model's use of evidence and improve easy-to-medium attack behavior.

The same ablation also marks the boundary of the adapter. PGD-50 accuracy improves only from 2.30% to 2.60%, so the result should not be read as a complete defense. GRPO obtains a positive soft semantic shift under PGD-50 but a negative hard semantic shift, suggesting that argmax policy decisions become unstable under strong attacks. The practical conclusion is plug-and-play compatibility, not plug-and-play robustness.

The frozen-backbone design is what makes this result interpretable. If full end-to-end retraining were allowed, any gain could be attributed to a stronger classifier. Here, only the trigger or policy is trainable. The outcome is deliberately mixed: the adapter improves clean and weak-attack behavior, but it cannot repair the frozen backbone's strong PGD-50 boundary. This supports the two-component interpretation more directly than a larger end-to-end model would.

Table 4 TinyImageNet-200 remote stress test under a shared WRN-28-4, 20k-subset, 120-epoch, PGD-3 protocol. RobustMemory does not match the best clean accuracy. Its 4.7k trainable memory/gate parameters reach the same PGD-3 range as adversarial-training baselines and produce a measurable Aha delta under this limited protocol.

Method	Trainable params	Best epoch	Best clean	Best FGSM	Best PGD-3	Aha delta
Aha-ResNet WRN-28-4	5.90M	120	13.85	1.95	2.75	–
RobustMemory student	4.7k	120	21.23	**23.95**	15.67	16.14
PGD-AT	5.90M	53	30.00	9.45	14.80	–
TRADES	5.90M	115	32.34	8.89	14.22	–
MART	5.90M	68	30.83	10.94	**16.28**	–
AWP	5.90M	113	**33.91**	9.12	15.21	–

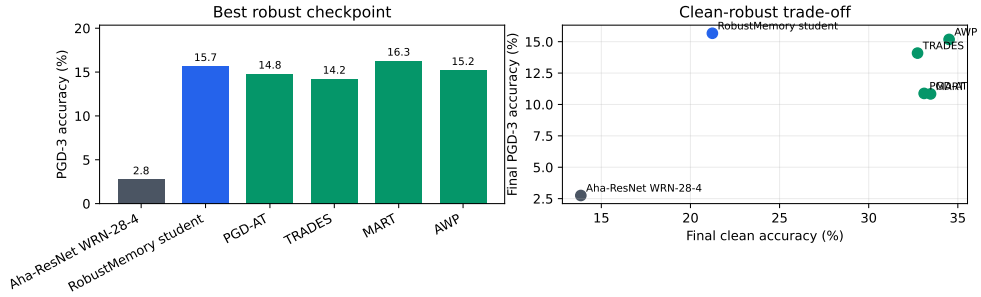


Fig. 4 TinyImageNet stress-test comparison. The result supports the paper's mechanism claim: low-capacity memory/gating can produce weak-attack robustness in the same range as several baselines under this protocol, while full robust classification still depends on the strength of the underlying decision boundary.

4.8 Cross-Dataset Comparison: TinyImageNet Stress Test

TinyImageNet tests whether the mechanism survives a more difficult visual regime. We ran the remote stress test on the B860 EXO GPU node with WRN-28-4 backbones, a 20k-image training subset, 120 epochs, and PGD-3 evaluation at $\epsilon = 8/255$. We compare against four closely related adversarial-training baselines: PGD-AT [19], TRADES [39], MART [31], and AWP [33]. The purpose is not to make an AutoAttack or SOTA claim. It is to ask whether a low-capacity memory/gate can still expose adversarial evidence conflict when the dataset has 200 classes, smaller per-class sample support, and more diverse backgrounds.

Figure 4 plots the same TinyImageNet comparison visually, making the trade-off between clean accuracy, PGD-3 accuracy, and low trainable-parameter adaptation easier to inspect.

Table 5 CIFAR-100 occurrence extension. Values are reported at the epoch with the best PGD-5 accuracy for each model, with final PGD-5 shown for late-training stability. The RobustMemory student uses only 4.7k trainable memory/gate parameters and its maximum PGD Aha delta across training is 15.57 percentage points.

CIFAR-100 model	Params	Best epoch	Clean	FGSM	PGD-5	Final PGD-5
Aha-ResNet teacher	5.87M	21	28.55	9.23	9.31	7.34
RobustMemory student	4.7k	10	**40.88**	**34.61**	**18.02**	**17.43**

Table 6 Cross-dataset comparison between the proposed occurrence student and adversarial-training baselines. CIFAR-100 uses PGD-5 and CelebA uses PGD-3. All runs use WRN-28-4 with 20k-image training subsets; CIFAR-100 is trained for 60 epochs and CelebA for 40 epochs. Best epoch denotes the epoch with the best adversarial accuracy for that method.

Dataset	Method	Best epoch	Best clean	Best FGSM	Best PGD-k	Final PGD-k
CIFAR-100	RobustMemory (ours)	10	40.88	**34.61**	18.02	17.43
CIFAR-100	PGD-AT	41	45.94	20.67	21.89	20.39
CIFAR-100	TRADES	59	43.65	19.11	20.30	19.73
CIFAR-100	MART	50	**48.13**	23.34	**24.40**	**23.00**
CIFAR-100	AWP	57	42.37	19.31	20.69	20.17
CelebA	PGD-AT	27	89.84	70.75	79.00	74.63
CelebA	RobustMemory (ours)	36	71.23	66.43	58.59	58.57
CelebA	TRADES	32	89.61	**73.37**	80.21	**77.86**
CelebA	MART	18	**91.91**	70.83	**80.72**	74.22
CelebA	AWP	25	89.63	69.61	78.66	76.74

4.9 Cross-Dataset Comparison: CIFAR-100 and CelebA

CIFAR-100 and CelebA separate two forms of generalization. CIFAR-100 increases label granularity while preserving the small natural-image regime. CelebA changes the problem into binary facial-attribute classification with larger aligned images. The CIFAR-100 occurrence run used WRN-28-4, a 20k-image training subset, 60 epochs, and PGD-5. The baseline queues used PGD-AT, TRADES, MART, and AWP under the same CIFAR-100 protocol, and a 40-epoch PGD-3 protocol on CelebA. The goal is not to declare a dataset-wide winner. It is to ask whether the diagnostic split between conflict detection and robust classification remains visible when either the label space or the image domain changes.

Table 5 shows that the occurrence student transfers beyond CIFAR-10 under a single-epoch reporting rule. At the epoch with the best PGD-5 accuracy, the 4.7k-parameter RobustMemory student improves clean accuracy from 28.55% to 40.88%, FGSM accuracy from 9.23% to 34.61%, and PGD-5 accuracy from 9.31% to 18.02% relative to the Aha-ResNet teacher. It also produces a maximum PGD Aha delta of 15.57 points over the 60-epoch run. This result matters because CIFAR-100 is a harder class-granularity test than CIFAR-10. A useful gate cannot merely learn that an input

is perturbed. It must preserve enough class-relevant information to avoid collapsing among 100 labels.

Table 6 gives the complementary baseline view under a best-adversarial-epoch reporting rule. On CIFAR-100, MART obtains the strongest best PGD-5 score at 24.40%, followed by PGD-AT at 21.89%, AWP at 20.69%, and TRADES at 20.30%. The RobustMemory student reaches 18.02% PGD-5 and 17.43% final PGD-5. It remains below the fully adversarially trained WRNs in best PGD-5 accuracy, but it produces the strongest FGSM accuracy in the table at 34.61% while training only the low-capacity memory/gate component on top of the evidence pathway. This comparison keeps the interpretation balanced: RobustMemory is parameter-efficient and diagnostic, whereas full adversarial training remains stronger for robust boundaries.

CelebA shows a different pattern. The RobustMemory student reaches 58.59% best PGD-3 accuracy and 58.57% at the final epoch, while using only 4.7k trainable parameters. Its maximum Aha delta is 27.17 percentage points. The student therefore exposes a strong evidence-conflict signal and remains stable late in training, but it does not match the fully adversarially trained baselines. Among those baselines, MART reaches the highest best PGD-3 accuracy at 80.72%, while TRADES has the best final PGD-3 accuracy at 77.86%. The best-final gap remains informative: MART peaks early on CelebA, whereas TRADES and AWP retain more late-training adversarial accuracy. Together, CIFAR-100 and CelebA reinforce the same interpretation as TinyImageNet. Adversarial training is the more direct route to robust boundaries, while the occurrence student provides a parameter-efficient measurement of evidence conflict.

4.10 Analysis Experiment: What Does Adversarial Training Fix?

The TinyImageNet, CIFAR-100, and CelebA comparisons clarify what existing adversarial-training methods fix. PGD-AT, TRADES, MART, and AWP mainly strengthen the adversarial decision boundary. On TinyImageNet, their best PGD-3 accuracies range from 14.22% to 16.28%, compared with 2.75% for the plain Aha-ResNet WRN-28-4 backbone. On CIFAR-100, their best PGD-5 accuracies range from 20.30% to 24.40%. On CelebA, their best PGD-3 accuracies range from 78.66% to 80.72%. These results support the standard view: adversarial training remains the most direct way to improve white-box robustness.

Our analysis adds a second axis rather than disputing that conclusion. The same results expose what adversarial training does not explain. First, all four baselines retain large clean–PGD gaps on every multi-class dataset, which means robust optimization reduces but does not remove reliance on fragile evidence. On TinyImageNet the gap is 14.55–18.70 percentage points; on CIFAR-100 it remains 21.68–27.74 points. Second, the training objective affects robustness stability. On TinyImageNet, PGD-AT and MART obtain their best PGD-3 scores at epochs 53 and 68, respectively, but lose 3.92 and 5.44 percentage points by epoch 120. TRADES and AWP peak much later and lose only 0.13 and 0.03 points.

The CIFAR-100 and CelebA extensions show why final robustness must be reported alongside best robustness. Methods with strong peaks can still give back adversarial

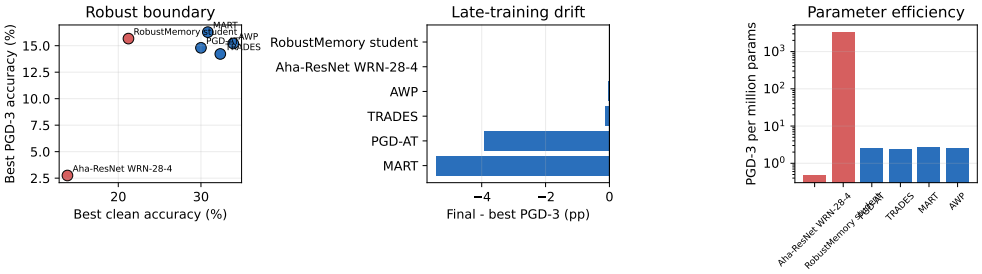


Fig. 5 Effectiveness and limits of adversarial training under the Aha lens. Left: adversarial training mainly improves the robust decision boundary. Middle: PGD-AT and MART show larger late-training PGD-3 drift than TRADES and AWP. Right: RobustMemory is not the strongest classifier, but its trainable-parameter efficiency and measurable Aha delta indicate that conflict detection is a distinct mechanism from full robust classification.

accuracy by the end of training. More importantly for this paper, none of the standard baselines directly reports whether the model detected an evidence conflict while the boundary was being crossed. RobustMemory fills this diagnostic role across both datasets. It reaches 18.02% best PGD-5 on CIFAR-100 and 58.59% best PGD-3 on CelebA with only 4.7k trainable memory/gate parameters; its maximum Aha deltas are 15.57 and 27.17 percentage points, respectively. These results indicate that a low-capacity module can learn a measurable conflict signal and remain reasonably stable late in training, even when full robust classification still depends on the underlying WRN decision boundary.

Figure 5 summarizes this two-axis interpretation: adversarial training improves the boundary, whereas RobustMemory exposes conflict-detection behavior at much lower trainable capacity.

4.11 Analysis Experiment: Tensor Evolution Under PGD

The PGD trace shows how conflict detection and misclassification develop together. Over PGD steps, the loss rises from 1.26 to 2.55, the average margin moves from 0.31 to -1.49, and the Aha gate rises from 0.66 to 0.74. Figure 6 shows that the gate responds along the same trajectory that crosses the decision boundary. The response is meaningful, but it is reactive: the gate increases as the attack pushes the sample toward a wrong class.

We further trace 50 paired CIFAR-10 tensors with the robust_ce120 AhaV2 checkpoint under deterministic PGD-20. The model correctly classifies 42/50 clean inputs but only 6/50 final adversarial inputs. Among the 36 samples that are initially correct and finally wrong, 33 cross the decision boundary within the first four PGD steps; the flip-step distribution is 5, 8, 7, and 13 samples at steps 1–4, respectively. The probability mass moves accordingly. The mean true-class probability falls from 0.7279 to 0.1955, the strongest wrong-class probability rises from 0.1728 to 0.6204, and the mean margin changes from 2.2380 to -2.1839 while ℓ_∞ remains bounded by 8/255. This

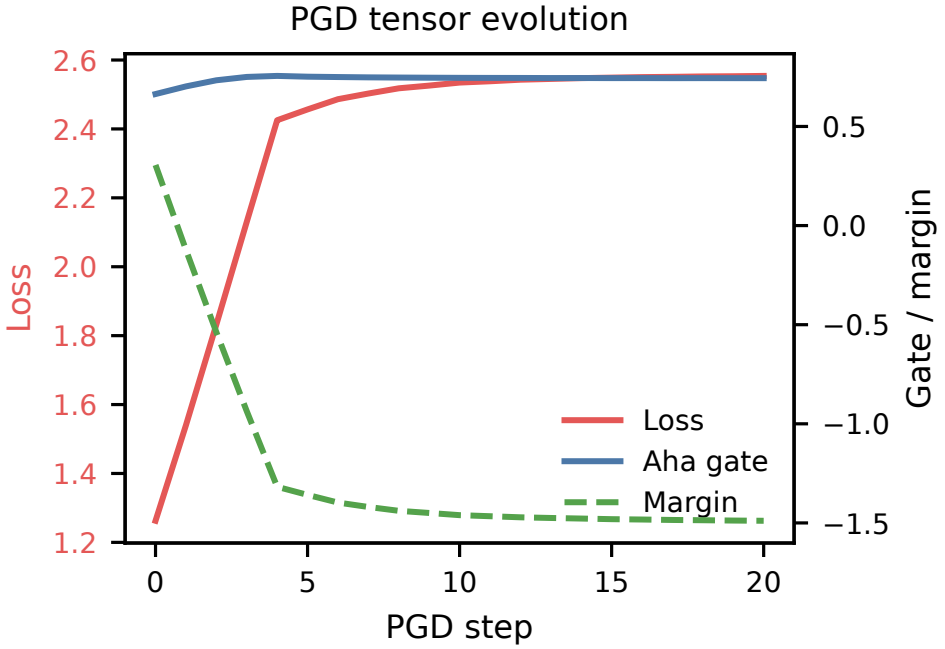


Fig. 6 PGD tensor evolution. As loss increases and the margin crosses the decision boundary, the Aha gate also increases.

sample-level trace is consistent with adversarial misclassification as a margin-collapse process rather than a visually obvious semantic change.

Figure 7 shows this sample-level process as probability transfer, margin crossing, and simultaneous gate increase under the same bounded perturbation.

5 Discussion

The experiments support a bounded interpretation. Aha moments are trainable: gates and policies can be made to respond more strongly to adversarial inputs than to clean inputs. They are also partly plug-and-play: a small AhaV2 trigger can be attached to a frozen PGD-5 backbone and improve clean and weak-attack behavior. The boundary of the result is equally important. Under PGD-50, a model can shift strongly toward semantic evidence while still making the wrong prediction. The large-model-assisted analysis helps formulate this failure mode, but the tensor experiments provide the evidence.

This division is the central methodological contribution. Large models are useful for formulating human-readable fragility hypotheses, such as high-frequency evidence injection or preserved visible structure. In the proposed pipeline, those hypotheses are converted into fixed view transformations and tensor statistics before training begins. The large model is therefore not used as a robustness judge, a runtime detector, or

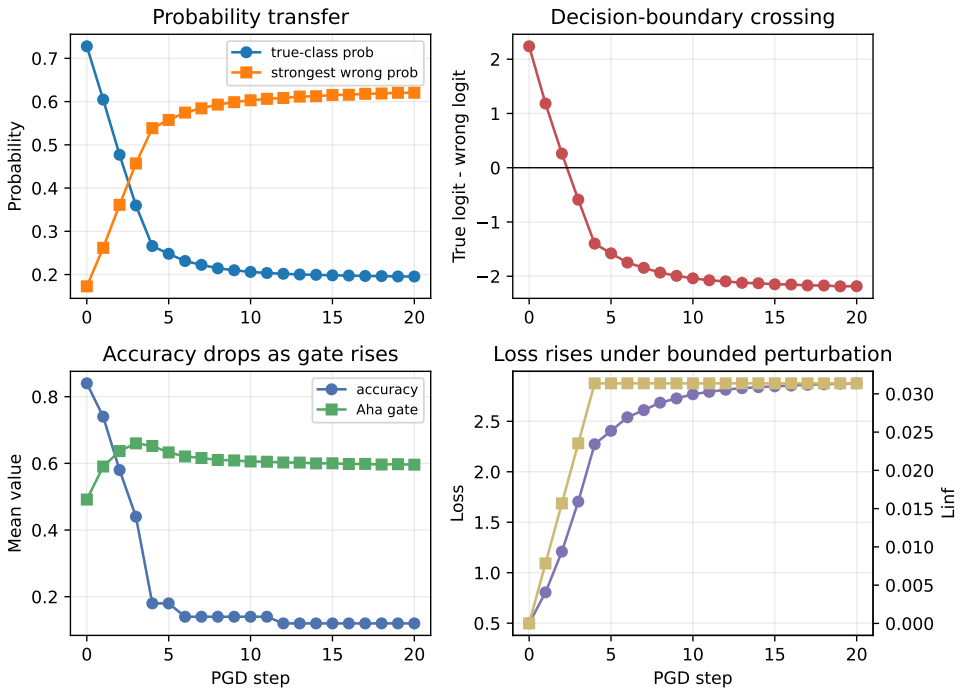


Fig. 7 Sample-level PGD misclassification process on 50 CIFAR-10 tensors. Left: probability mass transfers from the true class to the strongest wrong class. Middle: the average true-vs-wrong logit margin crosses zero, marking decision-boundary crossing. Right: accuracy drops as the Aha gate increases under the same bounded perturbation.

an adaptive defense component. Paired tensors and white-box attacks decide which hypotheses survive. This keeps the role of the large model deliberately narrow and makes the evidence auditable.

The adversarial-training comparison clarifies the same point from the opposite direction. PGD-AT, TRADES, MART, and AWP are effective because they directly optimize adversarial decision boundaries. This is why they dominate the plain Aha-ResNet backbone on TinyImageNet PGD-3 and why they exceed the CIFAR-100 occurrence student in best PGD-5 accuracy. Yet their remaining clean–PGD gaps, best-versus-final drift, and different behavior across CIFAR-100 and CelebA indicate that robust optimization does not automatically expose the internal evidence conflict. The Aha modules should therefore be viewed as diagnostic and complementary. They measure whether a model notices fragile evidence; adversarial training determines whether the model has a class boundary that can survive after noticing it.

The failure case is instructive rather than merely negative. Robust reasoning requires both conflict detection and a robust alternative evidence path. If the semantic, low-pass, or denoised view shares a weak decision boundary with the base model, then switching to that view does not prevent white-box attacks from optimizing through

the entire computation. This explains why high Aha shift can coexist with low PGD-50 accuracy.

An alternative explanation is that the Aha signal only reflects confidence, entropy, or a generic distribution-shift alarm. The current design reduces this risk in three ways. First, the Aha definition is pairwise, so a uniformly high trigger does not count as a successful event. Second, the tensor audit links the trigger to specific perturbation properties such as high-frequency energy, margin collapse, and preserved grayscale structure. Third, the plug-and-play experiment freezes the backbone and changes only the trigger adapter, separating the detector from full classifier retraining. These checks do not eliminate all confounds. Future adaptive attacks should directly target the trigger, and detector-only ablations should compare Aha shift against confidence, entropy, and calibration baselines.

Limitations. The present evidence is strongest for mechanism analysis and weaker for benchmark robustness. CIFAR-10 strong evaluation uses a 2,000-example subset, while TinyImageNet, CIFAR-100, and CelebA use 20k-image training subsets and PGD-3/5 rather than AutoAttack. CelebA is also a binary attribute task, so its high PGD-3 accuracies should not be compared directly with the 100- or 200-class settings. The large-model component is restricted to hypothesis generation; the paper does not establish that a particular large model can reliably detect adversarial examples. The next experimental step should remain aligned with the mechanism claim. Large models can propose more precise fragility hypotheses and view transformations, but those proposals should be tested through AutoAttack, adaptive attacks against the trigger and policy, full-data cross-dataset runs, and view-level adversarial training for the alternative evidence paths.

6 Conclusion

We introduced a large-model-assisted framework for studying adversarial vulnerability through measurable pairwise evidence switching. On CIFAR-10 with compact ResNet-style models, AhaV2 and GRPO-style policies produce trigger or semantic-action shifts under adversarial perturbation. Strong evaluation indicates that these shifts are meaningful but not sufficient: the best current PGD-50 accuracy is 6.65%, whereas the largest Aha shifts occur in models with lower PGD-50 accuracy. A PGD-5 plug-and-play study confirms compatibility with adversarially trained backbones while exposing the same strong-attack boundary. The cross-dataset results sharpen the same conclusion. TinyImageNet, CIFAR-100, and CelebA extensions indicate that adversarial training improves robust boundaries across datasets, but it does not by itself explain whether the model detects vulnerable evidence. Conversely, the low-capacity occurrence student can learn a measurable conflict signal without becoming the strongest robust classifier. The resulting lesson is practical: large models can help formulate human-readable hypotheses about vulnerable evidence, but robustness claims must be decided by paired tensor measurements and strong adaptive attacks.

Statements and Declarations

Funding

This work was partially supported by the Jiangsu Province Excellent Postdoctoral Program, and the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY223213)

Competing interests

The author has no relevant financial or non-financial interests to disclose.

Author contributions

The current contribution record attributes conceptualization, methodology, software, validation, formal analysis, investigation, visualization, writing of the original draft, and review and editing to Keji Han.

Ethics approval

Not applicable. This study uses publicly available image datasets and does not report experiments involving human participants, human material, or animals conducted by the author.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The experiments use the publicly available CIFAR-10, CIFAR-100, TinyImageNet-200, and CelebA datasets. Dataset access and the corresponding original sources are identified in the manuscript references. No new human-participant data were generated for this study.

Code availability

The implementation, experiment configurations, and supporting result summaries are available from the corresponding author upon reasonable request. This statement should be replaced with a versioned public repository record if one is created before publication.

Use of large models

Large-model assistance is part of the reported research method only as a hypothesis-generation stage. Candidate hypotheses were converted into deterministic tensor

audits and evaluated by the reported experiments. The large model was not used as a robustness judge, classifier, or defense. The author is responsible for the scientific claims, experimental design, results, and final manuscript content.

References

[1] Alkan AK, Sourav S, Jablonska M, et al (2025) A survey on hypothesis generation for scientific discovery in the era of large language models. arXiv preprint arXiv:250405496

[2] Bai Y, Mei J, Yuille AL, et al (2024) Improving adversarial robustness via robust contrastive learning. In: International Conference on Learning Representations

[3] Croce F, Hein M (2020) Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning, pp 2206–2216

[4] Croce F, Andriushchenko M, Sehwag V, et al (2021) RobustBench: A standardized adversarial robustness benchmark. In: Advances in Neural Information Processing Systems Datasets and Benchmarks Track

[5] Dapello J, Marques T, Schrimpf M, et al (2020) Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In: Advances in Neural Information Processing Systems

[6] Geirhos R, Rubisch P, Michaelis C, et al (2019) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations

[7] Gong Y, Wang S, Chen Z, et al (2024) Random entangled tokens for adversarially robust vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

[8] Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: International Conference on Learning Representations

[9] Gowda S, Zonooz B, Arani E (2024) Conserve-update-revise to cure generalization and robustness trade-off in adversarial training. In: International Conference on Learning Representations

[10] Hendrycks D, Mu N, Cubuk ED, et al (2020) AugMix: A simple data processing method to improve robustness and uncertainty. In: International Conference on Learning Representations

[11] Huang H, Erfani SM, Li Y, et al (2025) X-transfer attacks: Towards super transferable adversarial attacks on clip. In: Proceedings of the 42nd International Conference on Machine Learning, pp 25204–25234

[12] Ilyas A, Santurkar S, Tsipras D, et al (2019) Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems

[13] Kim H, Lee J, Park W, et al (2023) The fantastic robustness measures and where to find them. In: International Conference on Learning Representations

[14] Li L, Guan H, Qiu J, et al (2024) One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

[15] Li Q, Li Y, Chen J, et al (2024) Focus on hiders: Exploring hidden threats for enhancing adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

[16] Lin G, Li C, Zhang J, et al (2024) Adversarial training on purification: Advancing both robustness and generalization. In: International Conference on Learning Representations

[17] Lin WA, Chen JH, Chen CH, et al (2023) Eliminating catastrophic overfitting via abnormal adversarial examples regularization. In: Advances in Neural Information Processing Systems

[18] Losch M, Omran M, Stutz D, et al (2024) On adversarial training without perturbing all examples. In: International Conference on Learning Representations

[19] Madry A, Makelov A, Schmidt L, et al (2018) Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations

[20] Mei H, Dong M, Xu C (2025) Efficient image-to-image diffusion classifier for adversarial robustness. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 6081–6089, https://doi.org/10.1609/aaai.v39i6.32650

[21] Perez E, Huang S, Song F, et al (2022) Red teaming language models with language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp 3419–3448

[22] Qi X, Huang K, Panda A, et al (2023) Visual adversarial examples jailbreak aligned large language models. arXiv preprint arXiv:230613213

[23] Robey A, Latorre F, Pappas GJ, et al (2024) Adversarial training should be cast as a non-zero-sum game. In: International Conference on Learning Representations

[24] Sitawarin C, Chang J, Huang D, et al (2024) PubDef: Defending against transfer attacks from public models. In: International Conference on Learning Representations

[25] Sow D, Lyu L, Qi S, et al (2024) Doubly robust instance-reweighted adversarial training. In: International Conference on Learning Representations

[26] Szegedy C, Zaremba W, Sutskever I, et al (2014) Intriguing properties of neural networks. In: International Conference on Learning Representations

[27] Tang Y, Zhang W, Liu X, et al (2024) Ensemble diversity facilitates adversarial transferability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

[28] Villani F, Maljkovic I, Lazzaro D, et al (2025) Robust image classification with multi-modal large language models. Pattern Recognition Letters 194:1–7. https://doi.org/10.1016/j.patrec.2025.04.022

[29] Wang B, Yin P, Xu T, et al (2024) Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

[30] Wang D, Shelhamer E, Liu S, et al (2021) Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations

[31] Wang Y, Zou D, Yi J, et al (2020) Improving adversarial robustness requires revisiting misclassified examples. In: International Conference on Learning Representations

[32] Wu B, Dai D, Zhang P, et al (2024) Robust NAS under adversarial training: Benchmark, theory, and beyond. In: International Conference on Learning Representations

[33] Wu D, Xia ST, Wang Y (2020) Adversarial weight perturbation helps robust generalization. In: Advances in Neural Information Processing Systems

[34] Wu TH, Su HT, Chen ST, et al (2025) Revisiting semi-supervised adversarial training via noise-aware online robust distillation. In: ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models

[35] Xie P, Bie Y, Mao J, et al (2025) Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14679–14689, https://doi.org/10.1109/CVPR52734.2025.01394

[36] Xie Y, Zheng W, Huang H, et al (2025) Towards million-scale adversarial robustness evaluation with stronger individual attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 30702–30711, https://doi.org/10.1109/CVPR52734.2025.02899

[37] Xu Y, Zhang J, Zhang Q, et al (2023) Robust and efficient vision transformer via self-supervised frequency-domain pre-training. In: Advances in Neural Information Processing Systems

[38] Yue X, Wang Y, Huang H, et al (2024) Revisiting adversarial training under long-tailed distributions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

[39] Zhang H, Yu Y, Jiao J, et al (2019) Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning, pp 7472–7482

[40] Zhang M, Levine S, Finn C (2022) MEMO: Test time robustness via adaptation and augmentation. In: Advances in Neural Information Processing Systems

[41] Zhou Y, Hu X, Han J, et al (2021) High frequency patterns play a key role in the generation of adversarial examples. Neurocomputing 459:131–141. https://doi.org/10.1016/j.neucom.2021.06.078

[42] Zou A, Wang Z, Carlini N, et al (2023) Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:230715043