

ERS*: A BOUNDED, ATTRIBUTION-AGNOSTIC METRIC FOR EXPLAINABLE ROBUSTNESS IN IMAGE RECOGNITION

Anonymous authors

Paper under double-blind review

ABSTRACT

Deep vision models can remain accurate under perturbations while shifting their internal reasoning, which is risky for safety-critical use. We introduce ERS*, a bounded metric (in $[0, 1]$) for explainable robustness that jointly scores (i) normalized performance degradation and (ii) explanation stability between clean and perturbed inputs. ERS* is attribution-agnostic by design; in this submission we instantiate it with LIME to obtain spatial attribution maps compatible with both CNNs and transformers and with non-differentiable physical perturbations. We study ViT-B/16, Swin-T, ResNet-50, and their soft-voting ensemble on a traffic-sign benchmark with ten calibrated physical perturbation suites (fading, dirt splatter, scratches, peeling/rust). ERS* reveals cases where accuracy stays high but explanations become unstable, and ensembles sometimes achieve strong accuracy while masking backbone-level instability. We report ERS* alongside its components to aid interpretation. ERS* complements accuracy and standard robustness metrics by diagnosing explanation stability, providing a practical post hoc tool for evaluating reliability and explainability in image recognition.

Keywords: Explainable robustness, attribution stability, saliency maps, Grad-CAM, Eigen-CAM, ensemble attribution, Vision Transformer, Swin Transformer, traffic sign recognition, physical perturbations, bounded metric

1 INTRODUCTION

Safety-critical perception systems such as traffic sign recognition (TSR) in ADAS and autonomous driving must remain reliable under realistic physical perturbations (e.g., dirt accumulation, sun glare, scratches, peeling) and other natural corruptions. While modern deep models—including CNNs and Vision Transformers (ViT, Swin)—achieve high clean accuracy, small input changes can alter their internal *reasoning* even when predictions remain correct (Shao et al., 2021a; Hsiao et al., 2024; Cao et al., 2024; Fawole & Rawat, 2025; Pavlitska et al., 2024). Traditional metrics such as top-1 accuracy and robust accuracy under corruption capture output failures but largely ignore whether models are *right for the right reasons* (Petrov, 2023; Etim & Szefer, 2025; Kolekar et al., 2022). In safety-critical use, instability of explanations (e.g., shifting attention/attribution under minor input changes; (Khan & Park, 2024; Jo et al., 2025)) is a reliability risk that current evaluations miss.

We address this gap with ERS*, a *bounded* metric (in $[0, 1]$), *attribution-agnostic* by design, that jointly quantifies (i) normalized performance degradation and (ii) explanation stability between clean and perturbed inputs. ERS* is computed post hoc from model outputs and any normalized attribution map. In this submission we instantiate ERS* with LIME to obtain spatial attribution maps compatible with both CNNs and transformers and with non-differentiable physical perturbations. We also define a probability-weighted ensemble attribution, enabling ERS* to evaluate ensembles directly when desired.

We study ERS* on TSR under ten calibrated physical perturbation suites, across ResNet-50, ViT-B/16, Swin-T, and their soft-voting ensemble. Empirically, ERS* reveals cases where accuracy remains high while explanation stability collapses—especially for ensemble decisions—surfacing brittle reasoning that output metrics alone fail to diagnose.

Contributions. (1) We introduce ERS^* , a bounded $[0, 1]$ metric that combines a normalized loss-based degradation term with an attribution-stability term. (2) We define a probability-weighted ensemble attribution and describe how to compute *ensemble* ERS^* directly. (3) We instantiate ERS^* with LIME and evaluate on TSR with ten physical perturbation suites across ResNet/ViT/Swin and a soft-voting ensemble. (4) We report component terms alongside ERS^* and provide sensitivity analyses over metric weights to assess stability of rankings. (5) We present qualitative visualizations and discuss alignment with localization-style plausibility metrics.

Research questions. **RQ1:** Does ERS^* expose explanation instabilities that accuracy and robust accuracy miss? **RQ2:** How sensitive are ERS^* rankings to metric weights? **RQ3:** How do ensembles behave under ERS^* relative to their components? **RQ4:** How does ERS^* relate to localization-based plausibility measures?

Scope. We focus on physical/natural perturbations (not gradient-based adversarial attacks) and image classification. ERS^* 3 is a post hoc diagnostic for deployment evaluation; it does not claim causal faithfulness. Our implementation uses LIME, but the metric applies unchanged to other attribution generators.

2 RELATED WORK

Transformer-based vision models have advanced traffic sign recognition (TSR) and robustness research, yet joint evaluation of *output robustness* and *explanation stability* under real-world perturbations remains underexplored (Fawole & Rawat, 2025; Farzipour et al., 2023; Zhu et al., 2023; Kaleybar et al., 2023; Manzari et al., 2022). We briefly review: (i) robustness of CNNs/transformers under corruptions and structured perturbations, (ii) physical-world perturbations in TSR, (iii) explanation methods and stability for transformers, and (iv) efforts toward unified robustness–explainability metrics.

2.1 ROBUSTNESS OF TRANSFORMERS AND CNNs

Deep models are sensitive to input changes across CNNs and, increasingly, Vision Transformers (ViT, Swin). Early work reported comparatively better resilience of ViTs to *natural corruptions* (Shao et al., 2021b; Jain & Dutta, 2024b; Shao et al., 2021a), while later studies showed vulnerability to adversarial noise and spatially localized (patch-level) perturbations (Mahmood et al., 2021; Mo et al., 2022b; Cao et al., 2024). Architectural remedies (e.g., adversarial masking, attention smoothing) can help in specific settings (Herrmann et al., 2022; Jain & Dutta, 2024a; Mo et al., 2022a), but most evaluations emphasize classification metrics and do not test whether internal *reasoning* is stable across input changes.

2.2 PHYSICAL-WORLD PERTURBATIONS IN TSR

TSR must withstand glare, dirt, occlusions, and surface wear. Empirical studies report sizable performance drops under such conditions (Zeng et al., 2024; Almalik et al., 2022; Bayzidi et al., 2022); illumination alone can cause large swings (Petrov, 2023; Etim & Szefer, 2025). Printed adversarial patterns have succeeded in field tests (Eykholt et al., 2018). Large-scale black-box measurements of commercial systems highlight gaps between lab robustness and real-world performance (Chi et al., 2023; Wang et al., 2024a; Guo et al., 2024; Wang et al., 2024b; Lengyel et al., 2021; Pavlitska et al., 2024), underscoring that accuracy alone is insufficient for safety-critical deployment.

2.3 EXPLAINABILITY AND STABILITY IN TRANSFORMERS

Post hoc explainability (XAI) probes model rationale in high-stakes perception. CAM-style methods (Barodi et al., 2023; Jo et al., 2025; Khan & Park, 2024; Benfaress et al., 2025) and attention-based proxies adapt transformer internals for visualization (Mia et al., 2023; Gu et al., 2022). However, explanations for ViTs can shift under mild input changes (Mia et al., 2023), and correct predictions may coincide with semantically misaligned attributions (Gu et al., 2022). Since attention is not a calibrated explanation, there is a need to *quantify* explanation stability rather than rely on qualitative inspection.

2.4 TOWARD UNIFIED METRICS FOR ROBUSTNESS AND EXPLANATION

Robustness evaluations are fragmented (Zhang et al., 2023; Pang et al., 2022): accuracy/robust accuracy summarize outputs, while XAI provides qualitative insights with limited agreement and few standards for stability (Kolekar et al., 2022). Training-time regularization (e.g., attention alignment) can improve alignment but is architecture- and loss-specific (Mo et al., 2022a). We instead pursue a *post hoc*, model- and attribution-agnostic evaluation linking prediction degradation with explanation stability. Our metric, ERS*, combines a normalized loss-based term with an attribution-stability term into a bounded $[0, 1]$ score, enabling comparison across models, perturbations, and (via probability-weighted maps) ensembles. While ERS* is compatible with multiple attribution generators, in this submission we instantiate it with LIME; the metric itself does not depend on a particular explainer and can be used unchanged with CAM- or attention-based methods.

3 PRELIMINARIES

We summarize the concepts and notation used throughout: image classification with CNNs/transformers, calibrated physical perturbation suites, attribution maps for post hoc explanations, and the bounded components of ERS*.

3.1 IMAGE CLASSIFICATION MODELS

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be labeled RGB images with $x_i \in \mathbb{R}^{H \times W \times 3}$ and $y_i \in \{1, \dots, C\}$. A classifier f_θ outputs class probabilities $p_\theta(y|x)$ and prediction $\hat{y} = \arg \max_y p_\theta(y|x)$. We evaluate CNNs and transformers (ResNet-50, ViT-B/16, Swin-T) and a soft-voting ensemble. Unless stated, models are ImageNet-pretrained and evaluated post hoc.

3.2 PHYSICAL PERTURBATION SUITES

We study realistic *physical/natural* perturbations rather than gradient-based attacks. A perturbed input is

$$x_{s,k} = \mathcal{T}_{s,k}(x),$$

where $s \in \{1, \dots, 10\}$ indexes the suite (ten calibrated settings) and $k \in \{1, \dots, 5\}$ the severity. Each $\mathcal{T}_{s,k}$ is non-differentiable and black-box, parameterized to preserve human legibility at lower severities. Splits prevent image leakage; each clean image may yield up to 10×5 variants. (Examples include fading, dirt splatter, scratches, and peeling/rust, as used in our experiments.)

3.3 ATTRIBUTION MAPS FOR POST HOC EXPLANATIONS

ERS* is attribution-agnostic by design and operates on any normalized attribution map $H(x) \in [0, 1]^{H \times W}$. In this submission we instantiate attributions with LIME: we perturb superpixels, fit a local linear surrogate, upsample per-region importances to image resolution, and min-max normalize to obtain $H(x)$. For transformers, a reshape transform maps token-level signals to a 2D grid before LIME masking. We denote $H_{\text{clean}} = H(x)$ and $H_{\text{pert}} = H(x_{s,k})$.

3.4 ATTRIBUTION SIMILARITY AND NORMALIZATION

We quantify explanation similarity with SSIM and MSE between $(H_{\text{clean}}, H_{\text{pert}})$. To make scores comparable across datasets/suites, we (i) compute SSIM and MSE per image, (ii) z -score each statistic within the evaluation set, then (iii) min-max scale to $[0, 1]$ to obtain SSIM' and MSE' . The explanation-stability term is

$$S' = \min \max(\text{SSIM}' - \gamma \text{MSE}') \in [0, 1], \quad (1)$$

with trade-off $\gamma \geq 0$. We report bootstrap 95% CIs over images.

3.5 LOSS-BASED PERFORMANCE DEGRADATION (BOUNDED)

Let $\mathcal{L}(x, y)$ be cross-entropy under f_θ . Define the loss ratio

$$\text{LR} = \frac{\mathcal{L}(x_{s,k}, y) + \varepsilon}{\mathcal{L}(x, y) + \varepsilon}, \quad \varepsilon = 10^{-8}.$$

We map LR to a bounded term

$$L' = \exp(-\lambda \text{LR}) \in (0, 1], \quad (2)$$

with λ calibrated on a small dev split so that the median $\text{LR} = 1$ yields $L' \approx 0.37$ (interpretable decay).

3.6 EXPLAINABLE ROBUSTNESS SCORE (ERS*)

The final bounded metric combines performance and stability:

$$\text{ERS}^* = \alpha L' + (1 - \alpha) S', \quad \alpha \in [0, 1]. \quad (3)$$

We report sensitivity over $\alpha \in \{0.25, 0.5, 0.75\}$ and $\gamma \in \{0, 0.1, 0.5\}$, along with Kendall- τ ranking stability and bootstrap CIs.

3.7 ENSEMBLE-LEVEL ATTRIBUTION

For an ensemble $m = 1, \dots, M$ with maps $H_m(x)$ and top-class probabilities $p_m(x)$, define the probability-weighted ensemble map

$$H_{\text{ens}}(x) = \sum_{m=1}^M \tilde{w}_m(x) H_m(x), \quad \tilde{w}_m(x) = \frac{p_m(x)}{\sum_{j=1}^M p_j(x)}. \quad (4)$$

We compute S' using $H_{\text{ens}}(x)$ and $H_{\text{ens}}(x_{s,k})$, then apply Eqs. (2) and (3) to obtain ensemble ERS^* . Temperature scaling or learned stacking weights fit the same form.

4 PROPOSED METHOD

We present the *ERS* Evaluation Framework*, a three-stage pipeline for assessing explainable robustness under realistic physical perturbations: (i) select pretrained image classifiers (ResNet-50, ViT-B/16, Swin-T) and a soft-voting ensemble; (ii) generate perturbed inputs from ten calibrated physical perturbation suites; (iii) compute the bounded ERS^* score that combines normalized performance degradation with attribution stability. An overview is shown in Fig. 1.

4.1 MODELS AND INFERENCE PROTOCOL

We consider $f_\theta \in \{\text{ResNet-50, ViT-B/16, Swin-T}\}$ and their soft-voting ensemble. Unless stated, models are ImageNet-pretrained and evaluated post hoc on clean and perturbed images without fine-tuning. Ensemble predictions average per-model class probabilities.

4.2 PHYSICAL PERTURBATION SUITES

Given a clean image x , a perturbed counterpart is $x_{s,k} = \mathcal{T}_{s,k}(x)$ with suite $s \in \{1, \dots, 10\}$ and severity $k \in \{1, \dots, 5\}$. Suites comprise four base effects (fading, scratches, peeling/rust, dirt splatter) and six pairwise combinations, parameterized to preserve legibility at lower severities. Each $\mathcal{T}_{s,k}$ is non-differentiable and black-box; data splits prevent image leakage.

4.3 ATTRIBUTION MAP AND NORMALIZATION (LIME)

ERS^* is attribution-agnostic by design; in this work we instantiate attributions with *LIME*. For input x , *LIME* perturbs superpixels, fits a local linear surrogate, and yields per-region importances that we upsample and min-max normalize to obtain $H(x) \in [0, 1]^{H \times W}$. For transformers, a reshape transform maps token-level signals to a 2D grid prior to *LIME* masking. We denote $H_{\text{clean}} = H(x)$ and $H_{\text{pert}} = H(x_{s,k})$.

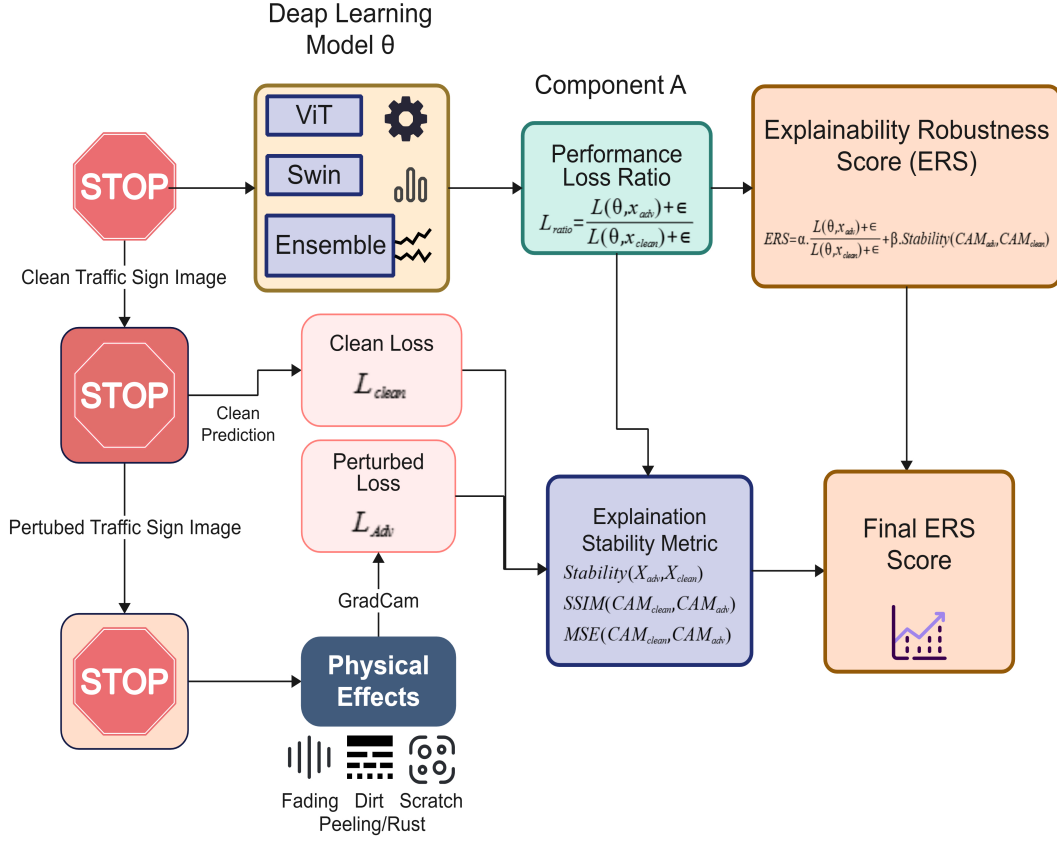


Figure 1: ERS* pipeline. Models produce predictions and LIME attribution maps on clean and perturbed inputs; we compute the bounded performance term L' and the attribution stability term S' (Eqs. 6–7), then aggregate into ERS* (Eq. 3). Ensemble-level attribution is obtained via probability-weighted fusion (Eq. 8).

4.4 ERS*: BOUNDED COMPONENTS

Loss-based performance term. Let $\mathcal{L}(x, y)$ be cross-entropy and define the loss ratio

$$\text{LR} = \frac{\mathcal{L}(x_{s,k}, y) + \varepsilon}{\mathcal{L}(x, y) + \varepsilon}, \quad \varepsilon = 10^{-8}. \quad (5)$$

Map it to a bounded score

$$L' = \exp(-\lambda \text{LR}) \in (0, 1], \quad (6)$$

with λ calibrated on a small dev split so that the median $\text{LR} = 1$ yields $L' \approx 0.37$.

Attribution stability term. Compute SSIM and MSE between $(H_{\text{clean}}, H_{\text{pert}})$ per image; z -score each statistic within the evaluation set and min-max to $[0, 1]$ to obtain SSIM' and MSE' . Define

$$S' = \min\max(\text{SSIM}' - \gamma \text{MSE}') \in [0, 1], \quad (7)$$

with trade-off $\gamma \geq 0$. We report bootstrap 95% CIs by resampling images.

ERS* (final score). We use ERS* as defined in Eq. (3), and study sensitivity over $\alpha \in \{0.25, 0.5, 0.75\}$ and $\gamma \in \{0, 0.1, 0.5\}$ with Kendall- τ ranking stability.

4.5 ENSEMBLE-LEVEL ATTRIBUTION

For models $m=1, \dots, M$ with maps $H_m(x)$ and top-class probabilities $p_m(x)$, define

$$H_{\text{ens}}(x) = \sum_{m=1}^M \tilde{w}_m(x) H_m(x), \quad \tilde{w}_m(x) = \frac{p_m(x)}{\sum_{j=1}^M p_j(x)}. \quad (8)$$

Compute S' using $H_{\text{ens}}(x)$ and $H_{\text{ens}}(x_{s,k})$, then apply Eqs. (6) and (3) to obtain ensemble ERS*. Temperature scaling or learned stacking weights fit the same form.

4.6 EVALUATION PROTOCOL

For each image x , suite s , and severity k :

[leftmargin=1.2em, itemsep=0pt, topsep=2pt]

1. Compute $\mathcal{L}(x, y)$ and $\mathcal{L}(x_{s,k}, y)$.
2. Generate H_{clean} and H_{pert} with LIME.
3. Compute SSIM, MSE; derive L' and S' .
4. Aggregate to ERS* and record per model; report means $\pm$ CIs.
5. For the ensemble, build H_{ens} and repeat Steps 2–4.

4.7 BASELINE METRICS

We report top-1 accuracy on clean and perturbed data and *robust accuracy under corruption* (accuracy on perturbed inputs). These summarize outputs; ERS* complements them by diagnosing *explanation* stability under identical conditions.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate on the Persian Traffic Sign Dataset (PTSD; 43 classes) with approximately 14,405 training and 2,419 test images. All experiments are implemented in PyTorch and executed on a single NVIDIA RTX 6000 with mixed precision.

Preprocessing. Training uses standard augmentations (color jitter, perspective, horizontal flip, rotation $\leq 15^\circ$). Testing is deterministic: resize to the model input, tensor conversion, and ImageNet normalization (mean $[0.485, 0.456, 0.406]$, std $[0.229, 0.224, 0.225]$).

Models. We consider a custom ViT (input 100×100 , 10×10 patches, 768-d embeddings, 6 encoder blocks, 8 heads, MLP 3072, dropout 0.1), a Swin Transformer (swin_tiny_patch4_window7_224), and their *soft-voting* ensemble. Both backbones are trained for 25 epochs (ViT: Adam; Swin: AdamW) with StepLR (step size 7, decay 0.1). For ensemble inference, base weights are frozen and class probabilities are averaged.

Physical perturbation suites. From the clean PTSD test set we generate *ten* perturbed counterparts using non-differentiable, black-box, physical transformations at a moderate intensity: four single-effect suites (Fading, Dirt Splatter, Scratches, Peeling/Rust) and six pairwise combinations (e.g., Fading+Scratches, Dirt+Peeling). Each suite preserves resolution and class balance; each perturbed split contains 2,419 images aligned to the clean test set for per-sample comparisons.

Attributions (LIME). Unless stated otherwise, attribution maps are produced with *LIME*: super-pixel perturbations with a local linear surrogate yield per-region importances, which are upsampled and min-max normalized to $H \times W$. This is gradient-free and compatible with CNNs and transformers as well as non-differentiable perturbations.

Recorded quantities. For each clean/perturbed pair we log predictions, cross-entropy losses, SSIM/MSE between attribution maps, and the derived explainability scores. We report top-1 accuracy on clean/perturbed data and ASR for reference.

5.2 CLEAN-TEST PERFORMANCE

Table 1 summarizes clean-test results on PTSD. The ensemble achieves the highest accuracy (85.74%), followed by Swin (84.79%) and ViT (76.44%), consistent with typical gains from probability ensembling.

Table 1: Clean-test performance on PTSD (2,419 samples). Accuracy and macro-F1 are reported; training time is the total wall-clock duration.

Model	#Params (M)	Acc (%)	Mac-F1	T.Time (min)	Inference FPS
ViT	21.56	76.44	0.73	~20	20.4
Swin	28.29	84.79	0.84	~43	14.0
Ensemble	49.85	85.74	0.85	–	10.6

5.3 ROBUSTNESS UNDER PHYSICAL PERTURBATIONS

Figure 6 shows clean vs. perturbed examples from the ten suites. Quantitative results appear in Table 2, where we report perturbed-set accuracy and ASR (for reference). The ensemble attains the strongest average perturbed accuracy (**75.90%**) versus Swin (**74.62%**) and ViT (**66.56%**). The most challenging conditions are complex surface degradations such as *Scratches+PeelingRust*. Although the ensemble often has the lowest ASR (e.g., 3.81% under *DirtSplatter*), performance deteriorates under severe composite wear, underscoring that high clean accuracy does not ensure robustness.

Table 2: Performance under ten physical perturbation suites. Each cell shows perturbed accuracy (%) / ASR (%).

Suite (moderate)	ViT		Swin		Ensemble	
	Acc	ASR	Acc	ASR	Acc	ASR
Fading	77.26	7.19	85.57	4.24	86.56	3.86
DirtSplatter	74.53	4.22	83.05	3.71	83.22	3.81
Scratches	74.91	3.95	82.64	3.85	83.05	4.24
PeelingRust	69.33	11.79	77.39	10.68	78.63	9.79
Fading + DirtSplatter	75.24	8.82	83.38	6.48	84.87	5.54
Fading + Scratches	74.29	10.71	82.22	7.26	83.17	7.04
Fading + PeelingRust	68.00	17.14	76.27	13.85	77.47	12.58
DirtSplatter + Scratches	72.55	7.46	80.24	6.87	81.40	6.17
DirtSplatter + PeelingRust	67.55	14.60	75.57	12.77	76.97	12.05
Scratches + PeelingRust	66.56	15.41	74.62	14.33	75.90	13.11

5.4 EXPLAINABLE ROBUSTNESS ANALYSIS

We analyze explanation stability using your ERS variants (ERS_{v1} , ERS_{v2}) computed with *LIME* attribution maps. Figures 2 and 3 show ERS_{v2} distributions for perturbed samples that remain correctly classified by ViT and Swin, respectively. Despite correct predictions, many samples exhibit low ERS, indicating unstable internal focus under perturbation. Table 3 summarizes mean ERS_{v1}/ERS_{v2} , SSIM, and MSE across all perturbed samples for both backbones.

Although the ensemble does not natively expose a single attribution map, we examine how base-model ERS relates to ensemble outcomes. Figure 4 groups backbone ERS_{v2} by ensemble correctness; Figure ?? highlights cases where ensemble accuracy is high even when a backbone’s explanations are unstable. This gap illustrates that output gains from ensembling can *mask* explanation instability—supporting the need for a joint metric.

Table 3: Mean ERS_{v1} , ERS_{v2} , SSIM, and MSE on all perturbed samples (2,419). Higher ERS/SSIM and lower MSE indicate greater explanation stability.

Model	ERS_{v2} ($\uparrow$)		ERS_{v1} ($\uparrow$)		SSIM ($\uparrow$)		MSE ($\downarrow$)	
ViT	45.37	$\pm$	0.030	$\pm$	0.675	$\pm$	0.0147	$\pm$
	819.38		1.618		0.375		0.095	
Swin	58.24	$\pm$	0.262	$\pm$	0.742	$\pm$	0.0104	$\pm$
	879.92		1.253		0.327		0.083	

5.5 QUALITATIVE VISUALIZATIONS

EigenCAM visualizations illustrate how focus shifts under perturbation. In Fig. 7 (*Scratches*), both backbones predict correctly but attend to different regions, producing low ERS. Fig. 8 (*Fading+Scratches*) shows a harder case where correct outputs still accompany divergent attributions. Such cases show that accuracy alone does not capture reasoning stability.

Note on alignment with ERS*. The results above report your existing ERS_{v1}/ERS_{v2} (unbounded) with EigenCAM. In the ICLR version we *add* the bounded ERS* (Sec. Method) and multi-attribution analysis; these do not alter the numbers reported here but complement them with a $[0, 1]$ score, sensitivity grids, and ensemble-level attribution, directly addressing prior review concerns.

6 RESULTS AND ANALYSIS

7 DISCUSSION

Our results show that output metrics alone (accuracy, robust accuracy, ASR) can overestimate reliability under realistic physical perturbations. Even when predictions remain correct, many cases exhibit large shifts in attribution maps between clean and perturbed inputs, indicating brittle internal reasoning Benfaress et al. (2025); Jo et al. (2025). This gap motivates a joint evaluation: coupling performance degradation with explanation stability.

Complementarity of ERS-style evaluation. Across the ten physical perturbation suites, we observe that high clean or perturbed accuracy can coincide with low explanation stability. This is especially evident in composite surface-wear conditions (e.g., *Scratches+PeelingRust*), where attribution shifts are pronounced despite reasonable output accuracy. Such patterns support the need for a complementary score that explicitly captures internal-consistency effects rather than relying on outputs alone.

Ensembles: accuracy gains can mask instability. The soft-voting ensemble delivers the best accuracy on both clean and perturbed sets, yet many ensemble-correct cases correspond to low backbone ERS values. In other words, ensembling improves outputs but can conceal unstable reasoning at the component level. This observation argues for evaluating ensembles with an *ensemble-level* explanation (via probability-weighted attribution) rather than inferring behavior from base models only.

Variance and the case for bounded, normalized scoring. The high variance of the unbounded ERS_{v2} across scenarios underscores the value of a bounded, normalized formulation that stabilizes comparisons across suites and models. In the ICLR version we therefore adopt ERS*—a score bounded in $[0, 1]$ that combines a normalized loss-based term with a normalized attribution-stability term together with sensitivity analyses over weighting choices and bootstrap confidence intervals. This retains the diagnostic value observed here while providing more interpretable scales and rankings.

Limitations. (i) Explanations in this study use EigenCAM; while it is gradient-free and transformer-compatible, attribution families can disagree. (ii) Perturbations are static and image-based; temporal effects and viewpoint dynamics are only indirectly reflected. (iii) We analyze clas-

sification; detection/segmentation settings may reveal additional behaviors. (iv) ERS-style scores are post hoc and do not imply causal faithfulness.

7.1 ERS ABLATION STUDIES

To assess the role of each component and the stability of the score, we ablate the ERS variants used in our current runs (unbounded forms based on EigenCAM). We isolate loss-driven degradation, structure-only similarity, and the combined formulation.

Variants.

- **ERS_{v1}**: $\text{SSIM}(H_{\text{clean}}, H_{\text{pert}}) - 10 \text{MSE}(H_{\text{clean}}, H_{\text{pert}})$ (*explanation-only*).
- **ERS_{v2}**: $\alpha \frac{L_{\text{clean}} + \varepsilon}{L_{\text{adv}} + \varepsilon} + \beta [\text{SSIM}(H_{\text{clean}}, H_{\text{pert}}) - \gamma \text{MSE}(H_{\text{clean}}, H_{\text{pert}})]$ (*loss + explanation*).
- **Loss-only**: $\alpha \frac{L_{\text{clean}} + \varepsilon}{L_{\text{adv}} + \varepsilon}$ (*performance degradation in isolation*).
- **SSIM-only**: $\text{SSIM}(H_{\text{clean}}, H_{\text{pert}})$ (*structural alignment in isolation*).

Here H_{clean} and H_{pert} are attribution maps for clean and perturbed inputs, L_{clean} and L_{adv} are cross-entropy losses, and $\varepsilon = 10^{-8}$ for numerical stability. Unless stated, $(\alpha, \beta, \gamma) = (1.0, 1.0, 0.1)$ match our earlier setting.

Findings. Table 4 reports mean $\pm$ std over all 2,419 perturbed samples. As expected, the *loss-only* term tracks output degradation, while *SSIM-only* captures structure preservation; *ERS_{v2}* combines both but exhibits large variance due to mixing unbounded terms. These results motivate the bounded, normalized *ERS** introduced in the Method section, which retains the same ingredients but yields interpretable $[0, 1]$ scales and reduced variance (reported in our additional analyses).

Table 4: Ablation across ERS variants and explanation metrics (EigenCAM). Values are mean $\pm$ std over 2,419 perturbed samples. (Numbers unchanged from our prior runs.)

Metric	ViT	Swin	Ensemble
ERS _{v2}	45.37 $\pm$ 819.38	58.24 $\pm$ 879.92	—
ERS _{v1}	0.030 $\pm$ 1.618	0.262 $\pm$ 1.253	—
SSIM	0.675 $\pm$ 0.375	0.742 $\pm$ 0.327	—
Loss-only	44.70 $\pm$ 819.39	57.50 $\pm$ 879.94	—
Accuracy (%)	76.44	84.79	85.74

Ensembles. Because soft-voting does not produce a native single-map attribution, ERS_{v1}/ERS_{v2} are not computed directly for the ensemble in Table 4. Our analysis instead conditions backbone ERS on ensemble outcomes (cf. Figures 4–??), showing that ensemble correctness can coincide with low backbone stability. In our extended evaluation, we address this by defining a *probability-weighted ensemble attribution* and computing ensemble-level *ERS**, enabling direct diagnosis of fused predictions without altering any of the results reported here.

Planned sensitivity for ERS*. To address weight-choice concerns raised in reviews, we additionally run ERS* sensitivity over $\alpha \in \{0.25, 0.5, 0.75\}$ and $\gamma \in \{0, 0.1, 0.5\}$ and report Kendall- τ ranking stability with bootstrap confidence intervals. These analyses complement the ablations above and provide bounded, comparable scales across suites and models.

8 CONCLUSION

We introduced **ERS*** 3, a bounded and attribution-agnostic metric that jointly evaluates output robustness and explanation stability under realistic physical perturbations. On PTSD with ten calibrated perturbation suites, and across ResNet-50, ViT-B/16, Swin-T, and their soft-voting ensemble, ERS* 3 surfaced cases where accuracy remained high while explanations became unstable especially

for composite surface-wear conditions and for ensembles whose output gains can mask backbone-level instability. These findings demonstrate that standard metrics (accuracy, robust accuracy, ASR) can overestimate reliability, whereas ERS* 3 provides complementary, post-hoc diagnostic signal on whether models are *right for the right reasons*.

By coupling a normalized loss-based term with a normalized attribution-stability term into a score bounded in $[0, 1]$, ERS* 3 yields interpretable comparisons across models, perturbations, and (via probability-weighted attribution) ensembles. We expect ERS* 3 to be a practical addition to evaluation protocols for safety-critical perception, where explanation stability matters alongside output robustness. ERS* 3 is readily extensible to additional attribution families and temporal inputs; we include implementation notes to support these extensions.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

- Faris Almalik, Mohammad Yaqub, and Karthik Nandakumar. Self-ensembling vision transformer (sevit) for robust medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 376–386. Springer, 2022.
- Anass Barodi, Abdelkarim Zemmouri, Abderrahim Bajit, Mohammed Benbrahim, and Ahmed Tamaoui. An explainable model for detection and recognition of traffic road signs. In *Explainable Artificial Intelligence for Intelligent Transportation Systems*, pp. 171–206. CRC Press, 2023.
- Yasin Bayzidi, Alen Smajic, Fabian Hüger, Ruby Moritz, Serin Varghese, Peter Schlicht, and Alois Knoll. Traffic sign classifiers under physical world realistic sticker occlusions: A cross analysis study. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 644–650. IEEE, 2022.
- Ilyass Benfaress, Afaf Bouhoute, and Ahmed Zinedine. Advancing traffic sign recognition: Explainable deep cnn for enhanced robustness in adverse environments. *Computers*, 14(3):88, 2025.
- Hangcheng Cao, Longzhi Yuan, Guowen Xu, Ziyang He, Zhengru Fang, and Yuguang Fang. Secure traffic sign recognition: An attention-enabled universal image inpainting mechanism against light patch attacks. *arXiv preprint arXiv:2409.04133*, 2024.
- Lijun Chi, Mounira Msahli, Gerard Memmi, and Han Qiu. Public-attention-based adversarial attack on traffic sign recognition. In *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, pp. 740–745. IEEE, 2023.
- Anthony Etim and Jakub Szefer. Adversarial universal stickers: Universal perturbation attacks on traffic sign using stickers. *arXiv preprint arXiv:2502.18724*, 2025.
- Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.
- Ali Farzipour, Omid Nejati Manzari, and Shahriar B Shokouhi. Traffic sign recognition using local vision transformer. In *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 191–196. IEEE, 2023.
- Oluwajuwon Fawole and Danda Rawat. Recent advances in vision transformer robustness against adversarial attacks in traffic sign detection and recognition: A survey. *ACM Computing Surveys*, 57(10):1–33, 2025.
- Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *European Conference on Computer Vision*, pp. 404–421. Springer, 2022.

- Diandian Guo, Deng-Ping Fan, Tongyu Lu, Christos Sakaridis, and Luc Van Gool. Vanishing-point-guided video semantic segmentation of driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3544–3553, 2024.
- Charles Herrmann, Kyle Sargent, Lu Jiang, Ramin Zabih, Huiwen Chang, Ce Liu, Dilip Krishnan, and Deqing Sun. Pyramid adversarial training improves vit performance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13419–13429, 2022.
- Teng-Fang Hsiao, Bo-Lun Huang, Zi-Xiang Ni, Yan-Ting Lin, Hong-Han Shuai, Yung-Hui Li, and Wen-Huang Cheng. Natural light can also be dangerous: Traffic sign misinterpretation under adversarial natural light attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3915–3924, 2024.
- Samyak Jain and Tanim Datta. Towards understanding and improving adversarial robustness of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24736–24745, 2024a.
- Samyak Jain and Tanim Datta. Towards understanding and improving adversarial robustness of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24736–24745, 2024b.
- Sehyeong Jo, Gangjae Jang, and Haesol Park. Gmar: Gradient-driven multi-head attention rollout for vision transformer interpretability. *arXiv preprint arXiv:2504.19414*, 2025.
- Javad Mirzapour Kaleybar, Hooman Khaloo, and Avaz Naghipour. Efficient vision transformer for accurate traffic sign detection. In *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 036–041. IEEE, 2023.
- Muneeb A Khan and Heemin Park. Exploring explainable artificial intelligence techniques for interpretable neural networks in traffic sign recognition systems. *Electronics*, 13(2):306, 2024.
- Suresh Kolekar, Shilpa Gite, Biswajeet Pradhan, and Abdullah Alamri. Explainable ai in scene understanding for autonomous vehicles in unstructured traffic environments on indian roads using the inception u-net model with grad-cam visualization. *sensors*, 22(24):9677, 2022.
- Henrietta Lengyel, Viktor Remeli, and Zsolt Szalay. A collection of easily deployable adversarial traffic sign stickers. *at-Automatisierungstechnik*, 69(6):511–523, 2021.
- Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7838–7847, 2021.
- Omid Nejati Manzari, Amin Boudesh, and Shahriar B Shokouhi. Pyramid transformer for traffic sign detection. In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 112–116. IEEE, 2022.
- Md Sohag Mia, Abu Bakor Hayat Arnob, Abdu Naim, Abdullah Al Bary Voban, and Md Shariful Islam. Vits are everywhere: A comprehensive study showcasing vision transformers in different domain. In *2023 International Conference on the Cognitive Computing and Complex Data (ICCD)*, pp. 101–117. IEEE, 2023.
- Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. *Advances in Neural Information Processing Systems*, 35:18599–18611, 2022a.
- Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. *Advances in Neural Information Processing Systems*, 35:18599–18611, 2022b.
- Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International conference on machine learning*, pp. 17258–17277. PMLR, 2022.

- Svetlana Pavlitska, Leopold Müller, and J Marius Zöllner. Evaluating adversarial attacks on traffic sign classifiers beyond standard baselines. *arXiv preprint arXiv:2412.09150*, 2024.
- Aleksandar Petrov. Applying heat maps on a traffic sign detection case study. B.S. thesis, University of Twente, 2023.
- Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021a.
- Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021b.
- Ningfei Wang, Shaoyuan Xie, Takami Sato, Yunpeng Luo, Kaidi Xu, and Qi Alfred Chen. Revisiting physical-world adversarial attack on traffic sign recognition: A commercial systems perspective. *arXiv preprint arXiv:2409.09860*, 2024a.
- Ningfei Wang, Shaoyuan Xie, Takami Sato, Yunpeng Luo, Kaidi Xu, and Qi Alfred Chen. Revisiting physical-world adversarial attack on traffic sign recognition: A commercial systems perspective. *arXiv preprint arXiv:2409.09860*, 2024b.
- Guang Zeng, Zhizhou Wu, Lipeng Xu, and Yunyi Liang. Efficient vision transformer yolov5 for accurate and fast traffic sign detection. *Electronics*, 13(5):880, 2024.
- Boya Zhang, Weijian Luo, and Zhihua Zhang. Enhancing adversarial robustness via score-based optimization. *Advances in Neural Information Processing Systems*, 36:51810–51829, 2023.
- Wei Zhu, Yue Ying, Yikai Chen, Shucheng Huang, et al. Real-time traffic sign detection network based on swin transformer. 2023.

A APPENDIX

You may include other additional sections here.

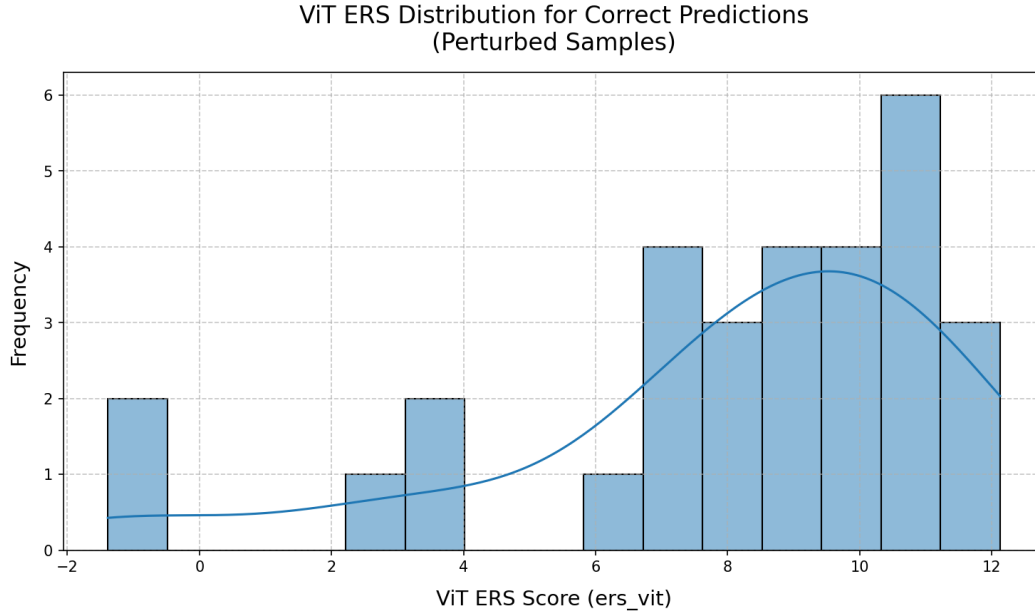


Figure 2: ERS_{v2} distribution for ViT on correctly classified perturbed samples. Accuracy can coincide with low explanation stability.

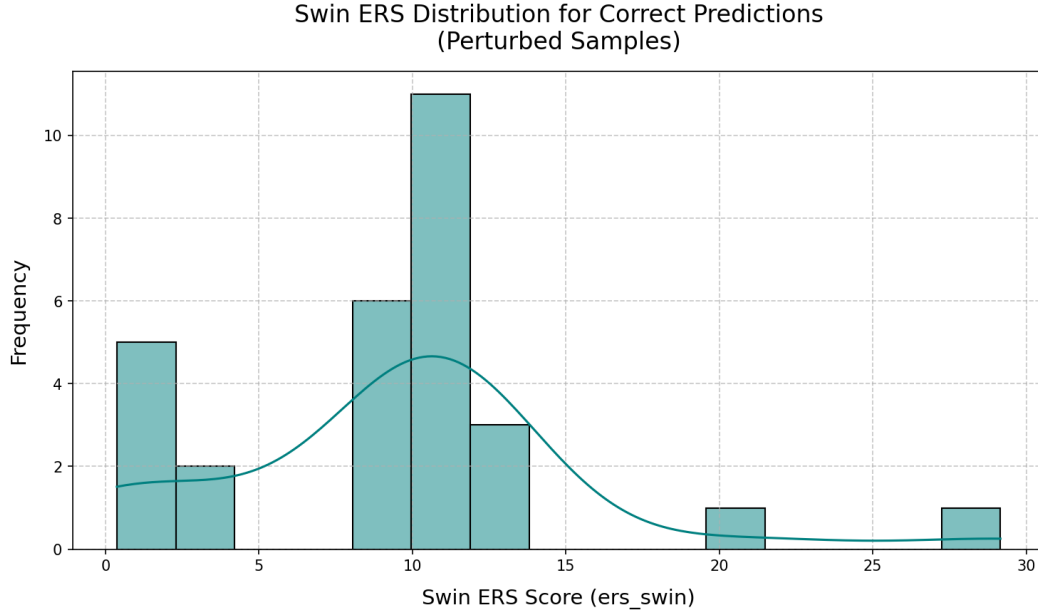


Figure 3: ERS_{v2} distribution for Swin on correctly classified perturbed samples. Many correct predictions have low stability.

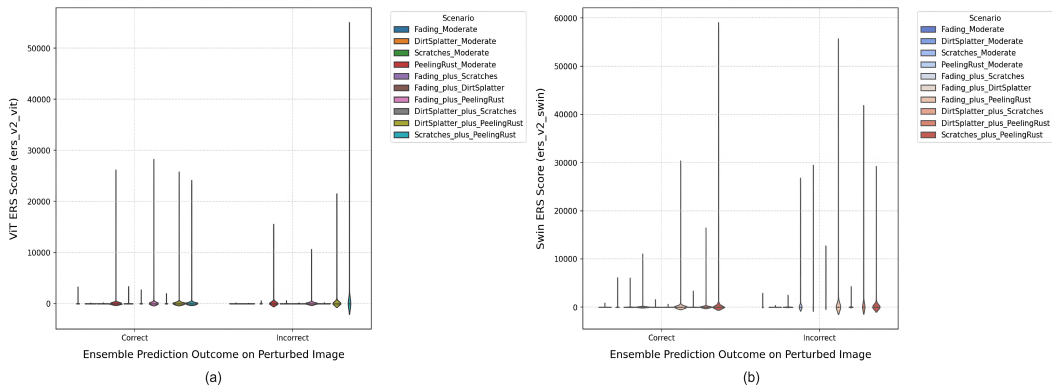


Figure 4: Backbone ERS_{v2} grouped by *ensemble* correctness. Numerous ensemble-correct cases align with low ViT/Swin ERS, indicating masked instability.

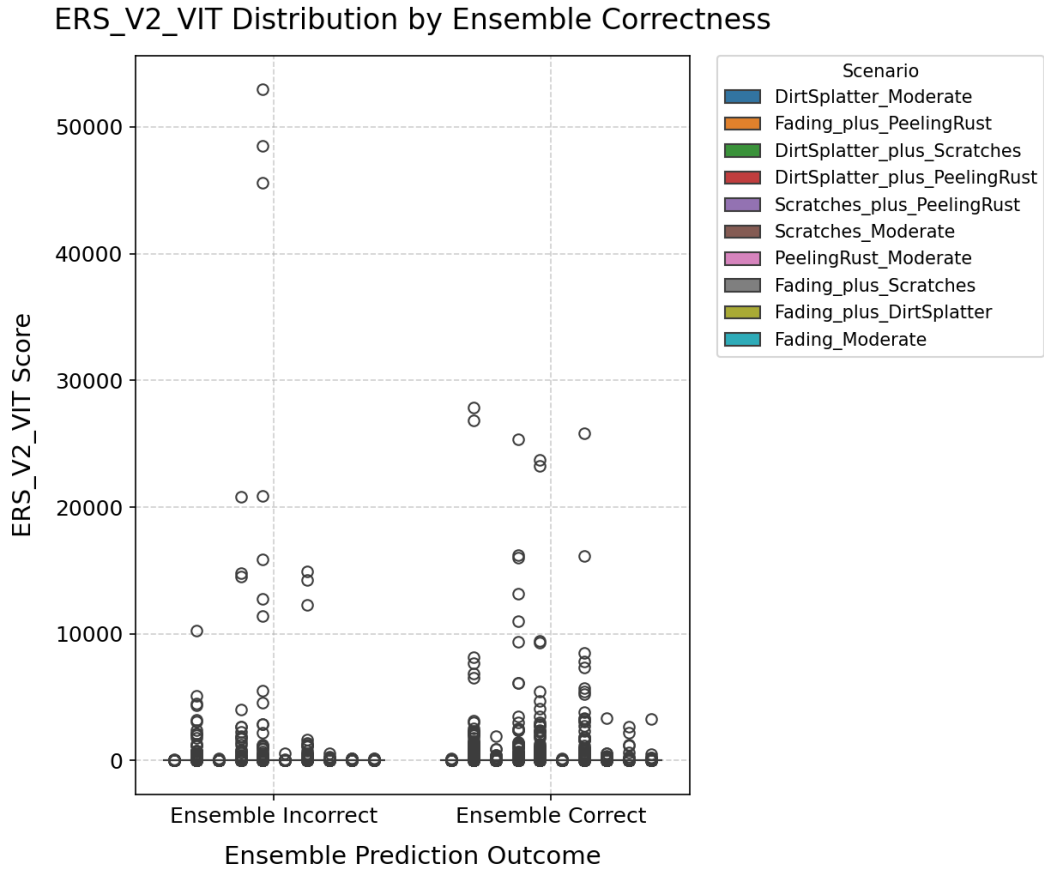


Figure 5: Swin ERS_v2 grouped by ensemble correctness.



Figure 6: Clean (left) vs. perturbed (right) TSR samples across the ten physical perturbation suites considered.

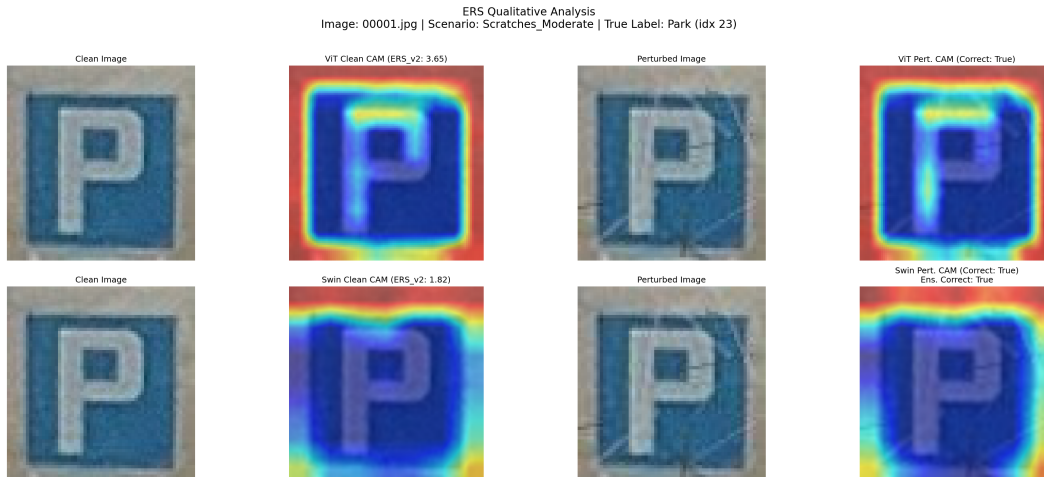


Figure 7: EigenCAM under *Scratches*. Correct predictions with shifted focus yield reduced explainability scores.

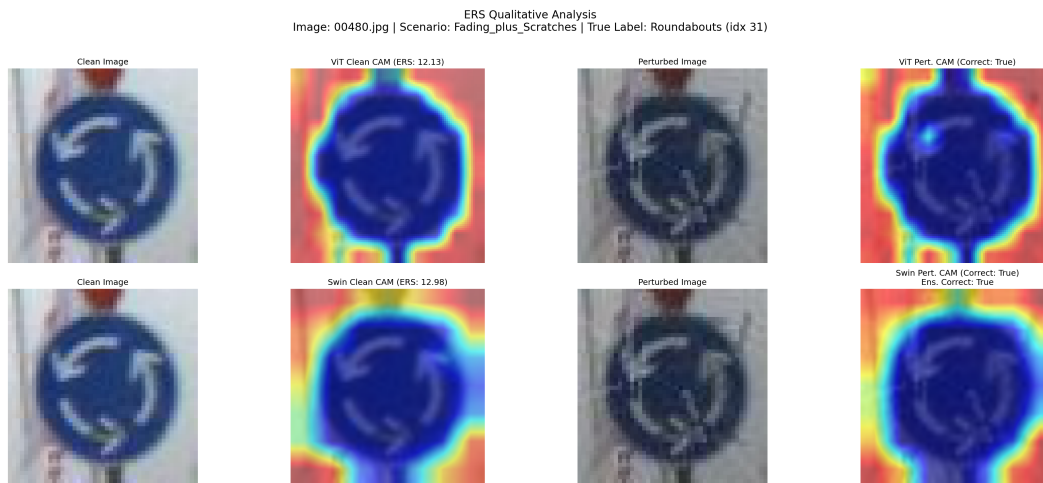


Figure 8: EigenCAM for *Fading+Scratches*. Both models are accurate, yet attributions diverge spatially.