PROTOTYPE MARGINS AND CLASSIFIER DECISIONS: A GEOMETRIC ANALYSIS OF DEEPFAKE ATTRIBUTION ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deepfake attribution models are vulnerable to image degradations, but whether prototype geometry reliably reflects classifier decisions remains unclear. We introduce **Attribution Margin Collapse (AMC)**, a diagnostic framework that examines class-center displacement and compares prototype margins with classifier logit margins. Two complementary statistics, AMC-S and AMC-F, summarize the mean prototype margin and the dataset-wide magnitude of negative prototype margins, respectively. Experiments on FaceForensics++ across four attribution models document robustness degradation, while JPEG analysis reveals similar aggregate trends in prototype and classifier margins. A complementary cross-domain study of binary deepfake detection on Celeb-DF reveals substantial disagreement: samples can retain positive prototype margins despite incorrect classifier predictions. These findings highlight both the utility and the limitations of prototype geometry for robustness analysis. AMC provides a framework for examining when geometric diagnostics reflect classifier behavior and when they offer an incomplete account of failure.

1 INTRODUCTION

Deepfake attribution identifies the manipulation method or generation source of forged content, extending binary real-versus-fake detection toward source-level forensic analysis Amerini et al. (2024); Jia et al. (2022). Despite strong performance on clean data, attribution models remain vulnerable to image degradations encountered during storage and transmission Liu et al. (2024); Baxevanakis et al. (2025). In our experiments, EfficientNetV2 achieves 93.0% accuracy on clean FaceForensics++ images, but only 63.3% under JPEG50 compression and 20.9% under Gaussian blur. Understanding these failures requires examining not only prediction accuracy but also changes in the representations on which predictions depend.

Existing analyses provide complementary perspectives on robustness. Perceptual metrics quantify image distortion, Grad-CAM localizes regions relevant to predictions, and calibration analysis examines the reliability of predictive confidence Zhang et al. (2018); Wang et al. (2004); Selvaraju et al. (2017); Guo et al. (2017). Prior attribution studies also investigate robustness under image transformations Yu et al. (2021). These perspectives motivate a more specific question: *when do changes in prototype-based feature geometry reflect changes in the decisions of an attribution classifier?*

Class prototypes offer a simple reference for studying representation changes. A sample's relative distances to its ground-truth prototype and competing prototypes define a geometric margin. However, this margin describes a nearest-prototype rule, which generally differs from the model's learned classification rule. A positive prototype margin therefore does not guarantee a correct classifier prediction, nor does a negative margin necessarily imply misclassification. Establishing the relationship between these two rules is essential before interpreting prototype geometry as an explanation of attribution failure.

We introduce **Attribution Margin Collapse (AMC)**, a diagnostic framework that examines class-center displacement and prototype-margin distributions alongside classifier logit margins. Two aggregate statistics, **AMC-S** and **AMC-F**, summarize the mean signed prototype margin and the av-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

erage negative-margin magnitude over the full evaluation set. These statistics describe prototype geometry; their relationship to classifier behavior is evaluated empirically rather than assumed.

Our experiments document robustness degradation across four attribution models. Under increasing JPEG compression, prototype and classifier margins exhibit similar aggregate trends. A complementary cross-domain study of binary deepfake detection on Celeb-DF reveals a contrasting pattern: positive prototype margins can coexist with incorrect classifier predictions. This comparison highlights the need to distinguish geometric diagnostics from actual decisions and to establish their correspondence within each evaluation setting.

Our contributions are threefold:

- We develop a diagnostic framework that explicitly distinguishes prototype margins from classifier margins when analyzing deepfake attribution robustness.
- We introduce AMC-S and AMC-F to summarize prototype-margin distributions, jointly describing average separation and the prevalence and magnitude of negative margins.
- We empirically examine the correspondence between prototype geometry and classifier behavior through five-class attribution experiments and a complementary binary cross-domain detection study, identifying both shared trends and disagreements.

2 Related Work

2.1 Deepfake Attribution and Robustness

Deepfake detection and multimedia forensics have been extensively studied in recent years Verdoliva (2020). Early research mainly focused on binary detection, while recent efforts have shifted toward *deepfake attribution*, which aims to identify the manipulation method or source model responsible for generating forged content Amerini et al. (2024); Yu et al. (2021); Jia et al. (2022). Existing approaches have explored forgery-specific artifacts Li et al. (2020), attribution techniques Amerini et al. (2019), and cross-domain generalization Baxevanakis et al. (2025); Shiohara & Yamasaki (2022). Attribution robustness under common image degradations has been documented Rossler et al. (2019); Wang et al. (2024), and robustness has been studied from both empirical and theoretical perspectives Meng et al. (2022). However, most prior work reports performance degradation rather than the geometric mechanism underlying attribution failure Zeiler & Fergus (2014); Nguyen-Le et al. (2024).

2.2 Feature Geometry and Explainability

Explainability methods such as frequency-domain analysis Frank et al. (2020); Dutta et al. (2025), perceptual similarity metrics, and Grad-CAM Selvaraju et al. (2017); Tjoa & Guan (2020) describe how inputs change but not how decisions fail.

Beyond explainability, the geometric structure of feature spaces has been extensively studied in metric learning, where distance functions are learned to pull semantically similar samples together and push dissimilar samples apart Bellet et al. (2013). Contrastive learning frameworks such as SimCLR further show that effective visual representations can be learned by maximizing agreement between augmented views in feature space Chen et al. (2020). These works establish that feature-space geometry encodes rich semantic information. However, what remains missing is an explanation framework that directly models the geometric evolution of feature representations. AMC fills this gap by analyzing how perturbations and domain shifts alter the learned geometry Cortes & Vapnik (1995); Kou et al. (2025).

3 Method

3.1 AMC Framework Overview

We investigate deepfake attribution robustness by examining how visual perturbations affect feature geometry and classifier decisions. Our framework, **Attribution Margin Collapse (AMC)**, uses

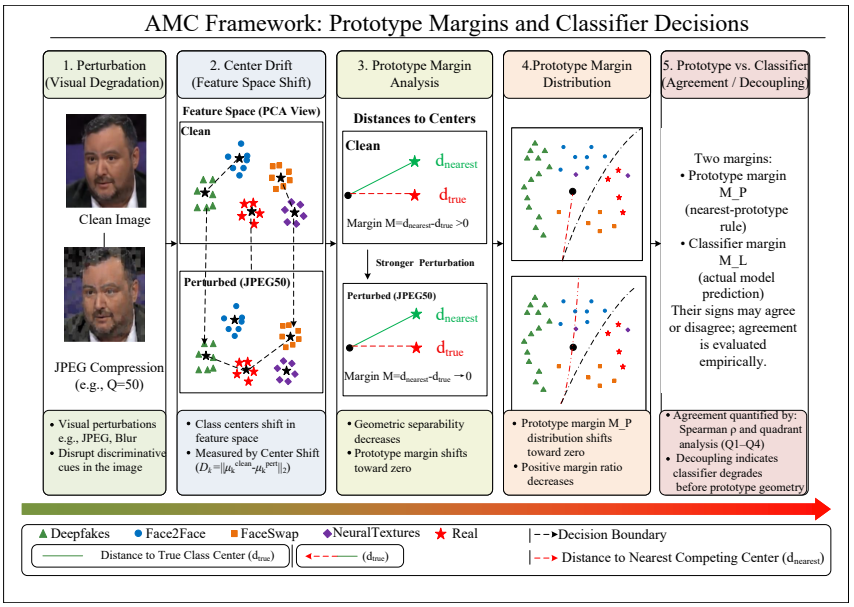


Figure 1: Overview of the proposed Attribution Margin Collapse (AMC) framework. Visual perturbations progressively distort the learned feature space, following the five-stage evolution: Perturbation → Feature Center Drift → Margin Collapse → Decision Boundary Crossing → Attribution Failure.

clean class prototypes as a fixed reference to characterize changes in learned representations and assesses whether these geometric changes correspond to changes in the classifier's predictions.

The analysis distinguishes two quantities. The *prototype margin* measures a sample's relative proximity to its ground-truth class prototype and the nearest competing prototype. The *classifier margin* measures the difference between the ground-truth class logit and the largest competing logit. These margins describe different decision rules: a negative prototype margin indicates an incorrect nearest-prototype assignment, whereas a negative classifier margin indicates an incorrect prediction by the original model. Their signs therefore need not agree.

AMC combines class-center displacement, sample-level prototype margins, and the aggregate statistics AMC-S and AMC-F to characterize perturbation-induced changes in prototype geometry. We then compare prototype and classifier margins to examine their agreement and disagreement across evaluation conditions. Margin collapse is treated as an empirical pattern to be evaluated, rather than a necessary consequence of feature-center drift. This formulation allows us to assess when prototype geometry reflects classifier behavior and when it provides an incomplete account of attribution failure.

3.2 Feature Space and Center Drift

Feature embeddings are L2-normalized for scale consistency. Given an input image $\mathbf{x}$, the network extracts

$$\mathbf{z} = \phi(\mathbf{x}) \in \mathbb{R}^D.$$

Each class is represented by a clean reference prototype

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i:y_i=k} \mathbf{z}_i,$$

and the geometric relationship of a sample to each class is measured by

$$d_k(\mathbf{x}) = \|\mathbf{z} - \boldsymbol{\mu}_k\|_2.$$

These clean prototypes remain fixed throughout evaluation, providing the reference geometry for center-drift and margin analysis.

Table 1: Euclidean distances between class prototypes in the L2-normalized feature space.

Category	Real	DF	F2F	FS	NT
Real	0.0000	1.3190	1.1888	1.2610	0.9686
DF	1.3190	0.0000	1.2891	1.3974	1.2309
F2F	1.1888	1.2891	0.0000	1.3176	1.1173
FS	1.2610	1.3974	1.3176	0.0000	1.2942
NT	0.9686	1.2309	1.1173	1.2942	0.0000

For brevity, Deepfakes, Face2Face, FaceSwap, and NeuralTextures are abbreviated as DF, F2F, FS, and NT. As shown in Table 1, all inter-prototype distances are bounded by 2, consistent with the L2-normalized feature space. Real and NeuralTextures are geometrically closest (0.9686), while Deepfakes and FaceSwap are the most separated (1.3974). These distinct inter-class separations provide a meaningful reference for subsequent Feature Center Drift and Attribution Margin analysis.

Perturbations first distort the learned space before causing failures. As images are degraded, their features gradually deviate from the clean ones, displacing the corresponding class prototypes Snell et al. (2017); Wen et al. (2016). For class k, let $\boldsymbol{\mu}_k^{\text{clean}}$ and $\boldsymbol{\mu}_k^{\text{pert}}$ denote the clean and perturbed prototypes. The perturbation-induced Feature Center Drift is

$$D_k = \left\| \boldsymbol{\mu}_k^{\text{clean}} - \boldsymbol{\mu}_k^{\text{pert}} \right\|_2 .$$

Larger D_k indicates stronger geometric distortion, implying reduced feature consistency and weaker class compactness Deng et al. (2019); Wang et al. (2018).

3.3 Attribution Margin and AMC Metrics

Center drift changes the geometric relationship between samples and prototypes. For a testing sample with feature $\mathbf{z}$, the distance to the ground-truth prototype is

$$d_{\text{true}}(\mathbf{x}) = \|\mathbf{z} - \boldsymbol{\mu}_y\|_2 ,$$

and the distance to the nearest competing prototype is

$$d_{\text{nearest}}(\mathbf{x}) = \min_{k \neq y} \|\mathbf{z} - \boldsymbol{\mu}_k\|_2 .$$

The attribution margin is defined as

$$M(\mathbf{x}) = d_{\text{nearest}}(\mathbf{x}) - d_{\text{true}}(\mathbf{x}).$$

A positive margin indicates that the sample is closer to its ground-truth prototype than to any competing prototype; a negative margin indicates that the prototype-induced decision boundary has been crossed. Within AMC, perturbations progressively reduce the margin through center drift, so margin evolution provides the intermediate link between structural feature degradation and attribution failure.

Two complementary metrics characterize the evolution of the decision boundary:

$$\text{AMC-S} = \frac{1}{N} \sum_{i=1}^{N} M(\mathbf{x}_i), \qquad \text{AMC-F} = \frac{1}{N} \sum_{i=1}^{N} \max\big(0,\, -M(\mathbf{x}_i)\big).$$

AMC-S captures the global evolution of decision boundaries. AMC-F measures failure severity over the entire testing set by averaging the magnitude of negative margins across all samples, thereby jointly accounting for both the fraction of samples that cross the boundary and the severity of their crossing. Unlike a formulation that averages only over negative-margin samples, this definition remains well-defined when no samples cross the boundary, in which case $\text{AMC-F} = 0$.

3.4 Theoretical Analysis

AMC explains attribution failure as a progressive geometric evolution rather than an abrupt error. The following proposition and corollary formalize the link between center drift and margin reduction.

Table 2: Attribution accuracy (%) under different perturbations.

Model	JPEG				Gaussian Blur	Motion Blur
	Clean	90	70	50		
EfficientNetV2	93.00	88.08	74.18	63.27	20.90	20.61
EfficientNetV2+FPN	92.08	87.69	72.78	59.41	23.55	25.24
EfficientNetV2+FPN+MLP	92.13	87.84	70.99	55.12	23.65	22.68
ResNet-50	72.97	72.78	68.34	61.53	33.78	38.03

Proposition 1. If perturbations increase the distance to the true-class prototype more than the distance to the nearest competing prototype, then the attribution margin decreases.

Proof Sketch. Let Δd_{true} and $\Delta d_{\text{nearest}}$ denote the perturbation-induced changes in the two distances. The perturbed margin is

$$M^{\text{pert}} = M^{\text{clean}} + \Delta d_{\text{nearest}} - \Delta d_{\text{true}}.$$

Hence if $\Delta d_{\text{true}} > \Delta d_{\text{nearest}}$, then $M^{\text{pert}} < M^{\text{clean}}$. This condition indicates that perturbation-induced distortion reduces the geometric separability between the target class and its competitors, providing the basis for Attribution Margin Collapse.

Corollary 1. Under the condition of Proposition 1, AMC-S decreases monotonically as perturbation strength increases, whereas the proportion of positive-margin samples decreases monotonically as well.

Proof Sketch. Proposition 1 establishes that the attribution margin decreases for samples whose true-class distance grows faster than their nearest-competing-class distance. Averaging over the testing set, the mean margin AMC-S therefore decreases monotonically with perturbation strength. Similarly, once $M(\mathbf{x})$ becomes negative for a sample, that sample crosses the prototype-induced decision boundary, so the proportion of positive-margin samples decreases monotonically as perturbation strength increases.

4 Experiments

4.1 Experimental Setup

Experiments use the FaceForensics++ (FF++) dataset Rossler et al. (2019) under the standard five-class attribution protocol (*Real*, *Deepfakes*, *Face2Face*, *FaceSwap*, *NeuralTextures*). Faces are aligned and resized to 224×224, and the official video-level split yields 15,948 training and 2,072 testing images.

Four attribution models are evaluated: ResNet-50, EfficientNetV2, EfficientNetV2+FPN, and EfficientNetV2+FPN+MLP. During evaluation, only the testing images are perturbed; all models are trained on the original clean training set. Three perturbations are considered: JPEG ($Q = \{90, 70, 50\}$), Gaussian blur, and motion blur. Gaussian blur is applied with kernel size 5 and $\sigma \in \{1.0, 2.0\}$, and motion blur uses a linear kernel of length 15 with angle 45°. Cross-domain evaluation uses Celeb-DF.

All models are implemented in PyTorch under identical settings. Penultimate-layer features are used to construct the feature space, from which prototypes, Feature Center Drift, margins, AMC-S, and AMC-F are computed.

4.2 AMC under Visual Perturbations

We first evaluate whether common visual perturbations degrade attribution performance (Table 2), and then validate whether AMC captures the underlying geometric evolution.

Table 2 shows that all models degrade consistently. JPEG compression produces a gradual decline, while Gaussian and motion blur cause substantially larger drops; EfficientNetV2, for instance, falls from 93.00% on clean images to roughly 21% under blur. Although ResNet-50 is relatively more robust, all models remain highly vulnerable.

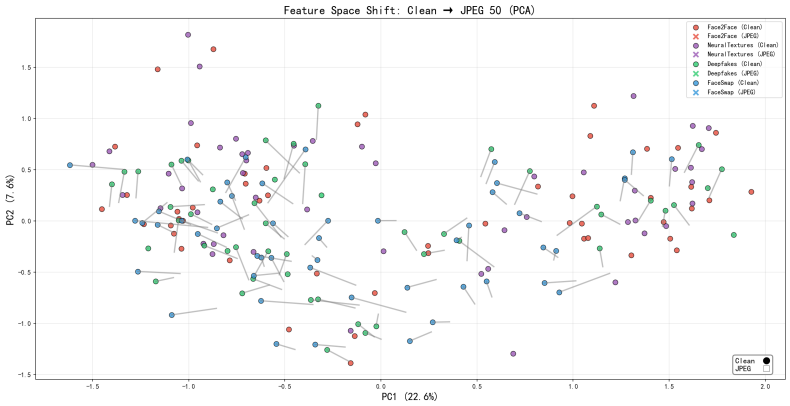


Figure 2: Feature embedding evolution under JPEG50 compression visualized by PCA. Each point represents a sample embedding, and arrows indicate the displacement of class prototypes from clean to perturbed conditions.

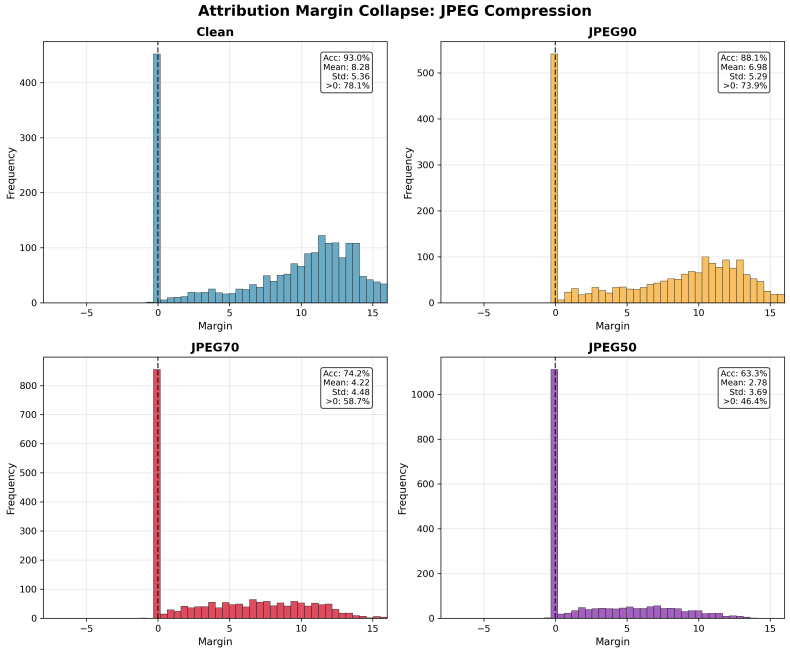


Figure 3: Evolution of attribution margin distributions under increasing JPEG compression. Increasing perturbation strength progressively reduces attribution margins and shifts their distributions toward the prototype-induced decision boundary ($M = 0$).

Figure 2 shows the evolution of sample embeddings and the displacement of class prototypes under JPEG50 compression, indicating that Feature Center Drift is a universal geometric phenomenon. Figure 3 shows that the margin distribution progressively shifts toward the prototype-induced decision boundary $M = 0$: under clean conditions most samples have large positive margins, whereas under JPEG50 a substantial fraction exhibits non-positive margins. These observations provide direct empirical evidence for the proposed AMC process.

Table 3 quantifies this evolution. AMC-S decreases monotonically from 0.5602 under clean conditions to 0.1273 under JPEG50, reflecting reduced geometric separability, while AMC-F increases from 0.0187 to 0.1157. The logit margin exhibits the same monotonic decline, from 7.6586 under clean conditions to 1.6852 under JPEG50, confirming that both the prototype geometry and the classifier decision space degrade consistently under JPEG compression. The proportion of positive-

Table 3: AMC metrics under different JPEG compression levels (EfficientNetV2).

Condition	Acc (%)	Proto AMC-S	Logit AMC-S	AMC-F	Margin >0 (%)
Clean	93.00	0.5602	7.6586	0.0187	93.24
JPEG90	88.08	0.4573	6.2202	0.0320	88.56
JPEG70	74.18	0.2515	3.5384	0.0736	75.39
JPEG50	63.27	0.1273	1.6852	0.1157	64.33

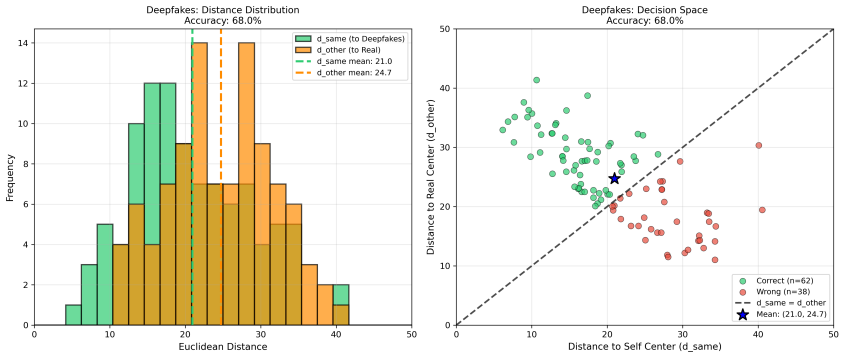


Figure 4: Decision boundary crossing under JPEG50 compression. Correctly classified samples remain inside the decision region, whereas misclassified samples cross the boundary toward competing class prototypes.

margin samples drops from 93.24% to 64.33%, closely tracking the degradation in attribution accuracy (93.00% to 63.27%). The alignment between positive-margin ratio and accuracy confirms that the prototype-induced margin is tightly coupled with the classifier decision behavior under JPEG compression.

Figure 4 illustrates the final stage: as perturbation strength increases, samples move from the correct decision region toward competing prototypes, and the attribution margin changes from $d_{\text{true}} < d_{\text{nearest}}$ to $d_{\text{true}} > d_{\text{nearest}}$, indicating that boundary crossing directly results in attribution failure.

4.3 Generalization Across Perturbations

To test generality, we repeat the evaluation under Gaussian blur and motion blur. The results show consistent geometric trends across all models: although the absolute values of AMC-S and AMC-F vary with perturbation type, AMC consistently characterizes feature-space degradation, demonstrating its generality across diverse perturbations.

4.4 Cross-Dataset Generalization

We next examine whether AMC extends beyond perturbations to cross-domain shift, in which the model is trained on FF++ but evaluated on a different forgery dataset. Because Celeb-DF provides only a binary real-vs-fake annotation, we evaluate a binary attribution model trained on FF++ and tested on Celeb-DF. Under this binary setting, each sample has two possible prototypes (Real and Fake), and the prototype margin is defined as the signed distance advantage of the ground-truth prototype over the competing one. Likewise, the logit margin is computed as the difference between the ground-truth class logit and the maximum competing logit.

On Celeb-DF, accuracy drops from 93.00% to 57.85%. Both margins shrink substantially: Prototype AMC-S decreases from 0.5487 to 0.0963, and Logit AMC-S from 7.6586 to 1.0631. The negative-margin ratio rises from 6.42% to 29.05%, indicating that a substantial fraction of samples has crossed the prototype-induced decision boundary.

To characterize the relationship between the two margins, we compute the Spearman rank correlation. Under FF++, the two margins are highly consistent ($\rho = 0.9031$). Under Celeb-DF, the correlation drops to 0.5246, indicating a partial decoupling between prototype geometry and classi-

Table 4: Cross-dataset summary on FF++ (in-domain) and Celeb-DF (cross-domain). The FF++ results use a five-class attribution model; the Celeb-DF results use a binary real-vs-fake model. Absolute margin values are not directly comparable across the two settings; we report the Spearman correlation and quadrant distribution, which are robust to the number of classes.

Metric	FF++ (5-class)	Celeb-DF (binary)
Accuracy (%)	93.00	57.85
Prototype AMC-S	0.5487	0.0963
Logit AMC-S	7.6586	1.0631
AMC-F	0.0173	0.0328
Prototype Neg. Ratio (%)	6.42	29.05
Spearman ρ (M_P vs. M_L)	0.9031	0.5246
Q2 (%)	0.8	24.0

Table 5: Quadrant analysis on Celeb-DF based on the signs of M_P and M_L. Q2 ($M_P > 0$, $M_L < 0$) accounts for 24.0% of samples, indicating that the classifier degrades before the prototype geometry.

Quadrant	Condition	Ratio (%)	Interpretation
Q1	$M_P > 0,\ M_L > 0$	46.9	Both spaces correct
Q2	$M_P > 0,\ M_L < 0$	24.0	Geometry correct, decision fails
Q3	$M_P < 0,\ M_L > 0$	10.9	Geometry fails, decision correct
Q4	$M_P < 0,\ M_L < 0$	18.1	Both spaces fail

fier decisions. This decoupling is considerably weaker than what we would expect from a complete collapse, but it is substantially weaker than the in-domain alignment.

Table 5 confirms this at the sample level. On Celeb-DF, Q1 accounts for 46.9% of samples, where both margins are positive. The most informative quadrant is Q2, which accounts for 24.0% of the samples: these samples remain geometrically correct in the prototype space ($M_P > 0$) but are misclassified by the classifier ($M_L < 0$). This indicates that the classifier degrades earlier than the prototype geometry, and that the prototype margin can serve as an early warning signal before the classifier collapses.

We further visualize this behavior. Figure 5 shows the M_P–M_L scatter plots on FF++ (left) and Celeb-DF (right). On FF++, samples are concentrated in Q1, indicating that the prototype geometry and the classifier decision space agree well in-domain. On Celeb-DF, a noticeable fraction of samples spreads into Q2, visually confirming the partial decoupling between the two spaces. Figure 6 shows the corresponding margin distributions: both M_P and M_L shift toward smaller values under cross-domain shift.

Overall, the cross-domain experiment reveals a partial decoupling between prototype geometry and classifier decisions. Unlike perturbations, which preserve the alignment between the two margins, domain shift weakens this alignment substantially (Spearman ρ drops from 0.9031 to 0.5246) and pushes 24.0% of samples into Q2, where geometry remains correct but decisions fail. This suggests that AMC can characterize intermediate failure modes that accuracy alone cannot capture.

5 Discussion

AMC characterizes prototype geometry without equating it with classifier behavior. Under JPEG compression, prototype and classifier margins exhibit similar aggregate trends; under cross-domain shift, positive prototype margins can coexist with incorrect predictions. These observations suggest that prototype geometry provides a partial account of robustness, rather than a universal explanation of attribution failure.

The five-class attribution and binary detection experiments address different tasks. Margin magnitudes are therefore not directly comparable across these settings. Spearman correlation and quadrant proportions summarize within-setting relationships but do not eliminate task differences. Cross-domain comparisons require a fixed binary classifier and fixed source-domain prototypes.

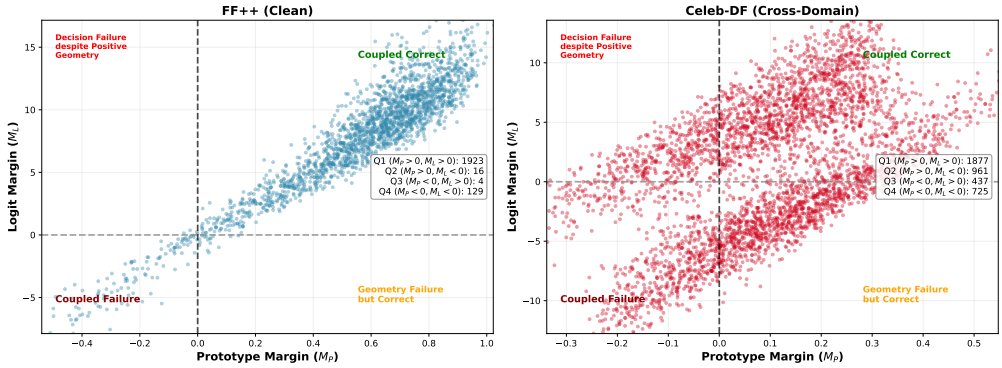


Figure 5: Scatter plots of Prototype Margin (M_P) vs. Decision Margin (M_L) on FF++ (left) and Celeb-DF (right). On FF++, samples are concentrated in Q1, indicating that prototype geometry and classifier decisions agree well in-domain. On Celeb-DF, a substantial fraction of samples spreads into Q2, visually confirming the partial decoupling between the two spaces under cross-domain shift.

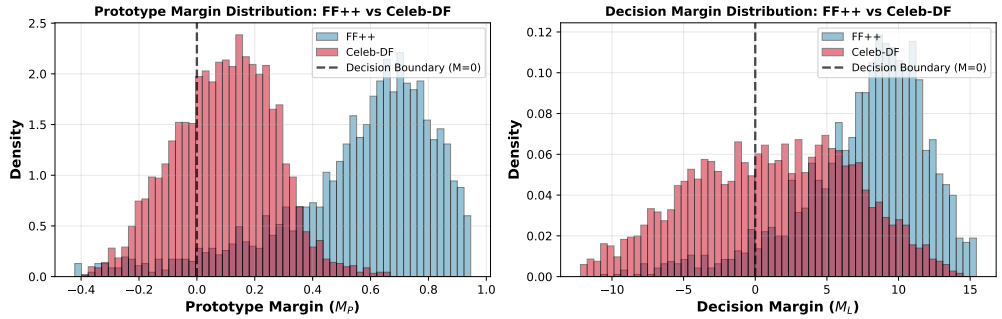


Figure 6: Margin distributions on FF++ and Celeb-DF. Both distributions shift toward smaller values under cross-domain shift, consistent with the partial decoupling observed in the quadrant analysis.

6 Conclusion

We presented Attribution Margin Collapse (AMC), a diagnostic framework for examining the relationship between prototype geometry and classifier decisions in deepfake attribution. The proposed statistics, AMC-S and AMC-F, summarize the mean signed prototype margin and the dataset-wide magnitude of negative margins. Our experiments document attribution degradation under visual perturbations, with JPEG analysis showing similar aggregate trends in prototype and classifier margins. A complementary cross-domain study of binary deepfake detection reveals that positive prototype margins can coexist with incorrect classifier predictions. These findings highlight both the utility and the limitations of prototype-based analysis: geometric margins characterize representation structure, but their correspondence with classifier behavior must be established empirically. AMC thus supports robustness analysis beyond accuracy while distinguishing geometric observations from claims about the causes of classification failure.

AI Use Statement

In this work, we used generative AI tools for language polishing and code debugging. We have reviewed all AI-assisted work. AI-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics Statement

This work does not involve human subjects, and all datasets used are publicly available and widely used in the research community. The proposed framework is intended for forensic analysis of manipulated media and is not designed for malicious use.

Reproducibility Statement

We provide detailed experimental settings in Section 4, including dataset splits, model architectures, and perturbation parameters. The class prototypes and margin computation follow the definitions in Section 3. Source code and pretrained checkpoints will be released upon publication.

References

Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Tania Sari Bonaventura, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, et al. Deepfake media forensics: State of the art and challenges ahead. In *International Conference on Advances in Social Networks Analysis and Mining*, pp. 33–48. Springer, 2024.

Spiros Baxevanakis, Manos Schinas, and Symeon Papadopoulos. Do deepfake attribution models generalize? In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation*, pp. 45–54, 2025.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Anurag Dutta, Arnab Kumar Das, Ruchira Naskar, and Rajat Subhra Chakraborty. Wavedif: Wavelet sub-band based deepfake identification in frequency domain. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6312–6321, 2025.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pp. 3247–3258. PMLR, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Shan Jia, Xin Li, and Siwei Lyu. Model attribution of face-swap deepfake videos. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2356–2360. IEEE, 2022.

Yi Kou, Peng Li, Hongjiang Ma, Jiliu Zhou, Zhan ao Huang, and Xiaojie Li. Sfiad: Deepfake detection through spatial-frequency feature integration and dynamic margin optimization. *Artificial Intelligence Review*, 58(7):217, 2025.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5001–5010, 2020.

P. Liu, Q. Tao, and J. Zhou. Robust deepfake detection by addressing generalization and trustworthiness challenges: A short survey. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, pp. 3–11, 2024. doi: 10.1145/3689090.3689386.

Mark Huasong Meng, Guangdong Bai, Sin Gee Teo, Zhe Hou, Yan Xiao, Yun Lin, and Jin Song Dong. Adversarial robustness of deep neural networks: A survey from a formal verification perspective. *IEEE Transactions on Dependable and Secure Computing*, 2022.

Hong-Hanh Nguyen-Le, Van-Tuan Tran, Dinh-Thuc Nguyen, and Nhien-An Le-Khac. Passive deepfake detection across multi-modalities: A comprehensive survey. *arXiv preprint arXiv:2411.17911*, 2024.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18720–18729, 2022.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2020.

Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*, 57(3):1–35, 2024.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pp. 499–515. Springer, 2016.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14448–14457, 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.