Computing and Artificial Intelligence 2025, 3(4), 4258.
doi: https://doi.org/10.59400/cai4258

Article

Temporal Explanation Drift: Measuring and analyzing temporal explanation drift in deep neural networks

Zaryab Rahman[1,*], **Mattia Ottoborgo**[2]

[1] Department of Computer Science and IT, Faculty of Computing Sciences and Engineering, University of Malakand, Chakdara 18800, Pakistan
[2] Trustpilot, 1112 Copenhagen, Denmark
*** Corresponding author:** Zaryab Rahman, zaryabrahman848@gmail.com

CITATION

Rahman Z, Ottoborgo M. Temporal Explanation Drift: Measuring and analyzing temporal explanation drift in deep neural networks. Computing and Artificial Intelligence. 2025; 3(4): 4258. https://doi.org/10.59400/cai4258

ARTICLE INFO

Received: 14 August 2025
Revised: 5 December 2025
Accepted: 8 December 2025
Available online: 26 December 2025

COPYRIGHT

Copyright © 2025 Author(s). *Computing and Artificial Intelligence* is published by Academic Publishing Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license. https://creativecommons.org/licenses/by/4.0/

Abstract: The evaluation of explainable artificial intelligence (XAI) methods has largely focused on the properties of a single, fully-trained model. This static view overlooks a critical dimension of reliability: the stability of an explanation throughout the training process. In this paper, we introduce a formal framework to measure and analyze Temporal Explanation Drift (TED), quantifying how a model's explanations for a consistent, correctly classified input change from one training epoch to the next. Our analysis on Vision Transformers reveals a fundamental decoupling of stability: the high-level semantic focus of an explanation stabilizes early, while its fine-grained structural representation continues to drift long after predictive accuracy has converged. We demonstrate that this is a general phenomenon that extends to Convolutional Neural Networks (CNNs), but find that its magnitude is critically dependent on the explanation method. Our experiments reveal that post-hoc, gradient-based methods like Grad-CAM are dramatically less stable than intrinsic methods like attention rollout. A final, controlled experiment scientifically dissects these effects, showing that the choice of explanation method is the primary driver of temporal instability, which is further modulated by a nuanced interaction with the model's architecture. Our work establishes temporal stability as a distinct and vital property for evaluating the trustworthiness of explanations and provides a practical framework for its measurement.

Keywords: explainable artificial intelligence (XAI); temporal stability; explanation drift; saliency methods; training dynamics

1. Introduction

Deep neural networks have become a dominant force in artificial intelligence, establishing new state-of-the-art benchmarks across a vast landscape of scientific and industrial domains. Their success is particularly prominent in computer vision, where they have revolutionized tasks ranging from medical image analysis for disease diagnosis [1, 2] and pathology detection [3] to safety-critical applications like autonomous vehicle perception [4, 5] and robotic control [6]. This remarkable performance, however, is coupled with a significant challenge: their inherent opacity. The complex, hierarchical, and non-linear nature of these "black box" models makes their internal decision-making processes exceptionally difficult for humans to comprehend. As these systems are increasingly integrated into high-stakes environments where accountability and trust are non-negotiable, this lack of transparency poses a significant barrier to their adoption and safe deployment [7, 8].

The field of Explainable AI (XAI) has emerged to directly confront this challenge, aiming to develop methods that can render the reasoning of complex models intelligible [9, 10]. A prominent family of techniques within XAI produces saliency maps, which visualize the importance of different input features—typically pixels or image regions—to a model's final prediction. These methods can be broadly categorized into intrinsic approaches, which leverage architectural components like the self-attention mechanism in Transformers [11, 12], and post-hoc approaches, such as the widely used gradient-based method Grad-CAM [13], which can be applied to any differentiable model. The ultimate goal of these techniques is not just to produce a plausible-looking heatmap, but to provide an explanation that is a faithful and reliable account of the model's true behavior [14].

Significant research has been dedicated to critically evaluating the properties of these explanations. This has led to a more rigorous understanding of their limitations, establishing key desiderata such as robustness to input perturbations [15, 16] and reliability across semantically similar inputs [17]. Foundational "sanity checks" have even revealed that some saliency methods are surprisingly insensitive to model parameters, operating more like sophisticated edge detectors and thus failing to reflect the model's learned knowledge [18, 19]. However, these vital evaluation paradigms have almost exclusively focused on a static snapshot of the model: a single, fully-trained checkpoint. This overlooks a crucial dimension of model behavior—its evolution over the course of training.

This raises a critical and unexplored question: what if an explanation, while seemingly coherent at the final epoch, was radically different just a few epochs prior, even when the model's prediction was consistently correct and confident? A highly volatile explanation for a stable prediction implies an unstable underlying reasoning process, which undermines the trustworthiness of the final explanation. In safety-critical domains—such as medical diagnosis or autonomous driving—a model whose internal rationale fluctuates unpredictably during training may not be considered reliable, even if its final accuracy appears sufficient. Such instability might suggest that the model's decision boundary remains sensitive to small changes in optimization dynamics, or that the explanation method itself is not a faithful instrument for interpreting the model's cognitive process.

This paper argues that the temporal stability of an explanation is a distinct, measurable, and vital property of its overall reliability, complementary to established desiderata like faithfulness and robustness. An explanation that remains consistent as a model converges on a solution is inherently more trustworthy than one that shifts erratically, because it indicates a consolidation of the model's internal feature attribution rather than a transient alignment. To formalize and investigate this concept, we introduce a framework for measuring Temporal Explanation Drift (TED). Our systematic analysis across different architectures and explanation methodologies makes the following contributions:

- We propose a rigorous methodology and a new set of metrics, including the Temporal Explanation Stability Score (TESS), to quantify the stability of saliency-based explanations across a model's entire training trajectory. Our

framework is carefully designed to isolate the dynamics of the explanation from changes in the model's predictive output by conditioning on a stable prediction subset.

- Our analysis of Vision Transformers [20] reveals a fundamental decoupling of stability. We show that the high-level semantic focus of an explanation (what the model is looking at) stabilizes relatively early in training, while its fine-grained structural representation (the precise shape and texture of the explanation map) remains in a state of continuous, subtle refinement long after model accuracy has converged.
- We demonstrate that this phenomenon is generalizable to Convolutional Neural Networks (CNNs) and find that post-hoc, gradient-based methods like Grad-CAM are dramatically less stable than intrinsic methods like attention rollout. Our final, controlled experiment scientifically dissects these effects, revealing that the choice of explanation method is the primary driver of temporal instability, with this effect being further modulated by a nuanced interaction with the model's architecture.

The remainder of this paper is structured as follows. We first review related work in XAI evaluation and the analysis of training dynamics. We then detail our methodology for quantifying temporal drift. Subsequently, we present the results of our three-part experimental investigation, before concluding with a discussion of our findings and their implications for developing more reliable and trustworthy AI systems.

2. Related work

Our research is positioned at the intersection of three key domains: the development of explanation methods, the critical evaluation of their properties, and the broader analysis of neural network training dynamics.

2.1. Methods in explainable AI (XAI)

The pursuit of model interpretability has given rise to a diverse array of explanation techniques. Our work engages with two prominent families of these methods for vision models. The first is intrinsic methods, which derive explanations directly from a model's architectural components. For Vision Transformers (ViTs) [20], which form the basis of our initial experiments with DeiT [21], the self-attention mechanism itself serves as a natural source of interpretability. Techniques like Attention Rollout [11] aggregate attention weights across layers to produce a saliency map, operating under the assumption that these weights reflect the model's feature importance priorities. Recent works have further refined attention-based explainability for ViTs, introducing methods that combine attention with gradient information or adapt rollout to deeper architectures [12, 22]. Beyond attention, intrinsic methods also include concept-based interpretability built directly into model architectures [23].

The second family is post-hoc, gradient-based methods, which are model-agnostic and can be applied to any differentiable architecture. These methods use backpropagated gradients as a proxy for feature importance. Grad-CAM [13] is arguably the most influential of these, using the gradients flowing into the final

convolutional layer to produce a coarse localization map that highlights important regions for a given prediction. Its widespread adoption, applied here to a MobileNetV3 architecture [24], makes it a critical subject for our stability analysis. While these methods are powerful, our work investigates a previously under-explored property: how their outputs change as the gradients themselves evolve throughout training. Notably, other major post-hoc families—perturbation-based (e.g., LIME [25], SHAP [26]) and axiomatic approaches (e.g., Integrated Gradients [27])—are beyond the scope of our experiments but represent important paths for future temporal stability evaluation. Recent surveys have systematically categorized these methods and their evaluation criteria [28,29].

2.2. The critical evaluation of explanations

Early XAI research focused primarily on generating explanations, but a more recent and critical line of inquiry has focused on evaluating their quality and reliability. This literature has established several key desiderata for a good explanation. One of the most important is faithfulness, which measures whether the explanation accurately reflects the model's internal reasoning process [14]. Another is robustness, which assesses the stability of an explanation in response to small, often imperceptible, perturbations to the input image [15,18]. Sanity checks proposed by Adebayo et al. [18] famously demonstrated that some explanation methods were not sensitive to model parameters, casting doubt on their faithfulness. Similarly, work on reliability has shown that explanations can be inconsistent across semantically similar inputs [17]. More recent evaluation frameworks have introduced standardized benchmarks and metrics for comparing saliency maps across a broad range of models and methods [30–32].

While these works have been foundational in establishing a more rigorous standard for XAI, they have predominantly focused on evaluating explanations for a static, fully-trained model. Our work introduces a new, orthogonal dimension to this evaluation paradigm: temporal stability. We argue that an explanation's consistency over the training trajectory is a distinct and vital property. An explanation that is faithful and robust at epoch 50 but was radically different at epoch 49, despite no change in predictive accuracy, may not be a trustworthy account of a stable reasoning process. This temporal perspective connects directly to growing concerns about the brittleness of explanations when models are deployed in non-stationary environments [29].

2.3. Analyzing neural network training dynamics

Our methodology is also inspired by a broader line of research that seeks to understand not just what neural networks learn, but how they learn it over time. Seminal works have investigated how internal representations evolve during training, using techniques such as Singular Vector Canonical Correlation Analysis (SVCCA) [33] or Centered Kernel Alignment (CKA) [34] to measure the similarity of learned representations across different layers, models, or training epochs. These studies have revealed that networks often converge on stable representations in a bottom-up fashion and that different architectures have distinct learning dynamics. Recent extensions have applied these similarity measures to study phenomena like grokking and the emergence

of structured representations in large models [35,36]. Furthermore, new approaches that use loss landscape geometry to understand training phases have also emerged [37,38].

This paper extends this "dynamics-oriented" perspective to the domain of explainability. Just as prior work has tracked the evolution of internal feature maps, we track the evolution of the final saliency maps. By doing so, we connect the field of XAI evaluation with the fundamental study of learning dynamics, using explanation drift as a novel lens through which to observe model convergence, consolidation, and the subtle onset of overfitting.

3. Methodology

To quantitatively investigate the temporal stability of explanations, we developed a formal framework for measuring drift over the course of model training, as illustrated in **Figure 1**. Our approach is designed to isolate the dynamics of the explanation mechanism itself, independent of changes in the model's predictive output.

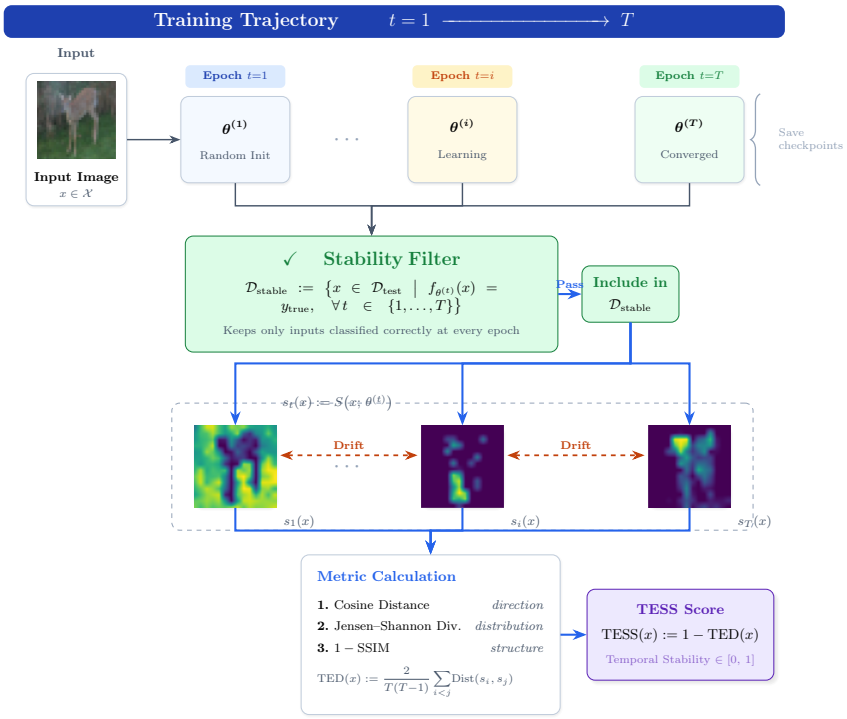


Figure 1. The Proposed Framework for Measuring Temporal Explanation Drift (TED). The diagram illustrates the extraction of the Stable Prediction Subset ($\mathcal{D}_{stable}$) and the calculation of stability metrics across training epochs.

3.1. Preliminaries and problem formulation

Let $\mathcal{X} \subset \mathbb{R}^d$ be the input image space and $\mathcal{Y}$ be the corresponding label space. We consider a deep neural network, parameterized by $\theta^{(t)}$ at training epoch $t \in \{1, \ldots, T\}$, which defines a mapping function $f_{\theta^{(t)}} : \mathcal{X} \to \mathcal{Y}$.

An explanation function, S, produces a saliency map for a given input $\mathbf{x} \in \mathcal{X}$ and model state $\theta^{(t)}$. This map is a vector or matrix in $\mathbb{R}^m$ that assigns an importance score to each of the m input features (e.g., patches or pixels). We define the explanation for

input $\mathbf{x}$ at epoch t as:

$$\mathbf{s}_t(\mathbf{x}) := S(\mathbf{x}; \theta^{(t)}) \in \mathbb{R}^m. \tag{1}$$

Our goal is to quantify the dissimilarity of the sequence of explanations $\{\mathbf{s}_1(\mathbf{x}), \mathbf{s}_2(\mathbf{x}), \ldots, \mathbf{s}_T(\mathbf{x})\}$ for a given input $\mathbf{x}$.

3.2. Isolating explanation dynamics: The stable prediction subset

A naive analysis of explanation drift can be confounded by changes in the model's prediction. An explanation will necessarily change if the model fundamentally alters its classification of an object. To disentangle the drift of the explanation from the drift of the prediction itself, we perform all analyses exclusively on a Stable Prediction Subset, denoted $\mathcal{D}_{\text{stable}}$. This subset is formally defined as:

$$\mathcal{D}_{\text{stable}} := \{\mathbf{x} \in \mathcal{D}_{\text{test}} \mid f_{\theta^{(t)}}(\mathbf{x}) = y_{\text{true}}, \quad \forall t \in \{1, \ldots, T\}\}. \tag{2}$$

By conditioning our analysis on this subset of consistently and correctly classified images, we ensure that any observed drift is attributable to changes in the model's internal reasoning process, not its final output.

3.3. Quantifying Temporal Explanation Drift (TED)

For any given input $\mathbf{x} \in \mathcal{D}_{\text{stable}}$, we define the Temporal Explanation Drift (TED) as the average pairwise dissimilarity between its explanations across all sampled epochs. This provides a single, aggregate measure of its overall stability throughout training:

$$\text{TED}(\mathbf{x}) := \frac{2}{T(T-1)} \sum_{i=1}^{T} \sum_{j=i+1}^{T} \text{Dist}(\mathbf{s}_i(\mathbf{x}), \mathbf{s}_j(\mathbf{x})) \tag{3}$$

where $\text{Dist}(\cdot, \cdot)$ is a chosen distance function between two saliency maps.

3.4. Distance metrics for saliency maps

To provide a holistic and robust measure of dissimilarity, we employ three complementary distance metrics, each capturing a different aspect of an explanation's properties. Let $\mathbf{A}$ and $\mathbf{B}$ be two saliency maps in $\mathbb{R}^m$.

- **Cosine distance:** Measures the change in the vector direction of the explanations, making it sensitive to shifts in the high-level focus of attention but invariant to uniform changes in magnitude. It is defined as:

$$\begin{aligned} \text{Dist}_{\cos}(\mathbf{A}, \mathbf{B}) &= 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} \\ &= 1 - \frac{\sum_{k=1}^{m} A_k B_k}{\sqrt{\sum_{k=1}^{m} A_k^2} \sqrt{\sum_{k=1}^{m} B_k^2}} \end{aligned} \tag{4}$$

  Cosine distance is widely used in XAI evaluation to compare explanation vectors [30].

- **Jensen–Shannon divergence (JSD):** After normalizing the saliency maps to sum

to one (treating them as probability distributions P_A and P_B), JSD quantifies the difference in the overall distribution of importance scores. It is a symmetrized and smoothed version of the Kullback–Leibler (KL) divergence, based on information-theoretic principles [39], defined as:

$$\text{JSD}(P_A \| P_B) = \frac{1}{2} D_{KL}(P_A \| M) + \frac{1}{2} D_{KL}(P_B \| M) \tag{5}$$

where $M = \frac{1}{2}(P_A + P_B)$ and the KL divergence is $D_{KL}(P\|Q) = \sum_{k=1}^{m} P_k \log\left(\frac{P_k}{Q_k}\right)$.

- **Structural similarity index measure (SSIM):** Used as a distance metric ($1 -$ SSIM), it measures perceptual changes in structure, texture, and luminance between two image-like maps. It is highly sensitive to fine-grained spatial reconfigurations. The SSIM index was originally proposed for image quality assessment [40] and is given by:

$$\text{SSIM}(\mathbf{A}, \mathbf{B}) = \frac{(2\mu_A\mu_B + c_1)(2\sigma_{AB} + c_2)}{(\mu_A^2 + \mu_B^2 + c_1)(\sigma_A^2 + \sigma_B^2 + c_2)} \tag{6}$$

where μ is the mean, σ^2 is the variance, σ_{AB} is the covariance of **A** and **B**, and c_1, c_2 are small constants to ensure stability.

3.5. The Temporal Explanation Stability Score (TESS)

For ease of interpretation, we invert the drift into a normalized stability score, the Temporal Explanation Stability Score (TESS), bounded between 0 and 1:

$$\text{TESS}(\mathbf{x}) := 1 - \text{TED}(\mathbf{x}). \tag{7}$$

A TESS score of 1 indicates perfectly stable explanations that do not change at all during training, while a score approaching 0 indicates maximal instability. The average TESS score over the dataset, TESS_{avg}, is used to compare the overall stability of different model-method combinations.

3.6. Analyzing fine-grained dynamics: Consecutive drift

While TED and TESS provide aggregate stability measures, they do not reveal the evolutionary dynamics of drift over time. To analyze this, we compute the Consecutive Drift, which measures the dissimilarity between explanations at adjacent epochs:

$$e\text{Drift}_{\text{consec}}(\mathbf{x}, t) := \text{Dist}(\mathbf{s}_t(\mathbf{x}), \mathbf{s}_{t+1}(\mathbf{x})) \tag{8}$$

By averaging this value across all images in $\mathcal{D}_{\text{stable}}$ for each epoch t, we can plot the evolution of drift throughout training. This visualization is crucial for identifying distinct phases of stability, such as the initial high-volatility reorganization phase and the later fine-tuning and convergence phase.

3.7. Theoretical motivation for decoupled drift

Before turning to the experiments, we articulate a theoretical rationale for the decoupled drift phenomenon that our framework is designed to capture. This rationale

gives a principled basis for expecting two distinct timescales in explanation evolution.

3.7.1. Early semantic stabilisation

In the initial phase of training, a deep network rapidly learns to separate the major class-conditioned modes of the data. Once the model has identified the coarse image region that discriminatively signals the target class, the direction of the explanation vector becomes anchored. From the perspective of the loss landscape, the network enters a wide basin where the dominant gradient component with respect to the input is aligned with that semantic region [38]. Consequently, the high-level attribution (what the model looks at) changes very little from that point onward.

3.7.2. Persistent structural refinement

Although the semantic focus is stable, the precise shape of the saliency map continues to evolve. Each stochastic gradient step perturbs the decision boundary locally, especially near the margins of the attended region. These tiny boundary adjustments manifest as fine-grained spatial reconfigurations in the explanation heatmap—blurring, sharpening, or slight translation—while the centre of mass remains unchanged. The structural similarity index measure (SSIM) is acutely sensitive to such local texture and luminance variations, which is why structural drift ($1 - \text{SSIM}$) remains elevated long after semantic metrics (Cosine, JSD) have plateaued.

3.7.3. Gradient-based methods amplify the effect

Post-hoc gradient-based explainers, such as Grad-CAM, differentiate through the loss surface directly. They therefore inherit the high-frequency noise present in stochastic gradients, which acts as an additional jitter on the saliency map [18]. Intrinsic methods, like attention rollout, read directly from the model's internal feature-attribution weights, which are of lower frequency and averaged over many forward passes. This qualitative difference explains why gradient-based explanations are inherently more fragile to temporal perturbations throughout training.

Taken together, these considerations motivate the use of a composite metric suite (Cosine, JSD, SSIM) and a consecutive-drift view: they jointly disentangle semantic convergence from structural wandering and isolate the extra instability that gradient-based methods introduce.

3.8. Computational cost and scalability

The TED framework operates offline and does not alter the training loop; however, its cost profile is worth characterising. For a model trained for T epochs with a stable subset of size N, the procedure requires:

- Storing T model checkpoints (or equivalently, computing saliency maps on the fly).
- Generating $N \cdot T$ saliency maps for each explanation method under study.
- Computing pairwise distances for aggregation and consecutive drift.

On our experimental setup—a DeiT-Tiny model trained on CIFAR-10 for $T = 50$ epochs with $N = 139$—the entire analysis (all three metrics, two explanation methods) completes in under one hour on a single mid-range GPU. Checkpoint storage

is negligible (< 1 GB total).

For substantially larger models or datasets, the cost can be kept manageable with two simple design choices:

1. **Anchor-epoch evaluation:** instead of saving every epoch, evaluate at a subset of $K \ll T$ evenly spaced epochs. The qualitative phase structure can still be recovered with $K \approx 10$–15.
2. **Incremental distance updates:** the consecutive drift at epoch t can be computed and immediately aggregated, so that only two successive saliency maps need to be held in memory at any time.

These options allow the framework to scale to larger settings without compromising its diagnostic power, making TED practical for research quality assurance and, potentially, for periodic monitoring during model development.

4. Experimental results

To study how explanations evolve during training, we fine-tuned a pre-trained DeiT-Tiny Vision Transformer on CIFAR-10 for 50 epochs and saved checkpoints after every epoch. For a fixed set of test images that were correctly classified throughout training (the stable prediction subset $\mathcal{D}_{\text{stable}}$), we extracted attention-based saliency maps at each epoch. Using our consecutive drift measures and the aggregated TESS score, we found that explanation drift is not uniform but rather unfolds in two qualitatively distinct stages.

4.1. Temporal dynamics reveal a decoupling of semantic and structural drift

Figure 2 shows the average consecutive drift between adjacent-epoch attention maps, measured with three different distances. Two clearly separated regimes emerge:

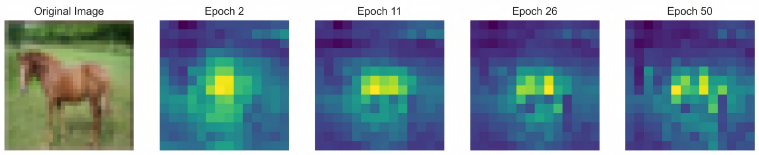
- **Phase 1: Early reorganization (Epochs 1–15).** All three metrics indicate high volatility in the first 15 epochs. The model, adapting from its pre-trained initialization to CIFAR-10, undertakes a rapid restructuring of feature representations. Structural drift (1– SSIM) is especially pronounced, signaling that the spatial layout of attention changes considerably during this period.
- **Phase 2: Semantic consolidation and structural refinement (Epochs 15–50).** After this initial phase, Cosine distance and JSD quickly drop and remain near zero. This means that the overall direction of the explanation—what part of the image the model focuses on—becomes stable relatively early. In contrast, structural drift (1– SSIM) settles at a higher baseline and continues to fluctuate. The model keeps refining the precise contour and texture of its attention maps even after its semantic focus has locked in.

This divergence between semantic and structural drift is the core empirical finding of our work. While the model's high-level attention target is fixed early, the detailed shape of the saliency map remains under continuous modification throughout the rest of training, long after prediction accuracy has plateaued.

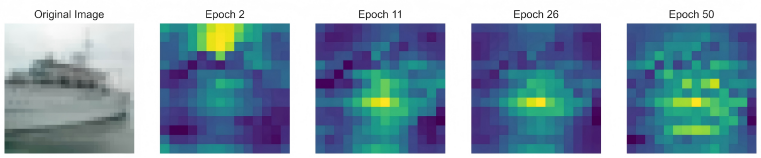
Figure 2. Average consecutive drift between explanation maps across adjacent epochs. The plot shows high initial drift across all metrics, followed by a divergence where semantic drift (Cosine, JSD) stabilizes near zero while structural drift ($1-$ SSIM) remains elevated.

4.2. Qualitative analysis of stability extremes

Figure 3 provides visual examples from both ends of the stability spectrum within $\mathcal{D}_{\text{stable}}$. The most stable example ($\text{TESS}_{\cos} = 0.975$) shows the model attending to the horse's torso already at Epoch 2, and this focus is maintained in all later epochs. Yet even here, subtle shifts in the sharpness and spread of the attention map are visible: at Epoch 26 the hotspot is more concentrated, while by Epoch 50 it has become more diffuse. Cosine distance remains low because the center of attention does not move, but SSIM picks up these local shape changes.



(a) Most stable example (top row).



(b) Least stable example (bottom row).

Figure 3. Evolution of attention maps for the most stable (top row) and least stable (bottom row) examples from $\mathcal{D}_{\text{stable}}$. The images show how semantic focus can lock in early, while fine-grained structural details continue to change.

The least stable example ($\text{TESS}_{\cos} = 0.930$) illustrates the opposite. Early attention targets the ship's superstructure (Epoch 2), then moves significantly to the hull by Epoch 11. This large spatial shift corresponds to the elevated drift across all metrics during the early reorganization phase. As training progresses, the attention map continues to reshape, aligning with the persistent structural drift seen in later epochs.

4.3. Aggregate stability over the full training trajectory

Table 1 summarizes the stability scores averaged over the entire 50-epoch trajectory. Out of 200 analyzed test images, 139 (69.5%) remained consistently correctly classified, forming a reasonably sized stable subset for analysis.

Table 1. Aggregate stability results.

Category	Metric	Value
Dataset statistics	Size of Analyzed Test Set	200
	Size of $\mathcal{D}_{\text{stable}}$	139
	Percentage Stable	69.50%
Aggregate stability scores	TESS_{avg} (Cosine)	0.9574 ± 0.0084
	TESS_{avg} (JSD)	0.9862 ± 0.0029
	TESS_{avg} (SSIM)	0.8257 ± 0.0475

Note: Dataset statistics and average Temporal Explanation Stability Scores (TESS) over $\mathcal{D}_{\text{stable}}$ across the full 50-epoch training trajectory. Directional/distributional metrics (Cosine, JSD) show high stability, while the structural metric (SSIM) is lower and more variable.

The average TESS values reinforce the patterns observed in the consecutive drift plot. Cosine and JSD scores are high (0.9574 and 0.9862, respectively), indicating that explanations are remarkably stable in terms of direction and distribution. The SSIM-based TESS is substantially lower (0.8257) and shows greater variance (standard deviation ± 0.0475). This confirms that while semantic positioning of attention stabilizes early, the fine spatial structure of the saliency map is the primary source of long-term explanation instability in a fine-tuned Vision Transformer.

Taken together, these results show that temporal explanation stability is a two-component phenomenon. Once a model has learned the basic class-discriminative features, its semantic attention stabilizes, whereas the structural details of its explanations continue to evolve, exhibiting a prolonged refinement phase that persists even after predictive accuracy converges.

5. Generalization and the fragility of gradient-based explanations

Having established the existence of temporal explanation drift in Vision Transformers, a critical question arises: is this phenomenon an idiosyncrasy of the attention mechanism, or is it a more fundamental property of deep neural network training? To investigate this, we extended our analysis to a different architectural class—a Convolutional Neural Network (CNN), and a different mode of explanation—the widely used gradient-based method, Grad-CAM (see **Figure 4**).

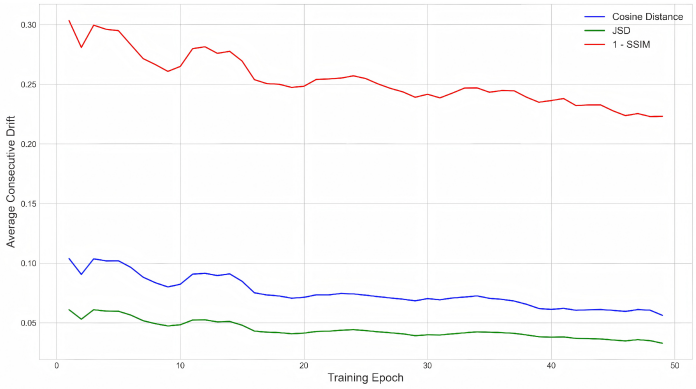


Figure 4. Average consecutive drift for Grad-CAM explanations from MobileNetV3. The pattern is similar to the Vision Transformer, but the magnitude of the drift is significantly higher across all metrics.

For this experiment, we trained a MobileNetV3 model on the CIFAR-10 dataset, following the same 50-epoch protocol. The results provide supporting evidence that temporal explanation drift generalizes across architectures, while also revealing notable differences in magnitude that warrant careful interpretation.

5.1. Architectural generalization and evidence of overfitting

The training dynamics of the MobileNetV3 model, shown in **Figure 5**, provide essential context for our stability analysis. While the model achieves high training accuracy (99%), a significant and persistent gap emerges between training and validation performance, indicating a more pronounced overfitting behavior compared to the DeiT model. This tendency to overfit, which is common in compact architectures trained on small datasets, may amplify the brittleness of gradient signals that Grad-CAM relies upon.

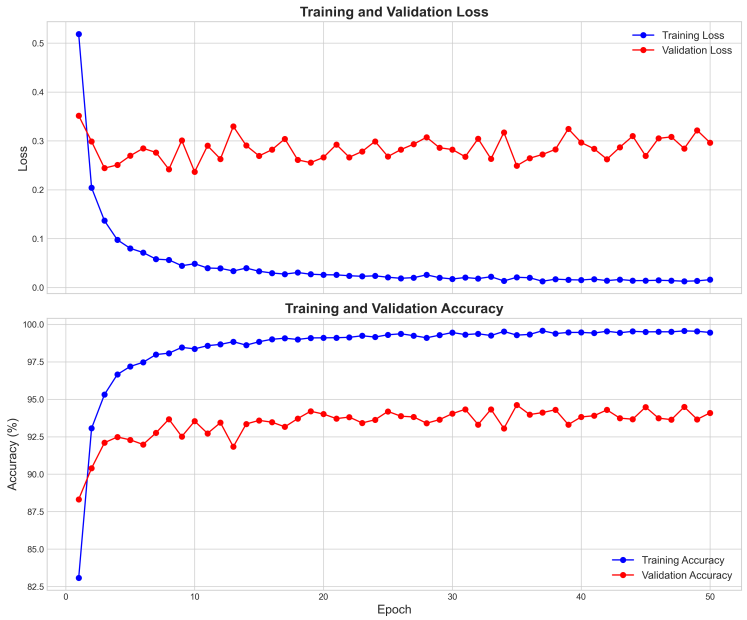


Figure 5. Training and validation accuracy for the MobileNetV3 model over 50 epochs. The divergence between the two curves suggests a degree of overfitting.

5.2. Amplified drift in a CNN architecture

The analysis of the MobileNetV3 explanations reveals a drift pattern qualitatively similar to that observed in the Vision Transformer: a turbulent early phase followed by a later phase where semantic metrics (Cosine, JSD) stabilize while structural drift (1−SSIM) persists. However, the absolute scale of this drift is notably larger. The semantic drift, as measured by Cosine distance, stabilizes at an average value of approximately 0.07—nearly three times higher than the 0.025 observed for the DeiT model's attention maps. The structural drift is similarly inflated.

This increase in instability is quantified in **Table 2**. The aggregate TESS scores for MobileNetV3 are substantially lower across all metrics compared to the DeiT model from Experiment 1. The average structural stability score, $TESS_{avg}$ (SSIM), drops from 0.826 to 0.670, and the directional stability, $TESS_{avg}$ (Cosine), falls from 0.957 to 0.886. These results, while limited to a single CNN architecture and training run, strongly

suggest that the Grad-CAM explanations for this model are more temporally erratic than the attention rolls for the ViT.

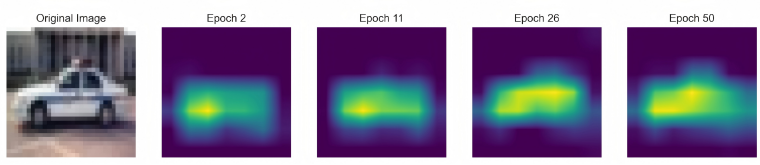
Table 2. Aggregate stability results for MobileNetV3 with Grad-CAM.

Category	Metric	Value
Dataset Statistics	Size of Analyzed Test Set	200
	Size of $\mathcal{D}_{\text{stable}}$	130
	Percentage Stable	65.00%
Aggregate Stability Scores	TESS_{avg} (Cosine)	0.8860 ± 0.0550
	TESS_{avg} (JSD)	0.9350 ± 0.0327
	TESS_{avg} (SSIM)	0.6697 ± 0.0635

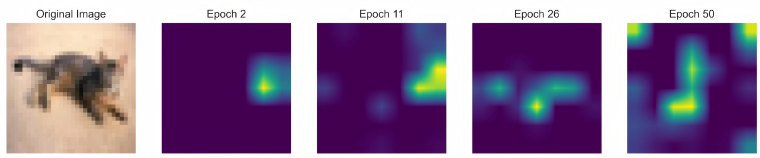
Note: Dataset statistics and average Temporal Explanation Stability Scores (TESS) over $\mathcal{D}_{\text{stable}}$. Compared to **Table 1**, all metrics are notably lower, with the sharpest decline in SSIM indicating severe structural drift.

5.3. Qualitative analysis: A case study in explanation collapse

The qualitative examples in **Figure 6** provide a stark visual testament to this instability. Even the "Most Stable" example ($\text{TESS}_{\text{cos}} = 0.971$) exhibits considerable structural flux, with the Grad-CAM heatmap expanding and shifting its focus across the car's chassis over time.



(a) Most stable example (top row).



(b) Least stable example (bottom row).

Figure 6. Evolution of Grad-CAM explanations for the most stable (top) and least stable (bottom) examples from MobileNetV3. The least stable example demonstrates a near-total collapse of the explanation's semantic coherence over time.

The "Least Stable" example ($\text{TESS}_{\text{cos}} = 0.482$) is a powerful illustration of a near-total explanation collapse. The model's explanation for its stable prediction becomes chaotic: focus shifts erratically from the bird's tail (Epoch 2) to its wing (Epoch 11), eventually dissolving into a fragmented pattern by Epoch 50.

This behavior illustrates that when a model exhibits overfitting, gradient-based explanations can become particularly brittle. Since Grad-CAM operates directly on backpropagated gradients, it may amplify local fluctuations that arise from memorization rather than robust feature learning. Consequently, the model may output the correct answer while its gradient-based justification varies substantially across training epochs.

In summary, this experiment provides initial evidence that the decoupling of semantic and structural drift may be a common phenomenon across architectures, but

that the severity of drift is highly dependent on the explanation method and the degree of overfitting. While intrinsic methods like attention rollout appear relatively stable, gradient-based post-hoc methods can be significantly more fragile, especially under conditions that promote overfitting. We caution that these observations derive from a single CNN architecture on a relatively simplistic dataset; broader validation is needed to establish general patterns.

6. Decomposing instability: Architectural vs. methodological effects

Our preceding experiments confirmed that temporal explanation drift generalizes across architectures but appears more severe in a CNN using Grad-CAM than in a ViT using attention rollout. This, however, conflates two variables: the model and the method. To scientifically dissect the sources of this instability, we conducted a final, crucial experiment: we applied the Grad-CAM explanation method to the original, well-behaved DeiT model checkpoints. This allows for a direct, controlled comparison between explanation methodologies on an identical model, and between architectures using an identical explanation method.

The results, summarized in **Figure 7** and **Table 3**, reveal that the choice of explanation method is not a minor detail but is, in fact, the primary driver of temporal instability.

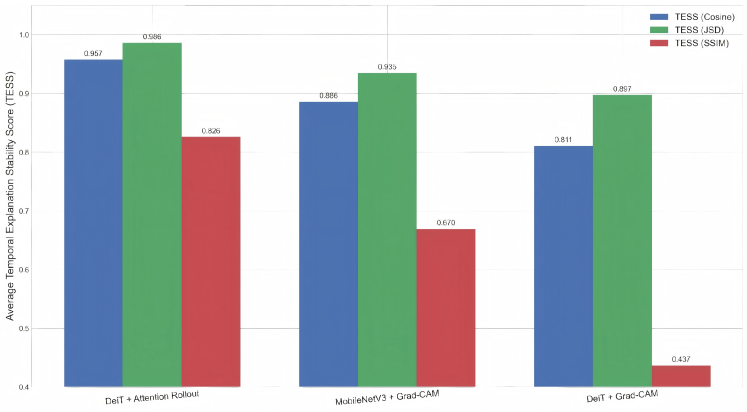


Figure 7. Comparative analysis of Temporal Explanation Stability Scores (TESS) across different model-method combinations. This chart highlights the dramatic drop in stability when switching from attention rollout to Grad-CAM on the same DeiT model.

Table 3. Aggregate stability results for DeiT with Grad-CAM.

Category	Metric	Value
Dataset statistics	Size of Analyzed Test Set	200
	Size of $\mathcal{D}_{\text{stable}}$	139
	Percentage Stable	69.50%
Aggregate stability scores	TESS_{avg} (Cosine)	0.8109 ± 0.0689
	TESS_{avg} (JSD)	0.8973 ± 0.0426
	TESS_{avg} (SSIM)	0.4369 ± 0.0715

Note: Dataset statistics and average Temporal Explanation Stability Scores (TESS) over $\mathcal{D}_{\text{stable}}$. Compared to **Table 1** (DeiT with Attention Rollout), all metrics drop sharply, with the largest declines in Cosine and SSIM, highlighting the instability introduced by Grad-CAM.

6.1. A Pronounced drop in stability with gradient-based methods

By applying Grad-CAM to the DeiT model, we observed a profound collapse in explanation stability. As shown in the comparative results in **Figure 7**, the structural stability score ($TESS_{avg}$ SSIM) plummeted from a robust 0.826 with attention rollout to a deeply unstable 0.437 with Grad-CAM—on the exact same set of saved model weights. The directional stability ($TESS_{avg}$ Cosine) likewise degraded significantly, from 0.957 to 0.811.

This finding provides unambiguous evidence that post-hoc, gradient-based methods like Grad-CAM are inherently more susceptible to temporal drift than intrinsic methods that read directly from the model's architectural state. While the model's internal attention stabilized early in training, its corresponding gradient landscape remained in a state of significant flux, which was then inherited by the Grad-CAM explanations.

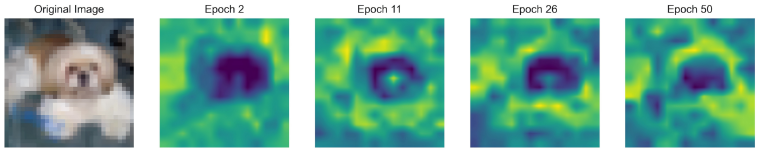
6.2. A nuanced architectural interaction

The comparison between architectures using the same explanation method also yielded a surprising insight. Contrary to the hypothesis that the more heavily over-fitting MobileNetV3 model would be the least stable, the DeiT model explained with Grad-CAM was demonstrably worse ($TESS_{avg}$ SSIM of 0.437 vs. 0.670 for MobileNetV3).

This suggests a more nuanced interaction between architecture and explanation. While the hierarchical and spatially-constrained nature of CNNs may inherently produce smoother gradient landscapes, the global, less-structured self-attention blocks of a ViT may result in sharper or more sensitive gradients. Consequently, when a ViT is probed by a gradient-based method, this underlying sensitivity is expressed as severe temporal instability.

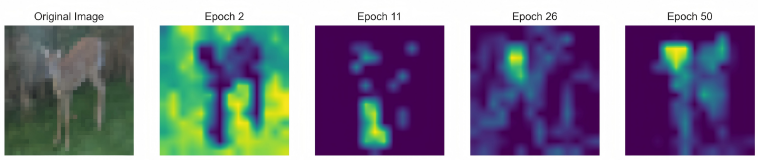
6.3. Visualizing the instability

The qualitative results in **Figure 8** visually cement this conclusion. The "Least Stable" example, with a $TESS_{cos}$ score of 0.630, is a textbook case of explanation chaos. The model's focus shifts without any discernible semantic logic from one epoch to the next, despite the model's prediction remaining correct and stable. This is the visual signature of an explanation method that is failing to provide a consistent or trustworthy account of the model's behavior. The "Most Stable" example, while quantitatively better, still shows notable changes in heatmap shape and intensity that were absent in the attention-based explanations.



(a) Most stable example (top row).

Figure 8. *Cont.*



(b) Least stable example (bottom row).

Figure 8. Qualitative examples of Grad-CAM explanations for the DeiT model. The "Least Stable" example (bottom) shows erratic and semantically incoherent shifts in focus across epochs, despite a stable and correct prediction.

In conclusion, our work reveals a clear hierarchy of stability. The most stable explanations are derived from intrinsic model properties (like attention). Stability degrades significantly when using post-hoc gradient methods, and this degradation is further modulated by the underlying architecture, with ViTs exhibiting a particular fragility to this mode of explanation. This underscores a critical, practical takeaway for the XAI field: an explanation's temporal stability is a distinct and vital property that is fundamentally tied to the method used to generate it.

7. Conclusion

Our investigation into the temporal dynamics of XAI reveals that explanation stability is not a monolithic property, but a multi-faceted characteristic of deep learning models. Across architectures and explanation methods, we consistently observed a decoupling: an explanation's high-level semantic focus stabilises early in training, while its fine-grained structural representation continues to drift long after predictive accuracy has converged.

Crucially, the choice of explanation method is the primary driver of temporal instability. Post-hoc, gradient-based methods like Grad-CAM are dramatically less stable than intrinsic methods like attention rollout. This instability is further modulated by a nuanced interaction with the model's architecture. The core practical takeaway is that temporal stability is a distinct and vital property of explanation trustworthiness, and it is fundamentally tied to the method used to generate the explanation.

The framework itself is theoretically grounded (Section 3.7) and computationally practical. A full TED analysis on a standard vision benchmark completes in under an hour on modest hardware, and lightweight protocols such as anchor-epoch evaluation and incremental drift updates are available for large-scale settings (Section 3.8).

We acknowledge several limitations of this initial study. Our experiments are limited to the CIFAR-10 dataset; the exact drift magnitudes may vary on more complex benchmarks (e.g., CIFAR-100, ImageNet) or in other modalities such as object detection and NLP. We examined only two representative explanation methods (attention rollout and Grad-CAM), while perturbation-based (LIME), axiomatic (SHAP, Integrated Gradients), and example-based approaches remain to be tested. The models used (DeiT-Tiny, MobileNetV3) are relatively small compared to modern foundation models, and scaling could introduce new stability dynamics. Finally, this study does not include statistical significance tests or replicate runs, which would be needed to formally quantify metric variability. Despite these limitations, the TED

framework is general and can be applied directly to any saliency map, architecture, or dataset.

Several immediate directions follow from this work. First, systematically evaluating the effect of common regularisation techniques (weight decay, dropout, data augmentation) on TESS will connect training practice to explanation stability. Second, extending the analysis to a broader family of explanation methods (LIME, SHAP, Integrated Gradients) and more complex datasets (CIFAR-100, ImageNet, and beyond) will map the full landscape of temporal stability. Third, practical mitigation strategies deserve dedicated investigation: one could design an explanation-aware training penalty that directly minimises consecutive SSIM drift, or apply post-hoc temporal ensembling of saliency maps from nearby checkpoints to smooth structural fluctuations. These directions shift the role of explanation stability from a passive diagnostic to an actively optimised property, paving the way toward models that are both accurate and transparently consistent in their reasoning.

Author contributions: Conceptualization, ZR and MO; methodology, ZR and MO; software, ZR; validation, ZR and MO; formal analysis, ZR; investigation, ZR; resources, MO; data curation, ZR; writing—original draft preparation, ZR; writing—review and editing, ZR and MO; visualization, ZR; supervision, MO; project administration, ZR. Both authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional review board statement: Not applicable. This study did not involve human participants or animal subjects and therefore did not require ethics approval.

Informed consent statement: Not applicable.

Data availability statement: The data used in this study are publicly available. The CIFAR-10 dataset can be accessed from its official repository.

Conflict of interest: The authors declare that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AI use statement: During the preparation of this work, the authors used AI-based language tools, mainly ChatGPT, to improve the clarity, grammar, and readability of the manuscript. No AI tools were used for data analysis, interpretation, or generation of scientific content. All scientific contributions, ideas, experiments, and conclusions are solely those of the authors. After using these tools, the authors carefully reviewed and edited the content and take full responsibility for the final version of the manuscript.

References

1. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017; 42: 60–88.

2. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542(7639): 115–118.
3. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017; 318(22): 2199–2210.
4. Bojarski M, Del Testa D, Dworakowski D, et al. End to end learning for self-driving cars. arXiv preprint. 2016. doi: 10.48550/arXiv.1604.07316
5. Grigorescu S, Trasnea B, Cocias T, et al. A survey of deep learning techniques for autonomous driving. Journal of Field Robotics. 2020; 37(3): 362–386.
6. Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies. Journal of Machine Learning Research. 2016; 17(1): 1334–1373.
7. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019; 1(5): 206–215.
8. Goodfellow IJ, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. arXiv preprint. 2014. doi: 10.48550/arXiv.1412.6572
9. Gunning D, Stefik M, Choi J, et al. XAI—Explainable artificial intelligence. Science Robotics. 2019; 4(37).
10. Arrieta AB, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020; 58: 82–115.
11. Abnar S, Zuidema W. Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 5–10 July 2020; Online. pp. 4190–4197.
12. Chefer H, Gur S, Wolf L. Transformer Interpretability Beyond Attention Visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 20–25 June 2021; Nashville, TN, USA. pp. 782–791.
13. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the International Conference on Computer Vision (ICCV); 22–29 October 2017; Venice, Italy. pp. 618–626.
14. Jacovi A, Goldberg Y. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 5–10 July 2020; Online. pp. 4198–4205.
15. Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence. 2019; 33(1): 3681–3688.
16. Dombrowski AK, Alber M, Anders CJ, et al. Explanations can be manipulated and geometry is to blame. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019); 8–14 December 2019; Vancouver, BC, Canada.
17. Kindermans PJ, Hooker S, Adebayo J, et al. The (un)reliability of saliency methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer; 2019. pp. 267–280.
18. Adebayo J, Gilmer J, Muelly M, et al. Sanity checks for saliency maps. In: Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018); 2–8 December 2018; Montréal, QC, Canada. pp. 9505–9515.
19. Nie W, Zhang Y, Patel AB. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In: Proceedings of the 35th International Conference on Machine Learning (ICML); 10–15 July 2018; Stockholm, Sweden. pp. 3809–3818.
20. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint. 2020. doi: 10.48550/arXiv.2010.11929
21. Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. Proceedings of the 38th International Conference on Machine Learning. 2021; 139: 10347–10357.
22. Khan S, Naseer M, Hayat M, et al. Transformers in Vision: A Survey. ACM Computing Surveys (CSUR). 2022; 54(10s): 1–41. doi: 10.1145/3505244
23. Koh PW, Nguyen T, Tang YS, et al. Concept Bottleneck Models. Proceedings of the 37th International Conference on Machine Learning. 2020; 119: 5338–5348.
24. Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3. In: Proceedings of the International Conference on Computer Vision (ICCV); 27 October–2 November 2019; Seoul, Republic of Korea. pp. 1314–1324.
25. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 13–17

August 2016; San Francisco, CA, USA. pp. 1135–1144.
26. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017); 4–9 December 2017; Long Beach, CA, USA. pp. 4768–4777.
27. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. Proceedings of the 34th International Conference on Machine Learning. 2017; 70: 3319–3328.
28. Samek W, Montavon G, Lapuschkin S, et al. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proceedings of the IEEE. 2021; 109(3): 247–278.
29. Bodria F, Giannotti F, Guidotti R, et al. Benchmarking and Survey of Explanation Methods for Black Box Models. Data Mining and Knowledge Discovery. 2023; 37(5): 1719–1778.
30. Hooker S, Erhan D, Kindermans PJ, et al. A Benchmark for Interpretability Methods in Deep Neural Networks. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 8–14 December 2019; Vancouver, BC, Canada.
31. Hedström A, Weber L, Bareeva D, et al. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. Journal of Machine Learning Research. 2023; 24: 1–11.
32. Agarwal C, Krishna S, Saxena E, et al. OpenXAI: Towards a transparent evaluation of model explanations. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022); 28 November–9 December 2022; New Orleans, LA, USA.
33. Raghu M, Gilmer J, Yosinski J, et al. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017); 4–9 December 2017; Long Beach, CA, USA. pp. 6076–6085.
34. Kornblith S, Norouzi M, Lee H, et al. Similarity of neural network representations revisited. Proceedings of the 36th International Conference on Machine Learning. 2019; 97: 3519–3529.
35. Nanda N, Chan L, Lieberum T, et al. Progress Measures for Grokking via Mechanistic Interpretability. arXiv preprint. 2023. doi: 10.48550/arXiv.2301.05217
36. Gururangan S, Marasović A, Swayamdipta S, et al. Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models. arXiv preprint. 2022. doi: 10.48550/arXiv.2208.03306
37. Frankle J, Dziugaite GK, Roy DM, et al. Linear Mode Connectivity and the Lottery Ticket Hypothesis. Proceedings of the 37th International Conference on Machine Learning. 2020; 119: 3259–3269.
38. Zhang C, Bengio S, Hardt M, et al. Understanding Deep Learning (Still) Requires Rethinking Generalization. Communications of the ACM. 2021; 64(3): 107–115.
39. Lin J. Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory. 1991; 37(1): 145–151.
40. Wang Z, Bovik AC, Sheikh HR, et al. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing. 2004; 13(4): 600–612.