NEUTRALIZING 0TH-ORDER RESIDUAL INTERFERENCES IN LAYER-WISE RELEVANCE PROPAGATION VIA GEOMETRIC TRAJECTORY REALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard Layer-wise Relevance Propagation (LRP) operates under the premise that 0th-order residual terms, even if not strictly eliminated, do not affect the essential properties of relevance propagation. However, this study reveals that when introducing a simple constant 0th-order term as a minimal 0th-order probe, the linearity of the distribution persists under constrained input methods, causing noise to accumulate during relevance propagation. This phenomenon induces a deviation (θ) within the α-β parameter space and misaligns geometric trajectories, thereby structurally violating LRP's conservation principle and ultimately leading to fatal signal collapse in deep layers. To empirically validate this theoretical distortion mechanism, we introduce the π-rule—a normalization mechanism that neutralizes the linear effects of the 0th-order terms by aligning geometric trajectory slopes to unity. Our analysis of CNN architectures (e.g., VGG, ResNet, DenseNet) confirms the validity of our hypothesis: the π-rule maintains a constant variance of θ, effectively suppressing distribution-induced noise and improving explanation fidelity. Conversely, the marginal improvements observed in Vision Transformers (ViTs) expose the limitations of the simplified constant 0th-order assumption in complex non-linear models. Despite this boundary of validity, our empirical study unequivocally proves that the distribution of 0th-order terms alone generates critical noise in LRP, establishing an essential theoretical foundation for designing novel rules to address non-linear residual dynamics in future Transformer architectures.

1 INTRODUCTION

Deep Neural Networks (DNNs) have demonstrated exceptional performance across various domains, yet their opaque nature has driven the need for reliable Explainable Artificial Intelligence (XAI) techniques Swartout & Moore (1993); Lapuschkin et al. (2019). Among these, Layer-wise Relevance Propagation (LRP) Bach et al. (2015) stands out as a powerful tool grounded in the "Divide-and-Conquer" paradigm Montavon et al. (2019), which decomposes a network's prediction into local, layer-wise sub-functions. Standard LRP operates under the premise that 0th-order residual terms, even if not strictly eliminated during relevance redistribution, do not compromise the essential properties of relevance propagation or the structural consistency of the modules Montavon et al. (2017).

To improve the explanation properties of LRP, constrained input methods have been proposed and achieved partial success Montavon et al. (2017). Despite these advancements, as networks become deeper and incorporate structures like skip-connections, a critical distortion phenomenon has been persistently reported across numerous datasets: the target signal progressively weakens and collapses to zero (signal collapse) Montavon et al. (2019), or the sign of the relevance flips abnormally (sign flip) Achtibat et al. (2024). The research community has largely attributed these relevance distortions to the complexity of DNNs or the inherent limitations of the LRP framework itself. The persistence of such destructive signal leakage and distortion, despite numerous countermeasures, implies that existing approaches have failed to address the fundamental structural contradiction hidden behind the divide-and-conquer assumption.

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.

This paper identifies that the root cause of this unresolved distortion lies in the very "0th-order terms" that have historically been dismissed as negligible. By excluding complex non-linearities, we analyzed the propagation dynamics by introducing a simple, constant 0th-order term as a *minimal 0th-order probe*. Our analysis reveals that under constrained input methods, the 0th-order term does not vanish; instead, it couples with the linearity of the neuron distribution to induce an unintended *linear effect*.

This linear effect stems from the size imbalance between the activator and inhibitor sets, creating a discrepancy of 0th-order terms within the $\alpha-\beta$ parameter space. We define this geometric trajectory deviation as θ. As the network deepens, the accumulating variance of θ distorts the directionality of relevance propagation. This *structurally violates* the global conservation principle of LRP, ultimately leading to signal collapse.

If our theoretical distortion mechanism holds true, realigning the misaligned geometric trajectories should eliminate the signal collapse phenomenon. To empirically validate this hypothesis, we introduce the π-rule, a mechanism that mathematically neutralizes the linear effects of the 0th-order terms by **aligning geometric trajectory slopes to unity**. Rather than being a mere heuristic tweak, this normalization technique serves as a rigorous testing ground to determine the *boundary of validity* of our theoretical assumption using the minimal probe in actual models.

Extensive experimental results confirm that while the variance of θ increases with depth under standard rules, the π-rule consistently suppresses it, preventing signal collapse in CNN architectures Simonyan & Zisserman (2015); He et al. (2016); Huang et al. (2017) and validating our mathematical hypothesis. Although the performance improvements observed in models like Vision Transformers (ViTs) Dosovitskiy et al. (2021) were marginal due to the limitations of the simplified constant 0th-order assumption, our study successfully demonstrates that the distribution of 0th-order terms alone generates critical noise in LRP.

Our main contributions are summarized as follows:

- **Identification of Structural Dependency:** Using a minimal 0th-order probe, we mathematically prove that under constrained input methods, distribution-induced noise causes geometric trajectory discrepancies (deviation θ), structurally violating LRP's conservation principle.
- **Empirical Validation via Geometric Alignment:** We introduce the π-rule to experimentally neutralize the linear effect of 0th-order terms by aligning geometric trajectory slopes to unity. This successfully prevents signal collapse and significantly enhances explanation fidelity across CNN architectures, including VGG, ResNet, and DenseNet.
- **Delineation of Boundary of Validity:** Through ViT experiments, we transparently analyze the limitations of the constant 0th-order assumption in non-linear structures. This establishes the critical importance of distribution-induced noise and provides an essential theoretical foundation for designing novel LRP rules to address complex non-linear residual dynamics in future Transformer architectures.

2 Related Work

LRP decomposes global predictions into local sub-functions. Various propagation rules (ϵ, γ, $\alpha\beta$, and composite LRP) have evolved to stabilize this process and enhance explanation fidelity by adjusting heuristic parameters Montavon et al. (2019). However, controlling explainability via constrained input methods (e.g., γ-rule and $\alpha\beta$-rule) relies on the strong premise that 0^{th}-order residual terms do not fundamentally alter propagation properties Montavon et al. (2017). Within the Deep Taylor Decomposition (DTD) framework, these 0^{th}-order biases are traditionally treated as negligible, either completely omitted or forcibly absorbed into adjacent nodes to maintain local numerical conservation. While axiomatic methods like Integrated Gradients Sundararajan et al. (2017) or DeepLIFT Shrikumar et al. (2017) account for baselines globally, they fundamentally abandon LRP's core divide-and-conquer modularity.

Conversely, standard LRP's local absorption strategy superficially satisfies numerical conservation but overlooks the structural noise that persistently propagates when residuals couple with constrained input methods. As neural architectures deepen Simonyan & Zisserman (2015); He et al.

(2016), the influence of these neglected residuals drastically amplifies. We mathematically demonstrate that even a minimal constant 0^{th}-order probe couples with the distributional linearity of constrained input spaces, generating persistent linear effects. This linear effects accumulates layer-by-layer, ultimately triggering the catastrophic signal collapse frequently observed in deep architectures Achtibat et al. (2024). Therefore, achieving robust relevance conservation in deep networks necessitates a fundamental geometric restoration to neutralize the linear structural noise induced by 0^{th}-order terms.

3 Theoretical Analysis: Structural Dependency via the Minimal Probe

3.1 DTD Expansion and the Minimal 0th-Order Probe

LRP decomposes the prediction of a neural network into local, layer-wise sub-functions and redistributes relevance to lower layers via first-order Taylor approximations.

Based on this DTD framework, let the activation function of neuron j located at the m-th layer be $f_j^m(x)$ for an input vector x. Expanding this as a Taylor series around a specific reference point $\tilde{x}$ yields:

$$f_j^m(x) = f_j^m(\tilde{x}) + \sum_i J_{ji}(\tilde{x})(x_i - \tilde{x}_i) + H \tag{1}$$

Here, $J_{ji}(\tilde{x})$ is the Jacobian matrix evaluated at the reference point $\tilde{x}$, and H represents the higher-order terms arising from non-linearities. LRP traditionally assumes that H and the reference point parts can be ignored or minimized. By rearranging the above equation into the linear contribution of pure input features and the remaining residual part, it can be expressed as:

$$f_j^m(x) = \sum_i J_{ji}x_i + \underbrace{f_j^m(\tilde{x}) - \sum_i J_{ji}\tilde{x}_i + H}_{\tilde{b}_j} \tag{2}$$

The bracketed term $\tilde{b}_j$ denotes the **residual term**, which integrates the network's bias, the shift of the reference point, and higher-order interactions. To satisfy the global relevance conservation principle ($\sum_i R_i = R_j$), the relevance R_j allocated to neuron j is redistributed to the lower layer in proportion to the activation function. If we define this proportionality constant as the relevance ratio $c = \frac{R_j}{f_j^m(x)}$, the relevance redistribution equation expands into an intuitive scaling format as follows:

$$R_j = c \cdot f_j^m(x) = c\left(\sum_i J_{ji}x_i + \tilde{b}_j\right) \tag{3}$$

Conventional Perspective: Treatment as a Cohesive Function
Historically, mainstream literature and standard methodologies dealing with LRP have treated the entire expression $\left(\sum_i J_{ji}x_i + \tilde{b}_j\right)$ as a *single cohesive function* to be distributed to the lower layer. Namely, under the divide-and-conquer paradigm, the residual term $\tilde{b}_j$ has not been interpreted independently from the 1st-order relevance term. It has been firmly presumed that even if the residual term is not strictly eliminated, it is simply absorbed by being scaled together by the relevance ratio c, exerting no detrimental effect on the structural consistency or modular independence.

Our Perspective: Structural Rupture Induced by the 0th-Order Probe
However, this study points out a fatal flaw in the conventional reductionist premise that treats $\left(\sum_i J_{ji}x_i + \tilde{b}_j\right)$ as a single bundled entity. When constrained input methods, such as the $\alpha\beta$-rule, are applied to enhance explanation fidelity, the complex residual term $\tilde{b}_j$ inside the parentheses is by no means passively absorbed as a simple scaling factor. Instead, during the process where

relevance is partitioned into activator and inhibitor sets, this residual acts as an independent noise generator, causing a **structural rupture** in the ideal parallel propagation trajectory of standard LRP.

To mathematically and rigorously dissect the nature of this rupture, we temporarily control the nonlinear variables that compound the network's complexity and isolate purely the influence of the 0th-order residual. To this end, under the assumption that the variance of the scaled residual vector occurring during the relevance scaling process is negligible within a layer, we introduce the following assumption.

Assumption 3.1 (Bias Term Constancy)**.** For a set of neurons N belonging to an arbitrary layer m, the scaled residual vector $c\tilde{b}_j$ varies negligibly among neurons $j \in N$. Therefore, we approximate it as a single constant vector b as follows:

$$c\tilde{b}_j \approx b, \quad \forall j \in N \tag{4}$$

In other words, the entire complex residual scaled by the relevance ratio is replaced and treated as a **minimal 0th-order probe** (b), which is a single constant vector acting uniformly across the layer.

We will algebraically prove in the next section how the simplified 0th-order probe (b) defined through Assumption 3.1 survives without vanishing within the constrained input space, generating linear noise.

3.2 Linear effect under Constrained Input Methods

Building upon the minimal probe assumption (Assumption 3.1) defined in Section 3.1, we expand the propagation rule of the $\alpha\beta$-rule—a representative constrained input method—to mathematically trace the impact of the 0th-order term on layer-wise relevance propagation.

The $\alpha\beta$-rule redistributes relevance by strictly separating the positive contributions, denoted by the activator set $\mathcal{P}$, and the negative contributions, denoted by the inhibitor set $\mathcal{M}$. Let us denote the normalized 1st-order relevance distribution term as contribution score $Z_{ji}^{ml} = \frac{J_{ji}x_i}{f(x_j)}$. By explicitly substituting the 0th-order probe b^m into the equation for the relevance R_i^l propagated from layer m to a neuron i in the lower layer l, we obtain:

$$R_i^l = \alpha \sum_{j \in \mathcal{P}(m)} \left(Z_{ji}^{ml} R_j^m + b^m\right) - \beta \sum_{j \in \mathcal{M}(m)} \left(Z_{ji}^{ml} R_j^m + b^m\right) \tag{5}$$

By performing an algebraic decomposition on the above equation to separate the ideal bias-free 1st-order relevance propagation from the residual part generated by the 0th-order probe, we can reformulate it as follows:

$$R_i^l = \underbrace{\left(\alpha \sum_{j \in \mathcal{P}(m)} Z_{ji}^{ml} R_j^m - \beta \sum_{j \in \mathcal{M}(m)} Z_{ji}^{ml} R_j^m\right)}_{\overline{R}_i^l \text{ (1st-order bias-free relevance)}} + \underbrace{\left(\alpha \sum_{j \in \mathcal{P}(m)} b^m - \beta \sum_{j \in \mathcal{M}(m)} b^m\right)}_{\text{0th-order noise}} \tag{6}$$

The first bracketed term represents the pure 1st-order relevance score, $\overline{R}_i^l$, which is the ideal target we aim to acquire when the 0th-order term is perfectly excluded. In contrast, the second bracketed term illustrates the accumulated consequence of the 0th-order probe within the partitioned modular structure. Since b^m is treated as a single constant vector within layer m under Assumption 3.1, let $P^m = |\mathcal{P}(m)|$ and $M^m = |\mathcal{M}(m)|$ denote the sizes of the activator and inhibitor sets, respectively. The final relevance R_i^l then simplifies to:

$$R_i^l = \overline{R}_i^l + (\alpha P^m - \beta M^m)\, b^m \tag{7}$$

This equation carries a physical implication that directly contradicts the conventional premise that the 0th-order residual is simply absorbed or vanishes. Constrained input methods inevitably entail a numerical imbalance between activator and inhibitor neurons ($P^m \neq M^m$). Under the highly

exceptional circumstance where $\alpha = \beta$ and $P^m = M^m$, the noise would be perfectly canceled out. However, in actual deep neural networks, distributional asymmetry constantly exists depending on the data and weights, and this distortion is often exacerbated by the user-defined heuristic ratio $(\alpha : \beta)$.

Consequently, the surviving residual $(\alpha P^m - \beta M^m)\, b^m$ is not merely a simple additive constant distributed to the lower layer. Rather, it is a **linear noise** created by the structural coupling of the 0th-order probe b^m with the linearity of the constrained input space (network topology P^m, M^m) and the heuristic parameters (α, β).

In the next section, we analyze how this noise term geometrically induces a trajectory misalignment within the $\alpha - \beta$ parameter space.

3.3 Geometric Misalignment in the $\alpha - \beta$ Parameter Space

Through the linear noise term derived in Section 3.2, we established that the 0th-order probe b^m is not merely an additive constant distributed to lower layers. Rather, it is a vector whose scale and directionality are determined by the neuron distribution sizes (P^m, M^m) and heuristic parameters (α, β).

To provide a geometric intuition for this linear noise, we project the relevance redistribution process onto a two-dimensional parameter space, with the user-defined heuristic parameters α and β as its axes. From the perspective of the probe model, let i^m denote the linear scalar coefficient multiplied by b^m, which is expressed by the equation $i^m = \alpha P^m - \beta M^m$. Rearranging this into a linear equation within the parameter space yields:

$$\beta = \left(\frac{P^m}{M^m}\right)\alpha - \frac{i^m}{M^m} \tag{8}$$

This equation reveals that the slope of the relevance trajectory is determined exclusively by the ratio of the neuron set sizes (P^m/M^m). The **ideal condition** for relevance propagation to maintain structural independence under the divide-and-conquer paradigm is for this trajectory to maintain a slope of one (parallel to the $\alpha = \beta$ direction). A slope of unity implies that the residual induces no structural asymmetry during the positive and negative redistribution process.

However, because constrained input methods inherently entail the imbalance $P^m \neq M^m$, the slope of the actual trajectory deviates from one. To mathematically quantify this phenomenon, we introduce the following concept.

Definition 3.2 (Geometric Misalignment Angle θ)**.** We define the **angular deviation** θ as the degree to which the actual relevance propagation trajectory (with slope P^m/M^m) deviates from the ideal parallel trajectory (with slope one). This is formulated as:

$$\theta = \arctan\left(\frac{P^m - M^m}{P^m + M^m}\right) \tag{9}$$

Here, $\theta \in [-\pi/4, \pi/4]$. This angle quantifies the magnitude of the **directional distortion** induced by the 0th-order probe within the $\alpha - \beta$ space.

In a single layer, the trajectory deviation caused by θ may appear as a mere angular tilt. However, in deeply stacked architectures such as VGG, the propagation modules of each layer—which should ideally remain independent—exhibit different θ values. Consequently, they become entangled in non-parallel (intersecting) trajectories.

As a result, Section 3.4 examines how this geometric misalignment accumulates across layers, ultimately violating the global conservation principle of LRP and triggering signal collapse.

3.4 Structural Violation and Signal Collapse

In Section 3.3, we established that the 0th-order probe induces a geometric trajectory misalignment by an angle of θ in the $\alpha - \beta$ parameter space. In a single layer, this angular deviation might merely cause a localized distortion. However, in deep neural networks with deeply stacked layers, such as

VGG architectures, this noise recursively accumulates along the relevance propagation pathway. In this section, we model this accumulation process as a one-dimensional random walk to prove how the accumulation of θ leads to a structural breakdown of the global conservation principle.

1. Recursive Noise Accumulation and Random Variable Definition
Let d denote the depth of the network. We define the linear noise generated at layer m as an independent random variable W_m. According to the derivation in Section 3.2, this noise depends on the layer's neuron width N and the activator/inhibitor ratio, meaning it can reach extreme amplitudes of $\pm Nb$. Consequently, the total accumulated noise S_d reaching the input layer (depth d) from the topmost layer can be expressed as a Markov chain (cumulative sum of the state space):

$$S_d = \sum_{m=1}^{d} W_m \tag{10}$$

2. Generalized Expectation and Variance via the Random Walk Model
To rigorously analyze this accumulation process, let us assume an asymmetric random walk where the probability of generating positive noise ($+Nb$) is p, and the probability of negative noise ($-Nb$) is $q = 1-p$ at each layer. After d propagation steps following independent and identically distributed (i.i.d.) variables, the expected value (mean) and variance of the accumulated noise S_d are generalized as follows:

$$\mathbb{E}[S_d] = d \cdot Nb(p - q) \tag{11}$$

$$\text{Var}(S_d) = 4dpq(Nb)^2 \tag{12}$$

If we assume that the network is sufficiently large and complex such that the occurrence of positive and negative deviations is symmetric ($p = q = 1/2$), the expected value of the total accumulated noise becomes exactly zero:

$$\mathbb{E}[S_d] = 0 \tag{13}$$

This result elucidates why previous studies mistakenly assumed that 0th-order residuals could be safely ignored. Because the "average bias" of the noise converges to zero from a macroscopic perspective, the conservation principle appears superficially preserved.

3. Variance Amplification and Signal Collapse
However, the fact that the expected value is zero does not imply the disappearance of noise. Substituting $p = q = 1/2$ into the variance formula, the volatility of the accumulated noise explosively increases in direct proportion to the depth d:

$$\text{Var}(S_d) = d(Nb)^2 \tag{14}$$

This mathematically proves that the relevance propagation trajectory fails to proceed in a straight line toward the target (input features), losing its way amidst random amplitudes as it passes through layers. As the depth d increases ($d \to \infty$), the accumulated noise variance $\text{Var}(S_d)$ completely overwhelms the signal strength of the pure 1st-order relevance ($\overline{R}^l$).

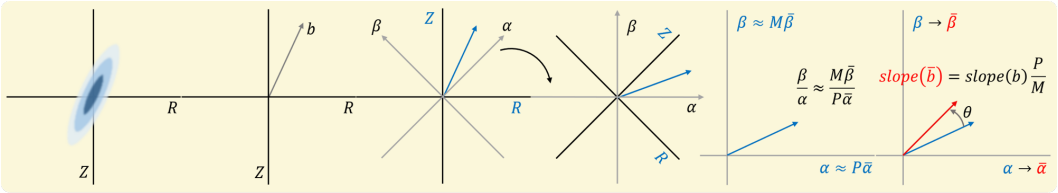
Consequently, the uncontrollably amplified noise variance caused by the geometric trajectory misalignment (θ) induces destructive interference between layers. This results in a **structural violation**, rather than a mere numerical one, of the Global Conservation Principle—the most fundamental premise of LRP. Ultimately, it constitutes the root cause of the **Signal Collapse** phenomenon, where the target signal vanishes to zero in deep layers.

To prevent this catastrophe and restore the conservation principle of LRP in deep neural networks, the sole mathematical solution is to suppress the angular deviation θ itself—the seed of the variance generated at each layer—by forcibly **aligning the trajectories to unity**.

4 Geometric Restoration via the π-rule

4.1 Motivation: Controlling the Seed of Variance

Through the Markov chain analysis in Section 3.4, we mathematically confirmed that the root cause of Signal Collapse in deep neural networks is the "Variance Amplification" driven by the recursive



Constant Bias Assumption | Coordinate Transform | Geometric Trajectory Realignment

Figure 1: Geometric interpretation of structural dependency via the minimal 0^{th}-order probe. (Left) The complex residual is reduced to a constant bias assumption. (Center) The redistribution process is projected into the α-β parameter space. (Right) The distributional imbalance ($P \neq M$) geometrically skews the trajectory, inducing an angular deviation θ from the ideal parallel propagation.

accumulation of the angular deviation θ. This seed of destructive variance, θ, ultimately originates from the cardinal imbalance ($P^m \neq M^m$) between the activator set ($\mathcal{P}$) and the inhibitor set ($\mathcal{M}$) within the constrained input space.

Therefore, to restore LRP's divide-and-conquer modularity and conservation principle even in deep layers, a mathematical normalization mechanism is essential. This mechanism must strictly compensate for the structural distortion caused by the disparity in neuron set sizes while preserving the user's intended explanatory properties (i.e., the heuristic ratio $\alpha : \beta$).

4.2 Mathematical Formulation of the π-rule

The fundamental limitation of the standard $\alpha\beta$-rule lies in its uniform application of fixed weights α and β to an uneven neuron population. To decouple the relevance magnitude from the set cardinalities, we denote the sizes of each set as $\pi_P = |\mathcal{P}(m)|$ and $\pi_M = |\mathcal{M}(m)|$, and introduce a global normalization factor π based on the Harmonic Mean $\pi = 2\pi_P\pi_M/(\pi_P + \pi_M)$.

This harmonic structure balances the influx of positive and negative contributions, playing a crucial role in perfectly maintaining the global relevance conservation law. Utilizing this factor, we define the newly adjusted parameters α_π and β_π as follows:

$$\alpha_\pi = \frac{\alpha\pi}{\pi_P(\alpha+\beta)}, \quad \beta_\pi = \frac{\beta\pi}{\pi_M(\alpha+\beta)} \tag{15}$$

Substituting these parameters into the relevance propagation formula yields the final formulation of the π-rule:

$$R_i^l = \alpha_\pi \sum_{j\in\mathcal{P}(m)} \left(Z_{ji}^{ml} R_j^m + b^m\right) - \beta_\pi \sum_{j\in\mathcal{M}(m)} \left(Z_{ji}^{ml} R_j^m + b^m\right) \tag{16}$$

By dynamically recalibrating the set imbalances, this formulation suppresses the macroscopic noise induced by the 0^{th}-order probe b, and propagates the originally intended $\alpha : \beta$ ratio to lower layers without structural distortion.

4.3 Theoretical Resolution: Variance Suppression via Trajectory Alignment

We now mathematically prove how the proposed π-rule resolves the geometric misalignment and variance explosion raised in Section 3.

By substituting the compensated α_π and β_π into the linear equation of the parameter space ($i = \alpha P^m - \beta M^m$), the π_P and π_M in the denominators exactly cancel out the actual set cardinalities P^m and M^m. Factoring out the common terms, the linear interaction bias induced by the 0^{th}-order term simplifies to $R_{bias} \propto (\alpha - \beta)b^m$.

This equation signifies a decisive geometric resolution. The threatening structural variable P^m/M^m, which previously dictated the trajectory slope, is completely eliminated from the equation, leaving only a linear combination of α and β.

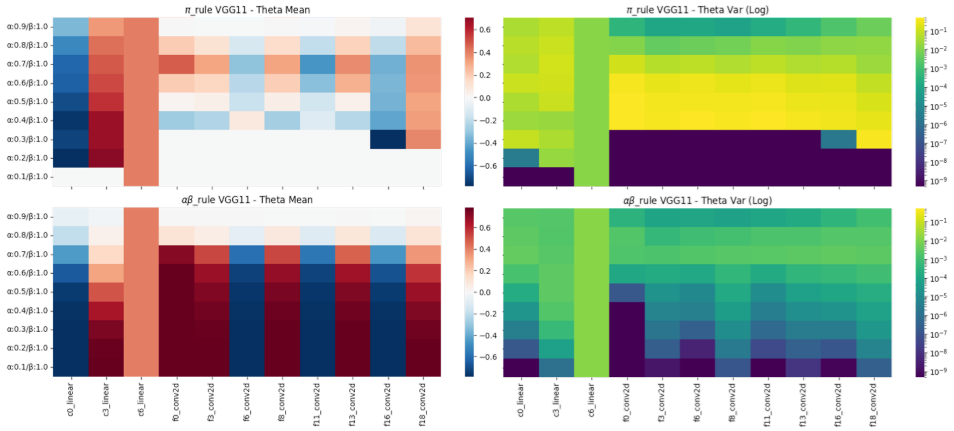


Figure 2: **Verification of Geometric Distortion in VGG.** Under the standard $\alpha\beta$-rule, unneutralized structural bias causes the negative operations of the inhibitor to simply invert the signals, exhibiting a relatively small variance. The π-rule effectively neutralizes this 0^{th}-order bias. (Zero variances are set to 5×10^{-10} for logarithmic visualization.)

Consequently, in the parameter space where the π-rule is applied, the slopes of all relevance propagation trajectories are forcibly **aligned to unity**, strictly independent of the neuron distribution. As the slope becomes one, the inconsistency angle defined in Section 3.3 vanishes ($\theta = 0$), and the trajectories recover a perfectly parallel state.

With θ aligned to zero, the directional distortion induced by the 0^{th}-order term is fully neutralized. Therefore, the cumulative noise variance $\text{Var}(S_d)$, which infinitely amplifies with depth d in the Markov chain model of Section 3.4, is fundamentally suppressed. We have theoretically proven that the π-rule structurally prevents Signal Collapse in deep sequential neural networks such as VGG. In the following Section 5, we empirically verify this variance suppression and signal restoration effect through large-scale ImageNet experiments.

5 Empirical Verification

5.1 Evaluation Protocol and LRP Configurations

The LRP algorithm is inherently deterministic. To fully leverage this and eliminate the need for statistical validation or sampling, we conducted a large-scale evaluation using the entire ImageNet validation dataset (50,000 images) Deng et al. (2009).

To systematically validate our theoretical claims, we bifurcate our experimental configurations:

Uniform LRP: To purely isolate and analyze the geometric dynamics and variance accumulation across layers (Section 5.2), we applied the standard rules uniformly across all layers Montavon et al. (2019).

Composite LRP: To evaluate practical explanation fidelity (Section 5.3), we employed composite strategies (LRP_{com} and LRP_{ϵ}) Montavon et al. (2019) that modify heuristic parameters or apply the ϵ-rule in the upper layers to enhance target-class sensitivity.

Detailed configurations for these baselines are provided in Appendix A.

5.2 Verification of Geometric Distortion: Variance Analysis

To empirically observe the variance accumulation predicted by our Markov chain model, we applied Uniform LRP to the VGG network Simonyan & Zisserman (2015). We swept the heuristic ratio $\alpha : \beta$ from $1.0 : 0.1$ to $0.1 : 1.0$, specifically focusing on the strongly inhibited regime ($\alpha < \beta$) to directly expose the influence of inhibitory contributions.

Table 1: **Quantitative Evaluation on Full Validation Set.** MoRF AUC scores on core architectures.

	$\text{LRP}^{\alpha\beta}_{com}$	LRP^{π}_{com}	$\text{LRP}^{\alpha\beta}_{(\epsilon)}$	$\text{LRP}^{\pi}_{(\epsilon)}$	$\text{CLRP}^{\alpha\beta}$	CLRP^{π}
VGG11	15.652	11.131	8.835	8.835	10.167	10.166
VGG19	20.676	14.770	9.012	9.012	10.424	10.398
ResNet34	21.793	9.168	8.721	8.718	20.441	19.755
ResNet101	28.401	12.992	12.629	12.695	26.619	26.219
DenseNet121	24.792	23.202	18.946	18.945	27.103	27.102
DenseNet169	25.486	24.491	21.425	21.424	26.841	26.835
ViT base	45.221	46.807	44.472	44.470	45.134	45.200
ViT large	48.183	49.641	47.055	47.054	47.919	47.936

As illustrated in Figure 2 (Theta Mean), the unneutralized bias under the standard $\alpha\beta$-rule causes the negative contributions of the inhibitor to flip the signals, exhibiting alternating red and blue patterns across layers. Crucially, in the variance plot (log scale), the standard $\alpha\beta$-rule exhibits a linear amplification of variance as depth increases, precisely corroborating the Markov chain derivation.

Conversely, the π-rule effectively suppresses this geometric distortion, maintaining a highly stable and constant variance across layers. We note a sharp variance spike for the π-rule in extreme inhibitory settings (e.g., $\alpha : \beta = 0.4 : 1.0$). This occurs because the π-rule neutralizes the initial positive bias, bringing the expected mean near zero; consequently, the network becomes highly sensitive to the extreme scaling of negative contributions. Extensive visual proofs of the resulting sign flips under Uniform LRP, along with zero-valued explanation maps in complex architectures like ResNet He et al. (2016), are provided in Appendix B.

5.3 Quantitative Fidelity and Boundary of Validity

To quantitatively assess explanation fidelity, we evaluated diverse architectures—including VGG, ResNet, DenseNet Huang et al. (2017), and Vision Transformer (ViT) series Dosovitskiy et al. (2021)—using perturbation-based metrics (MoRF, LeRF Samek et al. (2017), Insertion, Deletion Petsiuk et al. (2018), and Positive/Negative perturbations Gur et al. (2021)). For optimal target sensitivity, we compared the π-rule against Composite LRPs (LRP_{com}, LRP_{ϵ}), CLRP Gu et al. (2019), and SGLRP Iwana et al. (2019).

Table 1 presents a subset of the MoRF outcomes for the composite and CLRP baselines. The π-rule demonstrates consistently superior targeting performance across all evaluated CNN architectures. By neutralizing the structural linear noise, the π-rule prevents signal collapse and ensures robust fidelity regardless of architectural depth.

However, the improvements observed in the ViT series are relatively marginal. This delineates the boundary of validity for our mathematical framework: while the minimal constant bias assumption excellently models CNNs, it struggles to fully capture the highly non-linear residual dynamics introduced by components like GELU activations in Transformers Lee (2023).

Due to space constraints, the comprehensive suite of quantitative results and multi-class sign flip visualizations using Composite LRP are fully documented in Appendix C.

6 Conclusion

This paper challenges the fundamental "Divide-and-Conquer" premise in LRP. By introducing a Minimal 0^{th}-Order Probe, we mathematically demonstrate that uncontrolled residuals inherently couple with network topology, inducing geometric trajectory deviations (θ) and Markov chain-like variance amplification. This structural violation of LRP's conservation principle ultimately triggers signal collapse in deep networks. To resolve this, we propose the π-rule, a geometric restoration mechanism that neutralizes interaction noise by forcibly realigning trajectory slopes to unity. Extensive deterministic evaluations on ImageNet confirm that the π-rule successfully suppresses noise variance, prevents signal collapse across deep CNNs, and yields robust, parameter-invariant interpretations. While controlling higher-order interactions in complex non-linear attention mechanisms (ViTs) remains an open challenge, our geometric alignment framework establishes a vital theoretical foundation for designing robust interpretability methods in future architectures.

AI USE STATEMENT

In this work, we used generative AI tools (Large Language Models) primarily for language polishing, text compression to meet formatting constraints, and LaTeX formatting assistance. We have not used generative AI tools for generating the core mathematical proofs, formulating the theoretical hypotheses, or writing the experimental source code. We have thoroughly reviewed all AI-assisted text word-by-word to ensure it accurately reflects our original scientific intent and mathematical rigor. We take full responsibility for the final content of this work, including all text, claims, and formatting produced with the aid of generative AI.

ETHICS STATEMENT

This research focuses on the theoretical foundations of Explainable AI (XAI) and proposes a mathematical method to improve the reliability of Layer-wise Relevance Propagation (LRP). We utilize the standard, publicly available ImageNet dataset solely for empirical validation and do not involve human subjects, sensitive personal data, or crowdsourcing. By enhancing the fidelity and structural consistency of neural network explanations, this work fundamentally contributes to making deep learning systems more transparent, trustworthy, and accountable. We do not foresee any direct negative societal impacts, potential for malicious misuse, or ethical concerns arising from the methodologies proposed in this paper.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our theoretical and empirical claims, we have implemented several rigorous measures. First, the complete mathematical derivations for the Deep Taylor Decomposition (DTD) expansion and the detailed proofs for the Markov chain variance analysis are provided in Appendix A and B. Second, for the empirical validation, we deliberately employed a deterministic evaluation protocol over the entire ImageNet validation dataset to strictly preclude any statistical sampling bias or cherry-picking. Comprehensive experimental settings, including baseline configurations and evaluation metrics, are documented in Appendix C. Finally, the complete source code, environment specifications, and scripts required to faithfully reproduce all quantitative results and visual explanation maps have been provided as an anonymized downloadable zip file in the supplementary materials.

REFERENCES

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 135–168, 2024.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Jindong Gu, Yinchong Yang, and Volker Tresp. Understanding individual decisions of cnns via contrastive backpropagation. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pp. 119–134, 2019.

Shir Gur, Ameen Ali, and Lior Wolf. Visualization of supervised and self-supervised neural networks via attribution guided factorization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11545–11554, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Brian Kenji Iwana, Ryohei Kuroki, and Seiichi Uchida. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In *CVF International Conference on Computer Vision Workshop*, pp. 4176–4185, 2019.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

Minhyeok Lee. Gelu activation function in deep learning: a comprehensive mathematical analysis and performance. *arXiv preprint arXiv:2305.12073*, 2023.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 193–209, 2019.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, pp. 151, 2018.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations*, pp. 1–14, 2015.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328, 2017.

William R Swartout and Johanna D Moore. Explanation in second generation expert systems. In *Second Generation Expert Systems*, pp. 543–585. 1993.

A Detailed Experimental Settings

A.1 Dataset, Hardware, and Computational Efficiency

All experiments were conducted deterministically using the complete ImageNet (ILSVRC 2012) validation dataset, which consists of 50,000 high-resolution images across 1,000 classes. By evaluating the entirety of the validation set rather than relying on stochastic sampling or random subsets, we completely preclude any statistical sampling bias, ensuring that our evaluation reflects the fundamental structural properties of the models. All evaluations were executed using the PyTorch framework on a standard GPU workstation, guaranteeing full computational reproducibility.

Furthermore, the proposed π-rule incurs negligible computational overhead. The dynamic calculation of the activator and inhibitor set cardinalities (P and M) and the harmonic normalization factor (π) during the backward pass introduces only an approximate 0.2% to 1.0% increase in wall-clock execution time compared to the standard fixed-parameter $\alpha\beta$-rule. This confirms that our geometric restoration mechanism is not only theoretically rigorous but also highly efficient and scalable in practical deployments.

A.2 Evaluation Metrics and Masking Strategies

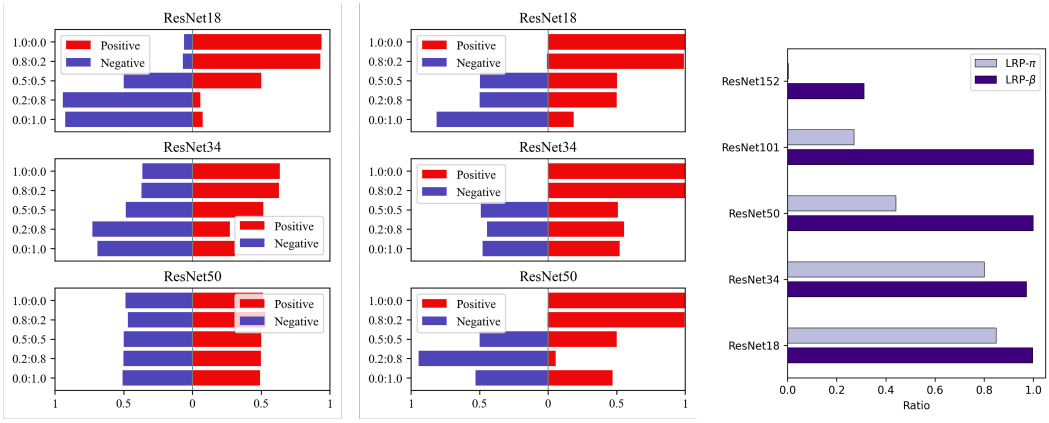
To quantitatively measure the explanation fidelity without statistical variance, we employed perturbation-based evaluation frameworks. We applied different masking strategies depending on the metric to accurately capture the structural integrity of the generated relevance maps:

- **MoRF and LeRF (Zero-masking):** For the Most Relevant First (MoRF) and Least Relevant First (LeRF) tests, we employed zero-masking. Pixels were iteratively masked to zero in descending (MoRF) or ascending (LeRF) order of their relevance scores.
- **Insertion and Deletion (Mean-masking):** To robustly measure the area under the perturbation curve (AUC) for the Insertion and Deletion metrics, we utilized mean-masking, where perturbed pixels are replaced by the channel-wise mean of the dataset.

A.3 LRP Configurations and Class-Discriminative Variants

To systematically evaluate the structural interactions and explanation fidelity, we implemented several propagation strategies based on established LRP variants.

- **Uniform LRP:** To explicitly isolate the geometric dynamics and variance accumulation across layers, we applied the standard $\alpha\beta$-rule uniformly across all layers of the network. Specifically, we swept the ratio from $\alpha : \beta = 1.0 : 0.1$ to $0.1 : 1.0$ to intentionally amplify the inhibitory contributions and observe the signal collapse phenomenon in deep residual architectures.
- **Composite LRP:** For a more stable and faithful explanation in practical scenarios, we employed composite strategies that assign different rules to different parts of the network.
  - **LRP$_{com}$:** This strategy applies the γ-rule (i.e., $\alpha : \beta = 1 : 0$) to the lower layers for stable, noise-free propagation. In the upper layers, it transitions to the $\alpha\beta$-rule with a $1 : 1$ ratio to effectively capture class-contrastive and conditional information.
  - **LRP$_{(\epsilon)}$:** This variant uses the ϵ-rule in the upper layers to prevent numerical instability near zero-activations, while similarly utilizing the γ-rule ($\alpha : \beta = 1 : 0$) in the lower layers to emphasize positive contributions. Both composite baselines were evaluated under the exact same layer-wise division for a fair comparison.
- **Contrastive LRP (CLRP):** To further extend our evaluation, we utilized CLRP, which initializes the relevance to highlight class-specific features by contrasting the target class against all other classes. The initial relevance for the target class t given a prediction c is defined as $R_c^d = 1$ if $c = t$, and $-1/(N_c - 1)$ otherwise, followed by the γ-rule in lower layers.
- **Softmax Gradient LRP (SGLRP):** SGLRP employs the gradient of the softmax function for the starting propagated signal to eliminate conflicting objects. The initialization is de-



(a) $\alpha\beta$-rule (b) π-rule (c) zero-valued maps ratio

Figure 3: **Analysis of Structural Interaction and Signal Collapse.** (a) The standard $\alpha\beta$-rule exhibits a severe dominance of inhibitor signals as network depth increases. (b) The π-rule effectively neutralizes this imbalance. (c) The propagation imbalance leads to a high rate of zero-valued maps (signal collapse) in deep networks under $\alpha : \beta = 1 : 1$, which the π-rule successfully mitigates.

fined as $R_c^d = \hat{y}_c(1 - \hat{y}_t)$ if $c = t$, and $-\hat{y}_t\hat{y}_c$ otherwise, where $\hat{y}_c$ denotes the model's prediction probability for class c.

A.4 Target Architectures

To thoroughly validate the geometric restoration across various structural patterns (e.g., sequential stacks, residual connections, dense connections, and attention mechanisms), we utilized the following deep learning architectures, all equipped with pre-trained weights from PyTorch's TorchVision library:

- **VGG Series:** VGG-11, VGG-13, VGG-16, and VGG-19 (including variants with Batch Normalization, denoted as VGG(b)).
- **ResNet Series:** ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152.
- **DenseNet Series:** DenseNet-121, DenseNet-161, DenseNet-169, and DenseNet-201.
- **Vision Transformer (ViT) Series:** ViT-Base and ViT-Large.

B Extended Evaluation: Uniform LRP Analysis

In this section, we provide extended empirical evidence to support the theoretical claims made in Section 3. By applying the standard $\alpha\beta$-rule uniformly across all layers, we explicitly isolate the geometric dynamics and variance accumulation caused by the 0^{th}-order residuals, independent of the stabilizing effects typically provided by composite propagation strategies.

B.1 Prevention of Signal Collapse in Deep Architectures

While geometric distortion manifests as variance accumulation in sequential networks, it inherently causes fatal structural collisions in architectures with parallel merging paths. To observe this, we evaluated the ratio of positive (activator) and negative (inhibitor) contributions at the final output of various ResNet architectures while varying the $\alpha : \beta$ ratio under Uniform LRP.

As illustrated in Figure 3(a), deep networks utilizing the standard $\alpha\beta$-rule exhibit a disproportionate accumulation of inhibitor signals. This structural imbalance culminates in signal collapse, where the explanation maps completely vanish (zero-valued maps) across the dataset. As shown in Figure 3(c), the occurrence rate of zero-valued maps is remarkably high under the standard rule, especially in deeper architectures.

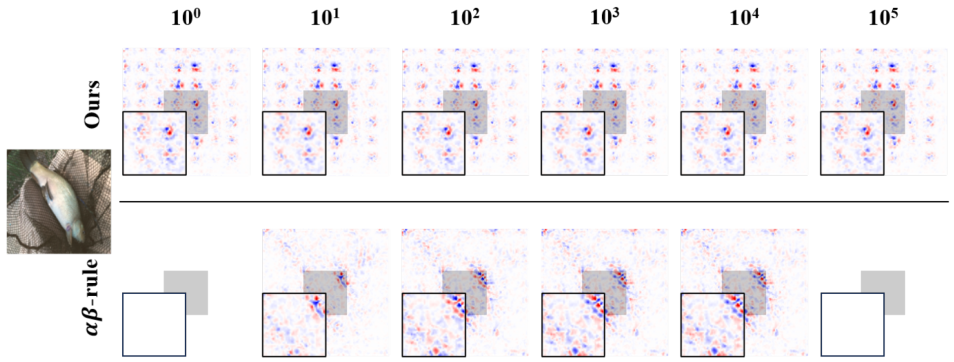


Figure 4: **Scale Invariance Test on ResNet-50.** The ratio is fixed at $\alpha : \beta = 1 : 1$ while the scale is exponentially increased. Under the standard rule ($\alpha\beta$-rule), the unneutralized bias shifts with the scale, causing severe sign flips and degraded explanations. The π-rule (Ours) exhibits scale invariance, preserving the structural integrity of the explanation map.

Conversely, as depicted in Figure 3(b), the π-rule effectively neutralizes this structural imbalance. It generates final activation maps that consistently correspond to the intended heuristic ratio, thereby substantially mitigating the occurrence of zero-valued maps despite the inherent constraints of the Uniform LRP setting.

B.2 Scale Invariance and Sign Flip Prevention

To further isolate the influence of structural noise from the heuristic ratio itself, we conducted a scale invariance test on the ResNet-50 architecture. In this experiment, the heuristic ratio was strictly fixed at $\alpha : \beta = 1 : 1$, while the absolute scale of the parameters was exponentially increased (e.g., from $10^0 : 10^0$ to $10^5 : 10^5$). Note that for a strictly fair comparison regarding scale sensitivity, the scalar normalization term $1/(\alpha + \beta)$ was intentionally omitted from the π-rule formulation.

Ideally, if the structural independence holds, multiplying both α and β by a constant scalar should not alter the shape or sign of the resulting explanation map. However, as depicted in Figure 4, merely altering the scale under the standard $\alpha\beta$-rule induces severe sign flips and eventual signal collapse (zero-valued maps). This visually proves that unneutralized structural bias couples with the parameter scale to unpredictably distort the explanation trajectory. In stark contrast, the π-rule maintains consistent explanation structures regardless of the scale magnitude, empirically demonstrating its geometric stability and robustness against parameter scaling.

C Extended Evaluation: Composite LRP Analysis

In Section 5.3 of the main text, we demonstrated that the π-rule yields superior explanation fidelity under composite configurations. In this section, we provide extended qualitative visualizations in multi-class scenarios and present the comprehensive suite of quantitative evaluation tables.

C.1 Multi-Class Target Sensitivity and Diagnostic Reliability

To intuitively understand the impact of geometric trajectory realignment on target sensitivity, we visualize explanation maps for images containing multiple classes using the ResNet-50 architecture. As illustrated in Figure 5, when conditioned on specific target classes (e.g., highlighting only the 'Dog' or 'Cat'), the standard composite rule ($\alpha\beta$-rule in upper layers) suffers from severe distortions. It frequently exhibits sign flips (blue regions appearing on the target object) or unconstrained over-activations that fail to distinguish between the adjacent classes. Conversely, the π-rule accurately isolates the features corresponding to the specified target class by preventing structural interaction noise.

This qualitative improvement carries a profound implication for the field of Explainable AI, particularly when contrasted with tasks like Weakly Supervised Object Localization (WSOL). The primary

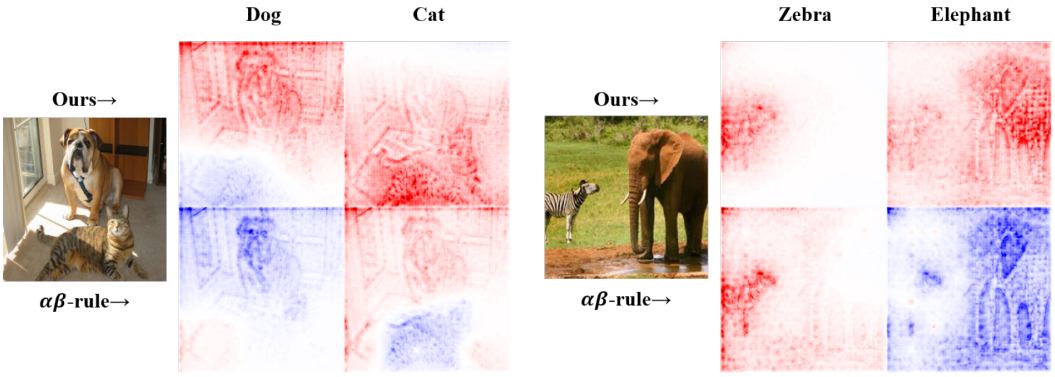


Figure 5: **Multi-Class Target Sensitivity on ResNet-50.** The standard composite rule (bottom rows) suffers from structural interaction noise, leading to fatal sign flips (e.g., blue regions on the Elephant and Dog) or over-activations that fail to separate the classes. The proposed π-rule (top rows) dynamically neutralizes these distortions, cleanly isolating the features the model utilized for the specific target class with high sensitivity.

objective of WSOL is to accurately discover bounding boxes from image-level labels; thus, the focus is solely on highlighting the target object by any means necessary. In contrast, LRP is fundamentally a *diagnostic tool* designed to faithfully reflect the inner workings of a black-box model.

Historically, when standard LRP produced sparse, sign-flipped, or insensitive relevance maps, researchers faced a critical diagnostic ambiguity: it was unclear whether the DNN itself was fundamentally flawed (e.g., poor feature entanglement) or if the explanation method was failing structurally. By geometrically restoring the structural independence of LRP modules, our π-rule resolves this ambiguity. It eliminates the explainer's inherent structural noise, ensuring that any generated relevance map is a true reflection of the model's actual learned features, thereby providing a highly reliable diagnostic methodology for deep neural networks.

C.2 Comprehensive Quantitative Results

Due to stringent page limits, the main text reported a subset of the MoRF (Most Relevant First) evaluations. Here, we present the exhaustive quantitative results across all evaluated CNN and ViT architectures. The following tables encompass a wide array of perturbation-based metrics to ensure a robust and multi-faceted assessment of explanation fidelity:

- **Table 2**: Full MoRF AUC scores (Zero-masking).
- **Table 3**: Full LeRF AUC scores (Zero-masking).
- **Table 4**: Negative Perturbation Test scores.
- **Table 5**: Positive Perturbation Test scores.
- **Table 6**: Insertion AUC scores (Mean-masking).
- **Table 7**: Deletion AUC scores (Mean-masking).

These comprehensive tables consistently corroborate our theoretical claims: the π-rule universally outperforms standard LRP variants across sequential and skip-connected CNN architectures by structurally decoupling explanation quality from unneutralized residual noise.

Table 2: **MoRF Using Zero-mask.**

	$\text{LRP}^{\alpha\beta}_{com}$	LRP^{π}_{com}	$\text{LRP}^{\alpha\beta}_{(\epsilon)}$	$\text{LRP}^{\pi}_{(\epsilon)}$	$\text{CLRP}^{\alpha\beta}$	CLRP^{π}	$\text{SGLRP}^{\alpha\beta}$	SGLRP^{π}
VGG11	15.652	11.131	8.835	8.835	10.167	10.166	17.805	17.807
VGG13	17.776	12.810	9.585	9.584	10.860	10.861	20.017	20.018
VGG16	19.652	13.520	9.106	9.106	10.435	10.435	21.213	21.211
VGG19	20.676	14.770	9.012	9.012	10.424	10.398	21.624	21.545
VGG(b)11	15.547	12.062	8.809	8.809	11.114	11.114	18.680	18.682
VGG(b)13	17.552	13.548	9.929	9.928	12.091	12.091	20.847	20.848
VGG(b)16	18.082	12.706	9.497	9.497	11.193	11.194	21.587	21.560
VGG(b)19	18.921	13.408	10.075	10.075	11.479	11.480	22.921	22.922
ResNet18	19.787	9.308	8.993	8.994	14.652	14.680	25.180	25.179
ResNet34	21.793	9.168	8.721	8.718	20.441	19.755	24.989	24.739
ResNet50	26.192	11.172	10.680	10.685	20.951	20.486	29.125	32.417
ResNet101	28.401	12.992	12.629	12.695	26.619	26.219	30.258	30.103
ResNet152	31.640	13.800	13.309	13.397	27.815	27.447	33.327	33.118
DenseNet121	24.792	23.202	18.946	18.945	27.103	27.102	29.666	29.666
DenseNet161	28.614	26.780	20.534	20.533	30.576	30.580	31.383	31.382
DenseNet169	25.486	24.491	21.425	21.424	26.841	26.835	31.885	31.885
DenseNet201	25.616	24.242	19.944	19.944	27.218	27.224	29.049	29.050
ViT base	45.221	46.807	44.472	44.470	45.134	45.200	45.337	45.384
ViT large	48.183	49.641	47.055	47.054	47.919	47.936	48.307	48.308

Table 3: **LeRF Using Zero-mask.**

	$\text{LRP}^{\alpha\beta}_{com}$	LRP^{π}_{com}	$\text{LRP}^{\alpha\beta}_{(\epsilon)}$	$\text{LRP}^{\pi}_{(\epsilon)}$	$\text{CLRP}^{\alpha\beta}$	CLRP^{π}	$\text{SGLRP}^{\alpha\beta}$	SGLRP^{π}
VGG11	33.896	40.195	37.362	37.363	36.781	36.779	29.341	29.337
VGG13	36.459	42.284	40.869	40.869	40.002	39.999	31.394	31.392
VGG16	39.280	45.089	43.115	43.115	42.313	42.314	33.144	33.148
VGG19	39.979	45.038	43.333	43.333	42.300	42.314	33.455	33.490
VGG(b)11	41.195	45.094	44.838	44.838	43.528	43.527	35.703	35.702
VGG(b)13	45.623	49.090	49.099	49.099	47.634	47.634	39.439	39.436
VGG(b)16	47.496	51.488	49.719	49.719	49.642	49.643	40.210	40.226
VGG(b)19	50.235	53.101	51.169	51.169	51.252	51.251	41.830	41.826
ResNet18	39.145	44.844	42.188	42.161	44.175	44.224	34.475	34.479
ResNet34	29.152	47.032	43.469	43.527	30.574	31.880	30.691	31.056
ResNet50	33.840	52.981	50.420	50.584	39.248	40.259	35.141	35.397
ResNet101	34.574	55.746	53.668	53.743	36.104	36.818	36.663	36.873
ResNet152	36.004	58.771	56.992	57.062	39.647	40.246	37.870	38.077
DenseNet121	37.297	38.717	41.318	41.320	34.069	34.071	31.826	31.826
DenseNet161	39.481	41.370	46.825	46.826	36.620	36.615	36.861	36.861
DenseNet169	37.756	38.639	41.095	41.095	35.656	35.662	31.375	31.376
DenseNet201	36.164	37.765	41.724	41.726	33.483	33.484	33.127	33.128
ViT base	46.157	44.545	47.166	47.160	46.314	46.268	46.029	46.036
ViT large	49.081	47.524	50.046	50.046	49.375	49.366	48.918	48.935

Table 4: **Negative Perturbation Test.**

	$\text{LRP}^{\alpha\beta}_{com}$	LRP^{π}_{com}	$\text{LRP}^{\alpha\beta}_{(\epsilon)}$	$\text{LRP}^{\pi}_{(\epsilon)}$	$\text{CLRP}^{\alpha\beta}$	CLRP^{π}	$\text{SGLRP}^{\alpha\beta}$	SGLRP^{π}
VGG11	28.653	34.236	26.597	26.594	25.684	25.674	21.006	21.018
VGG13	33.256	39.110	32.555	32.551	31.260	31.243	24.999	24.986
VGG16	37.958	43.920	36.812	36.815	36.162	36.159	28.666	28.654
VGG19	38.423	44.244	38.078	38.080	37.292	37.329	29.492	29.576
VGG(b)11	39.945	43.767	36.410	36.412	36.613	36.611	30.344	30.324
VGG(b)13	44.546	48.173	43.461	43.460	43.277	43.272	35.046	35.034
VGG(b)16	52.468	55.444	51.148	51.146	51.447	51.440	41.941	41.980
VGG(b)19	53.543	55.878	51.854	51.852	52.331	52.337	42.414	42.405
ResNet18	35.805	38.543	33.050	33.038	41.105	41.217	31.092	31.094
ResNet34	21.164	40.227	33.306	33.378	22.375	23.624	24.085	24.490
ResNet50	26.972	49.793	44.099	44.324	32.876	33.909	29.139	29.356
ResNet101	24.397	51.934	46.265	46.366	26.308	26.899	28.720	28.917
ResNet152	23.217	53.024	47.605	47.655	27.443	27.878	27.686	27.802
DenseNet121	30.939	32.333	34.512	34.519	27.276	27.291	25.321	25.309
DenseNet161	34.523	36.514	41.737	41.739	31.259	31.244	31.544	31.545
DenseNet169	32.130	32.945	34.672	34.674	29.876	29.869	25.560	25.565
DenseNet201	32.020	33.510	36.834	36.829	29.034	29.018	28.650	28.644
ViT base	29.023	26.594	30.440	30.421	29.307	29.291	29.139	29.061
ViT large	29.388	27.194	30.860	30.894	29.614	29.638	29.276	29.293

Table 5: **Positive Perturbation Test.**

	$\text{LRP}^{\alpha\beta}_{com}$	LRP^{π}_{com}	$\text{LRP}^{\alpha\beta}_{(\epsilon)}$	$\text{LRP}^{\pi}_{(\epsilon)}$	$\text{CLRP}^{\alpha\beta}$	CLRP^{π}	$\text{SGLRP}^{\alpha\beta}$	SGLRP^{π}
VGG11	11.578	7.338	2.360	2.362	3.607	3.606	11.303	11.291
VGG13	14.501	10.069	3.435	3.435	4.618	4.614	14.454	14.468
VGG16	17.873	11.926	3.615	3.617	4.981	4.993	17.005	17.010
VGG19	19.218	13.328	3.855	3.855	5.270	5.258	17.956	17.963
VGG(b)11	12.650	9.096	2.999	2.997	5.391	5.387	13.798	13.798
VGG(b)13	14.559	10.281	3.819	3.820	6.489	6.482	16.648	16.646
VGG(b)16	18.206	11.478	4.850	4.851	7.140	7.137	20.496	20.484
VGG(b)19	19.000	12.464	5.796	5.795	7.962	7.966	22.006	22.017
ResNet18	15.460	3.370	2.547	2.563	10.283	10.433	21.538	21.553
ResNet34	14.574	4.412	3.634	3.651	14.039	13.511	19.038	18.790
ResNet50	18.204	5.637	4.416	4.473	12.045	11.862	22.655	22.514
ResNet101	17.987	6.998	5.821	5.865	16.056	15.752	22.220	22.091
ResNet152	18.587	5.812	4.418	4.495	14.154	13.972	22.810	22.611
DenseNet121	18.927	17.272	12.433	12.427	21.027	21.018	23.612	23.599
DenseNet161	22.848	21.122	14.718	14.725	24.823	24.842	25.844	25.844
DenseNet169	19.746	18.802	15.718	15.696	20.896	20.885	26.072	26.066
DenseNet201	21.179	19.908	15.716	15.703	22.590	22.622	24.536	24.528
ViT base	27.962	30.095	26.601	26.602	27.670	27.676	27.995	28.084
ViT large	28.046	30.083	26.709	26.687	27.883	27.919	28.579	28.524

Table 6: **Insertion Using Mean-mask.**

	$\text{LRP}^{\alpha\beta}_{com}$	LRP^{π}_{com}	$\text{LRP}^{\alpha\beta}_{(\epsilon)}$	$\text{LRP}^{\pi}_{(\epsilon)}$	$\text{CLRP}^{\alpha\beta}$	CLRP^{π}	$\text{SGLRP}^{\alpha\beta}$	SGLRP^{π}
VGG11	43.965	48.946	46.617	46.617	45.011	45.013	37.548	37.540
VGG13	45.652	50.522	49.244	49.243	47.381	47.380	38.872	38.874
VGG16	46.599	51.672	50.218	50.218	48.768	48.767	39.432	39.431
VGG19	46.996	51.678	50.592	50.592	49.062	49.094	39.752	39.850
VGG(b)11	49.989	52.816	52.387	52.387	50.537	50.538	43.307	43.307
VGG(b)13	54.130	56.409	56.418	56.418	54.471	54.471	47.029	47.027
VGG(b)16	55.884	58.790	57.592	57.592	56.813	56.814	48.131	48.167
VGG(b)19	57.556	59.919	58.467	58.468	57.959	57.956	48.923	48.925
ResNet18	44.987	51.389	49.191	49.188	50.353	50.394	40.370	40.377
ResNet34	37.443	53.938	51.411	51.465	39.320	40.568	37.796	38.146
ResNet50	42.032	59.641	57.815	57.934	48.069	48.950	42.483	42.719
ResNet101	43.272	62.480	60.951	61.003	45.114	45.780	44.500	44.723
ResNet152	43.654	64.736	63.450	63.507	47.760	48.335	44.959	45.176
DenseNet121	44.051	45.612	48.670	48.670	40.843	40.845	38.617	38.618
DenseNet161	46.762	48.646	54.305	54.304	43.948	43.946	44.213	44.214
DenseNet169	44.894	45.872	48.628	48.630	42.886	42.891	38.411	38.413
DenseNet201	44.095	45.752	49.852	49.851	41.371	41.371	41.002	41.002
ViT base	47.539	45.632	48.616	48.607	47.585	47.529	47.312	47.300
ViT large	49.736	48.222	50.796	50.798	50.035	50.025	49.712	49.722

Table 7: **Deletion Using Mean-mask.**

	$\text{LRP}^{\alpha\beta}_{com}$	LRP^{π}_{com}	$\text{LRP}^{\alpha\beta}_{(\epsilon)}$	$\text{LRP}^{\pi}_{(\epsilon)}$	$\text{CLRP}^{\alpha\beta}$	CLRP^{π}	$\text{SGLRP}^{\alpha\beta}$	SGLRP^{π}
VGG11	20.160	14.879	12.189	12.189	13.739	13.738	22.425	22.429
VGG13	21.861	16.390	12.628	12.628	14.082	14.080	24.139	24.135
VGG16	23.217	16.786	12.126	12.126	13.544	13.545	24.985	24.983
VGG19	24.230	17.970	11.950	11.950	13.491	13.462	25.365	25.288
VGG(b)11	19.612	15.688	12.068	12.068	14.435	14.437	22.743	22.748
VGG(b)13	21.713	17.219	13.271	13.270	15.514	15.514	24.901	24.902
VGG(b)16	22.521	16.463	13.335	13.335	14.863	14.862	26.093	26.067
VGG(b)19	23.000	17.007	13.859	13.859	15.106	15.106	27.226	27.228
ResNet18	23.383	12.023	11.946	11.931	17.961	17.974	29.150	29.150
ResNet34	27.339	12.585	12.536	12.519	25.695	24.856	30.227	29.933
ResNet50	31.282	14.037	13.779	13.765	25.721	25.170	33.879	33.693
ResNet101	34.594	16.154	16.151	16.217	32.809	32.317	35.972	35.795
ResNet152	37.616	16.592	16.432	16.520	33.685	33.283	38.594	38.374
DenseNet121	29.645	27.842	23.460	23.461	32.161	32.160	34.794	34.793
DenseNet161	34.211	32.219	25.477	25.476	36.423	36.424	36.948	36.946
DenseNet169	30.445	29.371	26.156	26.155	32.030	32.024	37.248	37.247
DenseNet201	31.818	30.271	25.549	25.550	33.685	33.687	35.259	35.259
ViT base	44.979	46.639	44.140	44.158	44.940	44.972	45.179	45.226
ViT large	47.331	48.801	46.163	46.182	47.074	47.108	47.469	47.462