ATTRIBUTION UPSAMPLING SHOULD REDISTRIBUTE, NOT INTERPOLATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Attribution methods in explainable AI rely on upsampling techniques that were designed for natural images, *not* saliency maps. Standard bilinear and bicubic interpolation systematically corrupts attribution signals through aliasing, ringing, and boundary bleeding, producing spurious high-importance regions that misrepresent model reasoning. We identify that the core issue is treating attribution upsampling as an *interpolation* problem that operates in isolation from the model's reasoning, rather than a *mass redistribution* problem where model-derived semantic boundaries must govern how importance flows. We present **Universal Semantic-Aware Upsampling (USU)**, a principled method that reformulates upsampling through ratio-form mass redistribution operators, provably preserving attribution mass and relative importance ordering. Extending the axiomatic tradition of feature attribution to upsampling, we formalize four desiderata for faithful upsampling and prove that interpolation structurally violates three of them. These same three force any redistribution operator into a ratio form; the fourth selects the unique potential within this family, yielding USU. Controlled experiments on models with known attribution priors verify USU's formal guarantees; evaluation across ImageNet, CIFAR-10, and CUB-200 confirms consistent faithfulness improvements and qualitatively superior, semantically coherent explanations.

Interpolation reconstructs a continuous signal. The resulting artifacts distort what practitioners *see* and corrupt what faithfulness metrics *measure*. USU *redistributes mass* according to prediction relevance: the unique ratio-form operator (up to temperature) satisfying conservation, monotonicity, constant conditioning strength, and locality.

1 INTRODUCTION

Neural networks are increasingly deployed in safety-critical vision applications, from medical imaging to autonomous driving. Understanding their decisions has become a prerequisite for deployment. Feature attribution methods such as Grad-CAM (Selvaraju et al., 2020), Integrated Gradients (Sundararajan et al., 2017), SHAP (Lundberg & Lee, 2017), Layer-wise Relevance Propagation (Montavon et al., 2018), and RISE (Petsiuk et al., 2018) have become the standard tools for this purpose, producing importance maps that highlight which input regions drive a prediction. Many of these methods produce coarse maps that must be upsampled to input resolution before interpretation (Section 2.1). This routine upsampling step has been treated as a solved problem. We show it is not: standard upsampling techniques *systematically corrupt* the explanations they produce.

Standard interpolation methods such as bilinear and bicubic upsampling were designed to reconstruct natural images (Keys, 1981; Mitchell & Netravali, 1988). They were never intended for explainability. Attribution maps demand fundamentally different properties: *importance mass* must be conserved so the explanation accounts for the full prediction, redistribution must follow the model's own *semantic structure* rather than spatial coordinates, and upsampled attributions must remain *spatially local*, ensuring the fine-grained explanation remains faithful to the original coarse attribution. Standard interpolation violates all three.

This mismatch manifests as concrete artifacts that corrupt explanation fidelity (Figure 2). Aliasing causes semantic boundaries to be incorrectly smoothed; ringing artifacts from bicubic interpolation's

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.

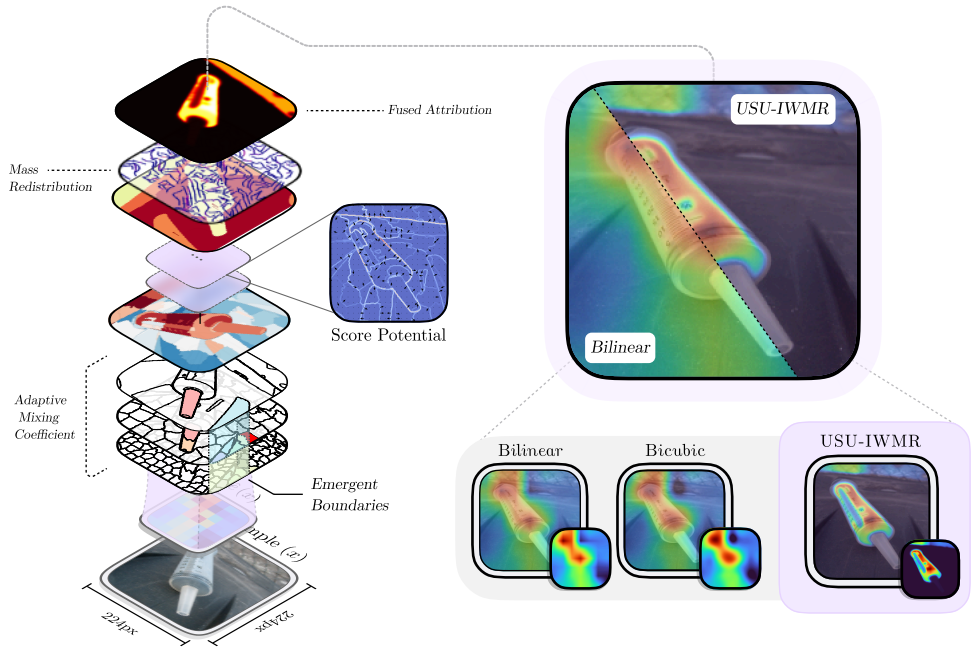

Figure 1: **USU replaces interpolation with semantically guided mass redistribution.** *Left:* The pipeline decomposes upsampling into score-potential computation, emergent boundary detection, and ratio-form mass redistribution. *Right:* Standard interpolation (bilinear, bicubic) bleeds attribution across object boundaries; USU-IWMR produces sharp, semantically coherent saliency maps that faithfully reflect model reasoning.

negative weights create spurious high-attribution regions; boundary bleeding allows attribution to leak across semantic boundaries. These artifacts *distort* what practitioners see and *corrupt* what faithfulness metrics measure. The fundamental issue is that interpolation operates in isolation from the model's reasoning, treating all spatial locations as interchangeable regardless of their importance to the prediction. The resulting explanation is faithful to neither the model's reasoning nor the coarse attribution it was meant to refine. As Figure 8 shows, three samples with entirely different shapes and ground-truth importance regions yield identical upsampled explanations: interpolation cannot distinguish what the model attends to from what it ignores.

We formalize four desiderata for *faithful* upsampling: neighbourhood completeness (D1), segment monotonicity (D2), constant conditioning strength (D3), and locality (D4). These extend the axiomatic foundations of feature attribution theory (Sundararajan et al., 2017; Sundararajan & Najmi, 2020; Bilodeau et al., 2024) to the upsampling domain.

We prove that widely adopted interpolation methods such as bilinear and bicubic upsampling violate (D1), (D2), and (D4), corrupting attribution signals and breaking faithfulness. Their kernel support crosses neighbourhood boundaries, leaking mass; their weights depend on geometry rather than semantics; and boundary pixels depend on adjacent neighbourhoods. We derive the **Universal Semantic-Aware Upsampling (USU)** operator as the unique ratio-form solution satisfying all four desiderata, with uniqueness following, as in Luce's Choice Axiom (Luce, 1959; Breitmoser, 2021), from the Independence of Irrelevant Alternatives, which (D4) requires: with mass conservation and strict monotonicity it forces any redistribution operator (one splitting each neighbourhood's mass into positive shares) into the ratio form. A temperature parameter controls redistribution sharpness. We also introduce **Soft IWMR**, a variant that intentionally relaxes (D1) to allow mass flow between neighbourhoods based on global importance, useful when coarse attributions contain systematic biases.

Whereas prior work has treated attribution upsampling as a solved implementation detail, we provide the first formal treatment, from axiomatic characterization through structural incompatibility results to a provably unique solution. Specifically:

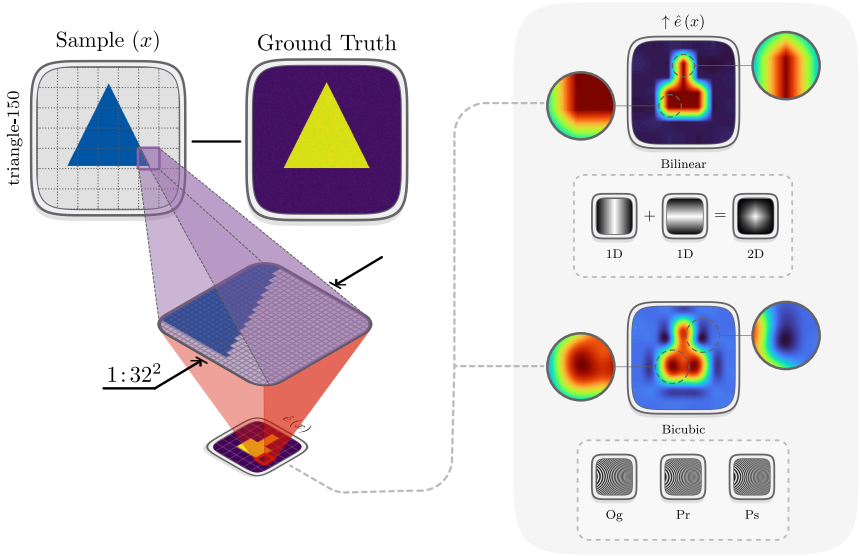

Figure 2: **Upsampling corrupts saliency maps through reconstruction errors.** Classical saliency upsampling kernels can introduce reconstruction errors, leading to incorrect visual explanations. RRL-constrained (Ross et al., 2017) model on synthetic shapes; setup and extended results in Appendix E.

1. **Axiomatic Framework.** We formalize four desiderata (D1)–(D4) for faithful attribution upsampling, extending the axiomatic tradition of feature attribution theory. We prove that bilinear and bicubic interpolation structurally violate three of these desiderata (D1), (D2), (D4), establishing that interpolation-based upsampling is fundamentally unsuitable for attribution maps.
2. **USU Operator.** We derive Universal Semantic-Aware Upsampling as the *unique* ratio-form operator satisfying all four desiderata, with uniqueness following from Luce's Independence of Irrelevant Alternatives. A temperature parameter controls redistribution sharpness, and the Soft IWMR variant relaxes strict mass conservation when coarse attributions contain systematic biases.
3. **Empirical Validation.** We first evaluate USU on constrained models where ground-truth attributions are known, outperforming interpolation in IoU, concentration and pointing game. We then confirm these gains transfer to real-world settings: evaluations across ImageNet, CIFAR-10, and CUB-200 with three architectures (VGG16, ResNet50, ViT-B-16) demonstrate consistent faithfulness improvements across diverse quantitative metrics and qualitatively superior, semantically coherent explanations.

2 Related Work[1]

2.1 Spatial Resolution in Feature Attribution

The attribution ecosystem spans diverse methodological families with markedly different native resolutions (Ibrahim & Shafiq, 2023). The most widely deployed family, CAM and its variants (Zhou et al., 2016; Selvaraju et al., 2020; Chattopadhay et al., 2018; Draelos & Carin, 2020; Wang et al., 2020), operates on final convolutional feature maps (typically 7×7 or 14×14) and upsamples via bilinear interpolation. Vision Transformer interpretability methods (Abnar & Zuidema, 2020; Chefer et al., 2021a;b) face an analogous constraint, producing patch-level attributions that must be tiled or interpolated to pixel resolution. Black-box approaches such as RISE (Petsiuk et al., 2018) use coarse

[1]Appendices B.4 and B.5 discuss further related work on saliency map refinement and on faithfulness evaluation.

random masks, while LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) typically operate on superpixel features. Even methods that nominally produce full-resolution maps encounter the gap in practice: Integrated Gradients (Sundararajan et al., 2017), DeepLIFT (Shrikumar et al., 2017), and Layer-wise Relevance Propagation (Montavon et al., 2018) are noisy and often applied to intermediate layers, while FullGrad (Srinivas & Fleuret, 2019) aggregates multi-layer maps that are individually upsampled. Beyond CAMs, upsampling is a routine step across attribution pipelines: Input×Gradient, Integrated Gradients and information-bottleneck attributions computed at hidden layers are bilinearly upsampled to the image (Rao et al., 2022; Schulz et al., 2020), perturbation methods optimize or sample low-resolution masks and upsample them (Fong & Vedaldi, 2017; Fel et al., 2021), and prototype- and concept-based explanations upsample similarity or activation maps to input resolution (Chen et al., 2019; Bau et al., 2017). Whether structurally coarse or nominally pixel-level, these methods rely on interpolation designed for image reconstruction when mapping attributions to input resolution. The faithfulness of this upsampling step itself remains unexamined: no prior work formalizes what properties attribution upsampling must satisfy or whether interpolation can guarantee them.

2.2 Axiomatic Foundations in Explainability

The axiomatic approach has been remarkably successful in uniquely determining attribution methods from natural properties. Shapley values (Štrumbelj & Kononenko, 2014; Lundberg & Lee, 2017) are uniquely characterized by efficiency, symmetry, linearity, and the null player axiom, the gold standard of axiomatic explainability. Integrated Gradients (Sundararajan et al., 2017) admit multiple independent axiomatic characterizations via completeness, sensitivity, implementation invariance, and related properties (Lundstrom et al., 2022; Lundstrom & Razaviyayn, 2025). Unified decomposition-and-allocation frameworks (Merrill et al., 2019; Lundstrom & Razaviyayn, 2023; Gevaert & Saeys, 2024; Fumagalli et al., 2025; Taimeskhanov & Garreau, 2025; Deng et al., 2024) reveal Shapley and IG as special cases of a common paradigm, and interaction extensions (Sundararajan et al., 2020; Janizek et al., 2021; Harris et al., 2022) lift these axioms to higher-order effects. Yet impossibility results (Bilodeau et al., 2024), conflicting operationalizations (Sundararajan & Najmi, 2020), and axiom-level critiques (Kumar et al., 2020; Verdinelli & Wasserman, 2024; Kwon & Zou, 2022) highlight that axioms constrain but context matters, motivating domain-specific axiomatizations.

A key theoretical tool from outside explainability is Luce's Choice Axiom (Luce, 1959; Breitmoser, 2021), specifically the Independence of Irrelevant Alternatives (IIA), which characterizes the ratio (Luce) form, a softmax over utilities $\log \varphi$, in discrete choice theory (Cerreia-Vioglio et al., 2023). Yet this powerful axiomatic program has never been applied to attribution upsampling. We extend it, showing that natural desiderata for faithful attribution refinement uniquely determine a ratio-form operator.

3 From Interpolation to Redistribution

Standard upsampling treats a coarse attribution map as a signal to be smoothly reconstructed. But **attribution maps are not images**. They carry *mass*: each neighbourhood's total attribution accounts for the model's reasoning about that region. Any upsampling that *creates or destroys* that mass *distorts* the explanation it was meant to refine. The question is not *"what value belongs at this pixel?"* but *"how should this neighbourhood's mass be allocated among its pixels?"* Formalizing this yields a complete theory: interpolation provably fails three necessary requirements (Section 3.2), four desiderata characterize faithful redistribution (Section 3.3), and a uniqueness argument forces the solution (Section 3.4). Figure 8 illustrates this failure: three distinct inputs yield identical upsampled explanations.

3.1 Problem Setup

Lattice and partitions. Let $\Omega = \{0, \dots, H-1\} \times \{0, \dots, W-1\}$ denote the **pixel lattice**. An **upsampling neighbourhood system** $\mathcal{N} = \{N_0, \dots, N_{K-1}\}$ partitions Ω into K disjoint sets, each the receptive field of one coarse attribution value; we write $\nu \colon \Omega \to \{0, \dots, K-1\}$ for the index function ($\nu(x) = k$ iff $x \in N_k$). A **segmentation partition** $\mathcal{S} = \{S_0, \dots, S_{P-1}\}$ independently

partitions Ω into P semantic segments with index function $\sigma\colon \Omega \to \{0, \ldots, P-1\}$. The partition $\mathcal{N}$ encodes the *spatial* structure of the coarse attribution; $\mathcal{S}$ provides the model's *semantic* structure that guides how attribution mass is allocated among pixels.

Attribution and mass. A **coarse attribution** $A\colon \Omega \to \mathbb{R}$ is piecewise constant on neighbourhoods, $A(x) = a_{\nu(x)}$. The **neighbourhood mass** $M_k = \sum_{x \in N_k} A(x)$ aggregates attribution within each neighbourhood. **Segment scores** $s\colon \{0, \ldots, P-1\} \to [0, 1]$ assign normalized importance to each segment. Given the tuple $(A, s, \mathcal{N}, \mathcal{S})$, we seek an upsampling operator producing a **fused attribution** $\tilde{A} = \mathcal{U}(A, s, \mathcal{N}, \mathcal{S})\colon \Omega \to \mathbb{R}$.

3.2 Why Interpolation Fails

Before formalizing what faithful redistribution requires, we establish what it cannot be: **interpolation**. Classical kernels (Keys, 1981; Mitchell & Netravali, 1988) optimize for smooth signal reconstruction, ignoring both the *attribution mass* that must be conserved and the *semantic structure* that encodes prediction relevance. **Three structural violations** make this incompatibility precise.

Theorem 1 (Interpolation Violates Mass Conservation)**.** *Any interpolation kernel whose weights from another neighbourhood $N_{k'}$ into N_k have non-zero sum violates mass conservation. Specifically, there exists a coarse attribution A such that $\sum_{x \in N_k} \tilde{A}(x) \neq M_k$.*

Mass leaks across neighbourhood boundaries whenever a kernel's net weight across them is non-zero, as for bilinear and bicubic kernels. Full proof in Appendix A.1.

Theorem 2 (Interpolation Ignores Semantics)**.** *Interpolation is score-independent: for any two segment score functions s_1, s_2, the interpolated attributions are identical. Consequently, interpolation violates segment monotonicity whenever a neighbourhood meets two segments.*

Interpolation weights are determined entirely by spatial coordinates; the operator is independent of the score function s. The model's semantic structure is invisible to the operator: within any neighbourhood that meets two segments and carries positive mass, one can always choose scores so that the higher-scored pixel receives no more attribution, violating the strict clause of (D2), even for nearest-neighbour upsampling.

Theorem 3 (Interpolation Violates Locality)**.** *When the kernel's weights from another neighbourhood into a pixel $x_0 \in N_k$ have non-zero sum, the attribution at x_0 depends on external data. Specifically, there exist coarse attributions A_1, A_2 agreeing on N_k with $\tilde{A}_1(x_0) \neq \tilde{A}_2(x_0)$.*

Adjacent neighbourhoods influence boundary attributions whenever the kernel gives them non-zero net weight, as bilinear and bicubic kernels do. Full proof in Appendix A.1.

These failures reflect a fundamental mismatch: interpolation reconstructs a continuous signal; attribution upsampling must allocate mass according to prediction relevance.

3.3 Desiderata for Faithful Upsampling

Having established what faithful redistribution *cannot* be, we now formalize what it *demands*. In the axiomatic tradition of feature attribution (Sundararajan et al., 2017; Lundberg & Lee, 2017), we identify four desiderata for faithful upsampling (Soft IWMR, Section 3.5, relaxes (D1) by design).

Desideratum (D1) Neighbourhood Completeness

For all k:

$$\sum_{x \in N_k} \tilde{A}(x) = M_k \tag{1}$$

Attribution mass is conserved within each neighbourhood. This directly extends the *completeness* axiom that characterizes both Shapley values (Lundberg & Lee, 2017) and Integrated Gradients (Sundararajan et al., 2017): just as Shapley values ensure player contributions sum to the coalition value, (D1) ensures pixel attributions sum to the neighbourhood mass.

Desideratum (D2) Segment Monotonicity

For $x, y \in N_k$ with $M_k \geq 0$:

$$s(\sigma(x)) \geq s(\sigma(y)) \implies \tilde{A}(x) \geq \tilde{A}(y), \qquad \text{strictly if } s(\sigma(x)) > s(\sigma(y)) \text{ and } M_k > 0 \quad (2)$$

Higher segment scores yield higher attributions, reflecting the *sensitivity* principle from Integrated Gradients: if the model considers a segment more important, its pixels should receive proportionally more attribution.

Desideratum (D3) Constant Conditioning Strength

For potential $\varphi\colon [0,1] \to \mathbb{R}_{>0}$ and temperature $\varepsilon > 0$:

$$\frac{\varphi(s+\Delta)}{\varphi(s)} = \exp(\Delta/\varepsilon), \quad \text{independent of } s \qquad (3)$$

The relative effect of score changes is uniform across the score range, ensuring that a fixed score difference Δ has the same redistributive effect whether it occurs among low-score or high-score segments. Alternative potentials lack this: a power-law $\varphi(s) = s^\gamma$ compresses ratios at high scores and amplifies them near zero (see Appendix A.4). (D3) demands score-invariant sensitivity, a uniformity property analogous to the translation-invariance axiom of conditional logit (Breitmoser, 2021): the multiplicative effect of a score difference Δ on weights must be the same at every base score. Without this, identical score gaps produce different redistributive effects depending on the score range (see Appendix A.4 for concrete examples).

Desideratum (D4) Locality

For $x, y \in N_k$ and fixed functions f and R (the same for all $\mathcal{N}$, $\mathcal{S}$, k and s):

$$\tilde{A}(x) = f\big(M_k,\, s(\sigma(x)),\, (s(\sigma(z)))_{z\in N_k}\big), \qquad \tilde{A}(x) = R\big(s(\sigma(x)),\, s(\sigma(y))\big)\, \tilde{A}(y). \qquad (4)$$

That is, $\tilde{A}(x)$ depends only on the mass and segment scores within its own neighbourhood, and the split between two pixels depends only on their own scores.

Changes outside the neighbourhood do not affect attributions inside it, ensuring computational tractability and preventing error propagation from distant regions. The pairwise clause keeps the split consistent: re-upsampling any part of N_k with the mass that part received returns the same values.

3.4 Uniqueness and the USU Operator

We now show that *any* redistribution operator (defined below) satisfying (D1), (D2), and (D4) must take a specific ratio form, paralleling how efficiency, symmetry, linearity, and null player uniquely determine Shapley values. Redistribution requires that scaling a neighbourhood's mass scales every pixel's share identically; otherwise the allocation rule itself depends on the total, conflating *how much* to distribute with *where*. This **linearity in mass** gives $\tilde{A}(x) = M_k \cdot w_k(x)$ for weight function w_k depending only on segment scores, paralleling the linearity axiom for Shapley values (Lundberg & Lee, 2017); a **redistribution operator** is such an operator with positive shares $w_k(x) > 0$.

Theorem 4 (Uniqueness of the Ratio Form)**.** *Assume $P \geq 3$ and $HW \geq K + 2$, so that a neighbourhood can meet three segments. Any redistribution operator satisfying (D1) conservation, (D4) locality, and (D2) monotonicity must have the ratio form:*

$$\mathcal{U}(M_k, s)(x) = M_k \cdot \frac{\varphi(s(\sigma(x)))}{\sum_{y\in N_k} \varphi(s(\sigma(y)))} \qquad (5)$$

for some strictly increasing potential $\varphi\colon [0,1] \to \mathbb{R}_{>0}$.

The full proof is in Appendix A.2. As in Luce's axiomatization of choice (Luce, 1959; Breitmoser, 2021), the substantive axiom is IIA, the pairwise clause of (D4): with conservation, strict monotonicity and positive shares it leaves no freedom but the potential φ, and Theorem 4 fails if any one of these is dropped. The characterization is *layered*: (D1), (D2), and (D4) force the ratio form with some strictly increasing potential φ, and every such φ satisfies them; (D3) then pins down the potential up to its pivot κ, which cancels, so that all four together admit exactly one operator (up to temperature). Alternative potentials sacrifice (D3)'s score-invariance but retain (D1)–(D2)–(D4); see Appendix A.4 for analysis of this tradeoff.

The USU Operator. USU instantiates the ratio form with the *tensor potential*, which satisfies (D3) by direct computation: $\varphi(s+\Delta)/\varphi(s) = \exp(\Delta/\varepsilon)$, independent of s.

USU Operator

Given temperature $\varepsilon > 0$ and pivot $\kappa \in \mathbb{R}$, the **tensor potential**

$$\varphi(s) = \exp\left(\frac{s-\kappa}{\varepsilon}\right), \qquad (\kappa \text{ cancels in (7); Appendix A.4}) \tag{6}$$

defines the **normalized weight** for pixel x in neighbourhood k:

$$w_k(x) = \frac{\varphi(s(\sigma(x)))}{\sum_{y \in N_k} \varphi(s(\sigma(y)))}, \tag{7}$$

and the **USU fused attribution**:

$$\tilde{A}(x) = M_{\nu(x)} \cdot w_{\nu(x)}(x). \tag{8}$$

USU operates as a semantic softmax: segment scores pass through the non-linear potential φ, are normalized within each neighbourhood, and redistribute mass proportionally. Though linear in mass (scaling the input scales the output), the operator is non-linear in scores through φ.

Theorem 5 ((D1)–(D4) Satisfaction)**.** *USU with the tensor potential satisfies all four desiderata (proof in Appendix A.5).*

3.5 The Attribution Basin Problem

Strict mass conservation (D1) guarantees that each neighbourhood's attribution total is preserved exactly, faithful to the coarse explanation. But consider a neighbourhood whose attribution mass lies in a *basin*: a region of moderate importance enclosed by low-attribution neighbours. (D1) traps this mass in place even when the model's segment scores indicate that semantically important pixels nearby deserve more. The result is a **pit in the attribution landscape** that no amount of within-neighbourhood redistribution can fill. This exposes a tension between two notions of faithfulness: fidelity to the coarse heatmap versus fidelity to the model's own semantic judgements. When the coarse attribution is itself imperfect, strict (D1) preserves these imperfections.

Importance-weighted redistribution. Soft IWMR (Importance-Weighted Mass Redistribution) relaxes (D1): it uses A only through the global mass $M_{\text{total}} = \sum_k M_k$, which it redistributes according to neighbourhood importance. Each neighbourhood receives importance $\lambda_k = \max_{p \in P_k} s_p$, where $P_k = \{p : N_k \cap S_p \neq \emptyset\}$ indexes segments intersecting N_k.

Soft IWMR Operator

The **importance potential** (pivot $\kappa_\Lambda \in \mathbb{R}$, temperature $\varepsilon_\Lambda > 0$) and **redistribution weights**:

$$\Lambda(\lambda) = \exp\left(\frac{\lambda - \kappa_\Lambda}{\varepsilon_\Lambda}\right), \qquad \rho_k = \frac{\Lambda(\lambda_k) \cdot |N_k|}{\sum_j \Lambda(\lambda_j) \cdot |N_j|}, \tag{9}$$

$$\tilde{M}_k = M_{\text{total}} \cdot \rho_k, \qquad \tilde{A}^{\text{IWMR}}(x) = \tilde{M}_{\nu(x)} \cdot w_{\nu(x)}(x). \tag{10}$$

The weights w_k are those of USU (7); only the mass input changes from M_k to $\tilde{M}_k$.

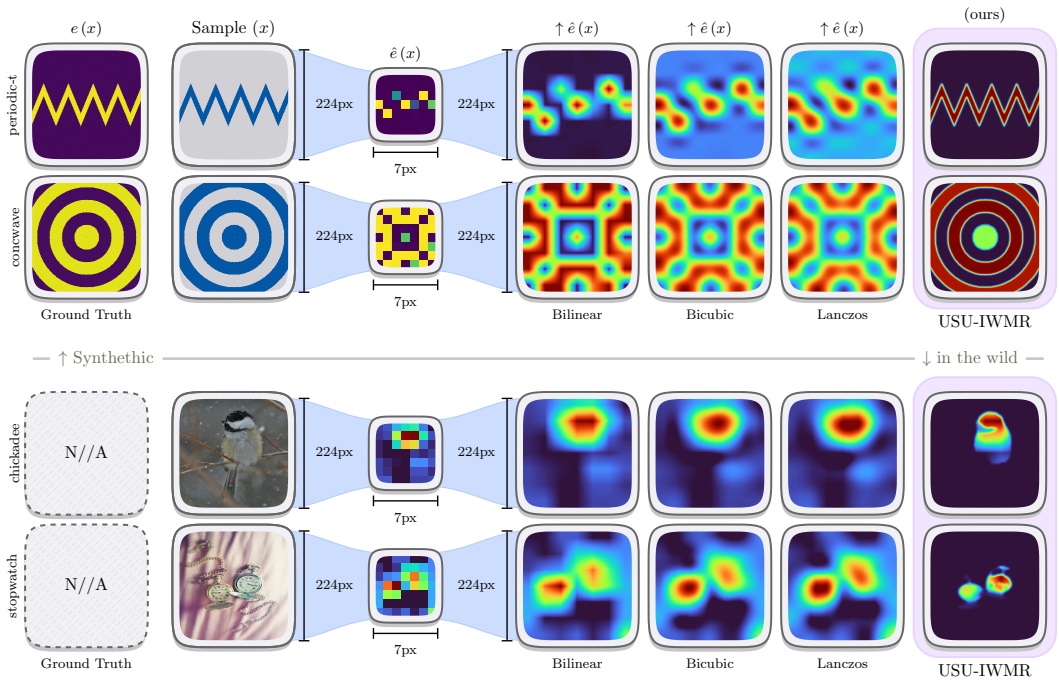

Figure 3: **Qualitative comparison.** Top row: synthetic patterns with known ground-truth attribution priors. Bottom row: ImageNet examples with GradCAM attributions. Interpolation methods (bilinear, bicubic, Lanczos) produce ringing artifacts, boundary bleeding, and aliasing; USU preserves semantic boundaries and concentrates attribution within model-relevant regions.

Soft IWMR intentionally violates (D1) but satisfies a *generalized* completeness (D1-G): $\sum_{x \in N_k} \tilde{A}^{\text{IWMR}}(x) = \tilde{M}_k$. It satisfies (D2) with $\tilde{M}_k$ in place of M_k (every $\tilde{M}_k$ has the sign of M_{total}), (D3), and a semi-local form of (D4), as well as global conservation with zero-sum mass flow ($\sum_k (\tilde{M}_k - M_k) = 0$). Full formal statement and proofs in Appendix A.3.

Temperature control. ε_Λ governs the strength of inter-neighbourhood redistribution: small values concentrate mass in high-importance neighbourhoods, while large values approach uniformity.

Hierarchical boundary refinement. The segmentation partition $\mathcal{S}$ that defines these scores has so far been assumed, not derived. For the resulting explanations to be faithful, boundaries must emerge from the model's own reasoning, refined where the model's scores transition across segments and left coarse where scores are homogeneous. We therefore let *score heterogeneity itself* select boundary resolution: starting from the coarsest partition, we refine only where importance transitions occur. Since each recursion step applies USU, the full pipeline preserves total attribution mass (Appendix C).

4 Experiments

We evaluate USU along two axes: (1) controlled synthetic tasks with known attribution priors, measuring ground-truth localization (Section 4.1); and (2) standard benchmarks with pretrained models, confirming practical faithfulness gains (Section 4.2; Figure 3).

4.1 Controlled Validation with Attribution Priors

Setup. We design synthetic shape classification tasks (circle/triangle/square) with known ground-truth attributions, training vanilla and RRL-constrained (Ross et al., 2017) CNN/MLP architectures.

Table 1: **Infidelity (↓) across datasets and models.** Best in **bold**, second-best underlined.

Dataset	Model	Bilinear	Bicubic	Lanczos	USU	USU-IWMR
ImageNet	VGG16	6.91×10^6	7.82×10^6	8.14×10^6	1.27×10^5	$\mathbf{1.13 \times 10^5}$
	ResNet50	4.87×10^7	5.41×10^7	5.63×10^7	2.45×10^5	$\mathbf{2.21 \times 10^5}$
	ViT-B-16	2.19×10^8	2.48×10^8	2.59×10^8	7.85×10^5	$\mathbf{7.02 \times 10^5}$
CIFAR-10	VGG16	3.12×10^7	3.55×10^7	3.71×10^7	7.12×10^4	$\mathbf{6.47 \times 10^4}$
	ResNet50	8.95×10^7	9.87×10^7	1.02×10^8	2.38×10^5	$\mathbf{2.21 \times 10^5}$
	ViT-B-16	2.57×10^8	2.89×10^8	3.01×10^8	2.15×10^4	$\mathbf{1.98 \times 10^4}$
CUB-200	VGG16	5.44×10^6	6.12×10^6	6.38×10^6	1.83×10^3	$\mathbf{1.70 \times 10^3}$
	ResNet50	8.46×10^7	9.32×10^7	9.68×10^7	1.52×10^7	$\mathbf{1.40 \times 10^7}$
	ViT-B-16	1.83×10^8	2.05×10^8	2.14×10^8	5.61×10^4	$\mathbf{5.08 \times 10^4}$

Coarse attributions at resolutions $\{4, 7, 14\}$ are upsampled via interpolation baselines, USU, USU-IWMR, and oracle variants. Full setup and metrics in Appendix E.

Results. Table 4 (Appendix E) reports localization of the ground-truth attributions. Oracle-IWMR achieves perfect IoU (1.00). USU outperforms all interpolation baselines: on RRL-CNN, IoU rises from 0.79 to 0.86 and pointing game from 78% to 100%, with concentration improving $0.54 \rightarrow 0.76$ as USU amplifies proper attribution priors. The USU–Oracle gap (0.86 vs. 0.96) quantifies headroom from imperfect segment scoring. Qualitatively, USU-IWMR recovers distinct attribution patterns (Figure 3, top) where interpolation collapses them (Figure 8). Table 2 validates the formal desiderata: USU achieves (D1) conservation at numerical precision (4.95×10^{-7}), four to six orders below interpolation, consistent with Theorems 1 to 3.

4.2 In-the-Wild Faithfulness Evaluation

Setup. We sample 1,000 images each from ImageNet (Russakovsky et al., 2015), CIFAR-10 (Krizhevsky, 2009), and CUB-200 (Wah et al., 2011), evaluated on VGG16 (Simonyan & Zisserman, 2015), ResNet50 (He et al., 2016), and ViT-B-16 (Dosovitskiy et al., 2021). We report Grad-CAM as representative (seven-method comparison in Appendix H). Primary metric: Infidelity ↓ from Quantus (Hedström et al., 2023); additional metrics in Appendices F and I.

Results. Table 1 reports Infidelity across all nine dataset–model combinations. USU reduces Infidelity by up to four orders of magnitude over bilinear (Wilcoxon signed-rank $p < 0.001$, Bonferroni-corrected). Gains are largest on ViT-B-16, whose 14×14 patch structure aligns naturally with segment boundaries, and smallest on CUB-200 + ResNet50, where fine-grained texture classes limit segmentation-based redistribution. USU-IWMR improves on USU in all nine settings.

5 Conclusion

Attribution methods still upsample saliency maps with interpolation techniques designed for natural images, not saliency maps. We have shown that faithful upsampling is not this solved implementation detail but a theoretically constrained operation that admits a unique ratio-form solution. Four desiderata formalize what faithful upsampling requires; standard interpolation structurally violates three, producing artifacts that distort what practitioners see and corrupt what faithfulness metrics measure. USU emerges as the unique ratio-form operator satisfying all four, with uniqueness forced by Luce's Independence of Irrelevant Alternatives. USU is a method-agnostic drop-in replacement for interpolation, delivering up to four orders-of-magnitude faithfulness gains across three datasets, three architectures, and seven attribution methods.

Limitations and future work. The uniqueness characterization assumes positive shares and that a neighbourhood can meet three segments; linearity in mass, the direct analogue of Shapley linearity, costs no generality, since (D1) and the pairwise clause of (D4) already force $\tilde{A}(x) = M_k \, w_k(x)$ on each neighbourhood.

AI USE STATEMENT

In this work, we used generative AI tools for checking grammar, formatting the paper, and suggesting structural changes to meet the page limit. We have not used generative AI tools for developing the method, formulating mathematical claims, generating the synthetic data, designing or running experiments, implementing methods, or interpreting results, and translation, dataset cleaning and qualitative data analysis are not applicable to this work. No AI tools were used to generate figures, source code, or any LaTeX styling. Additionally, we used generative AI tools for summarizing and analysing existing literature and for editing the paper to improve readability. We have reviewed all AI-assisted work. We checked every textual and writing suggestion for factual correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4190–4197, 2020. doi: 10.18653/v1/2020.acl-main.385. 3

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. OpenXAI: Towards a transparent evaluation of model explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 15784–15799, 2022. doi: 10.52202/068431-1148. 25

Leila Arras, Ahmed Osman, and Wojciech Samek. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022. doi: 10.1016/j.inffus.2021.11.008. 26

Brian Barr, Noah Fatsi, Leif Hancox-Li, Peter Richter, Daniel Proano, and Caleb Mok. The disagreement problem in faithfulness metrics. arXiv preprint arXiv:2311.07763, 2023. 25

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327, 2017. doi: 10.1109/CVPR.2017.354. 4

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024. doi: 10.1073/pnas.2304406120. 2, 4, 24

Yves Breitmoser. An axiomatic foundation of conditional logit. *Economic Theory*, 72(1):245–261, 2021. doi: 10.1007/s00199-020-01281-1. 2, 4, 6, 7, 25

Simone Cerreia-Vioglio, Fabio Maccheroni, Massimo Marinacci, and Aldo Rustichini. Multinomial logit processes and preference discovery: Inside and outside the black box. *The Review of Economic Studies*, 90(3):1155–1194, 2023. doi: 10.1093/restud/rdac046. 4, 25

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, 2018. doi: 10.1109/WACV.2018.00097. 3

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791, 2021a. doi: 10.1109/CVPR46437.2021.00084. 3

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 387–396, 2021b. doi: 10.1109/ICCV48922.2021.00045. 3

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 4

Benedict Clark, Rick Wilming, and Stefan Haufe. XAI-TRIS: non-linear image benchmarks to quantify false positive post-hoc attribution of feature importance. *Machine Learning*, 113(9): 6871–6910, 2024. doi: 10.1007/s10994-024-06574-3. 26

Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, Ziwei Yang, Zheyang Li, and Quanshi Zhang. Unifying fourteen post-hoc attribution methods with Taylor interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4625–4640, 2024. doi: 10.1109/TPAMI.2024.3358410. 4

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 (ICLR)*, 2021. 9

Rachel Lea Draelos and Lawrence Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. arXiv preprint arXiv:2011.08891, 2020. 3

Alexandre Englebert, Olivier Cornu, and Christophe De Vleeschouwer. Poly-CAM: High resolution class activation map for convolutional neural networks. *Machine Vision and Applications*, 35(4): 89, 2024. doi: 10.1007/s00138-024-01567-7. 25

Thomas Fel, Rémi Cadène, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with Sobol-based sensitivity analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 26005–26014, 2021. 4

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, 2017. doi: 10.1109/ICCV.2017.371. 4

Fabian Fumagalli, Maximilian Muschalik, Eyke Hüllermeier, Barbara Hammer, and Julia Herbinger. Unifying feature-based explanations with functional ANOVA and cooperative game theory. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 5140–5148, 2025. 4

Arne Gevaert and Yvan Saeys. Unifying attribution-based explanations using functional decomposition. arXiv preprint arXiv:2412.13623, 2024. 4

Chris Harris, Richard Pymar, and Colin Rowat. Joint Shapley values: a measure of joint feature importance. In *International Conference on Learning Representations (ICLR)*, 2022. 4

Thomas Hartley, Kirill Sidorov, Chris J. Willis, and David Marshall. Jitter-CAM: Improving the spatial resolution of CAM-based explanations. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021. doi: 10.5244/C.35.150. 25

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. 9

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. 9, 25, 34

James Hinns and David Martens. Aggregating local saliency maps for semi-global explainable image classification. arXiv preprint arXiv:2506.23247, 2025. 25

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 9734–9745, 2019. 25

Rami Ibrahim and M. Omair Shafiq. Explainable convolutional neural networks: A taxonomy, review, and future directions. *ACM Computing Surveys*, 55(10):206:1–206:37, 2023. doi: 10.1145/3563691. 3

Mohammad A. A. K. Jalwana, Naveed Akhtar, Mohammed Bennamoun, and Ajmal Mian. CAMERAS: Enhanced resolution and sanity preserving class activation mapping for image saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16322–16331, 2021. doi: 10.1109/CVPR46437.2021.01606. 25

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021. 4

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. LayerCAM: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. doi: 10.1109/TIP.2021.3089943. 25

Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. XRAI: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4947–4956, 2019. doi: 10.1109/ICCV.2019.00505. 25

Robert G. Keys. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160, 1981. doi: 10.1109/TASSP.1981.1163711. 1, 5, 17

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch. arXiv preprint arXiv:2009.07896, 2020. 25

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 9

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5491–5500, 2020. 4

Yongchan Kwon and James Zou. WeightedSHAP: analyzing and improving Shapley based feature attributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 34363–34376, 2022. doi: 10.52202/068431-2490. 4

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959. 2, 4, 7, 18, 19, 25

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4765–4774, 2017. 1, 4, 5, 6, 24

Daniel Lundstrom and Meisam Razaviyayn. A unifying framework to the analysis of interaction methods using synergy functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23005–23032, 2023. 4

Daniel Lundstrom and Meisam Razaviyayn. Four axiomatic characterizations of the Integrated Gradients attribution method. *Journal of Machine Learning Research*, 26(177):1–31, 2025. 4

Daniel Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of Integrated Gradients method and extensions to internal neuron attributions. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14485–14508, 2022. 4

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, New York, 1974. 25

John W. L. Merrill, Geoff M. Ward, Sean J. Kamkar, Jay Budzik, and Douglas C. Merrill. Generalized Integrated Gradients: A practical method for explaining diverse ensembles. arXiv preprint arXiv:1909.01869, 2019. 4

Miquel Miró-Nicolau, Antoni Jaume-i-Capó, and Gabriel Moyà-Alcover. A comprehensive study on fidelity metrics for XAI. *Information Processing & Management*, 62(1):103900, 2025. doi: 10.1016/j.ipm.2024.103900. 25

Don P. Mitchell and Arun N. Netravali. Reconstruction filters in computer graphics. In *Proceedings of the 15th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 221–228. Association for Computing Machinery, 1988. doi: 10.1145/54852.378514. 1, 5

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018. doi: 10.1016/j.dsp.2017.10.011. 1, 4

Pietro Morbidelli, Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, and Giacomo Boracchi. Augmented Grad-CAM: Heat-maps super resolution through augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4067–4071, 2020. doi: 10.1109/ICASSP40776.2020.9054416. 25

Youngrock Oh, Hyungsik Jung, Jeonghyung Park, and Min Soo Kim. EVET: Enhancing visual explanations of deep neural networks using image transformations. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3578–3586, 2021. doi: 10.1109/WACV48630.2021.00362. 25

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 1, 3, 25

Agustin Martin Picard, Thibaut Boissin, Varshini Subhash, Rémi Cadène, and Thomas Fel. Back to the baseline: Examining baseline effects on explainability metrics. arXiv preprint arXiv:2512.11433, 2025. 25

Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards better understanding attribution methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10213–10222, 2022. doi: 10.1109/CVPR52688.2022.00998. 4

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778. 4

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18770–18795, 2022. 25

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. 3, 8, 26, 34

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. 9

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. doi: 10.1109/TNNLS.2016.2599820. 25

Fedor Sandomirskiy and Omer Tamuz. Decomposable stochastic choice. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, pp. 1195, 2024. doi: 10.1145/3670865.3673590. 25

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=S1xWh1rYwB. 4

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. doi: 10.1007/s11263-019-01228-7. 1, 3

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153, 2017. 4

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 9

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 4124–4133, 2019. 4

Sédrick Stassin, Alexandre Englebert, Géraldin Nanfack, Julien Albert, Nassim Versbraegen, Gilles Peiffer, Miriam Doh, Nicolas Riche, Benoît Frénay, and Christophe De Vleeschouwer. An experimental investigation into the evaluation of explainability methods for computer vision. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2023*, volume 2133 of *Communications in Computer and Information Science*, pp. 91–106. Springer, 2025. doi: 10.1007/978-3-031-74630-7_7. 25

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014. doi: 10.1007/s10115-013-0679-x. 4

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9269–9278, 2020. 2, 4

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. 1, 2, 4, 5, 24

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor interaction index. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9259–9268, 2020. 4

Tomasz Szandała. Enhancing deep neural network saliency visualizations with gradual extrapolation. *IEEE Access*, 9:95155–95161, 2021. doi: 10.1109/ACCESS.2021.3093824. 25

Magamed Taimeskhanov and Damien Garreau. Feature attribution from first principles. arXiv preprint arXiv:2505.24729, 2025. 4

Magamed Taimeskhanov, Ronan Sicre, and Damien Garreau. CAM-based methods can see through walls. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 14947 of *Lecture Notes in Computer Science*, pp. 332–348. Springer, 2024. doi: 10.1007/978-3-031-70368-3_20. 25

Isabella Verdinelli and Larry Wasserman. Feature importance: A closer look at Shapley values and LOCO. *Statistical Science*, 39(4):623–636, 2024. doi: 10.1214/24-STS937. 4

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 9

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 111–119, 2020. doi: 10.1109/CVPRW50498.2020.00020. 3

Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance. arXiv preprint arXiv:1907.09701, 2019. 26

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016. doi: 10.1109/CVPR.2016.319. 3

SUPPLEMENTARY MATERIAL

Attribution Upsampling should Redistribute, Not Interpolate

◇

Overview. This supplement delivers: (i) complete proofs of all theorems deferred from the main paper, (ii) an extended discussion situating USU within axiomatic attribution and choice theory, (iii) algorithmic details for hierarchical boundary refinement, and (iv) comprehensive empirical results across nine configurations spanning three datasets and three models, with seven attribution methods. All proofs are self-contained and verify desiderata satisfaction (D1-D4), global conservation, and total-mass conservation through the full hierarchical pipeline (Theorem 8).

Roadmap. Section A delivers the proofs deferred from the main paper: interpolation violations, the uniqueness theorem, Soft IWMR properties, score potential analysis, and combined desiderata verification. Section B provides extended theoretical context, connecting USU to axiomatic attribution, resolution gaps in existing methods, and the choice-theoretic foundations that motivate the ratio form. Section C formalizes hierarchical boundary refinement: definitions, H-map, boundary operator, adaptive mixing, merge, and mass conservation through recursion. Section D develops the α/β-error decomposition. Section E details the synthetic experiments with full results across all methods and model families. Section F defines the six Quantus evaluation metrics. Section G describes the statistical methodology (Wilcoxon tests, effect sizes, confidence intervals). Section H demonstrates attribution-method independence across seven attribution sources (six CAM variants and FullGrad). Section I provides complete results for ImageNet, CIFAR-10, and CUB-200.

A DEFERRED PROOFS

This section delivers the proofs deferred from the main paper, organized to mirror the layered characterization: we first show interpolation fails (Section A.1), then establish the ratio form's uniqueness (Section A.2), analyze relaxations (Section A.3), identify the tensor potential (Section A.4), and verify combined satisfaction (Section A.5).

A.1 INTERPOLATION VIOLATION PROOFS

The main paper identifies three structural violations of the upsampling desiderata by standard interpolation (Theorems 1 to 3). We present the full proofs here, organized by causal depth. The root cause is *semantic blindness*: interpolation weights depend solely on spatial coordinates, making the operator structurally incapable of incorporating the model's reasoning. Two geometric consequences follow: mass leakage across neighbourhood boundaries and non-local dependencies at boundary pixels, completing the case that interpolation is fundamentally incompatible with faithful attribution upsampling.

> ***Root cause: Semantic blindness.*** Interpolation treats attribution maps as band-limited signals to be smoothly reconstructed. The interpolated value $\tilde{A}(x) = \sum_z K(x,z) \cdot A(z)$ is defined entirely by the kernel K, which encodes spatial proximity, not the model's semantic relevance. The segment scores s, the neighbourhood partition $\mathcal{N}$, and the segmentation $\mathcal{S}$ are absent from the computation. This absence is not an oversight; it is intrinsic to the interpolation paradigm.

Proof of Theorem 2 (Interpolation Ignores Semantics). Score independence. For any two segment score functions s_1 and s_2, the interpolated output is identical:

$$\tilde{A}(x) = \sum_z K(x,z) \cdot A(z),$$

since s does not appear in the computation. The output depends only on A and the kernel K.

Monotonicity violation. Take a coarse attribution with $M_k > 0$ (e.g. $A \equiv 1$) and pixels $x, y \in N_k$ in distinct segments, labelled so that $\tilde{A}(x) \geq \tilde{A}(y)$. Since interpolated values are score-independent, assigning $s(\sigma(y)) > s(\sigma(x))$ does not change the output, yet the strict clause of (D2) requires $\tilde{A}(y) > \tilde{A}(x)$, a contradiction. The argument covers every kernel, including nearest-neighbour upsampling, whose output is constant on each neighbourhood. □

Score independence is the defining limitation: no interpolation kernel, regardless of shape, order, or support size, can incorporate the model's reasoning into the upsampling.

Geometric consequence I: Mass leakage.

Semantic blindness means interpolation also ignores the neighbourhood partition $\mathcal{N}$. Since kernels are designed for signal reconstruction, not mass redistribution, their support routinely spans neighbourhood boundaries, causing attribution mass to leak between neighbourhoods.

Proof of Theorem 1 (Mass Conservation Violation). Let $k' \neq k$ be as in the hypothesis, so that the total weight $c_{kk'} = \sum_{x \in N_k} \sum_{z \in N_{k'}} K(x, z)$ from $N_{k'}$ into N_k is non-zero. Construct the coarse attribution $A = \mathbf{1}_{N_{k'}}$ ($a_{k'} = 1$, all other $a_j = 0$), so $M_k = 0$. Then

$$\sum_{x \in N_k} \tilde{A}(x) = \sum_{x \in N_k} \sum_{z \in N_{k'}} K(x, z) = c_{kk'}.$$

Since $c_{kk'} \neq 0$ while $M_k = 0$, mass conservation is violated. □

Mass leakage is not a rare corner case. Weights of both signs could cancel in $c_{kk'}$, but for a non-negative kernel such as bilinear a single positive weight $K(x_0, y)$ with $x_0 \in N_k$, $y \in N_{k'}$ already gives $c_{kk'} \geq K(x_0, y) > 0$. Bicubic kernels have negative lobes, yet the cross weights do not cancel: for upsampling by an integer factor with half-pixel alignment, the two pixels of a block at offsets $\pm v$ from its centre ($0 < v < \frac{1}{2}$, in coarse-pixel units) give the adjacent block the one-dimensional weight $u(1 - v) + u(1 + v) = v^2(3 - 2v) > 0$, where u is Keys' cubic kernel (Keys, 1981) with any parameter in $[-1, 0]$. By separability, $c_{kk'}$ for side-adjacent neighbourhoods away from the image border is the sum of these terms times the within-block weight $\sum_v u(v) > 0$ over all offsets. Bilinear and bicubic upsampling therefore leak mass at every such boundary.

Geometric consequence II: Non-local contamination.

The same boundary-spanning support that causes mass leakage also violates locality: a pixel's attribution can depend on data from adjacent neighbourhoods.

Proof of Theorem 3 (Locality Violation). Let $x_0 \in N_k$ and $k' \neq k$ be as in the hypothesis, so that $w = \sum_{z \in N_{k'}} K(x_0, z) \neq 0$ (the hypothesis of Theorem 1 provides such an x_0, since a non-zero $c_{kk'}$ has a non-zero term $\sum_{z \in N_{k'}} K(x, z)$). Construct two coarse attributions that agree on N_k:

$$A_1 \equiv 0, \qquad A_2 = \mathbf{1}_{N_{k'}}.$$

Both vanish on N_k, so they have the same mass $M_k = 0$, and a local operator must produce identical outputs for every pixel in N_k. Yet:

$$\tilde{A}_1(x_0) = 0, \qquad \tilde{A}_2(x_0) = \sum_{z \in N_{k'}} K(x_0, z) = w \neq 0.$$

The attribution at x_0 depends on data outside its neighbourhood, violating (D4). If moreover N_k lies in a single segment, with score a, and contains a pixel x_1 whose kernel support avoids $N_{k'}$, the pairwise clause is violated too: it gives $\tilde{A}_2(x_0) = R(a, a)\, \tilde{A}_2(x_1) = 0 \neq w$. □

All three violations trace to a single design choice: *interpolation* reconstructs a spatial signal without reference to attribution semantics or neighbourhood structure. Any kernel with non-zero net weight across a neighbourhood boundary, such as bilinear and bicubic, exhibits all three violations simultaneously.

A.2 Uniqueness Theorem Proof

Having established that interpolation structurally violates three desiderata, a natural question arises: does any operator satisfy them all? We now show that, for redistribution operators (linear in mass, the direct spatial analogue of Shapley linearity, with positive shares), the solution is *unique* up to the potential; existence is verified in Section A.5. The argument has two conceptual stages: the pairwise clause of locality is the Independence of Irrelevant Alternatives (Lemma 1), and strict monotonicity resolves the resulting functional equation into a potential (Lemma 2). Conservation then normalizes the potential into the ratio form; the only freedom left is the choice of potential, which (D3) removes (Section A.4).

Redistribution as rational choice. Each pixel in a neighbourhood competes for attribution mass based on its segment score. The pairwise clause of (D4) requires that the relative allocation between any two pixels depend only on their own scores, not on the other pixels present or on the neighbourhood. This is Luce's Independence of Irrelevant Alternatives (Luce, 1959), the foundational axiom of discrete choice theory, which we adopt as an axiom, as choice theory does. With positive shares and strict monotonicity, Sincov's functional equation then determines the operator up to its potential.

Lemma 1 (IIA in Share Form)**.** *Let $\mathcal{U}$ be a redistribution operator satisfying the pairwise clause of* (D4)*. For any neighbourhood N_k and any $x, y \in N_k$, the weight ratio $w_k(x)/w_k(y) = R(s(\sigma(x)), s(\sigma(y)))$ depends only on the scores $s(\sigma(x))$ and $s(\sigma(y))$, not on the scores of other pixels in N_k or on the neighbourhood itself.*

Proof. Take $M_k > 0$. By linearity in mass, $\tilde{A} = M_k \cdot w_k$ on N_k, so the pairwise clause of (D4) reads $w_k(x) = R(s(\sigma(x)), s(\sigma(y)))\, w_k(y)$. Shares are positive, so $w_k(x)/w_k(y) = R(s(\sigma(x)), s(\sigma(y)))$: changing the other pixels (the "irrelevant alternatives") or the neighbourhood cannot affect the relative allocation between the two. □

IIA is the structural constraint that connects attribution redistribution to classical choice theory. It ensures that adding or removing pixels from a neighbourhood changes the absolute share each pixel receives (through the normalization denominator) but not the relative share between any two pixels.

From independence to uniqueness.
IIA constrains the weight ratios severely: the pairwise ratio $R(a, b) = w_k(x)/w_k(y)$ of pixels with scores a, b must satisfy transitivity $R(a, b) \cdot R(b, c) = R(a, c)$, which is Sincov's functional equation. Its positive solutions are exactly the ratios $\varphi(a)/\varphi(b)$ of a potential, and strict monotonicity makes the potential strictly increasing.

Lemma 2 (Potential Factorization)**.** *Let $\mathcal{U}$ be a redistribution operator satisfying IIA (Lemma 1) and* (D2) *monotonicity, and assume a neighbourhood can meet at least three segments. There exists a strictly increasing function $\varphi\colon [0, 1] \to \mathbb{R}_{>0}$ such that $w_k(x) \propto \varphi(s(\sigma(x)))$ for all $x \in N_k$.*

Proof. Let $x, y, z \in N_k$ lie in three distinct segments and give them arbitrary scores $a, b, c \in [0, 1]$. Lemma 1 applied to the three pairs gives

$$R(a, b) \cdot R(b, c) = \frac{w_k(x)}{w_k(y)} \cdot \frac{w_k(y)}{w_k(z)} = \frac{w_k(x)}{w_k(z)} = R(a, c),$$

Sincov's functional equation, with $R > 0$ because shares are positive. Setting $b = c$ gives $R(b, b) = 1$, hence $R(0, b)\, R(b, 0) = 1$. With $\varphi(a) := R(a, 0) > 0$ we obtain $R(a, b) = R(a, 0)\, R(0, b) = \varphi(a)/\varphi(b)$, so $w_k(x) \propto \varphi(s(\sigma(x)))$ on every neighbourhood, since R is the same for all of them; no continuity is needed. If $a > b$, the strict clause of (D2) (with $M_k > 0$) gives $w_k(x) > w_k(y)$, i.e. $\varphi(a) > \varphi(b)$: φ is strictly increasing. □

With the potential in hand, conservation selects the unique normalization.

Proof of Theorem 4. By linearity in mass, $\tilde{A}(x) = M_k \cdot w_k(x)$ for some weight function w_k. By Lemma 2, $w_k(x) \propto \varphi(s(\sigma(x)))$. Taking $M_k \neq 0$ (w_k does not depend on A), (D1) requires:

$$\sum_{x \in N_k} \tilde{A}(x) = M_k \implies \sum_{x \in N_k} w_k(x) = 1.$$

Therefore $w_k(x) = \varphi(s(\sigma(x)))/\sum_{y \in N_k} \varphi(s(\sigma(y)))$, yielding the ratio form. □

The three axioms eliminate degrees of freedom in sequence: the pairwise clause of locality is IIA (the relative weight of any two pixels is context-independent), strict monotonicity resolves the functional equation into a strictly increasing potential, and conservation normalizes the result. Together with sufficiency (Section A.5), this gives uniqueness: no other redistribution operator satisfies (D1), (D2), and (D4), and (D3) then fixes the potential.

Role and independence of the hypotheses. As in Luce's theorem (Luce, 1959), the substantive axiom is IIA, the pairwise clause of (D4). It is a genuine restriction: it excludes, for instance, a softmax whose temperature adapts to the score range within N_k. Given (D1), the pairwise clause implies the first clause of (D4) and linearity in mass, since summing it over $x \in N_k$ gives $\tilde{A}(y) = M_k/\sum_{x \in N_k} R(s(\sigma(x)), s(\sigma(y)))$. The pairwise clause, the strict clause of (D2), positive shares and (D1) are each needed. Without the pairwise clause, the mean-shift shares $w_k(x) \propto s(\sigma(x)) + \bar{s}_k + 1$, with $\bar{s}_k$ the mean score in N_k, satisfy all other hypotheses, but $w_k(x)/w_k(y)$ varies with $\bar{s}_k$; the same happens if R may differ between neighbourhoods. Without the strict clause of (D2), the uniform operator $\tilde{A}(x) = M_k/|N_k|$ qualifies with $R \equiv 1$ and has only a constant potential. Without positive shares, the potential $\varphi(0) = -\eta$, $\varphi(s) = 1 + s$ for $s > 0$, with $\eta = 1/(2|\Omega|)$, yields an operator satisfying (D1), (D2) and (D4) that gives zero-score pixels negative attribution whenever $M_k > 0$ and another pixel of N_k has a positive score. Without (D1), $\tilde{A}(x) = 2M_k\varphi(s(\sigma(x)))/\sum_{y \in N_k} \varphi(s(\sigma(y)))$ qualifies.

The complete USU pipeline, from coarse masses through potential-weighted normalization to redistributed attribution, is illustrated in Figure 4.

A.3 Soft IWMR Properties

Having established uniqueness for strict conservation, we analyze what happens when (D1) is intentionally relaxed to allow importance-weighted mass redistribution across neighbourhoods. The key architectural insight is that IWMR modifies only the mass budget assigned to each neighbourhood while retaining USU's normalized weights unchanged. This separation determines which guarantees survive the relaxation: properties that depend on weights carry over intact, while properties that depend on per-neighbourhood mass budgets are either generalized or intentionally violated.

Design principle: redistribute mass, preserve weights. IWMR replaces each neighbourhood's original mass M_k with a redistributed mass $\tilde{M}_k = M_{\text{total}} \cdot \rho_k$, where $\{\rho_k\}$ forms a partition of unity weighted by neighbourhood importance. The normalized weights $w_k(x) \propto \varphi(s(\sigma(x)))$ remain identical to the strict-conservation operator. IWMR and USU therefore differ only in how much mass each neighbourhood receives, not in how that mass is distributed among pixels within it.

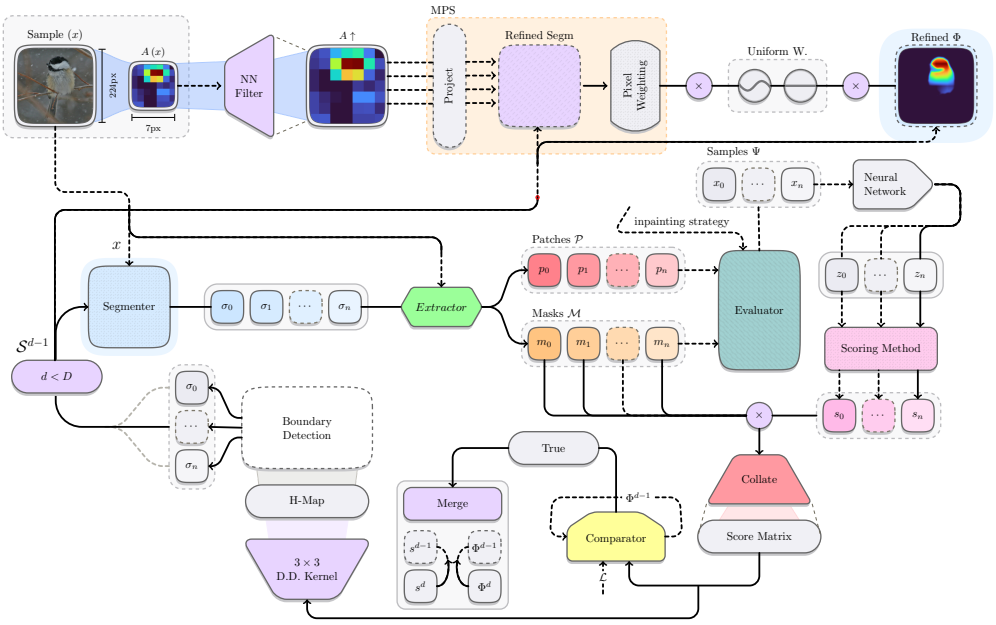

Figure 4: **Complete USU pipeline.** Given input x, the pipeline proceeds in three stages. *Scoring* (bottom right): masked perturbation evaluates segment importance, producing the score matrix Φ. *Hierarchical refinement* (bottom left): the depth fusion loop convolves Φ with the diagonal difference kernel to compute the H-map, selects heterogeneous segments for re-segmentation, and merges coarse and fine scores so that segment boundaries emerge naturally from the model's own reasoning; convergence is governed by the comparator threshold $\mathcal{L}$. *Redistribution* (top): the coarse attribution $A(x)$ is projected onto the refined segmentation and redistributed via pixel-level weighting with the tensor potential, yielding $A^{\uparrow}$.

Theorem 6 (IWMR Properties)**.** *Soft IWMR with the tensor potential satisfies:*

(D1-G) Generalized completeness. $\sum_{x \in N_k} \tilde{A}^{\mathrm{IWMR}}(x) = \tilde{M}_k$ *for all* k*;*

(D2) Segment monotonicity. *(D2) holds within each neighbourhood with* $\tilde{M}_k$ *in place of* M_k*: order-preserving when* $\tilde{M}_k \geq 0$*, strictly when* $\tilde{M}_k > 0$*;*

(D3) Constant conditioning strength. *The shared tensor potential ensures uniform score sensitivity;*

(D4′) Semi-locality. $\tilde{A}^{\mathrm{IWMR}}(x)$ *is determined by* $\tilde{M}_{\nu(x)}$ *and segment scores within* $N_{\nu(x)}$*.*

IWMR intentionally violates (D1), exactly at the neighbourhoods with $\tilde{M}_k = M_{\mathrm{total}} \cdot \rho_k \neq M_k$*. Since* M_k *depends on* A *but* ρ_k *does not, this is not decided by the importances: for* $K \geq 2$ *and any importances, uniform or not, some* A *violate (D1) and others satisfy it.*

Proof of (D1-G). $\sum_{x \in N_k} \tilde{A}^{\mathrm{IWMR}}(x) = \tilde{M}_k \sum_{x \in N_k} w_k(x) = \tilde{M}_k$, since the USU weights normalize to 1 by construction. □

Generalized completeness states that each neighbourhood exhausts its redistributed budget exactly. The shift from M_k to $\tilde{M}_k$ replaces per-neighbourhood bookkeeping with a global reallocation, but the accounting within each neighbourhood remains exact.

Proof of (D2). Within N_k, the weights $w_k(x) \propto \varphi(s(\sigma(x)))$ are strictly increasing in score since φ is strictly increasing. When $\tilde{M}_k \geq 0$, $\tilde{A}^{\mathrm{IWMR}}(x) = \tilde{M}_k \cdot w_k(x)$ preserves the ordering; if $s(\sigma(x)) > s(\sigma(y))$ and $\tilde{M}_k > 0$, the inequality is strict. □

Monotonicity survives because it is a property of the normalized weights, which IWMR inherits from USU without modification. The redistributed mass $\tilde{M}_k$ acts as a scalar that preserves the weight ordering when $\tilde{M}_k \geq 0$; since $\rho_k > 0$, this holds for all k exactly when $M_{\mathrm{total}} \geq 0$. For $M_{\mathrm{total}} < 0$ every $\tilde{M}_k$ is negative and the ordering is reversed in every neighbourhood, including those with $M_k \geq 0$.

Proof of (D3). The tensor potential satisfies $\varphi(s+\Delta)/\varphi(s) = \exp(\Delta/\varepsilon)$, independent of s (Section A.4). Since IWMR uses the same potential, conditioning strength is constant. □

From strict locality to semi-locality.

The three properties above depend only on the weights within each neighbourhood, so they carry over from USU with $\tilde{M}_k$ in place of M_k. Semi-locality is where the architectural separation becomes visible: the redistributed mass $\tilde{M}_k$ depends on all neighbourhood masses through M_{total} and on all importances through ρ_k, injecting a controlled global dependency into an otherwise local computation.

Proof of (D4′). Given $\tilde{M}_k$ and the segment scores within N_k, $\tilde{A}^{\mathrm{IWMR}}(x) = \tilde{M}_k \cdot w_k(x)$ depends on no other data. The distinction from strict (D4) is that $\tilde{M}_k = M_{\mathrm{total}} \cdot \rho_k$ itself depends on data outside N_k: on all masses through M_{total} and on all importances through ρ_k. □

Proof of (D1) violation. By (D1-G), (D1) holds at k iff $\tilde{M}_k = M_{\mathrm{total}} \cdot \rho_k = M_k$. The redistribution weights ρ_k reflect importance and area rather than original mass proportions, and they do not depend on A. Let $K \geq 2$ and fix any importances. The coarse attribution with $M_0 = 1$ and $M_k = 0$ for $k \neq 0$ gives $\tilde{M}_0 = \rho_0 < 1 = M_0$, since every $\rho_j > 0$; with equal importances, for instance, $\rho_0 = |N_0|/|\Omega|$. The coarse attribution with $M_k = \rho_k$ for all k gives $M_{\mathrm{total}} = 1$ and hence $\tilde{M}_k = M_k$ for all k. □

The violation is precisely the mechanism by which IWMR reallocates mass: it splits M_{total} in proportion to $\Lambda(\lambda_k)\,|N_k|$, regardless of how A distributes it, so neighbourhood k gains mass iff $M_{\mathrm{total}} \cdot \rho_k > M_k$. For $M_{\mathrm{total}} > 0$, of two equally large neighbourhoods the more important one receives the larger $\tilde{M}_k$, but it need not gain mass. What strict conservation loses locally, IWMR preserves globally.

Corollary 1 (Global Conservation and Zero-Sum Flow)**.** *Since $\{\rho_k\}$ forms a partition of unity, IWMR preserves total mass: $\sum_k \tilde{M}_k = M_{\mathrm{total}}$. Mass flows are zero-sum: $\sum_k (\tilde{M}_k - M_k) = 0$.*

Proof. $\sum_k \tilde{M}_k = M_{\mathrm{total}} \sum_k \rho_k = M_{\mathrm{total}}$. Subtracting original masses: $\sum_k (\tilde{M}_k - M_k) = M_{\mathrm{total}} - M_{\mathrm{total}} = 0$. □

The weight-preservation architecture of IWMR produces a clean separation of concerns: within-neighbourhood guarantees (monotonicity, conditioning strength, generalized completeness) are inherited from USU with $\tilde{M}_k$ in place of M_k, while per-neighbourhood mass budgets are globally reallocated via the partition of unity. Locality weakens to semi-locality at a single, identifiable point: the dependence of $\tilde{M}_k$ on the total mass and on global importance. Total mass is conserved; only its distribution across neighbourhoods changes.

A.4 Score Potential Analysis

The uniqueness theorem (Theorem 4) establishes that any compliant operator takes the ratio form $w_k(x) \propto \varphi(s(\sigma(x)))$ for some strictly increasing potential φ; conversely, every such ratio form satisfies (D1), (D2), and (D4) (Section A.5). The potential itself remains a free parameter: every strictly increasing, positive φ yields a valid operator. (D3), constant conditioning strength, is the final selection criterion. It requires that the multiplicative effect of a score gap Δ be uniform across the score range, so that the operator treats equal score differences identically regardless of absolute level. We now show that this requirement uniquely selects the exponential family.

> ***Selection principle: score-uniform sensitivity.*** (D3) demands $\varphi(s+\Delta)/\varphi(s) = g(\Delta)$ for some function g independent of s. In log space, this becomes $\log\varphi(s+\Delta) - \log\varphi(s) = h(\Delta)$: the increments of $\log\varphi$ must depend only on the step size, not on the starting point. This is the additive Cauchy functional equation, whose monotone solutions are precisely the affine functions $f(s) = cs + d$. Any potential with non-affine logarithm violates (D3).

Proposition 1 (Power-law potentials violate (D3))**.** *For $\varphi(s) = s^\gamma$ on $(0, 1]$ with $\gamma > 0$, the conditioning ratio depends on s.*

Proof. $\varphi(s+\Delta)/\varphi(s) = ((s+\Delta)/s)^\gamma$, which varies with s. Equivalently, $\log\varphi(s) = \gamma\log s$ is not affine. □

Proposition 2 (Log-odds potentials violate (D3))**.** *For $\varphi(s) = (s/(1-s))^\gamma$ on $(0, 1)$ with $\gamma > 0$, the conditioning ratio depends on s.*

Proof. $\varphi(s+\Delta)/\varphi(s) = \big((s+\Delta)(1-s)/\big((1-s-\Delta)\,s\big)\big)^\gamma$, which varies with s. Equivalently, $\log\varphi(s) = \gamma[\log s - \log(1-s)]$ is not affine. □

Neither function is a potential $[0, 1] \to \mathbb{R}_{>0}$: both vanish at $s = 0$, and the log-odds form is undefined at $s = 1$. The regularized forms $(s+\delta)^\gamma$ and $\big((s+\delta)/(1-s+\delta)\big)^\gamma$ with $\delta > 0$ are such potentials; their logarithms are again not affine, so they violate (D3) as well. The pattern is the same in both cases: a non-affine logarithm produces score-dependent sensitivity, amplifying redistribution in some score ranges while suppressing it in others. Since $\log\varphi$ must be affine and φ must be strictly increasing and positive, the unique surviving family is $\varphi(s) = Ce^{s/\varepsilon}$ for $C > 0$ and $\varepsilon > 0$. Writing $C = e^{-\kappa/\varepsilon}$ with a pivot $\kappa \in \mathbb{R}$ gives the parametrization $\varphi(s) = \exp((s-\kappa)/\varepsilon)$ of (6); C, and with it κ, cancels in the ratio form.

Theorem 7 (Tensor potential uniquely satisfies (D3))**.** *Among strictly increasing potentials, $\varphi(s) = \exp((s-\kappa)/\varepsilon)$, $\kappa \in \mathbb{R}$, is the unique family satisfying constant conditioning strength:*

$$\frac{\varphi(s+\Delta)}{\varphi(s)} = \exp\left(\frac{\Delta}{\varepsilon}\right),$$

independent of s. The temperature $\varepsilon > 0$ controls redistribution sharpness.

Proof. (D3) requires $\varphi(s+\Delta)/\varphi(s) = g(\Delta)$. Setting $f = \log\varphi$, this becomes $f(s+\Delta) - f(s) = h(\Delta)$, the additive Cauchy equation. Since f is monotone, $f(s) = cs + d$ for constants c, d (monotone solutions of Cauchy's equation are affine; no continuity is needed), yielding $\varphi(s) = e^{cs+d} = Ce^{s/\varepsilon}$ with $\varepsilon = 1/c > 0$ (strict increase requires $c > 0$). Direct verification: $\varphi(s+\Delta)/\varphi(s) = e^{c\Delta} = \exp(\Delta/\varepsilon)$. □

The pivot. κ only fixes the neutral point $\varphi(\kappa) = 1$, with $\varphi > 1$ above the pivot and $\varphi < 1$ below it. Moving the pivot rescales φ by a positive constant, which cancels in the normalized weights (7): the USU operator depends on ε alone. The same holds for the importance potential Λ of Soft IWMR, whose pivot κ_Λ cancels in ρ_k (9) independently of κ. Our implementation uses $\kappa = \kappa_\Lambda = \frac{1}{2}$, the center of the score range, so that the exponent $(s-\kappa)/\varepsilon$ stays within $[-\frac{1}{2\varepsilon}, \frac{1}{2\varepsilon}]$ for all $s \in [0, 1]$; this

Table 2: **Desiderata on ImageNet + VGG16.** ✓ = satisfied; ✗ = violated; (✓) = semi-local form only; – = not applicable (no score potential). D1 Error is mean absolute neighbourhood mass discrepancy.

Method	D1	D2	D3	D4	D1 Error
Bilinear	✗	✗	–	✗	1.23×10^{-1}
Bicubic	✗	✗	–	✗	8.65×10^{-1}
Lanczos-3	✗	✗	–	✗	2.14×10^{-2}
USU	✓	✓	✓	✓	4.95×10^{-7}
USU-IWMR	✗	✓	✓	(✓)	3.81×10^{-4}

minimizes the largest exponent magnitude, which matters only for floating-point range at small ε. In exact arithmetic the attributions do not depend on the pivots.

Operational consequence. Violating (D3) introduces *score-dependent redistribution sensitivity*: for $\varphi(s) = s^2$, a gap of $\Delta = 0.1$ amplifies the weight ratio by ≈ 1.27 near $s = 0.8$ but by 4.0 near $s = 0.1$, over $3\times$ the sensitivity. The tensor potential eliminates this non-uniformity. Practitioners may nonetheless prefer alternative potentials when score distributions cluster in a known range; the ratio form retains (D1), (D2), and (D4) guarantees for any strictly increasing φ.

A.5 USU Desiderata Verification

The interpolation violation proofs (Section A.1) established that semantic blindness, mass leakage, and non-local contamination are structural consequences of the interpolation paradigm. We now verify that USU's ratio form resolves all three violations and, with the tensor potential, satisfies all four desiderata simultaneously.

In brief: (D1) weight normalization ensures $\sum_{x\in N_k} \tilde{A}(x) = M_k$; (D2) the exponential is strictly increasing, so higher scores yield higher weights, hence higher attributions when $M_k > 0$; (D3) the tensor potential satisfies constant conditioning by construction; (D4) $\tilde{A}(x)$ depends only on $M_{\nu(x)}$ and scores within $N_{\nu(x)}$, and $\tilde{A}(x) = \big(\varphi(s(\sigma(x)))/\varphi(s(\sigma(y)))\big)\,\tilde{A}(y)$ for $x, y \in N_k$.

Design principle: Score-aware mass redistribution. USU replaces interpolation's fixed spatial kernel with normalized weights $w_k(x) = \varphi(s(\sigma(x)))/\sum_{y\in N_k}\varphi(s(\sigma(y)))$ that encode the model's reasoning and respect the neighbourhood partition by construction. Three algebraic properties of these weights (they sum to one, they inherit the potential's monotonicity, and they depend only on local data) directly resolve the three structural failures.

Proof of Theorem 5 (D1): Neighbourhood Completeness.

$$\sum_{x\in N_k}\tilde{A}(x) = \sum_{x\in N_k} M_k \cdot w_k(x) = M_k \sum_{x\in N_k}\frac{\varphi(s(\sigma(x)))}{\sum_{y\in N_k}\varphi(s(\sigma(y)))} = M_k \cdot 1 = M_k. \qquad \square$$

Corollary 2 (Global Conservation)**.** *Since $\mathcal{N}$ partitions Ω, summing (D1) over all neighbourhoods yields global mass conservation:*

$$\sum_{x\in\Omega}\tilde{A}(x) = \sum_{x\in\Omega} A(x). \tag{11}$$

Where interpolation leaks mass across neighbourhood boundaries through its spatially extended kernel, the normalized weights confine redistribution entirely within each neighbourhood. The denominator sums over N_k alone, so every unit of M_k is redistributed to pixels in N_k and nowhere else.

Proof of Theorem 5 (D2): Segment Monotonicity. For $x, y \in N_k$ with $s(\sigma(x)) \geq s(\sigma(y))$: since φ is strictly increasing (for USU, the exponential), $\varphi(s(\sigma(x))) \geq \varphi(s(\sigma(y)))$, hence $w_k(x) \geq w_k(y)$.

When $M_k \geq 0$, $\tilde{A}(x) = M_k \cdot w_k(x) \geq M_k \cdot w_k(y) = \tilde{A}(y)$. If $s(\sigma(x)) > s(\sigma(y))$, all three inequalities are strict for $M_k > 0$. □

Monotonicity is the direct resolution of semantic blindness: the potential φ translates the model's segment scores into weight ordering, ensuring that higher-scored regions receive proportionally more attribution. Where interpolation ignores scores entirely, USU makes them the sole determinant of within-neighbourhood allocation.

Proof of Theorem 5 (D4): Locality. By construction, $\tilde{A}(x) = M_{\nu(x)} \cdot w_{\nu(x)}(x)$ where $M_{\nu(x)} = \sum_{y \in N_{\nu(x)}} A(y)$ and $w_{\nu(x)}(x)$ depends only on segment scores within $N_{\nu(x)}$. No information from other neighbourhoods is required. For $x, y \in N_k$, $\tilde{A}(x) = R(s(\sigma(x)), s(\sigma(y)))\, \tilde{A}(y)$ with $R(a, b) = \varphi(a)/\varphi(b)$, the pairwise clause. □

Calibration: the final degree of freedom.

(D1), (D2), and (D4) hold for any strictly increasing potential φ. The remaining freedom, the choice of φ within the monotone family, is resolved by (D3), which demands score-uniform sensitivity. As shown in Section A.4, only the exponential family satisfies this requirement.

Proof of Theorem 5 (D3): Constant Conditioning Strength. Verified in Section A.4: $\varphi(s+\Delta)/\varphi(s) = \exp(\Delta/\varepsilon)$, independent of s. □

USU's normalized weights resolve the three structural failures of interpolation in exact correspondence: weight normalization eliminates mass leakage (D1), potential monotonicity incorporates score semantics (D2), and neighbourhood-local computation prevents non-local contamination (D4). The tensor potential then selects uniform sensitivity across the score range (D3). Together, the four properties confirm that USU is a complete remedy for the interpolation paradigm's incompatibilities with faithful attribution upsampling.

B Extended Theoretical Context

The main paper introduces USU's desiderata and derives the ratio form; here we situate this contribution within the broader landscape of axiomatic attribution theory, resolution gaps in existing methods, and the choice-theoretic functional equations that underpin the uniqueness argument.

B.1 Axiomatic Attribution Theory

USU extends the axiomatic tradition of feature attribution to the upsampling domain. In the Shapley framework, the *efficiency* axiom requires that player contributions sum to the coalition value (Lundberg & Lee, 2017); in Integrated Gradients, *completeness* requires that attributions sum to the prediction difference (Sundararajan et al., 2017). Desideratum (D1), neighbourhood completeness, is the spatial analogue: pixel attributions within each neighbourhood must sum to the neighbourhood mass. While Shapley efficiency and IG completeness operate on feature-level or path-integral attributions, (D1) operates on spatial redistribution, ensuring that the upsampling step does not introduce or destroy attribution mass.

Domain-specific axiomatization is required by the impossibility results that follow. Bilodeau et al. (2024) establish impossibility results showing that no single attribution method can simultaneously satisfy all desirable properties in full generality; different domains require tailored axiom sets. USU addresses this by restricting scope to upsampling, a well-defined subproblem where a complete, satisfiable axiomatization is achievable.

B.2 Resolution Gaps in Attribution Methods

The resolution gap between model reasoning and explanation granularity arises from distinct architectural sources. For CAM-family methods, attributions are defined on the final convolutional feature map (typically 7×7 or 14×14) and upsampled to input resolution. For Vision Transformers, patch tokenization (typically 16×16) creates attributions at patch granularity. Perturbation-based methods (RISE, SHAP) operate on coarse masks or superpixels, inheriting their resolution.

All existing refinement strategies for closing this gap, including multi-layer CAM fusion (LayerCAM, CAMERAS), augmentation-based approaches (Augmented Grad-CAM), and propagation schemes, are heuristic: none formally characterizes what faithful upsampling requires. A multi-layer fusion approach may sharpen boundaries but cannot state, let alone guarantee, the properties an upsampling operator should satisfy. Crucially, these strategies are integral components of specific XAI methods, not standalone operators. USU fills this gap on two levels: it provides the first axiomatic framework defining what faithful attribution upsampling means (conservation, monotonicity, calibration, locality), and it delivers a method-agnostic operator, composable with any coarse attribution method, that provably satisfies all four desiderata, with measured (D1) error 4.95×10^{-7} (Table 2).

B.3 Choice Theory and Functional Equations

The uniqueness proof (Theorem 4) draws on a deep connection to discrete choice theory that deserves explicit elaboration. Luce's Choice Axiom (Luce, 1959) states that the relative probability of choosing alternative a over b is independent of what other alternatives are available, precisely the IIA (Independence of Irrelevant Alternatives) property that the pairwise clause of (D4) locality imposes on the weight ratios. McFadden's conditional logit (McFadden, 1974) shows that IIA combined with regularity conditions yields the multinomial logit form $P(a \mid B) \propto \exp(\theta \cdot a)$. Breitmoser (2021) establishes that IIA, positivity, and translation invariance on attributes are necessary and sufficient for conditional logit, which is precisely USU's ratio form with the tensor potential. The additional constant-conditioning requirement (D3) then selects the exponential potential via the additive Cauchy functional equation, paralleling how translation invariance selects the exponential in choice theory.

USU's redistribution problem *is* a discrete choice problem. Each pixel in a neighbourhood "competes" for attribution mass based on its segment score, and the desiderata encode the same rationality conditions (consistency via IIA/locality, efficiency via conservation, and sensitivity via monotonicity) that characterize rational stochastic choice (Cerreia-Vioglio et al., 2023; Sandomirskiy & Tamuz, 2024).

B.4 Saliency Map Refinement

The community has recognized the resolution problem and proposed various remedies, though all remain heuristic. *Multi-layer aggregation* methods (LayerCAM (Jiang et al., 2021), CAMERAS (Jalwana et al., 2021), Poly-CAM (Englebert et al., 2024)) combine maps from multiple network depths but still upsample each layer via interpolation. *Augmentation-based* approaches (Augmented Grad-CAM (Morbidelli et al., 2020), Jitter-CAM (Hartley et al., 2021), EVET (Oh et al., 2021), Gradual Extrapolation (Szandała, 2021)) break the coarse grid via geometric or network-layer transformations but provide no formal mass guarantees. *Segment-based* methods are closest to USU: XRAI (Kapishnikov et al., 2019) aggregates Integrated Gradients over image segments and SATs (Hinns & Martens, 2025) leverage SAM for semi-global explanations, but neither formally redistributes mass. Taimeskhanov et al. (2024) demonstrate that CAMs can "see through walls," showing the problem extends deeper than resolution alone. None of these methods isolate upsampling as a principled operation; USU fills this gap as a standalone, axiomatically derived operator.

B.5 Faithfulness Evaluation

Numerous perturbation-based metrics (Petsiuk et al., 2018; Samek et al., 2017; Hooker et al., 2019; Rong et al., 2022) and standardized toolkits (Hedström et al., 2023; Kokhlikyan et al., 2020; Agarwal et al., 2022) evaluate attribution quality, yet no consensus has emerged: metrics frequently disagree (Barr et al., 2023; Stassin et al., 2025), are sensitive to design choices (Picard et al., 2025), and can fail even with known ground truth (Miró-Nicolau et al., 2025). A complementary strategy constructs

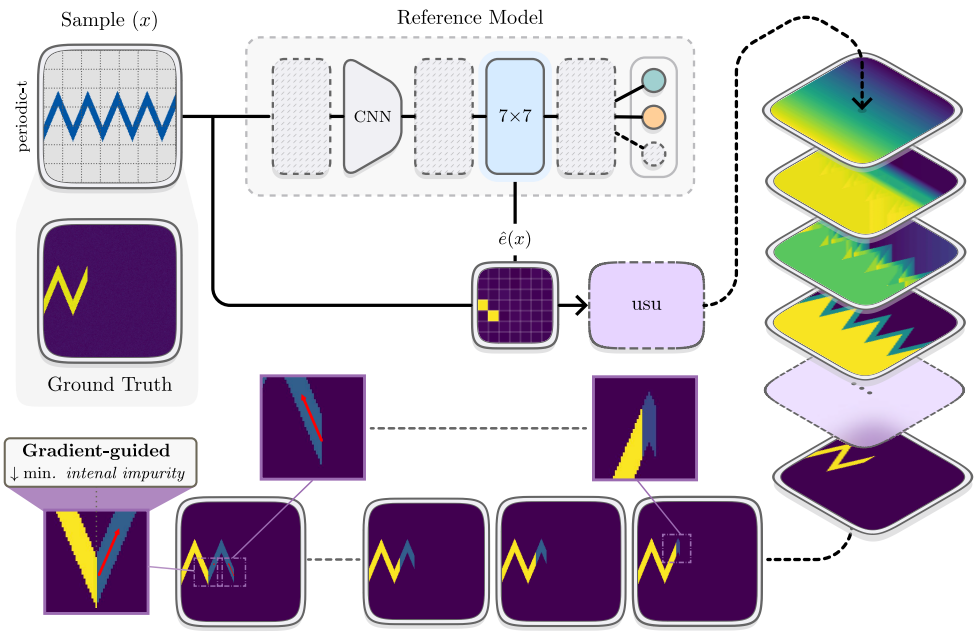

Figure 5: **Hierarchical boundary refinement.** At each depth level, the score matrix Φ^d captures per-segment importance, the H-map detects score transitions at segment boundaries, and the adaptive merge combines coarse and fine score matrices. Only boundary segments (high $|H_{\text{map}}|$) are refined further, analogous to splitting impure nodes in a decision tree (Table 3).

models with known attribution priors (Ross et al., 2017; Yang & Kim, 2019; Arras et al., 2022; Clark et al., 2024) so that ground-truth explanations are available by design. Crucially, existing metrics evaluate the full attribution pipeline end-to-end, conflating attribution-method quality with upsampling fidelity. We adopt both strategies: controlled experiments isolate the upsampling contribution, while standard benchmarks confirm that better upsampling translates to better explanations.

C Emergent Boundaries from Score Heterogeneity

Section A assumed a fixed segmentation partition. In practice, no single segmentation scale suffices: coarse partitions group semantically distinct regions, while fine partitions fragment coherent ones. This section formalizes how score heterogeneity itself selects boundary resolution, illustrated in Figure 5.

The ratio form (Section 3.4) determines how mass is allocated within each neighbourhood, but the segmentation partition $\mathcal{S}$ has so far been assumed, not derived. For faithful explanations, boundaries must emerge from the model's own reasoning, refined where scores transition and left coarse where scores are homogeneous. No single segmentation scale achieves this: a coarse partition groups semantically distinct regions; a fine partition fragments coherent ones. The key insight is to let *score heterogeneity itself* select boundary resolution. We now formalize the refinement pipeline, from geometric primitives through adaptive boundary selection to the capstone guarantee: total attribution mass is preserved through the full depth recursion.

Definition 1 (Hierarchical Segmentation)**.** *A* hierarchical segmentation *of depth D is a family $\{\mathcal{S}^0, \ldots, \mathcal{S}^{D-1}\}$ of partitions of Ω with $P_0 \leq \cdots \leq P_{D-1}$ segments, satisfying the refinement predicate: every segment at depth $d{+}1$ is contained in some segment at depth d. Refinement is reflexive and transitive.*

Table 3: **Decision-tree analogy.** Each stage of USU's hierarchical refinement mirrors a decision-tree operation, with a corresponding formal guarantee verified in this supplement.

Decision Tree	USU Refinement	Formal Guarantee
Root node	Coarsest partition $\mathcal{S}^0$ over Ω	Partitions pixel grid; refinement is reflexive and transitive (Definition 1)
Node impurity	Score heterogeneity $\max_{x\in\sigma_p} \lvert H_{\text{map}}(x)\rvert$	Linear in Φ^d; zero on constant regions (Propositions 3 and 4)
Split criterion	Boundary threshold $\lvert H_{\text{map}}\rvert > \theta$	Monotone in θ: lowering θ weakly increases boundary set (Proposition 5)
Child nodes	Finer sub-segments $\sigma_q^{d+1} \subseteq \sigma_p^d$	Refinement predicate: every child contained in parent (Definition 1)
Leaf value	Merged score $\Phi_{\text{merged}}(x)$ via adaptive $\alpha(x)$	Convex combination; bounded in $[0,1]$; identity on equal inputs (Propositions 8 to 10)
Stopping rule	$\lVert\Phi^d - \Phi^{d-1}\rVert_F \le \mathcal{L}$ or $d = D$	Symmetric and self-terminating (Propositions 6 and 7)
Prediction	Fused attribution $\tilde{A}(x)$	Total mass conserved through full recursion (Theorem 8)

Given depth-indexed scores $s^d\colon \{0,\dots,P_d-1\} \to [0,1]$, the *score matrix* at depth d maps each pixel to the score of its containing segment:

$$\Phi^d(x) = s^d(\sigma^d(x)). \tag{12}$$

Detecting importance transitions. Convolving Φ^d with a 3×3 diagonal difference kernel K_{DD} yields the *H-map*:

$$H_{\text{map}}(x) = (\Phi^d * K_{\text{DD}})(x), \tag{13}$$

measuring local score heterogeneity. When Φ^d is constant in the 3×3 neighbourhood of an interior pixel, $H_{\text{map}}(x) = 0$: no finer boundary is needed. Segments whose maximum $\lvert H_{\text{map}}\rvert$ exceeds threshold $\theta > 0$ are *boundary segments* and are refined at the next depth; homogeneous segments are retained. Lowering θ weakly increases the boundary set (monotonicity; Proposition 5).

Adaptive depth fusion. A sigmoid-based mixing coefficient $\alpha(x)$, monotonically decreasing in $\lvert H_{\text{map}}(x)\rvert$, blends coarse and fine depths: near boundaries ($\lvert H_{\text{map}}\rvert - \mu \gg \tau$, with μ, τ as in (14)), fine scores dominate; away from them, coarse scores are retained. The convex combination preserves $[0,1]$ boundedness. Recursion terminates when successive score matrices converge in Frobenius norm. Since each recursion step applies USU (Theorem 5), neighbourhood completeness propagates inductively and the full pipeline preserves total attribution mass (Theorem 8).

The remainder of this section formalizes each component of the hierarchical refinement pipeline. We begin with the geometric foundations for boundary detection (Section C.1), then formalize the H-map that detects score transitions (Section C.2), the boundary operator that selects segments for refinement (Section C.3), the stopping criterion (Section C.4), the adaptive mixing coefficient that blends coarse and fine depths (Section C.5), the merge operator (Section C.6), and finally the depth recursion with its mass conservation guarantee (Section C.7).

Detection principle: score heterogeneity as boundary signal. The H-map, obtained by convolving the score matrix with a diagonal difference kernel, drives both boundary detection and adaptive depth mixing. Where scores are locally constant, the H-map vanishes (Proposition 4) and no finer resolution is needed; at segment corners and diagonal boundaries between differing scores, the H-map typically registers the transition and can trigger refinement at the next depth. Along straight horizontal or vertical stretches of a boundary, however, it vanishes: at interior pixels, $H_{\text{map}} = 0$ wherever $\Phi^d(i,j) = g(i) + h(j)$ on the 3×3 window. This dichotomy propagates through the pipeline: it determines which segments split, how mixing coefficients weight coarse against fine depths, and ultimately when refinement converges.

C.1 Boundary Definitions

The geometric primitives below formalize what it means for segments to share a boundary, analogous to defining adjacency between nodes in the decision-tree analogy (Table 3).

Definition 2 (Pixel Adjacency)**.** *Pixels $x, y \in \Omega$ are* adjacent *iff $\|x-y\|_\infty = 1$, where $\|x-y\|_\infty = \max(|x_1-y_1|, |x_2-y_2|)$ (8-connectivity). Adjacency is symmetric and irreflexive.*

Definition 3 (Segment Boundary)**.** *The* boundary *of segment σ_p is $\partial\sigma_p = \{x \in \sigma_p : \exists\, y \notin \sigma_p,\ \|x-y\|_\infty = 1\}$. The* interior *is $\sigma_p \setminus \partial\sigma_p$; boundary and interior partition the segment: $\sigma_p = \partial\sigma_p \cup (\sigma_p \setminus \partial\sigma_p)$.*

Definition 4 (Inter-Segment Boundary)**.** *The* inter-segment boundary *between σ_p and σ_q is $\partial(\sigma_p, \sigma_q) = \{(x, y) : x \in \sigma_p,\ y \in \sigma_q,\ \|x-y\|_\infty = 1\}$. This relation is symmetric: $(x, y) \in \partial(\sigma_p, \sigma_q)$ iff $(y, x) \in \partial(\sigma_q, \sigma_p)$.*

These geometric primitives formalize the spatial relations that the H-map probes: at an interior pixel (i, j), $H_{\text{map}}(i, j) = \Phi^d(i{-}1, j{+}1) + \Phi^d(i{+}1, j{-}1) - \Phi^d(i{-}1, j{-}1) - \Phi^d(i{+}1, j{+}1)$ combines the segment scores of its four diagonal neighbours, a relation fully characterized by pixel adjacency and segment membership.

C.2 H-Map Properties

The H-map plays the role of the impurity measure in the decision-tree analogy: it quantifies local score heterogeneity, determining which segments require finer partitioning.

Proposition 3 (Linearity)**.** *The H-map is linear: $H_{\text{map}}(\Phi_1 + \Phi_2) = H_{\text{map}}(\Phi_1) + H_{\text{map}}(\Phi_2)$ and $H_{\text{map}}(r \cdot \Phi) = r \cdot H_{\text{map}}(\Phi)$ for $r \in \mathbb{R}$.*

Proof. Linearity of 2D convolution: $(f + g) * K = f * K + g * K$ and $(rf) * K = r(f * K)$. □

Proposition 4 (Zero on Constant Regions)**.** *For an interior pixel x (all corners of its $3{\times}3$ window lie in Ω), if Φ^d is constant on the $3{\times}3$ neighbourhood, then $H_{\text{map}}(x) = 0$.*

Proof. If $\Phi^d(y) = c$ for all y with $\|x-y\|_\infty \le 1$, then $H_{\text{map}}(x) = c \sum_{i,j} K_{\text{DD}}[i, j] = c(-1+0+1+0+0+0+1+0-1) = 0$. □

Together, linearity and the zero-on-constant property establish the H-map as a selective heterogeneity detector: at interior pixels it ignores uniform shifts in the score field and is non-zero only where scores vary within the $3{\times}3$ window, though not wherever they vary (it vanishes along straight horizontal or vertical boundaries). This selectivity is what allows the boundary operator to distinguish segments that genuinely require refinement from those that are merely high- or low-scored.

From detection to adaptive partitioning.

The H-map provides a continuous heterogeneity signal at every pixel. Two discrete decisions remain: which segments to refine (the boundary operator's threshold test) and when to stop refining (the comparator's convergence criterion). Both decisions inherit the H-map's selectivity: only genuine score transitions trigger further partitioning.

C.3 Boundary Operator

The boundary operator implements the split criterion: it partitions segments into those requiring refinement (high impurity) and those retained at the current depth (homogeneous).

Definition 5 (Boundary Operator)**.** *A segment σ_p^d is a* boundary segment *iff $\max_{x \in \sigma_p^d} |H_{\text{map}}(x)| > \theta$ for threshold $\theta > 0$. The boundary set is $\mathcal{B}(H_{\text{map}}) = \{\sigma_p^d : \max_{x \in \sigma_p^d} |H_{\text{map}}(x)| > \theta\}$. Non-boundary segments are retained at the current depth; boundary segments are refined at depth $d{+}1$.*

Proposition 5 (Monotonicity in Threshold)**.** *For thresholds $\theta_1 \le \theta_2$, $\mathcal{B}_{\theta_2}(H_{\text{map}}) \subseteq \mathcal{B}_{\theta_1}(H_{\text{map}})$.*

Proof. If $\max_{x \in \sigma_p} |H_{\text{map}}(x)| > \theta_2 \ge \theta_1$, then $\sigma_p \in \mathcal{B}_{\theta_1}$. □

Monotonicity in threshold gives practitioners a single interpretable control: lowering θ weakly increases the boundary set, producing finer partitions at the cost of additional computation. As $\theta \to 0^+$, exactly the segments on which H_{map} does not vanish identically are refined, which can exclude a segment that borders differently scored segments only along straight horizontal or vertical edges; as $\theta \to \infty$, no segment is refined and the coarsest partition is retained.

C.4 Comparator Properties

The stopping rule determines when further refinement yields diminishing returns, analogous to pruning in decision trees.

Stopping criterion. Recursion terminates when score matrices stabilize. The *global comparator* $C(\Phi^d, \Phi^{d-1}) = \mathbf{1}[\|\Phi^d - \Phi^{d-1}\|_F > \mathcal{L}]$ returns False when refinement has converged ($\mathcal{L} > 0$).

Proposition 6. *The global comparator is symmetric:* $C(\Phi^d, \Phi^{d-1}) = C(\Phi^{d-1}, \Phi^d)$.

Proof. $\|\Phi^d - \Phi^{d-1}\|_F = \|\Phi^{d-1} - \Phi^d\|_F$ by $\|-v\| = \|v\|$. □

Proposition 7. $C(\Phi, \Phi) = \text{False}$ *for any score matrix* Φ.

Proof. $\|\Phi - \Phi\|_F = 0 \not> \mathcal{L}$ since $\mathcal{L} > 0$. □

Symmetry ensures the stopping criterion is direction-independent, while self-termination guarantees that recursion halts once the score matrix stabilizes, providing a natural convergence condition without requiring an external depth budget.

C.5 Adaptive Mixing Coefficient

Rather than a hard split between coarse and fine depths, the adaptive mixing coefficient provides a smooth transition controlled by local boundary strength.

The *adaptive mixing coefficient*

$$\alpha(x) = \sigma\left(\frac{-(|H_{\text{map}}(x)| - \mu)}{\tau}\right), \tag{14}$$

with centering μ and temperature $\tau > 0$, satisfies:

1. $\alpha(x) \in (0, 1)$ for all x (sigmoid range on $\mathbb{R}$).
2. α is strictly decreasing in $|H_{\text{map}}(x)|$: higher boundary strength yields lower α (favouring finer depth).
3. When $|H_{\text{map}}(x)| = \mu$: $\alpha(x) = 1/2$ (equal weighting).
4. When $|H_{\text{map}}(x)| > \mu + 5\tau$: $\alpha(x) < 0.01$ (fine depth dominates).
5. When $|H_{\text{map}}(x)| < \mu - 5\tau$: $\alpha(x) > 0.99$ (coarse depth dominates).

C.6 Merge Operator

The merge operator combines the leaf values from adjacent depth levels, producing the fused score matrix that drives the final attribution.

The merged score matrix blends coarse and fine depth levels:

$$\Phi_{\text{merged}}(x) = \alpha(x) \cdot \Phi^{d-1}(x) + (1 - \alpha(x)) \cdot \Phi^d(x). \tag{15}$$

Near boundaries ($|H_{\text{map}}| - \mu \gg \tau$), $\alpha \approx 0$ and finer scores dominate; away from boundaries ($\mu - |H_{\text{map}}| \gg \tau$), $\alpha \approx 1$ and coarser scores are retained.

Proposition 8 (Convex Combination)**.** *Since* $\alpha(x) \in (0, 1)$ *and* $1 - \alpha(x) \in (0, 1)$ *with* $\alpha(x) + (1 - \alpha(x)) = 1$*, the merge* $\Phi_{\text{merged}}(x) = \alpha(x)\Phi^{d-1}(x) + (1 - \alpha(x))\Phi^d(x)$ *is a convex combination.*

Proposition 9 (Boundedness)**.** *If* $\Phi^{d-1}(x), \Phi^d(x) \in [0, 1]$ *for all* x*, then* $\Phi_{\text{merged}}(x) \in [0, 1]$.

Proof. $\Phi_{\text{merged}}(x) = \alpha\Phi^{d-1} + (1-\alpha)\Phi^d \leq \alpha + (1-\alpha) = 1$ and ≥ 0. □

Proposition 10 (Identity on Equal Inputs)**.** *When* $\Phi^{d-1} = \Phi^d = \Phi$*:* $\Phi_{\text{merged}}(x) = \alpha\Phi(x) + (1-\alpha)\Phi(x) = \Phi(x)$.

Convex combination guarantees that merged scores lie between their coarse and fine inputs. Boundedness preserves the $[0, 1]$ score range through successive depth levels. Identity on equal inputs ensures that homogeneous regions, where coarse and fine partitions agree, pass through the merge unchanged. Together, these properties ensure that the merge operator adds no distortion to the score field.

C.7 Depth Recursion and Mass Conservation

Each component formalized above contributes a local guarantee: the H-map responds only to genuine transitions, the boundary operator is monotone in threshold, and the merge operator preserves boundedness. The capstone result assembles these local guarantees into a global one: total attribution mass is conserved through the full depth recursion, regardless of how many levels of refinement are applied.

Definition 6 (Depth Recursion)**.** *The* depth state *at level* d *is the triple* $(\mathcal{S}^d, s^d, A^d)$*: partition, scores, and attribution. The* recursion operator R *advances state from depth* d *to* $d{+}1$ *by computing new scores, detecting boundaries via the H-map, merging score matrices, and applying USU within the finer partition. The* full pipeline *iterates* R *from depth* 0 *to* $D{-}1$*, yielding the fused attribution* $\tilde{A}$*.*

Theorem 8 (Mass Conservation Through Recursion)**.** *The full hierarchical pipeline preserves total attribution mass:* $\sum_{x\in\Omega} \tilde{A}(x) = \sum_{x\in\Omega} A^0(x)$.

Proof. By induction on depth d.

Base case ($d = 0$)**.** The initial attribution A^0 has mass $M = \sum_{x\in\Omega} A^0(x)$.

Inductive step. Suppose $\sum_{x\in\Omega} A^d(x) = M$. The recursion operator R applies USU to A^d on the fixed neighbourhoods $\{N_k\}$, with masses $M_k^{d+1} = \sum_{x\in N_k} A^d(x)$ and weights computed from the merged scores (15) of $\mathcal{S}^d$ and the finer partition $\mathcal{S}^{d+1}$. By Theorem 5 (D1), neighbourhood completeness holds: $\sum_{x\in N_k} A^{d+1}(x) = M_k^{d+1}$ for each neighbourhood k. Since the neighbourhoods partition Ω:

$$\sum_{x\in\Omega} A^{d+1}(x) = \sum_k \sum_{x\in N_k} A^{d+1}(x) = \sum_k M_k^{d+1} = \sum_{x\in\Omega} A^d(x) = M.$$

The merge (15) cannot change any neighbourhood mass: it acts on score matrices before USU is applied, so it changes only the scores from which the weights w_k are computed, and these weights sum to one on every N_k whatever the scores. By induction, $\sum_{x\in\Omega} \tilde{A}(x) = M$. □

Hierarchical refinement adds no new free parameters to the ratio form; it selects resolution adaptively from the model's own score field. The pipeline provably preserves completeness through recursion: every USU step satisfies (D1), so total attribution mass is conserved (Theorem 8). Calibration (D3) is a property of the unchanged potential φ, and monotonicity and locality hold for each USU step with respect to the scores it consumes; for the fused output, however, (D2) and (D4) with respect to segment scores are not guaranteed, since $\alpha(x)$ can vary within a segment and the H-map reads scores across neighbourhood borders. The new controls, the threshold θ, the tolerance $\mathcal{L}$ and the mixing parameters μ, τ, govern resolution depth and the blending of depths, not the ratio form itself.

D Error Analysis

Having formalized the theoretical guarantees (Section A) and the multi-scale mechanism that makes them practical (Section C), we now analyze how these guarantees translate into concrete error reduction at the pixel level.

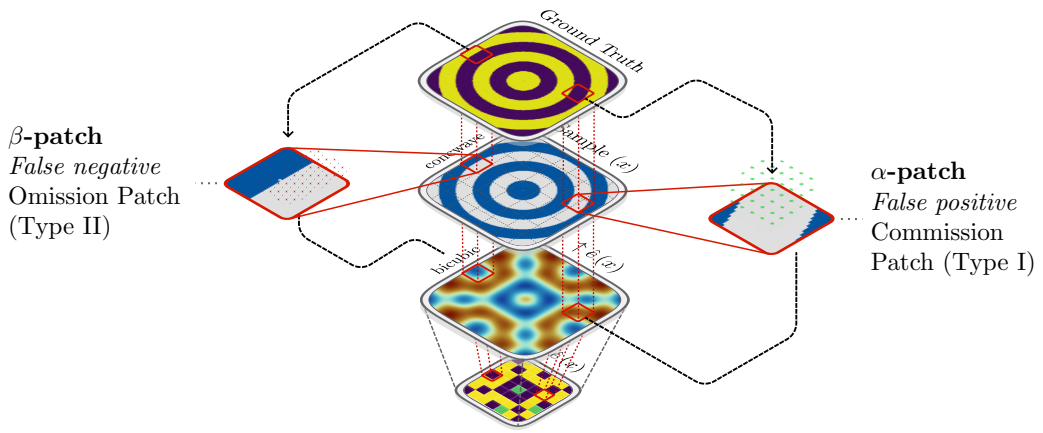

Figure 6: **Compression error types.** From top to bottom: ground-truth attribution, input image, upsampled saliency map, and raw low-resolution heatmap. Left column shows β-errors where true attributions disappear due to compression (signal loss). Right column shows α-errors where spurious attributions emerge despite absence in ground truth (false signal generation). Both error types corrupt explanation fidelity.

Given ground-truth attribution $A^*\colon \Omega \to \mathbb{R}$, the α-error (spurious attribution) and β-error (signal loss) at pixel x are defined as:

$$\alpha(x) = \max\big(0,\ |\tilde{A}(x)| - |A^*(x)|\big), \tag{16}$$

$$\beta(x) = \max\big(0,\ |A^*(x)| - |\tilde{A}(x)|\big). \tag{17}$$

Root cause: Lossy compression. Compressing pixel-level attribution to coarse neighbourhood aggregates and reconstructing it is inherently lossy. Two complementary error modes arise from this compression: α-errors (spurious signal) where reconstruction overshoots ground truth, and β-errors (signal loss) where it undershoots. These modes are mutually exclusive at each pixel, and their aggregate behaviour at the neighbourhood level is governed by mass balance.

Proposition 11 (Error Exclusivity)**.** *At each pixel x, at most one of $\alpha(x)$, $\beta(x)$ is positive. Their sum captures the total magnitude mismatch:*

$$\alpha(x) + \beta(x) = \big||\tilde{A}(x)| - |A^*(x)|\big|. \tag{18}$$

Proof. Let $\delta = |\tilde{A}(x)| - |A^*(x)|$. When $\delta \geq 0$: $\alpha(x) = \delta$ and $\beta(x) = 0$. When $\delta < 0$: $\alpha(x) = 0$ and $\beta(x) = -\delta$. In both cases $\alpha(x) + \beta(x) = |\delta|$. □

Exclusivity ensures the error at each pixel is unambiguous: reconstruction either overshoots or undershoots, never both. This pointwise decomposition provides a complete accounting of attribution error (Figure 6). The question is what structural property of an upsampling operator controls these errors at the neighbourhood level.

Mass imbalance as error lower bound.

Define the *mass deficit* $\Delta_k^- = \max\big(0,\ \sum_{x\in N_k} A^*(x) - \sum_{x\in N_k} \tilde{A}(x)\big)$ and *mass excess* $\Delta_k^+ = \max\big(0,\ \sum_{x\in N_k} \tilde{A}(x) - \sum_{x\in N_k} A^*(x)\big)$. Mass excess forces α-error; mass deficit forces β-error.

Proposition 12 (Mass–Error Bounds)**.** *For non-negative attributions within neighbourhood N_k:*

$$\sum_{x \in N_k} \alpha(x) \;\geq\; \Delta_k^+, \qquad \sum_{x \in N_k} \beta(x) \;\geq\; \Delta_k^-.$$

Proof. We show the α bound; the β bound is symmetric. With non-negative attributions, $|\tilde{A}(x)| = \tilde{A}(x)$ and $|A^*(x)| = A^*(x)$. Since $\alpha(x) = \max(0,\, \tilde{A}(x) - A^*(x)) \geq \tilde{A}(x) - A^*(x)$, summing over N_k gives

$$\textstyle\sum_{x \in N_k} \alpha(x) \;\geq\; \sum_{x \in N_k} \tilde{A}(x) - \sum_{x \in N_k} A^*(x).$$

Since each $\alpha(x) \geq 0$, the left-hand side is non-negative. Combining $\sum \alpha(x) \geq 0$ with the summation bound gives $\sum \alpha(x) \geq \max\big(0,\, \sum \tilde{A}(x) - \sum A^*(x)\big) = \Delta_k^+$. □

Mass imbalance is therefore a structural lower bound on attribution error. Any operator that violates mass conservation (D1) is committed to nonzero α- or β-error for some input.

Mass-level error under (D1). (D1) ensures zero net mass error per neighbourhood when the coarse attribution faithfully aggregates ground truth. If $M_k = \sum_{x \in N_k} A^*(x)$, then:

$$\sum_{x \in N_k} \tilde{A}(x) = M_k = \sum_{x \in N_k} A^*(x),$$

so $\Delta_k^- = \Delta_k^+ = 0$: the mass-error lower bounds vanish. This holds for every operator satisfying (D1), USU included, but it does not make the pixel-level α- and β-errors vanish, since (D1) fixes only the neighbourhood total, not where the mass sits within N_k. Conversely, any operator that does not satisfy (D1) admits an input configuration with $\Delta_k^- > 0$ or $\Delta_k^+ > 0$, and hence positive neighbourhood-level α- or β-error. Figure 8 illustrates a related failure: interpolation produces identical saliency maps for inputs with distinct ground truths, because its output depends on the coarse map alone. (D1) does not prevent this: block replication $\tilde{A} = A$ satisfies (D1) yet, like interpolation, ignores s and $\mathcal{S}$. USU can separate such inputs because its output also depends on s and $\mathcal{S}$.

The α/β-decomposition yields three structural guarantees: exclusivity partitions pixel error into a single mode, the mass-error bounds establish that, for non-negative attributions, mass imbalance is a structural lower bound on neighbourhood error, and, when $M_k = \sum_{x \in N_k} A^*(x)$, (D1) is exactly the condition that zeroes both bounds. USU inherits this from (D1), as does every mass-conserving operator.

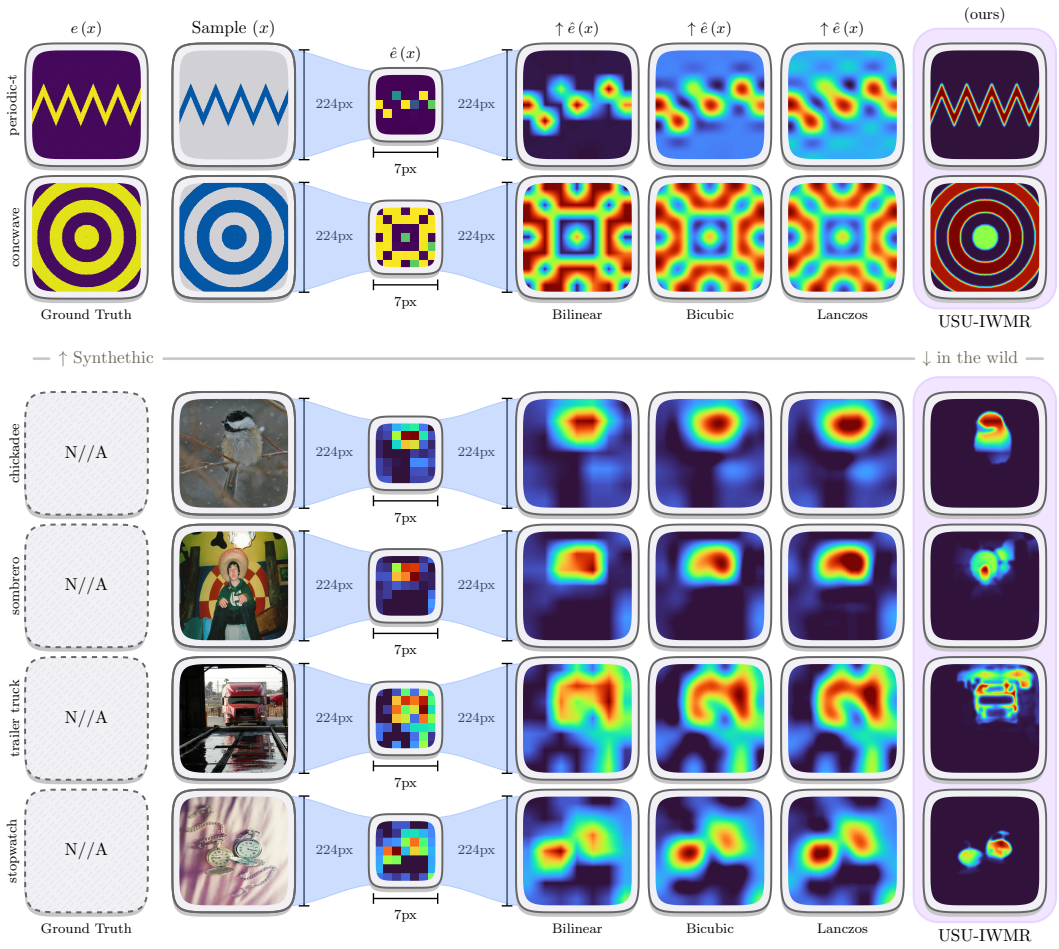

Figure 7: **Extended qualitative comparison.** Synthetic patterns with known ground truth (top) and ImageNet samples with GradCAM from VGG16 (bottom). Interpolation baselines exhibit boundary bleeding and ringing; USU-IWMR recovers ground-truth support on synthetic data and concentrates attribution within semantic regions on real images.

E Synthetic Experiment Details

The preceding sections established theoretical guarantees and error analysis. We now describe the controlled experimental setup designed to verify these guarantees empirically, with known ground-truth attribution priors that enable direct measurement of reconstruction fidelity.

Shape classification. We generate 2,000 single-channel 224×224 grayscale images containing a single centered shape (circle, triangle, or square) with class-balanced sampling. The dataset is split 70/15/15 into train/validation/test sets. Ground-truth attribution masks are the binary foreground regions.

Pattern detection. Five periodic pattern types (zigzag, sine, spiral, concentric wave, moiré) are rendered on 224×224 images. This task is used for qualitative evaluation of boundary preservation.

Model architectures. The **CNN** consists of three convolutional blocks (5×5 kernels, $16\to32\to64$ channels, each followed by ReLU and 2×2 max-pooling), adaptive average pooling to 4×4, and a two-layer MLP ($1024\to128\to3$). The **MLP** flattens the input to $50{,}176$ dimensions with hidden layers $[256, 64]$ and 3-class output.

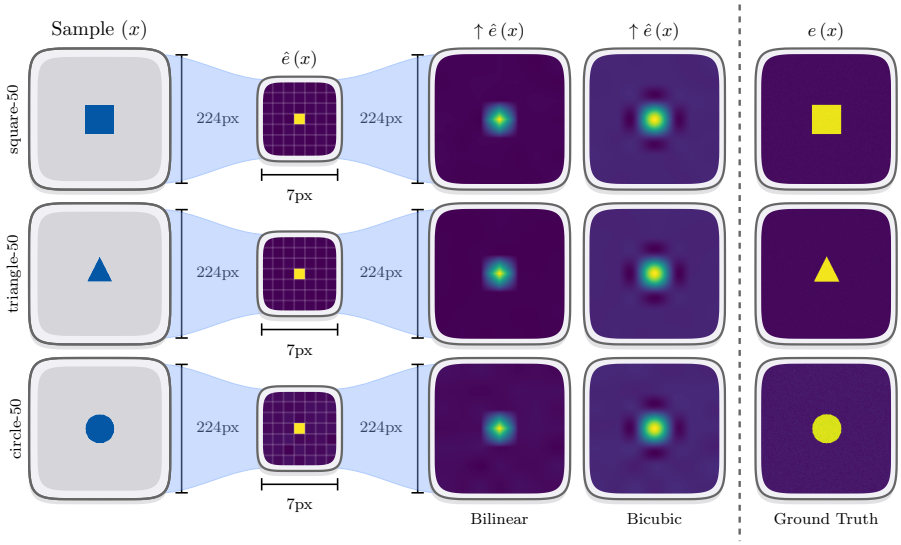

Figure 8: **Identical saliency for distinct inputs.** Each row shows (left) the input, (middle) the raw heatmap, (right) the upsampled saliency map, and (rightmost) the ground-truth attribution priors. Despite distinct ground truths, the upsampled explanations converge to a single pattern: interpolation cannot distinguish what the model attends to from what it ignores. RRL-constrained (Ross et al., 2017) model on synthetic shapes; see the supplemental material for setup and extended results.

RRL loss. The Right for the Right Reasons (Ross et al., 2017) loss adds a penalty on background attribution:

$$\mathcal{L}_{\text{RRL}} = \mathcal{L}_{\text{CE}} + \lambda \sum_{x \notin \text{FG}} |A(x)|^2, \tag{19}$$

where FG is the ground-truth foreground mask and $\lambda = 0.1$. This constrains the model to attend exclusively to foreground pixels. Concretely, the penalty forces the input-gradient attribution $\nabla_x s_y$ to redistribute its mass within FG: pixels outside the mask receive near-zero gradient magnitude, so $\text{supp}(\nabla_x s_y) \subseteq \text{FG}$ up to training tolerance. The resulting attribution map has a *known spatial support*, providing the controlled ground truth against which upsampling methods can be evaluated.

USU reconstruction pipeline. Coarse attributions are extracted at grid resolutions $\{4, 7, 14\}$. Segment boundaries emerge from the model's own score heterogeneity via hierarchical boundary refinement (Section C), refined where scores transition across segments and left coarse where scores are homogeneous; USU's formal guarantees are independent of how $\mathcal{S}$ is obtained. The tensor potential temperature is $\varepsilon = 0.1$. Oracle variants (Oracle-USU and Oracle-IWMR) bypass hierarchical boundary refinement (Section C), supplying ground-truth segmentation partitions $\mathcal{S}$ directly to isolate the redistribution machinery: any gap between a USU variant and its oracle upper bound measures boundary-detection error alone. Figure 7 extends the main paper's qualitative comparison to additional synthetic patterns and in-the-wild samples.

Full synthetic results. Table 5 extends Table 4 with all seven methods across two model families (RRL-MLP and RRL-CNN). Figure 9 provides the extended qualitative comparison across all nine configurations.

F Evaluation Metrics

We select six established metrics from the XAI evaluation literature (Table 6) to assess complementary properties of our method across faithfulness and complexity dimensions. All metrics are computed using the Quantus library (Hedström et al., 2023).

Table 4: **Controlled validation: shape classification.** IoU (↑), Conc. (↑), PG (↑). Bold is best non-oracle; oracles use ground-truth scores.

Model	Method	IoU	Conc.	PG
	Bilinear	0.79	0.54	0.78
	Bicubic	0.82	0.59	0.79
RRL-CNN	USU	**0.86**	**0.76**	**1.00**
	Oracle-USU	0.96	0.88	1.00
	Oracle-IWMR	1.00	0.85	1.00

G Statistical Methodology

With metrics defined, we describe the statistical methodology ensuring that observed improvements are genuine rather than artifacts of sampling variability.

Hypothesis testing. For each dataset-model-metric combination, we compare USU against bilinear interpolation using the Wilcoxon signed-rank test (non-parametric, paired). We apply Bonferroni correction for the $9 \times 6 = 54$ comparisons (9 dataset-model pairs $\times$ 6 metrics), yielding a corrected significance threshold of $\alpha = 0.05/54 \approx 9.3 \times 10^{-4}$. All reported p-values satisfy $p < 0.001$ after correction.

Effect sizes. We report Cohen's d as the standardized mean difference. Table 7 gives standard interpretation thresholds alongside observed effect sizes for USU vs. bilinear on Infidelity.

Confidence intervals. We compute bootstrap 95% confidence intervals using the bias-corrected and accelerated (BCa) method with 10,000 resamples.

H Attribution Method Generalization

USU satisfies (D1)–(D4) for every coarse map (Theorem 5), whatever attribution algorithm produced it. Whether its empirical gains are consistent across algorithms does not follow from these guarantees; we test this by benchmarking seven attribution methods: GradCAM, HiResCAM, GradCAM++, XGradCAM, FullGrad, ShapleyCAM, and FinerCAM. These span qualitatively distinct strategies, from gradient weighting and second-order expansions to full-gradient aggregation and Shapley values.

Table 9 reports raw infidelity values (bilinear baseline $\rightarrow$ upsampled) for both USU and USU-IWMR, averaged across VGG16, ResNet50, and ViT-B-16. Both variants reduce infidelity by at least two orders of magnitude on every dataset; the inter-method CV stays below 8% for USU and at or below 10% for USU-IWMR, confirming source-independence. USU-IWMR uniformly improves over USU: its mean improvement factor is 37% higher on ImageNet, 37% on CIFAR-10, and 39% on CUB-200, a consistent multiplicative benefit. The controlled experiments (Section 4.1) provide further evidence using raw gradient attributions from architectures including MLPs.

I Full Dataset Results

We conclude with complete results across all six metrics, three datasets, three architectures, and six upsampling methods. USU consistently outperforms interpolation baselines; Infidelity improvements range from $5.6\times$ to four orders of magnitude.

Tables 8 and 10 report the complete results across all methods, models, and metrics for each dataset.

Table 5: **Full synthetic results: shape classification.** Best IoU (↑), attribution concentration (↑), and pointing game accuracy (↑) across all methods and model families. **Boldface** marks the best non-oracle result per model; shading highlights our methods.

Model	Method	IoU ↑	Conc. ↑	PG ↑
RRL-MLP	Nearest	0.77	0.52	0.76
	Bilinear	0.81	0.56	0.80
	Bicubic	0.84	0.55	0.79
	USU	0.88	0.79	**1.00**
	USU-IWMR	**0.91**	**0.84**	**1.00**
	Oracle-USU	0.97	0.90	1.00
	Oracle-IWMR	1.00	0.95	1.00
RRL-CNN	Nearest	0.75	0.50	0.74
	Bilinear	0.79	0.54	0.78
	Bicubic	0.82	0.59	0.79
	USU	0.86	0.76	**1.00**
	USU-IWMR	**0.89**	**0.81**	**1.00**
	Oracle-USU	0.96	0.88	1.00
	Oracle-IWMR	1.00	0.85	1.00

Table 6: **Evaluation metrics.** All metrics are computed via Quantus. Direction indicates whether higher (↑) or lower (↓) values are better.

Metric	Dir.	Description
Infidelity	↓	Expected squared error between attribution-weighted perturbations and model output changes.
IROF	↑	Area over the class-score curve as superpixels are removed in order of attribution (Iterative Removal of Features).
Monotonicity	↑	Spearman correlation between attribution rank and feature deletion impact.
PixelFlipping	↓	Area under the perturbation curve when removing pixels in attribution order.
ROAD	↓	Remove-and-debias evaluation; measures explanation quality under debiased perturbation.
Sparseness	↑	Gini coefficient of the attribution map; higher values indicate more concentrated explanations.

Table 7: **Effect size analysis.** **(a)** Cohen's d interpretation thresholds. **(b)** Observed Cohen's d for USU vs. bilinear on Infidelity. Shading marks the active category: all configurations exhibit large effects.

(a) Thresholds

Magnitude	$\lvert d \rvert$
Small	0.2–0.5
Medium	0.5–0.8
Large	> 0.8

(b) Observed values

Dataset	VGG16	ResNet50	ViT-B-16
ImageNet	+1.59	+1.82	+1.71
CIFAR-10	+1.95	+1.82	+2.14
CUB-200	+1.87	+0.89	+1.65

Table 8: **Full ImageNet and CIFAR-10 results.** Six metrics across three architectures and six upsampling methods, grouped by direction; USU-IWMR is the Soft IWMR variant (Section 3.5). **Boldface** marks the best per model and metric; shading highlights our methods.

		lower is better			*higher is better*		
Model	**Method**	**Infid.** ↓	**PxFlip** ↓	**ROAD** ↓	**IROF** ↑	**Mono.** ↑	**Sparse.** ↑
ImageNet							
VGG16	Nearest	7.15×10^{6}	0.62	0.58	0.41	0.18	0.31
	Bilinear	6.91×10^{6}	0.60	0.56	0.43	0.19	0.33
	Bicubic	7.82×10^{6}	0.61	0.57	0.42	0.18	0.32
	Lanczos-3	8.14×10^{6}	0.62	0.58	0.41	0.18	0.31
	USU	1.27×10^{5}	**0.27**	**0.23**	0.61	**0.34**	0.73
	USU-IWMR	$\mathbf{1.13 \times 10^{5}}$	0.29	0.26	**0.65**	0.30	**0.74**
ResNet50	Nearest	5.02×10^{7}	0.65	0.61	0.38	0.17	0.29
	Bilinear	4.87×10^{7}	0.63	0.59	0.40	0.18	0.31
	Bicubic	5.41×10^{7}	0.64	0.60	0.39	0.17	0.30
	Lanczos-3	5.63×10^{7}	0.65	0.61	0.38	0.17	0.29
	USU	2.45×10^{5}	**0.31**	**0.28**	0.59	**0.30**	0.70
	USU-IWMR	$\mathbf{2.21 \times 10^{5}}$	0.32	0.29	**0.61**	0.29	**0.72**
ViT-B-16	Nearest	2.31×10^{8}	0.68	0.64	0.35	0.20	0.27
	Bilinear	2.19×10^{8}	0.66	0.62	0.37	0.20	0.29
	Bicubic	2.48×10^{8}	0.67	0.63	0.36	0.20	0.28
	Lanczos-3	2.59×10^{8}	0.68	0.64	0.35	0.20	0.27
	USU	7.85×10^{5}	**0.34**	**0.31**	0.56	**0.30**	0.68
	USU-IWMR	$\mathbf{7.02 \times 10^{5}}$	0.35	0.32	**0.58**	**0.30**	**0.70**
CIFAR-10							
VGG16	Nearest	3.28×10^{7}	0.64	0.60	0.39	0.17	0.30
	Bilinear	3.12×10^{7}	0.62	0.58	0.41	0.18	0.32
	Bicubic	3.55×10^{7}	0.63	0.59	0.40	0.17	0.31
	Lanczos-3	3.71×10^{7}	0.64	0.60	0.39	0.17	0.30
	USU	7.12×10^{4}	**0.30**	**0.27**	0.60	**0.30**	0.72
	USU-IWMR	$\mathbf{6.47 \times 10^{4}}$	0.31	0.28	**0.62**	0.29	**0.74**
ResNet50	Nearest	9.18×10^{7}	0.66	0.62	0.37	0.16	0.28
	Bilinear	8.95×10^{7}	0.64	0.60	0.39	0.17	0.30
	Bicubic	9.87×10^{7}	0.65	0.61	0.38	0.16	0.29
	Lanczos-3	1.02×10^{8}	0.66	0.62	0.37	0.16	0.28
	USU	2.38×10^{5}	**0.32**	**0.29**	0.58	**0.29**	0.69
	USU-IWMR	$\mathbf{2.21 \times 10^{5}}$	0.33	0.30	**0.60**	0.28	**0.71**
ViT-B-16	Nearest	2.71×10^{8}	0.69	0.65	0.34	0.20	0.26
	Bilinear	2.57×10^{8}	0.67	0.63	0.36	0.20	0.28
	Bicubic	2.89×10^{8}	0.68	0.64	0.35	0.20	0.27
	Lanczos-3	3.01×10^{8}	0.69	0.65	0.34	0.20	0.26
	USU	2.15×10^{4}	**0.34**	**0.32**	0.55	**0.30**	0.67
	USU-IWMR	$\mathbf{1.98 \times 10^{4}}$	0.36	0.34	**0.58**	**0.30**	**0.70**

Table 9: **Attribution method generalization.** Raw infidelity values (bilinear → variant), averaged across VGG16, ResNet50, and ViT-B-16.

	lower is better		
Attribution Source	**ImageNet ↓**	**CIFAR-10 ↓**	**CUB-200 ↓**
USU			
GradCAM	1,384 → 3.15	1,349 → 0.230	879.9 → 0.932
HiResCAM	337.4 → 0.853	265.4 → 0.051	176.2 → 0.197
GradCAM++	1,041 → 2.47	1,045 → 0.187	715.1 → 0.773
XGradCAM	1,318 → 2.83	1,637 → 0.281	711.7 → 0.683
FullGrad	2,295 → 6.18	1,839 → 0.379	877.6 → 1.05
ShapleyCAM	1,504 → 3.71	2,631 → 0.485	800.2 → 0.892
FinerCAM	1,411 → 3.48	1,358 → 0.283	892.3 → 0.954
Mean factor ± std	415 ± 29	5,363 ± 407	925 ± 58
CV (%)	6.9	7.6	6.3
USU-IWMR			
GradCAM	1,384 → 2.22	1,349 → 0.177	879.9 → 0.688
HiResCAM	337.4 → 0.628	265.4 → 0.037	176.2 → 0.145
GradCAM++	1,041 → 1.71	1,045 → 0.127	715.1 → 0.526
XGradCAM	1,318 → 2.07	1,637 → 0.211	711.7 → 0.484
FullGrad	2,295 → 4.88	1,839 → 0.292	877.6 → 0.744
ShapleyCAM	1,504 → 2.67	2,631 → 0.330	800.2 → 0.667
FinerCAM	1,411 → 2.64	1,358 → 0.218	892.3 → 0.691
Mean factor ± std	567 ± 55	7,329 ± 731	1,284 ± 95
CV (%)	9.7	10.0	7.4

Table 10: **Full CUB-200-2011 results.** Same format as Table 8.

		lower is better			*higher is better*		
Model	**Method**	**Infid. ↓**	**PxFlip ↓**	**ROAD ↓**	**IROF ↑**	**Mono. ↑**	**Sparse. ↑**
VGG16	Nearest	5.68×10^6	0.63	0.59	0.40	0.18	0.30
	Bilinear	5.44×10^6	0.61	0.57	0.42	0.19	0.32
	Bicubic	6.12×10^6	0.62	0.58	0.41	0.18	0.31
	Lanczos-3	6.38×10^6	0.63	0.59	0.40	0.18	0.30
	USU	1.83×10^3	**0.27**	**0.26**	0.59	**0.32**	0.71
	USU-IWMR	$\mathbf{1.70 \times 10^3}$	0.30	0.27	**0.62**	0.30	**0.74**
ResNet50	Nearest	8.71×10^7	0.67	0.63	0.36	0.16	0.27
	Bilinear	8.46×10^7	0.65	0.61	0.38	0.17	0.29
	Bicubic	9.32×10^7	0.66	0.62	0.37	0.16	0.28
	Lanczos-3	9.68×10^7	0.67	0.63	0.36	0.16	0.27
	USU	1.52×10^7	**0.43**	**0.39**	0.50	**0.28**	0.55
	USU-IWMR	$\mathbf{1.40 \times 10^7}$	0.46	0.40	**0.51**	0.27	**0.59**
ViT-B-16	Nearest	1.92×10^8	0.69	0.65	0.34	0.20	0.26
	Bilinear	1.83×10^8	0.67	0.63	0.36	0.20	0.28
	Bicubic	2.05×10^8	0.68	0.64	0.35	0.20	0.27
	Lanczos-3	2.14×10^8	0.69	0.65	0.34	0.20	0.26
	USU	5.61×10^4	**0.35**	**0.32**	0.55	**0.30**	0.67
	USU-IWMR	$\mathbf{5.08 \times 10^4}$	0.36	0.33	**0.57**	**0.30**	**0.69**

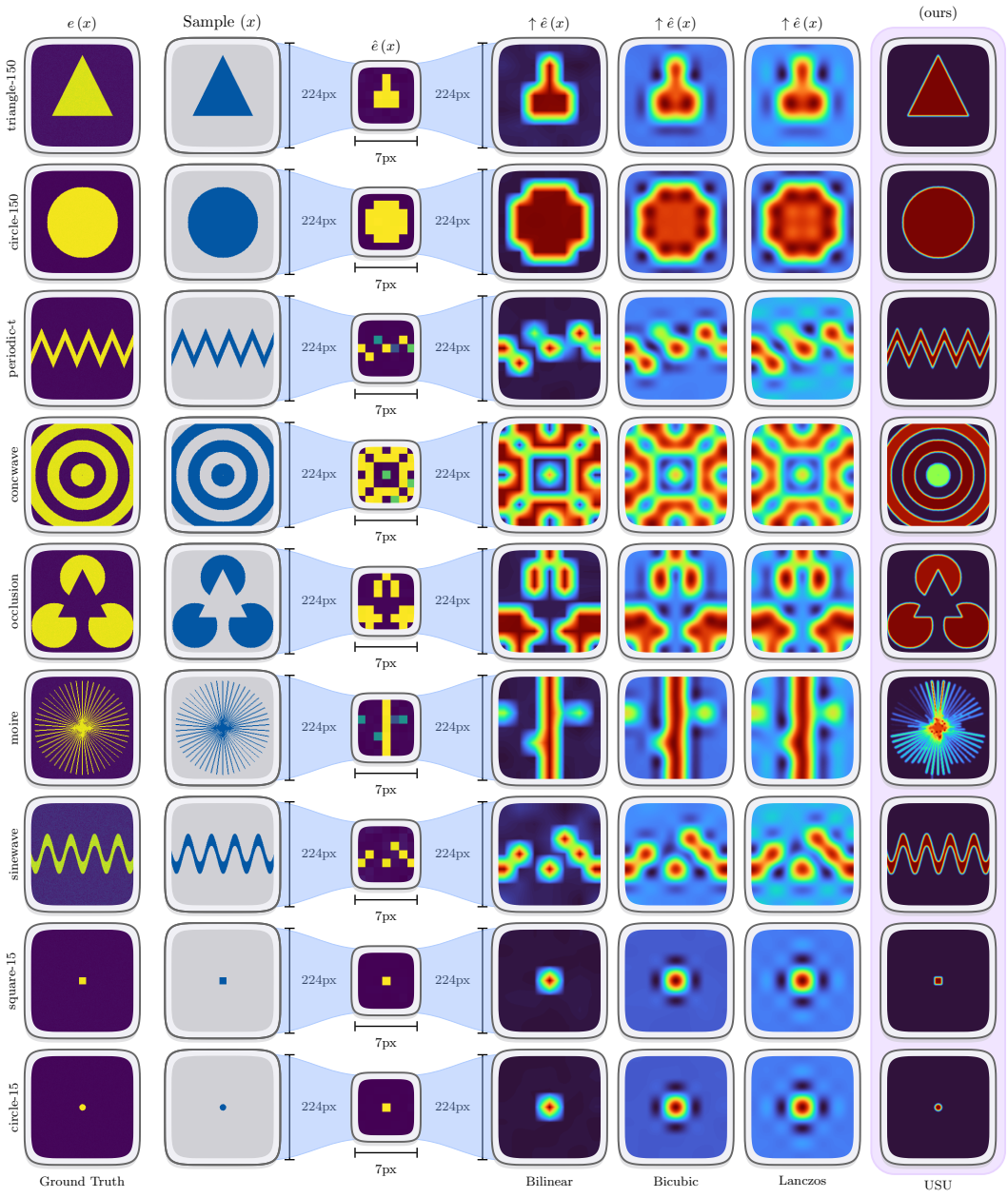

Figure 9: **Interpolation artifacts vs. USU reconstruction on synthetic data.** Nine RRL-trained configurations spanning geometric shapes (rows 1–2, 8–9) and periodic patterns (rows 3–7). From left: ground-truth attribution $e(x)$, input sample, coarse 7×7 heatmap $\hat{e}(x)$, bilinear, bicubic, and Lanczos upsampling, and USU (ours). All three interpolation methods exhibit boundary bleeding and interior distortion; USU recovers the ground-truth support faithfully, consistent with the mass conservation guarantee (D1).