Rethinking Explainer Trust: A Position on the Inconsistencies of Visual Explanations in Weakly Supervised Segmentation

Ayush Somani **Dilip K. Prasad**

Department of Computer Science, UiT The Arctic University of Norway

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Paper ID: eXCV-12

Attribution ≠ segmentation. Saliency maps highlight **decision evidence**, not **object extent**—treat them as validated cues.

POSITION → WHY IT MATTERS

Post-hoc maps explain **decision basis**, not **object extent**—repurposing them as WSSS labels is **fragile by design**. We advocate a **diagnose-then-assist** protocol.

Rephrase principle: **"Not an issue with explainers; an issue with using them as masks".**

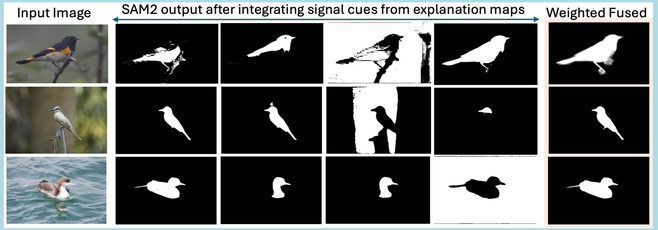

Figure 1. *Misalignment between explanation cues: example heatmap vs. SAM2 proxy mask using prompt signal cues from various explanation techniques.*

MOTIVATION

❖ **What's a property** (not a bug)

- **Discriminative focus** (small, decisive regions) → good for interpretability, insufficient for full masks.
- **Spurious/Context cues** → valuable for debugging, dangerous as labels.
- **Method disagreement** → expected from different priors; requires agreement checks, not cherry-picking.

❖ **Related work** (acknowledging the field)

- Evaluation/benchmarks of XAI (e.g., OpenXAI, Quantus, CLEVR-XAI, medical saliency audits).
- Alignment/robustness of explanations; human-in-the-loop weak supervision; explaining foundation models.

DATASETS PANEL

Table 1. *Comparison of datasets used in study.*

Dataset	#Classes	#Images	Domain
CUB-200-2011	200	11,788	Bird species (fine-grained)
Pascal VOC2012	20	1,450	General objects (natural)
USIS10K	7	10,632	Underwater scene (natural)
Sessile-Kvasir-SEG	1 (polyp)	1,000	Gastroenterology (medical)

EVIDENCE SNAPSHOT → PRACTICAL PROTOCOL

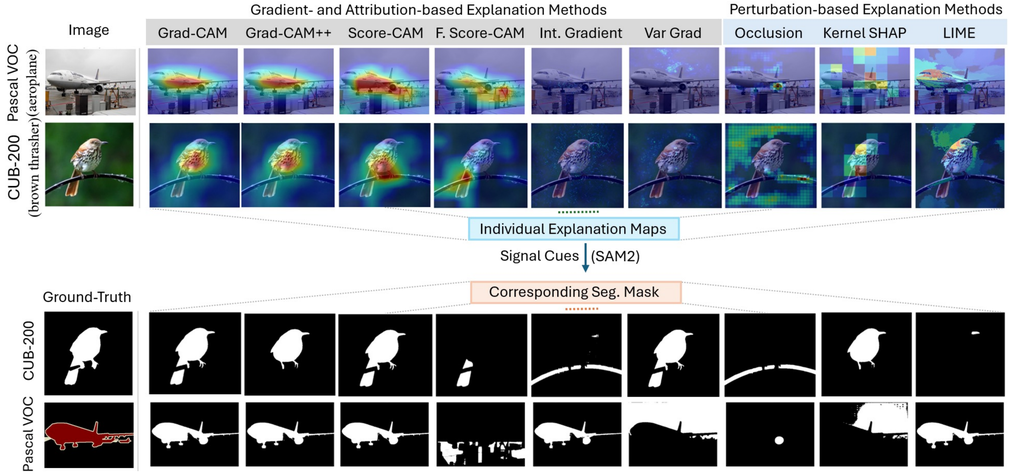

Figure 2. *Illustration of misalignment. Top: class-conditioned heatmaps overlayed for the "aeroplane" (Pascal VOC) and "brown thrasher" (CUB-200-2011); Bottom: proxy masks from SAM2 prompted by respective explanation cues; low pairwise overlap exposes unreliability for supervision.*

❖ **Metrics:**

- **Coverage:** |M∩GT|/|GT| on maps thresholded to binary support.
- **Spill:** |M\GT|/|M| on explanation maps.
- **Localization Accuracy (LA):** pointing-game hit rate (max map point ∈ GT).
- **mIoU:** mean IoU of final masks (after SAM prompting).
- **Effectiveness/Fidelity:** confidence drop when masking the attributed region.

❖ **What we observe (concise)**

- Coverage is often partial; Spill to background is common.
- Low inter-explainer overlap on the same image.
- Directly thresholding maps → unstable masks.

Flow: E_k → reliability weights → fused map → dual cues (FG/BG) → SAM2 prompt → mask.

❖ **How we safely use attribution (FM-assisted protocol):**

1. **Many maps E_k :** per-image from diverse explainers (Grad-CAM, Grad-CAM++, Score-CAM, IG, etc.).
2. **Per-image reliability w_k:** prefer **consensus** (IoU agreement), **focused** (low entropy), and **robust** (stable under Δ perturbations).
3. **Dual cues**: Fuse $E_\Sigma = \sum w_k E_k$; top-percentile → foreground points P, bottom-percentile → background points B.
4. **Prompt a segmenter** (e.g., SAM2) with (P,B) → refined mask $\hat{M}$ (no learnable prompts).
5. **Report separately:** Interpretability (Coverage/Spill, pointing game) v/s. Utility (mIoU, LA).

Ablate: single map vs. mean vs. confidence-weighted; with/without BG seeds; SAM-only vs. XAI→SAM.

RECOMMENDATIONS → DEBATE → SCOPE

❖ **Recommendations (checklist for the community)**

- Do not supervise with a single explainer; quantify agreement first.
- Prefer confidence-weighted fusion + dual cues over naïve thresholding/averaging.
- Declare FM-assisted scope (SAM2 prior) explicitly.

Interpretability ↔ Utility: the gap

Fidelity ≠ spatial completeness. Use segmentation quality as a **diagnostic** for explainer quality and pursue **task-aware faithfulness** (concept-aligned, robust).

Table 2. *Comparison of* explanation *methods—key characteristics, and GT-mask alignment (ResNet-50): mIoU ↑, Effectiveness ↑, Localization Accuracy ↑.*

Method	Strengths	Weaknesses	CUB-200-2011 (IoU / Eff. / L. Acc.)	Pascal VOC 2012 (IoU / Eff. / L. Acc.)	USIS10K (IoU / Eff. / L. Acc.)
Grad-CAM	Fast; stable; decent fidelity	Partial object coverage; coarse	0.70 / 0.84 / **0.96**	0.70 / 0.78 / 0.94	0.55 / 0.54 / 0.89
Grad-CAM++	Improved coverage for multiple objects	Still coarse; depends on conv. features	0.72 / 0.87 / 0.95	0.71 / 0.74 / **0.95**	**0.74** / 0.48 / **0.92**
Score-CAM	Sharp maps; high fidelity	Slow; Expensive; sensitive to masks	0.56 / 0.72 / 0.92	0.65 / 0.70 / 0.94	0.52 / 0.56 / 0.88
FasterScore-CAM	10× faster than Score-CAM	Skips minor features; still costly	0.49 / 0.59 / 0.94	0.64 / 0.72 / 0.92	0.54 / 0.59 / 0.85
Integrated Gradients	Pixel-level detail; theoretical completeness	Noisy attributions; multiple steps; gradient dilution	0.66 / 0.81 / 0.81	0.55 / 0.64 / 0.78	0.57 / **0.61** / 0.78
VarGrad	Stability via averaging gradients	Oversmoothing; sampling cost	0.40 / 0.44 / 0.62	0.40 / 0.62 / 0.60	0.42 / 0.46 / 0.76
Occlusion	Causal; model-agnostic	Slow; coarse heatmaps	0.48 / 0.71 / 0.65	0.44 / **0.81** / 0.80	0.66 / 0.50 / 0.87
KernelSHAP	Fair, model-agnostic, local	Heavy sampling; blocky maps	0.31 / 0.46 / 0.82	0.40 / 0.68 / 0.78	0.29 / 0.44 / 0.90
LIME	Broad part coverage; interpretable	blocky; heavy sampling; low fidelity	0.48 / 0.72 / 0.86	0.46 / 0.50 / 0.75	0.42 / 0.55 / 0.88
Fused-Weighted	Consistent; high utility	Requires fusion, added complexity	**0.77** / **0.91** / **0.96**	**0.78** / **0.81** / 0.84	0.55 / 0.60 / **0.92**

TAKEAWAY

- ✓ Coverage & Spill vs. SAM2 proxy show common failures: Grad-CAM ~50–60% coverage on VOC; LIME/SHAP ≥30% spill.
- ✓ Maps as prompts to SAM2 dramatically improves masks; cues reach ~78.4% mIoU on VOC test, rivalling fully supervised DeepLabV3—but success comes from SAM's correction, not from raw maps.
- ✓ Don't supervise with single explainers. Validate agreement, separate metrics, and—when needed—use explainers as guidance with transparent caveats about foundation-model priors.
- ✓ Explanations should support trust, not replace it. Raw saliency is insufficient for full masks. Assisted (XAI→dual-cues→SAM), masks improve because of the segmenter prior, not because maps "become" masks.

ACKNOWLEDGEMENT

This work was supported by the Research Council of Norway Project (nanoAI, Project ID: 325741).