

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

{firstname.lastname}@uit.no

Abstract

Post-hoc explainability methods (e.g., Grad-CAM, Integrated Gradients, LIME, SHAP) are widely adopted in weakly supervised semantic segmentation (WSSS) as pseudo-annotators or seed generators. This paper argues that these visual explanations, while convenient, often produce unreliable and misaligned cues for segmentation. Drawing on experiments with Pascal VOC 2012, CUB-200-2011, and USIS10K, we highlight how saliency-based explanations are frequently inconsistent across methods, focus on incomplete or spurious regions, and show poor correlation with actual object masks or downstream segmentation performance. We demonstrate qualitatively and quantitatively how individual explainers can fail—e.g., highlighting only small discriminative parts of an object or even background artifacts—and how different explainers often disagree with each other on the same image. Using a foundation segmentation model like Segment Anything Model as a proxy for high-quality segmentation, we reveal significant mismatches between explanation maps and object extents. These findings question the trustworthiness of saliency maps as supervision signals for segmentation. We further discuss the gap between interpretability vs. utility: an explanation may faithfully reflect a model’s prediction logic and yet be a poor proxy for the full object region needed in segmentation tasks. We conclude with challenges for the XAI and segmentation communities: Should we blindly trust visual explanation maps for pixel-level supervision? How do we ensure their faithfulness and consistency if repurposed for guiding segmentation models? Can we evaluate explainer quality using downstream segmentation performance? Our position advocates for a principled re-examination of explainer reliability in WSSS, aiming to foster methods that bridge the gap between human-interpretable explanations and effective segmentation cues.

1. Introduction

Position. *Post-hoc attribution maps are designed to reflect a classifier’s decision basis, not to delineate objects exhaustively. Treating them as segmentation surrogates in WSSS is therefore fragile by design. Our stance is not to “disprove” explainers, but to articulate a practical framework that (i) diagnoses the spatial completeness of explanations via a strong proxy oracle and (ii) repurposes explanations as guidance signals—not pseudo ground truth—for mask refinement.*

Weakly-Supervised Semantic Segmentation (WSSS) aims to learn pixel masks from cheap supervision (e.g. image labels) instead of dense pixel-level annotations [32]. A common pipeline first trains a classifier, extracts spatial localization cues via Class Activation Maps (CAMs) or related post-hoc explanations, and uses them as seeds [26, 28], linking the presence of a class to image regions [11]. This practice is effective yet conceptually mismatched: saliency emphasizes discriminative regions, whereas segmentation requires complete object extents. Decades of observation confirm partial-coverage and context-bias artifacts [2, 17, 21, 27]. For example, a CAM may strongly highlight the wings on an aeroplane or the belly of a bird but neglect the plane’s body or the bird’s crown shown in Figure 1. This well-known partial activation problem results in incomplete pseudo-labels, which in turn undermines the final segmentation performance.

Indeed, it is widely recognized that using only image-level supervision causes classifiers to focus on small salient parts of objects rather than their full extent [27]. WSSS pipelines must therefore cope with these imperfect seeds. A long line of WSSS research has proposed methods to refine CAM outputs—for instance, by enforcing consistency under image perturbations, propagating labels with affinity between pixels, or calibrating activation thresholds [3, 6, 32].

In parallel, Explainable AI (XAI) has developed a

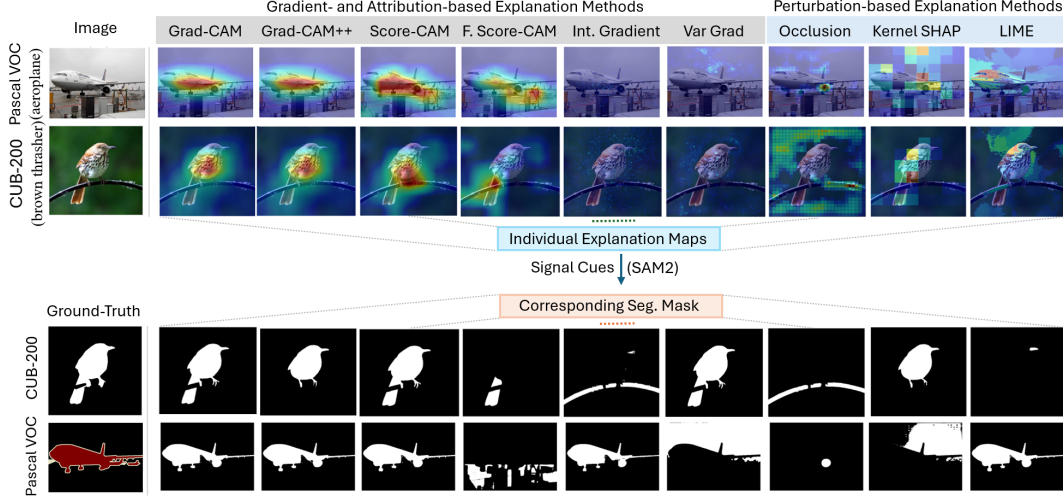


Figure 1. Illustration of misalignment. *Top*: class-conditioned heatmaps overlaid for the “aeroplane” (Pascal VOC) and “brown thrasher” (CUB-200-2011); *Bottom*: proxy masks from SAM2 prompted by respective explanation cues; misalignment is expected for attributions and quantified by our protocol.

range of post-hoc explanation techniques beyond CAM. Gradient-based methods like Grad-CAM, Grad-CAM++ and VarGrad; perturbation-based methods like Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP); and attribution methods like Integrated Gradients (IG) all aim to highlight “important” pixels for a classifier’s prediction [21]. In WSSS, some works have adopted or fused such explainers to improve segmentation cues [4, 23], assuming that complementary maps may better localize objects.

This, however, raises key concerns about consistency and trust: *if different explanation methods produce substantially different maps for the same input and model, which one (if any) should we trust?* Do they highlight the same important regions or present conflicting signals? Moreover, are these maps reliable enough to serve as segmentation surrogates? The community already knows attribution $\neq$ segmentation; what is missing is a lightweight, reproducible bridge that (i) measures the gap and (ii) uses explanations safely as assistance rather than supervision, with limitations spelled out.

We compile evidence across Pascal VOC [5], CUB-200 [24], USIS10K [15], and Kvasir-SEG [9] showing that maps—gradient, attribution, and perturbation based—are inconsistent and often incomplete when judged against object masks. Rather than seeking a single “correct” explainer, we (1) quantify inter-explainer disagreement and (2) use a frozen foundation segmenter (SAM2) [20] as a proxy oracle to assess coverage/spill and to refine fused explanation cues into full masks. This turns explanations into an actionable but validated signal. Related evaluation agendas (OpenXAI, Quantus) likewise argue for careful, task-aware assessment [1, 8].

Table 1. Comparison of four datasets used in this 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)

Fig. 1 illustrates the core tension: explanations often highlight only parts of an object, while the proxy mask covers the whole. Our contributions are:

- *Positioned diagnosis*: a simple protocol to evaluate explanation completeness with a proxy oracle (coverage/spill + pointing game), clarifying why raw saliency is brittle for WSSS.
- *Practice recipe*: per-image, confidence-weighted multi-explainer fusion and dual-threshold prompting that turns saliency into foreground/background guidance for SAM2—no learnable prompts, no FM fine-tuning.
- *Transparent scope*: results reported on a frozen ResNet-50 classifier (ImageNet-1k). Preliminary tests with DeiT-S/ViT-B and Swin-T on two datasets showed marginal gains; we keep them out of scope here to focus on the position and protocol.

Scope & Assumptions. Strict WSSS assumes no segmentation supervision. Using a frozen foundation model (FM) at inference as a refiner does not introduce task-specific mask labels but does import a powerful segmentation prior. We thus adopt the term *FM-assisted weak supervision* and explicitly report this limitation and its implications (Sec. 4).

2. Failure Modes of Visual Explanations

Despite their popularity in WSSS pipelines, post-hoc methods exhibit behaviors that are desirable for interpretability yet problematic for supervision. Saliency emphasizes features most decisive for the class (faithfulness), which may omit non-decisive object parts (completeness). We summarize representative behaviors—partial coverage, contextual cues (“Clever Hans”) [12], and instability across stochastic perturbations—together with simple quantitative probes that matter for WSSS (Table 2).

Incomplete Object Coverage: CAM-based methods typically focus on small, highly discriminative regions, ignoring less class-relevant regions. This leads to fragmentary localization. For instance, Score-CAM for the “bird” class may activate on the beak or belly but miss the rest of the body. We measured the overlap between explanation maps and ground-truth masks (using the Intersection-over-Union (IoU), after suitable thresholding of the maps). Our experiments show Score-CAM achieves only 0.4–0.6 on average across datasets (Table 2), confirming that a large portion of the object is left unexplained. This aligns with the latest findings that partial localization is a fundamental limitation of classifier-based saliency [30]. Even FasterScore-CAM or VarGrad offer only modest improvements, frequently missing substantial object parts. Some attribution methods, like IG, can produce even more sparse or diffuse maps with scattered pixel importance and weak coverage depending on how they accumulate importance. Conversely, Grad-CAM can occasionally overspread into the background when the discriminative region is broad or overlaps multiple objects, producing coarse masks. Thus, explainers may either under-cover or over-generalize.

Spurious Regions and Artifacts: Another issue is that explanation maps may highlight irrelevant background. Since explainers reflect what influences model predictions, they can surface contextual artifacts. For instance, in the CUB-200-2011, the classifier may rely on parts of a perch or background foliage that correlate with certain bird species, and saliency maps highlight those instead of the bird. In Pascal VOC, Score-CAM sometimes activates on photographic vignette edges, which obviously are not part of the object. In medical imaging, saliency maps have been shown to “focus on spurious correlations, e.g., patches, bubbles, or ruler marks introduced during image acquisition,” instead of actual pathology [18]. Such false positives in the explanation can mislead segmentation if used as seed annotations, causing the model to learn background regions. We also observed quantitatively that perturbation-based methods (e.g., LIME, SHAP) often exhibit low precision, with many highlighted pixels falling outside the true object.

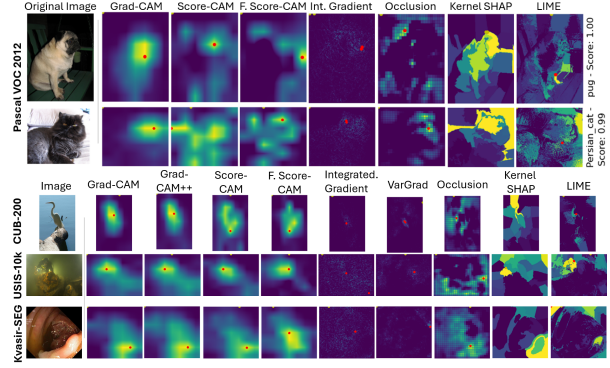


Figure 2. Qualitative comparison of post-hoc explanations across datasets. The red dot marks the global maximum of each map (used for Pointing-Game@1 and as a seed cue in our pipeline). These examples illustrate incomplete coverage and inter-explainer disagreement, which become more pronounced in cluttered (USIS-10K) and medical (Kvasir-SEG) scenes.

Noisy and Unstable Attributions: Different explanation techniques also show idiosyncratic noise patterns. For instance, Occlusion which blanks out patches to see the effect on output can produce noisy, grid-like saliency maps, with many false flickers on background regions. LIME perturbs superpixels, but highlighted regions may straddle object boundaries or include irrelevant areas. SHAP, especially with image baselines, assigns both positive and negative contributions, creating confusing or scattered maps. These methods are often unstable—small changes to the input or random seeds can shift results and would require extensive post-processing to be usable as segmentation seeds.

In summary, each explainer exhibits distinct features, as summarized in Table 2. If one were to treat these raw explanation maps as input prompts for training a segmentation model, models may learn incomplete or misleading object structures. In practice, WSSS pipelines apply additional steps (such as iterative expansion of seeds, graphical model smoothing like CRFs, or combining multiple maps) to correct for these incomplete or noisy localization cues [28]. The need for such correction underscores our position: *image classifiers (through explanations), in isolation, are not consistently trustworthy for segmentation*. See also systematic evaluations [1, 8, 16, 19] for complementary diagnoses.

3. Inter-Explainer Disagreement and Trust Implications

Post-hoc methods often provide different but valid perspectives on the same decision [25]. Accordingly, we treat cross-method agreement not as a ground truth but as a *stability proxy*: high agreement suggests robust, non-idiosyncratic evidence; sharp disagreement flags potential artifacts or narrow features [1, 8].

Table 2. Comparison of explanation methods: key characteristics, computational cost, strengths/weaknesses, and quantitative alignment (mIoU $\uparrow$, Effectiveness $\uparrow$ [31], Localization Accuracy $\uparrow$ [29]) between explanation maps and ground-truth mask (ResNet-50).

Method	Granularity	Cost	Strengths	Weaknesses	CUB-200-2011 (IoU / Eff. / L. Acc.)	Pascal VOC 2012 (IoU / Eff. / L. Acc.)	USIS10K (IoU / Eff. / L. Acc.)
Grad-CAM	Coarse	Low	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++	Coarse	Low–Moderate	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	Moderate	High	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	Moderate	Moderate	10 $\times$ 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	Fine	Moderate	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	Fine	Moderate–High	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	Patch-Level	High	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	Superpixel-Level	High	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	Superpixel-Level	High	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	Fine	Moderate	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

We evaluated multiple explainers on the same classifier across thousands of images and observed surprisingly low-to-moderate agreement. For example, Grad-CAM might highlight a highly discriminative region, while LIME, due to its superpixel approach, picks a completely different region—or several scattered ones. Figure 2 visualizes such divergence. Quantitatively, pairwise mIoU between Grad-CAM and LIME on Pascal VOC was often below 0.4, with similar low overlap between other methods (e.g., SHAP, IG, Score-CAM). Each explainer offers a distinct view, shaped by its assumptions—gradients, perturbations, or attribution—producing differing saliency maps. This inconsistency isn’t unique to image models. Prior studies (e.g., human activity recognition) have shown that SHAP and Grad-CAM may disagree even on feature importance rankings [22]. The underlying message holds: *explanation is method-dependent*.

In the absence of ground-truth explanations, cross-method agreement can serve as a proxy for stability: if diverse methods highlight the same region, it is likely meaningful, whereas outliers may reflect artifacts. Although some works compare saliency maps to segmentation masks (e.g., IoU), such evaluations are rare—only 21% of XAI papers report localization metrics [18]. Other approaches reward consensus (low entropy) or penalize outliers (e.g., KL divergence), but this remains problematic since classification lacks true pixel-level ground truth unlike segmentation.

From a trust standpoint, explanations should be

judged relative to use. For WSSS, stability and spatial completeness matter; for auditing, sensitivity/deletion may suffice. We therefore report both. Naïve averaging is suboptimal; confidence-weighted fusion balances sharpness and consensus. (Sec. 4).

For WSSS, the issue is more pressing: which explainer’s map should serve as the signal prompts? The lack of consistency should be a red flag—trust in explanations should be provisional, not absolute. Some approaches attempt to fuse maps, hoping to amplify consensus and suppress noise [3, 7]. While fusion (especially with learned confidence weights using Entropy-KL-IoU metric) improves explanation-driven segmentation outcomes quantitatively, sometimes exceeding any single explainer by $\geq +6$ mIoU, it also raises complexity in that one must decide how to weight or combine the maps. Also, requiring multiple explainers to concur is computationally expensive and not commonly done in practice. If done naïvely (e.g., an unweighted average of saliency maps), the result can be a blurry heatmap that still doesn’t align with object boundaries. Instead, many application pipelines still rely on a single explainer (often Grad-CAM due to its popularity and ease), ignoring disagreement.

In summary, visual explanation is not a ground-truth entity but rather an output heavily influenced by the choice of algorithm. In our experiments, explainers not only disagreed on pixel regions but also on shape and structure shown in fig. 2. Typical behaviors are visible: Grad-CAMs focus on coarse, highly discriminative

parts; Score-CAM captures more area but often spills into background; Integrated Gradients is sparse/edge-like; VarGrad oversmooths; Occlusion shows blocky grid artifacts; LIME/Kernel SHAP produce superpixel blocks that frequently include background. Such differences complicate fusion and interpretation. For WSSS, this creates variability in pseudo-labels and downstream performance. Until explainers become more consistent—or until we can reliably assess their accuracy, leveraging them as tools should be done with caution. A weighted fusion helps, as is evident in Table 2, but highlights the core issue: *no single explainer suffices; only by acknowledging their limitations and disagreements can we begin to use them meaningfully.*

4. Foundation Models as a Proxy for Ground Truth

Using a frozen segmenter such as SAM2 [20] at inference introduces a powerful segmentation prior. Under strict WSSS, this breaks the purist assumption. We therefore adopt the practical lens of FM-assisted weak supervision: (i) the classifier is supervised only by image labels, (ii) the foundation model is not fine-tuned or prompt-trained, and (iii) explanations act solely as guidance cues for refinement—akin to classical CRF post-processing, but stronger. We explicitly report this limitation.

Protocol. Given per-image saliency maps $\{E_k\}$ from diverse explainers, we compute per-image weights w_k by maximizing a simple criterion favoring *consensus & confidence*: high IoU with the running fusion, low entropy, and low KL divergence. The fused map $E_\Sigma = \sum_k w_k E_k$ is then turned into *dual cues* via adaptive percentiles: top-85% pixels $\rightarrow$ foreground points (P), bottom-10% pixels $\rightarrow$ background points (B). We feed (P, B) to SAM2 to obtain a refined mask $\hat{M}$ (no learnable prompts). We evaluate explanations by (Coverage,

Spill) against a reference mask M and evaluate masks by mIoU against GT where available.

Evaluating how well saliency-based explanations align with object boundaries is inherently challenging—especially in weakly supervised settings where pixel-level annotations are absent. To circumvent this, we leverage recent advances in foundation models like Segment Anything (SAM2) as a proxy for ground truth. SAM’s zero-shot segmentation abilities, trained on over 1 billion masks, make it a strong candidate for approximating high-quality segmentations from minimal prompts comparable to fully supervised models on standard datasets. [10]. These masks, while not manually annotated, closely resembled true object extents across several datasets (e.g., Pascal VOC, CUB-200), aligning with prior benchmarks showing SAM’s mIoU in the 70–90% range [10]. In fact, leveraging SAM has been considered as a “big leap” in segmentation, showing that a single model can adapt to segment almost anything with the right prompt [10]. We treat these SAM-generated masks as a proxy for ground truth to evaluate our explainers.

To evaluate explanation maps, we compared SAM2 generated masks against ground truth using two metrics:

1. **Coverage:** how much of the SAM2 mask overlaps with the explanation (i.e., recall of the explanation w.r.t. the object).
2. **Spill:** how much of the explanation lies outside the SAM2 mask (false positive rate).

Most popular explainers performed suboptimally by both metrics. For instance, Grad-CAM achieved only 50–60% coverage on average for Pascal VOC, often missing limbs or background objects, while LIME and SHAP showed high spill ($\geq 30\%$ in many cases), frequently activating irrelevant regions. Even the best individual method, Score-CAM or a tuned Grad-CAM (with aggressive thresholding to expand it slightly), left significant gaps.

Interestingly, when explanation maps were used as prompts for SAM2, performance improved dramatically. In other words, use the saliency map as an initial guess of foreground (and possibly background) for SAM2, and let it generate a refined mask as shown in fig. 3. Feeding SAM2 a fused, confidence-weighted saliency map produced segmentation masks that rival the quality of fully supervised models (e.g., 78.4% mIoU on Pascal VOC, matching a supervised DeepLabV3 trained with full annotations in Table 3). However, this success hinged on SAM’s ability to correct and extend the noisy cues—not the inherent quality of the explanation maps themselves. Single-explainer prompts led to significantly inferior masks, exposing their limitations as standalone pseudo-labels. Figure 3 conceptually shows such an example: a raw Grad-CAM vs. the SAM-refined

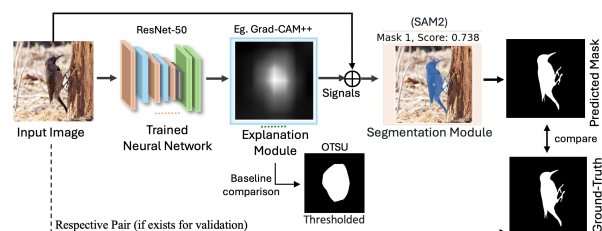


Figure 3. FM-assisted refinement pipeline. A frozen classifier (ResNet-50) yields a post-hoc map (e.g., Grad-CAM++), which we convert to dual-threshold cues to prompt a frozen SAM2 and obtain the predicted mask. An Otsu-thresholded saliency mask serves as a direct baseline; ground truth is used only for evaluation. SAM2 thus acts as a proxy oracle that refines incomplete saliency into segmentation-quality masks.

Table 3. Comparison of different methods w/o extra learnable prompts (all on VOC test). The image-level labels (I) and use of CLIP (C), Grounding DINO (D), and SAM (S) for training/inference are shown below. The † and ‡ indicate backbone pretrained on COCO ground-truth, or ImageNet-21k.

Method	Backbone	Supervision	mIoU
CLIP-ES+VPL _{AAAI'25}	ViT-B16	✓ I + C	77.8
Yang & Gong _{WACV'24}	R-101	✓ I + C + S	76.7
Yang & Gong _{WACV'24}	Swin-L _‡	✓ I + C + S	81.6
CLIP-ES _{CVPR'23}	R-101 _†	✓ I + C	73.9
ToCO _{CVPR'23}	ViT-B _‡	✗ I	72.2
CLIMS _{CVPR'22}	R-50 _†	✓ I + C	70.0
ViT-PCM _{ECCV'22}	R-101	✗ I	70.9
DeepLabV3+ _{CVPR'18}	R-101	Full Sup.	79.4
XAI-fused (Ours)	R-50	✗ I + S	78.4

mask. Grad-CAM alone provides incomplete localization; however, when used as input to SAM2, the resulting mask better captures the full object with fewer extraneous bits. Essentially, SAM acts as a robust proxy annotator, revealing the limitations of the original explanation.

Instead of fine-tuning or prompt-learning for SAM2, we adopt a zero-shot strategy: fusing diverse XAI maps (Grad-CAM, IG, Score-CAM) with per-image weights based on entropy, KL divergence, and IoU. The fused cues guide SAM2 to generate fine-grained masks—without extra training or external supervision. Using a frozen ResNet-50 backbone, our pipeline reaches 78.4 mIoU on VOC 2012, surpassing recent SAM-based pipelines (e.g., CLIP- or DINO-trained SAM pipelines [28]) and approaching fully supervised models.

SAM2 thus serves as a proxy to reveal and refine the limits of visual explanations. While single explainers fall short of segmentation reliability, their confidence-weighted fusion with SAM2 provides both semantic guidance and structural refinement. Ablations confirm that fusion improves mIoU by ≥ 5 points over the best single-explainer, showing gains are not due to SAM2 alone.

5. Pitfalls of Using Explanations as Segmentation Signals

The previous sections exposed how explanation maps can be inconsistent and often misalign with true object regions. We now examine the implications of using such maps as pseudo-labels in WSSS. Despite their convenience, this practice introduces several key pitfalls:

Seed Quality Limits Segmentation: In WSSS, the quality of initial pseudo-labels (seeds) strongly influences the final performance. Saliency-based seeds that

cover only a part of the object or include irrelevant regions can hinder the model’s ability to learn the full object extent demonstrated in fig. 4. While methods like CRFs, multi-view CAMs, or affinity propagation attempt to expand these seeds, they can only work with what the explanation provides. If crucial parts are entirely missing, they remain unrecoverable. Empirically, initial seed IoU correlates with final segmentation IoU—better seeds yield better outcomes. Thus, the initial explanation acts as a performance ceiling, raising the question: *should we improve the explainer or reconsider using it altogether?*

Bias Propagation: Saliency maps can expose model biases. If a classifier highlights background context (e.g., water for boats), a segmentation model trained on these cues may inherit and amplify such biases—labeling water as ‘boat’ simply due to pseudo-label supervision. In effect, the explanation’s mistake becomes the segmentation model’s mistake. This constitutes unfaithful supervision, where the explainer’s focus misaligns with the true object but is treated as ground truth. In worst cases, the segmentation model may localize objects less accurately than the classifier itself.

Lack of Feedback During Training: WSSS often lacks ground truth to validate seed quality during training. Consequently, models can overfit to flawed pseudo-labels, leading to biased or incomplete segmentation—issues that only surface at evaluation. A better approach would involve confidence-aware losses or mechanisms to assess explanation quality during training.

Over-Reliance on Heuristics: To address poor explanation quality, WSSS pipelines frequently rely on heuristics such as multi-threshold labeling, CRFs, objectness priors, or cross-image affinities. While effective, these increase complexity and require dataset-specific tuning. With the rise of models like SAM, a fundamental question arises: should explainers be used as ground truth at all, or should alternative weak supervision signals—e.g., text descriptions, sparse user input,

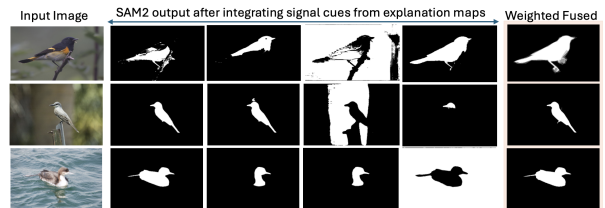


Figure 4. Illustration of inter-explainer variation on the CUB-200-2011 dataset. SAM2 outputs based on four different explanation maps show clear disagreement, each offering a distinct interpretation. The fused map, combining cues from multiple explainers, yields compelling performance.

or pre-trained segmentation outputs be explored?

Misaligned Objectives – Interpretability vs. Segmentation: Post-hoc explainers aim to justify a class prediction, not delineate object boundaries. Classifiers often rely on partial object cues (e.g., a cat’s face), whereas segmentation demands full object coverage. Forcing explainers to act as segmentation tools compromises their faithfulness and repurposes them beyond their design. This inherent tension between faithfulness and completeness has led some works to integrate spatial constraints during classifier training, effectively blending classification with segmentation.

A fundamental tension arises when explanation maps are used for supervision: *should we constrain model design to make explanations usable, even at the cost of task performance?* Training a classifier with standard cross-entropy and expecting its Grad-CAM to behave like a segmentation mask is inherently misaligned—the model was not optimized for that purpose. To obtain more useful explanations, the classifier itself may need to be modified for interpretability, potentially sacrificing predictive performance.

For instance, a saliency map highlighting only a bird’s belly in fig. 4 is faithful to the classifier’s reasoning but insufficient as a full-object supervision cue. Enhancing its utility—e.g., by introducing losses that promote spatial coverage or objectness—may produce better segmentation signals but alters the classifier’s internal reasoning. This compromises the original notion of explanation fidelity and blurs the line between interpreting a model and engineering it to be explainable. We argue that the community must more clearly acknowledge this disconnect: *explanations that are good for human interpretation or model debugging are not necessarily suitable for supervision.*

In summary, using explanation maps as supervision born out of necessity (lack of labels) can:

- Limit segmentation quality due to incomplete or noisy seeds,
- Inherit and propagate classifier biases,
- Obscure performance evaluation during training, and
- Require heavy reliance on heuristics.

These pitfalls underscore the fragility of post-hoc maps when used as training signals. Explanations are valuable for auditing a classifier; they become brittle as supervision. Our recipe treats them as assistance with explicit quality checks, not as labels. Future directions may focus on either developing explainers that are both faithful and complete or shifting toward alternative weak supervision paradigms that better align with the task. The next section explores how to rethink evaluation and trust in this setting, and outlines potential directions for progress.

6. Interpretability vs. Utility: The Segmentation Gap

A key tension lies in the gap between interpretability—how well an explanation reflects model reasoning, and utility—its effectiveness as supervision for segmentation. An explanation may be faithful to the model yet fail to capture the full object, or conversely cover the object but not reflect true decision cues.

Most XAI methods are judged by fidelity or stability, but if used for segmentation, spatial accuracy becomes essential. Poor overlap with object regions signals either a flawed explainer or a faithful revelation of biased model reasoning. Thus, segmentation quality itself can serve as a diagnostic for explanation quality. Addressing this requires task-specific notions of faithfulness, emphasizing semantic and spatial alignment. Such alignment might call for concept-based explanations or constrained explanation spaces, where saliency is guided to align with high-level object concepts rather than arbitrary features. While promising, these approaches remain underexplored in the WSSS context.

Hybrid approaches. Saliency maps may be optimized jointly for interpretability and completeness, e.g., by regularizing attention maps with objectness priors or foundation model outputs. Such dual objectives trade some accuracy for more transparent and reliable supervision. This is a conscious design trade-off—restricting a model from using spurious shortcuts to gain interpretability and supervision value. While it may reduce raw accuracy, it promotes transparency and downstream reliability.

This leads to a key proposition: Visual explanations should not be used as training signals without validation. Confidence metrics, explainer fusion, or agreement with trusted oracles like SAM can help ensure only reliable cues guide segmentation.

Rethinking Faithfulness and Consistency. Bridging the interpretability–utility gap requires new criteria for explanation quality, such as:

- Fusion-based confidence (e.g., entropy or inter-method agreement),
- Robustness benchmarking (e.g., sensitivity to noise),
- Contextualized faithfulness, where explanations align with semantic object concepts, not only model predictions.

A practical framework is to treat saliency maps as pseudo-annotators: applying refinements (e.g., Grab-Cut [14], random walks [13]) and measuring whether they yield usable segmentations. Strong segmentations suggest meaningful signals, providing a task-driven benchmark of utility.

Implications of Foundation Models. Foundation models like SAM reduce reliance on saliency cues by pro-

ducing masks from minimal prompts, yet explaining their behavior remains open. Outputs vary by prompt and region, and new interpretation tools are needed. Progress will require collaboration between XAI and vision communities—developing evaluation pipelines that account for downstream utility and supervision strategies that respect explainability constraints. Such co-design is vital for models that are both accurate and interpretable.

7. Conclusion and Future Directions

We examined the role of post-hoc visual explanations in weakly supervised semantic segmentation (WSSS) and showed that common methods—Grad-CAM, Score-CAM, IG, LIME, SHAP—often lack the consistency, completeness, and alignment needed for reliable supervision. Using SAM2 as a proxy, we demonstrated that these maps frequently cover only partial object regions and disagree with each other and with the true object extent.

While fusion and foundation models can mitigate some shortcomings, explanation maps should not be assumed suitable for training without validation. Segmentation requires pixel-level accuracy, whereas explainers are designed primarily for interpretability. We therefore argue for principled use: incorporating quality checks, iterative refinement, or models that integrate interpretability and segmentation from the outset.

This critique raises key questions for the community:

- Can visual explanations serve beyond inspection and debugging, or should alternative weak cues (e.g., clicks, text, foundation masks) replace them?
- How can we improve faithfulness and stability of saliency maps? Could architectural constraints, multi-task learning, or dedicated loss functions produce more complete, stable, and human-aligned saliency maps?
- Should segmentation utility influence how explanation methods are ranked?
- What role will foundation models play, and how can their transparency be ensured?

Addressing these questions will require closer collaboration between the XAI, human-computer interaction and vision communities. While visual explanations have advanced WSSS, they must be employed with skepticism and proper validation. Emerging approaches—such as explainer fusion, entropy-based trust estimation, and SAM-guided refinement offer a promising starting point for future research. Until interpretable representations are native to models, explanation maps must be treated with caution—validated, selectively applied, and continually improved.

Future Research Directions: Future work will stress-test the framework in domains such as medical imaging and endoscopy, where distribution shifts challenge both the proxy oracle and the explainers. To bridge the interpretability–utility gap, we highlight five avenues:

1. *Benchmarking:* Standardize localization metrics (IoU, pointing game) across datasets like Pascal VOC and COCO.
2. *Multi-Explainer Fusion:* Replace heuristic fusion with neural or graph-based models that learn consensus maps.
3. *Trust Estimation:* Identify reliable regions via entropy, agreement, or robustness probes.
4. *Human-in-the-Loop:* Use lightweight user corrections with SAM to refine explanation-based supervision.
5. *Explaining FMs:* Develop tools to interpret segmentation-focused foundation models such as SAM. Understanding their internal decision-making will be crucial for safe and effective use in weakly supervised or semi-supervised pipelines.

In closing, explanations should support trust, not replace it. Overreliance on unvalidated maps risks both interpretability and performance. By ensuring explanations are faithful, consistent, and task-aligned, we can advance segmentation systems that are accurate and meaningfully transparent.

Acknowledgements

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

References

- [1] 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. *Advances in neural information processing systems*, 35:15784–15799, 2022. [2](#), [3](#)
- [2] Piotr Borycki, Magdalena Tredowicz, Szymon Janusz, Jacek Tabor, Przemysław Spurek, Arkadiusz Lewicki, and Łukasz Struski. Epic: Explanation of pretrained image classification networks via prototype. *arXiv preprint arXiv:2505.12897*, 2025. [1](#)
- [3] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 969–978, 2022. [1](#), [4](#)
- [4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. [2](#)
- [5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. [2](#)
- [6] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10762–10769, 2020. [1](#)
- [7] Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. *Advances in neural information processing systems*, 33:11996–12007, 2020. [4](#)
- [8] 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. [2](#), [3](#)
- [9] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462. Springer, 2020. [2](#)
- [10] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36:29914–29934, 2023. [5](#)
- [11] Hyeokjun Kweon and Kuk-Jin Yoon. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19499–19509, 2024. [1](#)
- [12] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019. [3](#)
- [13] Changyang Li, Yuchen Yuan, Weidong Cai, Yong Xia, and David Dagan Feng. Robust saliency detection via regularized random walks ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2710–2717, 2015. [7](#)
- [14] Yubing Li, Jinbo Zhang, Peng Gao, Liangcheng Jiang, and Ming Chen. Grab cut image segmentation based on image region. In *2018 IEEE 3rd international conference on image, vision and computing (ICIVC)*, pages 311–315. IEEE, 2018. [7](#)
- [15] Shijie Lian, Ziyi Zhang, Hua Li, Wenjie Li, Laurence Tianruo Yang, Sam Kwong, and Runmin Cong. Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. In *Forty-first International Conference on Machine Learning*, 2024. [2](#)
- [16] Taro Makino, Stanisław Jastrzebski, Witold Oleszkiewicz, Celin Chacko, Robin Ehrenpreis, Naziya Samreen, Chloe Chhor, Eric Kim, Jiyon Lee, Kristine Pysarenko, et al. Differences between human and machine perception in medical diagnosis. *Scientific reports*, 12(1):6877, 2022. [3](#)
- [17] Woo-Jeoung Nam, Jaesik Choi, and Seong-Whan Lee. Interpreting deep neural networks with relative sectional propagation by analyzing comparative gradients and hostile activations. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11604–11612, 2021. [1](#)
- [18] Rosa YG Paccotacya-Yanque, Alceu Bissoto, and Sandra Avila. Are explanations helpful? a comparative analysis of explainability methods in skin lesion classifiers. In *2024 20th International Symposium on Medical Information Processing and Analysis (SIPAIM)*, pages 1–5. IEEE, 2024. [3](#), [4](#)
- [19] Sukrut Rao, Moritz Böhle, and Bernt Schiele. Better understanding differences in attribution methods via systematic evaluations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4090–4101, 2024. [3](#)
- [20] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. [2](#), [5](#)
- [21] Ayush Somani, Ludwig Alexander Horsch, Ajit Bopadikar, and Dilip Kumar Prasad. Propagating transparency: A deep dive into the interpretability of neural networks. 2024. [1](#), [2](#)
- [22] Felix Tempel, Daniel Groos, Espen Alexander F Ihlen, Lars Adde, and Inga Strümke. Choose your explanation: A comparison of shap and gradcam in human activity recognition. *arXiv preprint arXiv:2412.16003*, 2024. [4](#)

- [23] Osman Tursun, Simon Denman, Sridha Sridharan, and Clinton Fookes. Sess: Saliency enhancing with scaling and sliding. In *European Conference on Computer Vision*, pages 318–333. Springer, 2022. [2](#)
- [24] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. [2](#)
- [25] Matthew Watson, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed. Agree to disagree: When deep learning models with identical architectures produce distinct explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 875–884, 2022. [3](#)
- [26] Junyi Wu, Weitai Kang, Hao Tang, Yuan Hong, and Yan Yan. On the faithfulness of vision transformer explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10936–10945, 2024. [1](#)
- [27] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–998, 2022. [1](#)
- [28] Xiaobo Yang and Xiaojin Gong. Foundation model assisted weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 523–532, 2024. [1](#), [3](#), [6](#)
- [29] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. [4](#)
- [30] Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. Finer-cam: Spotting the difference reveals finer details for visual explanation. *arXiv preprint arXiv:2501.11309*, 2025. [3](#)
- [31] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021. [4](#)
- [32] Xiaoyu Zhu, Jeffrey Chen, Xiangrui Zeng, Junwei Liang, Chengqi Li, Sinuo Liu, Sima Behpour, and Min Xu. Weakly supervised 3d semantic segmentation using cross-image consensus and inter-voxel affinity relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2834–2844, 2021. [1](#)