Causal Quantification of the Sensitivity-Reliability Trade-Off in Semantic XAI: Comparing Object-Aware (SAM) and Texture-Aware (SLIC) Segmentation

Abhay Bhandarkar[1,2]

abhaybhandarkar@gmail.com

[1]Undergraduate, Ramaiah Institute of Technology, Bengaluru, India
[2]Intern, IIT Hyderabad, India

Abstract

Explainable AI (XAI) methods aiming to probe model internals for scientific discovery ("RED XAI") must move beyond correlational saliency maps. We address this by presenting a systematic comparison of segmentation methods within a causal attribution framework. We contrast an object-aware approach using the Segment Anything Model (SAM) against a texture-aware baseline using SLIC superpixels. Both are integrated into a pipeline utilizing Grad-CAM for saliency, CLIP for concept labeling, and a causal validation step quantifying concept importance via counterfactual interventions (blur masking) measured by raw confidence drop. Evaluating on 200 ImageNet images, we uncover a critical sensitivity-reliability trade-off: SAM-based object-centric concepts show significantly higher average causal impact (81.0% mean confidence drop vs. 37.7% for SLIC), demonstrating greater sensitivity, but suffer from segmentation failures in 9.5% of cases (181/200 successes). SLIC achieves perfect 100% reliability (200/200 successes) and lower impact variance, albeit with reduced sensitivity. This trade-off provides actionable guidance for domain scientists: SLIC's robustness is preferable for high-stakes, texture-reliant tasks (e.g., medical diagnostics), while SAM's sensitivity may benefit exploratory analysis of object-centric phenomena. Our work offers quantitative evidence of this trade-off, enabling more informed XAI method selection for reliable scientific insight.

Introduction

Deep neural networks (DNNs) offer unprecedented capabilities for analyzing complex scientific data, yet their inherent opacity often impedes trust, hinders debugging, and limits their utility for generating new scientific knowledge (Rudin 2019). Within Explainable AI (XAI), the pursuit extends beyond user-facing justifications ("BLUE XAI") towards "RED XAI"—methods designed to rigorously probe, decompose, and understand the internal mechanisms of models to facilitate scientific insight.

Standard post-hoc explanation techniques, such as saliency mapping via Grad-CAM (Selvaraju et al. 2017), highlight input regions correlated with a model's output. While useful, these methods face limitations for RED XAI: they do not elucidate the semantic concepts perceived by the model within salient areas, nor do they establish the causal necessity of these features for the prediction, as correlations can be spurious (Adebayo et al. 2018). Ante-hoc interpretable architectures like Concept Bottleneck Models (CBMs) (Koh et al. 2020) address concept identification but require access to labeled concept data and model retraining, precluding their application to the vast array of existing pre-trained models.

To overcome these challenges in the post-hoc setting, we propose and systematically evaluate an automated, post-hoc pipeline for generating causally validated, concept-based explanations from any pretrained vision classifier. Central to such pipelines is the segmentation of salient image regions into meaningful units. We posit that the choice of segmentation algorithm reflecting implicit assumptions about feature granularity (e.g., objects vs. textures) critically influences the resulting explanation's quality and reliability.

This paper presents the first systematic comparison, grounded in causal validation, of two distinct segmentation paradigms for semantic XAI:

1. **Object-Aware Segmentation** via the Segment Anything Model (SAM) (Kirillov et al. 2023).
2. **Texture-Aware Segmentation** via SLIC superpixels (Achanta et al. 2012).

These are integrated with Grad-CAM saliency, CLIP-based concept labeling (Radford et al. 2021), and a crucial causal validation stage employing counterfactual interventions (blur masking) to quantify explanation quality via raw causal impact. Our experiments on ImageNet reveal a fundamental sensitivity-reliability trade-off: SAM identifies object-centric concepts with higher average causal impact but exhibits lower robustness compared to the perfectly reliable, albeit less sensitive, SLIC baseline. We articulate the significant implications of this trade-off for applying XAI in diverse scientific domains, offering quantitative guidance for practitioners.

Related Work

Recent advances in explainable artificial intelligence (XAI) integrate several foundational themes: saliency mapping, concept-based explanations, causal inference, segmentation, and their application to sciences.

Saliency Methods: Saliency approaches offer insight into model decision processes via visual attribution. Grad-CAM

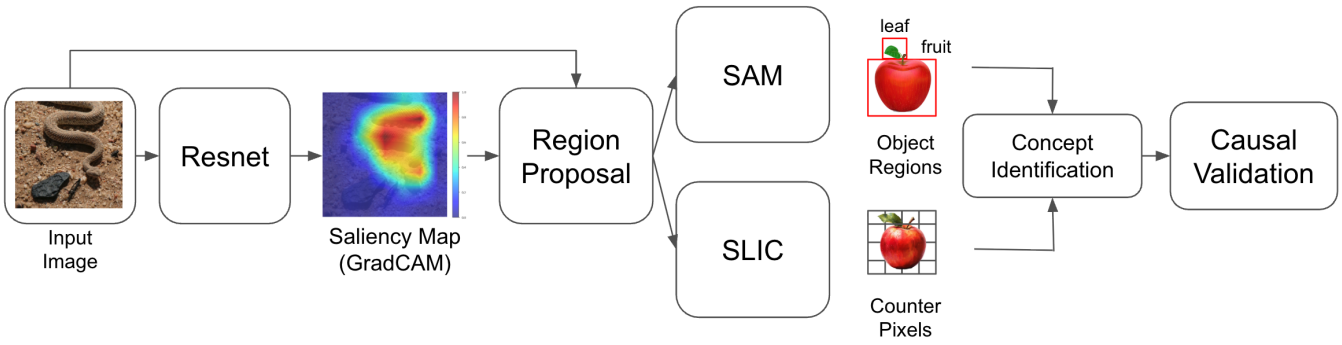

Figure 1: Overview of the 4-stage semantic attribution pipeline: (1) Saliency generation, (2) Segmentation into candidate regions (SAM vs. SLIC), (3) CLIP-based concept labeling, and (4) Causal validation via counterfactual intervention.

(Selvaraju et al. 2017), Grad-CAM++ (Chattopadhay et al. 2018), Integrated Gradients (Sundararajan, Taly, and Yan 2017), attention analysis with Transformers (Chefer, Gur, and Wolf 2021), and foundational attention visualization for RNNs are commonly used. However, their faithfulness is debated—as shown by rigorous sanity checks (Adebayo et al. 2018) and critiques of attention explanations (Jain and Wallace 2019). Expanded localization methods include counterfactual generation (Chang et al. 2019), feature attribution (Fong and Vedaldi 2017), and real-time saliency techniques.

Concept-based Explanations: Interpreting deep models by mapping representations to human-centric concepts is addressed with TCAV (Kim et al. 2018), Concept Bottleneck Models (CBMs) (Koh et al. 2020), concept-based explanations for CNNs (Yeh et al. 2019), and ACE (Ghorbani et al. 2019). Network Dissection quantifies interpretability in convolutional features (Bau et al. 2017). Advances in automated concept mapping harness vision-language models (VLMs) like CLIP (Radford et al. 2021); its use as a general-purpose concept extractor is highlighted by Kwon et al. (Kwon et al. 2022). Supervision-free approaches for biomedical and material domains build on datasets such as ImageNet (Deng et al. 2009) and HAM10000 (Tschandl, Rosendahl, and Kittler 2018).

Causal Inference in XAI: Correlation-based measures are limited in actionable trustworthiness. Causal inference via counterfactual and intervention-based reasoning, formalized by Pearl (Pearl 2009), underpins emerging machine learning explanations. Notable examples include intervention approaches (Goyal et al. 2019), critical reviews of counterfactual methods (Slack et al. 2021), the development of explainable reinforcement learning (Madumal et al. 2020), and GDPR-compliant transparency via counterfactuals (Wachter, Mittelstadt, and Russell 2017).

Segmentation for Explanations: Region-based explanations require robust segmentation, using superpixels such as SLIC (Achanta et al. 2012), and object-level methods as units of perturbation for LIME (Ribeiro, Singh, and Guestrin 2016). Deep models for similarity assessment (Chen et al. 2020), along with the Segment Anything Model (SAM) (Kirillov et al. 2023), have propelled the quality of neural segmentation. Reviews (Singla et al. 2023) compare SAM versus classical techniques for medical applications.

XAI for Science, Medicine, and Materials: Deploying XAI in high-stakes scientific contexts requires mechanistic and reliable explanations, as discussed for medicine (Holzinger et al. 2019; Tjoa and Guan 2020), climate science (Mamalakis et al. 2022), and materials design (Stanev et al. 2021; Zhang et al. 2021). Trust in machine learning is furthered by advocating interpretable models in critical settings (Rudin 2019). The sensitivity-reliability trade-off in model explanations continues to inform choice and deployment of XAI strategies.

Datasets and Tools: Large-scale datasets remain fundamental to benchmarking conceptual, causal, and region-based explanation methods, distinctly featuring ImageNet (Deng et al. 2009) and HAM10000 (Tschandl, Rosendahl, and Kittler 2018).

Stage 1: Saliency Map Generation

The first stage identifies regions of interest correlated with the model's prediction. For an input image $I \in \mathbb{R}^{H \times W \times 3}$, a classifier $f(\cdot)$, its top predicted class c, and the corresponding confidence $P(c|I)$, we generate a saliency map $S \in [0,1]^{H \times W}$.

We employ Grad-CAM (Selvaraju et al. 2017), a class-discriminative localization technique. Grad-CAM was chosen for its applicability to a wide range of CNN architectures without requiring architectural changes, making it ideal for a post-hoc pipeline. It uses the gradients of the predicted class c flowing into the final convolutional layer to produce a coarse localization map S. This map highlights regions that positively influence the prediction for class c. For our ResNet-50 backbone, we target the final block of `layer4`. The resulting map is upsampled to the input image size and normalized to $[0, 1]$.

Stage 2: Semantic Region Proposal

The raw, pixel-level saliency map S lacks semantic structure. This stage partitions the image into meaningful candidate regions $\{R_k\}$ which are then filtered based on saliency. We contrast two segmentation paradigms:

- **SAM:** This object-aware approach uses the Segment Anything Model (SAM) (Kirillov et al. 2023), a foundation model for segmentation. We first run SAM's automatic mask generator over I to produce a comprehensive set of candidate object and part masks $M = \{m_j\}$.
- **Baseline (SLIC):** This texture-aware baseline uses SLIC superpixels (Achanta et al. 2012), a classic algorithm that clusters pixels based on color and spatial proximity. It partitions I into a set of perceptually uniform superpixels $P = \{p_j\}$, which are ignorant of object boundaries and instead group similar textures.

For both methods, we filter the resulting candidate regions (masks m_j or superpixels p_j) using an identical process. We score each region by its mean saliency $\mathbb{E}_{p \in R_j}[S(p)]$. We retain only regions that meet two criteria:

1. **Saliency Threshold (τ_s):** The mean saliency must exceed τ_s (e.g., 0.3) to ensure the region is relevant to the prediction.
2. **Size Threshold (τ_{size}):** The pixel count $|R_j|$ must exceed τ_{size} (e.g., 500 pixels) to discard small, noisy regions.

From this filtered set, we select the top-K (e.g., $K = 5$) regions ranked by their mean saliency, yielding the final candidate regions $\{R_k\}$ and their corresponding bounding boxes $\{B_k\}$.

Stage 3: Concept Identification

This stage assigns a human-understandable text label C_k to each proposed region R_k. We leverage the zero-shot, open-vocabulary capabilities of CLIP (Radford et al. 2021).

First, for each region R_k, we extract the image patch defined by its bounding box B_k. To provide visual context, we expand B_k with a small margin (e.g., 20% of its size) before cropping from the original image I. This contextual patch is then passed through the CLIP image encoder E_{img} to produce a region embedding.

Separately, we pre-compute text embeddings for a fixed vocabulary V of ≈80 common concepts (e.g., "wheel", "fur texture", "sky"). Each concept $v \in V$ is formatted using a prompt, $E_{txt}(\text{prompt}(v))$, where prompt($v$) = "a photo of a {v}".

We then compute the cosine similarity between the region's image embedding and all pre-computed concept embeddings. The concept C_k with the highest similarity is assigned to the region R_k, as defined by the equation:

$$C_k = \arg\max_{v \in V} \frac{E_{img}(I(B_k)) \cdot E_{txt}(\text{prompt}(v))}{\|E_{img}(I(B_k))\| \|E_{txt}(\text{prompt}(v))\|}$$

Stage 4: Causal Validation

The saliency and CLIP similarity scores are correlational. This final, critical stage moves to causal inference by quantifying the necessity of the identified concept-region pair (C_k, R_k) for the model's prediction. We ask: "Does the model's confidence drop if we remove this concept?"

To answer this, we perform a counterfactual intervention. For each top-ranked region B_k, we "remove" its features by applying a strong Gaussian blur (e.g., 51x51 kernel) within the bounding box B_k. This intervention destroys high-frequency features (like texture and edges) associated with the concept while preserving the low-frequency color and structure, creating a neutral counterfactual image $I'_k = \text{Blur}(I, B_k)$.

We then feed this counterfactual image I'_k back into the classifier $f(\cdot)$ and measure the new confidence $P(c|I'_k)$ for the original predicted class c.

We define the raw causal impact I_k as the fraction of original confidence lost due to the intervention. This metric normalizes the confidence drop by the initial confidence, making it comparable across images:

$$I_k = \max\left(0, 1 - \frac{P(c|I'_k)}{P(c|I)}\right)$$

A high I_k (e.g., 0.9 or 90%) indicates a catastrophic confidence drop, implying the concept was causally necessary for the prediction. A low I_k (e.g., 0.0) implies the region was merely correlational, and the model used other features to maintain its confidence. Finally, all identified concepts are re-ranked by this raw causal impact I_k to produce the final explanation.

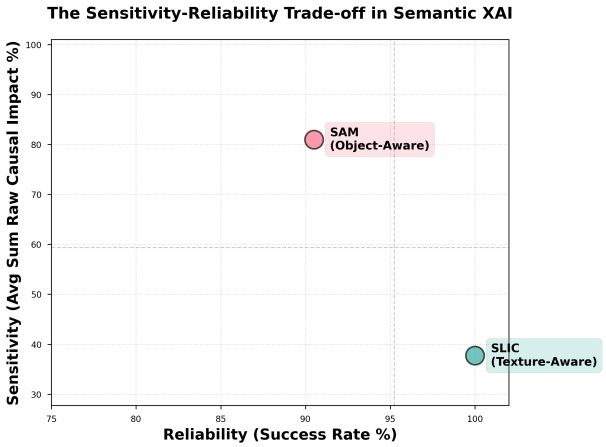

Figure 2: The Sensitivity-Reliability Trade-off. SAM achieves higher average causal impact (Sensitivity) but lower success rate (Reliability) compared to SLIC on 200 ImageNet images.

Experiments and Results

Experimental Setup

We used a pre-trained ResNet-50 from torchvision. Grad-CAM targeted the final block of layer4. SAM utilized the vit_h checkpoint (Kirillov et al. 2023). SLIC parameters were n_segments=100, compactness=20. CLIP used ViT-L/14 (OpenAI weights) (Radford et al. 2021) with a vocabulary of ≈80 terms. Blur intervention used a 51x51 kernel (σ=25). Qualitative analysis used one image; quantitative results are averaged over 200 ImageNet validation images streamed via Hugging Face datasets (Deng et al. 2009). All runs used an NVIDIA V100 GPU.

Qualitative Ablation: SAM vs. SLIC

Figure 3 provides a visual comparison on a 'beagle' image. Table 1 shows the top concepts and impacts. SAM identifies the 'collar' as a distinct region, whereas SLIC focuses on 'fur texture' patches. Notably, for this image, the texture patches identified by SLIC yielded a significantly higher total causal impact (60.5%) than the object-centric regions found by SAM (3.3%), suggesting the model may rely more on texture than object parts here. This highlights how causal validation reveals model strategy, independent of region semantics.

Table 1: Top concepts and raw causal impacts for the beagle image (Figure 3).

Pipeline	Top Concept	Causal Impact (Raw %)
SAM		
1.	fur texture	3.3%
2.	fur	0.0%
3.	collar	0.0%
Sum of Raw Impacts		**3.3%**
SLIC		
1.	fur texture	24.9%
2.	fur	24.0%
3.	fur texture	8.9%
Sum of Raw Impacts		**60.5%**

Quantitative Evaluation on ImageNet (N=200)

Table 2 summarizes the performance across 200 ImageNet images, focusing on the average sum of raw causal impacts and pipeline success rate. The SAM pipeline demonstrates significantly higher sensitivity, achieving an average impact sum of 81.0% compared to SLIC's 37.7%. This confirms that SAM's object-centric regions generally correspond to features more critical for the classifier's decisions. However, this sensitivity comes at the cost of reliability: SAM failed on 19 images (90.5% success rate), primarily due to finding no regions above the saliency threshold, whereas SLIC succeeded on all 200 images (100% reliability). Figure 2 visually represents this trade-off.

Discussion: Implications for XAI4Science

The quantified sensitivity-reliability trade-off (Figure 2, Table 2) provides critical guidance for scientists applying XAI. The choice between SAM and SLIC depends on the scientific goal and domain characteristics:

- **High-Stakes / Reliability-Critical Domains:** In fields like medical diagnosis, where consistent explanations are crucial and failures unacceptable, SLIC's 100% reliability and lower variance are advantageous. This is particularly relevant if diagnostic features are texture-based (e.g., skin lesion analysis (Tschandl, Rosendahl, and Kittler 2018), identifying interstitial lung patterns).
- **Exploratory / Sensitivity-Prioritized Domains:** For research exploring dominant model features or identifying object-centric phenomena (e.g., specific storm cells in climate data (Mamalakis et al. 2022), localizing defects in materials (Stanev et al. 2021; Zhang et al. 2021)), SAM's ability to isolate high-impact concepts may be preferred, provided occasional explanation failures can be tolerated or addressed through parameter tuning.
- **Hybrid Approaches:** Domains involving both object and texture features might benefit from running both pipelines or developing adaptive methods.

Our causal validation framework enables this quantitative comparison, moving RED XAI towards more principled methodological choices tailored to scientific needs.

Limitations and Future Work

While our evaluation on ImageNet ($N = 200$) successfully isolates the theoretical sensitivity-reliability trade-off, validating these findings in high-stakes scientific domains is the critical next step. Future work will extend this pipeline to specialized medical datasets, specifically HAM10000 (dermatoscopy) and RadImageNet, to confirm if SLIC's reliability advantage holds in clinical workflows. Additionally, relying on a predefined concept vocabulary limits discovery; we aim to integrate Large Multimodal Models (LMMs) for automated, open-ended concept generation to reduce human bias. Finally, while Gaussian blurring provides a first-order causal approximation, it introduces out-of-distribution artifacts. We propose replacing blurring with generative inpainting to maintain natural image statistics during causal auditing, alongside adaptive thresholding to mitigate SAM's failure modes.

Conclusion

We presented a systematic, causal evaluation comparing object-aware (SAM) and texture-aware (SLIC) segmentation for post-hoc semantic XAI. Our key contribution is the quantification of a sensitivity-reliability trade-off: SAM yields explanations with higher average causal impact but lower robustness than SLIC. This finding, enabled by our causal validation pipeline, provides crucial, actionable guidance for scientists selecting XAI methods. By highlighting the non-neutrality of segmentation choice and its domain-specific implications, this work advances the practice of RED XAI for more reliable scientific discovery from deep learning models.

Acknowledgements

The author would like to thank IIT Hyderabad for the computational resources provided for the research and development.

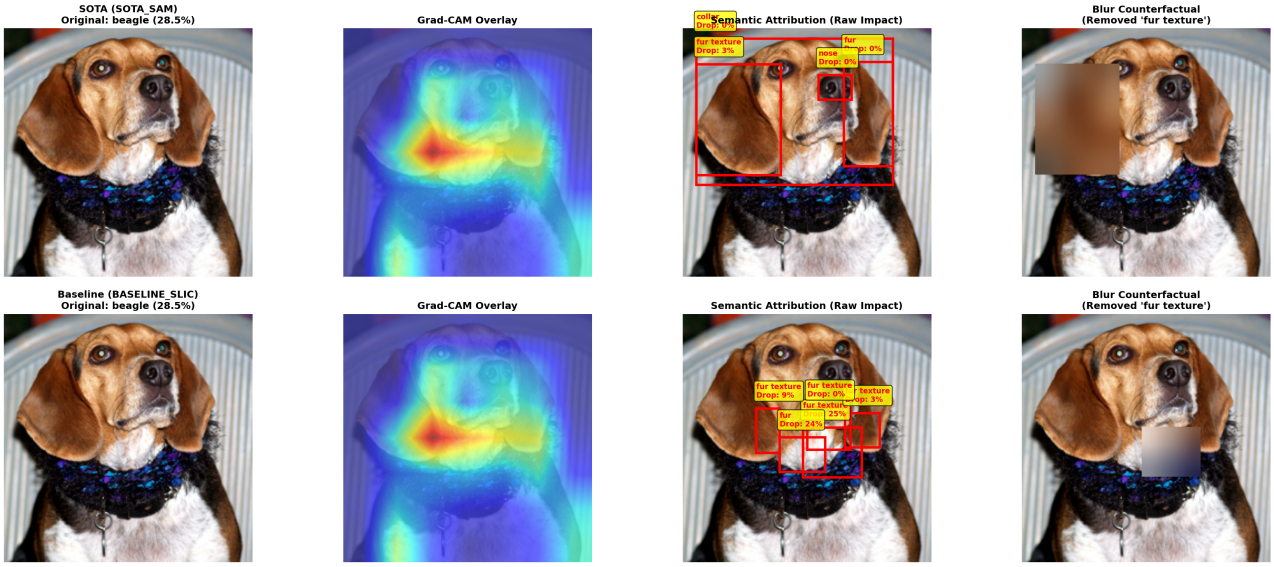

Figure 3: Qualitative comparison: (SAM, top row) vs. Baseline (SLIC, bottom row) on a 'beagle' image. Panels: (a) Original, (b) GradCAM, (c) Semantic Attribution (Raw Causal Impact %), (d) Blur Counterfactual.

Table 2: Quantitative comparison: Sensitivity vs. Reliability on 200 ImageNet images. Results show Mean ± Standard Deviation for the sum of raw causal impacts per image. N indicates the number of images where concepts with non-zero impact were found.

Method	**Avg. Sum Raw Causal Impact (%)**	**Success Rate (%)**	**Std Dev (%)**
SAM (Object-Aware)	**81.0** ± 57.6 (N=163)	181/200 (90%)	57.6
SLIC (Texture-Aware)	37.7 ± 34.9 (N=191)	**200/200 (100%)**	**34.9**
Note: SAM shows 115% higher sensitivity but 10% failure rate and higher variance.			

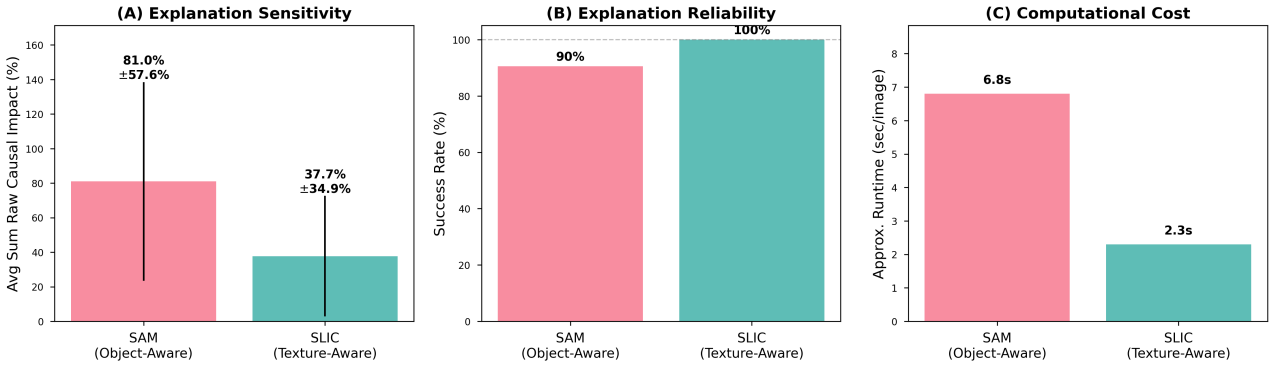

Figure 4: Comparative analysis of SAM (Object-Aware) and SLIC (Texture-Aware) pipelines on 200 ImageNet images, highlighting (A) Explanation Sensitivity (Avg. Sum Raw Causal Impact), (B) Explanation Reliability (Success Rate), and (C) Approximate Computational Cost per image.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11): 2274–2282.

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *Advances in neural information processing systems*, volume 31.

Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6541–6549.

Chang, C.-H.; Creager, E.; Goldenberg, A.; and Duvenaud, D.

2019. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*.

Chattopadhay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847.

Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 782–791.

Chen, L.; Wu, Z.; Milan, A.; Tommasi, T.; and Hauptmann, A. 2020. Looks like THIS: Deep learning for betraying appearance similarity. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 234–243.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Fong, R. C.; and Vedaldi, A. 2017. Interpreting deep neural networks with feature attribution: A comprehensive guide. volume 41, 2484–2497. IEEE.

Ghorbani, A.; Wexler, J.; Zou, J.; and Kim, B. 2019. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 32.

Goyal, Y.; Wu, Z.; Ernst, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Explaining classifiers: A causal-intervention approach. In *International Conference on Learning Representations*.

Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; and Müller, H. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4): e1312.

Jain, S.; and Wallace, B. C. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; and Sayres, R. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning*, 2668–2677. PMLR.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3879–3890.

Koh, P. W.; D'Amour, A.; Pierson, E.; Kim, B.; Taylor, J.; and Re, C. 2020. Concept bottleneck models. In *International conference on machine learning*, 5437–5447. PMLR.

Kwon, H.; Na, B.-S.; Chang, H.-I.; Kim, D.-G.; and Paik, J. 2022. CLIP as a concept extractor. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6715–6719.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 2493–2500.

Mamalakis, A.; B-H, E.; Yu, J.; Randerson, J. T.; and Pritchard, M. S. 2022. Explainable artificial intelligence (XAI) for climate science. *Artificial Intelligence for the Earth Systems*, 1(1): 2–1.

Pearl, J. 2009. *Causality*. Cambridge university press.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?": Explaining the predictions of any classifier.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Singla, T.; Saha, S.; Goyal, V.; and Batra, T. 2023. Understanding segment anything model: A review and analysis.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2021. Counterfactual explanations for deep learning: A critical review. In *Advances in Neural Information Processing Systems*, volume 34, 27163–27175.

Stanev, V.; Kusne, A. G.; Oses, C.; and Curtarolo, S. 2021. Explainable artificial intelligence for materials science. *Nature communications*, 12(1): 1–7.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Tjoa, E.; and Guan, C. 2020. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE transactions on neural networks and learning systems*, 32(11): 4793–4813.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL Tech.*, 31: 841.

Yeh, C.-K.; Kim, B.; Damour, A.; Tsipras, D.; Ilyas, A.; and Madry, A. 2019. Concept based explanations for CNNs. In *International Conference on Machine Learning*, 7063–7072. PMLR.

Zhang, Y.; Wang, C.; Gan, Z.; Shen, Y.; Wang, Z.; and Li, X. 2021. Explainable deep learning for material design. 3865–3873.

Appendix

SAM Failure Cases (N=19)

The SAM pipeline failed to produce explanations for 19 out of 200 ImageNet images (9.5% failure rate). In all recorded cases, the error indicated "No high-saliency regions found." This occurs when no SAM-generated mask meets both the average saliency threshold ($\tau_s = 0.3$) and the minimum size threshold ($\tau_{size} = 500$ pixels). Visual inspection of some failure cases suggests images with diffuse saliency maps or where salient objects were smaller than the size threshold. Further analysis or adaptive thresholding may mitigate these failures. Image IDs 8, 10, 11, 12, 25, 40, 48, 60, 73, 82, 105, 112, 119, 139, 143, 145, 161, 168, 182 are the ones that failed to get semantic attributions through the SAM pipeline.

Selected Visual Results on ImageNet (N=200)

Figures 5 through 6 present selected 4-panel visualizations (Original, Saliency, Attribution) produced by the SAM pipeline on diverse images from the ImageNet validation set run. These illustrate the method's behavior across various object classes and complexities.

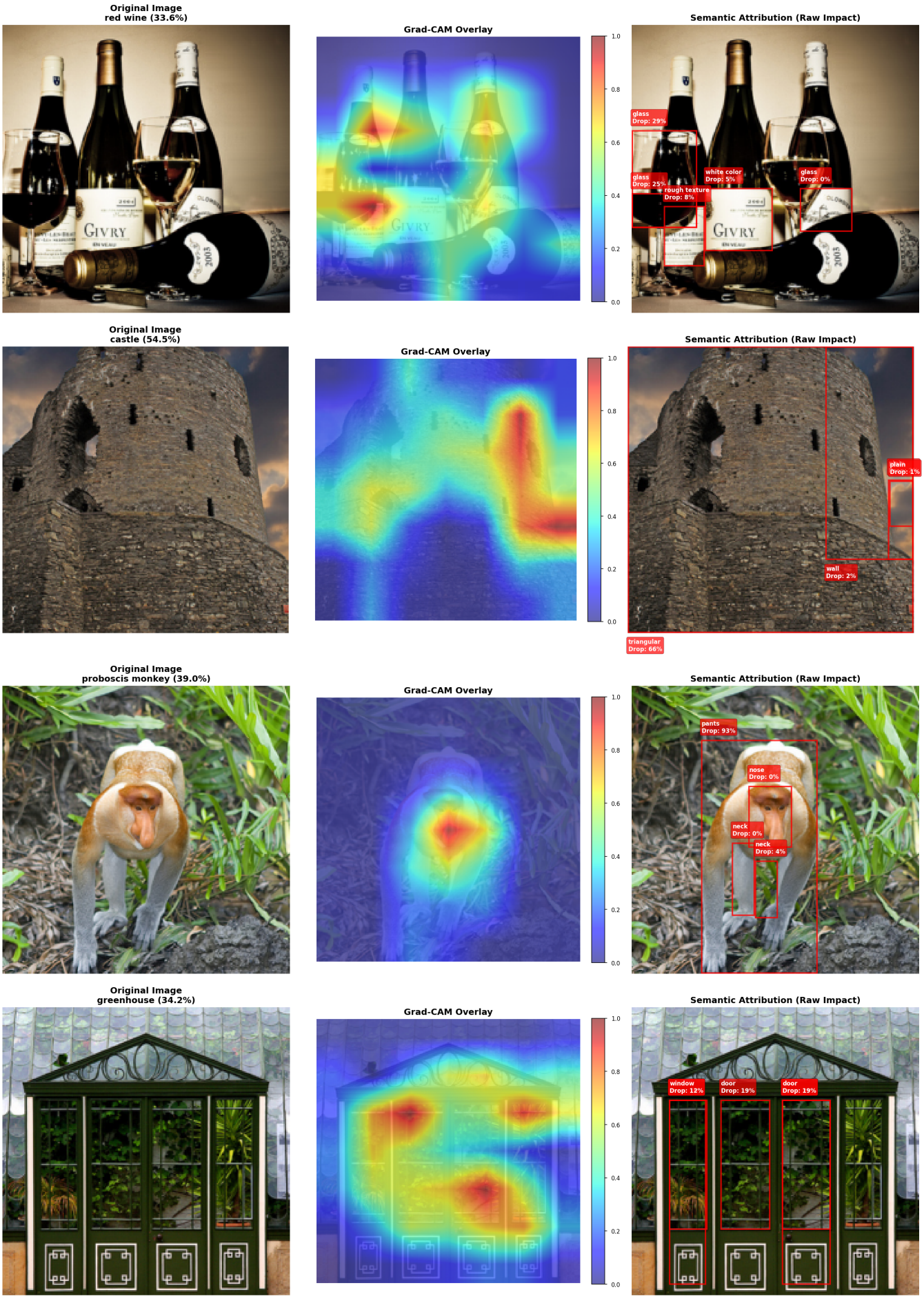

Figure 5: This figure shows the results of SAM pipeline results on ImageNet images

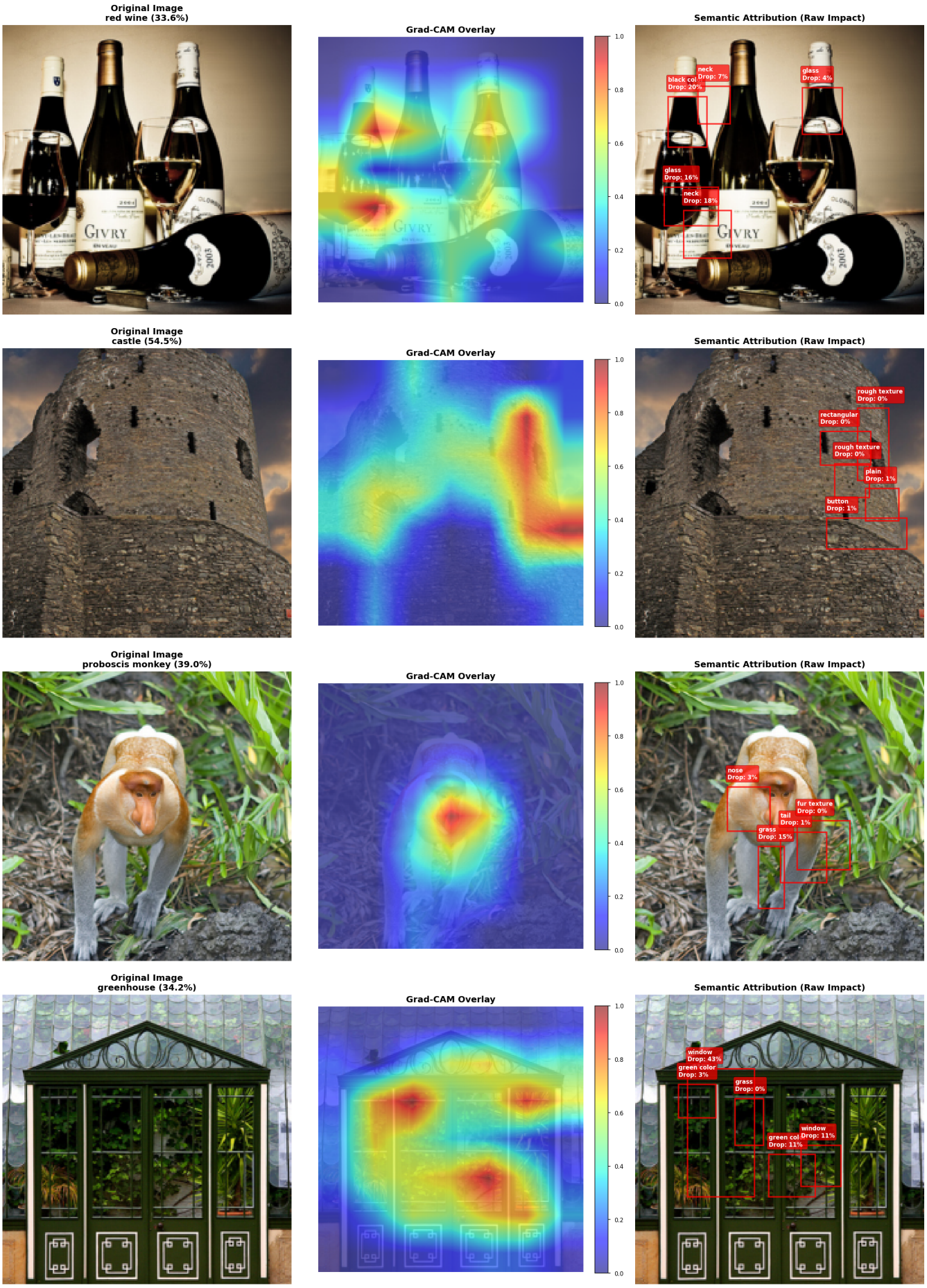

Figure 6: This figure shows the results of SLIC pipeline results on ImageNet images