TABLE 1

To compare how well different explainability methods (e.g., Grad-CAM, LIME, SHAP, Occlusion) highlight the influence of secondary objects in an image.

Method	Avg IoU w/ Sec. Mask ↑	Avg Confidence Drop ↑	Avg Attribution Score ↑
Grad-CAM	0.102	0.218	28.4%
IntegratedGrad	0.237	0.274	30.7%
GuidedBP	0.114	0.141	25.1%
RISE	0.143	0.274	35.9%

- Confidence drop and attribution score tend to correlate, suggesting that when a method assigns higher importance to an object, masking it results in a more significant prediction change. This reflects stronger alignment with model decision-making.
- IoU values remain relatively low across methods, highlighting that while explainers detect important regions, their spatial precision in aligning with object masks remains a challenge in multi-object scenarios.
- Some methods show high attribution and confidence change but low IoU, implying they capture the influence of secondary objects but not their exact location.
- Conversely, other methods achieve higher IoU but may underperform in attribution or fidelity, indicating better localization but weaker influence capture.

TABLE 2

To quantify how much the prediction of the top-1 class is affected when specific secondary objects are removed from the image.

=== Top 3 Results (Max Δ Conf) per Explainer === | Explainer | Objects | Full Conf | Masked Conf | Δ Conf | |—|—|—|—|—| | Grad-CAM | marine_mammal, coral, reptile | 0.66, 0.62, 0.73 | 0.07, 0.08, 0.28 | 0.59, 0.54, 0.45 | | IntegratedGrad | bird, coral, marine_mammal | 0.46, 0.57, 0.62 | 0.02, 0.14, 0.25 | 0.45, 0.43, 0.37 | | GuidedBP | coral, marine_mammal, invertebrates | 0.65, 0.67, 0.60 | 0.12, 0.28, 0.24 | 0.53, 0.39, 0.36 | | RISE | coral, marine_mammal, invertebrates | 0.65, 0.67, 0.59 | 0.01, 0.18, 0.14 | 0.64, 0.49, 0.45 |

=== Bottom 3 Results (Least Δ Confidence) per Explainer === | Explainer | Objects | Full Conf | Masked Conf | Δ Conf | |—|—|—|—|—| | Grad-CAM | anemone, equipment, plant | 0.63, 0.72, 0.63 | 0.58, 0.62, 0.47 | 0.05, 0.09, 0.16 | | IntegratedGrad | anemone, equipment, plant | 0.62, 0.74, 0.60 | 0.57, 0.67, 0.50 | 0.05, 0.08, 0.11 | | GuidedBP | anemone, equipment, plant | 0.62, 0.74,

0.59 | 0.53, 0.65, 0.47 | 0.09, 0.10, 0.12 | | RISE | anemone, equipment, plant | 0.63, 0.75, 0.60 | 0.59, 0.65, 0.46 | 0.04, 0.10, 0.14 |

Key Observations

- High Δ Confidence values indicate that the model's prediction is significantly influenced by the presence of certain secondary objects. This is commonly seen for objects like coral, marine_mammal, and reptile across multiple explainers.
- RISE and Integrated Gradients show the largest Δ Confidence values, suggesting that they are more sensitive to context features, and possibly better at surfacing dependencies on secondary objects.
- Low Δ Confidence values appear across all explainers for objects like anemone, equipment, and plant. These may represent cases where:
- The secondary object is not strongly correlated with the predicted class.
- The image is dominated by a single, clear primary object, reducing dependency on contextual cues.

Variability Across Scenes

- Higher variability in Δ Confidence is observed in multi-object scenes, where several secondary features jointly influence the prediction. In such cases, removing one object can disrupt contextual integrity, leading to a noticeable drop.
- Lower variability and smaller confidence drops occur in images with dominant primary objects and minimal background noise – the model relies less on secondary context in these scenarios.

Conclusion

- Fidelity-based evaluation reveals the contextual sensitivity of the model. Methods like RISE and Integrated Gradients tend to uncover strong dependencies on secondary features, which may help in identifying spurious correlations or overfitting to background elements. On the other hand, minimal Δ Confidence indicates either true model robustness or irrelevance of the masked region.

TABLE 3

To evaluate whether the explanation method not only highlights the correct region (via IoU), but also ranks secondary objects appropriately in terms of importance.

Most Significant Examples per Explainer (4 each) Showing objects with most extreme explanation weights

Explainer	Secondary Obj	Explanation Weight (%)	Sensitivity	Z-Score	RelToExplainer
IntegratedGrad	water	79.6%	0.13	2.3	+46.9%
IntegratedGrad	object	72.8%	0.11	2.0	+40.1%
IntegratedGrad	coral_reef	66.6%	0.17	1.7	+33.9%
IntegratedGrad	litter	4.8%	1.19	-1.3	-27.9%
GuidedBP	water	73.2%	0.13	2.0	+41.1%
GuidedBP	object	67.6%	0.11	1.7	+35.5%
GuidedBP	coral_reef	61.5%	0.17	1.4	+29.4%
GuidedBP	spines	57.0%	0.14	1.2	+24.9%
RISE	water	83.6%	0.13	2.5	+53.5%
RISE	object	75.8%	0.11	2.1	+45.7%
RISE	coral_reef	72.8%	0.17	2.0	+42.7%
RISE	litter	2.1%	1.19	-1.4	-28.0%
Grad-CAM	water	83.6%	0.13	2.5	+53.5%
Grad-CAM	object	75.8%	0.11	2.1	+45.7%
Grad-CAM	coral_reef	72.8%	0.17	2.0	+42.7%
Grad-CAM	litter	2.1%	1.19	-1.4	-28.0%

Key Statistics: - High sensitivity objects vary significantly across explainers - Z-score shows how extreme each weight is compared to all examples - RelToExplainer shows if weight is above/below explainer's average

Key Observations

- High-weight objects such as water, object, and coral_reef consistently receive elevated attribution scores across all methods, indicating strong visual or contextual relevance in the classifier's decision-making.
- Z-scores above 2.0 and high RelToExplainer values (e.g., +40–50%) confirm that these objects are exceptionally important, standing well above the explainer's average attribution for all objects.
- On the other end, objects like litter consistently show very low explanation weights (as low as 2.1%), negative Z-scores, and negative RelToExplainer percentages. These are likely considered irrelevant or even distracting by the model.
- The alignment of high-importance objects across all methods (e.g., water, object, coral_reef) suggests strong agreement on the core influential regions, reinforcing trust in those explanations.
- Sensitivity scores remain relatively low even for high-attribution objects, suggesting that explainers may capture relevance more sharply than actual output variation does – highlighting the importance of combining attribution with model fidelity tests.

Conclusion

- This table confirms that explainability methods are capable of ranking secondary objects meaningfully and consistently, especially in visually or

semantically dominant cases. The use of Z-scores and relative attribution deltas helps distinguish true signal from attribution noise, and supports deeper reliability analysis of saliency-based explainers.

TABLE 4

Unexpected Overdependence

Class Name	Original Confidence	Object Name	Masked Confidence Drop
leatherback_turtle	0.984	terrain	0.982
goldfish	0.909	water	0.933
isopod	0.894	object	0.890
sea_anemone	0.843	object	0.799
grey_whale	0.824	water	0.797
red-breasted_merganser	0.767	water	0.767
cabbage_butterfly	0.747	terrain	0.745
dugong	0.783	object	0.732
sea_snake	0.784	object	0.699
hermit_crab	0.834	fish	0.687
killer_whale	0.684	vehicle	0.681
loggerhead	0.840	object	0.666
chiton	0.948	object	0.627
stingray	0.735	terrain	0.579
sea_slug	0.863	fish	0.569
sturgeon	0.689	terrain	0.506
scuba_diver	0.864	light	0.406
lionfish	0.969	plant	0.386
tiger_shark	0.757	coral	0.314
rock_beauty	0.842	background	0.280
snorkel	0.915	terrain	0.273
electric_ray	0.779	terrain	0.269

Expected Overdependence

Class Name	Original Confidence	Object Name	Masked Confidence Drop
sea_urchin	0.946	invertebrate	0.900
sea_cucumber	0.862	invertebrate	0.808
jellyfish	0.910	animal	0.720
bottlecap	0.653	object	0.652
eel	0.815	fish	0.641

Class Name	Original Confidence	Object Name	Masked Confidence Drop
spiny_lobster	0.989	invertebrate	0.622
brain_coral	0.821	coral	0.604
anemone_fish	0.833	fish	0.555
coral_reef	0.857	coral	0.510
tench	0.866	coral	0.491
plastic_bag	0.730	object	0.441
starfish	0.865	coral	0.225

Independence

Class Name	Original Confidence	Object Name	Masked Confidence Drop
hammerhead	0.877	background	0.181
axolotl	0.934	terrain	0.171
puffer	0.924	terrain	0.138
king_crab	0.831	terrain	0.124
barracouta	0.605	coral	0.101
gar	0.917	object	0.021
terrapin	0.729	plant	0.017

Category Definitions

To assess how the model's predictions are influenced by secondary objects in the image, we categorize each instance based on the **semantic relationship** between the predicted class and the masked object, along with the **change in confidence** upon masking.

- **Unexpected Overdependence**: These are instances where the masked object is **semantically disjoint** from the predicted class (e.g., `terrain` affecting `leatherback_turtle`), yet its removal results in a **high confidence drop**. This indicates the model is **relying on irrelevant contextual cues**, potentially due to spurious correlations learned during training.
- **Expected Overdependence**: These represent cases where the masked object is **semantically related** to the predicted class (e.g., `coral` for `coral_reef`, `invertebrate` for `sea_urchin`), and the confidence drop upon masking is accordingly **high**. Such dependencies are considered **valid and expected**, as they reflect meaningful visual context.
- **Expected Independence**: In these cases, the masked object is **unrelated** to the predicted class, and masking leads to a **low change in**

confidence. This indicates the model is **correctly ignoring** irrelevant regions, which is desirable behavior.

- **Unexpected Independence**: These are cases where the masked object is **semantically similar or related** to the predicted class, yet its removal results in **negligible confidence change**. While this could imply model robustness, it may also indicate **under-attribution** or **explainer failure**. However, such cases were minimal in this study.

Distribution of Observed Behavior

Across the evaluated examples:

- **53%** of the cases fall under **unexpected overdependence**, indicating that in a significant portion of instances, the model relies heavily on secondary objects that are not semantically relevant to the prediction.
- **27%** of the cases correspond to **expected overdependence**, reflecting proper contextual reliance.
- **18%** exhibit **expected independence**, where the model demonstrates resilience to irrelevant contextual features.
- Only ~**2%** fall into the category of **unexpected independence**.

-> PIE CHART (53, 27, 18, 2), we can add this too

This low occurrence of unexpected independence is likely due to the fact that the **object masks were generated using SAM2**, and the classification labels were obtained from **top-1 predictions of a pretrained ResNet50 model**, which generally maintains semantic consistency. Furthermore, since the **saliency maps (from Grad-CAM, LIME, etc.) guided heatmap generation**, it is expected that few objects with high semantic relevance would be ignored. The small proportion of unexpected independence cases may be attributed to instances where **SAM2 masks captured spurious or background regions** that overlapped with **salient activations**, despite being semantically unrelated to the predicted class.

TABLE 5

This table is simply raw data to calculate variation in confidence values:

(NOT REQUIRED, but adding it just in case)

Class Name	Original Confidence	Top 4 Positive Influences	Top 3 Negative Influences
conch	0.997	object: 0.791	water: -0.002, terrain: -0.001

Class Name	Original Confidence	Top 4 Positive Influences	Top 3 Negative Influences
spiny_lobster	0.989	coral: 0.988, invertebrate: 0.622, terrain: 0.043, environment: 0.043	plant: -0.003
leatherback_turtle	0.984	reptile: 0.984, terrain: 0.982	plant: -0.000
whistle	0.971	vehicle: 0.970, toy: 0.970	object: -0.005
lionfish	0.969	fish: 0.607, plant: 0.386, coral: 0.310	terrain: -0.000
chiton	0.948	terrain: 0.948, object: 0.627	None
sea_urchin	0.946	plant: 0.939, invertebrate: 0.900, structure: 0.871, water: 0.730	fish: -0.005
axolotl	0.934	fish: 0.837, terrain: 0.171, plant: 0.064	None
puffer	0.924	fish: 0.916, terrain: 0.138, water: 0.002	background: -0.054, coral: -0.026
gar	0.917	fish: 0.860, object: 0.021, coral: 0.013	None
snorkel	0.915	human: 0.648, terrain: 0.273, reptile: 0.246, water: 0.132	fish: -0.028
jellyfish	0.910	light: 0.967, animal: 0.720, invertebrate: 0.701, object: 0.624	water: -0.057, fish: -0.000
goldfish	0.909	invertebrate: 0.968, water: 0.933, fish: 0.870, terrain: 0.637	plant: -0.132
isopod	0.894	material: 0.891, object: 0.890, background: 0.209	None
hammerhead	0.877	fish: 0.355, background: 0.181, water: 0.119, coral: 0.029	None
tench	0.866	fish: 0.863, coral: 0.491, plant: 0.491	None
starfish	0.865	invertebrate: 0.519, coral: 0.225, plant: 0.192, terrain: 0.177	object: -0.116

Class Name	Original Confidence	Top 4 Positive Influences	Top 3 Negative Influences
scuba_diver	0.864	background: 0.406, light: 0.406, human: 0.335, environment: 0.256	reptile: -0.074, terrain: -0.053, fish: -0.044
sea_slug	0.863	invertebrate: 0.764, fish: 0.569, object: 0.514, spines: 0.514	None
sea_cucumber	0.862	debris: 0.951, invertebrate: 0.808, spines: 0.775, fish: 0.657	animal: -0.098
coral_reef	0.857	environment: 0.921, coral: 0.510, water: 0.500, terrain: 0.485	material: -0.098, human: -0.097, bubble: -0.048
sea_anemone	0.843	animal: 0.815, object: 0.799, coral: 0.744, bubble: 0.683	fish: -0.027
rock_beauty	0.842	fish: 0.485, background: 0.280, terrain: 0.266, coral: 0.185	None
wreck	0.842	debris: 0.938, environment: 0.810, vehicle: 0.760, structure: 0.718	fish: -0.092, human: -0.020, equipment: -0.009
loggerhead	0.840	reptile: 0.791, object: 0.666, invertebrate: 0.348, light: 0.327	human: -0.008
hermit_crab	0.834	object: 0.958, fish: 0.687, terrain: 0.664, invertebrate: 0.413	coral: -0.078, animal: -0.028
anemone_fish	0.833	bubble: 0.992, fish: 0.555, coral: 0.538, plant: 0.115	None
king_crab	0.831	invertebrate: 0.694, terrain: 0.124, plant: 0.001	None
grey_whale	0.824	marine_mammal: 0.857, water: 0.797, object: 0.769, material: 0.031	None

Class Name	Original Confidence	Top 4 Positive Influences	Top 3 Negative Influences
brain_coral	0.821	bubble: 0.635, coral: 0.604, plant: 0.430, terrain: 0.396	fish: -0.030
eel	0.815	environment: 0.754, fish: 0.641, plant: 0.394, coral: 0.215	None
flatworm	0.815	fish: 0.802, invertebrate: 0.750, terrain: 0.315, coral: 0.304	plant: -0.022
chambered_nautilus	0.804	invertebrate: 0.899, fish: 0.610, coral: 0.259	plant: -0.053, terrain: -0.044
sea_snake	0.784	reptile: 0.934, object: 0.699, terrain: 0.621, fish: 0.527	None
dugong	0.783	shape: 0.771, debris: 0.732, marine_mammal: 0.628, fish: 0.434	environment: -0.214, coral: -0.152, terrain: -0.007
electric_ray	0.779	debris: 0.618, terrain: 0.269, invertebrate: 0.226, fish: 0.146	animal: -0.121, object: -0.080
red-breasted_merganser	0.767	equipment: 0.767, water: 0.767, terrain: 0.766	None
basketball	0.762	fish: 0.761, coral: 0.508	None
tiger_shark	0.757	water: 0.408, coral: 0.314, fish: 0.245, terrain: 0.087	bubble: -0.040
king_penguin	0.755	bird: 0.749	water: -0.045
cabbage_butterfly	0.747	object: 0.745, terrain: 0.745, background: 0.696	None
stingray	0.735	water: 0.610, terrain: 0.579, fish: 0.406, coral: 0.035	None
plastic_bag	0.730	terrain: 0.725, object: 0.441, plant: 0.349, background: 0.188	invertebrate: -0.141
terrapin	0.729	reptile: 0.727, plant: 0.017	None

Class Name	Original Confidence	Top 4 Positive Influences	Top 3 Negative Influences
sea_lion	0.711	marine_mammal: 0.706	terrain: -0.095, water: -0.095
sturgeon	0.689	fish: 0.681, terrain: 0.506	coral: -0.080
fiddler_crab	0.688	invertebrate: 0.688	coral: -0.214
killer_whale	0.684	equipment: 0.681, vehicle: 0.681, water: 0.253	None
feather_boa	0.678	coral: 0.677	None
golf_ball	0.665	terrain: 0.663, water: 0.663, reptile: 0.147	None
airship	0.662	equipment: 0.662	terrain: -0.187, coral: -0.140
candle	0.656	debris: 0.656	plant: -0.066
bottlecap	0.653	object: 0.652, debris: 0.652, terrain: 0.645	None
honeycomb	0.650	terrain: 0.667, fish: 0.449, material: 0.237, coral: 0.146	None
barracouta	0.605	fish: 0.550, coral: 0.101	terrain: -0.103