A. Implementation Details

A.1. Attribution Methods

LRP (Bach et al., 2015) Following (Montavon et al., 2017; Arias-Duart et al., 2021; Rao et al., 2022), we use the configuration that applies the ϵ-rule to the fully connected layers in the network, with $\epsilon = 0.25$, and the z^+-rule to the convolutional layers, except for the first convolutional layer where we use the z^B-rule.

Occlusion (Zeiler & Fergus, 2014) Occlusion uses a sliding window of size K and stride s over the input. Following (Rao et al., 2022), we use $K = 16$, $s = 8$ for the input layer and $K = 5$, $s = 2$ for the final layer.

RISE (Petsiuk et al., 2018b) We use a total of $N = 6000$ randomly generated masks per layer. Masks are initially generated on a low-resolution grid of size $s \times s$ (with default $s = 6$), where each cell is activated with probability 0.1. Each binary grid is bilinearly upsampled and cropped to match the input dimensions. The implementation supports CNN-like 4D tensors, following the setup in (Rao et al., 2022), as well as ViT-like 3D tensors.

A.2. Vision Transformer ViT-B/16

While evaluating attribution methods at the input layer for Vision Transformers is standard, evaluating them at the final layer requires modifications to the attribution logic. In CNNs, the final layer contains activations in the form of a 3D tensor of shape (C, H, W), where C is the number of channels and (H, W) correspond to spatial locations at a lower resolution compared to the input image. This structure aligns naturally with the input image and is usually interpolated (i.e., resized) to match its dimensions.

However, Vision Transformers (e.g., ViT-B/16 (Dosovitskiy et al., 2020)) process the image as a sequence of patches, with each patch treated as a token. The final feature representation prior to classification is typically of shape (N, D), where N is the number of tokens (including the [CLS] token), and D is the hidden dimension (e.g., 768 for ViT-B/16). For ViT-B/16 and a 224×224 input image, this results in $N = 197$ tokens: 1 [CLS] token and 196 patch tokens, where each patch corresponds to a 14×14 grid in the original image (a result of dividing 224 by 16).

To enable attribution on ViTs, we discard the [CLS] token and reshape the remaining 196 tokens into a $(14, 14)$ grid, analogous to CNN feature maps. This allows us to treat patch embeddings as spatial features and apply attribution methods at the final layer similarly to how they are applied in CNNs. We adapt our attribution logic to extract these reshaped features, which are then bilinearly interpolated to the original image size.

We have extended all attribution methods to be ViT-compatible, with the exception of LRP (Bach et al., 2015) and CAM (Zhou et al., 2016), due to the lack of direct extensions for these methods.

B. Evaluation of certified attributions on 5 models: ResNet-18, W-ResNet-50-2, ResNet-152, VGG-11 and ViT-B/16

B.1. Certified Robustness (%certified)

We assess attribution robustness on five different models using the %certified metric evaluated at the input and final layers across certified radii ($R = 0.10, 0.17, 0.22$) in Figure 9 and sparsification ($K = 50$, 30 and 10) in Figure 10. Note that ViT-B/16 is not evaluated on CAM and LRP. The general trend of reduced certification rates by increasing the certified radius R, as well as reduced certified top $K\%$ pixels by decreasing K holds across all five models.

Input layer LRP and RISE exhibit the highest robustness, as they also maintain a relatively high %certified scores across all three radii values in Figure 9. Interestingly, they also maintain a balance in certified top $K\%$ pixels by decreasing K in Figure 10.

Final layer Though Grad and GB still seem the most robust on CNNs in Figure 9, they localize very poorly, this is because they produce grid-like almost constant attributions at the final layer of the CNN architecture, due to max-pooling.

ViT-B/16 Gradient-based methods exhibit low robustness on ViT-B/16 comparable to their CNNs performance on both the input and final layers in Figure 9. Interestingly, RISE maintains the same performance of having high robustness across

both architecture types, with the highest at radius $R = 0.10$ at the final layer on ViT-B/16. Activation-based methods perform poorly on the transformer model compared to CNNs on the final layer. This is likely due to the lack of spatially structured convolutional features in ViTs, which affects the quality and stability of activation-based attributions when applied to token-based representations. We discuss this implementation detail in App. A.2.

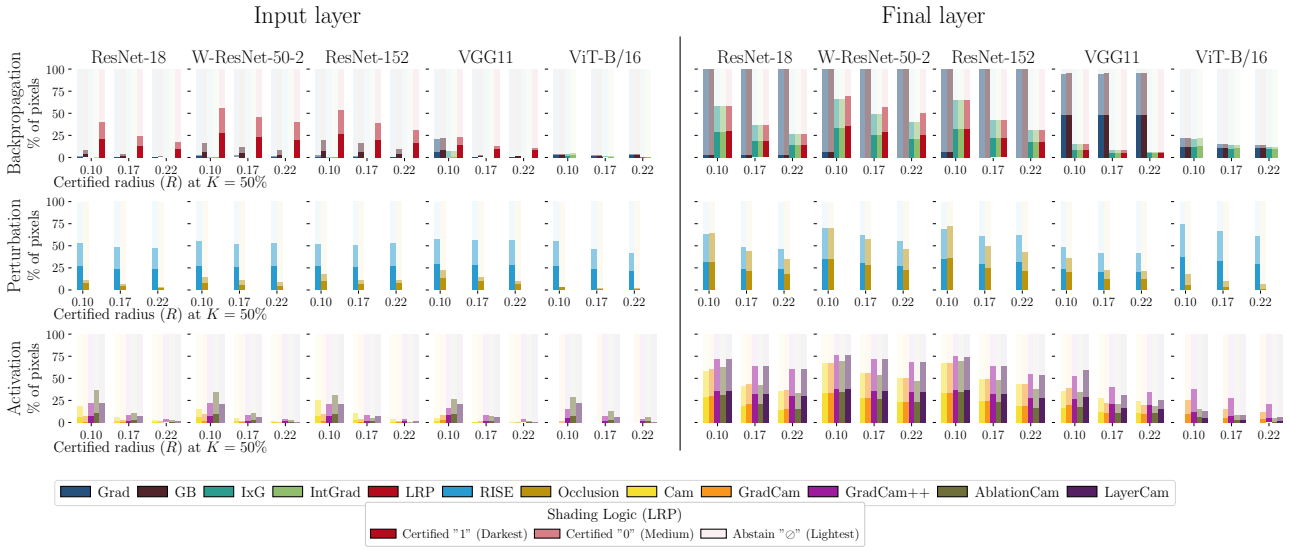


Figure 9: **Comparison of the per-pixel certification rate (%certified) against the certified radius R on all 5 models across backpropagation, activation and perturbation methods.** (*Left*) shows evaluation at the input and (*Right*) at the final layer. The darkest shades denote %certified pixels, while brightest denotes abstain ∅.

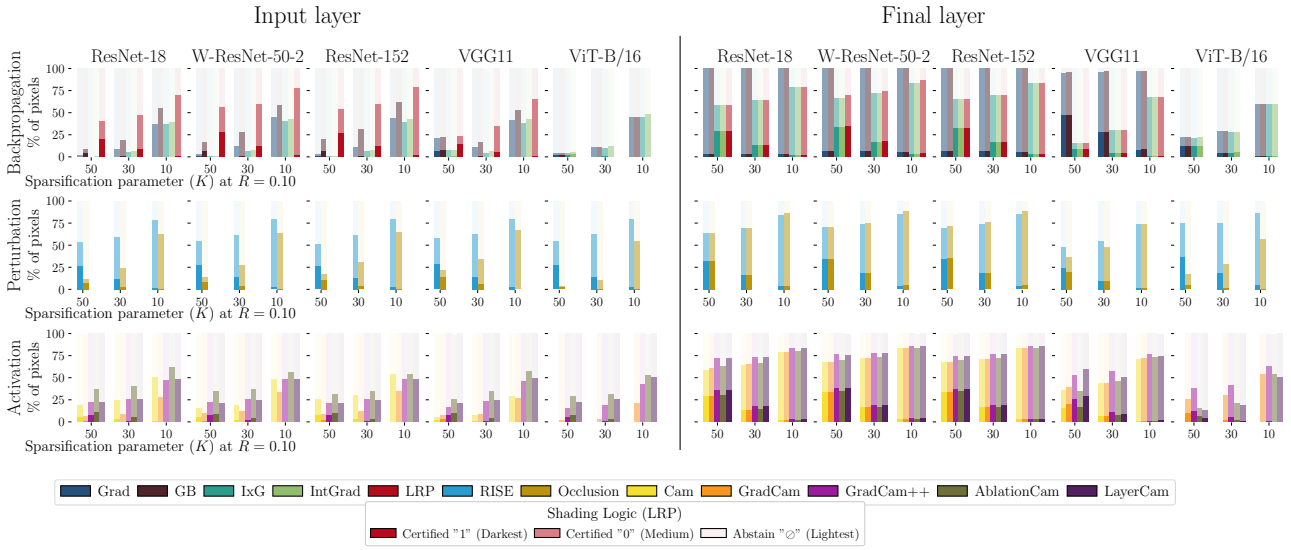


Figure 10: **Comparison of the per-pixel certification rate (%certified) against sparsification values K on all 5 models across backpropagation, activation and perturbation methods.** (*Left*) shows evaluation at the input and (*Right*) at the final layer.

B.2. Certified Localization (Certified GridPG)

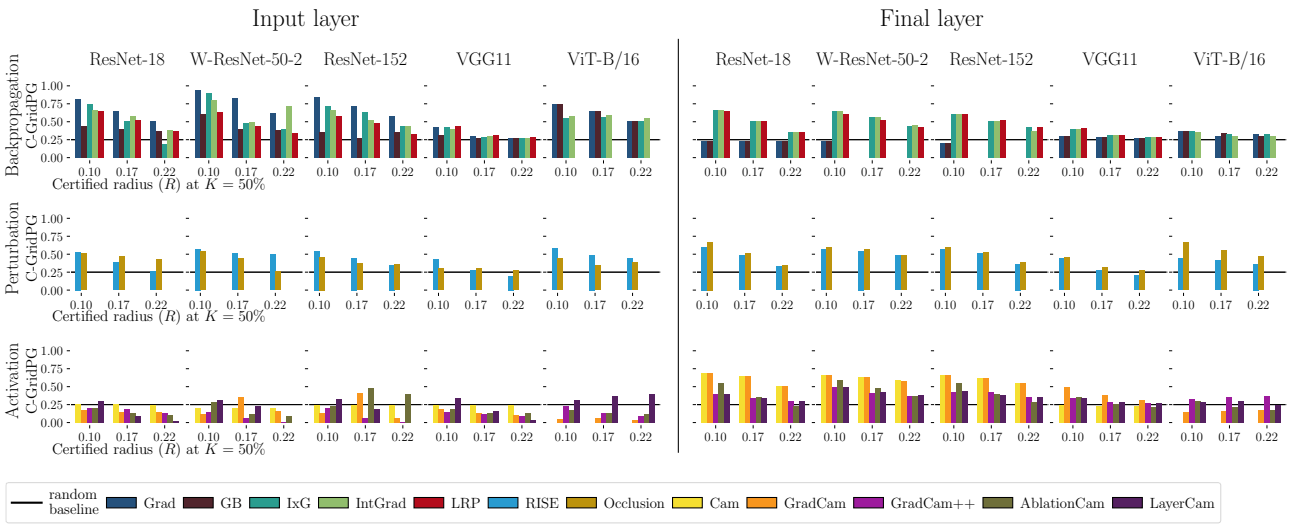


Figure 11: **Comparison of the certified localization (Certified GridPG) against the certified radius R on all 5 models across backpropagation, activation and perturbation methods.** (*Left*) shows evaluation at the input and (*Right*) at the final layer.

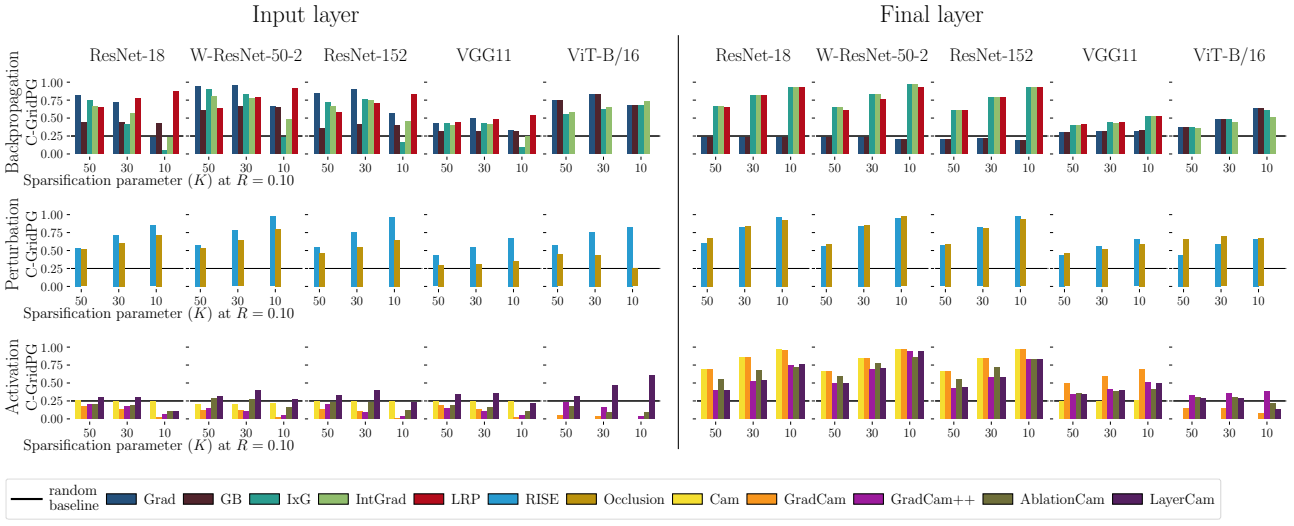


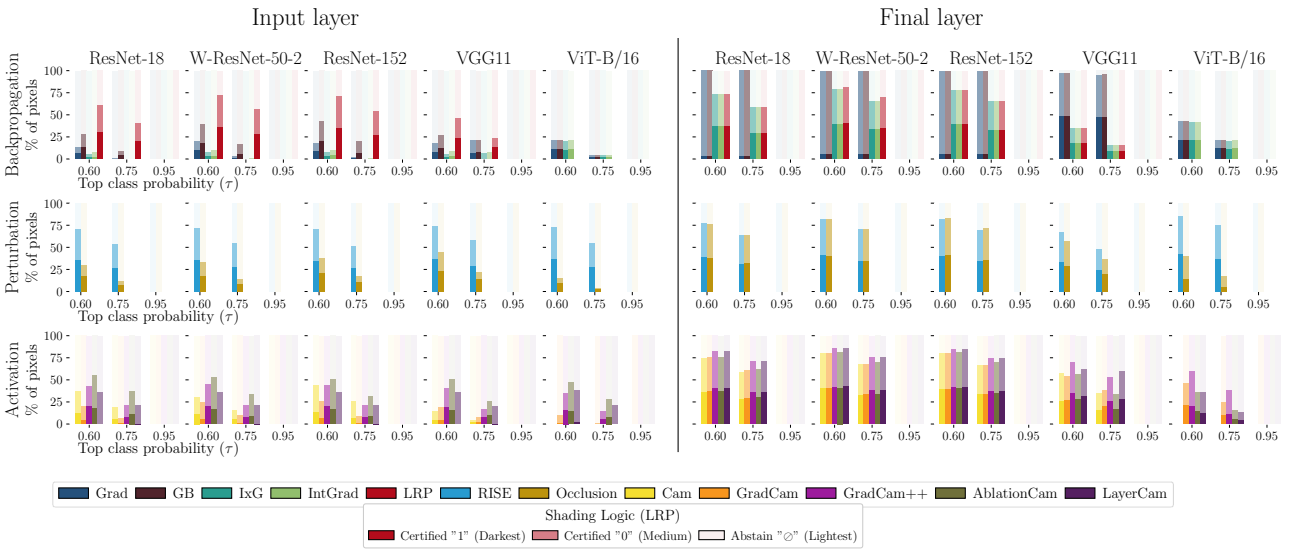
Figure 12: **Comparison of the certified localization (Certified GridPG) against the sparsification values K on all 5 models across methods.** (*Left*) shows evaluation at the input and (*Right*) at the final layer.

Input layer In Figures 11 and 12, backpropagation and perturbation methods demonstrate effective localization at the input layer across all models, except VGG-11. This is likely due to the absence of skip connections in VGG-11, which hampers gradient flow and impedes signal propagation to the input. In contrast, activation-based methods perform worse than random across all models, as they rely solely on high-level forward activations from the final layers, lacking direct correspondence with input pixels.

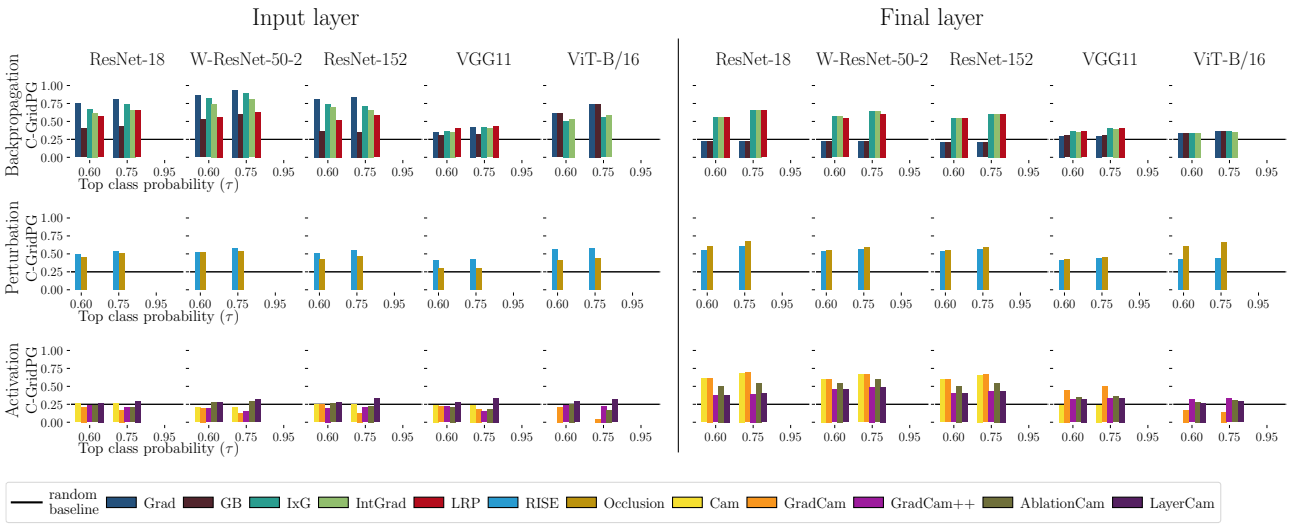
Final layer At the final layer, all attribution methods generally yield higher localization scores compared to the input layer, as shown in Figures 11 and 12. This improvement is particularly evident in activation and perturbation methods, which align better with the spatial structure of final-layer features. However, exceptions are observed in Grad and Guided Backpropagation (GB), which underperform relative to other methods.

C. Impact of varying the certification strictness (hyperparameter τ)

One of the hyperparameters in our certification setup is $\tau \in (0.5, 1]$, which sets a threshold for the top class probability of a pixel to certify it (discussed in Section 4.2 and defined in Eq. 5). The higher the value of τ, the more strict the certification condition is. We investigate the effect of increasing τ on %certified and Certified GridPG of all attribution methods in Figure 13.



(a) The certification rate (%certified) against the top class probability values τ.



(b) Certified GridPG against the top class probablity values τ.

Figure 13: The performance of attribution methods in terms of %certified (a) and Certified GridPG (b) by increasing the top class probability τ (making the certification more strict)

Effect of τ on %certified In Figure 13, we observe that increasing τ lowers the certification rate of all methods, except for Grad and GB at the final layer, since they produce almost constant grid-like patterns at that layer in ResNet18. At the highest value of $\tau = 0.95$, the performance of all methods drops to 0, indicating for the need of increasing the number of samples to be able to certify under the strict condition imposed by a high τ.

Effect of τ on Certified GridPG In Figure 13 (b), interestingly, increasing τ from 0.60 to 0.75 boosts the localization performance of all methods (with the exception of activation methods completely failing at the input layer). This indicates that by imposing a more strict certification setup, only the most confident pixels are certified to top K% in attribution output, which also localizes better. Hence, we use a default value of $\tau = 0.75$ throughout the paper.

D. GridPG and Certified GridPG

To understand how attribution methods maintain localization under input noise, we analyze the relationship between the original localization score (GridPG) and the certified localization score (Certified GridPG) in Figure 14. This comparison helps assess the robustness of each method's localization ability when subjected to noise during certification. At the input layer (Figure 14, top), LRP outperforms all methods, striking a balance in both metrics. Backpropagation-based and activation-based methods exhibit a higher Certified GridPG than their respective GridPG scores. This increase may be attributed to the certification process rejecting incorrect positive evidence located outside the correct subimage in the grid, thereby boosting the certified localization score. At the final layer (Figure 14, bottom), RISE is the only method that achieves a higher Certified GridPG than GridPG, demonstrating its ability to improve localization when subjected to input noise. Meanwhile, all other methods roughly lie on the diagonal, showing equal values of Certified GridPG and GridPG. GradCam achieves the highest scores in both GridPG and Certified GridPG (with the exception of ViT-B/16), indicating that its localization performance remains consistent between the raw and certified outputs.

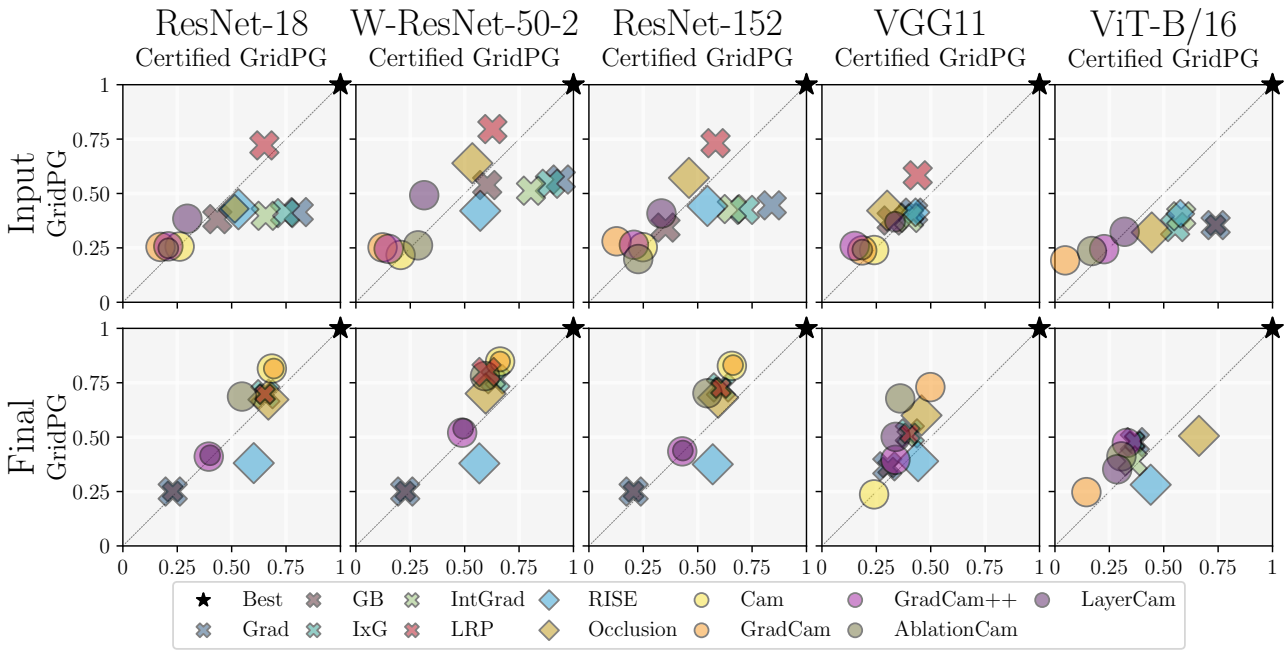


Figure 14: The performance of attribution methods on ResNet18 in terms of original GridPG score against Certified GridPG.

E. Impact of varying sparsification and certified radius on certified visuals

We present qualitative examples showing the certified visuals on ResNet-18 (Figures 15 and 16) and additionally ResNet-152 (Figures 17 and 18) of all attribution methods by varying the sparsification parameter ($K = 50\%, 25\%, 10\%$ and 5%) and certified radius ($R = 0.10, 0.17, 0.22$).

As the radius R increases with higher noise levels (σ), all methods abstain more and certify fewer top $K\%$ pixels. Amongst backpropagation-based methods, only LRP, Grad and GB show certified top K% pixels in the overlayed output, whilst IxG and IntGrad almost abstain from all top K% pixels at all radii. A notable observation is that activation-based methods show high-quality overlayed certified maps across all noise levels (certified radii) on both models ResNet-18 and ResNet-152. Perturbation-based methods maintian high-quality overlayed maps across all certified radii.

As K decreases, fewer pixels are certified within the top $K\%$, while more pixels fall within the lower $(100 - K)\%$ across methods. At smaller K values (e.g., $K = 5\%$), most methods highlight more distinctive features. Overlaying certified attribution maps across different K values offers insights into the relative pixel importance for each method.

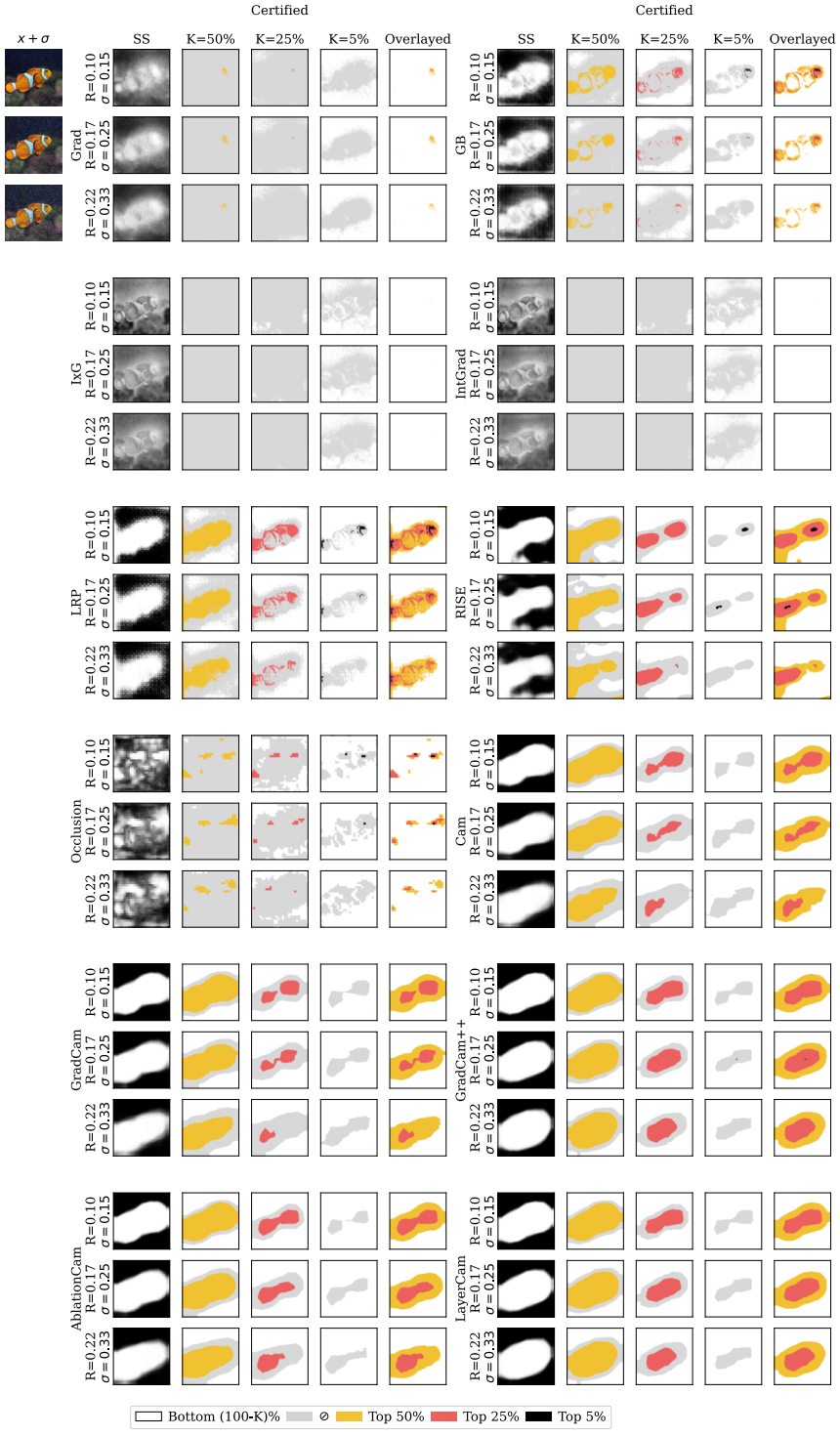


Figure 15: **Example image and its certified attribution maps on ResNet-18 of all methods at different sparsification (K) and certified radius K values**. SS (Smoothed Sparsified) is evaluated at $K = 50\%$. The "Overlayed" last column shows the certified top $K\%$ pixels from row-wise certified maps at different K values, with lower K taking precedence for each pixel.

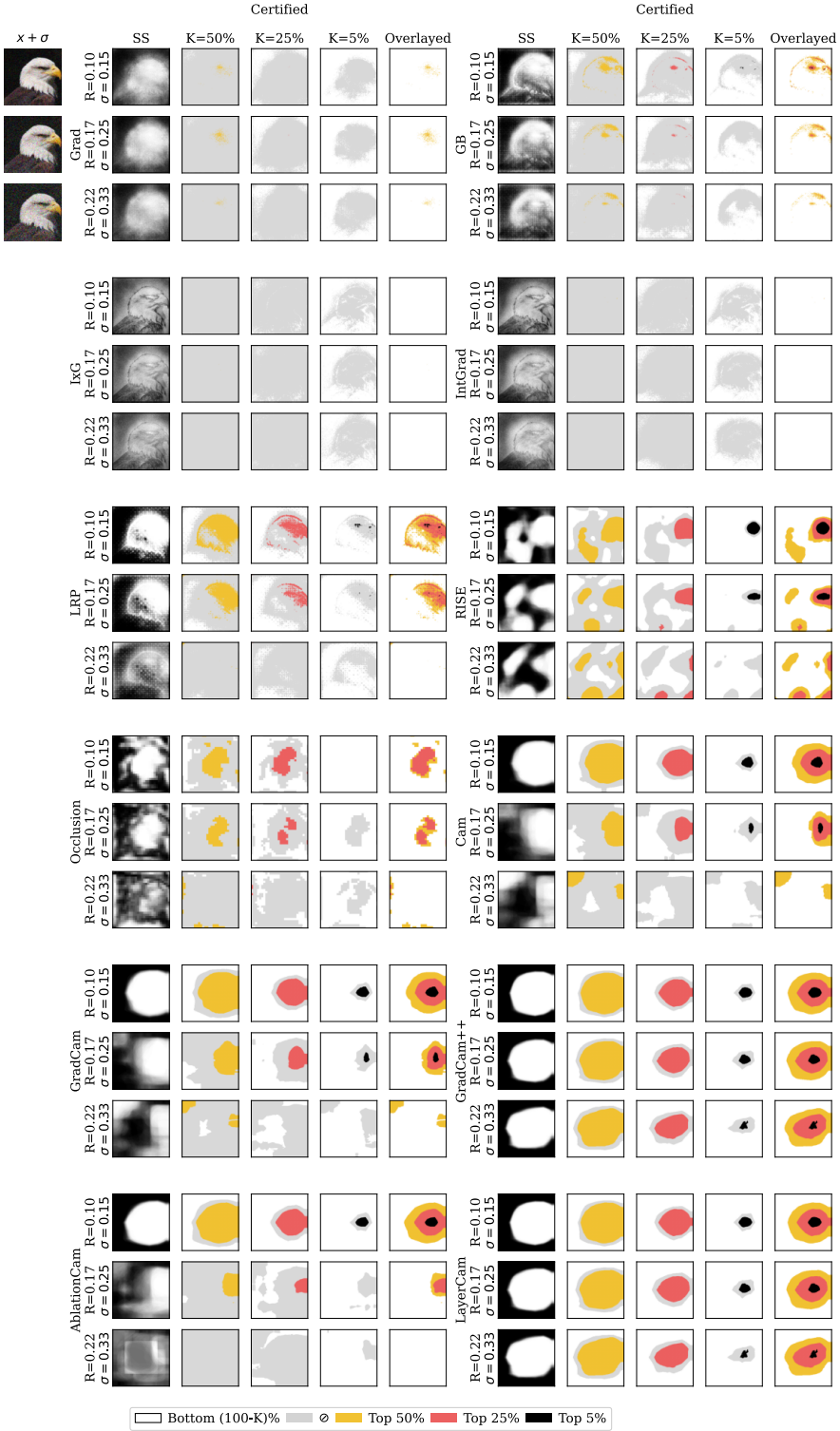
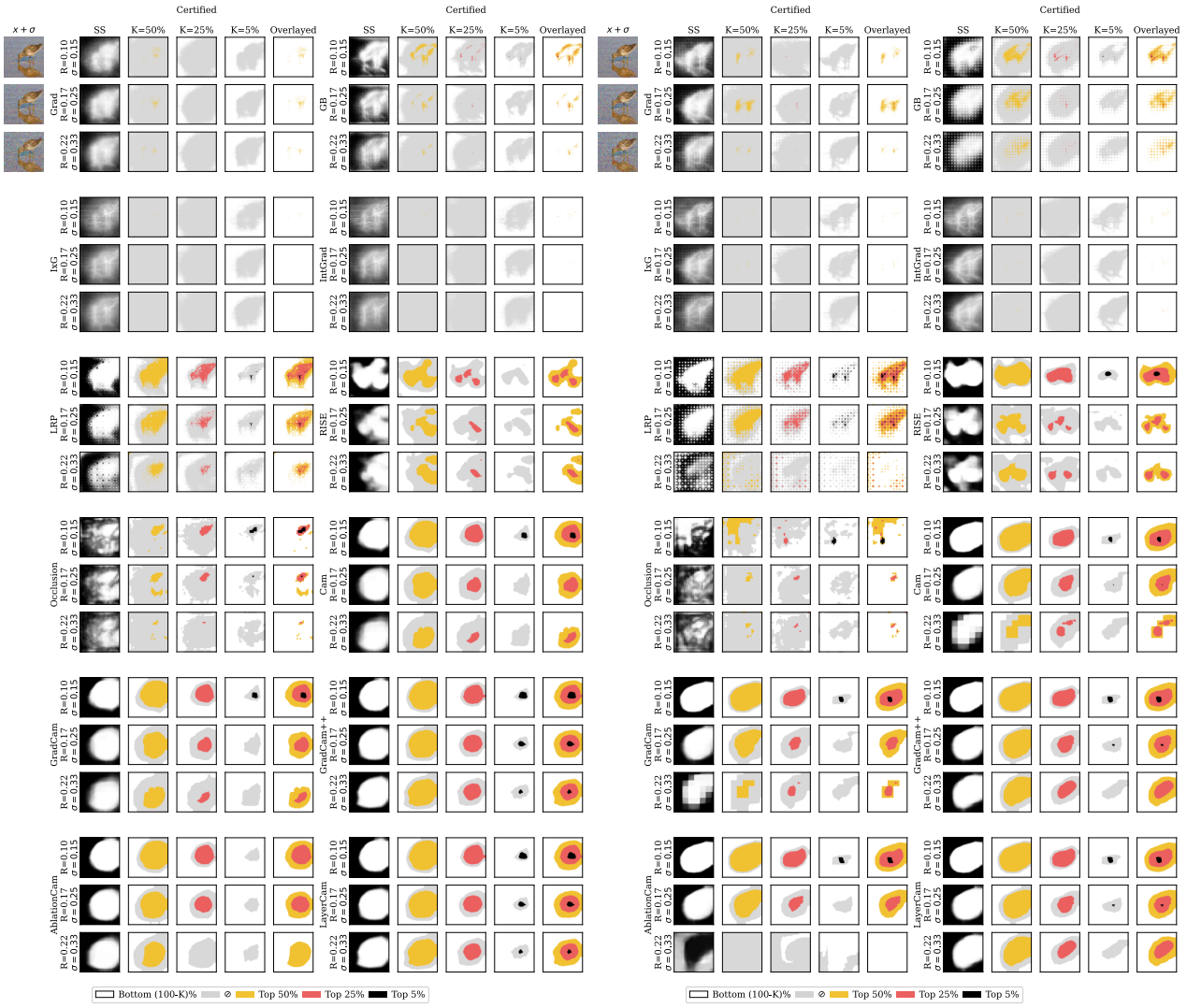


Figure 16: **Example image and its certified attribution maps on ResNet-18 of all methods at different sparsification (K) and certified radius K values**. SS (Smoothed Sparsified) is evaluated at $K = 50\%$. The "Overlayed" last column shows the certified top $K\%$ pixels from row-wise certified maps at different K values, with lower K taking precedence for each pixel.



(a) ResNet-18 (b) ResNet-152

Figure 17: **Certified attribution maps on ResNet-18 (a) and ResNet-152 (b) of all methods at different sparsification (K) and certified radius R values**. SS (Smoothed Sparsified) is evaluated at $K = 50\%$. The "Overlayed" last column shows the certified top $K\%$ pixels from row-wise certified maps at different K values, with lower K taking precedence for each pixel.

(a) ResNet-18 (b) ResNet-152

Figure 18: **Example image and its certified attribution maps on ResNet-18 (a) and ResNet-152 (b) of all methods at different sparsification (K) and certified radius K values**. SS (Smoothed Sparsified) is evaluated at $K = 50\%$. The "Overlayed" last column shows the certified top $K\%$ pixels from row-wise certified maps at different K values, with lower K taking precedence for each pixel.

F. Certified attribution visuals on ViT-B/16

Figure 19: **Certified attribution maps of ViT-B/16 of all methods at different sparsification parameter (K) values**. SS (Smoothed Sparsified) is evaluated on $n = 100$ noisy input samples per image and at $K = 50\%$. The "Overlayed" column shows certified top $K\%$ pixels per row, with lower K taking precedence.

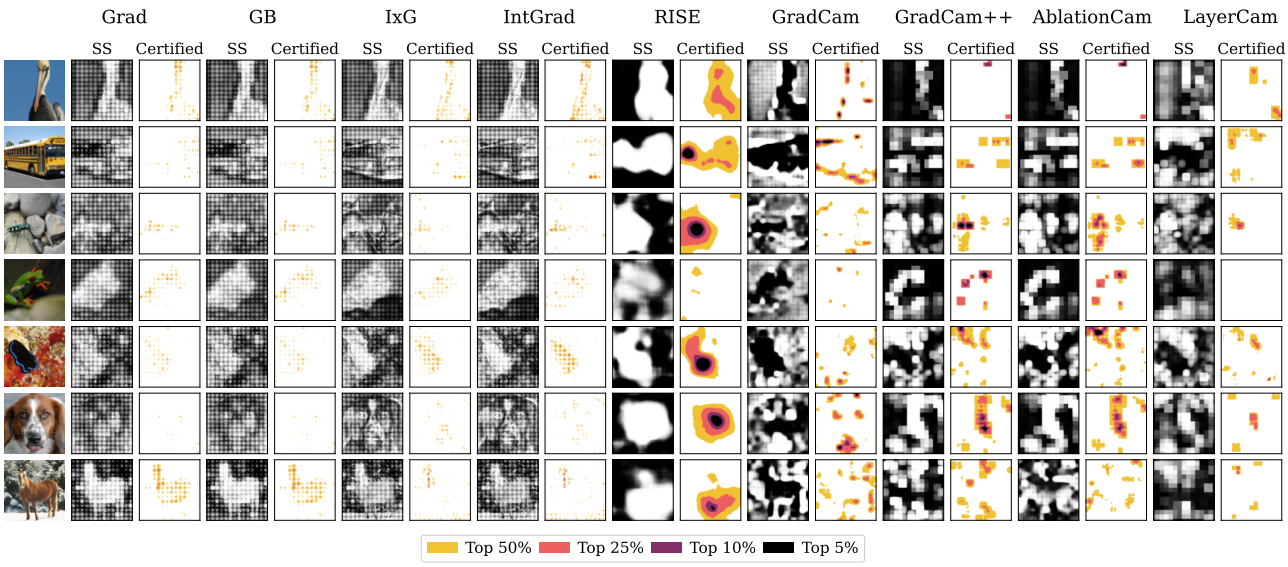


Figure 20: **Overlayed certified attributions on ViT-B/16 at different K values across methods .** SS (Smoothed Sparsified), which refers to the average of the sparsified attributions, is evaluated on $n = 100$ noisy input samples per image and at $K = 50\%$.

G. Qualitative Evaluation of Certified Localization

In this section, we present qualitative results using the AggAtt layout introduced by Rao et al. (2022), and adapting it for attribution methods evaluated on Certified GridPG at input and final layers. AggAtt (Rao et al., 2022) is a qualitative evaluation method that generates aggregate attribution maps by sorting maps based on their localization scores and grouping them into percentile bins. These bins, with smaller sizes for the top and bottom percentiles, highlight both the best and worst-case performance of attribution methods. This approach provides a comprehensive view of method performance across diverse inputs, emphasizing both consistent trends and distinct failure cases.

In Figures 21 and 22, we show an example from the median position of each AggAtt (Rao et al., 2022) bin for each attribution method at the input and final layers, respectively, evaluated on Certified GridPG at the top-left grid cell using ResNet-18 (He et al., 2016).

At the input layer, **backpropagation-based methods** show relatively better localization in the top-left grid cell. LRP (Bach et al., 2015) shows the best localization across these methods, followed by Grad (Simonyan, 2014). Meanwhile, GB (Springenberg et al., 2014) seems to mainly highlight edges in the input irrespective of the grid cell, and IxG (Shrikumar et al., 2017) and IntGrad (Sundararajan et al., 2017) only have very few pixels certified top $K\%$ pixels within the top-left grid cell. **Activation-based methods** show poor performance by almost abstaining from the entire grid. This aligns with the poor quantitative performance of these methods at the input layer in Figure 6.

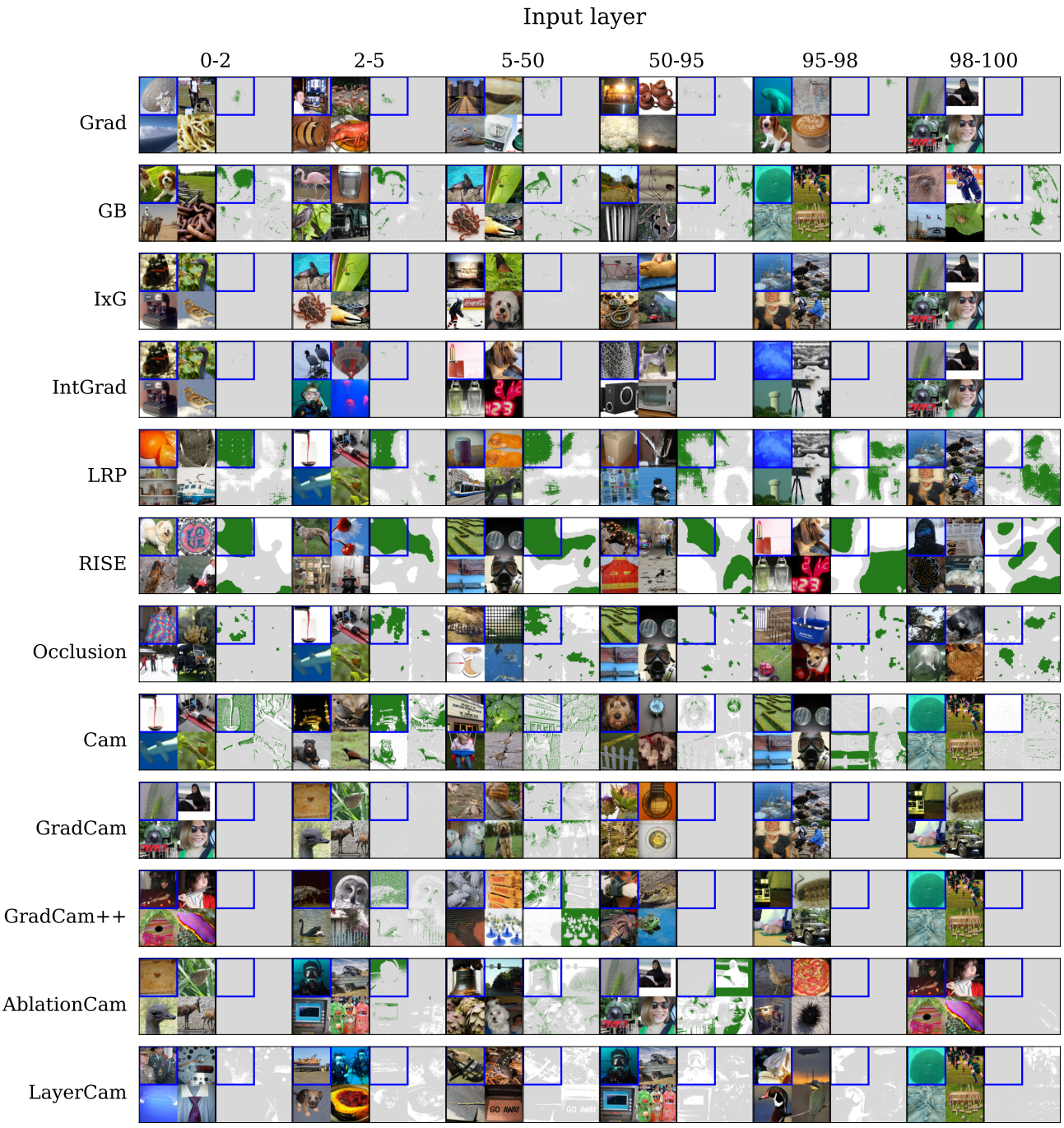


Figure 21: **Examples from each AggAtt for all methods at the input layer using Certified GridPG.** White denotes certified "0", black is certified "1" and gray is abstain ⊘. The percentile bin values are displayed at the top of every column. Default values: $K = 50\%$, $R = 0.10$, $\tau = 0.75$ and $n = 100$.

At the final layer (Figure 22), certified attributions from Grad (Simonyan, 2014) and GB (Springenberg et al., 2014) show a constant pattern where all pixels are certified as bottom $(100 - K)\%$. The localization of the rest of the methods improves considerably compared the input layer in Figure 21, which agrees with the quantitative results from Figure 6. All other methods show good localization, with the best example coming from Occlusion (Zeiler & Fergus, 2014), which achieves near perfect localization at the first $0 - 2$ AggAtt bin.

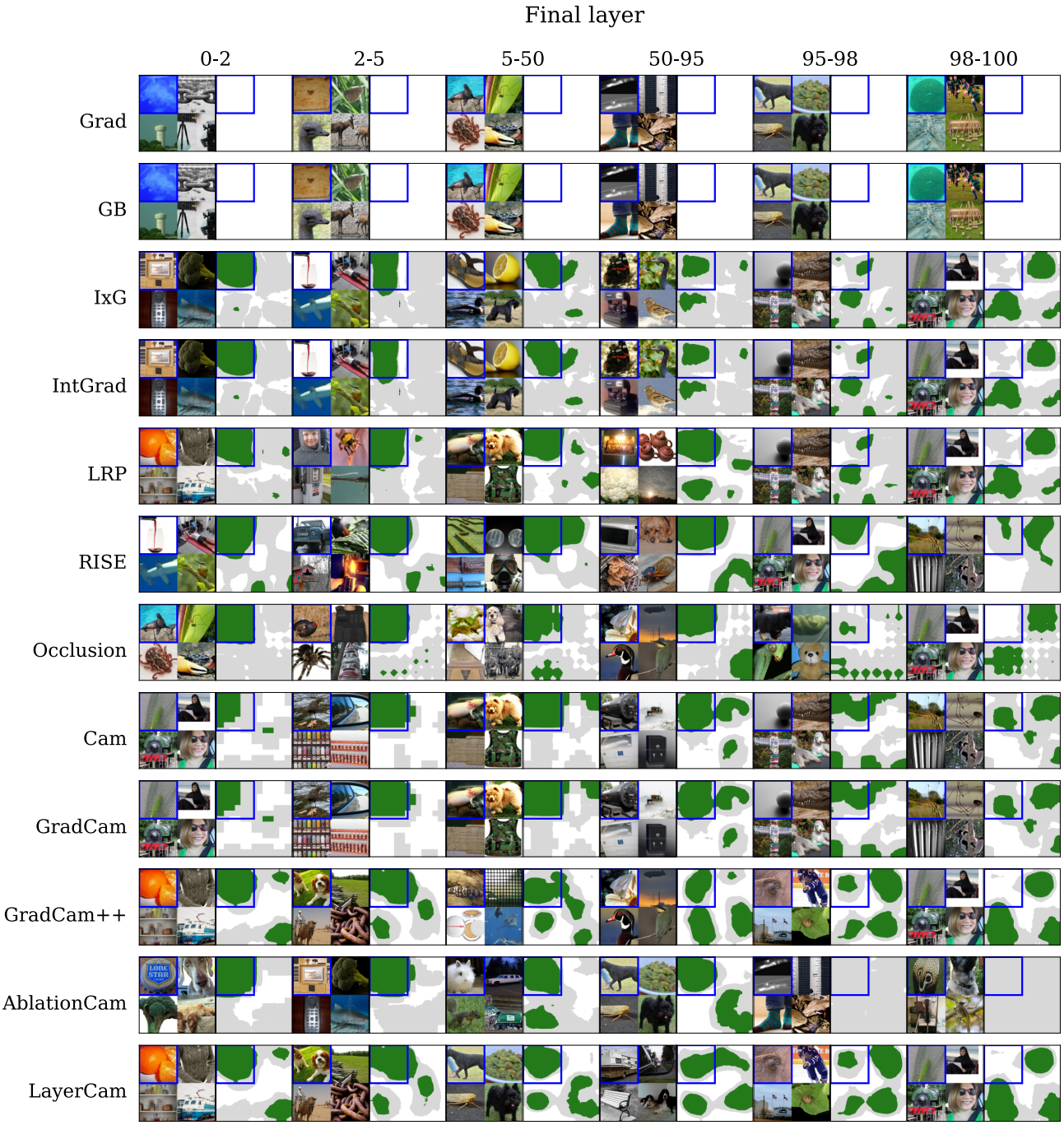


Figure 22: **Examples from each AggAtt for all methods at the final layer using Certified GridPG.** White denotes certified "0", black is certified "1" and gray is abstain ⊘. The percentile bin values are displayed at the top of every column. Default values: $K = 50\%$, $R = 0.10$, $\tau = 0.75$ and $n = 100$.

Finally, we show the AggAtt (Rao et al., 2022) bins for all methods on Certified GridPG at both layers at different sparsification values and certified radii in Figure 23. We see that the AggAtt bins in this comprehensive figure reflect the trends in the quantitative results in Figure 5 and Figure 6, as well as the qualitative results in Figure 21 and Figure 22.

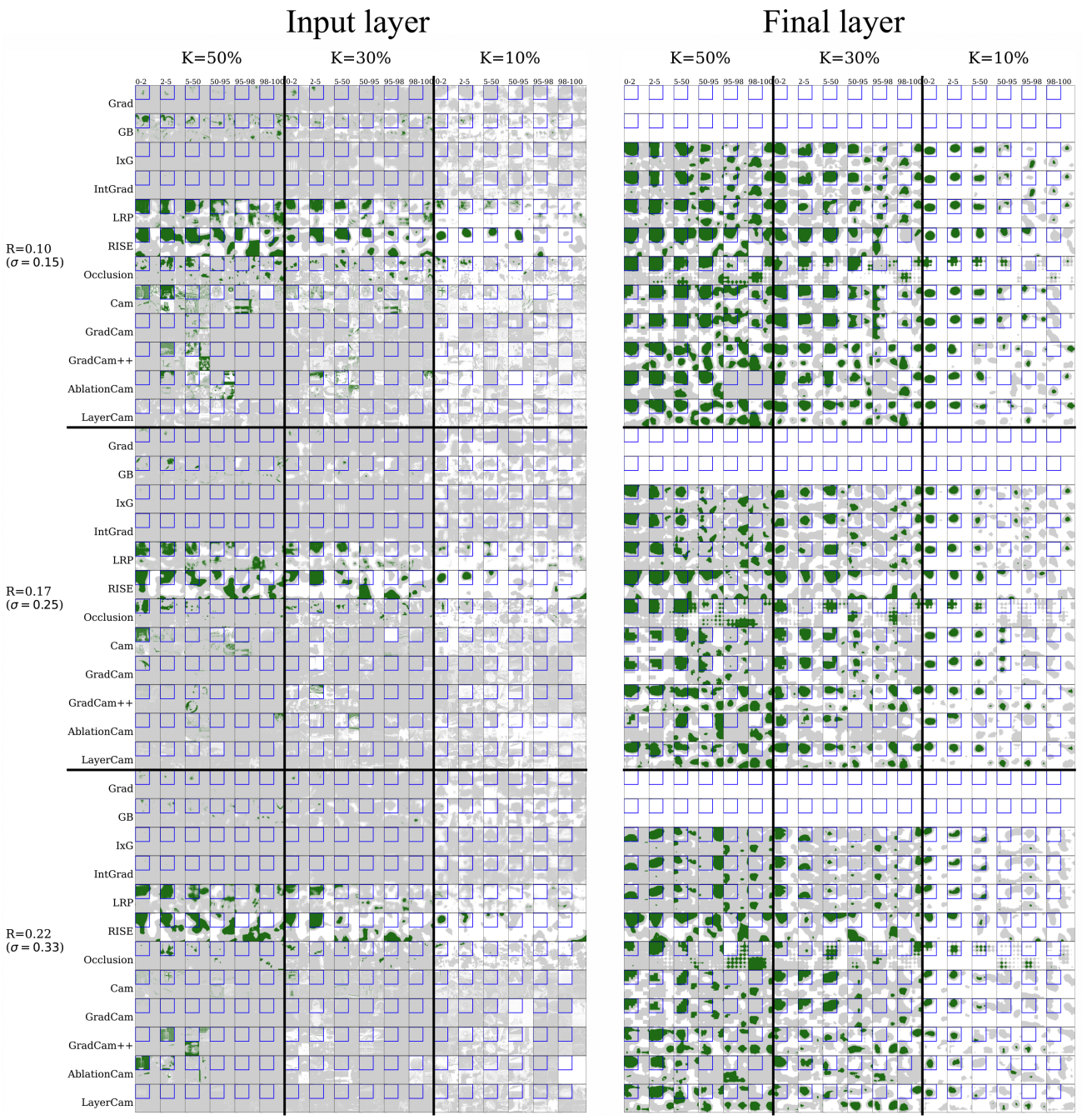


Figure 23: **AggAtt Evaluation on Certified GridPG for all attribution methods at the input and final layers** across sparsification parameters K and certified radii R values.