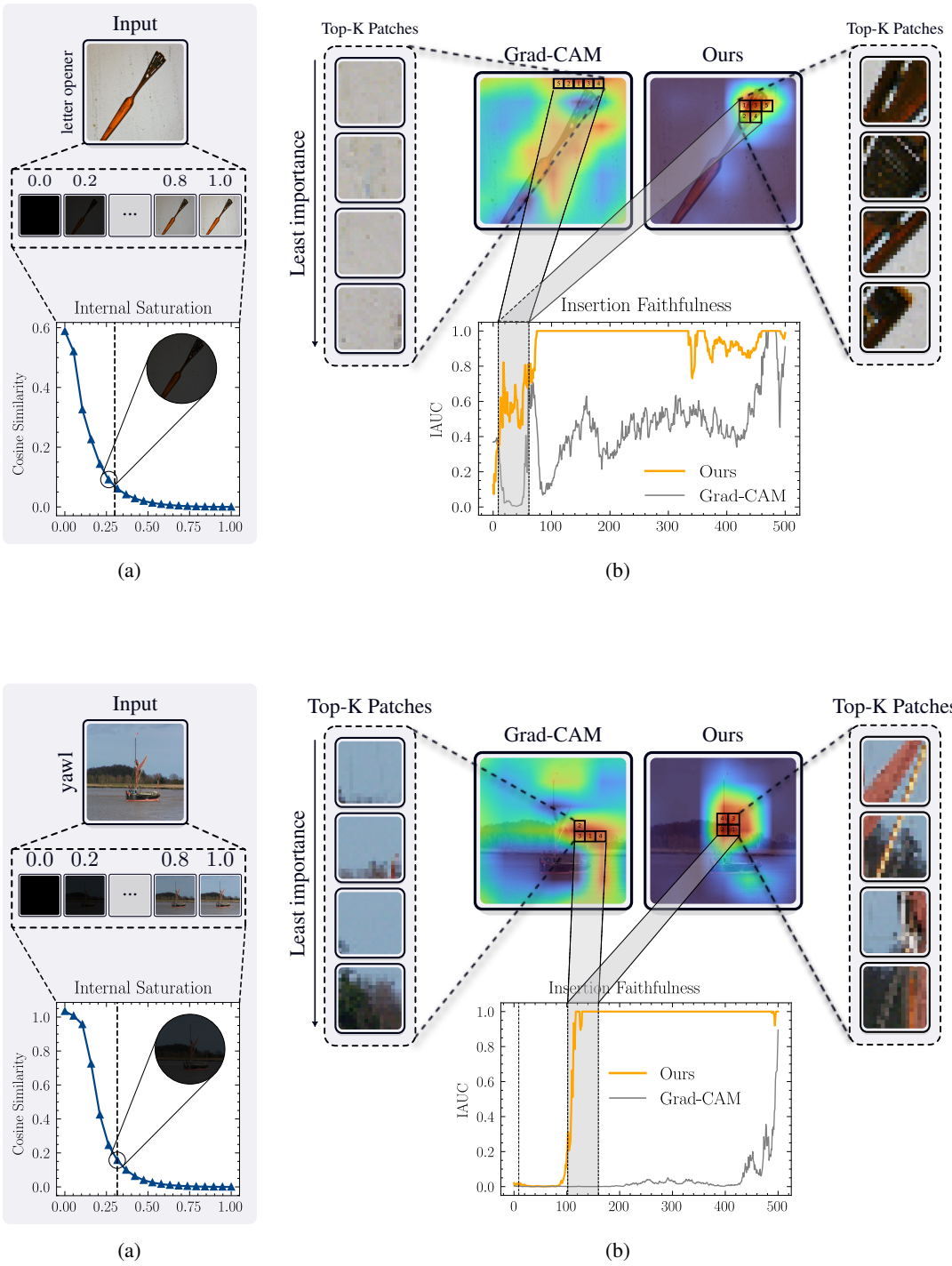

Figure 1: Comparison of the attribution maps under internal saturation conditions. In Figure 1a is shown the cosine similarity of the target layer's embeddings with respect to the interpolator parameter (α) (For more see Appendix **??**). Figure 1b shows the attribution maps of the different methods under the saturation condition. The internal saturation condition causes the baseline method to under-represent feature importances across saturating ranges. By extracting the top-4 most important features (fig. 1b) we can observe that the baseline method fails to capture the relevant discriminative regions, which produce low insertion AUCs (fig. 1b) as deemed not important by the model.