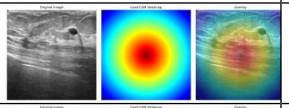
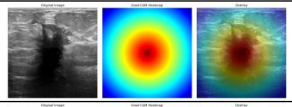
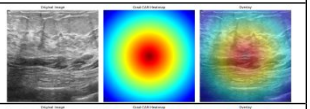
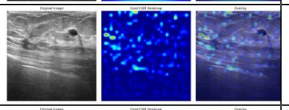
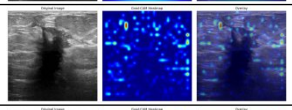
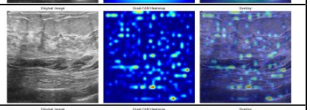
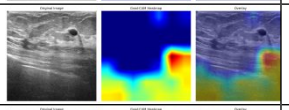
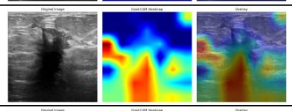
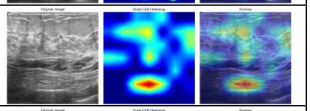
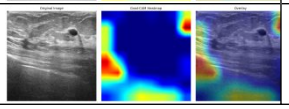
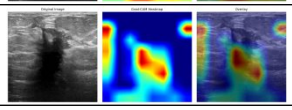
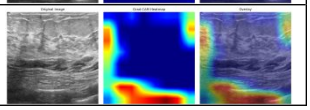
	Benign	Malignant	Normal
Origin			
Fabfrnet			
Hovertrans			
Freqtrans(ours)			

We thank the reviewers for their constructive comments regarding the need for stronger qualitative evidence and clearer interpretability of the proposed method. In response, we provide a consolidated qualitative analysis based on Grad-CAM visualizations, covering benign, malignant, and normal cases. All visualizations are generated using the same input images across different models to ensure a fair and consistent comparison.

Across the twelve Grad-CAM examples, several consistent patterns can be observed. For malignant cases, FreqTrans produces compact and well-localized activations that closely align with lesion boundaries and internal structures. In particular, the highlighted regions tend to follow irregular margins and low-contrast contours that are typical of malignant tumors in ultrasound images. By contrast, baseline methods such as FABRF-Net and HoVer-Trans often exhibit more diffuse responses, with attention spreading into surrounding background tissue or speckle-dominated regions. This difference suggests that explicit frequency-domain modeling helps suppress irrelevant low-frequency background variations while preserving high-frequency boundary cues.

For benign cases, FreqTrans shows moderate and spatially constrained activations that correspond to smoother lesion regions, without over-emphasizing surrounding tissue. Several baseline models, however, generate broader or fragmented activations, which may explain their higher false positive rates under fixed-recall evaluation. In normal cases, FreqTrans largely avoids strong activations, indicating that the model does not rely on spurious textural patterns for decision making. In contrast, some baselines still produce scattered responses in homogeneous tissue regions, likely influenced by speckle noise or acquisition artifacts.

In addition, a direct comparison between HoVer-Trans and FreqTrans reveals clearer differences in attention behavior. HoVer-Trans is designed to leverage anatomical priors and global context modeling, and its Grad-CAM visualizations generally highlight broader regions that often cover the lesion and part of the surrounding tissue. While this global awareness can be beneficial, it may also lead to less precise localization, especially in cases where lesion boundaries are blurred or embedded in heterogeneous background structures.

In contrast, FreqTrans produces more compact and sharply localized activation patterns. For malignant cases, the attention maps generated by FreqTrans are more consistently concentrated along irregular lesion margins and internal high-contrast regions, while suppressing responses in homogeneous background tissue. This suggests that explicit frequency-domain decomposition helps the model disentangle boundary-related high-frequency cues from low-frequency background variations, which are common in ultrasound imaging.

For benign and normal cases, HoVer-Trans occasionally exhibits residual activations in surrounding tissue regions, possibly due to its reliance on spatial correlations and global feature aggregation. FreqTrans, however, shows reduced activation strength in non-lesion areas, indicating improved robustness against speckle noise and textural artifacts. These differences are particularly evident when comparing the same input images across models, where FreqTrans

demonstrates more discriminative and anatomically meaningful focus.

We further compare FreqTrans with FABRF-Net, which is a representative frequency-aware model that explicitly fuses boundary and regional information for breast ultrasound analysis. FABRF-Net is effective at enhancing edge-related responses, and its Grad-CAM visualizations often emphasize strong boundary structures around lesions. In several malignant cases, FABRF-Net highlights sharp contour regions, which is consistent with its design philosophy of boundary-region fusion.

However, we observe that FABRF-Net tends to produce more fragmented or scattered activation patterns, particularly in cases with heterogeneous internal textures or weak lesion contrast. In some malignant samples, attention responses extend into adjacent background regions with similar frequency characteristics, which may contribute to increased false positive activations under high-sensitivity settings. This behavior aligns with the quantitative results, where FABRF-Net achieves strong area under the curve performance but exhibits lower recall compared with FreqTrans.

In contrast, FreqTrans demonstrates more coherent and structurally consistent attention maps. The proposed adaptive frequency fusion mechanism enables the model to dynamically balance low-frequency contextual information and high-frequency boundary cues, rather than emphasizing boundary responses alone. As a result, FreqTrans focuses not only on lesion contours but also on diagnostically relevant internal regions, leading to more stable and concentrated activations across different lesion appearances.

For benign and normal cases, FABRF-Net occasionally shows residual responses along anatomical structures with strong edges, whereas FreqTrans suppresses such activations more effectively. This difference suggests that the explicit separation and adaptive reweighting of frequency components in FreqTrans improves robustness against background artifacts and texture-induced false alarms.

Overall, the qualitative comparison indicates that while FABRF-Net provides strong boundary sensitivity, FreqTrans offers a more balanced and interpretable attention behavior by jointly modeling frequency-domain information and global context. This complementary design helps explain why FreqTrans achieves higher recall and lower false positive rates under clinically relevant operating points.

4These observations help explain the quantitative results reported in the fixed-recall evaluation, where FreqTrans achieves a lower false positive rate than HoVer-Trans under the same sensitivity constraint. The qualitative comparison therefore provides complementary evidence that the proposed frequency-aware design not only improves recall but also enhances spatial precision, which is critical for reducing unnecessary follow-up examinations in clinical screening scenarios.

These qualitative observations directly address the concerns raised by Reviewer 3430 and Reviewer 5DE0 regarding the lack of clinically meaningful visual evidence. They also respond to Reviewer 7AE9's request for better interpretation of which information the model leverages. While Grad-CAM does not provide a causal decomposition of frequency contributions, the visual patterns are consistent with the design of FreqTrans. The adaptive fusion of low-frequency global context and high-frequency structural details guides the model toward anatomically relevant regions rather than background textures.

Following Reviewer 49A5's suggestion, we present all Grad-CAM results in a single consolidated figure rather than splitting them into multiple panels. This design choice enables direct side-by-side comparison across models and classes, improving readability and interpretability while avoiding redundancy, which is particularly important under strict page constraints.

Overall, the qualitative results complement the quantitative improvements reported in recall, fixed-recall precision, and false positive rate. Together, they provide stronger evidence that FreqTrans not only improves numerical performance but also exhibits more clinically meaningful attention behavior in breast ultrasound image analysis.