

Appendix

727 Appendix A Extended Results

728 Due to space constraints, we report additional experimental results included in the main paper. Table
 729 7 and Table 8 extend Table 2 of the main paper. We investigate how many blocks of a standard ViT
 730 are used as the local encoder and the remaining as the global encoder for the fused tokens with all
 731 mentioned fusion strategies. Two key findings emerge from the results:

732 First, concatenation proves more robust to the choice of local/global ratio compared to the other
 733 fusion strategies. This robustness is expected, as concatenation preserves the information to the
 734 most extent and can fully utilize the global transformer blocks. Based on these results, we select
 735 concatenation as the default fusion strategy. Second, RTF generally enhances performance across all
 736 settings. The only exception occurs when using 25% blocks as the local encoder with CLS_{cat} . In this
 737 scenario, all spatial tokens are discarded at a very early stage, and only the two CLS tokens are sent to
 738 the global encoder, resulting in extremely low model capacity. Applying RTF in this situation harms
 739 performance, similar to the effects of aggressive regularization techniques on an already under-fitting
 740 model.

Table 7: Extended results on CBIS-DDSM, showing AUC performance depending on where the encoder is split for fusion.

Fusion	RTF Used	25% local	50% local	75% local
Average	No	0.753 ± 0.007	0.789 ± 0.014	0.803 ± 0.008
	Yes	0.756 ± 0.011	0.793 ± 0.006	0.809 ± 0.002
CLS_{cat}	No	0.711 ± 0.012	0.782 ± 0.007	0.802 ± 0.006
	Yes	0.709 ± 0.005	0.796 ± 0.001	0.811 ± 0.008
Concat	No	0.799 ± 0.002	0.799 ± 0.009	0.803 ± 0.003
	Yes	0.802 ± 0.001	0.810 ± 0.003	0.815 ± 0.001

Table 8: Extended results on CheXpert, showing AUC performance depending on where the encoder is split for fusion.

Fusion	RTF Used	25% local	50% local	75% local
Average	No	0.834 ± 0.004	0.845 ± 0.002	0.844 ± 0.004
	Yes	0.835 ± 0.003	0.849 ± 0.001	0.848 ± 0.002
CLS_{cat}	No	0.815 ± 0.003	0.841 ± 0.003	0.842 ± 0.006
	Yes	0.814 ± 0.003	0.844 ± 0.001	0.846 ± 0.001
Concat	No	0.842 ± 0.003	0.844 ± 0.003	0.843 ± 0.004
	Yes	0.845 ± 0.002	0.849 ± 0.001	0.849 ± 0.001

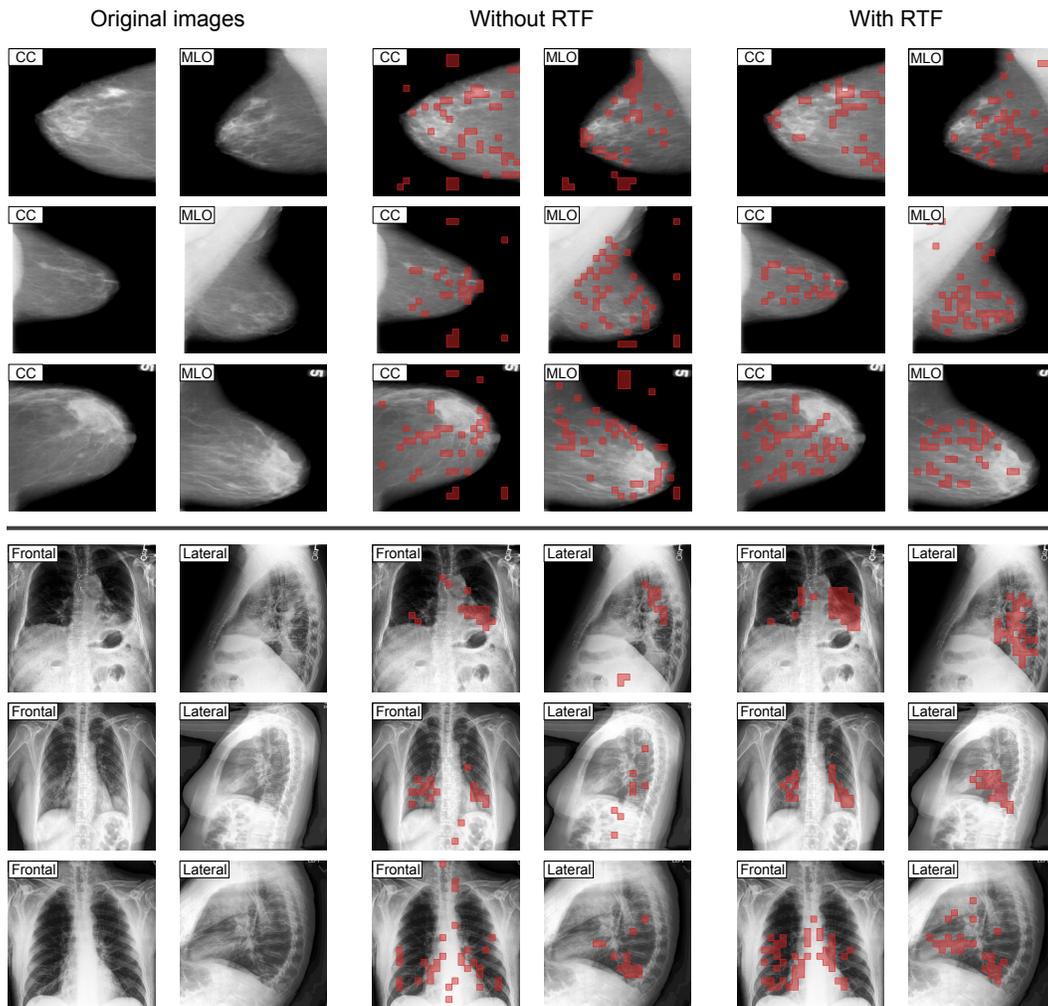


Figure 5: Extended results on CBIS-DDSM (**top**) and CheXpert (**bottom**), showing the model’s attention maps within the last block of the global encoder. RTF seems to address the issue of attention being allocated to uninformative areas, a common phenomenon observed in ViTs. It also encourages the model to focus on both views in many cases.