

## Appendix

### Appendix A Extended Results

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

First, concatenation proves more robust to the choice of local/global ratio compared to the other fusion strategies. This robustness is expected, as concatenation preserves the information to the most extent and can fully utilize the global transformer blocks. Based on these results, we select concatenation as the default fusion strategy. Second, RTF generally enhances performance across all settings. The only exception occurs when using 25% blocks as the local encoder with  $\text{CLS}_{\text{cat}}$ . In this scenario, all spatial tokens are discarded at a very early stage, and only the two CLS tokens are sent to the global encoder, resulting in extremely low model capacity. Applying RTF in this situation harms performance, similar to the effects of aggressive regularization techniques on an already under-fitting 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	$0.753 \pm 0.007$	$0.789 \pm 0.014$	$0.803 \pm 0.008$
	Yes	<b><math>0.756 \pm 0.011</math></b>	<b><math>0.793 \pm 0.006</math></b>	<b><math>0.809 \pm 0.002</math></b>
$\text{CLS}_{\text{cat}}$	No	<b><math>0.711 \pm 0.012</math></b>	$0.782 \pm 0.007$	$0.802 \pm 0.006$
	Yes	$0.709 \pm 0.005$	<b><math>0.796 \pm 0.001</math></b>	<b><math>0.811 \pm 0.008</math></b>
Concat	No	$0.799 \pm 0.002$	$0.799 \pm 0.009$	$0.803 \pm 0.003$
	Yes	<b><math>0.802 \pm 0.001</math></b>	<b><math>0.810 \pm 0.003</math></b>	<b><math>0.815 \pm 0.001</math></b>

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	$0.834 \pm 0.004$	$0.845 \pm 0.002$	$0.844 \pm 0.004$
	Yes	<b><math>0.835 \pm 0.003</math></b>	<b><math>0.849 \pm 0.001</math></b>	<b><math>0.848 \pm 0.002</math></b>
$\text{CLS}_{\text{cat}}$	No	<b><math>0.815 \pm 0.003</math></b>	$0.841 \pm 0.003$	$0.842 \pm 0.006$
	Yes	$0.814 \pm 0.003$	<b><math>0.844 \pm 0.001</math></b>	<b><math>0.846 \pm 0.001</math></b>
Concat	No	$0.842 \pm 0.003$	$0.844 \pm 0.003$	$0.843 \pm 0.004$
	Yes	<b><math>0.845 \pm 0.002</math></b>	<b><math>0.849 \pm 0.001</math></b>	<b><math>0.849 \pm 0.001</math></b>

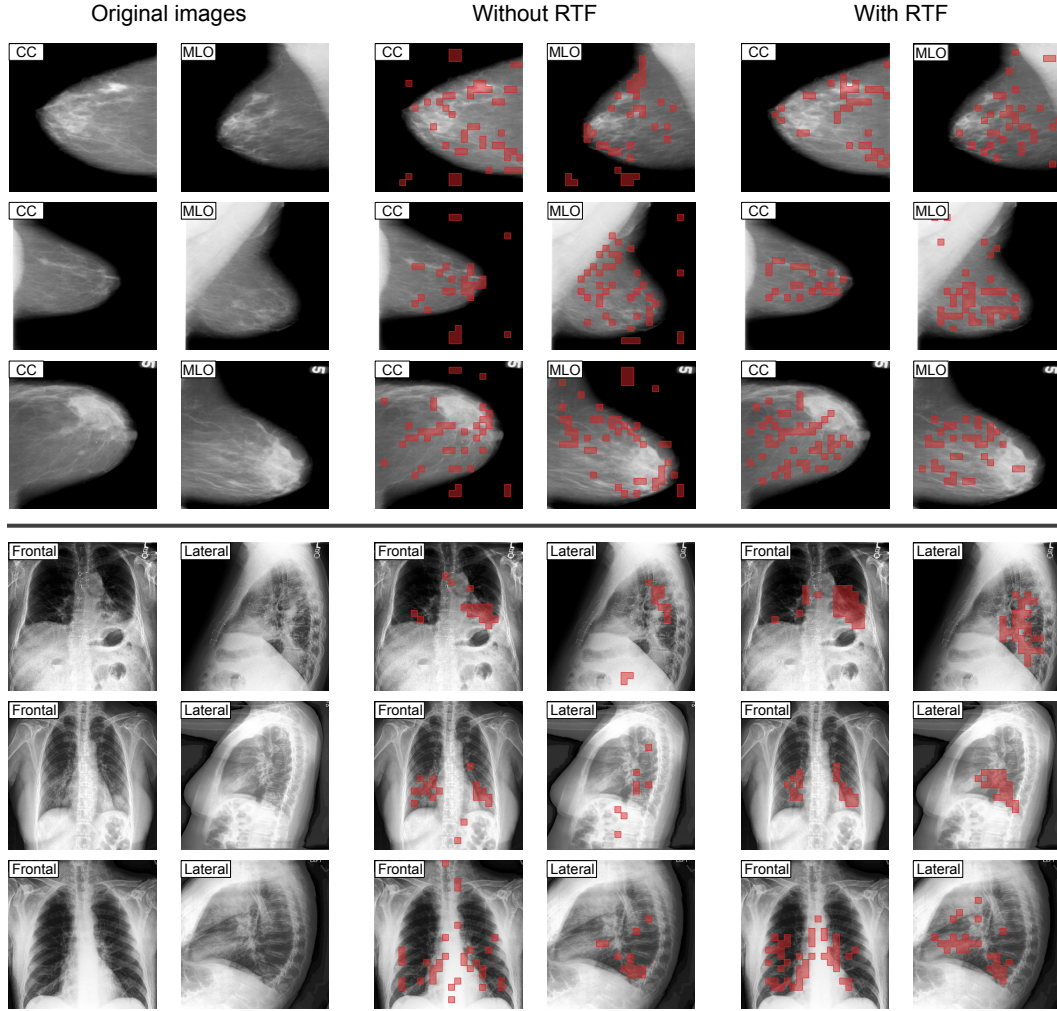


Figure 5: Extended results on CBIS-DDSM (**top**) and CheXpert (**bottom**), showing the model’s attention maps within the last block of the global encoder. RTF seems to address the issue of attention being allocated to uninformative areas, a common phenomenon observed in ViTs. It also encourages the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