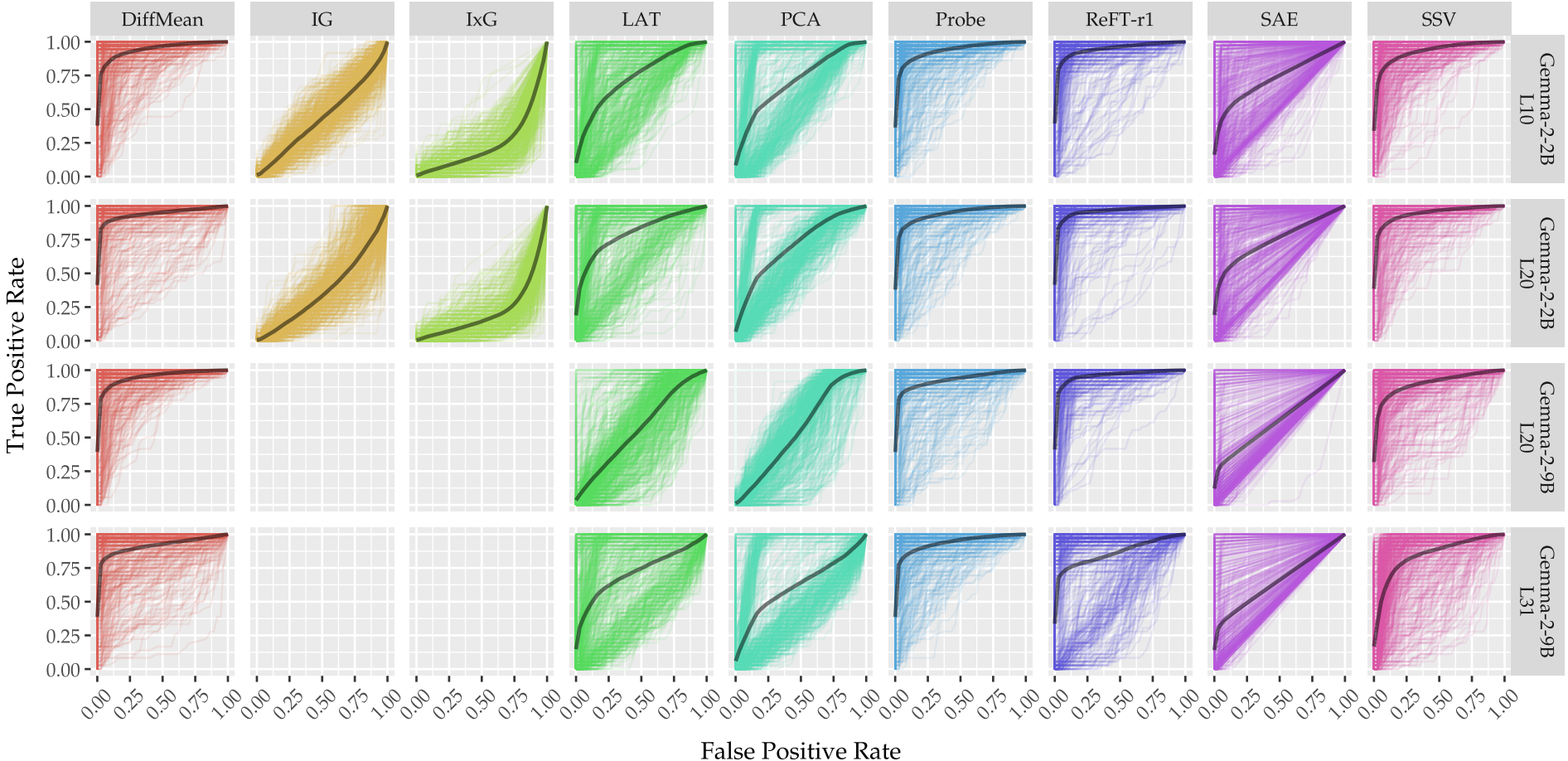
DiffMean
IG
IxG
LAT
PCA
Probe
ReFT-r1
SAE
SSV
Gemma-2-2B L10
Gemma-2-2B L20
Gemma-2-9B L20
Gemma-2-9B L31
True Positive Rate
False Positive Rate
0.00
0.25
0.50
0.75
1.00