Supplementary Materials

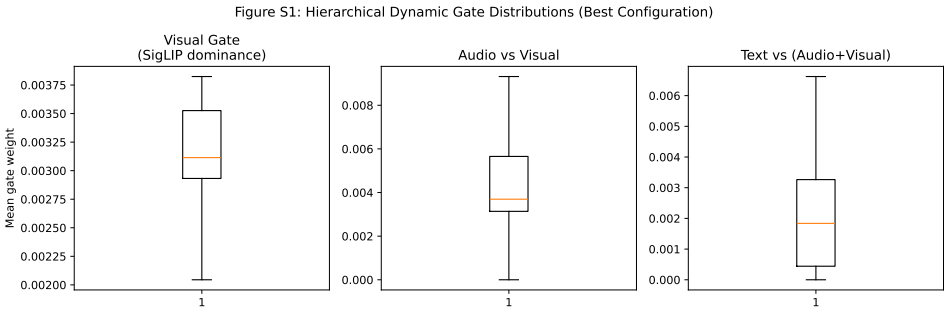


Figure S1: Hierarchical dynamic gate distributions under the best-performing cross-attention configuration. Left: intra-visual gate weighting between SigLIP semantic embeddings and head pose features. Center: audio-versus-visual cross-modal gate. Right: text-versus-(audio+visual) gate. Each point represents the session-level mean gate value averaged across utterances. The wide dispersion observed in cross-modal gates indicates subject-specific and context-dependent modality reweighting beyond static fusion assumptions.

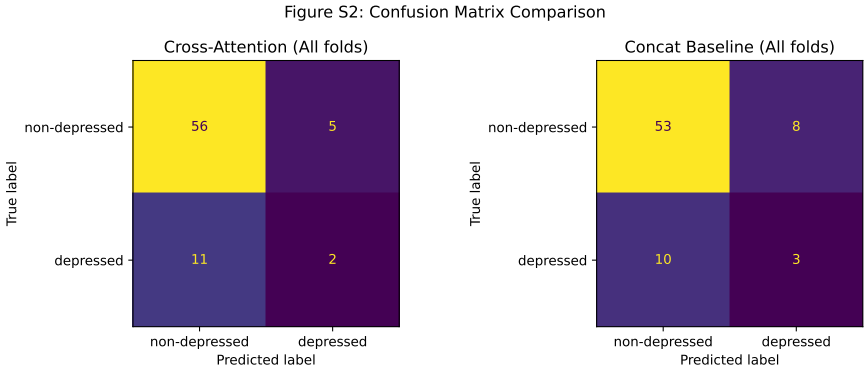


Figure S2: Confusion matrices aggregated across five folds for the best-performing cross-attention configuration (left) and the concatenation baseline (right). Cross-attention reduces false positives in the non-depressed group, while exhibiting a different error distribution for the depressed class. These differences reflect threshold-dependent trade-offs consistent with the PR-AUC analysis reported in Table 5.

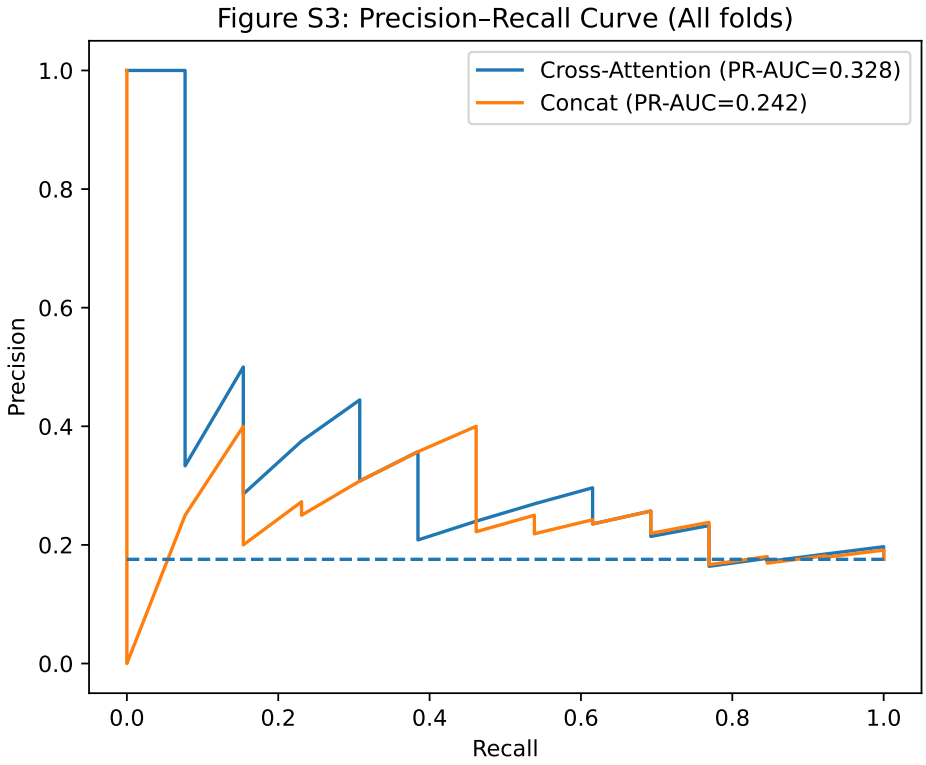


Figure S3: Precision–recall curves aggregated across five folds for the cross-attention model and the concatenation baseline. Cross-attention achieves a higher PR-AUC (0.328) than concatenation (0.242), indicating improved minority-class discrimination under class imbalance. The curve illustrates superior precision at moderate recall levels for the depressed class.