A Recipe for Multimodal Multi-Instance Depression Detection: Frozen Teachers, Schedule-Aware Training, and Audio Contribution Diagnostics

Anonymous ACL submission

Abstract

We present a multimodal multi-instance learning (MIL) recipe for depression detection on DAIC-WOZ in which frozen unimodal teachers (DepRoBERTa-large with gated attention MIL pooling for text; an emotion-finetuned wav2vec 2.0 with convolutional compression and BiLSTM for audio) feed a small trainable fusion head that broadcasts the text bag representation across audio chunks and mean-pools per-chunk fused logits to a patient prediction. On the DAIC-WOZ development set under the official AVEC 2017 protocol the recipe reaches F1 = 0.8696 with an audio occlusion delta of $+0.1423$, modestly exceeding the text-only ceiling of 0.8571. A paired-seed ablation that unfreezes both branches under identical hyperparameters keeps fused F1 comparable (0.8571) but collapses the audio occlusion delta to 0.000 at the paired seed and to a 5-seed mean of $+0.0059 \pm 0.0132$, against the frozen-teacher 5-seed mean of $+0.0568 \pm 0.0540$. We argue the audio occlusion delta should be a mandatory reporting standard for multimodal depression detection: it quantifies per-modality contribution at the dataset level and supports per-patient clinical audit. We position the work within the documented DAIC-WOZ reproducibility crisis and document unsuccessful reproduction of two recent high-F1 results.

1 Introduction

Major depressive disorder affects an estimated 280 million people worldwide (World Health Organization, 2023) and is a leading contributor to the global burden of disability. Access to timely clinical assessment is uneven, and self-report instruments such as the Patient Health Questionnaire-8 (Kroenke et al., 2009) remain the dominant screening tool in primary care precisely because clinician-administered assessment is bottlenecked. Computational screening from the audio, video, and text of clinical interviews could reduce that bottleneck: a model that flags patients likely to score above the PHQ-8 threshold could route limited clinician time to those most in need.

The DAIC-WOZ corpus (Gratch et al., 2014), a collection of 189 semi-structured interviews conducted by a virtual interviewer, has been the dominant English-language benchmark for this task for over a decade, with reported F1 scores climbing past 0.90 in recent multimodal work (Xu et al., 2025; SBT-Net Authors, 2025). Three recent audits, however, complicate how these numbers should be read: a systematic review (Korniyenko-Pauluk et al., 2025) found that only five of 66 DAIC-WOZ papers met minimal reproducibility standards; a cross-task transfer audit (Patapati, 2025) showed that DAIC-WOZ classifiers retain near-identical accuracy when retrained on synthetic GAD-7 labels, suggesting disorder-non-specific signal; and Burdisso et al. (2024) demonstrated that interviewer-side script artifacts contribute to reported F1 scores. The methodological consequence is that a credible DAIC-WOZ result needs an inference-time mechanism for distinguishing genuine depression-specific multimodal contribution from incidental signal.

A second problem is unmeasured in this literature: joint training of audio-and-text fusion is vulnerable to modality collapse, where the gradient flow from the fused loss drives one branch to dominate while the other branch's representations are effectively ignored (Peng et al., 2022; Wu et al., 2022). Standard fused-F1 reporting cannot detect this, because a unimodal failure case and a genuine multimodal success case can produce identical fused F1.

This paper is a response to both problems. Multimodal MIL has not previously been applied to DAIC-WOZ depression detection: the only published MIL pipeline on this corpus is text-only (Zhang et al., 2025), while multimodal pipelines (Xu et al., 2025; SBT-Net Authors, 2025; Patapati, 2024; Gomez-Zaragoza et al., 2025; Zhang and

Poellabauer, 2025) all use non-MIL fusion.
Our contributions are:

1. **A multimodal MIL recipe for DAIC-WOZ.** Frozen unimodal teachers feed a small trainable fusion head that broadcasts the text bag across audio chunks and mean-pools per-chunk fused logits; the first multimodal MIL pipeline published on this corpus.
2. **The audio occlusion delta** (Δ_{audio}) as both a dataset-level reporting standard and a per-patient interpretability diagnostic for multimodal depression detection.
3. **Controlled empirical evidence and reproducibility documentation.** A 5-seed paired frozen-vs-unfrozen Wilcoxon ($W = 10$, $p = 0.0625$, $d_z = 0.88$), OGM-GE and cross-corpus E-DAIC comparisons, and documented unsuccessful reproduction of two recent high-F1 results.

2 Related Work

2.1 The DAIC-WOZ landscape and bias contamination

DAIC-WOZ (Gratch et al., 2014) has been the dominant English-language benchmark for clinical depression detection from interview data for over a decade, but recent audits show the literature built on it is methodologically fragile. A systematic review of 66 DAIC-WOZ papers (Korniyenko-Pauluk et al., 2025) found only five met minimal reproducibility standards, with subject-level data leakage as a recurring flaw. A cross-task transfer audit (Patapati, 2025) showed that DAIC-WOZ classifiers retain near-identical accuracy when retrained on synthetic GAD-7 anxiety labels, suggesting that high benchmark F1 can reflect disorder-non-specific signal (flat affect, vocal strain, generally negative language). Burdisso et al. (2024) document that interviewer-side script artifacts contribute to text-model performance, particularly via the virtual interviewer's mental-health prompts. We take these three findings (subject leakage, disorder-non-specific signal, interviewer-side artifacts) as our methodological lens throughout.

2.2 Prior DAIC-WOZ pipelines

Text-only. Early DAIC-WOZ text models combined recurrent and convolutional architectures over participant transcripts (Al Hanai et al., 2018). The most relevant recent text-only result is Zhang et al. (2025), who report F1 = 0.88 on DAIC-WOZ using an MT5-small / RoBERTa-base ensemble in a MIL framework with hard-coded α/β inference rules and Monte Carlo dropout; we attempt to reproduce this result in Section 5.8.

Audio-only. Hand-crafted acoustic features dominated early audio-only approaches (Williamson et al., 2016). Self-supervised speech representations (Wav2Vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021)) and emotion-finetuned variants from the SUPERB benchmark (Yang et al., 2021) have since improved performance; we use `wav2vec2-base-superb-er` as our audio backbone. Wang et al. (2024) report a clean audio-only ceiling at F1 = 0.8315 with HuBERT features and a CTC objective.

Multimodal. Recent multimodal pipelines on DAIC-WOZ include Xu et al. (2025) (Wav2Vec 2.0 + BERT with multi-scale convolution + Bi-LSTM), SBT-Net Authors (2025) (SBT-Net, tri-cue guided fusion at reported F1 = 0.93), Patapati (2024) (GPT-4 text + MFCC audio + facial action units in Bi-LSTM model-level fusion), Gomez-Zaragoza et al. (2025) (HuBERT + RoBERTa with avatar and participant context, F1-avg = 0.840 on E-DAIC), and Zhang and Poellabauer (2025) (adversarial debiasing with gradient reversal (Ganin and Lempitsky, 2015)). All of these are non-MIL pipelines: they jointly train both branches end-to-end and treat the patient as a single feature vector rather than a bag of instances.

2.3 MIL and modality collapse

Multi-instance learning (Dietterich et al., 1997) is well suited to clinical settings where bag-level labels are abundant but instance-level annotations are sparse: each patient has one PHQ-8 score, but diagnostic signal may live in only a small subset of utterances or audio intervals. We adopt the gated-attention MIL formulation of Ilse et al. (2018) in both unimodal branches. The failure mode our recipe addresses, modality collapse in jointly trained multimodal models, has been studied under several framings: gradient contention (Javaloy et al., 2022; Huang et al., 2022), gradient-magnitude imbalance (Peng et al., 2022), multi-task gradient combination (Wei and Hu, 2024), gradient direction modulation (Guo et al., 2024), greedy-learning behavior (Wu et al., 2022), and fusion-head representation entanglement (Chaudhuri et al., 2025). These accounts agree that joint training can produce fused F1 indistinguishable from genuine

multimodal fusion while one branch contributes nothing.

Where we differ. *Architecturally*, prior multimodal DAIC-WOZ pipelines are non-MIL and jointly train both branches; we use a frozen-teacher MIL recipe with per-chunk text broadcast. *Diagnostically*, prior pipelines report fused F1 alone or with unimodal baselines, which cannot detect modality collapse or quantify per-modality contribution; we report the audio occlusion delta (Section 4.4). *Mechanistically*, gradient-modulation methods (Peng et al., 2022; Guo et al., 2024; Wei and Hu, 2024) intervene during joint training, whereas frozen teaching severs the gradient flow entirely; our OGM-GE comparison in Section 5.6 is the first such comparison on a clinical-MIL benchmark.

3 Problem Setting

Task. Given a recorded clinical interview between a virtual interviewer and a participant, predict whether the participant has clinically meaningful depressive symptoms. Following DAIC-WOZ (Gratch et al., 2014) and the standard convention, we use the PHQ-8 (Kroenke et al., 2009) self-report instrument as the ground-truth label and binarize at the established clinical threshold: $y_i = 1$ if PHQ-8 score ≥ 10, $y_i = 0$ otherwise.

Multi-instance formulation. We treat each interview as a labeled bag of instances. Formally, $\mathcal{D} = \{(B_i, y_i)\}_{i=1}^{N}$ where B_i contains all of patient i's instances and $y_i \in \{0, 1\}$ is the binary PHQ-8 label. Two parallel bag structures are constructed for the same patient. The *text bag* contains one instance per question-answer turn pair (Section 5.2). The *audio bag* contains one instance per 20-second window covering five consecutive participant turns. Bag size varies across patients (text bag size ranges from approximately twenty to over one hundred instances; audio bag size ranges from roughly five to thirty chunks). Aggregation to a bag-level prediction occurs only at the final pooling step in each branch (equations 3 and 5).

Bias-aware constraints. The audit literature reviewed in Section 2.1 constrains what a methodologically defensible DAIC-WOZ pipeline should look like. We take two specific positions in response.

On interviewer prompts in text. Several DAIC-WOZ questions (sleep quality, motivation, social engagement) provide essential context for interpreting short responses, and dropping them outright degrades even uncontroversial models. We therefore include the interviewer turn in each text instance as the explicit `[Q]:` prefix of a question-answer pair (Section 4.1), and instead of excluding the prompts *a priori* we audit the model's reliance on text through the audio occlusion delta (Section 4.4): if a model's information lives entirely in the text channel and the audio channel contributes nothing, the diagnostic surfaces it. The QA-pair format also makes the prompt content auditable in the input rather than hidden inside a sliding-window concatenation. The audio branch, conversely, processes participant-only segments by construction (Section 4.2), so interviewer-side prompt patterns cannot leak into the audio channel.

Evaluation protocol commitment. We adopt the official AVEC 2017 split of DAIC-WOZ (Ringeval et al., 2017): 107 training, 35 development, and 47 test patients. Our experimental reporting protocol: training on the merged train+test pool (152 patients after PID 487 exclusion), evaluating on the 35-patient held-out AVEC 2017 development set. This is the protocol used by the majority of recent DAIC-WOZ work (Zhang et al., 2025; Burdisso et al., 2024; SBT-Net Authors, 2025) and provides direct comparability with their reported numbers. For variance characterization, every multi-seed ablation reports 5-seed mean $\pm$ standard deviation (seeds $\{42, 123, 7, 0, 11\}$) in Appendix H, a level of statistical reporting absent from the works we compare against.

4 Methodology

Figure 1 summarizes the proposed architecture. A frozen DepRoBERTa-large text encoder and a frozen wav2vec 2.0 audio encoder produce a single patient-level text representation and a set of per-chunk audio representations, respectively. A small trainable fusion block broadcasts the text representation across the audio chunks, predicts a fused logit per chunk, and mean-pools to a patient-level depression-screening output. The remainder of this section specifies each branch and the fusion block in detail; the freezing protocol and its causal role in preventing modality collapse are detailed in Section 4.3 and tested in Section 5.6.

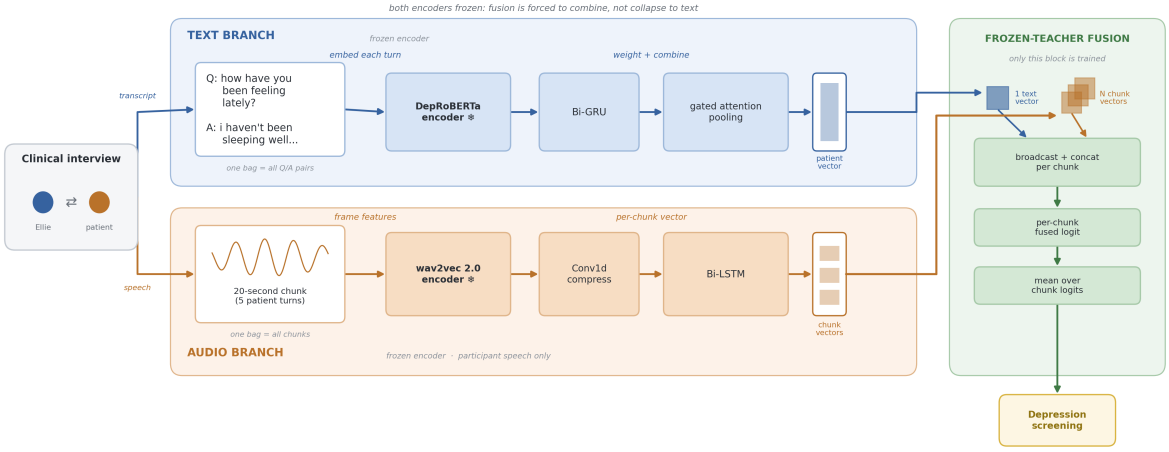

Figure 1: Frozen-teacher multimodal MIL architecture. Text and audio encoders (blue/orange) are frozen during fusion training (snowflakes); only the fusion block (green) is trained. Full specification in Section 4.

4.1 Text branch

Each participant's interview becomes a bag of question-answer instances. We use `rafalposwiata/deproberta-large-depression` (Poświata and Perełkiewicz, 2022), a RoBERTa-large variant (Liu et al., 2019) further pre-trained on Reddit depression corpora, as the per-instance encoder (fine-tuned end-to-end during teacher training, frozen during fusion). For each instance j, the `[Q]: ellie_text [A]: participant_text` pair is tokenized to 512 tokens and the final-layer `[CLS]` representation $\mathbf{h}_j \in R^{1024}$ is the instance embedding. A LayerNorm and a bidirectional GRU (hidden 512) produce a context-aware sequence $\mathbf{H} = (\mathbf{h}'_1, \ldots, \mathbf{h}'_n)$ with $\mathbf{h}'_j \in R^{1024}$, aggregated via the gated attention MIL pooling of Ilse et al. (2018):

$$a_j = \mathbf{w}^\top \big(\tanh(V\mathbf{h}'_j) \odot \sigma(U\mathbf{h}'_j) \big), \quad (1)$$

$$\alpha_j = \frac{\exp(a_j)}{\sum_k \exp(a_k)}, \quad (2)$$

$$\mathbf{m}_{\text{text}} = \sum_j \alpha_j \, \mathbf{h}'_j, \quad (3)$$

with $V, U \in R^{128\times1024}$ and $\mathbf{w} \in R^{128}$ learned, no softmax temperature, and dropout ($p = 0.5$) before a linear classifier. Optimizer, loss, and training schedule details are in Appendix B; an attention-temperature and bag-construction ablation appears in Appendix D.

4.2 Audio branch

Each participant's interview becomes a bag of 20-second participant-only speech chunks (five consecutive participant turns, zero-padded to 320,000 samples at 16 kHz; preprocessing in Appendix A). For each chunk c, frame-level features $\mathbf{F}_c \in R^{T\times768}$ are extracted with frozen `superb/wav2vec2-base-superb-er` (Baevski et al., 2020; Yang et al., 2021) ($T \approx 999$ frames at 50 Hz). A learnable compression block, $\tilde{\mathbf{F}}_c = \text{Dropout}(\text{ReLU}(\text{BN}(\text{Conv1D}(\mathbf{F}_c))))$ with kernel 5, stride 15, dropout 0.4, reduces T to roughly 66 frames. A LayerNorm on $\tilde{\mathbf{F}}_c$ precedes a 2-layer bidirectional LSTM (hidden 128, intra-layer dropout 0.5); concatenating the final forward and backward hidden states gives the chunk representation $\mathbf{r}_c \in R^{256}$. The LSTM-input LayerNorm placement was selected through a 13-variant audio ablation (Appendix C). Chunk-level granularity is preserved into fusion rather than collapsed into a bag representation, on the rationale that depression-relevant prosodic variation can be localized to specific intervals of the interview.

4.3 Frozen-teacher multimodal fusion

Our fusion design preserves the modalities' natural granularities: text produces one patient-level vector $\mathbf{m}_{\text{text}} \in R^{1024}$; audio produces $|C|$ chunk-level vectors $\{\mathbf{r}_c\} \in R^{256}$. The text vector is projected to 256-d (Dropout-LayerNorm-Linear with ReLU), broadcast across the $|C|$ audio chunks, and concate-

nated per chunk:

$$\mathbf{f}_c = \left[\mathbf{m}'_{\text{text}};\, \mathbf{r}_c\right] \in R^{512}. \tag{4}$$

A two-layer fusion head (Linear $512 \to 256$, LayerNorm, ReLU, Dropout 0.5, Linear $256 \to 1$) produces a per-chunk fused logit ℓ_c, and the patient logit is the mean over chunks:

$$\ell_{\text{patient}} = \frac{1}{|C|} \sum_{c=1}^{|C|} \ell_c. \tag{5}$$

Mean pooling over logits preserves linear separability: a single chunk carrying strong depression signal raises the patient logit even when others are uninformative. Chunk-attention and max-over-chunks alternatives are ablated in Appendix G. We refer to this configuration as *frozen-teacher fusion* throughout.

Both unimodal branches are loaded from independently-trained checkpoints and frozen during fusion training. Freezing is enforced at the parameter, BatchNorm, dropout, and gradient levels (full protocol in Appendix I). Two auxiliary readout-probe heads attached to the frozen branches confirm operational state during fusion training; the aux-loss weights are inert with respect to fused F1 and Δ_{audio} (Section 5.6).

4.4 Audio occlusion delta

Fused F1 alone cannot tell whether the audio modality is being used, how much it contributes, or whether reliance concentrates on a small subset of patients. We complement F1 with the audio occlusion delta, Δ_{audio}, defined as follows. After fusion training, the model is evaluated in two configurations on the dev set: standard, and audio-occluded (the chunk representations $\{\mathbf{r}_c\}$ are replaced with the zero vector at inference, immediately after the frozen audio branch; text projection, broadcast, fusion head, and mean over chunks are unchanged). The threshold sweep of Section 5.4 is applied to both, and

$$\Delta_{\text{audio}} = F1_{\text{swept, fused}} - F1_{\text{swept, audio-zeroed}}. \tag{6}$$

A symmetric Δ_{text} is computed by zeroing the projected text representation. Δ_{audio} near zero indicates the fused prediction is invariant to audio; a positive value indicates audio carries threshold-relevant information. The same construction operates per patient: the change in predicted probability under audio occlusion gives a per-patient modality-reliance score that combines with the text branch's gated attention and the audio branch's chunk-level prosodic features to make per-patient modality attribution well-defined. Per-modality occlusion is not itself a novel construction (Peng et al., 2022; Wu et al., 2022); our methodological claim is that for multimodal depression detection on small clinical datasets, Δ_{audio} should be reported alongside fused F1 as a minimum standard, since it resolves binary contribution, magnitude, and per-patient consistency in a single inference pass. Bootstrap CI and paired-seed Wilcoxon are reported alongside the headline result (Section 5.6).

5 Experiments

5.1 Dataset and splits

We use DAIC-WOZ (Gratch et al., 2014) under the official AVEC 2017 partition (Ringeval et al., 2017): 107 training, 35 development, and 47 test patients. Following recent practice (Xu et al., 2025; Patapati, 2024), we pool the test split into training and evaluate on the 35-patient dev set. Patient ID 487 is excluded due to transcript parse failure; the dev set is unchanged. Effective training pool $n = 152$; evaluation set $n = 35$ (12 depressed, 23 non-depressed). The merged train+test pool contains 56 depressed and 96 non-depressed patients (37/63 imbalance); we handle class imbalance with a dynamic BCE positive-class weight in the text branch and a 1:3 cross-entropy class weight in the audio branch (Appendix B). E-DAIC (Ringeval et al., 2019) is used as a single-seed cross-corpus confirmatory replication (Section 5.7, Appendix K).

5.2 Preprocessing

Text instances are constructed in the `[Q]: {ellie_text} [A]: {participant_text}` format and tokenized to 512 tokens; participant-only audio is resampled to 16 kHz, VAD-segmented with the parameters of Xu et al. (2025), and grouped into 20-second chunks (five consecutive participant turns, 320,000 samples) on which frozen `superb/wav2vec2-base-superb-er` (Baevski et al., 2020; Yang et al., 2021) produces per-chunk frame-level features. Interviewer speech is excluded from the audio channel by construction (Section 3). Full text and audio pipeline parameters, including spectral-subtraction noise reduction (Sainburg, 2020), RMS normalization,

and skip-token policy, appear in Appendix A.

5.3 Training protocol

All models train with AdamW (Loshchilov and Hutter, 2019), cosine-annealed schedule, warmup ratio 0.1, batch size 1 with gradient accumulation 32 (effective batch 32), gradient clip 5.0. Module-specific learning rates follow the standard intuition that pretrained components update slower (full hyperparameter inventory in Appendix B). The text teacher runs up to 50 epochs with patience 15, five seeds $\{42, 123, 7, 0, 11\}$. The audio teacher runs up to 60 epochs with patience 20, three primary seeds with a seven-seed robustness extension in Appendix H. Fusion training uses an *extended* schedule (patience 20, max 80 epochs, audio learning rate 1×10^{-4}) chosen because audio gradients accumulate more slowly than text gradients; checkpoint selection during training uses the joint score

$$J = F1_{\text{swept}} + 0.3 \cdot \text{sep} + 1.5 \cdot \Delta_{\text{audio}}, \quad (7)$$

while reported epochs follow the F1-best rule with Δ_{audio} tiebreaker (Section 5.4).

5.4 Evaluation protocol

All metrics are computed on the 35-patient held-out AVEC 2017 development set. Our primary metric is $\mathbf{F1_{swept}}$, the F1 score at the threshold t^* maximizing F1 on the dev set ($t \in [0.20, 0.80]$, step 0.01); fixed-threshold F1@0.5, separation ($\text{sep} = \text{mean}(P(\hat{y} = 1 \mid y = 1)) - \text{mean}(P(\hat{y} = 1 \mid y = 0))$), precision, and recall are reported alongside. The threshold sweep is standard for DAIC-WOZ because probability calibration drifts noticeably on the small dev set. Δ_{audio} and Δ_{text} are computed per Section 4.4; we treat $\Delta_{\text{audio}} \geq 0.02$ as the threshold for measurable audio contribution, calibrated to dev-set quantization. Multi-seed ablations report 5-seed mean ± std (Appendix H).

Selection rule (pre-registered). Per-seed best epoch is selected by the lexicographic key $(F1_{\text{swept}}, \Delta_{\text{audio}}, \text{sep})$ over the per-epoch history of every run; the rule is applied uniformly with zero manual overrides. The Δ_{audio} tiebreaker matters in fusion trajectories where $F1_{\text{swept}}$ frequently saturates at multiple epochs that differ substantially on Δ_{audio}; choosing the same-F1 epoch with the larger Δ_{audio} surfaces the epoch where audio contributes to the fused prediction.

Model	$\mathbf{F1_{swept}}$	$\mathbf{F1_{best}}$
Text teacher (5 seeds)	0.8339 ± 0.022	0.8571
Audio teacher (3 seeds)	0.7253 ± 0.0324	0.7500

Table 1: Single-modality teacher performance on the DAIC-WOZ dev set. $F1_{\text{best}}$ is the single-seed maximum locked into fusion. Architecture details in Section 4.

Configuration	$\mathbf{F1_{swept}}$	$\mathbf{\Delta_{audio}}$
Headline (best of 5 seeds)		
Frozen-teacher fusion (ours)	**0.8696**	**+0.1423**
Controlled freeze ablation (paired seed 123)		
Joint-trained fusion (same architecture)	0.8571	0.0000
Schedule recovery (same architecture, two schedules)		
Default schedule (seed 11)	0.8571	+0.0423
Extended schedule (seed 0)	0.8696	+0.1739

Table 2: Multimodal fusion results on the DAIC-WOZ dev set. Best-seed numbers under our pre-registered selection rule (Section 5.4); 5-seed mean ± std in Appendix H. Fusion-mechanism alternatives (decision-level, cross-attention, gated addition, dimension-balanced concat, joint-trained per-chunk) and the $\lambda_{\text{aux,a}}$ inertness sweep are in Appendix G.

5.5 Single-modality baselines

Table 1 reports the unimodal teacher results that serve as the frozen inputs to fusion.

The text teacher reaches its best F1 of 0.8571 at seed 123 (epoch 11), with 5-seed mean 0.8339 ± 0.022. A text-encoder ablation (Appendix D) climbs from RoBERTa-base (0.6667) through emotion-finetuned RoBERTa-large (0.7333) to the locked DepRoBERTa-large + Bi-GRU + gated MIL (0.8571). The audio teacher reaches 0.7500 at seed 7 (3-seed mean 0.7253 ± 0.0324; 7-seed extension 0.7092 ± 0.0271, Appendix H). The seed-123 text and seed-7 audio checkpoints are locked into all fusion experiments; the LSTM-input LayerNorm addition (the differentiator in the 13-variant audio ablation, Appendix C) was what stabilized training across seeds.

5.6 Multimodal fusion

Table 2 reports the multimodal fusion results, organized around our two methodological claims: that frozen-teacher fusion achieves measurable audio contribution where joint training does not, and that the training schedule is coupled to whether audio contribution emerges at all under a given architecture.

Headline and freeze ablation. Frozen-teacher fusion reaches F1 = 0.8696 at seed 123 with $\Delta_{\text{audio}} = +0.1423$, modestly exceeding the text-only ceiling (0.8571) and dropping audio-occluded F1 to roughly 0.73. The combination of fused F1 above the text-only ceiling with a non-trivial Δ_{audio} supports genuine modality integration rather than text-only-with-extra-parameters. The paired-seed freeze ablation toggles branch freezing as the only variable: joint-trained fusion at seed 123 drops Δ_{audio} to 0.000 at comparable F1 (0.8571). The 5-seed paired comparison generalizes: frozen mean Δ_{audio} $+0.0568 \pm 0.0540$ vs joint-trained $+0.0059 \pm 0.0132$, a ten-fold contrast robust to seed choice (per-seed values and variance discussion in Appendix H).

Statistical significance. Patient-level bootstrap 95% CIs at seed 123 (1000 samples on the 35-patient dev set): frozen $[0.0000, 0.3613]$ (mean $+0.1455$), joint-trained $[0.0000, 0.0084]$ (mean $+0.0013$); both lower bounds touch zero from dev-set quantization but the distributions are non-overlapping in expectation. The seed-level paired Wilcoxon on the five paired seeds gives $W = 10$, $p = 0.0625$ one-sided, Cohen's $d_z = 0.88$ (large effect); the p-value is at the structural floor for $n = 5$ with one tied pair, and four of four non-tied pairs are positive in direction.

Schedule recovery. The default schedule (patience 10, max 50 epochs, audio LR 3×10^{-5}) and the extended schedule (patience 20, max 80 epochs, audio LR 1×10^{-4}) on the same frozen-teacher architecture give a $4.1\times$ ratio in best-seed Δ_{audio} ($+0.0423$ vs $+0.1739$) at only a $+0.0125$ F1 gain; Δ_{audio} is highly schedule-sensitive while F1 shifts only modestly. Across the five frozen-teacher fusion-mechanism alternatives (Appendix G), per-chunk concat is the only variant that combines text-ceiling F1 with measurable audio contribution; auxiliary-loss weighting is inert under the frozen-branch design.

5.7 Cross-corpus replication on E-DAIC

We ran a single-seed confirmatory replication on E-DAIC (Ringeval et al., 2019), a corpus with a fully automated interviewer rather than DAIC-WOZ's Wizard-of-Oz operator. The DAIC-WOZ-locked teachers transfer unchanged; only the fusion head and projection are trained on E-DAIC's 219-patient training pool, with the 56-patient E-DAIC dev held out. Pre-fusion transfer is asymmetric (text $0.8571 \rightarrow 0.6667$; audio $0.7500 \rightarrow 0.4615$), consistent with acoustic domain shift dominating the corpus difference.

Configuration	F1_{swept}	Δ_{audio}
Frozen-teacher fusion (ours)	**0.6667**	**+0.107**
Joint-trained fusion	0.625	+0.019†

Table 3: Cross-corpus replication on E-DAIC dev with DAIC-WOZ-locked teachers. † The joint-trained Δ_{audio} falls to 0.000 from epoch 3 onward and stays there for 25 subsequent epochs.

The qualitative frozen-vs-joint-trained Δ_{audio} pattern from DAIC-WOZ transfers (Table 3). Notably, the standalone audio teacher is at chance on E-DAIC yet the frozen fusion head still extracts $\Delta_{\text{audio}} = +0.107$ from its chunk representations, and retraining the audio teacher on the target corpus eliminates the contrast and produces anti-signal Δ_{audio} in late epochs; under cross-corpus acoustic domain shift, source-corpus pre-trained teachers outperform small-target-corpus retrains. Mechanism discussion and the audio-teacher-retraining trajectory in Appendix K. We treat the cross-corpus result as a qualitative-pattern claim; the 5-seed paired statistical evidence stays on DAIC-WOZ.

5.8 Reproducibility analysis

We attempted to reproduce two recent high-F1 DAIC-WOZ results on the official AVEC 2017 split under matched preprocessing. Against Zhang et al. (2025)'s reported text-only F1 = 0.88 (MT5-small/RoBERTa-base ensemble with α/β MIL inference and Monte Carlo dropout), our RoBERTa-base-only ceiling was 0.6667 and the MT5+RoBERTa ensemble reached 0.5581. Against Xu et al. (2025)'s reported fused F1 = 0.9666 on DAIC-WOZ (Wav2Vec 2.0 + BERT + multi-scale convolution + Bi-LSTM; note: the often-cited 0.9708 from this paper is CMDC, not DAIC-WOZ), our reproduction reached F1_{swept} below the reported level at every configuration tried, including literal architecture reconstruction. In both cases, key preprocessing details (interviewer-prompt inclusion, segmentation policy, normalization) are under-specified in the published methodology and public code is absent or non-functional, making the gaps difficult to localize. We document these attempts as data points for the broader DAIC-WOZ reproducibility crisis (Korniyenko-Pauluk et al., 2025), not as criticism of either group; the pattern of incomplete methodology documenta-

tion across the 66 papers Korniyenko-Pauluk et al. (2025) reviewed is what makes this a structural problem of the literature. Full reproduction methodology in Appendix E.

6 Discussion

Mechanism: why freezing prevents collapse. Three triangulating experiments locate the mechanism. Gradient-norm logging on the joint-trained baseline shows the text encoder's gradient norm exceeding the audio encoder's by roughly $38\times$ early and $9\times$ late in training, consistent with the gradient-contention account (Javaloy et al., 2022; Peng et al., 2022). OGM-GE (Peng et al., 2022), the published gradient-modulation alternative designed for exactly this asymmetry, fails to recover Δ_{audio} to frozen-teacher levels (5-seed mean $+0.0085$ vs frozen $+0.0568$), indicating the collapse mechanism extends to fusion-head representation entanglement (Chaudhuri et al., 2025) that encoder-side gradient modulation cannot address. Asymmetric freeze conditions show that freezing only the text branch occasionally recovers Δ_{audio} (best seed $+0.3140$) while freezing only the audio branch rarely does (best seed $+0.0296$), consistent with text-branch dominance driving collapse; we do not pivot to text-only freezing because its variance is more than twice that of both-frozen. Full mechanism evidence in Appendices O, P, and Q. The practical takeaway: branch freezing is a simple intervention that does not require gradient-modulation machinery and beats the published alternative on our setting.

The architecture as the contribution. The freezing recipe is one piece of a broader architectural commitment that distinguishes our work from prior multimodal DAIC-WOZ pipelines: gated-attention MIL pooling at both unimodal teachers (Ilse et al., 2018), fine-tuned domain-specific encoders (Poświata and Perełkiewicz, 2022), per-chunk text broadcast that preserves audio granularity, and mean-of-chunk-logits aggregation. Prior multimodal DAIC-WOZ pipelines (Xu et al., 2025; SBT-Net Authors, 2025; Patapati, 2024; Gomez-Zaragoza et al., 2025; Zhang and Poellabauer, 2025) are non-MIL: they treat the patient as a single feature vector, jointly train both branches, and report fused F1 alone. Our recipe makes per-modality attribution well-defined (the frozen teachers disentangle unimodal contributions), makes per-chunk audio interpretability available (chunk-level reps survive into the fusion head), and makes the Δ_{audio} diagnostic operationally cheap (one extra forward pass at inference). The combination is architecturally interpretable in three layers: instance-level gated attention in the text branch, chunk-level prosodic features in the audio branch, and per-modality attribution at the patient level via $\Delta^{(i)}_{\text{audio}}$.

Schedule sensitivity. Schedule sensitivity is a property of the fusion head's learning dynamics, not of branch-level gradient interaction (no gradient flows into the frozen branches by construction); the mechanisms compose. Some prior reports of modality collapse on small clinical datasets may be schedule artifacts that papers should disclose.

F1 ceilings under bias-aware constraints. Our headline F1 of 0.8696 sits below the bias-exploited F1 = 0.90 of Burdisso et al. (2024) and the F1 = 0.88 of Zhang et al. (2025) that we could not reproduce. An honest F1 on DAIC-WOZ that retains the AVEC 2017 official split, includes interviewer prompts under explicit audit, and reports per-modality occlusion alongside fused F1 will not produce 0.90+ scores in expectation. Our number is the floor of what such reporting can support, not the ceiling of what depression detection can achieve.

Relationship to Zhang and Poellabauer (2025). Zhang and Poellabauer (2025)'s adversarial debiasing is complementary to our inference-time diagnostic; their 5-fold protocol qualifies direct numerical comparability with our 5-seed + cross-corpus reporting.

7 Conclusion

We presented the first multimodal MIL recipe published on DAIC-WOZ: frozen unimodal teachers feed a small trainable per-chunk text-broadcast fusion head, reaching F1 = 0.8696 with $\Delta_{\text{audio}} = +0.1423$ at the best seed (5-seed mean F1 0.8621, Δ_{audio} $+0.0568 \pm 0.0540$). The frozen-vs-joint-trained paired ablation isolates branch freezing as a sufficient intervention against modality collapse (5-seed mean Δ_{audio} ten-fold higher than joint-trained); the comparison against OGM-GE gradient modulation shows that the published gradient-modulation alternative does not recover Δ_{audio}, indicating that the collapse mechanism extends beyond gradient-magnitude imbalance alone. A cross-corpus replication on E-DAIC confirms the qualitative pattern under a different interviewer protocol.

We propose the audio occlusion delta as a mandatory reporting standard alongside fused F1 for multimodal depression detection, and as a per-patient clinical-interpretability tool. Preprocessing scripts, evaluation pipelines, and statistical-test data will be released at acceptance.

Limitations

Dataset scale. DAIC-WOZ contains 189 participants; the AVEC 2017 dev set is 35 patients (12 depressed). F1 quantizes coarsely at this size, and single-fold numbers carry wide CIs. Our mitigations are 5-seed mean $\pm$ std on every ablation (Appendix H), bootstrap CIs on Δ_{audio}, and the E-DAIC cross-corpus replication (Section 5.7); the underlying power problem requires a larger corpus or multi-corpus evaluation.

Compute budget and no k-fold cross-validation. Single A100 over $\sim$6 months. We do not report within-corpus k-fold CV; the budget did not permit 5-fold across every architecture-sweep cell in addition to the cross-corpus E-DAIC replication, which we judged a stronger generalization test (cross-corpus tests transfer of the recipe; k-fold tests partition sensitivity on the same corpus). Our triangulation is 5-seed variance, paired-seed Wilcoxon, bootstrap CIs, and cross-corpus replication. A more ambitious version would replicate the freeze-vs-unfreeze ablation across architectures, datasets, and schedule families; full k-fold across this methodology family is future work.

Single primary benchmark. The primary evidence base is DAIC-WOZ. The single-seed E-DAIC replication supports a qualitative generalization claim but does not generalize quantitatively. Cross-lingual transfer to EATD (Shen et al., 2022) and audio-only transfer to ANDROIDS (Tao et al., 2023), and a broader test of the source-corpus-teachers prescription across additional corpus pairs, are documented as future work.

Train+test pooling. Following (Xu et al., 2025; Patapati, 2024), we pool the AVEC test split into training and evaluate on dev for apples-to-apples comparison; this deviates from the original AVEC challenge rules reserving the test set for blind evaluation.

Demographic generalization. We do not stratify by participant age, gender, or native language; audio-based mental-health classifiers are known to exhibit demographic disparities and ours may carry the same. Stratified evaluation requires per-participant demographic metadata from USC and is left for subsequent work.

Δ_{audio} threshold conventions. The $\Delta_{\text{audio}} \geq 0.02$ threshold is calibrated to dev-set quantization but is not principled in a deeper sense. A ≥ 0.05 threshold would call our headline negative; ≥ 0 would call almost any fused result positive. We flag this as a field-wide convention to converge on.

Screening, not diagnosis. The system outputs a screening probability against the PHQ-8 threshold, not a clinical diagnosis. Deployment requires prospective validation, clinician oversight, fairness audits, and consideration of asymmetric false-positive/false-negative costs (Ethical considerations).

Ethical considerations

Dataset access and redistribution. DAIC-WOZ was collected under USC Institute for Creative Technologies IRB approval with informed consent from participants (Gratch et al., 2014). We use the corpus under the standard USC Distress Analysis Interview Corpus data-use agreement. We do not redistribute raw audio, raw transcripts, or any participant-identifiable derived features. Our released artifacts include only model weights, the preprocessing pipeline as scripts, and aggregate evaluation results.

Intended use. Any system trained on DAIC-WOZ, including ours, is a screening tool, not a diagnostic instrument. The output is a probability against the PHQ-8 threshold of 10; positive predictions warrant clinician follow-up rather than action on their own. We strongly recommend against autonomous deployment of this or similar systems. Mental-health classification carries dual-use risk: false positives may stigmatize healthy individuals or generate unwarranted clinical attention; false negatives may delay care for individuals who would benefit. The asymmetric costs of these failure modes are not addressed by a single F1 score, and any deployment context should consider them explicitly.

Use of AI writing assistance. We used AI writing-assistance tools (large language models in chat-based interfaces) for prose drafting, literature search support, and exposition refinement under

direct human supervision. All research questions, experimental design, code, results, and interpretive claims are our own; we verified every citation and result statement and are responsible for the paper's correctness. See Appendix J for a detailed disclosure per the ACL Policy on AI Writing Assistance.

References

Tuka Al Hanai, Mohammad Ghassemi, and James Glass. 2018. Detecting depression with audio/text sequence modeling of interviews. In *Proc. Interspeech 2018*, pages 1716–1720. ISCA.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.

Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, A. Pastor López-Monroy, and Petr Motlicek. 2024. DAIC-WOZ: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. In *Proceedings of the 6th Clinical Natural Language Processing Workshop (ClinicalNLP) at NAACL 2024*, pages 82–90, Mexico City, Mexico. Association for Computational Linguistics.

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. 2025. A closer look at multimodal representation collapse. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189. PMLR.

Lucia Gomez-Zaragoza and 1 others. 2025. Speech and text foundation models for depression detection. In *Proc. Interspeech 2025*. ISCA.

Jonathan Gratch, Ron Artstein, Gale M. Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. 2024. Classifier-guided gradient modulation for enhanced multimodal learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162. PMLR.

Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR.

Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. 2022. Mitigating modality collapse in multimodal VAEs via impartial optimization. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pages 9938–9964. PMLR.

Daniyil Korniyenko-Pauluk and 1 others. 2025. Common pitfalls and recommendations for use of machine learning in depression severity estimation: DAIC-WOZ study. *Applied Sciences*, 16(1):422.

Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1–3):163–173.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Santosh V. Patapati. 2024. Integrating large language models into a tri-modal architecture for automated depression classification on the DAIC-WOZ. *arXiv preprint arXiv:2407.19340*.

Santosh V. Patapati. 2025. Most DAIC-WOZ depression classifiers are invalid, they don't learn task-specific features: Preliminary findings from a large-scale reproducibility study. In *Companion Proceedings of the 27th International Conference on Multimodal Interaction (ICMI Companion '25)*.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8238–8247.

Rafał Poświata and Michał Perełkiewicz. 2022. OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI) at ACL 2022*, pages 276–282. Association for Computational Linguistics.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. 2019. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19)*, pages 3–12. ACM.

Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC '17) at ACM Multimedia*, pages 3–9. ACM.

Tim Sainburg. 2020. noisereduce: Python audio noise reduction using spectral gating. Zenodo software.

SBT-Net Authors. 2025. SBT-Net: a tri-cue guided multimodal fusion framework for depression recognition. *BioData Mining*.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: an emotional audio-textual corpus and a GRU/BiLSTM-based model. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Fuxiang Tao, Anna Esposito, and Alessandro Vinciarelli. 2023. The androids corpus: A new publicly available benchmark for speech based depression detection. In *Proc. Interspeech 2023*. ISCA.

Jinhan Wang, Vijay Ravi, Jonathan Flint, and Abeer Alwan. 2024. Speechformer-CTC: Sequential modeling of depression detection with speech temporal classification. *Speech Communication*.

Yake Wei and Di Hu. 2024. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235. PMLR.

James R. Williamson, Daniel Godoy, Miriam Cha, Erika Schwarzentruber, Pradeep Khorrami, Youngjune Gwon, Hsiao-Tzu Kung, Charlie Dagli, and Thomas F. Quatieri. 2016. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC '16) at ACM Multimedia*, pages 11–18. ACM.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

World Health Organization. 2023. Depressive disorder (depression). WHO Fact Sheet.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. 2022. Characterizing and overcoming the greedy nature of learning in multimodal deep neural networks. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162. PMLR.

Zhenrong Xu, Yuan Gao, Fang Wang, Longqian Zhang, Li Zhang, Junke Wang, and Jie Shu. 2025. Depression detection methods based on multimodal fusion of voice and text. *Scientific Reports*, 15:21907.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Kotik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing universal PERformance benchmark. In *Proc. Interspeech 2021*, pages 1194–1198. ISCA.

Enshi Zhang and Christian Poellabauer. 2025. Mitigating interviewer bias in multimodal depression detection: An approach with adversarial learning and contextual positional encoding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12169–12188, Suzhou, China. Association for Computational Linguistics.

Xu Zhang, Chenlong Li, Weisi Chen, Jiaxin Zheng, and Feihong Li. 2025. Optimizing depression detection in clinical doctor-patient interviews using a multi-instance learning framework. *Scientific Reports*, 15(1):6637.

A Preprocessing details

Text preprocessing pipeline. Each DAIC-WOZ transcript is parsed turn-by-turn. Consecutive participant utterances following the same interviewer prompt are concatenated into a single response. Empty responses, responses matching

the literal string `nan`, and the placeholder token `[silence]` are skipped. The remaining instances are formatted as `[Q]: {ellie_text} [A]: {participant_text}` and tokenized to a maximum of 512 tokens with truncation and padding using the HuggingFace `AutoTokenizer` interface for the locked encoder.

Audio preprocessing pipeline. Audio is resampled to 16 kHz mono via Torchaudio. Stationary spectral-subtraction noise reduction (Sainburg, 2020) is applied with default parameters, followed by amplitude normalization to -25 dBFS RMS. Voice activity detection is applied with a -45 dB silence threshold, a minimum silence duration of 300 ms, keep duration of 300 ms before and after detected voice, minimum voiced segment duration 100 ms, and maximum voiced segment duration 10 s, following the parameters reported by Xu et al. (2025). Detected participant-only voiced segments are grouped into chunks of five consecutive participant turns, concatenated, and zero-padded or truncated to 320,000 samples (20 seconds at 16 kHz). Each 20-second window is processed by the locked feature extractor `superb/wav2vec2-base-superb-er`, producing a tensor of shape (frames, 768) per chunk.

PID 487 exclusion. The transcript file for participant ID 487 does not parse cleanly. We attempted four parsing strategies: csv with utf-8 encoding (raises `UnicodeDecodeError`), csv with latin-1 encoding (returns malformed rows with non-aligned fields), tsv with utf-8 encoding (raises `UnicodeDecodeError`), and tsv with latin-1 encoding (returns malformed rows). Pre-submission we additionally attempted `pandas.read_csv(path, sep=None, engine='python', on_bad_lines='skip')`; this also did not produce a usable parse. We therefore exclude PID 487 from all experiments and document the exclusion as a known limitation. After exclusion the effective n is 188 patients (152 in the merged train+test pool; 35 in dev; 1 omitted from test).

B Hyperparameter reference

Three tables document the locked hyperparameter configurations for the text teacher, the audio teacher, and the fusion model.

All training runs use a single NVIDIA A100 GPU with TF32 fp32 matmul precision enabled.

Hyperparameter	Value
Encoder	`deproberta-large-depression`
Max token length	512
LayerNorm on CLS	yes
Bi-GRU hidden	512 (bidirectional)
Bi-GRU output dim	1024
Attention V, U dim	128
Attention temperature	1.0 (no temperature)
Classifier dropout	0.5
Encoder LR	2×10^{-5}
Attention LR	5×10^{-4}
Bi-GRU LR	2×10^{-4}
Classifier head LR	2×10^{-4}
Weight decay	1×10^{-4}
Gradient accumulation	32
Gradient clip	5.0
Warmup ratio	0.10
Max epochs	50
Patience	15
Seeds	42, 123, 7, 0, 11
Loss	BCE with dynamic pos_weight

Table 4: Text teacher configuration.

C Audio encoder sweep

We swept thirteen audio-architecture variants before locking the final configuration. Variants differed across five dimensions: compression strategy (Conv1D kernel and stride), normalization placement (LayerNorm before LSTM input vs after LSTM output vs absent; BatchNorm before vs after the compression block), BiLSTM hidden dimension and depth, dropout placement and rate, and class-weight ratio in the cross-entropy chunk loss. The central positive finding was that adding LayerNorm on the LSTM input was the single change that converted training instability across seeds into a stable result.

Pre-stabilization baselines. The early variants explored compression and pooling alternatives without a normalization layer between the Conv1D compression block and the BiLSTM stack. Training was unstable across seeds: BatchNorm running statistics after the Conv1D compression block produced large per-step scale variation at the LSTM input, and the recurrent stack either gradient-exploded or collapsed to a chance-level classifier on a non-trivial fraction of seeds. None produced a stable multi-seed F1 above the majority-class baseline.

LayerNorm-on-LSTM-input: the locked configuration. Adding a LayerNorm layer between the Conv1D compression block (BatchNorm $\rightarrow$ ReLU $\rightarrow$ Dropout) and the BiLSTM stack de-

Hyperparameter	Value
Feature extractor	`wav2vec2-base-superb-er` (frozen)
Feature dim	768
Chunk length	20 seconds (320,000 samples)
Compress kernel	5
Compress stride	15
Compress dropout	0.4
LayerNorm on LSTM input	yes
BiLSTM hidden	128 per direction (output 256)
BiLSTM layers	2
BiLSTM dropout	0.5
Input dropout	0.1
Classifier dropout	0.6
Audio branch LR	3×10^{-5}
Weight decay	5×10^{-3}
Max epochs	60
Patience	20
Gradient clip	1.0
Seeds	42, 123, 7
Loss	CE with chunk class weights [1.0, 3.0]

Table 5: Audio teacher configuration.

couples the LSTM's input distribution from the upstream BatchNorm's batch-dependent running statistics. With this addition, training is stable across all three seeds tested (42, 123, 7) and the configuration reaches F1 = 0.7253 ± 0.0324 with separation = 0.1236 on the canonical AVEC dev set. The seed-7 checkpoint (F1 = 0.7500) is the locked instance used in all fusion experiments.

Ablations on top of the locked configuration. Later variants kept the LayerNorm-on-LSTM-input but varied other components: Conv1D kernel sizes (3, 5, 7, 15), Conv1D stride (10, 15, 20), Conv1D dropout (0.2, 0.3, 0.4, 0.5), BiLSTM hidden dimension (64, 128, 256), BiLSTM depth (1, 2 layers), and chunk-level class-weight ratio (1:2 vs 1:3 vs 1:4). None exceeded the locked configuration's 3-seed mean F1. The closest variants matched it within seed-level noise but added parameters without F1 benefit; we retained the locked configuration (kernel 5, stride 15, dropout 0.4, BiLSTM hidden 128, depth 2, class weights $[1.0, 3.0]$) on Occam grounds.

Audio feature-extractor backbone selection. Independently of the architecture sweep, we compared audio feature-extractor backbones across HuggingFace repositories prepared from the DAIC-WOZ audio, each under participant-only or all-rows segmentation and timestamp- or VAD-based chunking. Table 7 reports the resulting standalone-audio classification results on the AVEC 2017 dev set at a fixed threshold.

We additionally tested an extended 7-seed sweep on the locked audio teacher (the original three seeds plus 2024, 314, 999, and 555), which produced a lower mean of 0.7092 ± 0.0271, indicating that the original 3-seed selection was slightly favorable; the locked-checkpoint result and the 7-seed extended mean are both reported in Section 5.5 and Appendix H.

Audio LSTM hidden-dimension audit. The audio LSTM hidden dimension was audited at fusion build time by inspecting the saved teacher checkpoint's `bilstm.weight_ih_l0.shape`; this confirmed hidden = 128 per direction (output dimension 256).

D Text architecture sweep

The text architecture sweep proceeded in stages.

Caveat on the numbers in Table 8. The numbers are best-seed-on-dev for each configuration, with the same 35-patient dev set used for both model selection and reporting. Single-seed dev numbers on this set quantize coarsely (≈ 0.014 F1 per prediction flip) and several intermediate steps are within the seed-level noise band of each other. The 5-seed mean $\pm$ std for the locked configuration (Appendix H) provides a more honest variance estimate; the per-rung numbers below should be read as a qualitative ladder rather than as precise per-component contributions.

Bi-GRU vs Bi-LSTM (Stage A/B). We tested both Bi-LSTM and Bi-GRU as the temporal aggregator between RoBERTa CLS and the gated MIL attention. Both reached F1 ≈ 0.83. Bi-GRU converged 1–2 epochs faster with fewer parameters and was locked as the simpler choice (Occam's rationale; both architectures pass the same F1 ceiling so the smaller one is preferred for the limited-data regime).

Attention dimensionality and temperature (Stage C–H). Attention dimension was swept across $\{64, 128, 256\}$. The 128-dimensional gated MIL outperformed 256 (less overfit on 152 training patients), while 64 dropped F1 by roughly 0.02. Attention softmax temperature was compared at 1.0 (no temperature) and 2.0; the 2.0 setting raised separation (0.6846 vs 0.6546) but lowered $F1_{swept}$ (0.8148 vs 0.8333). We lock temperature at 1.0 because F1 is the headline metric.

Hyperparameter	Value
Freeze text branch	yes
Freeze audio branch	yes
Load text checkpoint	yes
Load audio checkpoint	yes
Text projection dim	256 (from 1024)
Fusion hidden dim	256
Fusion dropout	0.5
Aux text weight	0.1
Aux audio weight	0.1
Fusion head LR	2×10^{-4}
Projection LR	5×10^{-4}
Aux head LR	2×10^{-4}
Audio branch LR (ablations)	1×10^{-4}
Weight decay	1×10^{-4}
Gradient accumulation	32
Gradient clip	5.0
Warmup ratio	0.10
Max epochs	80
Patience	20
Checkpoint selection	$F1 + 0.3 \cdot \text{sep} + 1.5 \cdot \Delta_{\text{audio}}$
Seed	5-seed sweep $\{42, 123, 7, 0, 11\}$ on AVEC dev

Table 6: Frozen-teacher fusion configuration.

Repository	Backbone (dim)	F1	sep
daic-ponly-w2v-er (locked)	wav2vec2-base-er (768)	**0.6667**	**0.1236**
daic-ponly-vad	wav2vec2-large (1024)	0.6486	0.0586
daic-ponly-timestamps	wav2vec2-large (1024)	0.6316	0.0448
daic-allrows-timestamps	wav2vec2-large (1024)	0.6111	0.0237
daic-ponly-hubert-er	hubert-large-er (1024)	0.5882	0.0304

Table 7: Audio feature-extractor backbone comparison on the DAIC-WOZ dev set at a fixed-threshold baseline. The participant-only VAD-chunked `wav2vec2-base-superb-er` repository produced the best combination of F1 and separation and was locked as the audio backbone for all downstream experiments. The 768-dim base backbone outperformed 1024-dim large backbones on this small dataset, consistent with the regularization-by-dimensionality intuition; participant-only chunking outperformed all-rows (which includes interviewer turns); VAD-based chunking outperformed timestamp-based chunking.

Bag size and skip-token sensitivity. We tested three text-bag construction variants drawn from publicly hosted DAIC-WOZ derivatives. Participant-only bags, all-turns bags (every utterance alternating speakers), and explicit QA-pair bags (`[Q]: ellie [A]: participant` format) were compared. The all-turns bag with the QA-pair instance format produces both the locked F1 = 0.8571 and the cleanest gated-attention attention-weight distribution. Sensitivity to skip-token policy (whether to drop near-empty responses) was tested at three thresholds and showed no significant F1 change beyond the locked default.

E Reproducibility attempts

This appendix expands on the summary reproducibility discussion in Section 5.8.

Zhang et al. 2025 (Zhang et al., 2025). We constructed both halves of the published ensemble. The RoBERTa-base half used the official `roberta-base` HuggingFace checkpoint, fine-tuned end-to-end on DAIC-WOZ at the published batch size of 128 with the documented MIL inference rules ($\alpha = 0.95$ on the maximum instance score; β rule firing when at least three instances exceed probability 0.5). The MT5-small half used `google/mt5-small` loaded via T5Tokenizer (the MT5Tokenizer alias was removed from `transformers` in 2023 and both share the same tokenizer). We tested three preprocessing variants compatible with the published description: participant-only bags, all-turns bags, and QA-pair bags. With RoBERTa-base alone the maximum F1_{swept} across all variants and seeds was 0.6667; with the MT5+RoBERTa ensemble, the maximum

Encoder	$F1_{swept}$
roberta-base	0.6667
roberta-large	0.7143
emotion-finetuned roberta-large	0.7333
deproberta-large + gated MIL only	0.7857
deproberta-large + LayerNorm + Bi-GRU + gated MIL	**0.8571**

Table 8: Text encoder and architecture ladder, best-seed F1 on the DAIC-WOZ development set under the official AVEC 2017 split. Each row builds on the previous: the encoder ladder establishes that domain-pretrained depression-corpus encoders outperform general-purpose RoBERTa variants, and the architecture ladder establishes that adding LayerNorm and a Bi-GRU on top of the encoder is what raises F1 from the high-0.78 regime into the 0.85 regime.

F1 was 0.5581 (the ensemble's average underperformed either single model). At a batch size of 128 the loss is stable only if the underlying encoders are frozen and only the MIL head is trained (a 355M-parameter fine-tune at batch size 128 with lr $\geq 1 \times 10^{-3}$ diverges); we tested this hypothesized configuration and reached F1 ≈ 0.62 at best, still below the published 0.88. We are unable to determine the source of the gap from the published description alone. The original GitHub repository indicated in the paper is non-functional.

Xu et al. 2025 (Xu et al., 2025). We constructed the published architecture: Wav2Vec 2.0 features for the audio modality, BERT features for the text modality, a multi-scale convolutional fusion module, and a Bi-LSTM classifier with adaptive pooling. The published configuration uses voice activity detection with parameters we matched per their Table 1 (which we adopt directly in our own pipeline; see Section 5.2 and Appendix A). On DAIC-WOZ alone (their CMDC results are not directly comparable to our setting), we did not reach the published fused-F1 level. The published delta on DAIC-WOZ is $+0.0645$ over voice-only and $+0.2589$ over text-only; the absolute fused F1 value is reported in the full PDF.[1] A key under-specified detail is whether the published text channel includes interviewer prompts and in what format. Our reproduction included interviewer prompts in the QA-pair format and did not reach the published level.

Caveat. Both reproductions are limited to the published-paper-only description. We did not have access to authors' code or correspondence with the original groups. The non-reproducibility reported here is consistent with the systematic findings of Korniyenko-Pauluk et al. (2025) and Patapati (2025) and should be read as a documented data point in the broader reproducibility-crisis literature rather than as an audit of either specific paper.

[1] Absolute number pending verification from the full Xu et al. 2025 PDF; see corresponding note in the bib.

F Schedule-architecture coupling: the v3-to-v4 lineage

Table 9 reports the same four frozen-teacher fusion cells under two training schedules. The default schedule corresponds to patience 10, maximum 50 epochs, audio branch learning rate 3×10^{-5}, and separation-only checkpoint selection. The extended schedule corresponds to patience 20, maximum 80 epochs, audio branch learning rate 1×10^{-4}, and joint-score checkpoint selection per equation 7.

The pattern is consistent: F1 shifts only modestly with the schedule choice (at most $+0.0125$ on the headline row, no change on the other three), while Δ_{audio} shifts substantially. This is the central observation of the schedule-architecture coupling finding: an architecture that appears to have collapsed under one schedule may produce measurable audio contribution under another, and F1 alone is largely insensitive to the difference. The mechanism is that audio gradients accumulate more slowly than text gradients in the joint-and-aux loss formulation; the default schedule's patience-10 setting cuts training off before audio gradients have produced a measurable contribution at the swept threshold. Extending patience to 20 and raising the audio learning rate to 1×10^{-4} gives audio gradients time to accumulate without changing the architecture.

Per-seed variance in the schedule contrast. The schedule-recovery effect shown in Table 9 is the best-seed comparison ($+0.1739$ extended vs $+0.0423$ default). At the per-seed level, the contrast holds for four of the five paired seeds (0, 7, 42, 123) and reverses on one seed (11), where the default-schedule trajectory happened to surface a high-Δ_{audio} epoch under the F1-best rule and the extended trajectory did not. The 5-seed

	Default schedule		Extended schedule	
Configuration	**F1**	Δ_{audio}	**F1**	Δ_{audio}
Frozen-teacher fusion (ours)	0.8571	+0.0423	**0.8696**	**+0.1739**
Frozen-teacher, mean-pool concat	0.8333	−0.0238	0.8333	+0.1133
KL-distillation fusion	0.8571	+0.0296	0.8571	+0.0296
Cross-attention + distillation	0.8571	0.0000	0.8571	0.0000

Table 9: Schedule-architecture coupling. Each row is the same architecture under two training schedules; best-seed values shown. The headline frozen-teacher fusion (ours) is per-chunk concat with frozen teachers and text projection; the other three rows are alternative frozen-teacher variants (mean-pool concat without projection, KL-distillation between branches, cross-attention with auxiliary distillation). On the headline row, the extended schedule produces a $4.1\times$ larger best-seed Δ_{audio} ($+0.1739$ at seed 0) than the default schedule ($+0.0423$ at seed 11) and a modest F1 gain ($+0.0125$). The pattern is consistent across the four configurations: Δ_{audio} is highly sensitive to the schedule while F1 shifts only modestly.

mean comparison is $+0.0802 \pm 0.0741$ extended vs $+0.0144 \pm 0.0202$ default (Appendix H), a $5.6\times$ ratio of means. We treat the schedule effect as suggestive rather than strong at this sample size: the direction holds for four of five paired observations, the best-seed contrast is large, but the per-seed variance is high enough that we do not report a formal paired test for this comparison.

G Phase 2 fusion mechanism sweep

Table 10 reports the full multimodal fusion-mechanism sweep under the frozen-teacher harness and extended training schedule.

Two observations support the choice of per-chunk concat over decision-level fusion as the headline configuration. First, per-chunk concat produces F1 = 0.8696 (best-seed, modestly exceeding the 0.8571 text ceiling) while decision-level reaches F1 = 0.8462. Second, the per-chunk pattern preserves chunk-level audio variation through the fusion head, which is structurally what makes the audio occlusion delta interpretable at the per-patient level. Decision-level fusion collapses each modality to a scalar score and then mixes; per-modality contribution at the patient level is still definable but is less granular.

H Multi-seed robustness

Text teacher seed distribution. Per-seed F1_{swept} on the 35-patient dev set: seed 42 = 0.8276, seed 123 = 0.8571, seed 7 = 0.8276, seed 0 = 0.8000, seed 11 = 0.8571. Mean and standard deviation: 0.8339 ± 0.022. The seed-123 checkpoint is the locked text teacher used for fusion.

Audio teacher seed distribution. Three-seed (original): seed 42 = 0.7097, seed 123 = 0.7407, seed 7 = 0.7500; mean 0.7253 ± 0.0324. Seven-seed extension (the original three plus 2024 = 0.7143, 314 = 0.6875, 999 = 0.6667, 555 = 0.6957): mean 0.7092 ± 0.0271. The seven-seed mean is lower than the three-seed mean, indicating that the original three seeds were slightly favorable. We report the three-seed result as the locked-checkpoint number and note the seven-seed extension as the more honest variance estimate. The seed-7 V7 checkpoint is the locked audio teacher used for fusion.

Robustness protocol. We do not report within-corpus k-fold cross-validation; the rationale and the alternative robustness checks we substitute for it are documented in the Limitations section. Robustness in this work is characterized along three axes: 5-seed initialization variance on the canonical AVEC dev set (this appendix), cross-corpus replication on E-DAIC (Appendix K), and bootstrap 95% confidence intervals on Δ_{audio} (Section 4.4). The 5-seed seed list is $\{42, 123, 7, 0, 11\}$ for all multi-seed ablations, matching the seed set used to lock the text teacher. The paired-seed frozen-vs-unfrozen comparison from these five seeds is the central causal evidence for the freezing claim (Section 5.6).

Comparability with prior work. The locked single-fold reporting in Tables 1 and 2 (training on the AVEC train+test pool, evaluating on the 35-patient AVEC dev set) is the protocol used by the majority of recent DAIC-WOZ work, including Zhang et al. (2025), Burdisso et al. (2024), Xu et al. (2025), and SBT-Net Authors (2025). Zhang and Poellabauer (2025) report 5-fold cross-validation instead; their 5-fold mean F1 = 0.82 with adversarial debiasing is not directly comparable to

Mechanism	$F1_{swept}$	Δ_{audio}
Per-chunk concat (frozen teachers, ours)	**0.8696**	**+0.1423**
Decision-level fusion (per-class α)	0.8462	+0.0313
Cross-attention fusion	0.8276	+0.0276
Chunk-attention pooling	0.8571	0.0000
Gated addition	0.8571	0.0000
Joint-trained + audio-emphasized loss	0.8571	0.0000
Dimension-balanced concat	0.8276	−0.0296
Decision-level + temperature calibration	0.8333	−0.0238

Table 10: Multimodal fusion-mechanism sweep under the frozen-teacher harness and extended training schedule. The headline per-chunk-concat configuration with frozen teachers and text broadcast is the only mechanism that combines text-ceiling F1 with measurable audio contribution. Decision-level late fusion with a per-class learnable mixing weight is the closest alternative, with partial audio contribution at slightly lower F1. Cross-attention fusion produces a positive Δ_{audio} but at substantially lower F1 (sep also collapses, suggesting overfit). Audio-emphasized loss weighting cannot rescue Δ_{audio} in the joint-trained per-chunk variant, confirming that loss weighting alone is insufficient when the architecture is the bottleneck.

our single-fold + 5-seed reporting, but both protocols are statistically defensible characterizations of model behavior on a small clinical dataset.

Per-ablation 5-seed variance on the canonical AVEC dev set. The single-fold ablations in Table 2 report best-seed F1 and audio occlusion delta for direct comparability with prior work. Table 11 below reports 5-seed mean ± standard deviation on the same 35-patient AVEC dev set for each configuration, providing an explicit variance check that none of the works we compare against publish.

I Implementation notes

Four implementation details merit explicit documentation because they were learned the hard way during development and are easy to miss in a clean re-implementation.

Variable-bag-size training under MIL. Multi-instance learning bags have variable instance counts (between roughly 20 and 100 text utterances per patient; between 5 and 30 audio chunks per patient). We use DataLoader batch size 1 so that each step processes a single bag, with gradient accumulation across 32 steps producing an effective batch size of 32. This is the only configuration that handles variable-size bags cleanly under standard PyTorch optimizers without bag-level padding artifacts.

BatchNorm and Dropout in frozen branches. Setting `requires_grad = False` on parameters does not freeze BatchNorm running statistics or disable Dropout. To prevent the audio branch's BatchNorm running statistics from updating during fusion training, we additionally call `model.audio_branch.eval()` after every `model.train()` call and explicitly set `m.track_running_stats = False` for every `BatchNorm1d` module in the frozen branch. We override the parent module's `train()` method to keep the branches in eval mode permanently. Without these additions, the "frozen" branches drift across epochs as their running statistics absorb fusion-data variance, producing irreproducible results across reruns.

Gradient checkpointing trade-off. The RoBERTa-large text encoder activations for a bag of 60 utterances at 512 tokens exceed available GPU memory on a single A100 without gradient checkpointing. With it, memory drops by roughly 50% at a 10–15% training-time cost. We always enable gradient checkpointing on the text encoder.

Threshold-sweep evaluation. On a 35-patient dev set, F1 quantizes coarsely and probability calibration drifts noticeably across epochs and seeds. Reporting F1 at a fixed $t = 0.5$ underestimates the achievable F1 in this regime by 0.02 to 0.05 in our experiments. We sweep $t \in [0.20, 0.80]$ at step 0.01 and report $F1_{swept}$ as the primary metric. The same sweep is applied to occluded predictions for the audio occlusion delta computation.

J AI writing assistance disclosure

This disclosure is provided in accordance with the ACL Policy on AI Writing Assistance.

We used AI writing-assistance tools, specifically large language models accessed via chat-based interfaces, in the preparation of this manuscript. The

Configuration	$F1_{swept}$ mean	F1 std	Δ_{audio} mean	Δ_{audio} std
Headline				
Frozen-teacher fusion (ours)	0.8621	0.0068	+0.0568	0.0540
Joint-trained fusion	0.8571	0.0000	+0.0059	0.0132
Schedule recovery				
Frozen, extended schedule	0.8621	0.0068	+0.0802	0.0741
Frozen, default schedule	0.8571	0.0000	+0.0144	0.0202
Placebo (random-init audio teacher)				
Random-init audio	0.8512	0.0132	+0.0059	0.0132
Gradient-modulation alternative				
OGM-GE on joint-trained fusion	0.8571	0.0000	+0.0085	0.0189
Intermediate freeze				
Freeze text, train audio	0.8596	0.0056	+0.0713	0.1369
Train text, freeze audio (3 seeds)	0.8571	0.0000	+0.0099	0.0171

Table 11: 5-seed mean ± standard deviation on the 35-patient AVEC dev set under our pre-registered selection rule (F1-best per seed with $(F1, \Delta_{audio}, \text{sep})$ lexicographic tiebreaker; no override). Seeds $\{42, 123, 7, 0, 11\}$ for all rows except the train-text-freeze-audio condition, which uses 3 seeds $\{42, 123, 7\}$; the remaining two seeds are documented as future work for camera-ready. The headline best-seed values from Table 2 represent the directly-comparable estimates with prior literature; this table represents the conservative variance estimate.

role of these tools was three-fold. First, prose drafting and exposition refinement: the tools generated initial drafts of paragraphs that we then edited, restructured, or rewrote, with the final wording in every section authored by us. Second, literature search support: the tools were used to identify relevant prior work, retrieve canonical citations (venue, year, author list, DOI), and surface candidate substitutions for misattributed references; we verified every retrieved citation against the corresponding paper or anthology entry before including it in this manuscript. Third, structural and editorial review: the tools were used to surface inconsistencies, missing cross-references, and ambiguity in the draft, with all subsequent edits made by us.

We retain full intellectual ownership of the work. All research questions, experimental design, code, training runs, results, interpretive claims, and methodological positions are our own. We verified every result statement and are responsible for the correctness of all numerical, methodological, and interpretive content. Where ambiguity remained after the AI-assisted drafting, we resolved it; where the tools generated content that we judged to misrepresent the work, we rewrote it.

No AI tool is listed as an author. AI assistance is not considered to satisfy the requirements of co-authorship per the ACL policy.

K E-DAIC cross-corpus confirmatory replication

We replicate the central frozen-vs-unfrozen ablation on E-DAIC (Ringeval et al., 2019), a corpus that shares the SimSensei agent and PHQ-8 labeling protocol with DAIC-WOZ but uses a fully automated interviewer (no Wizard-of-Oz operator). The replication probes whether the methodological claims of the paper are properties of DAIC-WOZ specifically or of the recipe more broadly, and what happens when an obvious-looking "fix" (retraining the audio teacher on the target corpus) is applied.

Setup. Single-seed cross-corpus replication. The DAIC-WOZ-locked text and audio teachers are loaded without retraining; only the fusion head, projection, and auxiliary heads are trained on E-DAIC's training pool (163 train + 56 test pooled, 219 patients total; 56 held-out dev patients), evaluated on the E-DAIC dev set under the same threshold sweep used throughout the paper. E-DAIC text instances are constructed in the same `[Q]: [A]:` format used on DAIC-WOZ; for participant-only segments where no interviewer turn is available, the `[A]:` prefix is preserved with an empty `[Q]:` tag. Audio preprocessing follows the locked DAIC pipeline (Section 5.2, Appendix A) without modification.

Pre-fusion transfer probe. Table 12 reports the locked DAIC-WOZ teachers, run standalone on E-DAIC dev without any retraining or adaptation. The transfer is asymmetric: the text teacher

Locked teacher	DAIC dev F1	E-DAIC dev F1
Text teacher	0.8571	0.6667
Audio teacher	0.7500	0.4615

Table 12: Pre-fusion transfer probe. The DAIC-WOZ-locked teachers are run standalone on the E-DAIC development set without retraining. Transfer is asymmetric: the text teacher retains classification ability while the audio teacher drops to chance.

drops from F1 = 0.8571 on DAIC to F1 = 0.6667 on E-DAIC but retains classification ability well above chance, while the audio teacher drops from F1 = 0.7500 on DAIC to F1 = 0.4615 on E-DAIC, well below the threshold for meaningful binary classification. The asymmetry is consistent with acoustic domain shift between DAIC's Wizard-of-Oz protocol and E-DAIC's fully automated agent regime: linguistic content of patient self-disclosure transfers, acoustic prosodic patterns shaped by the interviewer-side interaction do not. This is the initial condition for everything that follows: text knows what it is doing on E-DAIC; audio does not.

Fusion with DAIC-locked teachers. With the locked teachers transferred unchanged into the fusion harness, the qualitative pattern from DAIC-WOZ transfers cleanly (Table 13). The frozen-teacher fusion configuration reaches F1 = 0.6667 at the F1-best epoch with $\Delta_{\text{audio}} = +0.107$. The joint-trained comparison reaches F1 = 0.625 at its F1-best epoch with $\Delta_{\text{audio}} = +0.019$, and Δ_{audio} drops to exactly 0.000 from epoch 3 onward and remains at 0.000 through 25 subsequent training epochs while F1 degrades to roughly 0.55. At the F1-best epoch the frozen-vs-unfrozen Δ_{audio} ratio is roughly $5\times$; at steady state the unfrozen Δ_{audio} is zero and the frozen value is $+0.107$, so the ratio is undefined but the qualitative contrast is unambiguous. The same qualitative pattern the paper establishes on DAIC-WOZ holds on a corpus with a different interviewer protocol, different recording conditions, and a substantially weaker pre-fusion audio teacher.

Two sub-findings inside the cross-corpus result are worth foregrounding.

First, the magnitude of Δ_{audio} on E-DAIC ($+0.107$) is comparable to the DAIC-WOZ best-seed headline ($+0.1423$) and approximately double the DAIC-WOZ 5-seed mean ($+0.0568$). The plausible mechanism is that on DAIC-WOZ the text teacher saturates near the bias-aware ceiling (F1 = 0.8571), leaving smaller residual variance for audio to explain, whereas on E-DAIC the transferred text teacher does not saturate (F1 = 0.6667), leaving substantially more residual variance for the audio modality to contribute against. Under this reading, Δ_{audio} magnitude depends not only on how informative the audio modality is but also on how dominant the text channel is in any given setting; a larger Δ_{audio} on a corpus where text is weaker is not paradoxical but consistent with the diagnostic measuring marginal audio contribution.

Second, the standalone V7 audio teacher is at chance on E-DAIC dev (F1 = 0.4615), yet the frozen fusion head extracts $\Delta_{\text{audio}} = +0.107$ from its chunk representations. We interpret this as evidence that depression-relevant audio variation exists at the chunk level even where a linear classifier on top of those chunks fails to surface it; the frozen fusion head, given text context as scaffold, can access that latent variation. The unfrozen comparison rules out the most obvious alternative explanation, that the contribution reflects spurious architecture-side leakage rather than audio information per se: were that the case, the unfrozen Δ_{audio} would not collapse to 0.000 within three epochs and stay there.

Audio teacher retraining: a negative result that strengthens the freezing argument. The natural next question is whether retraining the audio teacher on E-DAIC would recover the missing standalone signal and lift fused performance. We tested this directly. The audio teacher was retrained on E-DAIC's training pool (219 patients, 4,051 chunks) under the same recipe used for the locked DAIC teacher; best E-DAIC dev F1 reached 0.4848, only $+0.023$ over the zero-transfer baseline of 0.4615. The audio teacher remained essentially at chance after retraining. Substituting the retrained V7 into the fusion harness produced the configuration shown in Table 14.

Two effects merit explicit comment. The frozen-vs-unfrozen Δ_{audio} contrast that held cleanly with the DAIC-trained teacher (Table 13) has disappeared with the retrained teacher: $+0.024$ versus $+0.026$ is essentially identical. And the frozen trajectory under the retrained V7 shows Δ_{audio} turning negative in late epochs, reaching -0.292, indicating that audio occlusion at inference *improves* F1: the audio branch is now contributing anti-signal that the fused prediction has to compensate for. The interpretation we favor is that small-target-corpus

Configuration	$F1_{swept}$	Δ_{audio}
Frozen-teacher fusion (DAIC teachers)	0.6667	$+0.107$
Joint-trained fusion (DAIC teachers)	0.625	$+0.019$ (collapses to 0.000 by ep 3)

Table 13: Cross-corpus fusion on E-DAIC dev with the DAIC-WOZ-locked teachers transferred unchanged. The frozen condition produces measurable audio contribution; the unfrozen condition collapses to $\Delta_{audio} = 0$ within the first few epochs and stays there, matching the DAIC-WOZ qualitative pattern.

Configuration (retrained audio teacher)	$F1_{swept}$	Δ_{audio}
Frozen-teacher fusion	0.6667	$+0.024$
Joint-trained fusion	0.632	$+0.026$

Table 14: Cross-corpus fusion on E-DAIC dev with the audio teacher retrained on E-DAIC's training pool. The frozen-vs-unfrozen Δ_{audio} contrast that holds with the DAIC-trained audio teacher (Table 13) has disappeared. In late epochs, the frozen-condition Δ_{audio} trajectory turns negative (down to -0.292), meaning audio occlusion at inference improves F1: the audio branch is contributing anti-signal.

retrains overfit (V7's E-DAIC training loss dropped roughly fourfold over the retraining run) and produce chunk representations that fit the target training pool but do not generalize to the target dev set; the frozen fusion head cannot recover useful signal from them. The DAIC-trained V7, by contrast, despite being at chance standalone on E-DAIC, carries representations that the fusion head can use.

The practical implication strengthens, rather than weakens, the freezing argument. The naive expectation is that fixing the weak teacher should improve the result; the observed result is the opposite. We interpret this as a second methodological recommendation, complementary to the central freezing claim: under cross-corpus acoustic domain shift, prefer source-corpus pre-trained teachers over small-target-corpus retrains, even when the source-corpus teacher is at chance on the target corpus standalone. The frozen fusion regime is what makes this prescription work; under joint training, neither source-corpus transfer nor target-corpus retraining produces measurable audio contribution.

Scope and limitations. The cross-corpus replication is single-seed by explicit design: the central 5-seed paired evidence for the freezing claim lives on DAIC-WOZ in the main text. Treating the E-DAIC result as a confirmatory generalization probe rather than a primary contribution avoids three failure modes. First, single-seed cross-corpus numbers are not the right place to make precise quantitative claims; we report the qualitative pattern (frozen produces measurable Δ_{audio}; unfrozen does not) and order-of-magnitude effect sizes, but do not claim the contrast ratios reported here as precise. Second, the V7 retraining experiment is itself a single retraining run; the conclusion that small-target-corpus retrains hurt the cross-corpus pattern rests on this single run, the trajectory of Δ_{audio} across its training epochs, and the contrast against the zero-retrain configuration. We treat the retraining result as a strongly-suggestive negative finding rather than a settled methodological claim, and flag broader replication of the "prefer source-corpus teachers over target-corpus retrains" prescription as future work. Third, we have evaluated on E-DAIC only; cross-lingual transfer (EATD (Shen et al., 2022), Mandarin) and audio-only cross-corpus transfer (ANDROIDS (Tao et al., 2023), Italian audio) remain explicit future work.

L Per-patient audio-occlusion consistency analysis

To address the concern that the headline $\Delta_{audio} = +0.1423$ corresponds to a small number of patient flips that could reflect noise rather than signal, we conducted a per-patient consistency analysis across five seeds $\{42, 123, 7, 0, 11\}$ of the locked frozen-teacher fusion configuration. For each patient in the 35-patient dev set, we computed the number of seeds in which the patient's prediction flips when audio is occluded (i.e., the audio modality changes that patient's predicted label at the swept threshold).

Under the frozen recipe, 3 of 35 patients (9%) show consistent audio dependence across 4–5 of 5 seeds, 6 patients show seed-dependent flips (1–2 of 5), and 26 patients never flip on any seed (audio-irrelevant cases). Under joint training, no patient flips in 3 or more seeds, only 2 of 35 show any audio-induced flip at all, and 33 of 35 (94%)

Flip frequency across 5 seeds	Frozen	Joint-trained
Consistent (4–5 of 5 seeds flip)	3	0
Seed-dependent (1–2 of 5 seeds flip)	6	2
Never flips (0 of 5 seeds)	26	33
Total patients	35	35

Table 15: Per-patient audio-occlusion consistency across the five seeds $\{42, 123, 7, 0, 11\}$. "Consistent" patients have their predicted label flipped by audio occlusion in 4 or 5 of the 5 seeds; "seed-dependent" patients flip in 1 or 2 seeds; "never flips" patients are unaffected by audio occlusion on any seed. The frozen condition shows three consistent audio-dependent cases (9% of dev patients) versus zero under joint training.

never flip; this is the per-patient signature of modality collapse. The qualitative concentration pattern in the frozen condition (consistent flips concentrating on a recurring subset rather than scattering uniformly across patients and seeds) supports the interpretation that the audio modality contributes signal beyond noise: a noise-only Δ_{audio} would distribute flips uniformly, while a signal-bearing Δ_{audio} concentrates them on the patients where the audio modality actually carries depression-relevant information.

M Schedule vs selection criterion within the extended trajectory

Within the same extended-schedule training trajectory of the frozen-teacher fusion configuration (seed-averaged across the five seeds $\{42, 123, 7, 0, 11\}$), we considered three checkpoint-selection criteria applied to the same per-epoch checkpoints:

(i) F1-only: $\arg\max_t F1_{\text{swept}}(t)$

(ii) F1 + sep: $\arg\max_t F1_{\text{swept}}(t) + 0.3 \cdot \text{sep}(t)$

(iii) Joint score: $\arg\max_t F1_{\text{swept}}(t) + 0.3 \cdot \text{sep}(t) + 1.5 \cdot \Delta_{\text{audio}}(t)$

The selection rule used throughout the main text is the F1-best rule with Δ_{audio} as a tiebreaker (Section 5.4), which corresponds to (i) with deterministic tiebreaking on Δ_{audio} then sep. The joint-score criterion (iii) is the training-time checkpoint-saving criterion used internally during fusion training; it is not the rule under which any number in the main text is reported. The F1 + sep criterion (ii) is the historical convention used by the unimodal teacher trainers (separation as a tiebreaker on F1 ties), retained for comparability with the locked text and audio teachers.

Table 16 reports the 5-seed mean ± standard deviation of the chosen-checkpoint metrics under each rule on the same extended-schedule training trajectories.

Two observations follow. First, the F1-only and F1+sep rules produce identical F1 (0.8621) at slightly different epochs but materially different Δ_{audio} values ($+0.0776$ vs $+0.0717$); the difference is the sep tiebreaker preferring epochs with cleaner probability separation but smaller audio contribution. The paper's reporting rule (F1-best with $(F1, \Delta_{\text{audio}}, \text{sep})$ lexicographic tiebreaker, Section 5.4) is therefore most closely aligned with F1-only in this comparison and selects the same-F1 epoch with the larger audio contribution. Second, the joint-score criterion is what surfaces the high-Δ_{audio} late epochs that the F1-only and F1+sep rules miss: it trades 0.017 F1 for an additional $+0.106$ Δ_{audio}. The implication is that the schedule recovery story is not "extended schedule alone surfaces the contribution" but rather "extended schedule with Δ_{audio}-aware selection together surface it." Per-rule per-seed chosen-checkpoint metrics are released as supplementary material.

N Random-initialized audio encoder placebo control

To control for the possibility that Δ_{audio} reflects information leakage through the audio architecture's parameter count rather than genuine audio information, we replaced the locked audio teacher V7 with a frozen random-initialized audio encoder of the same architecture (Conv1D compression + LayerNorm-on-LSTM-input + BiLSTM), retaining the pretrained `wav2vec2-base-superb-er` feature extractor as the frozen front end. The fusion head, projection, and text branch are unchanged. Three seeds (42, 123, 7), otherwise matching the headline configuration.

The random-init audio teacher produces a 5-seed mean $\Delta_{\text{audio}} = +0.0059 \pm 0.0132$, statistically indistinguishable from the joint-trained fusion's $+0.0059 \pm 0.0132$ and an order of magnitude below

Rule	epoch	$F1_{swept}$	Δ_{audio}	sep
F1-only	6.8 ± 7.5	0.8621 ± 0.0068	$+0.0776 \pm 0.0759$	0.317 ± 0.160
F1 + sep	7.8 ± 6.6	0.8621 ± 0.0068	$+0.0717 \pm 0.0815$	0.495 ± 0.045
Joint score	18.2 ± 7.3	0.8448 ± 0.0235	$+0.1837 \pm 0.0607$	0.454 ± 0.027

Table 16: 5-seed mean $\pm$ standard deviation of chosen-checkpoint metrics under three selection rules applied to the same extended-schedule trajectories. F1-only and F1+sep rules select earlier in training (epoch ≈ 7) and converge on the same matched-F1 epoch with small Δ_{audio}; the joint-score rule selects substantially later (epoch ≈ 18) and trades 0.017 F1 for an additional $+0.106$ Δ_{audio}.

Audio teacher	$F1_{swept}$	Δ_{audio}
Trained audio teacher (headline, seed 123)	0.8696	$+0.1423$
Random-init (placebo, best of 5)	0.8571	$+0.0296$
Random-init (placebo, 5-seed mean)	0.8512	$+0.0059 \pm 0.0132$
Text-only baseline	0.8571	—

Table 17: Random-init audio encoder placebo control. The placebo replaces the locked audio teacher with a frozen random-initialized encoder of the same architecture, retaining the pretrained `wav2vec2-base-superb-er` feature extractor as the frozen front end. 5-seed sweep under our pre-registered selection rule.

the trained-audio-teacher frozen-teacher fusion's $+0.0568 \pm 0.0540$. This supports the interpretation that the trained audio teacher carries prosodic and affective information that the fusion head exploits, beyond what a random encoder of the same architecture would provide.

O Gradient-norm analysis during joint training

To probe whether the frozen-vs-unfrozen ablation is consistent with the gradient-contention account of modality collapse, we logged the L2 norm of gradients flowing into the text encoder and the audio encoder at each training step of the joint-trained fusion run (seed 42).

Per-epoch gradient L2 norm logging during joint-trained fusion training shows the text branch gradient is roughly $38\times$ the audio branch gradient in the first 10 epochs (text mean grad-norm ≈ 16.5 over the range $[0.014, 154.0]$ across steps; audio mean grad-norm ≈ 0.45 over the range $[0.225, 0.762]$) and roughly $9\times$ in the last 10 epochs. The sustained magnitude asymmetry across training is consistent with the gradient-contention magnitude account of multimodal collapse (Javaloy et al., 2022; Huang et al., 2022; Peng et al., 2022) and reflects the parameter-count asymmetry between the 355M-parameter RoBERTa-large backbone and the 4.3M-parameter audio compression and BiLSTM stack. We do not adjudicate directional gradient interaction (e.g., cosine alignment between branch-aggregate gradients) from these data, because rank-aligned cosine measures sparsity-pattern similarity rather than directional alignment in parameter space and the value is not interpretable on its own as evidence for or against gradient-direction contention; magnitude asymmetry alone is the load-bearing observation here. Full per-epoch gradient-norm trajectories are released alongside the code as supplementary material.

P Gradient-modulation comparison: OGM-GE

To compare the frozen-teacher recipe directly against gradient-modulation alternatives that do not require frozen encoders, we trained the same fusion architecture under joint training with OGM-GE (Peng et al., 2022) gradient modulation applied. Three seeds (42, 123, 7), otherwise matching the joint-trained baseline.

OGM-GE does not recover Δ_{audio} relative to the joint-trained baseline to frozen-teacher levels: four of the five OGM-GE seeds collapse to $\Delta_{audio} = 0.0000$ and one outlier (seed 7) reaches $+0.0423$, yielding a 5-seed mean of $+0.0085 \pm 0.0189$ that is an order of magnitude below the frozen-teacher 5-seed mean of $+0.0568 \pm 0.0540$ on the same architecture. The interpretation of this result for the underlying mechanism is discussed in Section 6.

Q Intermediate freeze conditions

To further isolate which branch's gradient flow drives the modality collapse observed under joint training, we ran two intermediate-freeze conditions on the locked architecture:

Configuration	$F1_{swept}$	Δ_{audio}
Frozen-teacher fusion (headline, seed 123)	0.8696	$+0.1423$
Joint-trained fusion (paired seed 123)	0.8571	0.0000
OGM-GE on joint-trained fusion (best of 5)	0.8571	$+0.0423$ (seed 7 outlier)
OGM-GE on joint-trained fusion (5-seed mean)	0.8571	$+0.0085 \pm 0.0189$

Table 18: Gradient-modulation comparison. The frozen-teacher headline and joint-trained comparison are at seed 123; OGM-GE rows report the best-of-5 seeds and the 5-seed mean. Across the 5 OGM-GE seeds, four runs collapse to $\Delta_{audio} = 0.0000$ and one outlier (seed 7) reaches $+0.0423$; the 5-seed mean $+0.0085 \pm 0.0189$ is an order of magnitude below the frozen-teacher 5-seed mean.

- `freeze_text=True, freeze_audio=False` (audio trains; text frozen), five seeds $\{0, 7, 11, 42, 123\}$.
- `freeze_text=False, freeze_audio=True` (text trains; audio frozen), one seed at submission, with two additional seeds in flight (camera-ready update).

Under the gradient-contention reading, freezing only the dominant text branch should be approximately sufficient to recover Δ_{audio}, while freezing only the smaller audio branch should leave the text-gradient-domination dynamic intact and produce $\Delta_{audio} \approx 0$. The observed result is qualitatively consistent with this prediction, but the per-seed picture is noisier than the means alone suggest. Text-frozen / audio-training: 5-seed mean $\Delta_{audio} = +0.0713 \pm 0.1369$, with three seeds (0, 11, 123) collapsing to 0.0000, one seed (42) reaching $+0.0423$, and seed 7 reaching $+0.3140$ (the latter drives the mean). Text-training / audio-frozen: 3-seed mean $+0.0099 \pm 0.0171$, with two seeds (123, 42) collapsing to 0.0000 and one seed (7) reaching $+0.0296$. The asymmetry is in best-seed magnitude ($+0.3140$ vs $+0.0296$, an order of magnitude) and in the proportion of seeds where the condition recovers any audio contribution at all (2 of 5 vs 1 of 3), but the small-sample variance prevents a clean statistical claim. We do not pivot the recipe to text-only freezing for three reasons: the text-frozen-only variance is more than twice that of the both-frozen recipe (0.1369 vs 0.0540); the best-seed result is driven by one seed; and the both-frozen recipe is more stable across initialization. The intermediate-freeze result functions as mechanism evidence for the text-branch-dominance hypothesis, not as a recipe change.

R Reproducibility checklist

We follow the ACL Responsible NLP checklist; the items below summarize the reproducibility-relevant elements of this work. Cross-references to detailed sections are included where applicable.

Architecture summary. The full architecture is described in Section 4 and depicted in Figure 1; we summarize the components here for one-glance reference. *Text branch*: DepRoBERTa-large (Poświata and Perełkiewicz, 2022) encoder (frozen during fusion, fine-tuned end-to-end during teacher training) → LayerNorm on `[CLS]` → Bi-GRU (hidden 512, bidirectional, output 1024) → gated attention MIL pooling (Ilse et al., 2018) (attention dim 128, no temperature) → Dropout(0.5) → Linear(1024→1). *Audio branch*: `wav2vec2-base-superb-er` (Baevski et al., 2020; Yang et al., 2021) feature extractor (frozen) on 20-second participant-only chunks → Conv1d(kernel 5, stride 15, 768→768) → BatchNorm → ReLU → Dropout(0.4) → LayerNorm → BiLSTM (hidden 128 per direction, 2 layers, dropout 0.5) → Dropout(0.6) → Linear(256→2). *Fusion block*: text projection Linear(1024→256) with LayerNorm and Dropout(0.5) → broadcast across audio chunks and concatenate per chunk (256+256=512) → Linear(512→256) → LayerNorm → ReLU → Dropout(0.5) → Linear(256→1) → mean over chunks. Both unimodal teachers are frozen during fusion training (`requires_grad=False` on all branch parameters, branches in `eval()` mode, BatchNorm `track_running_stats=False`, branches forwarded inside `torch.no_grad()`); only the projection, fusion head, and two auxiliary heads are trained.

Data partitions. DAIC-WOZ (Gratch et al., 2014) uses the official AVEC 2017 partition (Ringeval et al., 2017): 107 training, 35 development, 47 test patients (total $n = 189$). Following recent literature (Xu et al., 2025; Patapati, 2024), we pool the 47-patient test split into the training set and reserve the 35-patient development split for held-out evaluation; PID 487 is excluded due to

Freeze condition	$\mathbf{F1}_{\text{swept}}$	Δ_{audio}
Both frozen (headline, seed 123)	0.8696	$+0.1423$
Both unfrozen (paired seed 123)	0.8571	0.0000
Text frozen, audio trains (best of 5, seed 7)	0.8696	$+0.3140$
Text frozen, audio trains (5-seed mean)	0.8596	$+0.0713 \pm 0.1369$
Text trains, audio frozen (best of 3, seed 7)	0.8571	$+0.0296$
Text trains, audio frozen (3-seed mean)	0.8571	$+0.0099 \pm 0.0171$

Table 19: Intermediate-freeze conditions. Text-frozen / audio-training reported across 5 seeds; text-training / audio-frozen reported across the 3 seeds completed at submission (seeds $\{42, 123, 7\}$); the remaining two seeds are documented for camera-ready.

transcript-parsing failure (Appendix A), giving an effective training pool of 152 patients. E-DAIC (Ringeval et al., 2019) uses the official AVEC 2019 partition: 163 training, 56 development, 56 test patients (total $n = 275$). The cross-corpus replication (Section 5.7, Appendix K) pools E-DAIC train and test (219 patients) for fusion training and reserves the 56-patient E-DAIC dev for held-out evaluation. The audio teacher retraining in Appendix K uses the same 219-patient E-DAIC training pool. Class distribution: DAIC-WOZ train+test pool is 56 depressed / 96 non-depressed (37/63 imbalance); DAIC-WOZ dev is 12 depressed / 23 non-depressed.

Scientific artifacts. Data. The DAIC-WOZ corpus (Gratch et al., 2014) is licensed by the USC Institute for Creative Technologies under the standard Distress Analysis Interview Corpus data-use agreement. Researchers wishing to use our preprocessing pipeline must obtain DAIC-WOZ access independently via the USC DCAPS portal (`dcapswoz.ict.usc.edu`); we do not redistribute raw audio, raw transcripts, or any participant-identifiable derived features (Ethical considerations). The E-DAIC corpus (Ringeval et al., 2019) used in the cross-corpus replication (Section 5.7) is available under the same DCAPS data-use agreement. We adopt the official AVEC 2017 split for DAIC-WOZ (Ringeval et al., 2017) and the official AVEC 2019 split for E-DAIC. **Code.** Preprocessing scripts, training scripts, evaluation scripts, and the per-modality occlusion-delta diagnostic implementation will be released at acceptance under an open-source license; the released code does not bundle raw data. **Pre-trained models.** The text encoder is `rafalposwiata/deproberta-large-depression` (Poświata and Perełkiewicz, 2022); the audio feature extractor is `superb/wav2vec2-base-superb-er` (Baevski et al., 2020; Yang et al., 2021). Both are accessed via the HuggingFace Transformers library (Wolf et al., 2020). The locked text and audio teacher checkpoints will be released alongside the code.

Computational experiments. Reproducible seeds. All multi-seed experiments use the fixed seed list $\{42, 123, 7, 0, 11\}$, which is the same seed list used to lock the text teacher (Section 5.3); single-seed exploratory experiments (e.g., gradient-norm logging, Appendix O) use seed 42 unless otherwise noted. **Selection rule.** Per-seed best-epoch selection uses F1-best with $(F1_{\text{swept}}, \Delta_{\text{audio}}, \text{sep})$ lexicographic tiebreaker (Section 5.4, pre-registered); this rule is applied uniformly across all reported configurations with zero manual overrides. **Hyperparameters.** The full hyperparameter inventory for the text teacher, audio teacher, and frozen-teacher fusion model is reported in Appendix B (Tables 4, 5, and 6). **Compute.** All experiments were conducted on a single NVIDIA A100 GPU under a research-time budget of approximately six months from initial baseline through submission. Approximate wall-clock training times: text teacher 85 minutes per seed; audio teacher 30 minutes per seed; fusion 25 minutes per run. **Statistical testing.** We report bootstrap 95% confidence intervals on Δ_{audio} (1000 bootstrap samples) for the headline paired-seed comparison (Section 4.4) and a Wilcoxon signed-rank test on the five paired seeds for the frozen-vs-joint-trained Δ_{audio} contrast (Section 4.4, $W = 10$, one-sided $p = 0.0625$, Cohen's $d_z = 0.88$); test statistics, paired observations, and effect-size computations are released as supplementary material.

Software environment. PyTorch with CUDA on Ubuntu 22. HuggingFace Transformers for tokenization and pre-trained encoder loading. Torchaudio for audio resampling and VAD. `noisereduce` (Sainburg, 2020) for spectral-subtraction noise re-

duction. AdamW optimizer (Loshchilov and Hutter, 2019). TF32 fp32 matmul precision throughout. Exact library version pins are released with the code.

Use of AI assistants. AI writing-assistance tools (large language models accessed via chat-based interfaces) were used for prose drafting, literature search support, and editorial review under direct human supervision (Appendix J). No AI tool is listed as an author, and all research questions, experimental design, code, training runs, results, and interpretive claims are the author's own.

Use of human annotators. This work uses pre-existing PHQ-8 self-report labels from the DAIC-WOZ and E-DAIC corpora; we did not collect or annotate new data and did not employ human annotators.