SEGMENT-AWARE TEMPORAL EVIDENCE GRAPHS FOR RISK-CONTROLLED MULTIMODAL DEPRESSION SCREENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal depression screening aims to assess depression risk using behavioral cues observed during clinical interviews, such as speech, facial movements, and head pose. Existing methods often compress an entire interview into a single global representation, potentially weakening stage-dependent evidence, and require a binary decision for every subject, even when uncertainty is high. To address these limitations, we propose a segment-aware temporal evidence graph with selective risk control. Our method treats each window–modality pair as a graph node and combines temporal encoding with typed graph attention to capture temporal dependencies and cross-modal interactions. Segment-aware pooling then aggregates global, early-stage, middle-stage, and late-stage representations, along with the difference between late-stage and early-stage representations, for subject-level prediction. To handle uncertain cases, we introduce a positive-priority, two-round selective screening strategy that combines probability calibration, asymmetric decision thresholds, and ensemble disagreement. Ablation studies on the DAIC-WOZ development set support the effectiveness of segment-aware pooling and the complementary contributions of audio and visual behavioral cues. Selective screening results further show that the proposed strategy explicitly balances automated prediction coverage and the risk of missing positive cases by abstaining from predictions for uncertain subjects. Code is available at: `https://anonymous.4open.science/r/rc-sateg-anonymous-3876/`.

1 INTRODUCTION

Early identification of depression risk can inform subsequent assessment and intervention. Multimodal machine learning can support interview-based depression screening by analyzing behavioral cues such as speech, facial movements, and head pose. Resources such as DAIC-WOZ and the AVEC benchmarks support research on depression screening using semi-structured clinical interviews (Gratch et al., 2014; DeVault et al., 2014; Valstar et al., 2016; Ringeval et al., 2017; 2019). More broadly, multimodal machine learning has been widely studied as a means of integrating heterogeneous behavioral signals (Baltrušaitis et al., 2019). This work addresses two related questions: how to construct a subject-level representation from temporal multimodal evidence and when to issue an automated screening prediction.

A key challenge is the mismatch between interview-level labels and temporally localized evidence. Depression-related cues may emerge only in specific time windows and vary over the course of an interview. Global pooling can dilute these local or stage-dependent cues. Moreover, speech and visual behaviors within the same time window may convey complementary information that is not explicitly preserved when each modality is pooled independently. An effective representation should therefore capture temporal dependencies, cross-modal interactions, and differences across coarse-grained interview stages.

Another challenge is prediction reliability. Standard binary classifiers typically assign a positive or negative label to every subject. In risk-sensitive screening, however, false negatives are particularly concerning. A screening system should therefore issue predictions when confidence is sufficient and abstain otherwise. Calibration and selective prediction offer tools for this purpose (Guo et al., 2017;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

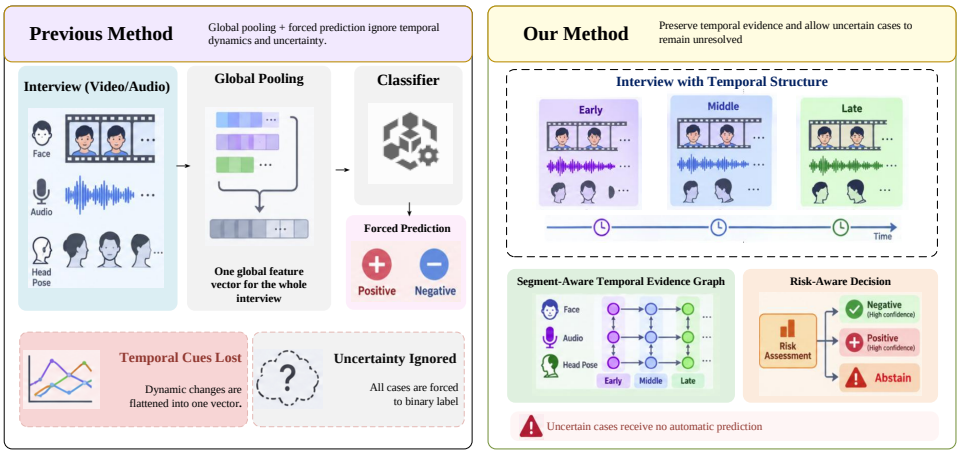

Figure 1: Motivation of the proposed framework. Left: Global pooling may obscure stage-dependent evidence, while forced binary decisions overlook uncertainty. Right: A segment-aware temporal evidence graph preserves multimodal dynamics across interview stages, and risk-aware selective screening classifies confident cases while abstaining on uncertain subjects.

Geifman & El-Yaniv, 2017; 2019), but must be adapted to account for the asymmetric costs of errors in mental health screening.

We propose a segment-aware temporal evidence graph with selective risk control. Our method constructs a graph with window–modality pairs as evidence nodes, combines temporal encoding with typed graph attention, and aggregates evidence through global, segment-level, and change-aware pooling. For decision making, we adopt a positive-priority, two-round selective screening procedure: a single model conducts the initial screening, and an ensemble reassesses unresolved cases in the second round. The system abstains from issuing automated predictions for cases that remain uncertain. Our contributions are as follows:

- We propose a segment-aware temporal evidence graph for multimodal depression screening, integrating temporal and cross-modal interactions with global, early-stage, middle-stage, late-stage, and late-minus-early representations for subject-level prediction.
- We develop a two-round selective screening strategy that combines calibration, asymmetric thresholds, and ensemble disagreement to assign automatic positive or negative predictions, while abstaining on unresolved subjects.
- Experiments on DAIC-WOZ demonstrate the benefits of segment-aware pooling and multimodal inputs, and show that two-round screening increases automatic decision coverage relative to one-round screening.

2 Related Work

Multimodal Depression Recognition. Multimodal depression recognition integrates speech, visual, and textual information to infer depression-related states. Research in this area draws on both task-specific screening systems and general multimodal representation learning frameworks (Baltrušaitis et al., 2019; Al Hanai et al., 2018; Tsai et al., 2019; Zadeh et al., 2018). The DAIC-WOZ dataset and depression-related AVEC challenge benchmarks are widely used for interview-based depression analysis (Gratch et al., 2014; DeVault et al., 2014; Valstar et al., 2013; 2016; Ringeval et al., 2017; 2019). Acoustic features characterize prosody and voice quality, facial movements and head pose capture visual behavior, and text conveys interview content. Prior work also suggests that long-term speech patterns and segment-level interview structure can aid depression detection. Recent work has also explored evidence-mediated reasoning for multi-disorder mental health screening (Anonymous Authors, 2026). Our work focuses on acoustic and visual behavioral features and ex-

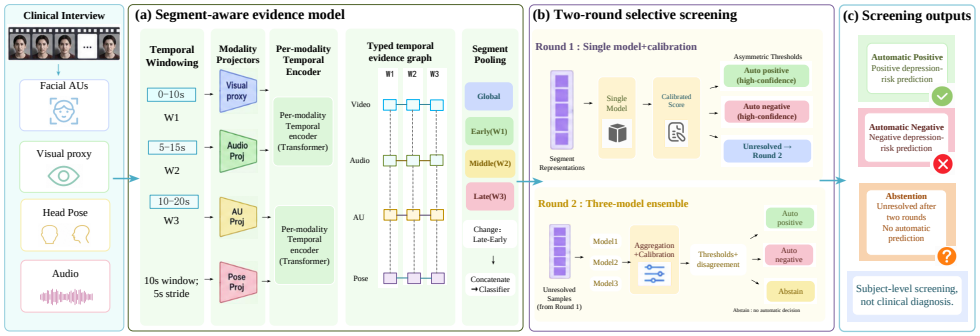

Figure 2: Overview of the proposed framework. **(a)** Temporal encoding, typed graph attention, and segment-aware pooling integrate multimodal evidence. **(b)** Calibrated single-model screening is followed by three-model ensemble re-evaluation of unresolved cases. **(c)** Outputs are automatic positive, automatic negative, or abstention.

plicitly models temporal windows and coarse-grained interview stages for binary depression risk screening.

Temporal Encoding and Graph Attention. Graph neural networks offer a natural framework for modeling structured dependencies. Graph attention networks aggregate features from neighboring nodes using learned attention weights (Veličković et al., 2018), while GATv2 enhances the expressiveness of the attention mechanism (Brody et al., 2022). Transformer-based sequence modeling is widely used for temporal representation learning (Vaswani et al., 2017). In our framework, temporal encoding and graph propagation serve complementary roles: temporal encoding contextualizes representations within each modality over time, while typed graph attention propagates evidence along temporal and cross-modal edges.

Calibration and Selective Prediction. Probability calibration assesses how closely predicted probabilities align with observed frequencies of positive labels. Post-hoc calibration methods, such as Platt scaling and temperature scaling, aim to improve this alignment (Platt, 1999; Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005; Guo et al., 2017). Selective prediction and classification with a reject option examine when models should abstain from making predictions and how to balance coverage and risk (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019). We adapt these ideas to positive-priority mental health screening through asymmetric decision thresholds, ensemble-based reassessment, and system-level metrics that separately quantify automated prediction coverage and the risk of missing positive cases. Our use of ensemble disagreement is also connected to research on predictive uncertainty estimation with deep ensembles and uncertainty evaluation under dataset shift (Lakshminarayanan et al., 2017; Ovadia et al., 2019).

3 Method

The proposed framework comprises two components: segment-aware temporal evidence modeling and selective risk-controlled screening. Given a clinical interview, we first divide it into fixed-length windows and extract multimodal behavioral features. We then construct a temporal evidence graph with window–modality pairs as nodes and aggregate node representations through segment-aware pooling to obtain a subject-level representation. Finally, a two-round selective screening procedure determines whether to issue an automated prediction.

3.1 Window-Level Inputs and Temporal Evidence Graph

Each interview is divided into overlapping 10-second windows with a 5-second stride. For each window t and modality m, we extract a feature vector $\mathbf{x}_{t,m}$, a quality score $q_{t,m}$, and a validity mask $v_{t,m}$. The feature groups comprise facial action features, gaze-related features, head pose features, and enhanced acoustic features. These features are related to those used in common facial

and acoustic analysis pipelines, such as OpenFace, openSMILE, and COVAREP (Baltrušaitis et al., 2018; Eyben et al., 2010; Degottex et al., 2014). During aggregation, valid nodes are weighted by their quality scores, and invalid nodes are masked out.

We construct a graph in which each node represents a window–modality pair (t, m). The graph contains three edge types: self-loops, temporal edges linking nearby windows within the same modality, and cross-modal edges linking different modalities within the same window. This typed graph structure captures temporal continuity and enables cross-modal evidence exchange.

3.2 Typed Graph Attention Propagation

For each modality, raw features are first projected into a shared hidden space. A temporal encoder then contextualizes representations within each modality across interview windows. Let $\mathbf{h}_{t,m}$ denote the resulting representation of node (t, m). Typed graph attention updates each node by aggregating information from its neighbors:

$$\mathbf{h}'_i = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_r(i)} \alpha^r_{ij}\mathbf{W}_r\mathbf{h}_j\right), \tag{1}$$

where $\mathcal{R}$ is the set of edge types, $\mathcal{N}_r(i)$ is the set of neighbors connected to node i by edges of type r, and α^r_{ij} is the learned attention weight. This design distinguishes temporal propagation from cross-modal fusion.

3.3 Segment-Aware Pooling and Prediction

After graph propagation, node embeddings are pooled into a subject-level representation. To complement global pooling, we partition each interview into early, middle, and late temporal segments and compute a quality-weighted representation for each segment. We also explicitly encode temporal change as the difference between the late- and early-segment representations:

$$\mathbf{z}_{change} = \mathbf{z}_{late} - \mathbf{z}_{early}. \tag{2}$$

The final representation concatenates the global, segment-level, and change-aware features:

$$\mathbf{z} = [\mathbf{z}_{global}; \mathbf{z}_{early}; \mathbf{z}_{middle}; \mathbf{z}_{late}; \mathbf{z}_{change}]. \tag{3}$$

A classifier uses $\mathbf{z}$ to predict depression risk. We evaluate the model on binary depression screening.

Figure 3 details how window-level features, quality scores, and validity masks are organized into evidence nodes and aggregated into a subject-level representation.

3.4 Two-Round Selective Screening

The selective screening module abstains from issuing automated predictions for highly uncertain cases. In the first round, a single model produces a calibrated score, and cases meeting the positive or negative decision threshold receive an automated prediction. Unresolved cases proceed to the second round, where a three-model ensemble produces a calibrated prediction and a disagreement score. Cases that still fail to meet the decision criteria are assigned an abstention output.

The decision rule uses separate thresholds for positive and negative predictions to account for the higher cost of missing positive cases in screening. Let $\hat{p}$ denote the calibrated probability and d the ensemble disagreement score. An automated prediction is issued only when the applicable confidence and disagreement criteria are met.

3.5 Threshold Selection and Metrics

We select thresholds on a calibration set by searching over candidate positive, negative, and disagreement thresholds to maximize automated prediction coverage, subject to empirical constraints on the risk of missing positive cases and positive predictive value.

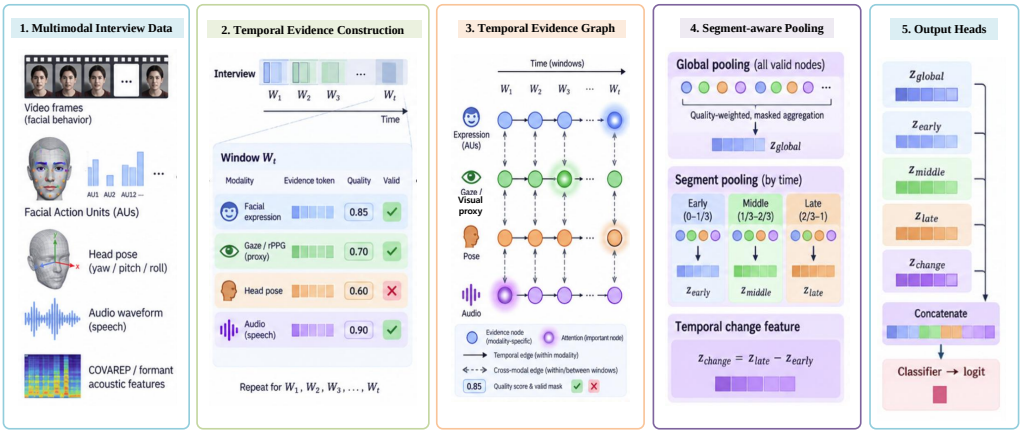

Figure 3: Detailed evidence representation pipeline. Window–modality features are organized into a temporal graph and aggregated through quality-weighted global, segment-level, and change-aware pooling for subject-level prediction. Quality values and attention highlights are illustrative.

Let N denote the total number of evaluated subjects and N_A the number receiving an automatic positive or negative decision. Automatic coverage is defined as

$$\text{Coverage} = \frac{N_A}{N}. \tag{4}$$

Among ground-truth-positive subjects, let TP denote the number automatically classified as positive, FN the number automatically classified as negative, and U_+ the number receiving no automatic decision. The system-level false-negative rate is defined as

$$\text{FNR}_{\text{sys}} = \frac{\text{FN}}{\text{TP} + \text{FN} + U_+}. \tag{5}$$

Precision is computed among automatic positive predictions. Accepted recall and accepted F1 are computed only on automatically decided subjects. In contrast, FNR_{sys} uses all ground-truth-positive subjects as its denominator, including those for whom the system abstains.

4 Experiments

4.1 Experimental Design

Dataset and Labels. We evaluate the proposed framework on DAIC-WOZ, which contains semi-structured interviews collected for automatic distress and depression analysis. Each subject is associated with a PHQ-8 score. Following the commonly used screening threshold, subjects with PHQ-8 scores greater than or equal to 10 are assigned positive labels, while the remaining subjects are assigned negative labels (Kroenke et al., 2001; 2009).

The experiments use the official training and development partitions. The training partition contains 107 subjects, including 30 positive and 77 negative subjects, while the development partition contains 35 subjects, including 12 positive and 23 negative subjects. The official test labels are not publicly available and are therefore not used in the reported evaluation. All predictions and metrics are computed at the subject level. Although each interview is represented by multiple temporal windows, each subject contributes one ground-truth label and one final screening outcome: positive, negative, or abstention.

Multimodal Evidence Construction. Each interview is divided into overlapping 10-second windows with a 5-second stride. Four behavioral feature groups are constructed for every window. The facial-action representation contains the intensities of AU01, AU02, AU04, AU12, and AU15. The

gaze-related representation contains five binocular gaze coordinates, and the head-pose representation contains three rotation parameters together with facial depth. The enhanced 54-dimensional acoustic representation combines four waveform-level statistics with window-level mean and standard deviation summaries derived from COVAREP and formant features.

In addition to the feature vector, every window–modality pair contains a validity mask and a quality score. Visual quality is derived from frame-level tracking confidence. Acoustic validity records the availability of the waveform, COVAREP, and formant streams. Invalid evidence nodes are excluded during aggregation, whereas valid nodes are weighted by their estimated quality. This prevents unavailable or unsuccessfully tracked intervals from being treated as ordinary behavioral observations. Feature extraction is performed independently for each subject, and no information from the development labels is used when constructing the temporal evidence.

Evaluation Protocol. Classification is evaluated using AUC, F1, and accuracy, reported as mean and standard deviation over four seeds, namely 1, 2, 3, and 42, under matched data splits, training settings, and checkpoint-selection rules. Selective screening is evaluated using coverage, abstention, system FNR, and abstained-positive rate; the latter two use all positive evaluation subjects as the denominator. Abstention is neither a correct classification nor a negative prediction.

Selective-Screening Setup. We perform five stratified 60/40 holdout splits of the 35 development subjects using split seeds 42–46. Each split contains 21 calibration subjects, including 7 positive and 14 negative subjects, and 14 evaluation subjects, including 5 positive and 9 negative subjects. Predictions from models trained with seeds 1–3 are fixed: round one uses a single model, and round two reassesses unresolved subjects using their three-model ensemble. Parameters are selected on the calibration subset with an empirical system FNR limit of 0.25, a minimum positive precision of 0.40, and cumulative coverage caps of 0.40 and 0.70. These repeated splits can overlap; because development data also informed checkpoint selection, the evaluation is exploratory rather than an independent test.

Implementation Details. All neural models are implemented in PyTorch. Modality-specific features are projected into a shared 64-dimensional hidden space before temporal encoding and graph propagation. The model uses two representation layers and a dropout rate of 0.2. The classification head is optimized using binary cross-entropy with a learning rate of 1×10^{-4} and a batch size of one interview. Training runs for at most 30 epochs, with early stopping after 10 epochs without improvement. Checkpoint selection follows an AUC-first rule on the development predictions. Each seed controls initialization and training randomness. All representation ablations share subject partitions, optimization settings, and stopping and checkpoint-selection rules, changing only the specified component or modality. Given the small development set, we report means and sample standard deviations across seeds. AUC measures discrimination independently of the decision threshold, while F1 and accuracy evaluate binary predictions. Seed variability informs our interpretation of the graph-propagation ablation.

4.2 Main Results

Depression-Risk Classification. The complete audio–visual model achieves an AUC of 0.67 ± 0.04, an F1 score of 0.64 ± 0.05, and an accuracy of 0.65 ± 0.06 across four random seeds. These results characterize subject-level classification performance on the DAIC-WOZ development partition. The contributions of individual components and input modalities are examined in the ablation study.

Selective Screening. Across the five repeated holdout splits, the two-round strategy achieves a mean automatic coverage of 51.43%, with an abstention rate of 48.57% and a system FNR of 4.00%. The abstained-positive rate is 64.00%, indicating that a substantial proportion of positive subjects remain unresolved rather than receiving an automatic negative prediction. The low system FNR must therefore be interpreted jointly with coverage and positive abstention, rather than as evidence that most positive subjects are automatically identified. Table 1 summarizes the main screening outputs.

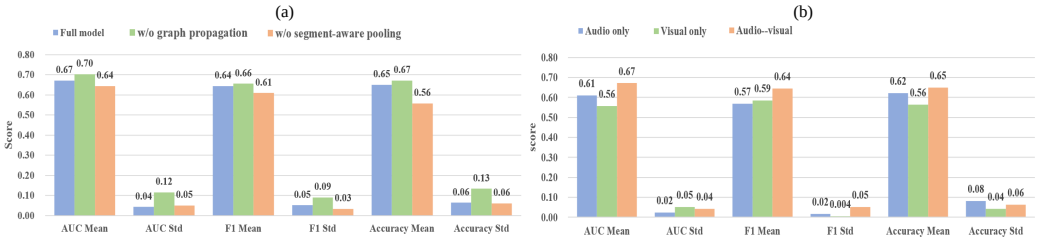

Figure 4: Ablation results on the DAIC-WOZ development set. (a) Component ablations evaluate the effects of typed graph propagation and segment-aware pooling. (b) Modality ablations compare audio-only, visual-only, and audio–visual inputs. Means and standard deviations across four random seeds are shown as separate bars.

Table 1: Two-round selective screening over five repeated holdout splits. Values are means and sample standard deviations. System FNR uses all positive evaluation subjects as its denominator.

Metric	Mean	Std.
Coverage	0.51	0.18
Abstention Rate	0.49	0.18
System FNR	0.04	0.08

The results represent a trade-off rather than an unconditional improvement in prediction performance. Abstention prevents uncertain subjects from receiving forced binary predictions, but it also reduces automatic coverage. The method should therefore be interpreted as a risk-aware screening mechanism rather than a clinical diagnostic system.

Qualitative Case Analysis. Figure 5 illustrates three selected subjects from the same repeated-holdout evaluation split generated using split seed 42. Case A receives an automatic positive decision in the first round, whereas Case B remains unresolved initially and receives an automatic negative decision after ensemble re-evaluation. Case C remains unresolved after both rounds and receives no automatic prediction. These examples illustrate the three decision paths of the framework. The accompanying behavioral summaries describe changes in the input features; they do not establish which features caused a particular prediction.

Screening Workflow and Outputs. Figure 6 illustrates how initial screening and ensemble reassessment lead to automatic positive decisions, automatic negative decisions, or abstention. Only unresolved subjects proceed to the second round. Subjects remaining unresolved after reassessment receive no automatic prediction and are excluded from coverage. The reported outputs summarize the same five repeated holdout splits used in the selective-screening analysis.

4.3 Ablation Studies

Table 2 summarizes component, modality, and selective-screening ablations under the experimental protocols described above.

Segment-Aware Pooling Ablation. As shown in Table 2 (a), removing segment-aware pooling decreases mean AUC from 0.67 to 0.64, F1 from 0.64 to 0.61, and accuracy from 0.65 to 0.56. The complete model also achieves higher AUC in all four paired runs, suggesting that the observed improvement is not driven by a single initialization. These results support preserving coarse temporal structure instead of summarizing the interview only through global pooling. Early, middle, and late representations retain information about when behavioral evidence occurs, while the late-minus-early representation describes changes across the interview. However, this ablation evaluates these representations jointly and does not isolate the contribution of the difference feature. The segments represent temporal positions rather than predefined topics or clinical stages.

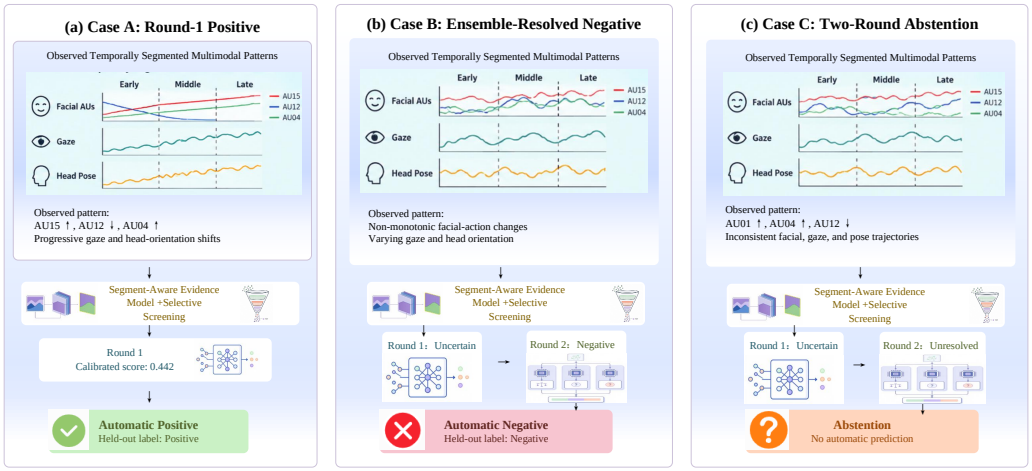

Figure 5: Selected DAIC-WOZ screening cases from the same repeated-holdout split. (a) First-round automatic positive. (b) Second-round automatic negative. (c) Abstention after both rounds. Behavioral summaries are derived from extracted features; curves are schematic rather than measured temporal traces. Reference labels are shown only for retrospective evaluation.



Figure 6: Two-round selective screening and aggregate outputs over five repeated holdout splits. Only unresolved subjects undergo ensemble reassessment; remaining unresolved subjects receive no automatic prediction. Error bars indicate sample standard deviations.

Graph Propagation Ablation. In Table 2 (a), we show that removing graph propagation yields higher mean AUC, F1, and accuracy than the complete model. The corresponding values are 0.70, 0.66, and 0.67 for the graph-free variant, compared with 0.67, 0.64, and 0.65 for the complete model. However, the AUC standard deviation increases from 0.04 to 0.12, indicating greater variability across the four training seeds. The complete model achieves higher AUC in three of four paired runs, whereas one high-scoring graph-free run substantially raises the graph-free mean. The average ranking and paired-run pattern therefore provide different perspectives on this comparison. Although graph propagation explicitly represents temporal and cross-modal interactions, the current results do not establish an average performance benefit. Its lower observed variability is descriptive evidence within this setting, rather than proof of greater robustness on unseen populations.

Modality Ablation. As shown in Table 2 (a), audio-only modeling achieves an AUC of 0.61, compared with 0.56 for visual-only modeling. Audio–visual modeling obtains the highest mean AUC, F1, and accuracy among the modality configurations. Using the values shown in the figure,

Table 2: Ablation studies of the representation model and selective screening strategy. (a) reports classification results over four random seeds on the DAIC-WOZ development set. (b) reports selective-screening results over five stratified repeated holdout splits. Arrows indicate whether higher or lower values are preferred. Abstained positive denotes the fraction of all positive evaluation subjects receiving no automatic decision. Nonempty splits counts splits with at least one automatic decision.

(a) Representation and modality ablations.

Analysis	Configuration	AUC ↑	F1 ↑	Accuracy ↑
Component	Full model	0.67 ± 0.04	0.64 ± 0.05	0.65 ± 0.06
	w/o graph propagation	$\mathbf{0.70 \pm 0.12}$	$\mathbf{0.66 \pm 0.09}$	$\mathbf{0.67 \pm 0.13}$
	w/o segment-aware pooling	0.64 ± 0.05	0.61 ± 0.03	0.56 ± 0.06
Modality	Audio only	0.61 ± 0.02	0.57 ± 0.02	0.62 ± 0.08
	Visual only	0.56 ± 0.05	0.59 ± 0.00	0.56 ± 0.04
	Audio–visual	$\mathbf{0.67 \pm 0.04}$	$\mathbf{0.64 \pm 0.05}$	$\mathbf{0.65 \pm 0.06}$

(b) Selective-screening ablation.

Setting	Coverage ↑	Abstention ↓	System FNR ↓	Abstained positive ↓	Nonempty splits ↑
One round	0.14 ± 0.20	0.86 ± 0.20	0.00 ± 0.00	0.84 ± 0.26	2/5
Two rounds	0.51 ± 0.19	0.49 ± 0.19	0.04 ± 0.09	0.64 ± 0.22	5/5

combining modalities improves AUC by 0.06 and F1 by 0.07 relative to audio alone. This pattern suggests that visual behavioral descriptors provide information complementary to acoustic evidence, even though visual-only discrimination is comparatively weak. It also distinguishes the value of a modality in isolation from its value when combined with other inputs. Nevertheless, the visual representation includes facial actions, gaze, and head pose together; this experiment cannot attribute the improvement to any single visual descriptor or establish that the same benefit will generalize beyond the current development set.

Selective-Screening Ablation. In Table 2 (b), we show that the two-round strategy increases mean coverage from 14.29% to 51.43% and reduces the abstained-positive rate from 84.00% to 64.00%. It produces at least one automatic decision in all five evaluation splits, compared with two splits for the one-round strategy. These gains are accompanied by an increase in mean system FNR from 0.00% to 4.00%, illustrating the trade-off between expanding automatic coverage and avoiding missed positive subjects. In particular, the zero FNR of the one-round variant must be interpreted alongside its high abstention rate, rather than as evidence of comprehensive positive identification. The two-round strategy still leaves a substantial fraction of positive subjects unresolved. This ablation evaluates the second-round mechanism as a whole, including ensemble reassessment and the increased cumulative coverage cap; it does not isolate the effect of ensembling alone.

5 Conclusion

We presented a segment-aware temporal evidence framework with two-round selective screening for multimodal depression screening. Ablations on the DAIC-WOZ development set support the effectiveness of segment-aware pooling and the complementarity of audio–visual evidence, while the contribution of graph propagation remains sensitive to training randomness. The two-round strategy increases automated prediction coverage and reduces unresolved positive cases, but entails a coverage–risk trade-off. These findings remain preliminary given the limited evaluation data. Future work will explore clinically grounded segmentation, more stable graph learning, external validation, and calibration across populations and recording conditions.

AI USE STATEMENT

The authors led the formulation of the research questions, development of the methodological framework, and experimental design, retaining authority over all final research decisions. Generative AI tools assisted with technical discussions, code drafting and debugging, interpretation of experimental outputs, manuscript drafting, translation and editing, and preliminary figure design. The authors conducted the experiments and take full responsibility for the implementation, reported results, interpretations, and final manuscript.

REFERENCES

Tuka Al Hanai, Mohammad Ghassemi, and James Glass. Detecting depression with audio/text sequence modeling of interviews. In *Proceedings of Interspeech*, pp. 1716–1720, 2018. doi: 10.21437/Interspeech.2018-2522.

Anonymous Authors. ECRL-MH: Evidence-conditioned reinforcement learning for explicit multi-disorder mental health screening, 2026. Anonymous manuscript.

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 59–66, 2018. doi: 10.1109/FG.2018.00019.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. COVAREP: A collaborative voice analysis repository for speech technologies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 960–964, 2014. doi: 10.1109/ICASSP.2014.6853739.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jonathan Gratch, Arno Hartholt, Margaux Lhommet, et al. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1061–1068, 2014.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Florian Eyben, Martin Wöllmer, and Björn Schuller. openSMILE: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462, 2010. doi: 10.1145/1873951.1874246.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Jonathan Gratch, Ron Artstein, Gale M. Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, and David Traum. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 3123–3128, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001. doi: 10.1046/j.1525-1497.2001.016009606.x.

Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1–3):163–173, 2009. doi: 10.1016/j.jad.2008.06.026.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, 1999. Manuscript dated March 26, 1999.

Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, and Nicholas Cummins. AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pp. 3–9, 2017.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, et al. 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*, pp. 3–12, 2019.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019. doi: 10.18653/v1/P19-1656.

Michel F. Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. AVEC 2013: The continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, 2013. doi: 10.1145/2512530.2512533.

Michel F. Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, 2016. doi: 10.1145/2988257.2988258.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.12021.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

A ADDITIONAL DETAILS

A.1 DATASET AND LABEL CONSTRUCTION

We use the DAIC-WOZ dataset, which contains semi-structured clinical interviews designed for depression analysis. Each subject is associated with a PHQ-8 score. Following common practice in DAIC-WOZ-based depression screening, we convert the PHQ-8 score into a binary label:

$$y = \begin{cases} 1, & \text{if PHQ-8} \geq 10, \\ 0, & \text{if PHQ-8} < 10. \end{cases} \tag{6}$$

Here, $y = 1$ denotes a positive depression-risk label and $y = 0$ denotes a negative label. The proposed method is evaluated as a screening model rather than a clinical diagnostic system.

Subject-Level Partitions. The official training partition contains 107 subjects, including 30 positive and 77 negative subjects. The development partition contains 35 subjects, including 12 positive and 23 negative subjects. The official test partition is not included in the reported evaluation because its labels are not publicly available. Interview windows are intermediate representations of a subject, not independent evaluation samples. Each subject has one reference label and one final screening output: positive, negative, or abstention.

Data Use in Model Selection and Screening. The training partition is used for model fitting, and development performance informs checkpoint selection. The selective-screening analysis subsequently partitions development subjects into calibration and evaluation subsets. Although these subsets are disjoint within each repetition, the evaluation subjects have already contributed to checkpoint selection. The resulting screening evaluation is therefore exploratory rather than an independent test of generalization. Feature construction does not use development labels.

A.2 TEMPORAL EVIDENCE CONSTRUCTION

Each interview is divided into overlapping 10-second windows with a 5-second stride. For each window, we construct four feature groups: 5-dimensional facial-action features, 5-dimensional gaze-related features, 4-dimensional head-pose features, and 54-dimensional enhanced acoustic features.

For each window–modality pair, the evidence file stores a feature vector, a quality score, and a validity mask. Invalid nodes are excluded during aggregation, while valid nodes are weighted by their quality scores.

Feature Groups. The facial-action features comprise the intensities of AU01, AU02, AU04, AU12, and AU15. The gaze-related representation contains five binocular gaze coordinates. The head-pose representation contains three rotation parameters and facial depth. The 54-dimensional acoustic representation combines four waveform-level statistics with within-window means and standard deviations of COVAREP and formant features. These four groups are treated as separate modalities in the graph, although the first three all describe visual behavior. No textual features are included in the reported audio–visual model.

Quality and Availability. For window t and feature group m, the stored evidence is $(\mathbf{x}_{t,m}, q_{t,m}, v_{t,m})$, where the three entries denote the feature vector, quality score, and validity mask, respectively. Visual quality is derived from frame-level tracking confidence. Acoustic validity records the availability of waveform, COVAREP, and formant streams. The mask determines whether evidence is eligible for aggregation, while the quality score controls the weight of valid evidence. Missing or unsuccessfully tracked inputs are therefore not interpreted as ordinary behavioral observations. All features are constructed independently for each subject.

A.3 Temporal Evidence Graph Details

The temporal evidence graph contains one node for each window–modality pair. Given T temporal windows and M modalities, the graph contains $T \times M$ nodes. We use three edge types: self-loops preserve node-specific information; temporal edges connect neighboring windows within the same modality; and cross-modal edges connect different modalities within the same temporal window.

Temporal edges are used to model behavioral continuity across time, while cross-modal edges allow evidence exchange between different feature groups. Typed graph attention is applied so that different edge types can use different transformations.

Node Representations and Edge Semantics. The four input feature groups have different dimensions and are projected into a shared 64-dimensional hidden space. Temporal encoding provides context within each group before graph propagation. In the full configuration, an interview with T windows therefore has $4T$ window–modality nodes before validity masking is applied during aggregation. A self-loop relates a node to itself, a temporal edge relates different windows of the same feature group, and a cross-modal edge relates different feature groups within one window. The cross-modal relation includes interactions among visual feature groups as well as interactions between acoustic and visual evidence.

In the graph update, $\mathbf{W}_r$ denotes the transformation associated with edge type r, and α_{ij}^r denotes the attention weight for a message from node j to node i under that relation. Distinguishing these relations allows temporal propagation and cross-modal exchange to use different parameters.

A.4 Segment-Aware Pooling Details

The interview is divided into early, middle, and late temporal segments according to window order. For each segment, we compute a quality-weighted pooled representation. Let $\mathcal{S}$ denote a segment and $\mathbf{h}_i$ denote the node embedding of node i. The pooled representation is:

$$\mathbf{z}_{\mathcal{S}} = \frac{\sum_{i \in \mathcal{S}} w_i \mathbf{h}_i}{\sum_{i \in \mathcal{S}} w_i + \epsilon}, \tag{7}$$

where w_i is derived from the corresponding quality score and validity mask.

The final subject-level representation concatenates global pooling, segment-level pooling, and the late-minus-early difference:

$$\mathbf{z} = [\mathbf{z}_{global}; \mathbf{z}_{early}; \mathbf{z}_{middle}; \mathbf{z}_{late}; \mathbf{z}_{late} - \mathbf{z}_{early}]. \tag{8}$$

This representation is then passed to a binary classifier.

Interpretation of Pooled Evidence. The node embeddings used in the pooling equation are the representations after graph propagation. The early, middle, and late segments are defined by temporal position in the interview, not by annotated topics or clinical phases. Global pooling summarizes the interview as a whole, while segment-level pooling retains a coarse indication of where evidence occurs. The late-minus-early term measures a difference between learned representations; it is not a direct measurement of a change in clinical state. Pooling retains coarse temporal information rather than the complete sequence of window-level observations.

A.5 Training Details

The temporal graph model is implemented in PyTorch and trained with binary cross-entropy loss for depression-risk classification (Paszke et al., 2019). The final reported temporal graph model uses

enhanced acoustic features and segment-aware pooling. Unless otherwise specified, the model is trained with a hidden dimension of 64, a learning rate of 1×10^{-4}, and early stopping based on development performance. We report results across multiple random seeds to characterize variation across runs.

Reported Training Configuration. The representation model uses two representation layers and a dropout rate of 0.2. The learning rate is 1×10^{-4}, and each batch contains one interview. Training runs for at most 30 epochs, with early stopping after 10 consecutive epochs without improvement. Checkpoint selection prioritizes development-set AUC. The reported classification experiments use training seeds 1, 2, 3, and 42, which control initialization and training randomness. For each configuration, results are summarized by the mean and sample standard deviation across these four runs.

Matched Representation Ablations. Component and modality configurations share the subject partitions, optimization settings, stopping criterion, and checkpoint-selection rule. Component ablations separately remove graph propagation or segment-aware pooling. Modality comparisons use audio-only, visual-only, and audio–visual inputs; the visual configuration includes facial actions, gaze, and head pose. Consequently, the modality comparison does not isolate the contribution of an individual visual feature group. Likewise, the pooling ablation assesses the pooled representations jointly rather than isolating the late-minus-early term.

A.6 Two-Round Selective Screening Details

The two-round selective screening module is designed to avoid forcing automatic decisions on uncertain subjects. The first round uses a single calibrated model for initial screening. Samples that cannot be confidently classified enter the second round, where a three-model ensemble is used for re-evaluation. If the second round still fails to satisfy the decision criteria, the sample receives no automatic prediction.

Let $\hat{p}$ denote the calibrated positive probability. A positive prediction is accepted when $\hat{p}$ is above the selected positive threshold. A negative prediction is accepted when $\hat{p}$ is below the selected negative threshold. For the ensemble round, we also use model disagreement as an uncertainty indicator. Samples exceeding the selected disagreement threshold remain unresolved.

Thresholds are selected on a calibration split by maximizing coverage under empirical constraints. In our experiments, the screening procedure uses positive-priority constraints, including a target bound on false-negative rate and a lower bound on positive precision. Figure 6 in the main text provides a visual summary of this two-round procedure and the reported screening outputs.

Repeated-Holdout Protocol. Screening is evaluated over five stratified holdout splits with split seeds 42–46. Each split assigns 21 development subjects to calibration (7 positive and 14 negative) and 14 to evaluation (5 positive and 9 negative). The split seed determines the calibration/evaluation partition; it is distinct from the training seed used to obtain a model. Predictions from models trained with seeds 1, 2, and 3 are fixed across these screening comparisons. The first round uses a single model, and the second uses their three-model ensemble to reassess only unresolved subjects.

Calibration Constraints. The reported parameter selection uses an empirical system-level FNR upper limit of 0.25 and a minimum positive precision of 0.40. The cumulative coverage caps are 0.40 for the first round and 0.70 for the two-round procedure. The second cap concerns cumulative coverage rather than an additional 70% of the subjects left unresolved by the first round. These are calibration settings, not guarantees that the same bounds hold on an evaluation subset or in another population. In particular, the selected decision rule is assessed on subjects not used to select its thresholds within that repetition.

Round-Wise Outputs. Let $\mathcal{A}_1$ contain subjects receiving a first-round prediction, $\mathcal{A}_2$ contain subjects first resolved in the second round, and $\mathcal{U}_2$ contain subjects unresolved after both rounds. These sets are disjoint and partition the evaluated subjects. The one-round procedure predicts only for $\mathcal{A}_1$, whereas the complete procedure predicts for $\mathcal{A}_1 \cup \mathcal{A}_2$. First-round decisions are retained; subjects

in $\mathcal{U}_2$ receive neither a positive nor a negative prediction. Thus, the extra subjects resolved by the second round contribute to cumulative coverage rather than replacing first-round predictions.

The one-round versus two-round comparison evaluates the complete second-round mechanism, including ensemble reassessment and the larger cumulative coverage cap. It does not isolate the causal effect of ensembling alone.

A.7 Selective Screening Metrics

We distinguish between metrics computed on automatically decided subjects and those computed over all evaluated subjects. Let N denote the total number of evaluated subjects and N_A the number receiving an automatic positive or negative decision. Coverage is defined as

$$\text{Coverage} = \frac{N_A}{N}. \tag{9}$$

The remaining $N - N_A$ subjects receive no automatic decision. The abstention rate is therefore

$$\text{Abstention} = \frac{N - N_A}{N} = 1 - \text{Coverage}\,. \tag{10}$$

Among ground-truth-positive subjects, let TP denote the number automatically classified as positive, FN the number automatically classified as negative, and U_+ the number receiving no automatic decision. These three counts partition the ground-truth-positive subjects. The system-level false-negative rate is defined as

$$\text{FNR}_{\text{sys}} = \frac{\text{FN}}{\text{TP} + \text{FN} + U_+}. \tag{11}$$

The abstained-positive rate, denoted by APR, measures the proportion of ground-truth-positive subjects for whom the system abstains:

$$\text{APR} = \frac{U_+}{\text{TP} + \text{FN} + U_+}. \tag{12}$$

Accepted precision, accepted recall, and accepted F1 are computed only on automatically decided subjects. In contrast, FNR_{sys} and APR use all ground-truth-positive subjects as their denominator, including those receiving no automatic decision. These metrics therefore describe different aspects of selective screening performance.

Accepted-Subset Metrics. Let FP and TN denote the numbers of negative subjects automatically classified as positive and negative, respectively, and let U_- denote the number of negative subjects for whom the system abstains. Then

$$N_A = \text{TP} + \text{FP} + \text{TN} + \text{FN}, \qquad N = N_A + U_+ + U_-. \tag{13}$$

When their denominators are nonzero, the accepted-subset metrics are

$$\text{Precision}_A = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall}_A = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{14}$$

$$\text{F1}_A = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{15}$$

These metrics exclude abstentions and therefore do not describe the proportion of all positive subjects identified by the system. An empty automatic decision set does not establish perfect classification, even when its system-level FNR is zero.

Accounting for All Positive Subjects. The fraction of all positive subjects automatically identified as positive satisfies

$$\frac{\text{TP}}{\text{TP} + \text{FN} + U_+} = 1 - \text{FNR}_{\text{sys}} - \text{APR}. \tag{16}$$

When accepted recall is defined, it also satisfies

$$\text{Recall}_A(1 - \text{APR}) = 1 - \text{FNR}_{\text{sys}} - \text{APR}. \tag{17}$$

The latter identity holds within each evaluation split. Products of separately averaged metrics need not satisfy the same identity. This distinction matters when comparing accepted-subset performance with system-level outcomes.

Aggregation and Nonempty Splits. For a metric g_k computed in repetition k, the reported mean and sample standard deviation are

$$\bar{g} = \frac{1}{K}\sum_{k=1}^{K} g_k, \qquad s_g = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(g_k - \bar{g})^2}. \tag{18}$$

Here $K = 4$ for the classification runs and $K = 5$ for the repeated screening splits. The nonempty-split count records the number of evaluation splits with $N_A > 0$; it does not indicate that an evaluation split satisfies every calibration constraint. A split with no automatic decisions has zero coverage and unit abstention, and contributes to the reported system-level summary. All screening evaluation subsets contain five positive subjects, so the denominators of system FNR and APR are nonzero.

A.8 Detailed Two-Round Screening Results

Table 3 reports four system-level screening metrics across five stratified repeated holdout splits. In addition to coverage, abstention, and system FNR, it reports the proportion of positive subjects receiving no automatic decision.

Table 3: Two-round selective screening results across five stratified repeated holdout splits of the DAIC-WOZ development partition. Values are means and sample standard deviations. System FNR and Abstained Positive Rate use all ground-truth-positive subjects as their denominator.

Metric	Mean	Std.
Coverage	0.51	0.18
Abstention Rate	0.49	0.18
$\mathrm{FNR}_{\mathrm{sys}}$	0.04	0.09
Abstained Positive Rate	0.64	0.22

Table 3 summarizes the two-round strategy under the same repeated-holdout protocol used in Table 2 (b). The system automatically decides 51.43% of evaluated subjects and abstains on 48.57%. The system-level false-negative rate is 4.00%, while 64.00% of ground-truth-positive subjects receive no automatic decision. The low false-negative rate must therefore be interpreted alongside the substantial abstention among positive subjects.

Resolution of Positive Cases. Using the reported mean system FNR of 0.04 and mean APR of 0.64, the mean fraction of all positive evaluation subjects receiving an automatic positive prediction is approximately $1 - 0.04 - 0.64 = 0.32$. The corresponding fraction for the one-round strategy is approximately $1 - 0.00 - 0.84 = 0.16$. These are derived system-level fractions, not accepted recall. They complement coverage by distinguishing positive cases correctly identified, misclassified as negative, and left unresolved.

Granularity and Repeated Subjects. Each screening evaluation subset contains only 14 subjects, including five positive subjects. Consequently, one additional automatic decision changes split-level coverage by $1/14$, or approximately 7.14 percentage points. One additional false negative changes split-level system FNR by $1/5$, or 20 percentage points. Repeated holdout subsets can overlap, so the five evaluations do not represent 70 distinct subjects. Variation across splits describes sensitivity to the partition within this development set, rather than uncertainty estimated from five independent datasets.

A.9 Discussion of Screening Trade-offs

The selective screening results should be interpreted as a trade-off between automatic coverage and screening risk. Higher coverage means that more subjects receive automatic decisions, but it may increase the chance of automatically misclassifying positive subjects as negative. Lower coverage can reduce forced decisions but leaves more subjects without an automatic prediction.

In this work, abstention does not mean that a subject is healthy or that no follow-up is needed. It only means that the automatic system does not provide a sufficiently reliable screening decision. This distinction is important for risk-sensitive applications such as mental health screening.

A.10 Scope of the Qualitative Examples

The three selected cases come from the evaluation subset generated with split seed 42. They illustrate first-round positive prediction, second-round negative prediction, and abstention after both rounds. The cases demonstrate possible paths through the procedure; they do not estimate how frequently each path occurs across the dataset. Behavioral summaries describe the extracted inputs. The curves in the main-text case illustration are schematic, whereas Figure A.1 reports recorded stage-wise means of selected features. Reference labels are shown for retrospective evaluation. Neither the summaries nor the schematic curves establish that a particular behavioral feature caused a prediction.

Figure A.1 presents stage-wise summaries of selected visual features for three subjects, together with their screening outcomes. Audio features are also used by the model but are not shown. These descriptive summaries do not establish causal relationships between individual features and predictions.

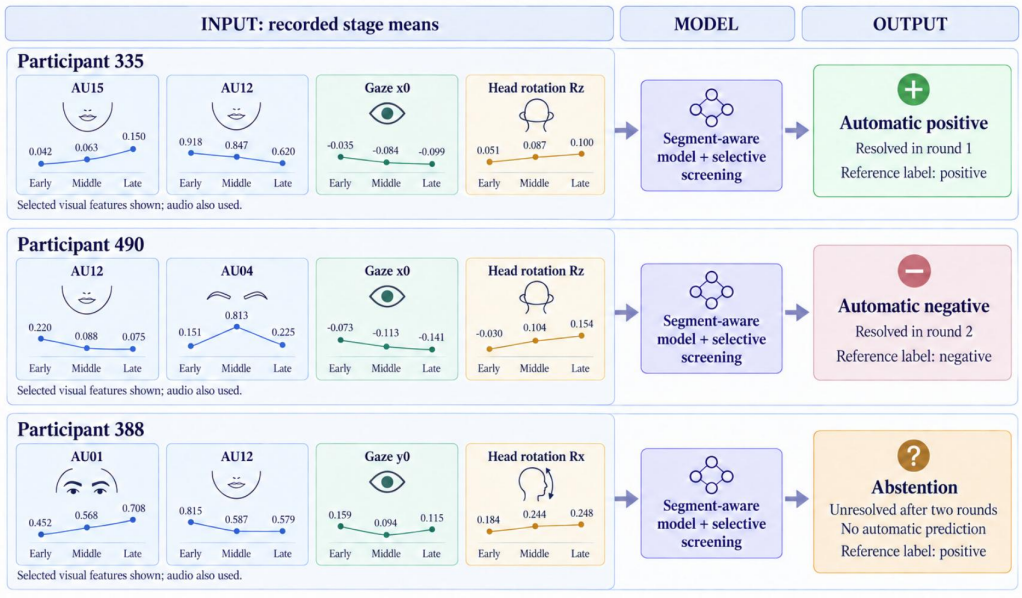

Figure A.1: Stage-wise means of selected visual features and corresponding screening outcomes. Participant 335 receives an automatic positive prediction in the first round, Participant 490 receives an automatic negative prediction in the second round, and Participant 388 remains unresolved after both rounds. The displayed features summarize early, middle, and late interview segments. Audio features are also used but are not shown. Reference labels are provided for retrospective evaluation.

A.11 Additional Limitations

The current study is preliminary and is evaluated only on the DAIC-WOZ development data. Its generalization should therefore be assessed on independent datasets. In addition, the early/middle/late division is based on temporal position rather than clinically annotated interview phases. The selective screening mechanism reduces forced decisions but does not provide a clinical diagnosis; abstained subjects still require appropriate follow-up. Future work will investigate clinically grounded temporal segmentation, stronger multimodal encoders, and external validation across datasets and recording conditions.

Empirical Risk Control and Interpretation. The screening constraints are evaluated on finite calibration subsets. They do not constitute a distribution-free guarantee of false-negative control on new subjects. Calibration, threshold selection, and checkpoint selection must also be distinguished: holding out subjects from threshold selection does not remove their earlier use in checkpoint selec-

tion. Independent validation is needed to assess both probability reliability and the coverage–risk trade-off outside the present development set.

The representation comparisons likewise have limited scope. Lower observed variability across four training seeds does not establish robustness to population or recording shifts. The component and modality ablations assess the tested configurations as a whole and do not establish feature-level causal explanations.