EVIDENCE-ORIENTED GRPO AND REFERRAL-AWARE SCREENING FOR MENTAL HEALTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal mental-health screening combines a high-stakes decision problem with heterogeneous audio, visual, and textual signals. Discriminative models can aggregate these signals effectively but usually expose only a score, whereas language models can consume textual evidence but need an interface that preserves the information learned by the discriminative stage. We introduce **EARN** (**E**vidence-oriented **A**nd **R**eferral-aware scree**N**ing using GRPO), which produces a disease probability and a 27-dimensional, symptom-named fingerprint from a shared representation. Prototype binding aligns the ordered fingerprint with a 12×27 evidence library. Fingerprint verbalization renders its entries as protocolized phrases for a text-only Qwen2.5-Omni-7B reader. Evidence-oriented GRPO optimizes the discriminative evidence head separately from the reader. On the E-DAIC official split, the verbalized reader attains AUROC 0.7406, whereas the same fingerprint supplied as raw floats attains 0.5611 after fine-tuning. The proposed audio-visual pipeline obtains AUROC 0.741 and F1 0.6286 on E-DAIC, and F1 0.9527 on D-Vlog under the reported protocols. A decoupled-decision analysis further shows the trade-off between referral rate and missed positives. On E-DAIC, referring 21.4% of cases reduces automatic-channel misses from five to two. Because this analysis contains only 17 positive cases, we present it as an exploratory operating curve rather than a clinical-performance claim. Code and the evidence library are available anonymously at `https://anonymous.4open.science/r/EARN-0F2D`.

1 INTRODUCTION

Multimodal foundation models have accelerated automated mental health screening. Clinical interviews for disorders such as depression and post-traumatic stress disorder (PTSD) contain heterogeneous temporal signals, including acoustic cues, facial behavior, and transcribed speech. Discriminative models can efficiently fuse these modalities and predict screening probabilities, whereas multimodal large language models (MLLM) reason over textual descriptions to produce screening decisions. However, these two paradigms remain poorly integrated. Discriminative models aggregate multimodal evidence effectively but typically expose only a scalar score, providing no symptom-level evidence for human auditing. Although existing adapters and projection layers can inject discriminative hidden states into MLLM as soft prompts, they mainly provide numerical compatibility and do not yield human-readable evidence.

Recent studies, including Psy-LLM, MentalGLM, and ECRL-MH (Dai et al., 2025, Zhai et al., 2025, Anonymous, 2026), have introduced large language models into mental health screening and achieved promising reasoning performance. Nevertheless, three limitations remain. **i) Unreadable intermediate semantics.** High-dimensional numeric fingerprints are not grounded in explicit symptom labels. Consequently, the evidence used by the model remains inaccessible to human reviewers, making it difficult to identify the specific symptom cues underlying a prediction. **ii)Transmittable but unauditable interfaces.**Projection and adapter layers can map discriminative representations tosoft prompt vectors consumable by language models. However, this mapping obscures the correspondence between the representations and symptom concepts, preventing users from auditing the evidence behind a screening decision. **iii)Limited human-in-the-loop decision support.** Existing end-to-end systems directly produce final predictions without explicitly identifying cases with am-

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.

biguous evidence or uncertain reasoning. This limits their applicability to high-risk workflows that require automated prescreening followed by human review.

To address these limitations, we propose EARN, a framework that preserves the multimodal fusion capability of a discriminative model while redesigning its interface to a language module. An availability-aware multimodal encoder processes the temporal signals and uses a shared representation to predict both the screening probability and a 27-dimensional, symptom-labeled feature fingerprint. A prototype-binding module aligns this fingerprint with a 12×27 evidence bank. Fingerprint textualization then converts the aligned fingerprint into standardized symptom phrases that can be consumed by a text-only large language model. We further optimize the evidence head separately using evidence-oriented Group Relative Policy Optimization (GRPO). Finally, a decoupled decision mechanism triggers referral based on reasoning disagreement and distance to the decision boundary. For controlled ablations, the underlying discriminative representation remains fixed while only the evidence presentation is varied, thereby separating evidence content from interface form.

During training, prototype contrastive learning enforces semantic alignment between the feature fingerprint and the evidence bank. A multidimensional evidence-quality reward further improves the evidence head with little additional training overhead. Consequently, EARN supports screening inference, interpretable evidence generation, and referral of uncertain cases within a unified framework.

Our contributions are summarized as follows.

- We propose a **structured evidence interface** that jointly learns a 27-dimensional symptom fingerprint, aligns it with a 12×27 evidence bank through prototype contrastive learning, and verbalizes the aligned coordinates as graded symptom phrases. This design exposes a reader-facing and human-auditable form of an intermediate discriminative representation while keeping the underlying decision path fixed.
- We develop an **evidence-aware optimization and referral framework**. Evidence-oriented GRPO independently trains the evidence head with grounding, attribution, consistency, pairwise, and contrastive-margin signals, while a decoupled decision rule refers cases using reasoning disagreement and distance to the decision boundary.
- Extensive experiments on E-DAIC, D-Vlog, MODMA, and PDCH demonstrate that EARN obtains AUROC 0.7406 for the verbalized reader and F1 scores of 0.6286 on E-DAIC and 0.9527 on D-Vlog under the reported protocols. Evidence-form, library, modality, and training ablations further support the contribution of the proposed components, while the small-sample referral analysis is presented as exploratory

2 Related Work

Multimodal Screening. Early work used handcrafted acoustic and visual statistics (Williamson et al., 2016, Song et al., 2018). Later studies introduced end-to-end models, including iterative attention (Jaegle et al., 2021), non-verbal cues (Gimeno-Gómez et al., 2024), sub-attentional fusion (Wei et al., 2022), temporal models (Yang et al., 2025a, He et al., 2024, Ye et al., 2025, Xing et al., 2024), multiple-instance and interaction networks (Shangguan et al., 2023, Zhou et al., 2023), and vlog-oriented models (Tao et al., 2024a), (Tao et al., 2024b). MMFformer (Haque et al., 2025) is a relevant recent multimodal baseline. We report E-DAIC, D-Vlog, MODMA, and PDCH results. Literature values in Table 1 are retained as context and should not be read as reimplementations under a common protocol.

LLMs in Mental Health. Generative systems use transcripts or structured descriptions as LLM inputs, including Psy-LLM (Lai et al., 2023), MentalGLM (Zhai et al., 2025), Psyche-R1 (Dai et al., 2025), HumanOmniV2 (Yang et al., 2025b), and multi-agent variants (Xu et al., 2025b, Ozgun et al., 2025). EARN is complementary to these approaches, because it studies how an intermediate discriminative representation should be rendered before an LLM reads it.

Evidence-grounded Decisions and RL. Decision-support systems often ground outputs in external knowledge or raw text. Here, evidence originates from an intermediate discriminative representation and is aligned to an ordered symptom library. Group-relative policy optimization (Shao et al., 2024) has been used for structured generation (Feng et al., 2025, Dai et al., 2025, Yang et al., 2025b).

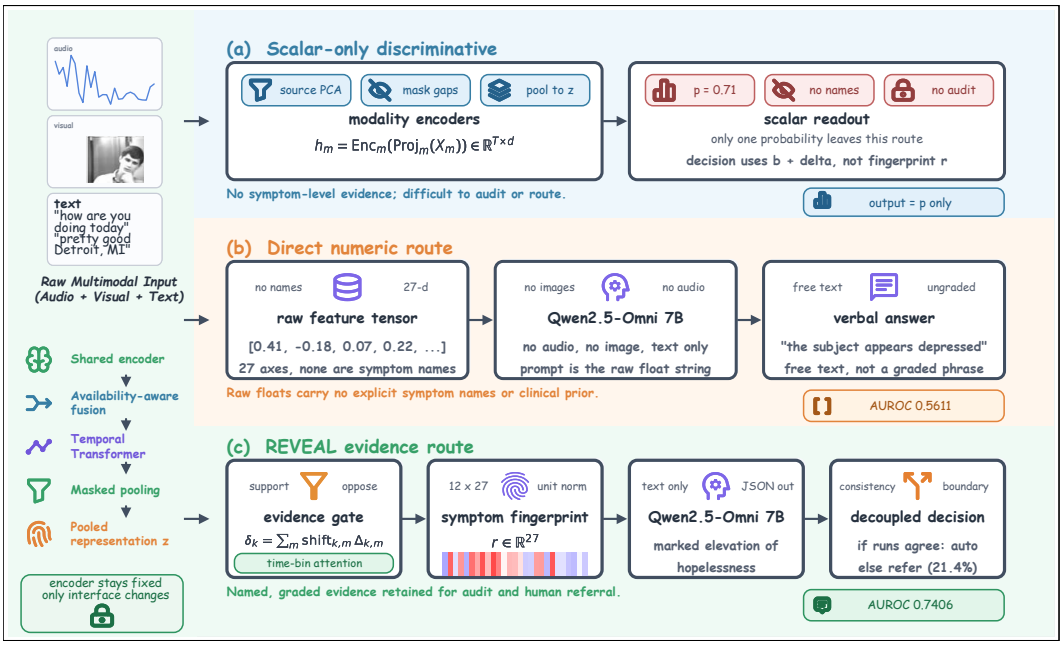


Figure 1. Three routes from the same multimodal input to a screening decision. (a) returns only a scalar probability, (b) passes 27 raw floats, and (c) keeps named phrases and can refer the case for human review.

EARN puts the group-relative reward on the discriminative evidence stage, and the language reader is trained in its own stage. The closest system, Anonymous (2026), uses explicit intermediate evidence and group-relative optimization but standardizes fixed handcrafted quantities. Our focus is the reader-facing form of a learned fingerprint, evaluated through the evidence-form and library ablations. Figure 1 places three routes side by side from the same multimodal input. The first route keeps only a scalar probability, the second copies 27 raw fingerprint values into the prompt, and EARN turns those values into symptom-named phrases before the reader decides. The figure separates what the discriminative stage computes from the text the reader is allowed to see.

3 Method

3.1 Overview

Given a participant's multimodal time series (audio, visual, and transcript), the discriminative network emits a disease probability p and a 27-dimensional fingerprint r. The fingerprint is rendered as structured, symptom-named phrases and supplied to an LLM reader adapted with LoRA and GRPO. An evidence library provides an ordered prototype space for r, and evidence-oriented GRPO trains the evidence head before reader adaptation (Section 3.5). Figure 2 follows one participant from the raw recording to the screening decision. The upper row encodes audio, visual, and text, applies the availability gate, and aligns a 27-dimensional symptom fingerprint with the evidence library. The lower row sends that fingerprint through language review to the answer. Solid arrows trace the data path, and the dashed arrow is the evidence-stage reward on the evidence head.

3.2 Encoding and Availability Gating

Because datasets differ in feature provenance, a per-source projector (PCA per source domain) maps the raw features of modality $m \in \{a, v, t\}$ into a shared latent space,

$$h_m = \mathrm{Enc}_m\big(\mathrm{Proj}_m(X_m)\big) \in \mathbb{R}^{T\times d}. \tag{1}$$

Fusion is availability-aware. With availability indicator $s \in \{0,1\}^3$, we mask the pre-softmax gate logits,

$$\ell = W_g[h_a, h_v, h_t, s] + b_g, \qquad \ell_m \leftarrow -10^4 \text{ if } s_m = 0, \qquad g = \mathrm{softmax}(\ell), \tag{2}$$

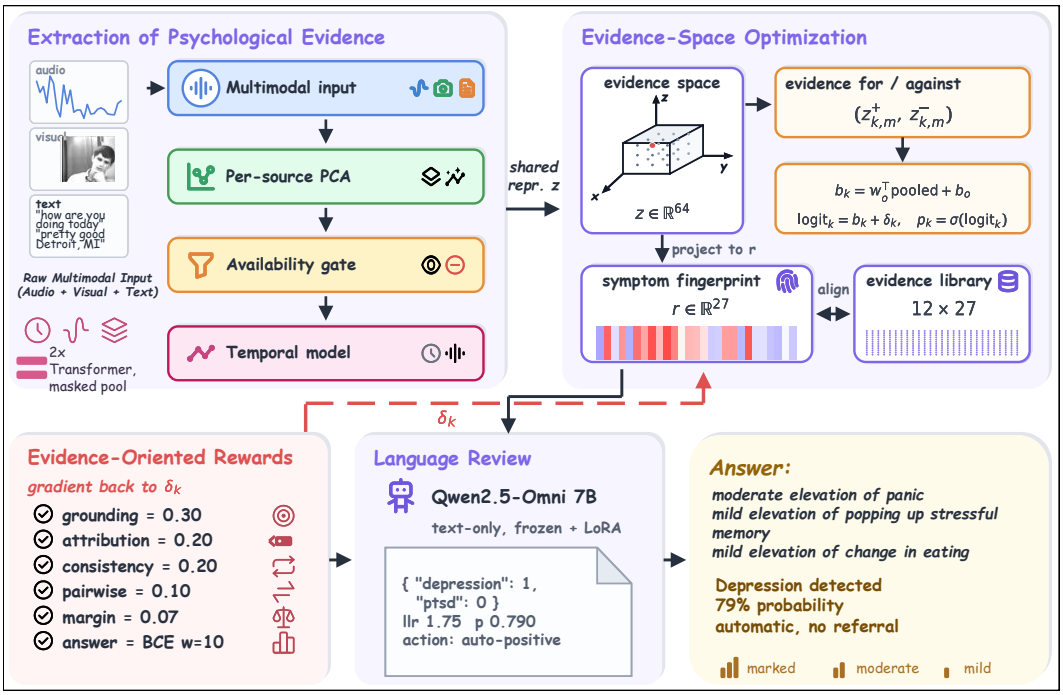


Figure 2. The architecture of EARN. Input passes through a symptom fingerprint and language review to the screening answer. Solid arrows mark the data path, and the dashed arrow marks the evidence-stage reward.

so missing modalities receive negligible weight. The fused stream is $z = \sum_m g_m \odot h_m + W_i u + b_i$, where u is an interaction term. Two temporal Transformer layers and masked average pooling produce the participant-level representation $\text{pooled} = \sum_t m_t z_t / \sum_t m_t \in \mathbb{R}^d$.

3.3 Evidence Gating for Support and Opposition

For each target condition k and each audio or visual evidence branch $m \in \{a, v\}$, we learn a pair of heads that produce evidence *supporting* and *opposing* that condition,

$$z^{\text{sup}}_{k,m} = f_{\text{sup}}(\text{pooled}) \in \mathbb{R}, \qquad z^{\text{con}}_{k,m} = f_{\text{con}}(\text{pooled}) \in \mathbb{R}. \tag{3}$$

Their difference is the net evidence for that modality. Learnable gate weights $\text{shift}_{k,m}$ are initialized to zero, allowing the model to begin from the base classifier.

$$\delta_k = \sum_m \text{shift}_{k,m}\big(z^{\text{sup}}_{k,m} - z^{\text{con}}_{k,m}\big), \qquad \text{logit}_k = b_k + \delta_k, \qquad p_k = \sigma(\text{logit}_k), \tag{4}$$

where $b_k = w_o^\top \text{pooled} + b_o$. The signed difference δ_k supports a contrastive objective between related conditions (Section 3.5). Each evidence head also uses per-bin temporal attention $\alpha_{k,m} = \text{softmax}(e_{k,m})$, where $e_{k,m} = g_\alpha(\text{Enc}_m(X_m)) \in \mathbb{R}^T$, as a temporal grounding signal.

Symptom Fingerprint. The same pooled representation feeds a projection head whose output is L2-normalized,

$$r = \text{Normalize}\big(W_s\,\text{pooled} + b_s\big) \in \mathbb{R}^{27}, \qquad \|r\|_2 = 1. \tag{5}$$

The decision uses $b_k + \delta_k$, not r. The fingerprint r is used for the prototype objective (Section 3.4) and verbalization (Section 3.6). This separation lets the evidence-form ablation change what the reader receives without changing the discriminative decision.

3.4 Disease-Prototype Contrastive Learning

The evidence library specifies 12 conditions (depression, PTSD, ADHD, ASD, OCD, ...) in a fixed order over 27 symptoms. $P = [p_1, \ldots, p_{12}] \in \mathbb{R}^{12\times 27}$, where $p_{c,i} = p(\text{symptom}_i \mid \text{condition}_c)$ and

each row is normalized. Cosine similarity is scaled by $\tau = 0.1$. For a positive prototype c^+ sampled from the condition-specific positive pool C^+, we minimize InfoNCE,

$$L_{\text{proto}} = -\log \frac{\exp\big(\text{sim}(r, p_{c^+})\big)}{\sum_{c=1}^{12} \exp\big(\text{sim}(r, p_c)\big)}, \qquad L = L_{\text{BCE}} + \lambda_{\text{proto}} L_{\text{proto}} + L_{\text{GRPO}}, \tag{6}$$

with $\lambda_{\text{proto}} = 0.1$. The constraint aligns r with the library's ordered symptom basis while leaving the decision path unchanged. It supplies a protocolized semantic interface. It does not by itself validate each coordinate as a clinical symptom measurement.

3.5 Evidence-oriented GRPO

Training runs for 20 BCE warm-up epochs and then for 70 epochs of evidence-oriented GRPO. In the evidence stage, the reward is defined on properties of the intermediate evidence rather than on the correctness of the final answer. For each target condition, G candidates are sampled per participant and assigned the reward

$$R_k = w_{\text{lab}} R_{\text{label}} + w_{\text{gr}} R_{\text{ground}} + w_{\text{at}} R_{\text{attrib}} + w_{\text{co}} R_{\text{consist}} + w_{\text{pa}} R_{\text{pair}} + \lambda_d R_{\text{margin}}, \tag{7}$$

where the terms measure attention-saliency agreement (grounding), agreement between the gate direction and the decision (attribution), stability across forward passes (consistency), and pairwise candidate agreement, together with the contrastive margin defined below. Group normalization of these rewards supplies the advantages A_i used in

$$L_{\text{GRPO}} = -\frac{1}{G} \sum_i A_i \cdot \frac{1}{|y_i|} \sum_t \log \pi_\theta\big(y_i^t \mid y_i^{<t}, x\big) - \beta_{\text{KL}} \text{KL}\big(\pi_\theta \| \pi_{\text{ref}}\big). \tag{8}$$

With $w_{\text{lab}} = 0$, the label term is left out of R_k. A separate BCE anchor of weight 10 carries the label signal and stabilizes ranking. The remaining terms shape the evidence head toward evidence that is grounded, consistent, and attributable.

Contrastive Margin. Reusing the net evidence of Section 3.3, we require the target disease to dominate any competitor,

$$\text{margin} = \text{clip}\big(\delta_{\text{target}} - \max_{k \neq \text{target}} \delta_k,\ -1,\ 1\big) \in [-1, 1], \tag{9}$$

and weight it by the direction of the sampled candidate, $q_{\text{ev}} = \sigma(\delta_k)$.

$$R_{\text{margin}} = \begin{cases} q_{\text{ev}} \cdot \text{margin}, & \hat{y} = 1, \\ (1 - q_{\text{ev}}) \cdot (-\text{margin}), & \hat{y} = 0. \end{cases} \tag{10}$$

A positive prediction is therefore encouraged to rest on evidence that is both strong and condition-specific. The margin term is kept subordinate through a development-set controller that constrains λ_d below the other evidence weights. The controller settings are given with the implementation details.

3.6 Verbalized Injection of Discriminative Evidence

We scan the fingerprint $r \in \mathbb{R}^{27}$ in descending order of $|r_i|$ and convert each retained coordinate into a tiered phrase.

$$\text{level}(v) = \begin{cases} \text{marked}, & |v| \geq 0.35, \\ \text{moderate}, & |v| \geq 0.25, \\ \text{mild}, & |v| \geq 0.15, \\ \text{discarded}, & \text{otherwise}, \end{cases} \tag{11}$$

The sign determines whether a phrase describes an elevation or reduction. The evidence document contains a fixed lead-in, the retained phrases in descending order (e.g., "marked elevation of hopelessness"), and a screening-only disclaimer. It is injected as plain text into Qwen2.5-Omni-7B (Xu et al., 2025a), fine-tuned with LoRA (Hu et al., 2021) ($r = 8$, $\alpha = 16$ on q/k/v/o projections) under

GRPO with the prompt *Use ONLY the protocolized evidence below to screen for depression and PTSD. Reply with exactly one JSON containing the keys depression and ptsd.*

Choice of Reader. The reader is text-only. Images and audio are empty, and the evidence document is its sole input. We use Qwen2.5-Omni-7B because its audio tower also provides the semantic features consumed by the discriminative network. This choice holds the feature and reader family fixed. A 3B checkpoint from the same family defines the weaker-representation regime in Table 6.

Readout. For a continuous, sampling-noise-free readout we use the log-likelihood ratio at the label position,

$$s = \log P_\theta(y = 1 \mid e) - \log P_\theta(y = 0 \mid e), \qquad e = \text{verbalized evidence}. \tag{12}$$

3.7 Decoupled Decision

Inference variation can arise from LLM sampling and from independently trained discriminative models. For S inference runs yielding $\hat{y}_1, \ldots, \hat{y}_S$, we define consistency as the share of the majority decision.

$$c = \frac{1}{S} \max_k \sum_s \mathbf{1}[\hat{y}_s = k], \qquad k \in \{0, 1\}. \tag{13}$$

If $c \geq \tau$, the sample takes the automatic channel. Otherwise, it is referred for human review. With $\tau = 1$, only unanimous runs pass automatically. The design prioritizes missed positives because false negatives are especially consequential in screening. A second signal, the evidence-boundary band, refers samples with the smallest $|s - \mathrm{th}|$, where th is the decision threshold. These signals respectively capture disagreement across runs and evidence near the threshold.

4 Experiments

4.1 Setup

We evaluate on four datasets spanning languages and protocols. E-DAIC, an English interview set with the official $163/56/56$ split, 56 test participants, and 17 positives, and D-Vlog, an English vlog set with split $155/27/56$, are the main domains. MODMA (Cai et al., 2020) and PDCH (Cao et al., 2025) are Chinese-interview indicators obtained by per-dataset training with five folds and an independent development split. All EARN ablations use the stated split and protocol. Literature values are not reimplementations. Available implementation details are reported after the experiments.

4.2 Main results

Table 1 shows EARN with ECRL (Anonymous, 2026) and literature values summarized there. On E-DAIC, the audio-visual evidence chain reaches AUROC 0.741 and F1 0.6286. Its recall is 0.647 and precision is 0.611. On D-Vlog, it obtains F1 0.9527 ± 0.0088 and AUROC 0.9422 ± 0.0242. These cross-paper rows provide context rather than a controlled comparison, because preprocessing, tuning, and evaluation details may differ. After per-dataset training, the Chinese-dataset F1 scores are 0.8679 (MODMA) and 0.7654 (PDCH, additional results). Table 1 reports the official E-DAIC and D-Vlog splits, contextual literature values, modality labels, and the bootstrap interval for the EARN E-DAIC result.

4.3 Ablation Studies

4.3.1 Evidence forms

Figure 3 compares reader AUROC on the E-DAIC official split when only the injected evidence form changes. Raw floats stay at 0.3228 zero-shot and 0.5611 after fine-tuning. The transcript, the learned projection, and the symptom-named phrases fall in 0.7330–0.7406 after fine-tuning, and the reported paired bootstrap does not separate them ($p \geq 0.50$). Verbalization therefore keeps an inspectable text surface while matching the other successful reader inputs in this comparison. Table 2 lists the same zero-shot and fine-tuned scores with the discriminative pipeline and readout held fixed.

Table 1. Main results on E-DAIC and D-Vlog. Comparison values are restricted to 2024–2026 literature.

			E-DAIC		D-Vlog
Method	Modality	AUROC	F1	(P / R)	F1
Gimeno-Gómez et al. (2024)	V	/	0.50 ± 0.08	0.58 ± 0.05 / 0.49 ± 0.07	/
Yang et al. (2025a)	V	/	0.417	0.714 / 0.294	/
Tao et al. (2024a)	/	/	/	/	0.7351
Ye et al. (2025)	/	/	/	/	0.7644
Xing et al. (2024)	/	/	/	/	0.7666
He et al. (2024)	/	/	/	/	0.5900
ECRL (Anonymous, 2026)	V	/	0.562	0.600 / 0.529	/
Gimeno-Gómez et al. (2024)	AV	/	0.56 ± 0.12	0.59 ± 0.07 / 0.58 ± 0.13	/
ECRL (Anonymous, 2026)	AV	/	0.595	0.550 / 0.647	0.9075
Haque et al. (2025)	AV	/	/	/	0.9092
Ours (LLM evidence chain)	**AV**	**0.741**	**0.6286**	**0.611 / 0.647**	**0.9527**

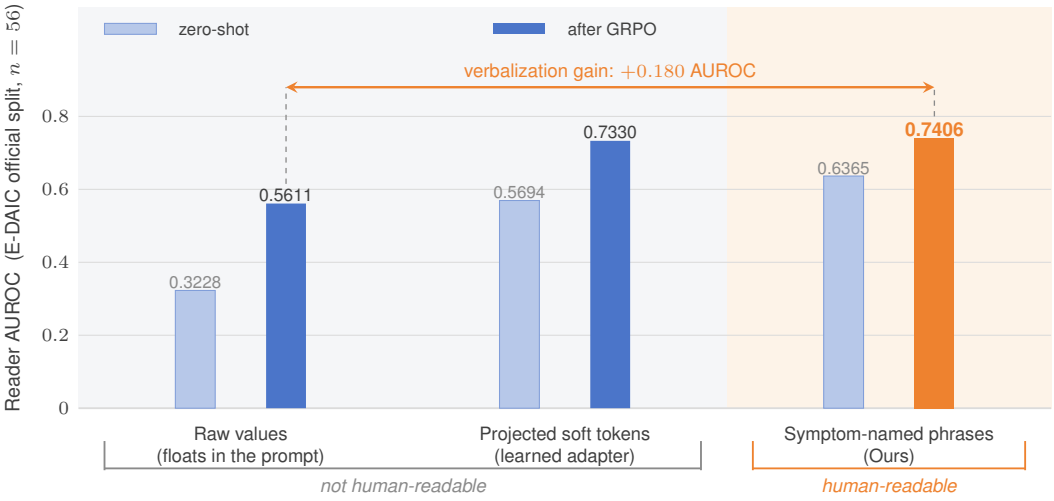


Figure 3. Reader AUROC by evidence-input form on the E-DAIC official split. Light bars are zero-shot, dark bars follow GRPO, and only symptom-named phrases are human-readable.

Table 2. Evidence-input comparison on E-DAIC.

Injected content	Readable	Zero-shot	Fine-tuned
The same 27-dimensional fingerprint, two injection channels			
Raw values (floats in the prompt)	no	0.3228	0.5611
Projected soft tokens (learned adapter)	no	0.5694	0.7330
What the language model is given instead			
Symptom-named phrases (Ours)	**yes**	0.6365	**0.7406**

Figure 4 shows the three surfaces for one E-DAIC participant. The left card is the raw 27-float fingerprint, the middle card is the verbatim ASR excerpt, and the right card renders the same fingerprint as symptom-named phrases. Only that card can be read as clinical evidence, and its language-AUROC label matches the fine-tuned column of Table 2.

4.3.2 Training Schedule

To assess the contribution of evidence-oriented optimization, we compare the full training schedule with a 20-epoch warm-up-only configuration in Table 3. The full schedule yields higher discriminative-head and reader AUROC, which is consistent with a benefit from the evidence-oriented stage. Because the comparison also changes the training duration, however, it should not be read as an isolated reward-component test.

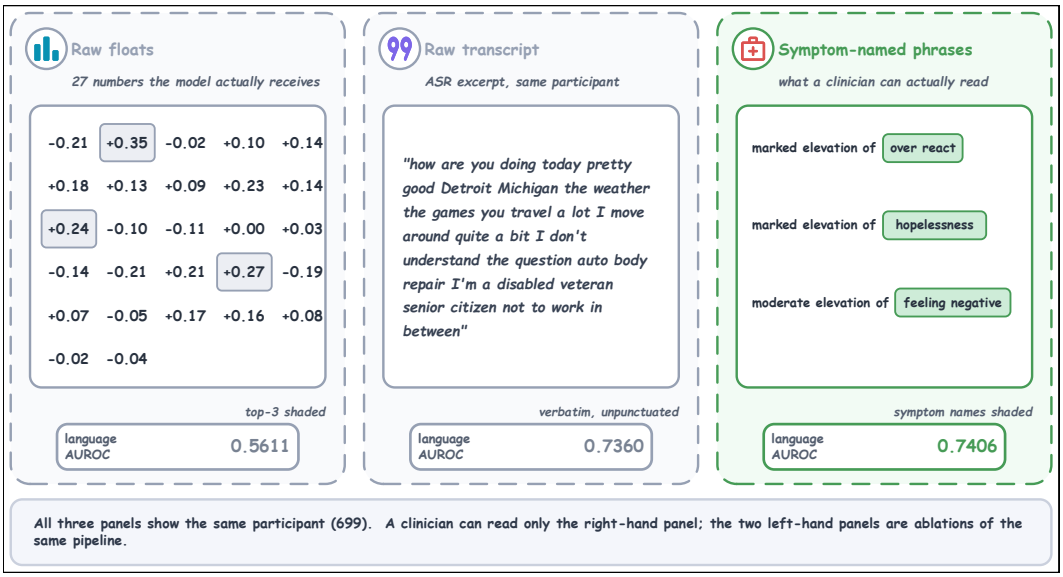


Figure 4. Evidence surfaces for one E-DAIC participant. Only the right card, written as symptom-named phrases, can be read as clinical evidence.

4.3.3 Evidence library

To determine whether the ordered prototype space contributes beyond the reader itself, Table 4 compares the full library with no-library and shuffled controls. Across the reported settings, the full library improves the reader score while leaving the decision path separate. This supports the use of the ordered prototype space for this reader, but does not independently validate individual phrase semantics.

4.3.4 Modality inputs

We further examine modality dependence in Table 5. Audio-visual input gives the highest reader AUROC (0.741) in this configuration, whereas vision-only and text-only inputs are lower. Audio-visual fusion does not improve the discriminative-head score over audio alone, separating the reader benefit from the discriminative score.

4.3.5 Representation regime

Table 6 examines how the component behaves across representation regimes. The gains for mel/OpenSMILE, Qwen-3B latent-32, and Qwen-3B latent-128 are $+0.133$, $+0.074$, and approximately zero, respectively. These observations suggest a representation-dependent trend rather than a fitted scaling relationship. Figure 5 lines up the audio energy, dominant fingerprint entries, verbalized evidence, and screening action for representative E-DAIC participants.

4.4 Decoupled decision

Figure 6 reports the operating points of the decoupled decision. By construction, referral only removes cases from the automatic channel, so its missed-positive count is monotonically non-increasing with the referral rate. On D-Vlog, 7.1% referral leaves zero automatic-channel misses. On E-DAIC, 21.4% referral reduces misses from five to two, and 53.6% referral reduces them to one. With only 17 positives, these reductions are not individually significant (exact McNemar $p = 0.25$ and $p = 0.125$). The curve is therefore exploratory, and its operating point requires external validation and a human-review budget. Misses fall as more cases are handed to review, which does not by itself establish a clinical benefit.

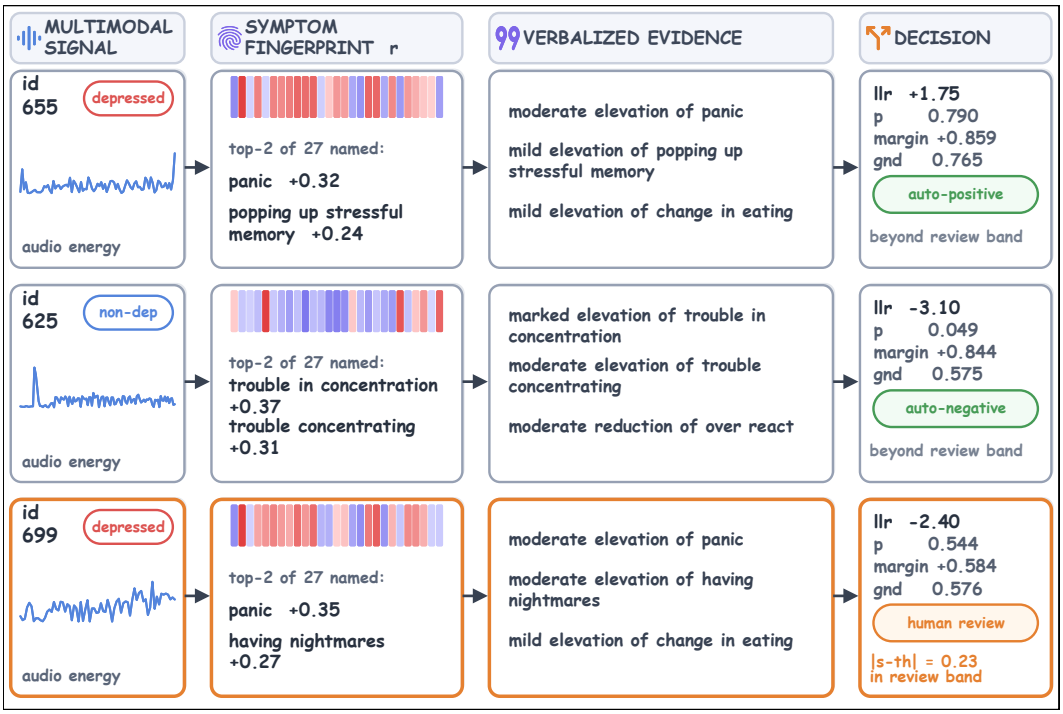


Figure 5. Case studies from the E-DAIC official test split. The first two cases are decided automatically, and the third is sent to human review.

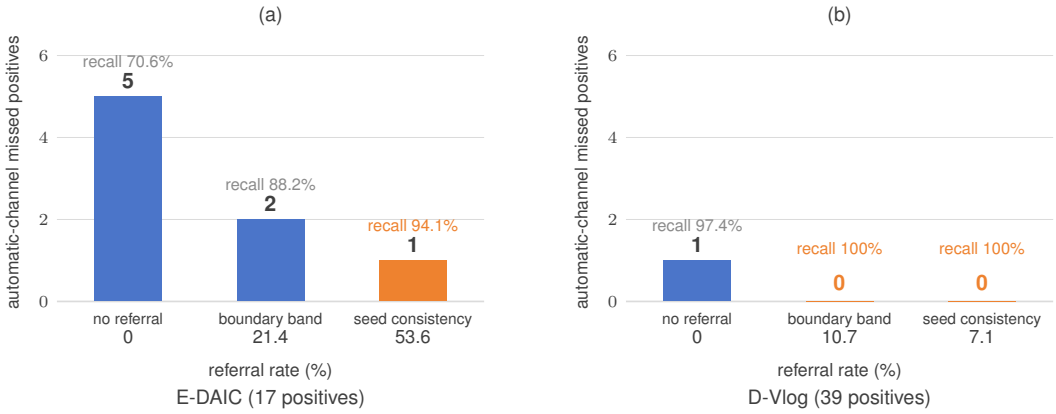


Figure 6. Automatic-channel misses versus referral rate. (a) On E-DAIC, misses fall from 5 to 1 as referral rises. (b) On D-Vlog, both referral settings leave zero misses.

5 Conclusion

In this work, we propose EARN, a screening framework that turns a shared multimodal representation into a 27-dimensional, symptom-named fingerprint and a referral-aware decision.By aligning that fingerprint with an ordered evidence library and training the evidence head with evidence-oriented GRPO, the reader receives protocolized phrases drawn from that fingerprint. Under the reported protocol, EARN obtains F1 0.6286 on the E-DAIC official split and 0.9527 on D-Vlog, and the verbalized reader reaches AUROC 0.7406 where the same fingerprint supplied as raw floats reaches 0.5611. Ablations support verbalization, the evidence library, and the decoupled referral rule, which keeps confident cases on the automatic channel and lowers automatic-channel misses by sending uncertain cases to human review.In conclusion, EARN gives the reader an auditable symptom-named record and a practical path for human review, with strong screening performance on the reported E-DAIC and D-Vlog protocols.

AI USE STATEMENT

Generative AI tools assisted prose restructuring and experiment-orchestration scripts. They did not generate the method or reported measurements. The authors reviewed the assisted text and code and take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The training and evaluation scripts are configuration driven. Reported hyperparameters, verbalization thresholds, and the evidence-library format are listed under implementation details. Further source data are recorded with the additional results. The public datasets are used under their original licences. Anonymous code and the evidence library are available at `https://anonymous.4open.science/r/EARN-5E5B`.

ETHICS STATEMENT

This work uses de-identified, publicly available research datasets and collects no new human-subject data. MODMA data are used with acknowledgment to the Gansu Provincial Key Laboratory of Wearable Computing, Lanzhou University, China, as the source of data. PDCH is distributed with its HAMD-17 assessments and is used here for screening research only. It is a screening research system, not a diagnostic device or deployment recommendation. Verbalized outputs include a screening-only disclaimer, and uncertain cases are referred rather than automatically released. We do not deploy the system on any population. Before release, authors must verify that every reproduced participant example is permitted by the relevant dataset terms and cannot facilitate re-identification.

REFERENCES

Anonymous. ECRL-MH: Evidence-conditioned reinforcement learning for explicit multi-disorder mental health screening, 2026. Under review, ICML 2026 (our baseline; anonymous submission, cited as such).

Hanshu Cai, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Huiyu Xiao, Jing Li, Zhenyu Yang, Xiaowei Li, Qinglin Zhao, Zhenyu Liu, Zhijun Yao, Mi Yang, Hong Peng, Jie Zhu, Xiaowei Zhang, Xiping Hu, and Bin Hu. MODMA dataset: a multi-modal open dataset for mental-disorder analysis, 2020. Data source: Gansu Provincial Key Laboratory of Wearable Computing, Lanzhou University, China.

Pengfei Cao, Yuanzhe Zhang, Chenxiang Zhang, et al. A multimodal depression consultation dataset of speech and text with HAMD-17 assessments. *Scientific Data*, 12:1577, 2025. doi: 10.1038/s41597-025-05817-9.

Cheng Dai, Jin Hu, Hao Shi, et al. Psyche-R1: Towards reliable psychological LLMs through unified empathy, expertise, and reasoning, 2025.

Kaituo Feng, Kaixiong Gong, Bohao Li, et al. Video-R1: Reinforcing video reasoning in MLLMs, 2025.

David Gimeno-Gómez, Ana-Maria Bucur, Adrian Cosma, Carlos-D. Martínez-Hinarejos, and Paolo Rosso. Reading between the frames: Multi-modal depression detection in videos from non-verbal cues. In *European Conference on Information Retrieval*, pp. 191–209. Springer, 2024.

Md. Rabiul Haque, Md. Mehedi Islam, S. M. Raju, Hamdi Altaheri, Lamiaa Nassar, and Fakhri Karray. MMFformer: Multimodal fusion transformer network for depression detection, 2025.

Lang He, Zhen Li, Prayag Tiwari, Chunpu Cao, Jing Xue, Fei Zhu, and Di Wu. DepressFormer: Leveraging video swin transformer and fine-grained local features for depression scale estimation. *Biomedical Signal Processing and Control*, 96:106490, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models, 2021.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664. PMLR, 2021.

Tin Lai, Yansen Shi, Zizhao Du, et al. Psy-LLM: Scaling up global mental health psychological services with AI-based large language models, 2023.

Melih C. Ozgun, Jiahuan Pei, Koen Hindriks, Lucia Donatelli, Qing Liu, and Jun Wang. Trustworthy AI psychotherapy: Multi-agent LLM workflow for counseling and explainable mental disorder diagnosis. In *Proceedings of the 34th ACM CIKM*, pp. 2263–2272, 2025.

Zizhuo Shangguan, Xinyu Li, Yuning Dong, and Xin Yuan. Automatic depression detection using attention-based deep multiple instance learning. In *International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, pp. 40–51. Springer, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024.

Siyang Song, Linlin Shen, and Michel Valstar. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 158–165. IEEE, 2018.

Yuwen Tao, Mingyue Yang, Hui Li, Yu Wu, and Bin Hu. DepMSTA: Multimodal spatio-temporal attentional transformer for depression detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):2956–2966, 2024a.

Yuwen Tao, Mingyue Yang, Yu Wu, Kyungeun Lee, Adrienne Kline, and Bin Hu. Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer. *Digital Communications and Networks*, 10(3):577–585, 2024b.

Ping-Cheng Wei, Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Multi-modal depression estimation based on sub-attentional fusion. In *European Conference on Computer Vision*, pp. 623–639. Springer, 2022.

James R. Williamson, Elizabeth Godoy, Michael Cha, et al. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 11–18, 2016.

Tao Xing, Yong Dou, Xin Xie, Jun Zhou, Xiao Chen, and Shuai Peng. Emo-mamba: Multimodal selective structured state space model for depression detection. In *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 2726–2731. IEEE, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, et al. Qwen2.5-omni technical report, 2025a.

Yiwen Xu, Jiahao Hu, Zhiyuan Zhao, Zheng Duan, Xiang Sun, and Xin Yang. MultiAgentESC: A LLM-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of EMNLP*, pp. 4665–4681, 2025b.

Mingyue Yang, Yi Liu, Yuwen Tao, and Bin Hu. Spike memory transformer: An energy-efficient model in distributed learning framework for autonomous depression detection. *IEEE Internet of Things Journal*, 2025a.

Qize Yang, Shimin Yao, Weiming Chen, et al. HumanOmniV2: From understanding to omni-modal reasoning with context, 2025b.

Junjian Ye, Jun Zhang, and Hongming Shan. DepMamba: Progressive fusion mamba for multimodal depression detection. In *IEEE ICASSP*, pp. 1–5. IEEE, 2025.

Wenbo Zhai, Nan Bai, Qing Zhao, et al. MentalGLM series: Explainable large language models for mental health analysis on Chinese social media. In *Proceedings of EMNLP*, pp. 13599–13614, 2025.

Lingyun Zhou, Zhi Liu, Xin Yuan, Zizhuo Shangguan, Yunpeng Li, and Bin Hu. CAIINet: Neural network based on contextual attention and information interaction mechanism for depression detection. *Digital Signal Processing*, 137:103986, 2023.

A IMPLEMENTATION DETAILS

Architecture. The language reader is Qwen2.5-Omni-7B (Xu et al., 2025a) in text-only mode. The evidence document is its sole input. The same checkpoint supplies the audio-tower features for the discriminative network. The discriminative network uses latent widths 128/256/384, a temporal hidden size of 64, two Transformer layers, and masked mean pooling. Per-source PCA projectors are fitted on the source domain and restored at inference. Each condition/modality pair has separate support and opposition heads. The shift $\mathrm{shift}_{k,m}$ is initialized to zero.

Training. The discriminative stage uses 20 BCE warm-up epochs and 70 evidence-oriented GRPO epochs. The learning rate is 2×10^{-4} for E-DAIC and 3×10^{-3} for smaller domains. Reward weights are $w_{\mathrm{gr}} = 0.3$, $w_{\mathrm{at}} = 0.2$, $w_{\mathrm{co}} = 0.2$, and $w_{\mathrm{pa}} = 0.1$, plus $\lambda_d R_{\mathrm{margin}}$ (Section 3.5). The BCE anchor has weight 10 and $w_{\mathrm{lab}} = 0$. We use $\lambda_{\mathrm{proto}} = 0.1$ and $\tau = 0.1$. The reader uses LoRA with $r = 8$ and $\alpha = 16$ on q/k/v/o projections, GRPO with $G = 4$, max-new-tokens 24, learning rate 2×10^{-5}, and three epochs. Smaller domains are warm-started from the E-DAIC reader. Verbalization thresholds are $0.15/0.25/0.35$.

Reward stages. Grounding, attribution, consistency, pair, and margin rewards are computed from discriminative-network outputs and are applied during the 70 evidence-stage epochs. The language channel does not modify the fingerprint. The language reader is adapted separately. Its sampled candidate receives 1.0 when the emitted JSON matches the label, plus 0.05 when the output is parseable, before group-normalized advantages are computed.

Margin-weight control. Because the margin is conditioned on sampled candidates and shares parameters with the discriminative objective, it is constrained by a development-set controller. The controller bounds λ_d in $[0.02, 0.08]$, caps each update at 5×10^{-4}, shrinks it toward 0.05, and does not increase it when its label correlation is absent. The reported values are 0.066 for depression and 0.038 for PTSD.

Evidence library. The JSON library covers 12 conditions and 27 symptom keys. Each entry contains $p(\text{symptom} \mid \text{condition})$, a specificity weight, and a prevalence delta. The fixed order of the 27 keys defines the coordinate order of r. The prototype matrix is the row-normalized condition profile. Positive pools include clinical subtypes of the target condition.

Evaluation protocol. Metrics are computed on each test split. The reader score is the log-likelihood ratio at the label position. The same reported threshold procedure is used for the main results and the referral curve. MODMA and PDCH use five-fold cross-validation with an independent development split within each training fold and multi-seed soft voting. Their test folds contain 10–13 participants, so a single participant changes AUROC by 0.024–0.042. Missed positives are true positives called negative within the automatic channel. Referred participants are excluded by construction.

B ADDITIONAL RESULTS AND SOURCE DATA

Table 3. Evidence-stage training ablation on E-DAIC. The table compares the full warm-up plus evidence-GRPO schedule with the warm-up-only control, which also uses fewer training epochs.

Configuration	Disc. head	Δ	LLM AUROC	Δ
Full (warm-up + evidence GRPO)	**0.6968**	n/a	**0.7406**	n/a
w/o GRPO (warm-up only)	0.5264	**−0.1704**	0.6154	**−0.1252**

Table 4. Evidence-library ablation on E-DAIC. FULL uses the ordered library, NOPROTO removes it, and SHUFFLE scrambles symptom values.

latent	Variant	Disc. head	LLM AUROC	ΔLLM vs. NOPROTO
128	FULL	0.686	**0.662**	**+0.052**
128	NOPROTO	0.699	0.610	n/a
256	FULL	0.683	**0.744**	**+0.101**
256	SHUFFLE	0.694	0.721	+0.078
384	FULL	0.668	**0.703**	**+0.042**
384	SHUFFLE	0.640	0.644	−0.017

Table 5. Modality ablation on E-DAIC. Results use latent width 384 and average two seeds.

Modalities	Disc. head	LLM readout	F1	P	R	ΔLLM vs. audio
audio	0.7421	0.6546	0.5490	0.412	0.824	n/a
visual	0.5068	0.5701	0.4828	0.341	0.824	−0.0845
text	0.5445	0.4344	0.4658	0.304	1.000	−0.2202
audio+visual	0.6968	**0.741**	**0.6286**	**0.611**	**0.647**	**+0.0864**
audio+visual+text	**0.6983**	0.6998	0.6222	0.500	0.824	+0.0452

Table 6. Component-scope analysis on E-DAIC. The table reports the component effect across three representation regimes.

Representation	w/o component	**with component**	**Gain**
mel/OpenSMILE (hidden 64)	0.619	**0.752**	**+0.133**
Qwen-3B @ latent 32	0.742	**0.816**	**+0.074**
Qwen-3B @ latent 128	0.930	0.919	≈ 0 (saturated)

Table 7. Per-dataset results on MODMA (Cai et al., 2020) and PDCH (Cao et al., 2025). Results use five folds, an independent development split, five seeds, and soft voting. Evaluation is within-dataset rather than cross-domain.

Dataset	Participants	Disc. head	Single-seed mean	**Soft vote**	Precision	Recall / **F1**
MODMA	55	0.9230	0.8645	**0.8866**	0.7667	**1.0000 / 0.8679**
PDCH	62	0.7539	0.7244	**0.7468**	0.6596	0.9118 / **0.7654**

Soft voting increases the reported single-seed mean by +0.0221 on MODMA and +0.0224 on PDCH. Because the datasets are small, these values should be interpreted with seed-level uncertainty rather than as stable deployment estimates.

Table 8. Evidence-representation ablation on E-DAIC. Qwen-derived features are compared with OpenSMILE eGeMAPS plus OpenFace descriptors.

Evidence representation	Disc. head	LLM AUROC	F1	Precision	Recall
Qwen-derived features	**0.697**	**0.741**	**0.564**	**0.500**	**0.647**
Handcrafted (eGeMAPS + OpenFace)	0.661	0.656	0.487	0.450	0.529
Δ	+0.036	+0.085	+0.077	+0.050	+0.118

Qwen-derived features are higher on the reported metrics in this configuration. The comparison does not identify whether the difference arises from feature pretraining, representation capacity, or alignment with the verbalized interface.

Table 9. Evidence-library ablation on the E-DAIC self-split. The representation is fixed while FULL and NOPROTO are compared at the reader level.

Variant	Fingerprint separability	LLM F1	Precision	Recall	ΔF1
FULL (with library)	0.974	**0.842**	0.800	0.929	**+0.175**
NOPROTO (no library)	0.963	0.667	0.583	0.833	n/a