INERTIAMPDD: MULTIMODAL PERSONALIZED DEPRESSION DETECTION VIA INDIVIDUAL EMOTIONAL-INERTIA MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing omni-modal depression detection largely infers depression representations shared across individuals, rather than latent variables of individual difference inferred from each person's behavioral signals. Yet depressive expression superimposes transient states on stable individual traits: unless traits are modeled and conditioned on, they confound the assessment of states, while conditioning on annotated individual attributes does not scale because such annotations are scarce. We propose InertiaMPDD, which reframes detection as estimating each person's deviation of current expression from a counterfactual healthy baseline, turning individual differences from externally supplied labels into latent variables inferred from behavioral signals. The baseline is an individual-conditioned latent dynamics predictor trained exclusively on healthy dynamics, so prediction error measures departure from the learned healthy reference. Rather than regenerating a "healthy" counterpart, we realize the counterfactual by single-step rollout: under the person's own trait conditioning, the model advances the observed state one step according to healthy dynamics, and the deviation supplies an individual-level screening feature. Prediction and observation share the same trajectory, context and recording conditions, while inferred subject characteristics condition the reference. Three complementary, annotation-free quantities follow: the residual R (position), the AR(1) restoring force $\mathrm{force}_{\mathrm{AR}}$ (latent-trajectory decay), and the spectral radius $\rho(J)$ of the predictor's Jacobian at the individual operating point (local predictive sensitivity). On the in-the-wild vlog corpora D-Vlog and LMVD, InertiaMPDD attains F1 of 98.41 and 84.98, and the joint system surpasses either pathway alone. The residual is significant with a consistent direction on both corpora.

1 INTRODUCTION

Depression is among the leading causes of disability worldwide, and its screening still relies on self-report scales and clinical interviews, which are limited by subjectivity and by the scarcity of trained professionals (World Health Organization, 2023). Detection built on large language models and omni-modal large models has substantially improved the joint understanding of language, speech and facial behavior (Sadeghi et al., 2024; Zheng et al., 2025; Zhu et al., 2026), and multimodal screening has been reviewed systematically (Wang et al., 2025b). Mainstream methods, however, cast the task as inferring a depression representation shared across individuals (Figure 1a): they learn from annotated samples of many people what depression "looks like" at the population level and score a new sample against it. This does not match the structure of depressive expression, in which a transient state is superimposed on a stable individual trait. The same slow speech, flattened prosody or reduced facial activity may be the habitual baseline of a reserved speaker yet a salient state signal in an expressive one. Matching a person's static features against a population template compresses trait and state into a single comparison, so the trait enters the decision as a confound instead of being modeled. Two limitations follow.

Limitation 1: individual differences are treated as explicit inputs, leaving the task an interpersonal classification. Work that recognizes individual differences has largely introduced them as additional inputs (Figure 1a, bottom). One line customizes inference per person through re-

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.

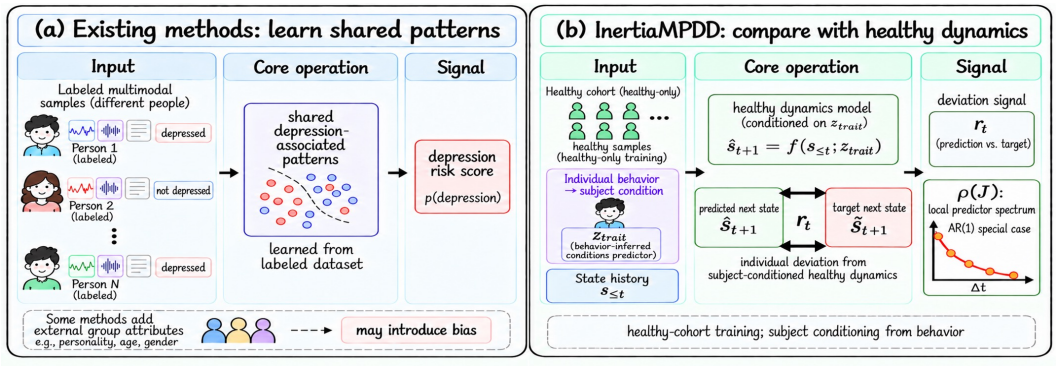


Figure 1: **Two modeling orientations. (a)** Existing methods learn a population-level depression representation, optionally taking external attributes z_i as extra input. **(b)** InertiaMPDD compares against healthy dynamics: a predictor trained on a healthy cohort and conditioned on an inferred trait produces $\hat{s}_{t+1} = f(s_{\leq t}; z_{\text{trait}})$, whose discrepancy with the target next state $\tilde{s}_{t+1}$ gives r_t. The comparison stays within the subject's own trajectory.

trieval augmentation, generating personalized queries from retrieved individual-related social knowledge (Zhang et al., 2025); another performs personality-guided representation learning, turning discrete individual features into contextual descriptions for personalized encoding (Fu et al., 2025b); personality-enhanced detection for the elderly and the MPDD challenge feed personality directly as input or supervision (Fu et al., 2025a; Wang et al., 2025a). In all of these, individual difference is an auxiliary feature z_i to be annotated or retrieved, while the discriminative objective remains classification across people. Such conditioning does not explicitly compare the observed trajectory with a subject-conditioned healthy reference, leaving trait–state confounding a concern.

Limitation 2: attribute-annotation conditioning does not scale and carries fairness risk. Conditioning on additional per-person attributes, such as age, personality, developmental background or social environment, is constrained by the general absence of such annotations in existing datasets. Writing demographic priors into the model may moreover reintroduce the population-level bias that recent counterfactual debiasing work seeks to remove (Hu et al., 2026). What is missing is not more individual annotation but a mechanism that absorbs individual differences during modeling without having them supplied externally.

Our reformulation. InertiaMPDD rewrites detection from matching a population template to comparing against a law of healthy dynamics (Figure 1b). Instead of learning what depression looks like, we learn how healthy emotion evolves over time and measure, on each subject's own trajectory, how far the observed dynamics deviate from what that law predicts. An emotional-dynamics predictor is trained on a healthy cohort and never sees depressed dynamics, with prediction error used to quantify departure from the learned reference; it is conditioned on a trait latent z_{trait} inferred from the subject's behavioral signals rather than annotated. The counterfactual thus concerns the *evolution law*, namely what the next instant should look like if this stretch of dynamics followed healthy dynamics, rather than a hypothetical "healthy version" of the person.

The comparison takes place *between adjacent instants of the subject's own trajectory*, not between this person and others. Prediction and observation share the same trajectory, context and recording conditions, and individual characteristics condition the reference rather than being supplied as attribute labels. We investigate whether subject-conditioned predictive deviations from a healthy reference model provide complementary information to multimodal discriminative depression screening. The full system combines a *Nomothetic Screening Pathway* (NSP), which learns cross-individual regularities and emits a depression tendency score, with an *Idiographic Inertia Pathway* (IIP), which supplies this deviation, the two being summed at the decision level. The former answers whether a sample looks like depression, the latter how far its dynamics depart from the healthy evolution law.

Complementary individual-level quantities. IIP yields three annotation-free quantities: the residual R, the departure of the observed dynamics from the healthy law (position); the AR(1)

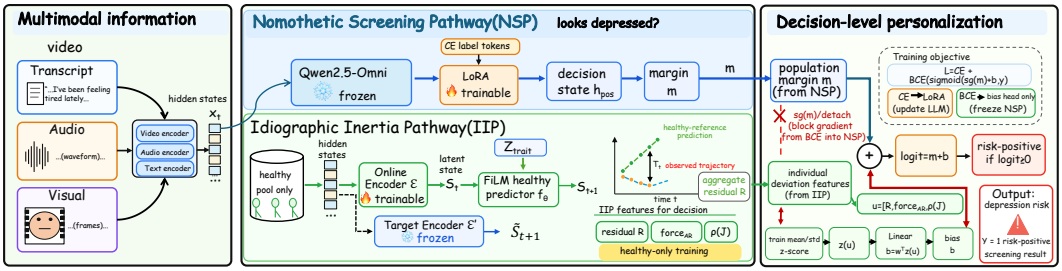


Figure 2: **Overall framework. Left:** the omni-modal reader encodes transcript, audio and frames into x_t. **Top (NSP):** a frozen backbone with LoRA, supervised by cross-entropy on label tokens, gives the margin m at the decision state h_{pos}. **Bottom (IIP):** E projects x_t into s_t, and a healthy predictor FiLM-modulated by z_{trait} gives $\hat{s}_{t+1}$ against the frozen target $\tilde{s}_{t+1}$, trained on the healthy pool only; the deviation yields R, force_{AR} and $\rho(J)$. **Right:** these are standardized and mapped to a bias b, giving $\text{logit} = m + b$; CE updates only LoRA, BCE only the bias head, and $\text{sg}(m)$ blocks BCE from reaching NSP.

restoring force force_{AR}, a latent-trajectory decay rate motivated by emotional inertia in affective science (Kuppens et al., 2010; Koval et al., 2013); and the spectral radius $\rho(J)$ of the predictor's Jacobian at the subject's operating point, a local spectral descriptor of the healthy predictor in the region the subject occupies (van de Leemput et al., 2014). Position indicates departure and restoring force summarizes latent temporal persistence; together they give the method its name. All three are computed on the subject's own trajectory under the inferred trait, so individual difference is borne by the quantities themselves.

We validate InertiaMPDD on two in-the-wild vlog corpora, the English D-Vlog (Yoon et al., 2022) and the Chinese LMVD (He et al., 2026). Our contributions are as follows. **Formulation.** We recast depression detection from inter-personal classification to within-person counterfactual deviation estimation, turning individual differences from external labels into latent variables inferred from behavioral signals and thereby obtaining personalization without attribute annotation. **Method.** We realize the counterfactual by single-step rollout and derive three complementary individual-level quantities, the position R and the restoring-force pair force_{AR} and $\rho(J)$, connecting annotation-free dynamical features to an emotional-inertia modeling hypothesis. **Results.** InertiaMPDD attains F1 of 98.41 on D-Vlog and 84.98 on LMVD, and the joint system outperforms either pathway alone; the residual is significant with a consistent direction on both corpora.

2 Method

2.1 Overview and the two pathways

Notation is summarized in Section B. NSP emits a cross-sample depression tendency score m by learning population-level multimodal regularities (nomothetic). IIP compares adjacent instants of the subject's own trajectory (idiographic), measures the deviation of the observed dynamics from the healthy evolution law and converts it into a decision-level bias b. The decision is $\text{logit} = m + b$, with $\text{logit} \geq 0$ declared risk-positive. The naming reflects the central distinction: NSP generalizes across people, IIP compares a person with the healthy evolution law along that person's own trajectory, and the two meet only at the decision. Figure 2 gives the overall framework and Figure E.1 in Section E details the two core modules.

2.2 Paralinguistic latent states

A frozen omni-modal LLM is run over each sample; its hidden states are pooled in time to 1 Hz and concatenated across the audio and visual channels to give x_t, which the online encoder projects into $s_t = E(x_t) \in \mathbb{R}^{256}$, so that each video forms a trajectory in a 256-dimensional space. The trait latent $z_{\text{trait}} \in \mathbb{R}^{16}$ is pooled from the first- and second-order statistics of the subject's paralinguistic hidden states and injected into the predictor by per-channel FiLM gain and bias with identity initialization, so that the predictor is conditioned on subject characteristics in the recording. It is read from the

frozen backbone and passes through no label-supervised path, consistent with a predictor that has only seen healthy samples. For a predictor activation h, FiLM applies $\gamma(z) \odot h + \beta(z)$, with initial gain one and bias zero. The same recording-level z conditions every prediction in a video, while $s_{\leq t}$ supplies its changing state history. Thus, conditioning adapts a shared predictor without fitting a separate model per subject.

2.3 THE HEALTHY-BASELINE EMOTIONAL-DYNAMICS PREDICTOR

Inspired by world models that predict the next state from the current one in a latent space (LeCun, 2022; Assran et al., 2023), we train on the healthy pool $\mathcal{H}$ a trait-conditioned emotional-dynamics predictor $f_\theta(\cdot; z)$ that characterizes how healthy emotion evolves (Figure E.1b). It is learned by a self-supervised single-step rollout objective,

$$\min_{\theta, E} \ \mathbb{E}_{\mathcal{H}} \big\| f_\theta(s_{\leq t}; z) - \mathrm{sg}(\tilde{s}_{t+1}) \big\|_2^2 \ + \ \lambda \, \mathcal{R}_{\mathrm{VICReg}}(E), \tag{1}$$

where sg denotes stop-gradient, the frozen semi-orthogonal target encoder E' provides the reference $\tilde{s}_{t+1} = E'(x_{t+1})$ for the true next state, the variance term $\sum_j \max(0, 1 - \mathrm{std}(E_j))$ of $\mathcal{R}_{\mathrm{VICReg}}$ prevents collapse and its covariance term prevents dimensional collinearity (Bardes et al., 2022). Modeling the target state as $\tilde{s}_{t+1} = f^\star(s_{\leq t}; z) + \varepsilon_t$ with conditionally zero-mean noise of covariance $\Sigma_\varepsilon(s_{\leq t}, z)$, and since the training distribution is healthy only, f_θ approximates the healthy conditional expectation $f^\star_{\mathcal{H}}(s_{\leq t}; z) = \mathbb{E}[\tilde{s}_{t+1} \mid s_{\leq t}, z, \mathrm{healthy}]$, and $\hat{s}_{t+1} = f_\theta(s_{\leq t}; z)$ is the healthy-dynamics prediction at that instant. The comparison therefore measures prediction error relative to a subject-conditioned healthy reference; its sensitivity to depression is evaluated empirically.

2.4 POSITION: THE COUNTERFACTUAL RESIDUAL

The per-step residual and its temporal mean (Figure E.1b) are

$$r_t = \|\tilde{s}_{t+1} - \hat{s}_{t+1}\|_2, \qquad R = \frac{1}{T-1} \sum_{t=1}^{T-1} r_t, \tag{2}$$

so that R summarizes the average departure of the whole trajectory from the healthy evolution law. Averaging over valid transitions prevents a longer video from receiving a larger score merely because it contains more prediction steps; it does not remove differences in content or estimation uncertainty associated with recording length.

Proposition 1 (Bias–noise decomposition of the residual)**.** *Let $\tilde{s}_{t+1} = g(s_{\leq t}; z) + \varepsilon_t$, with $\mathbb{E}[\varepsilon_t \mid s_{\leq t}, z] = 0$ and conditional covariance $\Sigma_\varepsilon(s_{\leq t}, z)$. Then*

$$\mathbb{E}[r_t^2 \mid s_{\leq t}, z] = \underbrace{\|g(s_{\leq t}; z) - f_\theta(s_{\leq t}; z)\|_2^2}_{\textit{departure term } b_t^2} + \mathrm{tr}(\Sigma_\varepsilon(s_{\leq t}, z)). \tag{3}$$

When $g \approx f_\theta$, the residual approaches the conditional noise floor; a larger departure term increases the expected squared residual at a fixed noise level.

The proof is given in Section B. The decomposition separates predictive mismatch from conditional noise; group differences in R can reflect both terms and are assessed in Section 3.3.

2.5 RESTORING FORCE: EMOTIONAL INERTIA

At the individual operating point $\bar{s} = \mathbb{E}[s_t]$, the Jacobian of the predictor and its spectral radius are

$$J = \left. \frac{\partial f_\theta(\cdot; z)}{\partial s_t} \right|_{s_t = \bar{s}}, \qquad \rho(J) = \max_i |\lambda_i(J)|, \tag{4}$$

and $\rho(J)$ describes the local spectrum of the healthy predictor, with history and z held fixed (Figure E.1b). Fitting a first-order autoregression $s_{t+1} - \bar{s} = \varphi(s_t - \bar{s}) + \eta_t$ to the subject's latent trajectory further gives the restoring force $\mathrm{force}_{\mathrm{AR}} = -\ln \varphi$ at the 1 Hz sampling rate. For $0 < \varphi < 1$, a slower return yields a smaller $\mathrm{force}_{\mathrm{AR}}$. $\rho(J)$ is read from the predictor and $\mathrm{force}_{\mathrm{AR}}$ from the observed trajectory; their dynamical interpretations coincide only under additional assumptions.

Two properties describe the connection to scalar dynamical models; proofs are given in Section B.

Proposition 2 (AR(1) special case)**.** *When $d = 1$ and the predictor is a transition in a common state coordinate, $f_\theta(s) = c + \varphi(s - c)$, we have $J = \varphi$ and $\rho(J) = |\varphi|$. For a stationary AR(1) process with $0 \le \varphi < 1$, this equals its lag-one autocorrelation, the emotional-inertia measure of Kuppens et al. (2010).*

Proposition 3 (Relaxation time and critical slowing down)**.** *For a first-order transition in common state coordinates, let $\bar{s}$ be a fixed point and $\delta_t = s_t - \bar{s}$. To first order $\delta_{t+k} \approx J^k \delta_t$. For diagonalizable J with $0 < \rho(J) < 1$, its slowest linearized decay has characteristic time $\tau = -1/\ln \rho(J)$, which diverges as $\rho(J) \to 1^-$.*

Propositions 2 and 3 motivate the relation to inertia and critical slowing down (van de Leemput et al., 2014). In our history-conditioned predictor, J maps online-state changes to target-state predictions at the sample operating point, so we use $\rho(J)$ as a local model descriptor. The additional assumptions for a recovery-time interpretation are discussed in Section B; cross-session stability remains a testable hypothesis.

2.6 The screening pathway and the margin

NSP adapts the frozen omni-modal backbone with LoRA (Hu et al., 2022) and casts the risk judgment as the generation of a label token (Figure E.1a). Let h_{pos} be the decision state, the hidden state one step before the label is written, and let $W \in \mathbb{R}^{d \times 2}$ give the scores of the two risk tokens. The label $y \in \{0, 1\}$ is supervised by cross-entropy,

$$\mathcal{L}_{\mathrm{CE}} = -\log p(\langle \mathrm{risk}_y \rangle \mid h_{\mathrm{pos}}), \tag{5}$$

the adapter's only label-supervised path. The margin $m = h_{\mathrm{pos}}^\top (W_1 - W_0)$ increases with risk; its readout is shared by all individuals, so m captures the cross-individual judgment of whether a sample looks depressed. The exact decision position is specified in Section C.

2.7 Decision-level personalization and joint fusion

Decision-level personalization merges the three IIP features with the NSP margin (Figure 2, right). Writing $u = [R,\ \mathrm{force}_{\mathrm{AR}},\ \rho(J)]$, we standardize it with training-set statistics into $z(u)$, map it linearly to the bias $b = w^\top z(u) + c$, and form $\mathrm{logit} = m + b$. The joint loss is

$$\mathcal{L} = \mathrm{CE}(\mathrm{labels}) + \mathrm{BCE}\big(\sigma(\mathrm{sg}(m) + b),\ y\big). \tag{6}$$

The stop-gradient on m decouples the two objectives. The decision can equivalently be written as $m \ge -b$: positive b lowers the required screening margin, whereas negative b raises it. The terms $w_j z(u)_j$ expose each feature's signed contribution to this correction. IIP alone instead predicts from u without m, using its own supervised head; it is the three-feature row of Table 3.

Proposition 4 (Role of the stop-gradient on the margin)**.** *Without a stop-gradient on m, the BCE term contributes an additional gradient to the LoRA adapter; with $\mathrm{sg}(m)$ the adapter gradient equals that of a CE-only recipe.*

The gradient of BCE with respect to m is $\sigma(m + b) - y$; without sg it flows into the adapter and changes the joint update. The stop-gradient assigns label-token learning to the adapter and margin correction to the bias head. With $\mathrm{sg}(m)$ the adapter is supervised by CE alone and the bias head corrects only at the decision level, so personalization is added without altering the screening pathway (proof in Section B). Algorithm 1 in Section F summarizes the full procedure.

3 Experiments

3.1 Experimental setup

Datasets. D-Vlog (Yoon et al., 2022) consists of English YouTube vlogs and LMVD (He et al., 2026) of Chinese vlogs, both with binary depression labels. D-Vlog is evaluated on its official train/validation/test split and LMVD on a 7:1:2 train/validation/test split. The healthy-dynamics predictor of IIP is trained only on healthy samples of the training split, with no uploader shared between training and evaluation.

Table 1: Results on D-Vlog and LMVD (%). Depression is the positive class for our models. Published values are taken from their original papers; "–" indicates not reported. Our LMVD results are means over five random seeds.

		D-Vlog				LMVD			
Method	Year	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Depression Detector (Yoon et al., 2022)	2022	–	65.40	65.57	63.50	–	–	–	–
TAMFN (Zhou et al., 2022)	2022	–	66.02	66.50	65.82	–	–	–	–
Attention-MIL (Shangguan et al., 2023)	2023	–	67.27	67.77	66.64	–	–	–	–
CAIINET (Zhou et al., 2023)	2023	–	66.56	66.98	66.55	–	–	–	–
STST (Tao et al., 2024b)	2024	70.70	72.50	77.67	75.00	–	–	–	–
DepMSTAT (Tao et al., 2024a)	2024	–	71.53	75.60	73.51	–	–	–	–
MDAVIF (Ling et al., 2024)	2024	–	74.25	76.00	75.25	–	–	–	–
Depressformer (He et al., 2024)	2024	65.00	64.00	54.00	59.00	–	–	–	–
DepMamba (Ye et al., 2025)	2025	68.87	68.19	86.99	76.44	72.13	70.18	76.56	73.20
CMF-Mamba (Zhou et al., 2025a)	2025	–	–	–	–	74.86	73.20	78.02	75.53
STCM-Mamba (Zhou et al., 2025b)	2025	72.64	73.72	82.11	77.69	75.41	72.55	76.68	76.49
Xception (He et al., 2026)	2026	–	–	–	–	71.38	71.94	71.38	71.19
ViT (He et al., 2026)	2026	–	–	–	–	73.03	73.52	73.03	72.90
MSI-Net (Shi et al., 2026)	2026	70.25	78.91	69.63	73.91	74.34	78.80	75.29	77.06
MDDFormer (He et al., 2026)	2026	–	–	–	–	76.88	77.02	76.88	76.85
BHDD (Ma et al., 2026)	2026	–	75.98	75.96	75.94	78.11	78.39	78.11	78.06
DepressionAgent (Zhu et al., 2026)	2026	95.26	95.24	96.15	95.69	82.57	80.48	85.76	83.04
NSP (alone)	–	97.14	94.90	**100.00**	97.38	74.30	76.55	72.81	74.57
IIP (alone)	–	63.91	63.64	76.92	69.65	73.70	72.40	79.78	75.88
InertiaMPDD	–	**98.29**	**96.88**	**100.00**	**98.41**	**84.00**	**82.72**	**87.44**	**84.98**

Metrics, baselines and implementation. For our models, we report accuracy, precision, recall and F1 with depression as the positive class; LMVD results are means over five random seeds. Group differences of the individual quantities and the significance of detection performance against random labels are both assessed by permutation tests (Section D). We compare against methods reported in recent literature on both corpora, including Mamba-based sequence models, transformer-based multimodal fusion and the LLM-based agent DepressionAgent, taking values from the respective original papers in the comparison format of Zhu et al. (2026). NSP trains a LoRA adapter on a frozen Qwen2.5-Omni backbone (Qwen Team, 2025) with cross-entropy as its only label-supervised path; the IIP quantities are computed as defined in Section 2, and the decision head maps the standardized features to a bias. IIP alone uses the three features $[R, \mathrm{force}_{\mathrm{AR}}, \rho(J)]$; duration is an auxiliary control. Hyperparameters, input limits and the extraction of the decision state are given in Section C.

3.2 Main results

Table 1 summarizes the comparison. InertiaMPDD attains accuracy 98.29 and F1 98.41 on D-Vlog and accuracy 84.00 and F1 84.98 on LMVD, with the reported DepressionAgent results at (95.26 / 95.69 and 82.57 / 83.04). The margin over the non-LLM baselines is larger: on LMVD they reach F1 between 71.19 and 78.06, while the three-feature IIP alone reaches 75.88. Its predictor is trained only on healthy samples, and its decision head is label-supervised.

The two pathways complement each other differently on the two corpora. On D-Vlog, NSP alone already reaches F1 97.38 with a recall of 100.00, so the gain from combination comes entirely from precision ($94.90 \rightarrow 96.88$), corresponding to the correction of two false positives while recall is preserved. IIP alone reaches F1 69.65 (permutation $p = 0.0005$), below NSP, while its contribution to the joint system is chiefly to correct healthy subjects that NSP flags as depressed. On LMVD the pathways are closer: IIP alone (F1 75.88) is 1.31 points above NSP alone (74.57), and the combination raises F1 to 84.98, gains of 9.10 and 10.41 points over IIP and NSP, respectively. Relative to IIP, both precision ($72.40 \rightarrow 82.72$) and recall ($79.78 \rightarrow 87.44$) improve. On both corpora the joint system exceeds either pathway, suggesting complementary information in the population-level judgment and the individual dynamical features.

Table 2: Group differences of the individual-level quantities (group means, healthy vs. depressed, with permutation-test p values; settings in Section D).

		D-Vlog			LMVD		
Quantity	Split	Healthy	Depressed	p	Healthy	Depressed	p
R	test	6.707	6.901	**0.0003**	12.510	12.887	**0.0040**
	train	6.820	6.916	0.078	12.487	12.966	<**0.0001**
$\mathrm{force}_{\mathrm{AR}}$	test	2.103	1.964	0.998	1.559	1.409	**0.0002**
	train	2.090	1.998	0.897	1.547	1.435	**0.0010**
$\rho(J)$	test	0.716	0.710	0.467	0.938	0.731	<**0.0001**
	train	0.718	0.709	0.421	0.898	0.750	<**0.0001**

Table 3: Ablation of the IIP features; p is a permutation test against random labels. LMVD values are means over five random seeds. Duration is the audio length in seconds, used as a confound control.

	D-Vlog					LMVD			
Feature set	Acc.	Prec.	Rec.	F1	p	Acc.	Rec.	F1	p
R	60.36	68.75	48.35	56.77	**0.0015**	57.80	60.39	59.69	0.507
$\mathrm{force}_{\mathrm{AR}}$	54.44	61.29	41.76	49.67	0.092	54.40	60.52	57.93	0.023
$\rho(J)$	46.75	50.55	50.55	50.55	0.877	68.30	69.74	69.47	<**0.0005**
$R + \mathrm{force}_{\mathrm{AR}} + \rho(J)$	63.91	63.64	76.92	69.65	**0.0005**	73.70	79.78	75.88	–
$R + \mathrm{force}_{\mathrm{AR}} + \rho(J)$+duration	62.72	60.45	89.01	72.00	0.001	74.70	74.04	75.18	<0.0005
Duration	56.21	55.86	89.01	68.64	0.148	53.50	49.38	52.32	0.559

3.3 Group differences of the individual-level quantities

The residual is consistent across corpora. On the test splits the residual is significantly higher in the depressed group on both corpora ($p = 0.0003$ on D-Vlog, $p = 0.004$ on LMVD), and the LMVD training split agrees ($p < 0.0001$). This is consistent with predictive mismatch in Proposition 1: the residual separates the groups in the same direction across the two corpora, despite their different languages and platforms.

$\mathrm{force}_{\mathrm{AR}}$ **is significant on LMVD.** Healthy subjects have the higher group mean on both corpora, corresponding to a lower fitted latent decay rate in the depressed group, in the direction motivating the emotional-inertia hypothesis (Kuppens et al., 2010; Koval et al., 2013). The difference is significant on both LMVD splits ($p = 0.0002$ and $p = 0.001$) and not on D-Vlog.

$\rho(J)$ **is corpus-dependent.** On LMVD, $\rho(J)$ is higher in the healthy group (0.938 vs. 0.731, $p < 0.0001$ on both splits); on D-Vlog the groups do not differ (test $p = 0.467$). Section 4.1 interprets the three quantities jointly.

3.4 Feature ablation and the duration confound

Table 3 compares the individual IIP features and their combinations. The ranking of single features is reversed between the corpora. On D-Vlog only the residual is significant on its own ($p = 0.0015$), while $\mathrm{force}_{\mathrm{AR}}$ and $\rho(J)$ are not; $R + \mathrm{force}_{\mathrm{AR}} + \rho(J)$ reaches accuracy 63.91 ($p = 0.0005$) and F1 69.65, so the individual quantities carry discriminative information that is real but limited on this corpus, consistent with the role of IIP there as a corrector of NSP. On LMVD, $\rho(J)$ is the strongest single feature ($p < 0.0005$), $\mathrm{force}_{\mathrm{AR}}$ is marginally significant ($p = 0.023$), the residual alone is not ($p = 0.507$), and $R + \mathrm{force}_{\mathrm{AR}} + \rho(J)$ reaches F1 75.88. Although the strongest single feature differs, the combination outperforms the best single feature by 12.88 F1 points on D-Vlog and 6.41 on LMVD, supporting their use as complementary predictive features.

Duration as a confound. Adding audio duration on D-Vlog raises recall from 76.92 to 89.01 and F1 to 72.00 ($p = 0.001$), but duration alone also reaches a recall of 89.01 (F1 68.64, $p = 0.148$),

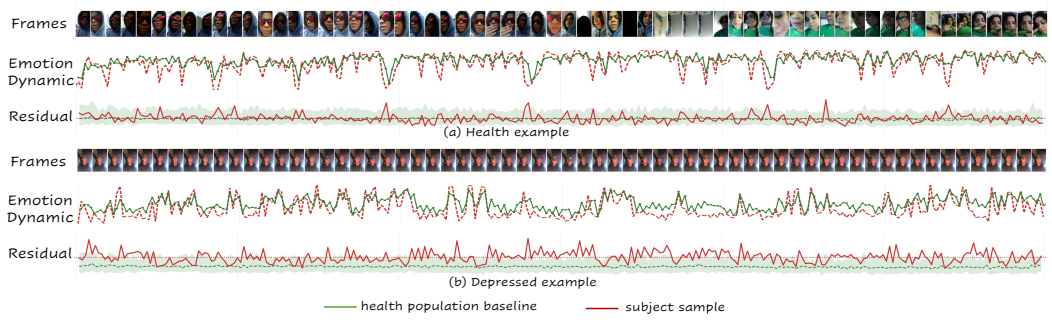


Figure 3: **Case visualization. (a) Healthy; (b) depressed subject.** Rows show video frames, latent dynamics and the per-step residual. Green is the healthy-dynamics rollout, red the observed trajectory; the shaded band is the 5–95% range of healthy residuals.

suggesting sensitivity to sample length. The increased recall is accompanied by lower precision ($63.64 \rightarrow 60.45$) and accuracy ($63.91 \rightarrow 62.72$), indicating a changed operating point. The duration-only permutation result does not establish a group association, so this control alone does not identify the source of the gain. On LMVD, duration alone is not significant in the permutation test (F1 52.32, $p = 0.559$) and adding it lowers F1 from 75.88 to 75.18. Duration is therefore excluded from the main results.

3.5 Case visualization

Figure 3 shows IIP on one healthy and one depressed video. Each case displays video frames, the latent dynamics, in which the healthy-dynamics rollout is compared with the observed trajectory, and the per-step residual r_t against the 5–95% range of healthy residuals. For the healthy subject (Figure 3a), scene, posture and expression vary considerably and the dynamics fluctuate markedly, including several deep dips, yet the observed trajectory stays close to the healthy prediction and returns close to the prediction after each dip, so the residual remains within the healthy 5–95% range almost throughout. For the depressed subject (Figure 3b), the frames barely change and expression appears flattened; the trajectory departs from the healthy prediction frequently and stays below it for extended stretches, and the residual repeatedly exceeds the 95th percentile with dense spikes spread across the whole video rather than concentrated at a few instants. This example shows sustained predictive disagreement rather than an isolated residual spike. The contrast illustrates the residual pattern in Section 3.3; the vertical latent coordinate itself does not measure mood severity. Because r_t localizes departures in time, a detection can also be traced to specific segments of the video.

4 Analysis and Discussion

4.1 The three quantities seen as emotion dynamics

The residual is the only individual quantity significant with a consistent direction on the test splits of both corpora, which supports the use of a healthy reference for predictive deviation features (Proposition 1). Its stability across corpora that differ in language, platform and recording conditions suggests a recurring association between predictive deviation and depression labels across these datasets. The other two quantities summarize different dynamical operations: $\mathrm{force}_{\mathrm{AR}}$ is fitted to the observed latent trajectory, while $\rho(J)$ describes the healthy predictor at the subject's operating point.

Group separation and screening performance address different questions. On LMVD, R has a significant test-set mean difference ($p = 0.004$), yet residual-only classification has F1 59.69 and permutation $p = 0.507$. A shift in group averages can coexist with substantial overlap and poor classification at the learned threshold. Conversely, the three-feature decision can use their joint configuration, explaining why a consistent group-level association need not identify the best standalone screening feature.

The healthy group has higher $\mathrm{force}_{\mathrm{AR}}$ (LMVD $p = 0.0002$), corresponding to faster fitted latent decay under the stable AR(1) interpretation (Kuppens et al., 2010; Koval et al., 2013). The healthy group also has higher $\rho(J)$ (LMVD 0.938 vs. 0.731). If both measured the same stable transition, the latter would instead imply slower decay; here they are computed from the observed sequence and a conditional predictor, respectively. We therefore interpret their joint value in terms of complementary predictive information, rather than attributing the spectral difference to flattened mood or fluctuation amplitude. The residual measures departure from the healthy reference, and $R + \mathrm{force}_{\mathrm{AR}} + \rho(J)$ outperforms any single quantity (Section 3.4). This pattern rests mainly on LMVD; on D-Vlog the group means of $\mathrm{force}_{\mathrm{AR}}$ and $\rho(J)$ point in the same direction without reaching significance. The distributions in Section G further show the overlap and spread behind these group summaries.

4.2 Pathway complementarity and the route to personalization

The two corpora display different relative strengths of the two pathways. On D-Vlog, NSP is close to its ceiling and IIP contributes by correcting healthy individuals whose presentation resembles depression; on LMVD, IIP and NSP have similar standalone F1, but IIP has higher recall (79.78 vs. 72.81) and lower precision (72.40 vs. 76.55). In both cases the combination exceeds either pathway, supporting the view that whether a sample looks like depression and how far its dynamics depart from the healthy law are complementary kinds of evidence.

How individual information enters the decision also matters. The subject condition shapes the healthy predictor, while the three resulting features determine a sample-specific correction to the screening margin. The predictor and bias-head parameters remain shared across subjects; personalization comes from the reference prediction and feature values computed for each recording. This separates the estimation of predictive deviation from its supervised use in screening. The stop-gradient preserves this division during optimization: CE updates the adapter, while BCE learns how the dynamical features should modify its margin (Proposition 4).

4.3 Limitations

Latent states are taken from omni-modal backbone hidden states rather than frame-level paralinguistic features, the counterfactual is a single-step rollout, and z_{trait} is estimated from the current video and may therefore contain state information. Separating a stable trait from the transient state relies on identifiability results under multi-view or auxiliary-variable conditions (Hyvärinen et al., 2019; Schölkopf et al., 2021); these conditions are not established by a single short video per person, and the cross-session stability of $\rho(J \mid z)$ requires multi-session data to test. The group differences of $\mathrm{force}_{\mathrm{AR}}$ and $\rho(J)$ are significant only on LMVD, and duration-stratified evaluation on D-Vlog is left to future work.

5 Conclusion

We presented InertiaMPDD, which reformulates multimodal depression detection from inter-personal classification into the estimation of within-person deviation from a law of healthy emotional dynamics. A trait-conditioned predictor trained only on healthy dynamics supplies the counterfactual through single-step rollout, and because prediction and observation are compared between adjacent instants of the subject's own trajectory, individual differences are carried by the resulting quantities rather than supplied as annotated attributes. From this comparison we derived three annotation-free individual-level quantities, the residual R, the restoring force $\mathrm{force}_{\mathrm{AR}}$ and the spectral radius $\rho(J)$, with the latter two motivated by emotional-inertia models and evaluated as latent dynamical features.

Combined at the decision level with an omni-modal screening pathway, InertiaMPDD attains F1 of 98.41 on D-Vlog and 84.98 on LMVD, and the joint system exceeds either pathway alone. The residual is significant with a consistent direction on both corpora, and on LMVD their combination improves over the individual features. These results suggest that personalization in depression detection can come from modeling how each person's emotion evolves rather than from annotating who each person is. Extending the counterfactual to multi-step rollouts, and testing the cross-session stability of the individual quantities on multi-session data, are the natural next steps.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Changzeng Fu, Jiacheng Dong, Yikai Su, Yuliang Zhao, Siyang Song, Zelin Fu, Qi Zhang, Wenbo Shi, Shiqi Zhao, Chaoran Liu, Xinhe Kuang, Kaifeng Su, Junfeng Yao, Jiadong Wang, Yuichiro Yoshikawa, Björn Schuller, and Hiroshi Ishiguro. The first MPDD challenge: Multimodal personality-aware depression detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025a. arXiv:2505.10034.

Changzeng Fu, Shiwen Zhao, Yunze Zhang, Zhongquan Jian, Shiqi Zhao, and Chaoran Liu. Personality-guided public-private domain disentangled hypergraph-former network for multimodal depression detection. *arXiv preprint arXiv:2511.12460*, 2025b.

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

Lang He, Zheng Li, Prayag Tiwari, Cui Cao, Jingjing Xue, Feng 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. doi: 10.1016/j.bspc.2024.106490.

Lang He, Kai Chen, Jie Zhao, Yang Wang, Ercheng Pei, Hui Chen, Jie Jiang, Shuo Zhang, Jinhui Zhang, Zhongmin Wang, Tao He, and Prayag Tiwari. LMVD: A large-scale multimodal vlog dataset for depression detection in the wild. *Information Fusion*, 126:103632, 2026. doi: 10.1016/j.inffus.2025.103632.

Marlies Houben, Wim Van Den Noortgate, and Peter Kuppens. The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, 141(4): 901–930, 2015. doi: 10.1037/a0038822.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Mingxuan Hu, Ziqi Liu, Hongbo Ma, Jiaqi Liu, Xinlan Wu, and Yangbin Chen. Mitigating gender bias in depression detection via counterfactual inference. *arXiv preprint arXiv:2512.01834*, 2026.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 859–868, 2019.

Peter Koval, Madeline L Pe, Kristof Meers, and Peter Kuppens. Affect dynamics in relation to depressive symptoms: variable, unstable or inert? *Emotion*, 13(6):1132, 2013.

Peter Kuppens, Nicholas B. Allen, and Lisa B. Sheeber. Emotional inertia and psychological maladjustment. *Psychological Science*, 21(7):984–991, 2010. doi: 10.1177/0956797610372634.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf.

Tao Ling, Daoxin Chen, and Bo Li. MDAVIF: A multi-domain acoustical-visual information fusion model for depression recognition from vlog data. In *ICASSP 2024 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8115–8119, 2024. doi: 10.1109/ICASSP48485.2024.10446185.

Meng Ma, Chao Wang, Shuai Liu, Wei Zhao, and Xin Jiang. BHDD: Facial behaviors modeling and hierarchical multimodal fusion for depression detection from vlog data. *IEEE Transactions on Affective Computing*, 2026.

Qwen Team. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Misha Sadeghi, Robert Richer, Bernhard Egger, Lena Schindler-Gmelch, Lena H. Rupp, Farnaz Rahimi, Matthias Berking, and Bjoern Eskofier. Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*, 3(1): 66, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Zixuan Shangguan, Xiaoxiao Li, Yanjie Dong, and Xiaoyan Yuan. Automatic depression detection using attention-based deep multiple instance learning. In *International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine)*, pp. 40–51. Springer, 2023.

Min Shi, Xin Qiao, Lang He, and Rui Gao. An interaction network based on multiscale spatiotemporal feature fusion for depression prediction. *IEEE Transactions on Affective Computing*, 2026.

Yuchen Tao, Minqiang Yang, Huiru Li, Yizhu Wu, and Bin Hu. DepMSTAT: Multimodal spatio-temporal attentional transformer for depression detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):2956–2966, 2024a. doi: 10.1109/TKDE.2024.3350071.

Yuchen Tao, Minqiang Yang, Yizhu Wu, Kevin 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. doi: 10.1016/j.dcan.2023.03.007.

Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. AVEC 2014: 3D dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, 2014. doi: 10.1145/2661806.2661807.

Ingrid A. van de Leemput, Marieke Wichers, Angélique O. J. Cramer, Denny Borsboom, Francis Tuerlinckx, Peter Kuppens, Egbert H. van Nes, Wolfgang Viechtbauer, Erik J. Giltay, Steven H. Aggen, Catherine Derom, Nele Jacobs, Kenneth S. Kendler, Han L. J. van der Maas, Michael C. Neale, Frenk Peeters, Evert Thiery, Peter Zachar, and Marten Scheffer. Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, 111(1):87–92, 2014. doi: 10.1073/pnas.1312114110.

Honghong Wang, Jing Deng, and Rong Zheng. Personality-enhanced multimodal depression detection in the elderly. *arXiv preprint arXiv:2510.08004*, 2025a.

Lei Wang, Chao Wang, Chen Li, Tetsuya Murai, Yan Bai, Zhen Song, Shuo Zhang, Qian Zhang, Yan Huang, and Xiaojun Bi. AI-assisted multi-modal information for the screening of depression: A systematic review and meta-analysis. *npj Digital Medicine*, 8(1):523, 2025b.

World Health Organization. Depressive disorder (depression). WHO Fact Sheet, 2023. URL `https://www.who.int/news-room/fact-sheets/detail/depression`.

Jiaxin Ye, Junjie Zhang, and Hongming Shan. DepMamba: Progressive fusion mamba for multimodal depression detection. In *ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-Vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12226–12234, 2022. doi: 10.1609/aaai.v36i11.21483.

Linhai Zhang, Ziyang Gao, Deyu Zhou, and Yulan He. Explainable depression detection in clinical interviews with personalized retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9927–9944, 2025.

Wenjie Zheng, Qianyu Xie, Zhicheng Wang, Jianfei Yu, and Rui Xia. Towards explainable multimodal depression recognition for clinical interviews. *arXiv preprint arXiv:2501.16106*, 2025.

Bowen Zhou, Marc-André Fiedler, and Ayoub Al-Hamadi. Cross-modal fusion mamba for multimodal depression detection. In *2025 14th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 225–230. IEEE, 2025a.

Bowen Zhou, Marc-André Fiedler, and Ayoub Al-Hamadi. STCM-Mamba: Multimodal spatio-temporal cross-modal mamba for depression detection. *IEEE Access*, 13:189494–189505, 2025b.

Licai Zhou, Zhenyu Liu, Zixuan Shangguan, Xiaoyan Yuan, Yang Li, and Bin Hu. TAMFN: Time-aware attention multimodal fusion network for depression detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:669–679, 2022. doi: 10.1109/TNSRE.2022.3224135.

Licai Zhou, Zhenyu Liu, Xiaoyan Yuan, Zixuan Shangguan, Yang Li, and Bin Hu. CAIINET: Neural network based on contextual attention and information interaction mechanism for depression detection. *Digital Signal Processing*, 137:103986, 2023. doi: 10.1016/j.dsp.2023.103986.

Fangjie Zhu, Haifeng Lu, Sicheng Zhao, Runhao Zeng, and Xiping Hu. DepressionAgent: Reading, listening, seeing, and deliberating multimodal evidence for depression risk assessment. *arXiv preprint arXiv:2608.13891*, 2026.

A Extended Related Work

Omni-modal depression detection. Early work recognized depression from hand-crafted paralinguistic features with shallow fusion, and established evaluation benchmarks on clinical interview corpora such as DAIC-WOZ (Gratch et al., 2014) and the AVEC challenge series (Valstar et al., 2014). With the development of large language models and omni-modal large models, recent methods exploit their joint understanding and reasoning over language, speech and facial behavior, and use their generative capacity to provide symptom-level explanations (Zhu et al., 2026). Such methods have advanced the characterization of population-level depression representations, but generally cast detection as classification across individuals, so that individual difference does not enter the framework as an object of modeling. Our IIP complements population-level screening by learning a healthy reference and measuring predictive departure on each subject's own trajectory.

Counterfactual and causal perspectives. Counterfactual inference has been introduced into depression detection, but mostly for debiasing at the population level, for instance by cutting the spurious causal path from gender to prediction in order to mitigate gender bias in acoustic models (Hu et al., 2026). Our counterfactual instead targets the evolution law rather than a population prediction: it asks what the next instant should be if this stretch of dynamics followed healthy dynamics, and the comparison is completed between adjacent instants of the subject's own trajectory. Prediction and observation share the same trajectory and conditions, with inferred subject characteristics conditioning the reference. Here counterfactual denotes a model-based healthy reference, rather than an identified causal intervention. This perspective turns the data deficiency of missing individual annotation into an inference problem solvable by self-supervised prediction constrained to healthy samples.

World models and latent-space prediction. Predicting the next state from the current one in a latent space rather than in signal space is an effective paradigm for dynamics modeling: it avoids modeling unpredictable signal detail, and variance and covariance regularization prevent representational collapse (LeCun, 2022; Assran et al., 2023; Bardes et al., 2022). We take inspiration from this idea without claiming to build a general world model: our predictor is a lightweight single-step rollout module specific to healthy emotional dynamics and conditioned on the individual trait, whose purpose is to characterize emotional evolution under the healthy baseline so as to support the computation of the counterfactual residual and the inertia quantities.

Emotional inertia and emotion dynamics. In affective science, emotional inertia characterizes the strength of emotional autocorrelation through a first-order autoregressive coefficient, reflecting how quickly a person returns to baseline after a perturbation, and has been shown to relate to depression (Kuppens et al., 2010; Koval et al., 2013). A meta-analysis of short-term emotion dynamics and psychological well-being further relates both emotional variability and inertia to psychological maladjustment (Houben et al., 2015). Critical slowing down, in which autocorrelation and variance rise together and recovery becomes markedly slower as a system approaches a tipping point, has been used as an early warning of depressive transitions (van de Leemput et al., 2014). This construct motivates our latent AR feature and the spectral descriptor of the individually conditioned predictor. Their connection to recovery is exact in the scalar transition special case below; clinical interpretation in multimodal latent space requires additional evidence.

Identifiability. Separating trait from state relies on identifiability results for representation learning and nonlinear ICA under conditions such as multiple views and auxiliary variables (Hyvärinen et al., 2019; Schölkopf et al., 2021). Whether these conditions hold delimits the extent to which the present method can cleanly separate a stable trait from the transient state, and we state this explicitly when discussing limitations.

B Theoretical Properties and Proofs

This section summarizes the notation of Section 2 and gives the proofs of its propositions.

Notation. Let $s_t \in \mathbb{R}^d$ ($d = 256$) denote the individual latent state sequence and $z = z_{\text{trait}} \in \mathbb{R}^{16}$ the individual trait. We model the subject's emotional evolution as $\tilde{s}_{t+1} = f^\star(s_{\le t}; z) + \varepsilon_t$ with

$\mathbb{E}[\varepsilon_t \mid s_{\le t}, z] = 0$ and conditional covariance $\Sigma_\varepsilon(s_{\le t}, z)$, where $f^\star$ is the one-step transition map. An online encoder E gives $s_t = E(x_t)$, and a frozen, semi-orthogonal target encoder E' gives the reference $\tilde{s}_t = E'(x_t)$. We write m for the discriminative margin and b for the individual bias.

Proof of Proposition 1 (bias–noise decomposition). Write the observation as $\tilde{s}_{t+1} = g(s_{\le t}; z) + \varepsilon_t$ with $\mathbb{E}[\varepsilon_t \mid s_{\le t}, z] = 0$ and conditional covariance $\Sigma_\varepsilon(s_{\le t}, z)$, and the prediction as $\hat{s}_{t+1} = f_\theta(s_{\le t}; z)$. Then

$$\begin{aligned}\mathbb{E}\big[r_t^2 \mid s_{\le t}, z\big] &= \mathbb{E}\big[\|g(s_{\le t}; z) + \varepsilon_t - f_\theta(s_{\le t}; z)\|_2^2 \mid s_{\le t}, z\big] \\ &= \|g - f_\theta\|_2^2 + 2\,\mathbb{E}[\varepsilon_t \mid s_{\le t}, z]^\top \big(g - f_\theta\big) + \mathbb{E}\big[\|\varepsilon_t\|_2^2 \mid s_{\le t}, z\big] \\ &= \|g - f_\theta\|_2^2 + \mathrm{tr}(\Sigma_\varepsilon(s_{\le t}, z)), \end{aligned} \tag{B.1}$$

where the cross term vanishes because $\mathbb{E}[\varepsilon_t \mid s_{\le t}, z] = 0$ and g, f_θ are deterministic given $s_{\le t}, z$. For a healthy subject $g \approx f^\star_{\mathcal{H}} \approx f_\theta$, so the departure term tends to zero and the residual reduces to the irreducible noise; otherwise it is elevated by exactly that amount. □

Proof of Proposition 2. For $d = 1$ and $f_\theta(s) = c + \varphi(s - c)$ the Jacobian is the scalar $J = \partial f_\theta / \partial s = \varphi$, so $\rho(J) = |\varphi|$. Under a stationary AR(1) model with white innovations and $|\varphi| < 1$, $s_{t+1} - c = \varphi(s_t - c) + \eta_t$, φ is exactly the lag-one autocorrelation of the centered series, which is the operational definition of emotional inertia; it equals $\rho(J)$ only for $\varphi \ge 0$. □

Proof of Proposition 3. Under the common-coordinate first-order transition assumptions of Proposition 3, a Taylor expansion about the fixed point $\bar{s}$ gives $\delta_{t+1} \approx J\delta_t$, hence $\delta_{t+k} \approx J^k \delta_t$. For diagonalizable $J = V\Lambda V^{-1}$ we have $\|J^k\| \le \kappa(V)\,\rho(J)^k$ with $\kappa(V)$ the condition number of V, which bounds the linearized decay when $0 < \rho(J) < 1$; the slowest mode attains this rate. Solving $\rho(J)^\tau = e^{-1}$ gives $\tau = -1/\ln \rho(J)$. □

Proof of Proposition 4. Writing $\ell_{\mathrm{BCE}} = -y \log \sigma(m + b) - (1 - y) \log(1 - \sigma(m + b))$, we have $\partial \ell_{\mathrm{BCE}} / \partial m = \sigma(m + b) - y$. Without a stop-gradient this term reaches the LoRA parameters ϕ through m. With $\mathrm{sg}(m)$, BCE reaches only the bias-head parameters, so $\partial \mathcal{L} / \partial \phi = \partial \mathcal{L}_{\mathrm{CE}} / \partial \phi$ exactly. This identity isolates the two gradients while retaining the dependence of the bias-head update on the screening margin. □

Remark on the non-normal case. When J is non-normal, that is $JJ^\top \ne J^\top J$, a perturbation may be transiently amplified even if $\rho(J) < 1$. For a discrete-time transition, one-step amplification is governed by the largest singular value of J; spectral radius instead controls asymptotic linear decay. In our implementation, the operating point is a sample mean rather than an established fixed point, and J is a partial derivative of a history-conditioned online-to-target predictor. It is therefore a local predictive descriptor, not an identified physiological recovery coefficient. These distinctions allow $\mathrm{force}_{\mathrm{AR}}$ and $\rho(J)$ to show different empirical group patterns.

C Implementation and Hyperparameters

Table C.1 lists the implementation settings. The three-input bias head is kept in fp32 with its own optimizer parameter group and learning rate. The target encoder is frozen and semi-orthogonal, supplying a fixed target-state reference; the VICReg term regularizes the variance and covariance of the online representation in Equation (1).

Extraction of the decision state. Only the assistant segment is supervised, and labels are not shifted. Let `valid` be the set of positions with labels $\ne -100$ and `_PREFIX_LEN` $= 4$. Then

$$\mathrm{pos} = \min\Big(\max\big(\texttt{valid}[0],\ \texttt{valid}[0] + \texttt{_PREFIX_LEN} - 1\big),\ \texttt{valid}[-1]\Big), \tag{C.1}$$

that is the hidden state one step before the model writes out the label.

Table C.1: Implementation and hyperparameters.

Item	Value
Backbone	Qwen2.5-Omni-3B thinker (bf16, frozen)
LoRA	rank 8, trained from scratch, lr 1×10^{-4}, batch 1 $\times$ grad-accum 8, linear schedule
Bias head	Linear($3 \to 1$), fp32, AdamW lr 1×10^{-2} (separate parameter group), $\mathcal{N}(0, 0.1)$ init
IIP	online encoder $4096 \to 256$; FiLM healthy predictor (3-layer causal Transformer); target encoder semi-orthogonal and frozen
Losses	NSP: CE; bias head: BCE (stop-gradient on the margin); IIP: MSE + VICReg
Input limits	audio ≤ 300 s (25 tok/s); video frames ≤ 32
Latent state	1 Hz pooled concatenation of audio 2048 and visual 2048 ($x_t \in \mathbb{R}^{4096}$) $\to$ $s_t \in \mathbb{R}^{256}$

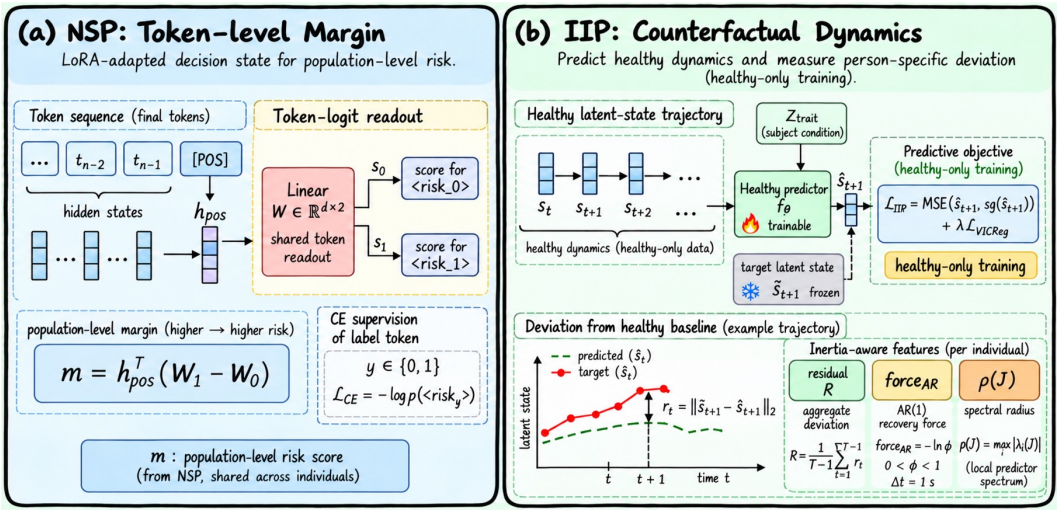


Figure E.1: **Core modules. (a) NSP: token-level margin.** The decision state h_{pos} at [POS] passes through a LoRA-adapted head $W \in \mathbb{R}^{d\times 2}$; cross-entropy supervises the label token and $m = h_{\text{pos}}^{\top}(W_1 - W_0)$ is the shared risk score. **(b) IIP: counterfactual dynamics.** The healthy predictor f_θ, conditioned on z_{trait}, gives $\hat{s}_{t+1}$, supervised by the frozen target state through $\mathcal{L}_{\text{IIP}} = \text{MSE}(\hat{s}_{t+1}, \text{sg}(\tilde{s}_{t+1})) + \lambda\mathcal{L}_{\text{VICReg}}$ on healthy samples only. The distance between observation and healthy prediction is r_t, giving $R = \frac{1}{T-1}\sum_{t=1}^{T-1} r_t$, $\text{force}_{\text{AR}} = -\ln\varphi$ and $\rho(J) = \max_i |\lambda_i(J)|$.

D Statistical Testing Protocol

Group differences are assessed by permutation tests, with the residual tested one-sided under the pre-specified direction that depressed subjects are higher. The significance of detection performance against random labels is assessed by a label permutation test. LMVD results are means over five random seeds. The feature ablation (Table 3) evaluates individual quantities, their combination and the audio-duration control.

E Additional Figures and Tables

This section collects the module-level figure referenced from the main text. Figure E.1 expands the two pathways of Figure 2 into their computational details: panel (a) shows where the margin is read out of the token sequence, which matters because the decision state must be taken exactly one step before the label token is written (Section C); panel (b) shows the predictive objective of IIP and the point at which the three individual quantities are derived from a single trajectory. Reading the two panels together makes explicit that LoRA is supervised by CE and the decision-bias head by BCE. The IIP prediction objective is self-supervised within a training pool selected using healthy labels.

Algorithm 1 InertiaMPDD: training and inference

Require: Training set $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}$, healthy pool $\mathcal{H} \subset \mathcal{D}$ with $y = 0$, frozen omni-modal backbone Φ, frozen target encoder E'

Ensure: LoRA parameters ϕ, online encoder E, healthy predictor f_θ, bias head (w, c)

Stage 1: healthy-dynamics pretraining (healthy pool).
for each sample $x \in \mathcal{H}$ **do**
  $x_{1:T} \leftarrow$ hidden states of $\Phi(x)$, pooled to 1 Hz and concatenated across modalities
  $s_t \leftarrow E(x_t)$; $\tilde{s}_t \leftarrow E'(x_t)$ {E' frozen, semi-orthogonal}
  $z_{\text{trait}} \leftarrow$ first- and second-order statistics of $x_{1:T}$ pooled to $\mathbb{R}^{16}$
  $\hat{s}_{t+1} \leftarrow f_\theta(s_{\le t}; z_{\text{trait}})$ {FiLM conditioning}
  Update θ, E by Equation (1)
end for
Stage 2: feature extraction (all samples, f_θ and E frozen).
for each sample $x \in \mathcal{D}$ **do**
  $r_t \leftarrow \|\tilde{s}_{t+1} - \hat{s}_{t+1}\|_2$; $R \leftarrow \frac{1}{T-1}\sum_{t=1}^{T-1} r_t$
  $\varphi \leftarrow$ AR(1) coefficient of $\{s_t\}$ about $\bar{s}$; $\text{force}_{\text{AR}} \leftarrow -\ln \varphi$
  $J \leftarrow \partial f_\theta / \partial s_t|_{\bar{s}}$; $\rho(J) \leftarrow \max_i |\lambda_i(J)|$
  $u \leftarrow [R,\ \text{force}_{\text{AR}},\ \rho(J)]$
end for
Compute the training-set mean and standard deviation of u for standardization $z(\cdot)$
Stage 3: joint training.
for each minibatch $(x, y) \in \mathcal{D}$ **do**
  $h_{\text{pos}} \leftarrow$ decision state of $\Phi_\phi(x)$; $m \leftarrow h_{\text{pos}}^\top (W_1 - W_0)$
  $b \leftarrow w^\top z(u) + c$
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}}(y) + \text{BCE}(\sigma(\text{sg}(m) + b), y)$ {CE $\to \phi$; BCE $\to (w, c)$ only}
  Update ϕ and (w, c)
end for
Inference. $\hat{y} \leftarrow \mathbb{1}[\, m + b \ge 0\,]$

F Full Algorithm

Algorithm 1 gives the full procedure. The three stages are deliberately separated rather than trained end to end. Stage 1 uses labels to select healthy training samples and then fits a self-supervised prediction objective. This defines the healthy reference against which predictive deviation is measured. Stage 2 freezes the predictor so that the three quantities are fixed features rather than moving targets, which also makes the standardization statistics well defined. Stage 3 trains the adapter and the bias head under the two losses, with the stop-gradient keeping them on separate parameters.

G DISTRIBUTIONS OF THE DYNAMICAL QUANTITIES

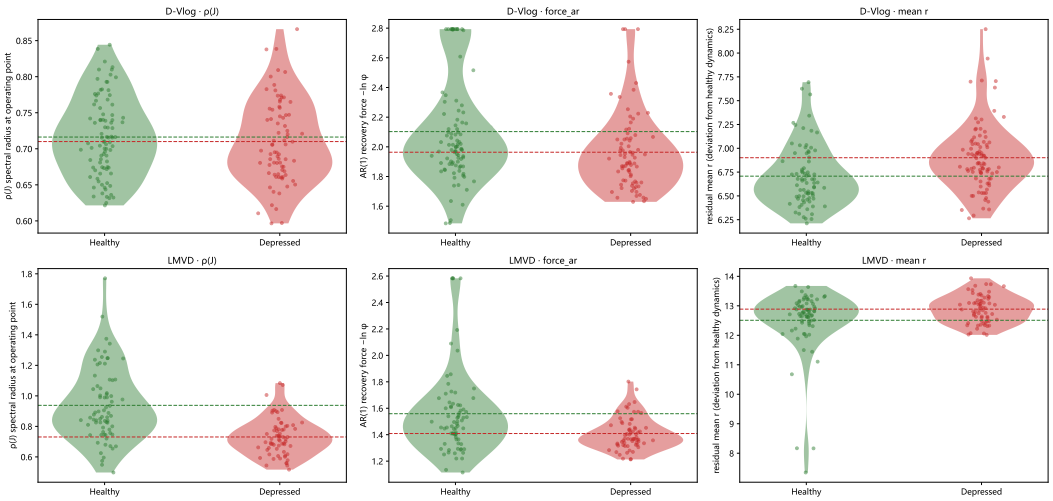


Figure G.1: Distributions of the three IIP quantities on D-Vlog (top) and LMVD (bottom). Columns show $\rho(J)$, $\mathrm{force}_{\mathrm{AR}}$, and the mean residual R (labeled mean r in the plot), respectively. Green and red denote healthy and depressed samples. Violins show distributional shape, dots show individual samples, and dashed lines mark the displayed group reference levels. Each panel uses its own vertical scale.

Figure G.1 shows the spread, overlap and tails underlying the individual-level quantities. These distributions complement the group summaries in Table 2 and the predictive comparisons in Table 3.

Residual R: overlapping distributions and asymmetric tails. On both corpora, the depressed group has the higher residual reference level, consistent with Table 2. On D-Vlog, healthy samples concentrate at lower residuals, while the depressed distribution shifts upward and has a longer upper tail; the two groups still overlap substantially. On LMVD, their central portions also overlap, while healthy samples show a pronounced lower tail and depressed samples concentrate at larger residual values. These asymmetric tails show why a group-average difference need not imply a clean sample-level threshold. Each point summarizes the temporal mean residual of a recording rather than an individual prediction step.

AR restoring force: shared ranges with different concentration. The D-Vlog groups occupy similar ranges of $\mathrm{force}_{\mathrm{AR}}$, with considerable overlap and upper-end observations in both groups. On LMVD, the depressed distribution is concentrated near the lower part of the healthy distribution, whereas healthy samples extend further into the upper tail. Under the stable AR(1) interpretation, a larger $-\ln\varphi$ corresponds to faster fitted latent decay. The overlap shows why this quantity is more useful as one component of the feature set than as a universal cutoff.

Spectral radius: a stronger distributional contrast on LMVD. The D-Vlog $\rho(J)$ distributions overlap closely around 0.7. On LMVD, healthy samples span a broader range and extend to larger values, while depressed samples concentrate near the lower part of that range. Some LMVD values exceed one, outside the stable-decay regime of Proposition 3; $\rho(J)$ is therefore read as a local predictor descriptor. Its broader healthy-group distribution and the AR decay pattern characterize different operations, rather than two interchangeable measures of emotional recovery. Together, the six panels motivate evaluating the joint feature configuration rather than relying on a single group-level direction.