# PRISM: A Frozen-Backbone Framework for Cross-Distribution Generalization in Longitudinal Behavioral Data

**Ahatsham Hayat**[1] **Mohammad Rashedul Hasan**[2]
[1]Department of Anesthesiology, Brigham and Women's Hospital,
Harvard Medical School, Boston, MA, USA
[2]Department of Electrical and Computer Engineering,
University of Nebraska-Lincoln, NE, USA
afnu@bwh.harvard.edu, hasan@unl.edu

## Abstract

Cross-distribution generalization remains a major challenge in longitudinal behavioral modeling, as models trained on one cohort often perform poorly on new populations. Existing approaches based on feature-distribution alignment, increased model capacity, or generative reformulations have achieved limited improvements. We investigate this limitation from the perspective of representational constraints during training. Our approach uses a frozen language model to impose a fixed coherence criterion on representations during training. We introduce PRISM, a frozen-backbone framework that integrates temporal, spectral, and semantic representations through directed cross-attention with instance-specific gating. A dual-path training objective combines discriminative supervision with narrative-generation constraints on the shared representation. On GLOBEM, a cross-cohort benchmark for behavioral health prediction, PRISM achieves 79.93% out-of-distribution (OOD) accuracy, improving the previous classification-based benchmark by 27.13 percentage points and the strongest LLM-based result by 12.53 points, with an in-distribution to OOD (ID-to-OOD) accuracy gap of 1.24 points. PRISM also achieves the highest OOD accuracy on three additional datasets spanning anxiety, educational engagement, and schizophrenia symptom forecasting, with consistently smaller ID-to-OOD gaps than fine-tuned VLM baselines. A frozen-LLM direct-prediction control trails PRISM's generation path by 8.47–14.27 points, while comparisons of PRISM's two paths show an additional inference-path advantage for the generative route in this setup.

## 1 Introduction

Longitudinal behavioral data, combining subjective self-reports with objective sensor measurements collected from individuals over extended periods, underpins predictive modeling in mental health monitoring, educational engagement, and workplace well-being (Saeb et al., 2015; Xu et al., 2023; Patharkar et al., 2024). The central obstacle in this domain is **cross-distribution generalization**. Models trained on one cohort or time period consistently degrade when applied to new populations or contexts. On GLOBEM (Xu et al., 2023), the largest multi-year benchmark designed for cross-cohort evaluation of behavioral health prediction, the best published out-of-distribution (OOD) accuracy for binary depression forecasting has remained at 52.8% for several years, barely above chance. Classical machine learning baselines, deep learning architectures, and domain generalization methods from the broader literature (Zhou et al., 2022) have not substantially improved this result. Recent large language model (LLM)-based reformulations of the task, which replace classification with

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

prospective narrative generation (Hayat et al., 2025), have narrowed the gap to 67.40% but have not closed it. Prior work combining data from two longitudinal datasets for cross-cohort prediction found that combined-dataset training helped only when the training and target distributions were similar (Adler et al., 2022). The persistence of this ceiling across methodologically diverse approaches suggests the limitation is not one of model capacity. Existing approaches primarily address differences between training and test distributions or increase model capacity, without explicitly constraining the representations learned from longitudinal behavioral data.

We diagnose the problem as one of insufficient representational constraints during training. Under a standard classification objective, the representation receives sparse supervision per sample, providing no guidance on how the encoding should be organized beyond whether the final output was correct. This permits shortcut representations that generalize poorly. Reframing the forecasting task as prospective narrative generation (Hayat et al., 2025) provides a richer training signal and accounts for most of the progress above the classification ceiling on GLOBEM. The remaining gap stems, in part, from the training procedure. Fine-tuning the LLM on the training distribution can cause its narrative generation to reflect distribution-specific patterns. The coherence standard that shapes the representation is consequently defined by a model trained on the same data. This failure mode is consistent with the information bottleneck principle (Tishby et al., 2000; Tishby and Zaslavsky, 2015). When the generative model co-adapts with the representation, both can converge on distribution-specific sufficient statistics that do not transfer.

To prevent this co-adaptation, we propose keeping the language model frozen during training. Narrative outputs are evaluated against a fixed pre-trained language prior that the representation cannot influence. Training encourages embeddings that support coherent continuations under the fixed decoder, without adapting its weights to distribution-specific patterns. Language thus serves as a constraint on the representation, not as a learned output channel. This formulation places representational constraints at the center of cross-distribution generalization.

Learning representations that satisfy this constraint requires capturing complementary aspects of behavioral structure. Raw multivariate sequences preserve exact values and cross-feature correlations at each time step but leave the temporal-frequency structure of behavioral change implicit. Spectral decompositions resolve when and at what scales behavioral patterns shift, although the choice of spectral format matters. Vision-language approaches that convert time-series signals into Fourier-based visual encodings assume stationary periodic structure (Zhong et al., 2025), an assumption that may not hold for behavioral trajectories. A spectral representation for this domain needs to maintain joint time-frequency localization. Neither raw sequences nor spectral decompositions explicitly encode the contextual meaning of behavioral patterns needed for narrative reasoning. This motivates a multimodal decomposition of each trajectory into complementary representations, each capturing a different aspect of behavioral structure.

To implement this approach, we introduce PRISM (**P**rospective **R**easoning through **I**ntegrated **S**pectral-temporal **M**ultimodal Learning), a frozen-backbone framework that decomposes each behavioral trajectory into three complementary streams, each encoding a different aspect of its behavioral structure. A temporal stream preserves the raw multivariate sequence through a trainable transformer encoder. A spectral stream extracts time-frequency structure using Continuous Wavelet Transform (CWT) scalograms (Mallat, 1999), processed by a frozen vision encoder with a trainable input-side domain bridge. This representation encodes when behavioral patterns change and at which frequency scales. A semantic stream processes LLM-generated meta-narratives through a frozen text encoder with a trainable injection layer. Directed cross-attention integrates the three streams using instance-specific gating. The fused representation supports two output paths. A prediction head generates outcome forecasts. A soft-prompt projection maps the fused embedding into the input space of a frozen LLM decoder, which generates a prospective behavioral narrative from which outcome predictions are extracted. The prediction loss offers sparse categorical supervision, while the generation loss imposes dense token-level supervision governed by the frozen LLM's pre-trained weights.

On GLOBEM, PRISM achieves 79.93% OOD accuracy, improving the previous classification-based state-of-the-art (SOTA) by 27.13 percentage points and the strongest LLM-based prior result by 12.53 points, with an in-distribution to OOD (ID-to-OOD) accuracy gap of only 1.24 points. PRISM also achieves the highest OOD accuracy on three additional longitudinal datasets spanning anxiety (LifeSnaps), educational engagement (MFAFY), and schizophrenia symptom forecasting

(CrossCheck), reaching 84.97% on the CrossCheck cross-cohort split. Across all four datasets, PRISM maintains consistently smaller ID-to-OOD gaps than fine-tuned VLM baselines. The frozen-backbone design requires only a few hours of training on a single GPU. Ablation results indicate that combining the prediction and generation losses under a frozen coherence constraint is critical to the observed cross-distribution generalization.

Our contributions are as follows:

- **I.** We formulate cross-distribution generalization in longitudinal behavioral modeling as a problem of representational constraints, using a frozen generative model to impose a fixed coherence criterion during training.
- **II.** We introduce PRISM, a frozen-backbone framework that integrates complementary temporal, spectral, and semantic representations through directed cross-attention and instance-specific gating. A dual-path training objective combines discriminative supervision with a coherence constraint imposed by the frozen generative model.
- **III.** We evaluate PRISM on four longitudinal datasets, showing substantial improvements in OOD accuracy and consistently smaller generalization gaps compared with existing baselines. Ablation studies examine the contributions of the dual-path objective and individual input modalities.

## 2 Related Work

**Longitudinal behavioral modeling.** Longitudinal passive sensing from smartphones and wearables has been applied to depression detection (Saeb et al., 2015; Wang et al., 2018; Chikersal et al., 2021), stress monitoring (Sano et al., 2018), and academic performance forecasting (Wang et al., 2016b), with most prior methods training and evaluating within a single population. LLM-based reformulations convert behavioral trajectories into text through prompted or fine-tuned models (Kim et al., 2024; Thach et al., 2025; Hayat et al., 2025), adapting either the input format (prompt design) or the model parameters (fine-tuning) to the behavioral domain. Research on cross-study generalization has addressed the population-transfer problem by combining heterogeneous longitudinal datasets and aligning their feature distributions (Adler et al., 2022). Where these approaches adapt the model or its training procedure to the behavioral domain, PRISM keeps the foundation model frozen and forces the learned representation to satisfy a criterion that the training process cannot shift.

**Frozen foundation models.** Frozen or partially fine-tuned LLMs have been repurposed for time series analysis as end-to-end sequence processors. Examples include reprogramming time series into text prototypes (Jin et al., 2024), frozen pre-trained transformers as time series analyzers (Zhou et al., 2023), and frozen VLMs for stationary periodic signals such as electricity and traffic (Zhong et al., 2025). A separate line uses frozen models as embedded modules within larger architectures, including soft prompting of frozen language models for multimodal conditioning (Tsimpoukelli et al., 2021) and frozen LLM blocks acting as visual information filters (Pang et al., 2024). **PRISM's use of the frozen language model differs from both lines**. It uses the frozen model as a fixed coherence criterion the learned representation has to satisfy. This is analogous to physics-informed neural networks (Raissi et al., 2019), where fixed physical laws constrain the solution and prevent convergence onto training-data-specific shortcuts. The design also differs from standard multi-task learning (Kendall et al., 2018; Ruder, 2017), where auxiliary task heads are trainable and co-adapt with the primary task, offering no structural reason to prevent distribution-specific overfitting.

**Spectral representations for non-stationary signals.** The CWT provides joint time-frequency localization for signals whose frequency content varies over time (Mallat, 1999; Torrence and Compo, 1998) and has been combined with pre-trained CNNs for ECG classification (Wang et al., 2021) and fault detection (Siddique et al., 2025). These approaches fine-tune the vision backbone on scalogram images, which risks overfitting when training data is limited. PRISM instead keeps the vision encoder frozen and uses trainable input-side adaptation to bridge the domain gap. Adapting the pre-trained encoder to behavioral training data would dissolve the distribution-invariance property the frozen encoder is being used to establish.

Additional discussion of related literature is provided in Appendix A.3.

# 3 Methodology

## 3.1 Problem Formulation

Let $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$ denote a dataset of longitudinal behavioral trajectories, where each $\mathbf{X}_i \in \mathbb{R}^{T \times d}$ records $d$ behavioral features over $T$ time steps, and $y_i \in \mathcal{Y}$ is the outcome label. Each feature vector at time step $t$ combines objective sensor measurements (physical activity, sleep, phone usage) with subjective self-reports (mood, stress, engagement). The experiments use binary targets $\mathcal{Y} = \{0, 1\}$. The architecture extends to multi-class, ordinal, and continuous targets by modifying the output heads.

Under cross-distribution generalization, we train $f : \mathbb{R}^{T \times d} \to \mathcal{Y}$ on a distribution $\mathcal{D}_T$ drawn from population $\mathcal{P}$ during time period $T$ and evaluate on $\mathcal{D}_{T'}$ drawn from a potentially different population $\mathcal{P}'$ during $T'$, with the prediction task fixed and $\mathcal{D}_{T'}$ unseen during training. Each trajectory $\mathbf{X}_i$ is paired with a meta-narrative $\mathbf{M}_i$, a natural language interpretation of the behavioral patterns generated offline following the protocol in (Hayat et al., 2025). The meta-narrative generation procedure is described in Appendix A.4.

## 3.2 PRISM Framework

Figure 1 illustrates how PRISM realizes the representational-constraint principle through three components. A modality decomposition provides the three input streams on which the coherence constraint operates. A directed fusion mechanism integrates these representations while respecting their functional asymmetry (Section 3.2.1). A dual-path training objective applies the coherence constraint through a frozen language prior (Section 3.3). All foundation model encoders remain frozen throughout training. Full architectural and training specifications are detailed in Appendix A.1.

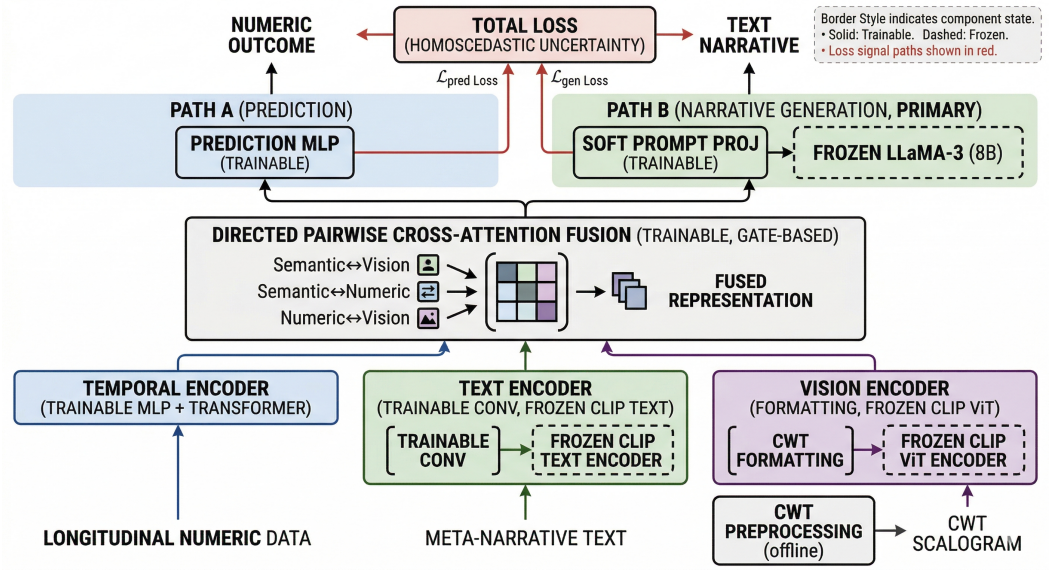


Figure 1: The PRISM framework. Three input streams (left, longitudinal numeric data; center, CWT scalogram; right, meta-narrative text) are processed through modality-specific encoders that produce token-level sequence representations. The temporal encoder is fully trainable; the CLIP vision and text encoders are frozen, with trainable input-side adaptation layers. Directed pairwise cross-attention with learned sigmoid gates produces a compact fused representation $\mathbf{F}_{\text{fused}}$, which feeds two output paths. Path A uses a prediction MLP for fast outcome forecasting; Path B projects the fused embedding into soft prompt tokens consumed by a frozen LLaMA-3 decoder for prospective narrative generation. Both paths train jointly through homoscedastic task uncertainty weighting. Solid borders indicate trainable components; dashed borders indicate frozen components.

**Temporal Stream.** The temporal stream preserves the raw multivariate sequence. A learned linear projection embeds each trajectory into $d_{\text{model}} = 256$ dimensions, and a two-layer transformer encoder (Vaswani et al., 2017) with sinusoidal positional encoding produces the token sequence

$\mathbf{Z}_{\text{temporal}} \in \mathbb{R}^{T \times d_{\text{model}}}$, in which each position retains the contextualized representation of its time step. This stream is trained from scratch on behavioral data, which lets it learn representations matched to the domain's feature statistics.

**Spectral Stream.** The Continuous Wavelet Transform (CWT) (Mallat, 1999; Torrence and Compo, 1998; Grossmann and Morlet, 1984) provides joint time-frequency localization, producing a scalogram that represents the trajectory as a two-dimensional surface over time and frequency. Unlike the Fourier transform, which reports only which frequencies are present across the full window, CWT preserves when each frequency is active.

The time-frequency structure of CWT scalograms differs substantially from the natural image distributions on which pre-trained vision encoders are trained. Trainable multi-scale convolutions transform the scalogram into a representation better suited to the vision encoder's pre-trained feature detectors. During training, gradients propagate from the downstream losses through the frozen encoder to these convolutions, which learn to express temporal-spectral patterns through visual features such as edges, gradients, and textures. This adaptation bridges the domain gap without modifying the vision encoder's weights, limiting the number of trainable parameters and reducing the risk of overfitting on small behavioral datasets. The frozen CLIP vision encoder (Radford et al., 2021) then extracts a sequence of patch tokens $\mathbf{Z}_{\text{vision}} \in \mathbb{R}^{P \times d_{\text{clip}}}$.

**Semantic Stream.** The meta-narrative $\mathbf{M}_i$ situates behavioral observations within contextual meaning (Fillmore, 1982). The same behavioral pattern (a simultaneous decline in physical activity and increase in sleep duration) has different implications for a student approaching final exams than for a post-surgical patient. A trainable injection MLP adapts the tokenized meta-narrative embeddings to the behavioral domain before they reach the frozen CLIP text encoder (Radford et al., 2021). We use the CLIP text encoder because its text and vision branches were co-trained to produce aligned representations, providing alignment for directed cross-attention between semantic tokens and spectral patches in the fusion stage. A language-only text encoder would lack this alignment. The frozen text encoder produces $\mathbf{Z}_{\text{text}} \in \mathbb{R}^{L \times d_{\text{clip}}}$.

### 3.2.1 Directed Cross-Modal Fusion

The three modality sequences reside in different latent spaces and encode functionally asymmetric information. Standard fusion approaches (concatenation, symmetric self-attention) treat all cross-modal relationships as interchangeable. The three streams play different functional roles. Those roles impose a directed structure on how their tokens interact. Semantic tokens query spectral patches to identify which visual disruption patterns the current interpretation points to. Semantic tokens query temporal positions to ground interpretation in specific behavioral changes. Temporal positions query spectral patches to anchor numeric measurements to their frequency signatures. Information flows asymmetrically from spectral and temporal streams (evidence) into semantic and numeric representations (interpretation).

Fusion proceeds in two stages. In Stage 1, three directed multi-head cross-attention blocks compute enriched sequence representations $\mathbf{H}_{S \leftarrow V}$, $\mathbf{H}_{S \leftarrow N}$, and $\mathbf{H}_{N \leftarrow V}$, where the subscript $X \leftarrow Y$ denotes stream $X$ querying stream $Y$'s tokens. Each enriched sequence is mean-pooled along its token dimension to produce a compact vector. In Stage 2, instance-specific sigmoid gates combine the three pooled vectors. The gates and fused embedding are computed as

$$\mathbf{g} = \sigma(\mathbf{W}_g \left[\mathbf{E}_{S \leftarrow V}; \mathbf{E}_{S \leftarrow N}; \mathbf{E}_{N \leftarrow V}\right] + \mathbf{b}_g), \quad \mathbf{F}_{\text{fused}} = \textstyle\sum_{j=1}^{3} g_j \odot \mathbf{E}_j, \tag{1}$$

where $\mathbf{g} = [g_1, g_2, g_3]$ weights each cross-modal pathway by its reliability for the current input and $\odot$ is element-wise multiplication. The gates allow the model to suppress unreliable pathways on a given input (e.g., spectral pathways on a trajectory with minimal behavioral periodicity) and expose which cross-modal relationships the model relies on. Appendix A.1.4 provides full cross-attention equations and pooling details.

## 3.3 Training Objective and Inference

The training objective applies the coherence constraint through a frozen language prior. The fused embedding $\mathbf{F}_{\text{fused}}$ feeds two output paths in parallel. Path A is a prediction head that generates an outcome forecast. Path B is a soft-prompt projection into a frozen LLaMA-3 8B decoder (Grattafiori et al., 2024) that generates a prospective behavioral narrative from which the forecast is extracted.

The two paths share all parameters up to and including $\mathbf{F}_{\text{fused}}$ and differ only in how the forecast is generated.

**Prediction Path.** A two-layer MLP computes a logit $\hat{y}_i$ from $\mathbf{F}_{\text{fused}}$, trained with standard binary cross-entropy $\mathcal{L}_{\text{pred}}$.

**Generation Path.** A trainable projection maps $\mathbf{F}_{\text{fused}}$ into $k = 32$ soft prompt tokens in the embedding space of the frozen decoder. These tokens are concatenated with a fixed textual template and passed to the frozen LLaMA-3 8B decoder (Grattafiori et al., 2024) for autoregressive narrative generation. The generation loss $\mathcal{L}_{\text{gen}}$ computes next-token cross-entropy against a target prospective narrative prepared offline. With the decoder weights fixed, gradients propagate through the soft prompt projection to the gated fusion module and modality encoders. The frozen decoder offers a fixed coherence criterion for evaluating prospective behavioral narratives. Minimizing the generation loss encourages the upstream representation to encode the structured information needed to satisfy this criterion. Full equations for the soft prompt projection and generation loss appear in Appendix A.1.5.

**How Dual-Path Training Promotes Distribution-Invariant Representations.** Path A provides a per-sample categorical training signal, while Path B contributes token-level gradients across hundreds of positions, reflecting how well the shared representation supports specific behavioral observations under the frozen decoder. These complementary objectives constrain the representation in different ways. The prediction loss encourages outcome separability, while the generation loss promotes compatibility with the fixed language prior. With the decoder weights frozen, the coherence criterion remains unchanged throughout training. Optimizing both objectives jointly encourages representations that support accurate predictions and coherent behavioral narratives, reducing their reliance on distribution-specific shortcuts.

**Loss Combination.** The prediction and generation losses are measured in incommensurable units. A fixed weighting coefficient would impose an arbitrary exchange rate requiring per-dataset tuning. To avoid this, we adopt the homoscedastic task uncertainty framework (Kendall et al., 2018), giving the total loss $\mathcal{L}_{\text{total}} = e^{-s_1}\mathcal{L}_{\text{pred}} + s_1 + e^{-s_2}\mathcal{L}_{\text{gen}} + s_2$, where $s_1, s_2$ are learnable log-variance parameters. The $e^{-s}$ factor inverse-weights each loss by its estimated noise, and the linear $+s$ term prevents collapse. The converged $s_1, s_2$ values are interpretable as the model's learned estimate of each task's relative reliability and yield the dataset-specific weighting ratios discussed in Section 4.4.

**Inference.** At inference, the generation path serves as the primary prediction mechanism. The fused embedding is projected into soft prompt tokens, which condition the frozen decoder to generate a prospective behavioral narrative. The outcome forecast is then extracted using the parsing procedure in (Hayat et al., 2025). This path yields both a forecast and a narrative describing the behavioral patterns associated with the predicted outcome. The prediction path computes an outcome forecast directly from the fused representation, without invoking the decoder. It serves as a baseline for assessing the contribution of the generation loss to the shared representation (Section 4.4.1) and as a fallback when autoregressive generation is computationally impractical.

# 4 Experiments

We evaluate PRISM on four longitudinal behavioral datasets, comparing against time-series, LLM-based, and VLM-based baselines, with implementation details in Appendix A.5.

## 4.1 Datasets and Evaluation Protocol

GLOBEM (Xu et al., 2023) forecasts depression across years 1–2 (train) and 3–4 (OOD) over multiple universities, a widely used cross-cohort benchmark for behavioral health prediction. LifeSnaps (Yfantidou et al., 2022) forecasts anxiety from four months of Fitbit and survey data, with the first two months for training and the last two for OOD. MFAFY (Hayat et al., 2024a) forecasts educational engagement from longitudinal categorical self-reports of undergraduate STEM students, evaluated cross-year. CrossCheck (Wang et al., 2016a) forecasts mental health symptoms from passive smartphone sensing of individuals with schizophrenia spectrum disorders, evaluated on a temporal cohort split (2015 enrollees as training, 2016–2017 as OOD). Detailed dataset statistics, processing protocols, and OOD split definitions are provided in Appendix A.2.

Models are trained on $\mathcal{D}_T$ and evaluated on a held-out ID split and the OOD test set $\mathcal{D}_{T'}$. The metrics are accuracy, precision, recall, and F1. The primary criterion is OOD accuracy. We quantify distribution invariance through the ID-to-OOD accuracy gap.

### 4.2 Baselines

We compare against three baseline categories. *Time-series models*, PatchTST (Nie et al., 2022) and iTransformer (Liu et al., 2024c), process numerical trajectories without language or vision components. *LLM-based models* serialize behavioral data into text; LLM-Meta (Hayat et al., 2025) is the strongest prior text-only approach. *VLM-based models* include Time-VLM (Zhong et al., 2025), which uses FFT-based visual encodings developed for stationary periodic signals, and a Bimodal VLM baseline that fine-tunes LLaVA-NeXT 7B (Liu et al., 2024a) with LoRA on interleaved per-feature line charts and meta-narratives. As a prior-only control, frozen LLaMA-3 8B (Grattafiori et al., 2024) predicts outcomes zero-shot from the same behavioral meta-narratives, without soft prompts, multimodal fusion, or task fine-tuning. We report both PRISM inference paths as **PRISM$_{\text{pred}}$** (Path A, prediction MLP) and **PRISM$_{\text{gen}}$** (Path B, narrative generation).

### 4.3 Main Results

Table 1: Cross-distribution generalization results ($\mathcal{D}_T \rightarrow \mathcal{D}_{T'}$). Bold indicates best OOD performance per dataset. Shaded cells indicate PRISM results.

| Dataset | Method | In-Distribution (ID) | | | | Out-of-Distribution (OOD) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc | P | R | F1 | Acc | P | R | F1 |
| **GLOBEM** | PatchTST (Nie et al., 2022) | 53.58 | 53.73 | 53.58 | 53.01 | 49.88 | 49.24 | 49.88 | 49.16 |
| | iTransformer (Liu et al., 2024c) | 54.61 | 54.61 | 54.61 | 54.61 | 51.06 | 51.07 | 51.06 | 51.07 |
| | LLM-Meta (Hayat et al., 2025) | 73.99 | 75.93 | 71.93 | 73.87 | 67.40 | 68.81 | 70.00 | 69.40 |
| | Time-VLM (Zhong et al., 2025) | 57.34 | 55.92 | 56.48 | 56.20 | 51.43 | 49.87 | 50.21 | 50.04 |
| | Bimodal VLM | 79.37 | 83.33 | 80.77 | 82.03 | 72.86 | 75.58 | 74.28 | 74.92 |
| | PRISM$_{\text{pred}}$ | 70.85 | 76.39 | 70.10 | 68.74 | 67.08 | 73.49 | 65.40 | 63.22 |
| | **PRISM$_{\text{gen}}$** | 81.17 | 83.39 | 80.75 | 80.69 | **79.93** | **82.53** | **79.06** | **79.14** |
| **LifeSnaps** | PatchTST | 47.83 | 47.83 | 47.83 | 47.83 | 43.75 | 41.23 | 43.75 | 40.47 |
| | iTransformer | 52.17 | 46.81 | 52.17 | 48.83 | 48.44 | 48.14 | 48.44 | 47.99 |
| | LLM-Meta | 64.71 | 77.78 | 63.64 | 70.00 | 67.19 | 63.89 | 74.19 | 68.66 |
| | Time-VLM | 49.28 | 47.11 | 47.86 | 47.48 | 47.36 | 45.02 | 45.71 | 45.36 |
| | Bimodal VLM | 82.35 | 85.71 | 75.00 | 80.00 | 71.88 | 83.33 | 71.43 | 76.92 |
| | PRISM$_{\text{pred}}$ | 75.00 | 85.71 | 66.67 | 66.67 | 68.75 | 74.97 | 67.94 | 66.07 |
| | **PRISM$_{\text{gen}}$** | 87.50 | 87.50 | 90.00 | 87.30 | **81.25** | **83.46** | **80.84** | **80.78** |
| **MFAFY** | PatchTST | 67.39 | 63.66 | 67.39 | 64.34 | 50.57 | 49.68 | 50.57 | 44.31 |
| | iTransformer | 53.26 | 51.77 | 53.26 | 52.47 | 44.29 | 43.05 | 44.29 | 42.42 |
| | LLM-Meta | 70.49 | 65.22 | 60.00 | 62.50 | 64.86 | 61.11 | 67.48 | 64.14 |
| | Time-VLM | 53.41 | 51.76 | 52.33 | 52.04 | 48.27 | 46.14 | 46.82 | 46.48 |
| | Bimodal VLM | 77.05 | 75.76 | 80.65 | 78.12 | 66.57 | 64.47 | 60.87 | 62.62 |
| | PRISM$_{\text{pred}}$ | 69.81 | 67.74 | 65.91 | 66.35 | 65.71 | 62.82 | 59.63 | 59.40 |
| | **PRISM$_{\text{gen}}$** | 75.47 | 74.18 | 72.42 | 73.01 | **73.71** | **78.03** | **66.50** | **66.77** |
| **CrossCheck** | PatchTST | 90.21 | 87.81 | 88.22 | 88.01 | 68.52 | 50.35 | 50.65 | 48.78 |
| | iTransformer | 51.55 | 52.44 | 53.00 | 49.61 | 44.01 | 51.37 | 53.16 | 39.01 |
| | LLM-Meta | 85.56 | 81.92 | 85.53 | 83.26 | 72.99 | 65.21 | 84.69 | 64.28 |
| | Time-VLM | 77.83 | 77.19 | 83.43 | 76.57 | 65.47 | 62.22 | 79.11 | 57.83 |
| | Bimodal VLM | 91.23 | 90.68 | 87.29 | 88.76 | 79.61 | 68.13 | 87.78 | 70.12 |
| | PRISM$_{\text{pred}}$ | 82.98 | 79.23 | 83.18 | 80.52 | 69.41 | 63.43 | 81.90 | 60.98 |
| | **PRISM$_{\text{gen}}$** | 92.78 | 93.06 | 88.92 | 90.68 | **84.97** | **71.35** | **88.86** | **75.19** |

PRISM$_{\text{gen}}$ achieves the highest OOD accuracy on every dataset (Table 1). On GLOBEM, the 79.93% accuracy improves the prior SOTA of 52.8% (Xu et al., 2023) by 27.13 percentage points and exceeds the strongest text-only result, LLM-Meta at 67.40% (Hayat et al., 2025), by 12.53 points. PRISM$_{\text{gen}}$ reaches 81.25% on LifeSnaps, 9.37 points above the next-best baseline. On MFAFY, the 73.71% accuracy exceeds Bimodal VLM by 7.14 points despite a purely categorical feature space that gives the spectral stream less to decompose. The CrossCheck result reaches 84.97%, the highest of the four datasets and 5.36 points above Bimodal VLM. The same dataset has been used in prior cross-study generalization work that combined CrossCheck with StudentLife

and aligned feature distributions (Adler et al., 2022). PRISM operates within CrossCheck alone, achieving cross-distribution generalization through representational constraint instead of feature-level alignment.

**A frozen language prior alone does not explain PRISM's gains.** We evaluated frozen LLaMA-3 8B as a direct zero-shot predictor on the same meta-narratives used by PRISM. As shown in Table 2, it is close to LLM-Meta: it scores 1.74 points lower on GLOBEM and 0.38 points higher on MFAFY, with modest gains of 1.68 and 1.97 points on LifeSnaps and CrossCheck. $\text{PRISM}_{\text{gen}}$ exceeds this direct predictor by 8.47–14.27 points across datasets. Thus, the pretrained language prior by itself is insufficient to explain PRISM's results; the learned multimodal architecture, training objectives, and generative inference path contribute beyond direct prompting.

Table 2: OOD accuracy (%) of a frozen LLaMA-3 8B direct-prediction control and $\text{PRISM}_{\text{gen}}$. The frozen model predicts zero-shot from PRISM's meta-narratives without soft prompts, fusion, or task fine-tuning.

| Dataset | Frozen LLaMA-3 | $\text{PRISM}_{\text{gen}}$ | Gain (pp) |
|---|---|---|---|
| GLOBEM | 65.66 | 79.93 | 14.27 |
| LifeSnaps | 68.87 | 81.25 | 12.38 |
| MFAFY | 65.24 | 73.71 | 8.47 |
| CrossCheck | 74.96 | 84.97 | 10.01 |

**The generative path outperforms the prediction path on every dataset.** $\text{PRISM}_{\text{gen}}$ exceeds $\text{PRISM}_{\text{pred}}$ by 12.85 points on GLOBEM, 12.50 on LifeSnaps, 8.00 on MFAFY, and 15.56 on CrossCheck. Both paths share the fused representation $\mathbf{F}_{\text{fused}}$; the difference is in how each path uses it. Autoregressive narrative generation through the frozen decoder accesses structure in $\mathbf{F}_{\text{fused}}$ that a single MLP forward pass cannot. The generation path provides an inference-time benefit beyond its role as a training-time regularizer.

**Domain-agnostic multimodal methods trail by large margins.** Time-VLM achieves 51.43% on GLOBEM, 47.36% on LifeSnaps, 48.27% on MFAFY, and 65.47% on CrossCheck. Its FFT-based visual encoding was designed for stationary periodic signals; behavioral trajectories are non-stationary. The first three datasets show the same pattern, consistent with this design mismatch. Time-VLM's CrossCheck accuracy is higher but still trails $\text{PRISM}_{\text{gen}}$ by 19.50 points. Even when the FFT representation captures partial signal, the absence of joint time-frequency localization and a coherence constraint both limit cross-distribution performance.

**Fine-tuning a VLM produces strong OOD accuracy but widens the generalization gap.** The Bimodal VLM baseline, which fine-tunes LLaVA-NeXT 7B with LoRA, achieves the strongest OOD accuracy among baselines on every dataset (72.86% on GLOBEM, 71.88% on LifeSnaps, 66.57% on MFAFY, 79.61% on CrossCheck). $\text{PRISM}_{\text{gen}}$'s ID-to-OOD gap is smaller on every dataset (Table 3), with reduction ratios of 5.25× on GLOBEM, 1.68× on LifeSnaps, 5.95× on MFAFY, and 1.49× on CrossCheck. This comparison isolates the effect of fine-tuning versus freezing under matched modalities. The Bimodal VLM fine-tunes its foundation model on the training distribution; PRISM holds its foundation models frozen and applies the coherence constraint through the generation loss. The wider gap under fine-tuning is consistent with documented modality suppression in VLM fine-tuning (Sim et al., 2025; Peng et al., 2022), while the narrower gap under freezing supports the representational-constraint framing.

Table 3: ID-to-OOD accuracy gap (percentage points). Smaller gap indicates more distribution-invariant representations.

| Method | GLOBEM | LifeSnaps | MFAFY | CrossCheck |
|---|---|---|---|---|
| Bimodal VLM | 6.51 | 10.47 | 10.48 | 11.62 |
| $\text{PRISM}_{\text{gen}}$ | **1.24** | **6.25** | **1.76** | **7.81** |

**Efficiency.** PRISM trains in under three hours on a single NVIDIA A40 GPU (48GB) with approximately 18M trainable parameters, compared to over a day on 8×A40 GPUs with substantially more trainable parameters for the Bimodal VLM baseline. The frozen-backbone design eliminates gradient computation through the vision and language foundation models, which accounts for most of the difference. Inference cost differs by path: $\text{PRISM}_{\text{gen}}$ is substantially slower and uses more memory

than $\text{PRISM}_{\text{pred}}$ because it requires autoregressive decoding. Under batch-size-one evaluation on a single A40, $\text{PRISM}_{\text{gen}}$ remains faster and uses less peak memory than the fine-tuned Bimodal VLM baseline (Table 4).

Table 4: Inference latency and peak memory at batch size 1 on one NVIDIA A40 GPU.

| **Method** | **Latency (ms/sample)** | **Peak memory (GB)** |
|---|---|---|
| $\text{PRISM}_{\text{pred}}$ | 8 | 4.1 |
| $\text{PRISM}_{\text{gen}}$ | 340 | 9.8 |
| Bimodal VLM | 410 | 18.6 |

### 4.4 Ablations

The central ablation varies the training objective while holding the architecture fixed, testing whether the dual-loss with a frozen coherence constraint is the mechanism responsible for distribution invariance. Appendix A.6 provides full per-metric breakdowns.

#### 4.4.1 The Dual-Path Training Objective

Table 5: Dual-path ablation. OOD accuracy (%) and ID-to-OOD gap under different training objectives. *Pred* and *Gen* denote the prediction and generative inference paths. The gap is computed for the trained path; for joint objectives, it is reported for the generation path. Blue values indicate an output head that receives no direct training signal; these are reported for setup verification, not performance comparison.

| | | **GLOBEM** | | | **LifeSnaps** | | | **MFAFY** | | | **CrossCheck** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Objective** | Pred | Gen | Gap | Pred | Gen | Gap | Pred | Gen | Gap | Pred | Gen | Gap |
| (a) | $\mathcal{L}_{\text{pred}}$ only | 65.10 | 59.12 | 7.10 | 64.06 | 64.06 | 10.94 | 64.57 | 55.71 | 7.13 | 66.91 | 61.90 | 10.93 |
| (b) | $\mathcal{L}_{\text{gen}}$ only | 55.86 | 71.33 | 3.11 | 64.06 | 70.31 | 4.69 | 52.29 | 68.29 | 5.29 | 62.97 | 71.02 | 13.51 |
| (c) | Fixed $\alpha = 0.5$ | 68.26 | 75.73 | 1.85 | 70.31 | 76.56 | 4.69 | 67.43 | 71.71 | 1.87 | 66.19 | 78.53 | 12.34 |
| (d) | Learned (Eq. 9) | 67.08 | **79.93** | **1.24** | 68.75 | **81.25** | **6.25** | 65.71 | **73.71** | **1.76** | 69.41 | **84.97** | **7.81** |

**Bidirectional regularization between the two losses.** Comparing rows (b) and (d), removing the prediction loss from the generation path drops OOD accuracy by 8.60 points on GLOBEM (79.93%→71.33%), 10.94 on LifeSnaps, 5.42 on MFAFY, and 13.95 on CrossCheck. The generation loss alone yields a representation too diffuse for strong generalization. Comparing rows (a) and (d), removing the generation loss from the prediction path drops OOD accuracy by smaller but consistent amounts (1.98 points on GLOBEM, 4.69 on LifeSnaps, 1.14 on MFAFY, 2.50 on CrossCheck). The two losses regularize the shared representation in complementary ways; removing either degrades the path the other supervises. The asymmetry (5–14 points vs 1–5 points) is consistent with the information-theoretic difference between a categorical supervision signal and dense token-level supervision from a frozen language prior.

**The full configuration achieves the strongest cross-distribution generalization.** Across all four datasets, the full configuration (row d) reaches the highest OOD accuracy and, on three datasets (GLOBEM, MFAFY, CrossCheck), the smallest ID-to-OOD gap. Prediction-only training achieves reasonable ID accuracy but generalizes poorly, suggesting reliance on distribution-specific classification shortcuts. Generation-only training encourages narrative coherence but lacks sufficient discriminative supervision to support accurate outcome forecasting. The full configuration combines both objectives under a fixed coherence constraint, delivering higher OOD accuracy while maintaining competitive ID-to-OOD gaps. Neither objective alone achieves this combination.

**Learned weighting reflects the generation signal's informational advantage.** The converged $s_1, s_2$ values correspond to effective weights that favor the generation loss by factors of 2.0× on MFAFY, 6.2× on GLOBEM, 6.8× on CrossCheck, and 6.9× on LifeSnaps. The three sensor-rich datasets (GLOBEM, LifeSnaps, CrossCheck) cluster between 6.2× and 6.9×. The categorical-only MFAFY sits much lower at 2.0×, consistent with the generation path drawing more independent signal from sensor-based trajectories than from categorical self-reports. Appendix A.6 provides the full $s_1, s_2$ values.

### 4.4.2 Modality Contribution

Each stream contributes to OOD accuracy across all four datasets (Table 6). The contribution ordering is consistent. Removing the semantic stream causes the largest drop (12–15 points), removing the spectral stream causes the second largest (9–11 points), and removing the temporal stream causes the smallest (6–9 points). The semantic stream's dominance and the temporal stream's smaller contribution align with the frozen-backbone principle. Pre-trained priors hold more transferable information than representations learned from scratch on small behavioral datasets. The spectral stream's non-trivial contribution on MFAFY, despite its categorical features, indicates that CWT captures structure the temporal stream does not encode as effectively.

Table 6: Modality removal. OOD accuracy (%) on the generative path with each stream removed; $\Delta$ is the drop from the full system.

| | **GLOBEM** | | **LifeSnaps** | | **MFAFY** | | **CrossCheck** | |
|---|---|---|---|---|---|---|---|---|
| **Configuration** | Acc | $\Delta$ | Acc | $\Delta$ | Acc | $\Delta$ | Acc | $\Delta$ |
| Full system | 79.93 | — | 81.25 | — | 73.71 | — | 84.97 | — |
| w/o Text | 65.62 | −14.31 | 68.75 | −12.50 | 61.14 | −12.57 | 69.77 | −15.20 |
| w/o Vision | 68.81 | −11.12 | 71.88 | −9.37 | 64.29 | −9.42 | 74.24 | −10.73 |
| w/o Temporal | 71.13 | −8.80 | 73.44 | −7.81 | 67.71 | −6.00 | 78.71 | −6.26 |

## 5 Discussion

The Bimodal VLM baseline fine-tunes foundation models on the same underlying data that PRISM processes using frozen backbones, but exhibits consistently larger ID-to-OOD gaps across all four datasets (Table 3). The dual-path ablation supports the role of the generation loss within the frozen-backbone system. Configurations without the generation loss exhibit larger gaps, and neither objective alone reaches the same generalization performance. On CrossCheck, where prior cross-study generalization work relied on feature-distribution alignment, PRISM achieves 84.97% OOD accuracy without combining data across studies or explicitly aligning feature distributions.

The representational-constraint principle may extend to other domains where learned representations adapt to training-distribution specificities in ways that impair cross-cohort transfer. This requires that the underlying data can be meaningfully expressed in language, so that a frozen language model can assess the coherence of representations through narrative generation. Domains with stable category structures, such as object recognition with a fixed class vocabulary, or primarily numerical tasks without meaningful linguistic descriptions, such as stationary time-series forecasting, may benefit less from this approach. In these settings, a language-based coherence constraint offers limited additional supervision.

Behavioral health prediction presents deployment risks, particularly when strong cross-distribution performance encourages applications in screening, hiring, or decisions about access to services. The small ID-to-OOD gaps observed in this research establish generalization only across the evaluated cohorts and should not be interpreted as evidence of readiness for untested populations. Clinical and educational deployment requires prospective validation in the intended population, regardless of performance on existing cross-distribution benchmarks.

**Limitations.** The semantic stream relies on GPT-4o-generated meta-narratives, introducing reproducibility and computational cost concerns. Ablations substituting Llama-3-8B-Instruct achieve over 95% BERTScore similarity with GPT-4o narratives and OOD accuracy within 1–2 points across datasets, and text-only baselines with maximal detail do not outperform the template-constrained meta-narratives. These findings suggest that performance is not solely dependent on the choice of narrative generator (Appendix A.7). The experiments are limited to binary classification; extending the prediction head and training objectives to multi-class, ordinal, and continuous outcomes remains untested. Results are reported from single training runs with fixed random seeds, applied consistently across methods for fair comparison. Multi-seed evaluation was computationally prohibitive for the fine-tuned VLM baselines, which required over a day on eight GPUs per run. The main comparisons show large effect sizes, though variance estimates would strengthen comparisons with smaller margins or limited OOD samples, especially LifeSnaps, which contains only 64 OOD samples. The directed cross-attention fusion is architecturally motivated but has not been compared against symmetric alternatives. The contribution of asymmetric fusion remains to be established experimentally.

## Acknowledgments

This research was supported by a standard grant from the U.S. National Science Foundation (NSF DUE 2142558).

## References

Adler, D. A., Wang, F., Mohr, D. C., and Choudhury, T. (2022). Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE*, 17(4):e0266516.

Canzian, L. and Musolesi, M. (2015). Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 1293–1304, New York, NY, USA. Association for Computing Machinery.

Chang, C., Peng, W.-C., and Chen, T.-F. (2023). Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms.

Chikersal, P., Doryab, A., Tumminia, M., Villalba, D. K., Dutcher, J. M., Liu, X., Cohen, S., Creswell, K. G., Mankoff, J., Creswell, J. D., Goel, M., and Dey, A. K. (2021). Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection. *ACM Trans. Comput.-Hum. Interact.*, 28(1):3:1–3:41.

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., Kamath, J., Russell, A., Bamis, A., and Wang, B. (2016). Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *2016 IEEE Wireless Health (WH)*, pages 1–8.

Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria,

A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. (2024). The llama 3 herd of models.

Grossmann, A. and Morlet, J. (1984). Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis*, 15(4):723–736.

Hayat, A., Khan, B., and Hasan, M. (2024a). Improving transfer learning for early forecasting of academic performance by contextualizing language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 137–148, Mexico City, Mexico. Association for Computational Linguistics.

Hayat, A., Khan, B., and Hasan, M. R. (2024b). Leveraging language models for analyzing longitudinal experiential data in education. *arXiv:2503.21617*.

Hayat, A., Khan, B., and Hasan, M. R. (2025). ConText-LE: Cross-distribution generalization for longitudinal experiential data via narrative-based LLM representations. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15335–15360, Suzhou, China. Association for Computational Linguistics.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. (2023). Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

Herzig, J., Nowak, P. K., Müller, T., Piccinno, F., and Eisenschlos, J. M. (2020). Tapas: Weakly supervised table parsing via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs].

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., and Wen, Q. (2024). Time-llm: Time series forecasting by reprogramming large language models.

Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics.

Kim, Y., Xu, X., McDuff, D., Breazeal, C., and Park, H. W. (2024). Health-llm: Large language models for health prediction via wearable sensor data.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2023). Large language models are zero-shot reasoners.

Li, X., Zhu, X., Zhu, X., Ji, Y., and Tang, X. (2020). Student academic performance prediction using deep multi-source behavior sequential network. In Lauw, H. W., Wong, R. W., Ntoulas, A., Lim, E.-P., Ng, S.-K., and Pan, S., editors, *Advances in Knowledge Discovery and Data Mining. PAKDD 2020*, volume 12084 of *Lecture Notes in Computer Science*, pages 570–582. Springer, Cham.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. (2024a). Llava-next: Improved reasoning, ocr, and world knowledge.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024b). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., and Long, M. (2024c). itransformer: Inverted transformers are effective for time series forecasting.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Academic Press, 2nd edition.

Nemati, E., Xu, X., Nathan, V., Vatanparvar, K., Ahmed, T., Rahman, M. M., McCaffrey, D., Kuang, J., and Gao, A. (2022). Ubilung: Multi-modal passive-based lung health assessment. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 551–555. IEEE.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. *ArXiv*, abs/2211.14730.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Pang, Z., Xie, Z., Man, Y., and Wang, Y.-X. (2024). Frozen transformers in language models are effective visual encoder layers.

Patharkar, A., Cai, F., Al-Hindawi, F., and Wu, T. (2024). Predictive modeling of biomedical temporal data in healthcare applications: review and future directions. *Frontiers in Physiology*, 15:1386760.

Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. (2022). Balanced Multimodal Learning via On-the-fly Gradient Modulation . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8237, Los Alamitos, CA, USA. IEEE Computer Society.

Rabbi, M., Klasnja, P., Choudhury, T., Tewari, A., and Murphy, S. (2019). Optimizing mhealth interventions with a bandit. In Baumeister, H. and Montag, C., editors, *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*, Studies in Neuroscience, Psychology and Behavioral Economics, pages 277–291. Springer International Publishing, Cham.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Raissi, M., Perdikaris, P., and Karniadakis, G. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., and Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, 17(7):e175.

Salles, R. S. and Ribeiro, P. F. (2023). The use of deep learning and 2-d wavelet scalograms for power quality disturbances classification. *Electric Power Systems Research*, 214:108834.

Sano, A., Taylor, S., McHill, A. W., Phillips, A. J. K., Barger, L. K., Klerman, E., and Picard, R. (2018). Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study. *Journal of Medical Internet Research*, 20(6):e210.

Siddique, M. F., Saleem, F., Umar, M., Kim, C.-H., and Kim, J.-M. (2025). A hybrid deep learning approach for bearing fault diagnosis using continuous wavelet transform and attention-enhanced spatiotemporal feature extraction. *Sensors*, 25(9):2712.

Sim, M. Y., Zhang, W. E., Dai, X., and Fang, B. (2025). Can VLMs actually see and read? a survey on modality collapse in vision-language models. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, Vienna, Austria. Association for Computational Linguistics.

Thach, N. T., Habecker, P., Eisenbraun, A. R., Mason, W. A., Tyler, K. A., Khan, B., and Chan, H. (2025). Muhboost: Multi-label boosting for practical longitudinal human behavior modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Accepted. Available at `https://openreview.net/pdf?id=BAelAyADqn`.

Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method.

Tishby, N. and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5.

Torrence, C. and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61 – 78.

Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S. M. A., Vinyals, O., and Hill, F. (2021). Multimodal few-shot learning with frozen language models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. arXiv:1706.03762 [cs].

Wahle, F., Kowatsch, T., Fleisch, E., Rufer, M., and Weidt, S. (2016). Mobile sensing and support for people with depression: A pilot trial in the wild. *JMIR mHealth and uHealth*, 4(3):e111.

Wang, R., Aung, M. S. H., Abdullah, S., Brian, R., Campbell, A. T., Choudhury, T., Hauser, M., Kane, J., Merrill, M., Scherer, E. A., Tseng, V. W. S., and Ben-Zeev, D. (2016a). Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, page 886–897, New York, NY, USA. Association for Computing Machinery.

Wang, R., Hao, P., Zhou, X., Campbell, A. T., and Harari, G. (2016b). SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students. *GetMobile: Mobile Computing and Communications*, 19(4):13–17.

Wang, R., Wang, W., daSilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., and Campbell, A. T. (2018). Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):43:1–43:26.

Wang, T., Lu, C., Sun, Y., Yang, M., Liu, C., and Ou, C. (2021). Automatic ecg classification using continuous wavelet transform and convolutional neural network. *Entropy*, 23(1):119.

Wei, H., Wu, J., Yang, J., Leung, H., and Drew, S. (2025). Spear: Soft prompt enhanced anomaly recognition for time series data.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs].

Xu, X., Chikersal, P., Dutcher, J. M., Sefidgar, Y. S., Seo, W., Tumminia, M. J., Villalba, D. K., Cohen, S., Creswell, K. G., Creswell, J. D., Doryab, A., Nurius, P. S., Riskin, E., Dey, A. K., and Mankoff, J. (2021a). Leveraging collaborative-filtering for personalized behavior modeling: A case study of depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):41:1–41:27.

Xu, X., Liu, X., Zhang, H., Wang, W., Nepal, S., Sefidgar, Y., Seo, W., Kuehn, K. S., Huckins, J. F., Morris, M. E., Nurius, P. S., Riskin, E. A., Patel, S., Althoff, T., Campbell, A., Dey, A. K., and Mankoff, J. (2023). GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):190:1–190:34.

Xu, X., Nemati, E., Vatanparvar, K., Nathan, V., Ahmed, T., Rahman, M. M., McCaffrey, D., Kuang, J., and Gao, J. A. (2021b). Listen2cough: Leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(1).

Yfantidou, S., Karagianni, C., Efstathiou, S., et al. (2022). Lifesnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild. *Scientific Data*, 9(1):663.

Zhong, S., Ruan, W., Jin, M., Li, H., Wen, Q., and Liang, Y. (2025). Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–20.

Zhou, T., Niu, P., Wang, X., Sun, L., and Jin, R. (2023). One Fits All:Power General Time Series Analysis by Pretrained LM. arXiv:2302.11939 [cs].

# A Appendix

## A.1 Detailed Methodology: Architecture and Training Specifications

This subsection specifies the architectural and training details referenced in Section 3.

### A.1.1 Temporal Stream Architecture

Each trajectory $\mathbf{X}_i \in \mathbb{R}^{T \times d}$ is projected into a latent space through a learned linear transformation $\mathbf{H}^{(0)} = \mathbf{X}_i \mathbf{W}_{\text{proj}} + \mathbf{b}_{\text{proj}}$ with $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d \times d_{\text{model}}}$, where $d_{\text{model}} = 256$. Sinusoidal positional encodings $\mathbf{P} \in \mathbb{R}^{T \times d_{\text{model}}}$ are added to preserve temporal order. The resulting sequence is processed through $L = 2$ transformer encoder layers (Vaswani et al., 2017) with multi-head self-attention (4 heads), position-wise feed-forward networks, residual connections, and layer normalization.

### A.1.2 Spectral Stream Processing

The CWT is defined as

$$W(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\, \psi^* \left( \frac{t - b}{a} \right) dt, \tag{2}$$

where $a > 0$ is the scale parameter (inversely related to frequency), $b$ is the translation parameter, and $\psi$ is the Morlet wavelet, a complex sinusoid modulated by a Gaussian envelope. The resulting scalogram $|W(a, b)|^2 \in \mathbb{R}^{S \times T}$, with $S$ scales, represents the trajectory as a two-dimensional surface over time and frequency. For each trajectory, the $d$ features are concatenated along the time axis to form a univariate signal of length $T \cdot d$, and the CWT is computed over logarithmically spaced scales from 2 to $T/2$, covering daily through multi-week periodicity. The univariate reduction lets the spectral stream capture temporal-frequency patterns across the full trajectory. Per-feature multivariate structure remains in the temporal stream. The scalogram is normalized to $[0, 255]$ via min-max scaling. Trainable multi-scale convolutional layers then transform the normalized scalogram into a $3 \times 224 \times 224$ RGB image formatted for the vision encoder's expected input. The frozen CLIP vision encoder (Radford et al., 2021) produces the patch token sequence $\mathbf{Z}_{\text{vision}} \in \mathbb{R}^{P \times d_{\text{clip}}}$, with $P$ spatial patch tokens and $d_{\text{clip}} = 512$.

### A.1.3 Semantic Stream Adaptation

The meta-narrative $\mathbf{M}_i$ is tokenized and embedded through a standard tokenizer, then passed through a trainable injection MLP (two-layer, ReLU activation, input and output dimension $d_{\text{clip}}$) before the frozen CLIP text encoder. The injection MLP provides input-side adaptation that shifts the embedding toward the behavioral domain without modifying the encoder's internal weights. The frozen text encoder produces the token sequence $\mathbf{Z}_{\text{text}} \in \mathbb{R}^{L \times d_{\text{clip}}}$, where $L$ is the token count.

### A.1.4 Fusion Equations

The CLIP token sequences are first projected to match the temporal dimension through learned per-token linear projections $(\cdot)^{\downarrow} : \mathbb{R}^{d_{\text{clip}}} \to \mathbb{R}^{d_{\text{model}}}$. Three directed multi-head cross-attention blocks compute

$$\mathbf{H}_{S \leftarrow V} = \text{MHA}(\mathbf{Z}^{\downarrow}_{\text{text}}, \mathbf{Z}^{\downarrow}_{\text{vision}}, \mathbf{Z}^{\downarrow}_{\text{vision}}), \tag{3}$$

$$\mathbf{H}_{S \leftarrow N} = \text{MHA}(\mathbf{Z}^{\downarrow}_{\text{text}}, \mathbf{Z}_{\text{temporal}}, \mathbf{Z}_{\text{temporal}}), \tag{4}$$

$$\mathbf{H}_{N \leftarrow V} = \text{MHA}(\mathbf{Z}_{\text{temporal}}, \mathbf{Z}^{\downarrow}_{\text{vision}}, \mathbf{Z}^{\downarrow}_{\text{vision}}), \tag{5}$$

where $\text{MHA}(Q, K, V)$ denotes multi-head attention (Vaswani et al., 2017) with $h = 4$ heads and head dimension $d_k = d_{\text{model}}/h = 64$. Each enriched sequence is mean-pooled along its token dimension to produce vectors $\mathbf{E}_{S \leftarrow V}, \mathbf{E}_{S \leftarrow N}, \mathbf{E}_{N \leftarrow V} \in \mathbb{R}^{d_{\text{model}}}$.

The three pooled vectors are combined through the gate equation given in Eq. 1 of the main paper. Token-level cross-attention before pooling allows each query position to attend to the full key and value sequence; the mean pool then compresses the enriched sequence into a compact vector that retains the retrieved cross-modal information.

### A.1.5 Training Loss Specifications

The frozen LLaMA-3 8B decoder (Grattafiori et al., 2024) is loaded in 8-bit quantization to reduce memory footprint during training.

**Prediction loss.** The prediction head outputs a single logit $\hat{y}_i$ for binary targets, and the loss is

$$\mathcal{L}_{\text{pred}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \sigma(\hat{y}_i) + (1-y_i)\log(1-\sigma(\hat{y}_i))\right]. \tag{6}$$

**Soft prompt projection and generation loss.** The fused embedding is projected into $k$ soft prompt tokens in the decoder's embedding space:

$$\mathbf{P}_{\text{soft}} = \text{MLP}_{\text{prompt}}(\mathbf{F}_{\text{fused}}) \in \mathbb{R}^{k \times d_{\text{llm}}}, \tag{7}$$

with $k = 32$ and $d_{\text{llm}} = 4096$. The generation loss is next-token cross-entropy against a target prospective narrative $\mathbf{M}_i^{\text{target}}$ generated offline:

$$\mathcal{L}_{\text{gen}} = -\frac{1}{|\mathbf{M}_i^{\text{target}}|}\sum_{j=1}^{|\mathbf{M}_i^{\text{target}}|}\log p(m_j \mid m_{<j}, \mathbf{P}_{\text{soft}}, \mathbf{T}_{\text{template}}). \tag{8}$$

**Combined loss.** The total loss combines prediction and generation losses through the homoscedastic task uncertainty framework (Kendall et al., 2018),

$$\mathcal{L}_{\text{total}} = e^{-s_1}\mathcal{L}_{\text{pred}} + s_1 + e^{-s_2}\mathcal{L}_{\text{gen}} + s_2, \tag{9}$$

where $s_1, s_2$ are learnable log-variance parameters. The $e^{-s}$ factor inverse-weights each loss by its estimated noise; the linear $+s$ term prevents the trivial solution of pushing $s \to \infty$ to ignore a task.

## A.2 Dataset Details

This subsection describes each evaluation dataset in detail, covering data collection, features, outcome labels, the cross-distribution split, and preprocessing. Table 7 reports summary statistics.

Table 7: Dataset summary. $|\mathcal{P}|$ and $|\mathcal{P}'|$ denote participants in the training and OOD test distributions. $|\mathcal{D}_T|$ and $|\mathcal{D}_{T'}|$ denote sample counts.

| Dataset | Domain | $\lvert\mathcal{P}\rvert$ | $\lvert\mathcal{P}'\rvert$ | $\lvert\mathcal{D}_T\rvert$ | $\lvert\mathcal{D}_{T'}\rvert$ | Feat. | OOD Shift |
|---|---|---|---|---|---|---|---|
| GLOBEM | Mental health | 344 | 317 | 2226 | 2023 | 15 | Cross-institution/year |
| LifeSnaps | Anxiety | 26 | 13 | 112 | 64 | 8 | Cross-participant/temporal |
| MFAFY | Education | 61 | 35 | 610 | 350 | 11 | Cross-year |
| CrossCheck | Mental health | 36 | 22 | 970 | 559 | 18 | Cross-cohort temporal |

**Shared Evaluation Protocol.** Train/test splits separate data from different cohorts or time periods. We reserve 15% of the training-period ($T$) data as an in-distribution (ID) test set; all data from the distinct period ($T'$) form the OOD test set. Preprocessing (feature scaling, mean imputation for missing values) is applied consistently across datasets before generating CWT scalograms and meta-narratives.

### A.2.1 GLOBEM

GLOBEM (Xu et al., 2023) is a benchmark for longitudinal human behavior modeling and cross-distribution generalization. The dataset contains data from 497 participants across two institutions over four academic years (Years 1 & 2 from Institution A; Years 3 & 4 from Institution B), totaling 661 person-years after preprocessing. Institutions A (pre-COVID) and B (post-COVID) represent distinct cohorts and time periods. Surveys include PHQ-4, BDI-II, and PANAS for depression assessment.

**Features.** We use 15 features based on prior work (Xu et al., 2023; Thach et al., 2025; Kim et al., 2024), derived from mobile sensing data. The categories are Location (variance, entropy, travel distance, duration of stay), Phone Usage (unlock counts, usage statistics), Bluetooth (scan counts, unique devices), Call (duration statistics, missed call count), Physical Activity (steps, active/sedentary duration), and Sleep (duration, episode statistics).

**Task and Sequence Construction.** The task is binary mental health prediction based on a threshold applied to depression survey scores, resulting in a nearly balanced distribution. Each person-year of data represents a 10-week observation period from which 6 sequences are generated using a 4-week sliding window predicting the subsequent week.

**OOD Split.** Data from Years 1 & 2 at Institution A form the training distribution ($|\mathcal{P}| = 344$ person-years, $|\mathcal{D}_T| \approx 2226$ sequences). Data from Years 3 & 4 at Institution B form the OOD test distribution ($|\mathcal{P}'| = 317$ person-years, $|\mathcal{D}_{T'}| \approx 2023$ sequences). This split evaluates cross-institution and cross-temporal generalization simultaneously, as the training and test data come from different institutions collected before and after the COVID-19 pandemic, respectively. The prior published best cross-distribution accuracy on this benchmark is 52.8% $\pm$ 0.024 (Xu et al., 2023).

### A.2.2 LifeSnaps

LifeSnaps (Yfantidou et al., 2022) is a longitudinal dataset collected from 71 participants over 4 months. Data sources include Fitbit sensing data (activity, sleep, stress, heart rate), ecological momentary assessments (mood, context), and validated psychological surveys. The dataset includes over 35 distinct data types.

**Features.** We use 8 features from the Fitbit sensing and EMA modalities, aligned with the feature dimensionality of the other datasets. Features include daily-aggregated heart rate statistics, step count, sleep duration and quality metrics, activity minutes, and self-reported mood and context indicators.

**Task and Sequence Construction.** We formulate a binary anxiety level forecasting task, predicting anxiety in the week subsequent to a $k = 1$ week observation window. After preprocessing, including filtering participants with significant missing values, a subset of participants is retained for the evaluation splits.

**OOD Split.** Training uses data from 26 participants collected during the first 2 months of the study period. Testing uses data from 13 disjoint participants collected during the last 2 months, assessing cross-temporal and cross-participant generalization within the study cohort. Using a $k = 1$ week window over these approximately 8-week periods yields a training set of approximately 112 sequences and a test set of approximately 64 sequences. No published OOD baseline exists for this task. This dataset validates cross-distribution generalization using a different dataset structure, population, and data collection protocol from GLOBEM.

### A.2.3 MFAFY

The Messages From A Future You (MFAFY) dataset (Hayat et al., 2024a,b) tracks first-year college students' academic experience across three consecutive semesters spanning two academic years (Year 1, Semesters 1 & 2; Year 2, Semester 3). Data were collected through a smartphone-based ecological momentary assessment platform. The dataset is high-dimensional, with non-cognitive factors (28 dimensions covering motivation, engagement, science identity), cognitive factors (41 dimensions covering assessment scores), and background factors (9 dimensions covering static academic meta-information).

**Features.** For forecasting behavioral engagement, input features use only the relevant non-cognitive dimensions ($d = 11$), capturing self-reported motivation (intrinsic and extrinsic), engagement (behavioral, emotional, and cognitive), and course-related attitudes.

**Task and Sequence Construction.** The task is predicting a student's lecture-related engagement status (binary, high/low) in the subsequent week, using a $k = 4$ week observation window. The binary target is derived by comparing the average of relevant non-cognitive features during weeks

$s$ through $s + k - 1$ with their average during week $s + k$, resulting in a nearly balanced binary distribution. Each subject's data represents a 15-week observation period from which 10 sequences are generated using the 4-week sliding window.

**OOD Split.** Year 1 data (Semesters 1 & 2) form the training distribution ($|\mathcal{P}| = 61$ subjects, $|\mathcal{D}_T| \approx 610$ sequences). Year 2 data (Semester 3) form the OOD test distribution ($|\mathcal{P}'| = 35$ subjects, $|\mathcal{D}_{T'}| \approx 350$ sequences). This split evaluates cross-year generalization.

**Feature Characteristics.** MFAFY differs from GLOBEM and LifeSnaps in that its features are purely categorical self-reports with no continuous sensor data. There are no heart rate, step count, or sleep-stage features. This makes MFAFY an informative test case for PRISM's spectral stream. CWT scalograms on categorical responses lack the continuous temporal dynamics that wavelet analysis is designed to capture; the modality ablation in Section 4.4.2 reports the actual spectral contribution on this dataset.

#### A.2.4 CrossCheck

The CrossCheck dataset (Wang et al., 2016a) is a longitudinal smartphone passive sensing study of individuals with schizophrenia spectrum disorders, conducted between 2015 and 2017 at a psychiatric hospital in the northeast U.S. (Adler et al., 2022). The study captures passive sensor streams (activity, audio, calls, mobility, phone usage, sleep) and ecological momentary assessment (EMA) self-reports of mental health symptoms.

**Features.** We use 18 daily-aggregated behavioral features spanning activity (on-foot, still, in-vehicle), social interaction (audio conversation duration and count, voice activity, ambient amplitude), communication (incoming and outgoing call counts, call duration, incoming and outgoing SMS counts), mobility (location distance, unique location visits), phone usage (unlock count and duration), environment (ambient light), and sleep (duration). All features use the full-day summary in the public CSV. The 18-feature count is comparable to the other datasets (GLOBEM, LifeSnaps, MFAFY).

**Task and Sequence Construction.** The task is binary mental health risk forecasting. Daily data are transformed into 28-day windows with a 7-day stride and a 7-day prediction horizon. Labels are derived from ecological momentary assessment (EMA) self-reports, using the composite EMA score (sum of positive items minus sum of negative items, range $-15$ to $+15$) averaged over the 7-day prediction horizon. Windows with mean EMA score $\leq 2$ are labeled at-risk; windows with mean EMA score $> 2$ are labeled stable. Higher EMA scores indicate better self-reported mental state. The clinician-rated BPRS scores used in some prior CrossCheck analyses are not in the public release; the EMA-derived label follows prior cross-study generalization work on this dataset (Adler et al., 2022).

**OOD Split.** We use a temporal cohort split within CrossCheck. Participants who enrolled in 2015 form the training cohort ($|\mathcal{P}| = 36$ participants, $|\mathcal{D}_T| = 970$ sequences with 29.6% at-risk). Participants who enrolled in 2016 and 2017 form the OOD test cohort ($|\mathcal{P}'| = 22$ participants, $|\mathcal{D}_{T'}| = 559$ sequences with 11.8% at-risk). The temporal cohort split simulates deployment to populations recruited under different study conditions and possible changes in clinical practice. The class-balance shift across cohorts (29.6% to 11.8% at-risk) is itself a feature of the cross-distribution challenge.

**Distinction from Other Datasets.** CrossCheck differs from the other three datasets in two respects. First, the clinical population (schizophrenia spectrum disorders) is more specialized than the university-student (GLOBEM, MFAFY) or general-adult (LifeSnaps) populations. Second, the same dataset has been used in prior cross-study generalization work (Adler et al., 2022) that combined CrossCheck with StudentLife and aligned feature distributions; we use a temporal cohort split within CrossCheck alone, evaluating PRISM's representational-constraint mechanism on the same underlying data through a different cross-distribution protocol.

### A.3 Extended Related Work

This subsection provides extended discussion of the literatures PRISM engages with.

### A.3.1 Longitudinal Behavioral Data Modeling

Modeling longitudinal behavioral data from smartphones and wearable devices has applications in mental health monitoring (Saeb et al., 2015; Wang et al., 2018; Chikersal et al., 2021; Xu et al., 2021a; Wahle et al., 2016; Farhan et al., 2016; Canzian and Musolesi, 2015), physical health (Nemati et al., 2022; Rabbi et al., 2019; Xu et al., 2021b), and educational engagement (Wang et al., 2016b; Li et al., 2020). Traditional approaches use handcrafted features from sensor data and train classifiers such as SVMs, random forests, or shallow neural networks. These methods prioritize in-distribution performance and have not generalized well across cohorts, a limitation documented systematically by the GLOBEM benchmark (Xu et al., 2023).

Several LLM-based approaches to longitudinal behavioral forecasting include prompted and fine-tuned LLMs for health prediction from wearable data (Kim et al., 2024), multi-label boosting methods leveraging zero-shot LLMs (Thach et al., 2025), and meta-narrative textualization of behavioral trajectories that frames prediction as prospective narrative generation (Hayat et al., 2025). LLM-based forecasting has also been applied to educational engagement using the MFAFY dataset (Hayat et al., 2024a,b). These approaches improve generalization over purely numerical methods but share a limitation, serializing the data's two-dimensional *Features* $\times$ *Time* structure into one-dimensional token sequences, which loses spatial co-occurrence structure (Hegselmann et al., 2023; Herzig et al., 2020; Liu et al., 2024b).

### A.3.2 Cross-Study Generalization

Research on cross-study generalization combines data from multiple longitudinal sensing studies (Adler et al., 2022). This approach measures Proxy-A Distance between training and target feature distributions and finds that combined training improves generalization only when that distance decreases, with feature-nearest-neighbor personalization providing limited additional benefit. The finding locates the limitation at the feature level, suggesting that the behavior-to-outcome relationship varies across populations in ways distribution-matching cannot fully absorb. PRISM operates at a different level. The coherence constraint through a frozen language prior does not depend on feature-distribution alignment; it holds the representation to a criterion defined by broad pre-training, which is invariant across behavioral data distributions.

### A.3.3 Cross-Distribution Generalization Theory

Domain adaptation (Pan and Yang, 2010) and domain generalization (Zhou et al., 2022) aim to improve model performance on target distributions different from the training distribution through techniques including invariant representation learning, meta-learning, and data augmentation. Success on longitudinal human behavioral data has been limited (Xu et al., 2023). The varied shifts that occur across human cohorts, including demographic differences, temporal context changes, and institutional culture, pose challenges beyond typical domain shifts in image classification. Under the information bottleneck view (Tishby et al., 2000; Tishby and Zaslavsky, 2015), representations trained for narrow classification objectives may converge to minimal sufficient statistics tied to the training distribution. PRISM introduces a second criterion, narrative coherence from a frozen LLM, that the representation has to satisfy in addition to outcome sufficiency.

### A.3.4 Foundation Models for Time Series

Foundation models for time series include frozen pre-trained transformers as general-purpose time series analyzers (Zhou et al., 2023) and time-series-to-text-prototype reprogramming for frozen LLM backbones (Jin et al., 2024). Two-stage fine-tuning approaches for adapting LLMs to time series have also been proposed (Chang et al., 2023). Soft prompts for adapting small frozen LLMs to time series anomaly detection (Wei et al., 2025) show that soft prompting can substitute for full fine-tuning. These approaches treat the LLM as a sequence processor. PRISM uses the frozen LLM as a constraint on the representation, not as an end-to-end predictor.

In the multimodal direction, time-series forecasting through a frozen VLM (Zhong et al., 2025) combines temporal, visual (FFT-based synthetic images), and textual (statistical summaries) modalities, targeting stationary periodic signals in electricity and traffic domains. PRISM shares the high-level structure of multimodal time series processing but differs in the visual representation (CWT scalograms for non-stationary signals), the text representation (contextual meta-narratives

instead of statistical summaries), and the architectural role of the frozen LLM (coherence constraint instead of end-to-end predictor).

### A.3.5 Frozen Models as Encoders and Constraints

Frozen foundation models in non-language tasks include frozen language models conditioned through learned visual input embeddings (Tsimpoukelli et al., 2021), establishing soft prompting as a paradigm for multimodal learning without fine-tuning. Frozen LLM transformer blocks have been shown to act as effective visual encoders across diverse vision tasks, with an "information filtering hypothesis" proposing that the blocks suppress uninformative visual tokens and amplify discriminative ones (Pang et al., 2024).

A conceptually related principle appears in physics-informed neural networks (Raissi et al., 2019), where known governing equations impose fixed constraints on learned neural network solutions. The physics cannot adapt to the training data, preventing the optimizer from converging on solutions that fit the data but violate physical law. In PRISM, the frozen LLM's language prior serves the analogous function. It defines what counts as a coherent behavioral narrative. This definition is external to the behavioral training data, and the optimizer cannot modify it. The representation is constrained to satisfy a criterion the training process cannot shift, which is the structural mechanism for distribution invariance.

This frozen-constraint design differs from standard multi-task learning (Kendall et al., 2018; Ruder, 2017), where all task heads are trainable and co-adapt with the shared representation. In standard MTL, the auxiliary task's definition of correctness evolves during training and can encode distribution-specific biases. In PRISM, the coherence constraint is immutable, providing a structural property standard MTL lacks.

### A.3.6 Wavelet-Based Representations for Deep Learning

The Continuous Wavelet Transform provides joint time-frequency localization for non-stationary signals (Mallat, 1999; Torrence and Compo, 1998). Converting time series into CWT scalograms and processing them through pre-trained CNNs has been applied in medical signal classification (Wang et al., 2021), industrial fault detection (Siddique et al., 2025), power quality analysis (Salles and Ribeiro, 2023), and environmental monitoring. These methods typically fine-tune the vision backbone on scalogram images, which risks overfitting on small datasets. PRISM keeps the vision encoder frozen and introduces a trainable input-side domain bridge (multi-scale convolutions) that adapts the scalogram to the encoder's expected input distribution without modifying internal weights. This input-side adaptation aligns with findings that frozen encoders with lightweight adaptation can match or exceed fine-tuned alternatives on limited training data (Jin et al., 2024; Tsimpoukelli et al., 2021).

### A.3.7 Multi-Task Learning and Multi-Resolution Supervision

Multi-task learning trains a shared representation on multiple objectives simultaneously, with auxiliary tasks providing implicit regularization (Ruder, 2017). Homoscedastic uncertainty weighting for combining losses of different scales (Kendall et al., 2018) is the formulation PRISM adopts for balancing prediction and generation objectives. The distinction between PRISM and standard multi-task learning is twofold. First, the two losses in PRISM address the same forecasting task at different supervision resolutions. The prediction head provides sparse supervision; the generation path provides dense token-level supervision. These are not two separate tasks. Second, the generation path's definition of correctness is fixed by a frozen model, making it distribution-invariant. The combination of multi-resolution supervision with a frozen constraint surface is, to our knowledge, novel in the behavioral forecasting literature.

## A.4 LLM Prompt for Behavioral Meta-Narrative Generation

For the Meta-Narrative approach, we employ a two-stage prompting process to transform longitudinal behavioral signals into a coherent textual representation, building on structured reasoning paradigms used in prior work (Wei et al., 2023; Kojima et al., 2023).

1. **Feature-Level Behavioral Analysis**: In the first stage, GPT-4o analyzes each behavioral signal independently across time. These signals include patterns such as daily walking activity, sleep duration and regularity, and time spent on phone usage. The model identifies temporal trends (e.g., increasing, decreasing, stable), fluctuations, and irregularities within each feature. This stage focuses on extracting localized temporal structure from individual behavioral trajectories, enabling precise characterization of how each signal evolves over time.

2. **Contextual Behavioral Synthesis**: In the second stage, GPT-4o integrates these feature-level observations into a unified narrative. This synthesis emphasizes relationships across behaviors (e.g., reduced physical activity co-occurring with increased phone usage or irregular sleep patterns) and interprets them within a broader behavioral context. The resulting narrative captures both temporal dynamics and cross-feature dependencies, producing a semantically coherent description of the user's overall behavioral trajectory.

This two-stage transformation converts multi-dimensional time-series behavioral data into a concise, contextually grounded narrative. By leveraging the LLM's ability to reason about temporal patterns and behavioral relationships, the Meta-Narrative representation preserves meaningful structure that may otherwise be lost through direct serialization of raw data.

System Prompt – Meta-Narrative

You are an expert behavioral analyst tasked with evaluating a user's daily behavioral patterns over a 4-week period. The data includes signals such as physical activity (e.g., walking patterns), sleep duration and consistency, phone usage, and routine behaviors.
Your objective is to analyze how the user's behavior evolves across the 4 weeks. In your response:

- Identify and describe specific behavioral trends, such as changes in activity levels, sleep patterns, or phone usage.
- Reference specific weeks (e.g., "In Week 1…", "By Week 3…").
- Use precise language to describe changes, such as "physical activity increased by Week 2", "sleep duration decreased from Week 1 to Week 4", or "phone usage remained stable until Week 3".
- Avoid vague terms like "overall" or "in general" to ensure analytical precision.
- Provide a concise, natural, and evidence-based analysis in 3–4 sentences.
- Do not include any personal or identifying information.

### A.5 Implementation Details

We describe the implementation details for all models evaluated in this work, including time-series baselines, LLM-based approaches, bimodal VLM configurations, and the proposed PRISM architecture.

**Time-Series Baselines** We implement strong time-series baselines using transformer-based architectures, including PatchTST and iTransformer. Both models operate on 28-day sequences with feature-wise temporal inputs. Data preprocessing includes mean imputation for missing values followed by standardization using statistics computed from the training split. Models are trained using cross-entropy loss with class-balanced weighting, AdamW optimization, and gradient clipping. Detailed implementation follows standard practices for sequence modeling, with patch-based tokenization for PatchTST and feature-wise tokenization for iTransformer.

**LLM-Based Configuration** For text-only baselines, we utilize Llama 3.1 8B Instruct (Grattafiori et al., 2024). The model is fine-tuned using parameter-efficient LoRA adaptation (Hu et al., 2021), applied to all attention modules. We use a rank of 32 and scaling factor $\alpha = 16$. Training is performed using a causal language modeling objective with teacher forcing. Optimization is carried out using paged AdamW (8-bit) with a learning rate of $1 \times 10^{-5}$ and cosine decay schedule, with a warmup ratio of 0.1. The model is trained for 20 epochs with a batch size of 8 under bfloat16 mixed precision.

**Bimodal VLM Configuration and Training** For multimodal experiments, we adopt LLaVA-NeXT 7B (Liu et al., 2024a) as the base vision-language model, consisting of a SigLIP vision encoder, a Qwen1.5-7B-Chat language model, and an MLP-based projection module that maps visual tokens into the language embedding space. We fine-tune the model using parameter-efficient LoRA adaptation with rank 64 and scaling factor $\alpha = 128$, applied to all attention modules in the language model while keeping the vision encoder frozen. Training is performed using a causal language modeling objective with teacher forcing, optimized using AdamW with a learning rate of $1 \times 10^{-4}$, cosine decay schedule, 0.1 warmup ratio, and weight decay of 0.01. The model is trained for 20 epochs with a batch size of 4 using bfloat16 mixed precision.

During inference, we employ deterministic decoding using greedy search (temperature = 0) with a maximum generation length of 2000 tokens to ensure stable and reproducible outputs.

**PRISM Architecture and Training** PRISM is a multimodal architecture designed to integrate temporal, spectral, and textual representations of longitudinal behavioral data. Raw sequences, with dataset-specific window lengths and feature counts, are transformed using the Continuous Wavelet Transform (CWT) with a Morlet wavelet to generate time-frequency scalograms, preserving localized non-stationary patterns. The model employs three parallel encoders: (i) a Transformer-based temporal encoder projecting raw sequences into a 256-dimensional latent space, (ii) a vision encoder that converts CWT scalograms into RGB inputs and extracts 512-dimensional embeddings using a frozen CLIP backbone, and (iii) a text encoder that processes static contextual information via a trainable injection module followed by a frozen CLIP text encoder. These heterogeneous representations are aligned through a pairwise cross-attention fusion module, where modality interactions (text–vision, text–temporal, temporal–vision) are modeled and combined using learnable gating, producing a unified 256-dimensional representation.

**Dual-Path Learning and Optimization** The fused representation is used in two complementary pathways: a direct classification head for binary prediction and a soft prompting mechanism that expands the representation into a sequence of virtual tokens, which are prepended to a fixed instruction and processed by a frozen LLaMA-3 model for generative reasoning. The model is trained using a dual-loss objective that jointly optimizes classification (binary cross-entropy) and generation (cross-entropy) losses with learnable uncertainty weighting, enabling dynamic balancing between the two tasks. Since the language model is frozen, gradients from the generative objective propagate through the soft prompting module and upstream encoders, enforcing alignment between numerical prediction and linguistic reasoning.

### A.6 Full Ablation Results

This subsection reports per-metric results for the ablation studies across all four datasets, including the CrossCheck fixed-$\alpha$ dual-path and modality-removal conditions.

#### A.6.1 Dual-Path Training Objective Ablation

Tables 8–11 report the per-metric results across the four datasets, organized by inference path (Generation, Prediction) and evaluation condition (ID, OOD).

**Note on prediction path performance.** On the prediction path (Tables 10–11), the fixed equal weighting ($\alpha = 0.5$) outperforms the learned weighting on OOD evaluation by 1.18 to 1.72 points across the three datasets where the fixed-$\alpha$ run is available. The learned weighting assigns 2.0$\times$ to 6.9$\times$ more effective weight to the generation loss across datasets (Section 4.4), benefiting the generation path at a small cost to the prediction path. The generation path achieves 8.00 to 15.56 points higher OOD accuracy across the four datasets and is the primary inference mechanism, which makes this trade-off favorable for the overall system.

#### A.6.2 Modality Removal Ablation

Tables 12–15 report the modality removal results, organized in the same path/condition structure as the dual-path tables.

Table 8: Dual-path ablation. OOD results for the generation path across all datasets (Acc / P / R / F1, %).

| Dataset | Objective | Acc | P | R | F1 |
|---|---|---|---|---|---|
| GLOBEM | $\mathcal{L}_{\text{pred}}$ only | 59.12 | 62.79 | 57.20 | 53.00 |
| | $\mathcal{L}_{\text{gen}}$ only | 71.33 | 77.39 | 72.64 | 70.36 |
| | Fixed $\alpha = 0.5$ | 75.73 | 80.17 | 74.53 | 74.15 |
| | Learned | **79.93** | **82.53** | **79.06** | **79.14** |
| LifeSnaps | $\mathcal{L}_{\text{pred}}$ only | 64.06 | 66.55 | 63.39 | 61.97 |
| | $\mathcal{L}_{\text{gen}}$ only | 70.31 | 71.97 | 69.84 | 69.41 |
| | Fixed $\alpha = 0.5$ | 76.56 | 80.23 | 76.00 | 75.55 |
| | Learned | **81.25** | **83.46** | **80.84** | **80.78** |
| MFAFY | $\mathcal{L}_{\text{pred}}$ only | 55.71 | 67.81 | 63.65 | 54.67 |
| | $\mathcal{L}_{\text{gen}}$ only | 68.29 | 67.46 | 61.24 | 60.79 |
| | Fixed $\alpha = 0.5$ | 71.71 | 74.63 | 64.29 | 64.12 |
| | Learned | **73.71** | **78.03** | **66.50** | **66.77** |
| CrossCheck | $\mathcal{L}_{\text{pred}}$ only | 61.90 | 61.01 | 76.43 | 54.91 |
| | $\mathcal{L}_{\text{gen}}$ only | 71.02 | 64.47 | 83.57 | 62.69 |
| | Fixed $\alpha = 0.5$ | 78.53 | 67.10 | 85.86 | 68.73 |
| | Learned | **84.97** | **71.35** | **88.86** | **75.19** |

Table 9: Dual-path ablation. ID results for the generation path across all datasets (Acc / P / R / F1, %).

| Dataset | Objective | Acc | P | R | F1 |
|---|---|---|---|---|---|
| GLOBEM | $\mathcal{L}_{\text{pred}}$ only | 61.43 | 66.05 | 60.52 | 57.35 |
| | $\mathcal{L}_{\text{gen}}$ only | 74.44 | 79.30 | 73.78 | 72.97 |
| | Fixed $\alpha = 0.5$ | 77.58 | 79.21 | 77.19 | 77.08 |
| | Learned | **81.17** | **83.39** | **80.75** | **80.69** |
| LifeSnaps | $\mathcal{L}_{\text{pred}}$ only | 68.75 | 67.95 | 61.67 | 61.35 |
| | $\mathcal{L}_{\text{gen}}$ only | 75.00 | 75.00 | 70.00 | 70.91 |
| | Fixed $\alpha = 0.5$ | 81.25 | 83.33 | 85.00 | 81.18 |
| | Learned | **87.50** | **87.50** | **90.00** | **87.30** |
| MFAFY | $\mathcal{L}_{\text{pred}}$ only | 58.49 | 69.90 | 65.68 | 57.75 |
| | $\mathcal{L}_{\text{gen}}$ only | 73.58 | **76.05** | 66.97 | 67.46 |
| | Fixed $\alpha = 0.5$ | 73.58 | 72.63 | **73.86** | 72.80 |
| | Learned | **75.47** | 74.18 | 72.42 | **73.01** |
| CrossCheck | $\mathcal{L}_{\text{pred}}$ only | 66.49 | 72.91 | 76.61 | 66.17 |
| | $\mathcal{L}_{\text{gen}}$ only | 84.53 | 81.30 | 87.93 | 82.11 |
| | Fixed $\alpha = 0.5$ | 86.59 | 83.74 | 86.42 | 85.24 |
| | Learned | **92.78** | **93.06** | **88.92** | **90.68** |

**Modality contribution on the prediction path.** The contribution ordering on the prediction path (Table 14) is mostly consistent with the generation path. On GLOBEM and MFAFY, removing text causes the largest OOD drop ($-5.98$ and $-6.85$), with vision and temporal removals smaller. LifeSnaps shows a different pattern. Removing vision causes the largest drop ($-9.37$), text close behind ($-7.81$), and temporal smallest ($-6.25$). Across the three datasets, the text and vision streams account for the largest accuracy drops, supporting the claim that the modality ordering reflects the information content of each stream, independent of the generative output mechanism.

### A.7 GPT-4o Robustness Analysis

The semantic stream's meta-narratives and target narratives are generated by GPT-4o using fixed template-based prompts with structured slots. The template design constrains output variance by specifying format, content organization, and abstraction level, guiding any language model to produce semantically equivalent narratives that focus on the same structural elements.

Table 10: Dual-path ablation. OOD results for the prediction path across all datasets (Acc / P / R / F1, %).

| Dataset | Objective | Acc | P | R | F1 |
|---|---|---|---|---|---|
| GLOBEM | $\mathcal{L}_{\text{pred}}$ only | 65.10 | 73.70 | **66.71** | 62.85 |
| | $\mathcal{L}_{\text{gen}}$ only | 55.86 | 67.08 | 58.03 | 50.27 |
| | Fixed $\alpha = 0.5$ | **68.26** | **75.22** | 66.60 | **64.57** |
| | Learned | 67.08 | 73.49 | 65.40 | 63.22 |
| LifeSnaps | $\mathcal{L}_{\text{pred}}$ only | 64.06 | 66.55 | 63.39 | 61.97 |
| | $\mathcal{L}_{\text{gen}}$ only | 64.06 | 71.47 | 63.10 | 59.73 |
| | Fixed $\alpha = 0.5$ | **70.31** | **76.04** | **69.55** | **68.06** |
| | Learned | 68.75 | 74.97 | 67.94 | 66.07 |
| MFAFY | $\mathcal{L}_{\text{pred}}$ only | 64.57 | 61.50 | 56.77 | 55.21 |
| | $\mathcal{L}_{\text{gen}}$ only | 52.29 | **66.55** | 60.65 | 50.47 |
| | Fixed $\alpha = 0.5$ | **67.43** | 65.71 | **60.70** | **60.34** |
| | Learned | 65.71 | 62.82 | 59.63 | 59.40 |
| CrossCheck | $\mathcal{L}_{\text{pred}}$ only | 66.91 | 63.15 | 81.24 | 59.27 |
| | $\mathcal{L}_{\text{gen}}$ only | 62.97 | 62.09 | 79.01 | 56.18 |
| | Fixed $\alpha = 0.5$ | 66.19 | 62.68 | 80.18 | 58.55 |
| | Learned | **69.41** | **63.43** | **81.90** | **60.98** |

Table 11: Dual-path ablation. ID results for the prediction path across all datasets (Acc / P / R / F1, %).

| Dataset | Objective | Acc | P | R | F1 |
|---|---|---|---|---|---|
| GLOBEM | $\mathcal{L}_{\text{pred}}$ only | 72.20 | 75.09 | 71.63 | 71.02 |
| | $\mathcal{L}_{\text{gen}}$ only | 56.95 | 63.23 | 57.95 | 52.68 |
| | Fixed $\alpha = 0.5$ | **72.65** | 75.43 | **72.10** | **71.54** |
| | Learned | 70.85 | **76.39** | 70.10 | 68.74 |
| LifeSnaps | $\mathcal{L}_{\text{pred}}$ only | 75.00 | 75.00 | 76.67 | 74.60 |
| | $\mathcal{L}_{\text{gen}}$ only | 62.50 | 58.33 | 56.67 | 56.36 |
| | Fixed $\alpha = 0.5$ | **81.25** | 80.91 | **78.33** | **79.22** |
| | Learned | 75.00 | **85.71** | 66.67 | 66.67 |
| MFAFY | $\mathcal{L}_{\text{pred}}$ only | **71.70** | **72.08** | 65.45 | 65.81 |
| | $\mathcal{L}_{\text{gen}}$ only | 56.60 | 69.00 | 64.17 | 55.59 |
| | Fixed $\alpha = 0.5$ | **71.70** | 70.18 | **67.42** | **68.01** |
| | Learned | 69.81 | 67.74 | 65.91 | 66.35 |
| CrossCheck | $\mathcal{L}_{\text{pred}}$ only | 77.84 | 74.80 | 79.59 | 75.58 |
| | $\mathcal{L}_{\text{gen}}$ only | 67.01 | 63.46 | 65.99 | 63.52 |
| | Fixed $\alpha = 0.5$ | 81.95 | **80.21** | **86.28** | **80.61** |
| | Learned | **82.98** | 79.23 | 83.18 | 80.52 |

To assess the method's sensitivity to the specific generator choice, we conducted preliminary ablations on GLOBEM replacing GPT-4o with Llama-3-8B-Instruct for narrative generation. BERTScore similarity between the two generators exceeded 95%, consistent with the template-based design's role in constraining output variance.

Separately, text-only baselines with maximal detail (serializing the full trajectory into prose) do not outperform the template-constrained meta-narratives, and all text-only variants underperform multimodal models. This pattern suggests that PRISM's performance gains come from complementary visual-textual reasoning on the shared representation, and not from any specific fluency or richness of GPT-4o's generation.

These two lines of evidence support the method's reproducibility and generator-agnostic character, subject to the template-based prompt design.

Table 12: Modality removal. OOD results for the generation path across all datasets (Acc / P / R / F1, %).

| Dataset | Config. | Acc | P | R | F1 |
|---|---|---|---|---|---|
| GLOBEM | Full | **79.93** | **82.53** | **79.06** | **79.14** |
| | w/o Text | 65.62 | 72.93 | 66.47 | 63.33 |
| | w/o Vision | 68.81 | 75.61 | 67.17 | 65.32 |
| | w/o Temporal | 71.13 | 77.20 | 69.64 | 68.40 |
| LifeSnaps | Full | **81.25** | **83.46** | **80.84** | **80.78** |
| | w/o Text | 68.75 | 74.97 | 69.50 | 67.18 |
| | w/o Vision | 71.88 | 74.23 | 72.34 | 70.85 |
| | w/o Temporal | 73.44 | 77.95 | 74.05 | 72.63 |
| MFAFY | Full | **73.71** | **78.03** | 66.50 | **66.77** |
| | w/o Text | 61.14 | 70.96 | 67.76 | 60.73 |
| | w/o Vision | 64.29 | 72.45 | **70.28** | 64.10 |
| | w/o Temporal | 67.71 | 65.44 | 62.28 | 62.45 |
| CrossCheck | Full | **84.97** | **71.35** | **88.86** | **75.19** |
| | w/o Text | 69.77 | 60.81 | 77.42 | 58.94 |
| | w/o Vision | 74.24 | 64.39 | 81.65 | 64.28 |
| | w/o Temporal | 78.71 | 67.26 | 85.14 | 68.63 |

Table 13: Modality removal. ID results for the generation path across all datasets (Acc / P / R / F1, %).

| Dataset | Config. | Acc | P | R | F1 |
|---|---|---|---|---|---|
| GLOBEM | Full | **81.17** | **83.39** | **80.75** | **80.69** |
| | w/o Text | 67.71 | 71.60 | 67.00 | 65.62 |
| | w/o Vision | 69.96 | 76.40 | 69.15 | 67.44 |
| | w/o Temporal | 72.65 | 78.15 | 71.93 | 70.80 |
| LifeSnaps | Full | **87.50** | **87.50** | **90.00** | **87.30** |
| | w/o Text | 75.00 | 73.33 | 73.33 | 73.33 |
| | w/o Vision | 81.25 | 80.91 | 78.33 | 79.22 |
| | w/o Temporal | 81.25 | 80.16 | 81.67 | 80.57 |
| MFAFY | Full | **75.47** | **74.18** | **72.42** | **73.01** |
| | w/o Text | 64.15 | 72.55 | 70.23 | 63.95 |
| | w/o Vision | 67.92 | 65.53 | 63.41 | 63.74 |
| | w/o Temporal | 71.70 | 70.87 | 66.44 | 67.00 |
| CrossCheck | Full | **92.78** | **93.06** | **88.92** | **90.68** |
| | w/o Text | 79.38 | 80.16 | 76.91 | 77.52 |
| | w/o Vision | 83.51 | 84.27 | 80.44 | 81.92 |
| | w/o Temporal | 87.11 | 88.64 | 84.27 | 85.61 |

Table 14: Modality removal. OOD results for the prediction path across all datasets (Acc / P / R / F1, %).

| Dataset | Config. | Acc | P | R | F1 |
|---|---|---|---|---|---|
| GLOBEM | Full | **67.08** | **73.49** | **65.40** | **63.22** |
| | w/o Text | 61.10 | 69.17 | 58.98 | 53.93 |
| | w/o Vision | 61.94 | 71.13 | 63.73 | 59.01 |
| | w/o Temporal | 62.98 | 72.42 | 64.75 | 60.20 |
| LifeSnaps | Full | **68.75** | **74.97** | **67.94** | **66.07** |
| | w/o Text | 60.94 | 65.62 | 61.73 | 58.67 |
| | w/o Vision | 59.38 | 67.07 | 58.26 | 52.73 |
| | w/o Temporal | 62.50 | 68.57 | 63.34 | 60.00 |
| MFAFY | Full | **65.71** | 62.82 | 59.63 | 59.40 |
| | w/o Text | 58.86 | 69.88 | 65.93 | 58.20 |
| | w/o Vision | 60.86 | 70.83 | 67.53 | 60.41 |
| | w/o Temporal | 61.43 | **71.10** | **67.99** | **61.04** |
| CrossCheck | Full | **69.41** | **63.43** | **81.90** | **60.98** |
| | w/o Text | 62.79 | 58.42 | 72.18 | 54.91 |
| | w/o Vision | 65.83 | 60.37 | 76.35 | 57.49 |
| | w/o Temporal | 67.62 | 61.91 | 79.36 | 59.26 |

Table 15: Modality removal. ID results for the prediction path across all datasets (Acc / P / R / F1, %).

| Dataset | Config. | Acc | P | R | F1 |
|---|---|---|---|---|---|
| GLOBEM | Full | **70.85** | **76.39** | **70.10** | **68.74** |
| | w/o Text | 63.23 | 71.99 | 64.18 | 60.04 |
| | w/o Vision | 63.68 | 71.93 | 62.67 | 58.90 |
| | w/o Temporal | 65.02 | 72.97 | 64.06 | 60.82 |
| LifeSnaps | Full | **75.00** | **85.71** | 66.67 | 66.67 |
| | w/o Text | 68.75 | 66.36 | 65.00 | 65.37 |
| | w/o Vision | 68.75 | 67.95 | 61.67 | 61.35 |
| | w/o Temporal | 68.75 | 77.27 | **75.00** | **68.63** |
| MFAFY | Full | **69.81** | 67.74 | 65.91 | **66.35** |
| | w/o Text | 60.38 | **70.79** | 67.20 | 59.86 |
| | w/o Vision | 62.26 | 62.04 | 62.80 | 61.59 |
| | w/o Temporal | 64.15 | 70.16 | **69.24** | 64.10 |
| CrossCheck | Full | **82.98** | **79.23** | **83.18** | **80.52** |
| | w/o Text | 76.29 | 72.64 | 77.35 | 73.41 |
| | w/o Vision | 78.87 | 75.39 | 79.74 | 75.92 |
| | w/o Temporal | 80.93 | 77.86 | 81.62 | 78.34 |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the main claims regarding cross-distribution generalization, the frozen coherence constraint, and the PRISM framework. These claims are supported by experimental results (Section 4.3) and ablation studies (Section 4.4), with scope clarified in Section 5.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 includes a dedicated **Limitations** subsection covering reliance on GPT-4o meta-narratives, the binary classification target restriction, single training runs without variance estimates, and the directed cross-attention fusion not being compared against symmetric alternatives. CrossCheck results are included for the fixed-$\alpha$ dual-path and modality-removal ablations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theorems or proofs. Theoretical concepts (e.g., information bottleneck) are used for interpretation, not formal derivation.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of datasets, OOD splits, model architecture, training objectives, and evaluation metrics (Sections 3 and 4), with full implementation details, preprocessing steps, and ablation configurations in the appendix.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and model checkpoints are not released at submission time. All datasets used are publicly available and cited, and the paper and appendix provide sufficient implementation detail to enable reproduction.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies dataset splits, model components, training objectives, evaluation metrics, and ablation settings (Sections 3 and 4), with additional implementation details in the appendix.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Results are reported from single training runs with fixed random seeds due to computational constraints. This is acknowledged in the Limitations subsection of Section 5. The main comparisons show large effect sizes (27.13-point improvement on GLOBEM, gap reductions of 1.49× to 5.95× across datasets) unlikely to be explained by run-to-run variance, while smaller-margin comparisons would benefit from variance estimates.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports compute resources including GPU type (NVIDIA A40), approximate training time, and parameter scale (Section 4.3, Efficiency paragraph).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available datasets and standard modeling practices. The paper discusses deployment risks and responsible use considerations in behavioral health contexts (Section 5).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential risks in behavioral health prediction, including misuse in screening and decision-making contexts, and emphasizes the need for prospective validation before deployment (Section 5).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new models or datasets that pose high risk for misuse. It uses existing publicly available datasets and standard pretrained models.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and pretrained models are cited in their respective sections. Datasets include GLOBEM, LifeSnaps, MFAFY, and CrossCheck, used under their original release terms. Pretrained models include LLaMA-3 8B (Meta Llama 3 Community License), CLIP (MIT License), and LLaVA-NeXT 7B (Llama 2 Community License) for the Bimodal VLM baseline.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets or model checkpoints with the submission. Code release is planned upon acceptance.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve new data collection, crowdsourcing, or direct interaction with human subjects. It uses previously collected and publicly available datasets.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human subjects research is conducted. All datasets used were collected in prior studies with their own ethical oversight.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are a core component of the methodology. The paper describes the use of frozen LLaMA-3 8B as the coherence-criterion decoder (Section 3) and GPT-4o for generating semantic meta-narratives (Section 3.1 and Appendix A.4), with reproducibility analysis across LLM providers in Appendix A.7.