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 · [2] Department of Electrical and Computer Engineering, University of Nebraska–Lincoln

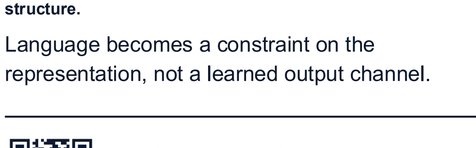
NEURAL INFORMATION PROCESSING SYSTEMS 2026

01 · THE PROBLEM

Why do behavior models break on new people?

Train on one cohort, test on the next, and accuracy collapses. On GLOBEM depression forecasting, the best published result sat at 52.8% for years.

GLOBEM · out-of-distribution accuracy (%) · coin flip = 50



02 · THE DIAGNOSIS

A class label is a weak teacher

One sparse signal per sample. It says whether the answer was right, never how to organize the representation. Shortcuts win.

03 · THE IDEA

What if the judge can't be moved?

Representation trainable → **LLM** fine-tuned

Both adapt to the cohort, and agree on its shortcuts.

Representation trainable → **LLM** frozen

Only the representation can move. It has to carry real structure.

Language becomes a constraint on the representation, not a learned output channel.

THE IDEA IN ONE LINE

Freeze the LLM. Make the representation earn a coherent story.

79.93%

OOD accuracy on GLOBEM

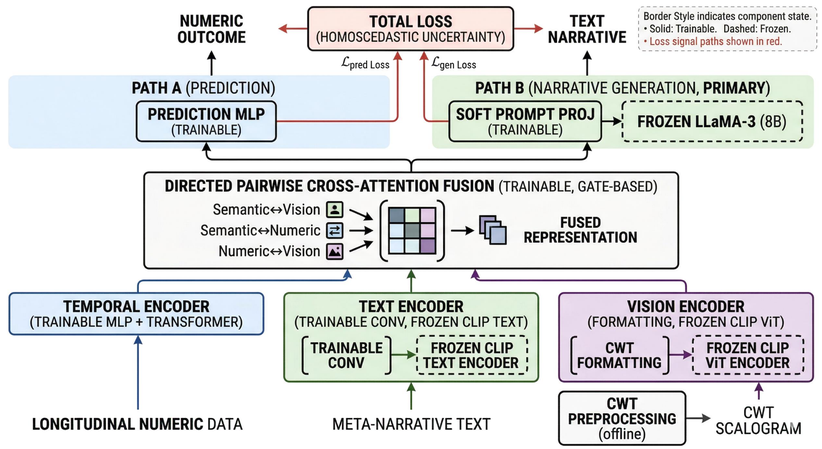
+27.13 pts over the classification benchmark. ID-to-OOD gap: 1.24 pts.

04 · ARCHITECTURE

Three views of one trajectory

Temporal · what happened | **Spectral** · when, at what scale | **Semantic** · what it means

Fused by directed, gated cross-attention, then sent down two paths.



05 · TRAINING

Two paths out: one sparse, one dense

PATH A · PREDICTION MLP

One label per sample

Sparse. Keeps outcomes separable.

PATH B · FROZEN LLaMA-3 8B

Hundreds of tokens per sample

Dense. 32 soft prompts must yield a coherent narrative.

Learned uncertainty weighting favors the generation loss 6.2–6.9× on sensor-rich datasets, 2.0× on categorical MFAFY.

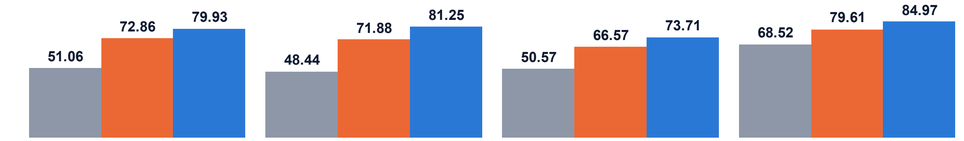
06 · MAIN RESULT

Does it hold on people it never saw?

Yes. Highest OOD accuracy on all four datasets.

Best time-series model · Fine-tuned VLM (LLaVA-NeXT + LoRA) · **PRISM (frozen)**

OOD accuracy, % · + pts: PRISM over the best baseline

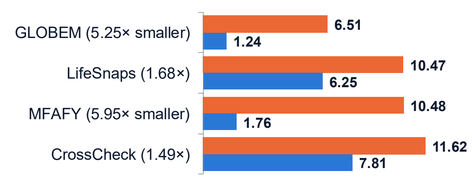


07 · GENERALIZATION GAP

Freezing shrinks the gap

A smaller ID-to-OOD gap on every dataset, by up to 5.95×.

Fine-tuned VLM · **PRISM** gap, pts



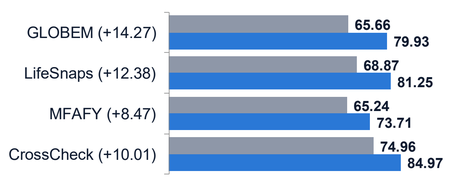
Trains in under 3 h on one A40 with ~18M trainable parameters. The fine-tuned VLM: over a day on 8 × A40.

08 · CONTROL

Is it just the LLM?

No. The same frozen LLaMA-3, prompted directly, trails by 8–14 points.

Frozen LLaMA-3, zero-shot · **PRISM** OOD accuracy, %



Same meta-narratives as input. No soft prompts, no fusion, no task training.

09 · ABLATION: LOSSES

Both losses are needed

Neither gets there alone.



GLOBEM OOD accuracy (%), on the path each objective trains.

10 · ABLATION: STREAMS

Meaning matters most

But every view earns its place.



Drop in GLOBEM OOD accuracy (pts) with one stream removed. Same ordering on all four datasets.

Ahatsham Hayat
ahatsham.github.io
afnu@bwh.harvard.edu

Supported by the U.S. National Science Foundation (NSF DUE 2142558)

HAL 2.0 · Human-First AI Lab
mrhasan-ai.github.io
hasan@unl.edu