ML Lab Research Proposal

CLD-Trans: Causal-Lagged Dynamic Transformer for Biosignals

Team:	Ankur (solo)
Date:	April 25, 2026
Project:	CLD-Trans: Causal-Lagged Dynamic Transformer for Biosignals
Repo:	https://github.com/coolguy46/CLD-Trans
Target Compute:	8x AMD Instinct M1300X (single node), approx. 2 weeks

---

1. PROBLEM STATEMENT

Existing biosignal foundation models (BIOT, BENDR, NeuroLM, ECG-FM) treat EEG/ECG channels as either independent or coupled by a static attention pattern, even though clinicians know that signals propagate through the body in continuous time—seizures spread from a focal lead to its neighbors over hundreds of milliseconds, and ECG depolarization reaches V6 a few milliseconds after V1. Existing causal-discovery methods (DYNOTEARS, Rhino) recover lagged graphs but assume integer-sample lags and acyclic dependencies, which is biologically wrong. I want to test whether a model built around continuous-valued, possibly cyclic propagation lags can recover clinically meaningful structure (e.g. seizure focal leads) with no task labels.

2. PROPOSED APPROACH

I am building CLD-Trans, which combines three components under a single non-Gaussian Lagged-Delay Structural Equation Model (LD-SEM) objective:

1. A Motif VQ-VAE that tokenizes raw waveforms into a 512-entry codebook of physiological primitives (spike-wave, QRS, K-complex, etc.).
2. A differentiable fractional-lag operator (FFT phase shift, modules/fractional_delay.py) that learns sub-sample per-channel-pair delays tau_ij in the range 0 to tau_max.
3. A graph-conditioned Neural ODE that integrates latent state through the inferred time-varying adjacency A(t).

I picked this stack over a plain Transformer FM because the lag operator gives me an interpretable, identifiability-friendly target—the closest thing in the literature is Rhino, but Rhino assumes integer lags and acyclic graphs, neither of which holds for EEG/ECG. The LD-SEM objective is also the same object I want to use in the identifiability theorem on the theory side, so method and theory share one optimization target.

3. DATASET

All datasets are public and pulled from PhysioNet open-data S3 buckets via scripts/download_datasets_aws.sh. No DUA, no PHI handling.

Pretraining (no labels used):

Corpus	Source	Raw Size	Notes
MIMIC-IV-ECG v1.0	s3://physionet-open/mimic-iv-ecg/1.0/	approx 90 GB	approx 800k 12-lead, 10 s, 500 Hz ECGs

EEG Motor Movement/Imagery (EEGMMIDB)	s3://physionet-open/eegmmidb/1.0.0/	approx 3 GB	109 subjects, 64-ch, 160 Hz

Downstream evaluation:

Dataset	Source	Raw Size	Use
CHB-MIT	PhysioNet	approx 40 GB	Zero-shot focal-lead localization (headline) + few-shot
PTB-XL	PhysioNet	approx 3 GB	Few-shot 5-superclass arrhythmia
Sleep-EDF	PhysioNet	approx 8 GB	Few-shot 5-stage sleep scoring

Preprocessing

- Resample to a per-modality common rate (160 Hz for EEG, 500 Hz for ECG).
- Bandpass filter (0.5-40 Hz EEG, 0.05-150 Hz ECG), z-score per channel.
- Cache fixed-length windows as memory-mapped npy shards for fast data-parallel loading.

Realistic on-disk footprint: approx 145 GB raw + roughly the same again for the npy window cache, so I plan for approx 300 to 400 GB of scratch, well under a 1 TB budget.

4. EVALUATION METRICS

Experiment	Metric
Zero-shot focal-lead localization (CHB-MIT, headline)	top-1/top-3/top-5 accuracy of argmin over i of (sum over j of tau_ij) vs. clinician annotation
Few-shot CHB-MIT seizure detection	AUROC, AUPRC at 1 percent/10 percent/100 percent label budgets
Few-shot PTB-XL arrhythmia	macro-AUROC over 5 superclasses
Few-shot Sleep-EDF	macro-F1 and Cohen's kappa over 5 stages
Synthetic LD-SEM identifiability	tau recovery MAE, edge-support F1 vs. ground truth
Ablations (no-VQ, no-lag, integer-lag, no-ODE)	Same primary metric per dataset, 5-seed mean plus or minus 95 percent CI

Seeds: 42, 123, 7, 0, 256.

5. COMPUTE REQUIREMENTS

Estimates below are scaled from a measured single-GPU smoke run (python main.py mode=stage1 train.max_steps=1) on a development GPU. Padded by 25% for M1300X.

Phase	Hardware	Estimated Wall-Clock
Stage 1 Motif VQ-VAE + LD-SEM pretraining (bf16, DDP across 8 GPUs, eff. batch 256, approx 7.7k steps/epoch, approx 60 epochs)	8x M1300X	approx 3 to 4 days
Stage 2 Graph-ODE fine-tune	8x M1300X	approx 1 day

Downstream eval (3 datasets x 3 budgets x 5 seeds)	1 to 2 M1300X per run	approx 1.5 days aggregate
Ablations + synthetic identifiability sweeps	1 to 8 M1300X	approx 2 days
Total	**8x M1300X node**	**approx 8 to 10 wall-clock days inside a 2-week window**

GPU memory: the largest activation tensor is the dense C-by-C-by-T lag tensor for 64-channel EEG plus the ODE adjoint state. On a single 192 GB MI300X this should fit without activation checkpointing; on an 80 GB H100 it would force checkpointing or tensor parallelism. That is the main reason I am asking for M1300X specifically rather than just 'any 8-GPU node'.

Storage: approx 400 GB scratch (datasets + npy cache + checkpoints).

Software / Dependencies

- ROCm 6.x + the official PyTorch ROCm wheel.
- pip install -r requirements.txt (stock PyTorch, torchdiffeq, numpy, scipy, mne, wfdb, hydra-core).
- AWS CLI for the one-time PhysioNet download.
- No CUDA-only kernels (no flash-attn v2, no xformers, no custom Triton CUDA kernels). FFT lag op is plain torch.fft.
- No Docker required; pure Python venv.

6. EXPECTED DELIVERABLES

- A trained Stage-1 motif VQ-VAE checkpoint on EEGMMIDB + MIMIC-IV-ECG, plus the trained Graph-ODE on top.
- Headline experiment table: zero-shot focal-lead localization on CHB-MIT vs. BIOT, BENDR, EEG-GCNN, DYNOTEARS, Rhino baselines.
- Few-shot transfer numbers on CHB-MIT, PTB-XL, Sleep-EDF at 1%/10%/100% label budgets.
- Synthetic LD-SEM identifiability table (tau recovery error and edge-support F1 with and without the non-Gaussianity assumption).
- A NeurIPS-format draft with plots and tables generated end-to-end by scripts/make_figures.sh, plus per-seed JSON metrics in results/ and logs in logs/.
- The full repo remains MIT-licensed and public, with an M1300X/ROCm-specific section in the README and the reproducibility statement.

I am intentionally not promising 'SOTA on every benchmark'—the goal is a clean, reproducible test of the identifiability + zero-shot focal-lead hypothesis, with the few-shot numbers as supporting evidence.

7. RISK & MITIGATION

Risk	Likelihood	Mitigation
Stage-1 pretraining doesn't converge in 60 epochs on the larger MIMIC-IV-ECG corpus	Medium	Early-stop on validation reconstruction; the headline experiments only need a converged motif codebook, not a fully converged loss curve. Can also subset MIMIC-IV-ECG to approx 200k recordings without changing the story.

ROCm-specific PyTorch op gap (e.g. torch.fft edge case, torchdiffeq adjoint quirk)	Medium	The lag op and ODE solver are isolated behind two modules with unit tests; CPU fallback exists for the FFT operator, and I can swap dopri5 for rk4 if the adjoint misbehaves on ROCm.
Identifiability theorem doesn't go through cleanly in time	Medium-High	Fallback: empirical lag-recovery on synthetic LD-SEM + zero-shot CHB-MIT result still stand on their own as an empirical paper. The theory becomes 'supporting' rather than headline.
Zero-shot focal-lead result is weak	Medium	Fall back to a self-supervised framing: pretrain + lag-only readout with light fine-tuning on a few seizure subjects. Still novel relative to BIOT/ BENDR.
Wall-clock overrun on the M1300X node	Low-Medium	Stage-1 schedule has a train.max_steps cap; can drop to 3 seeds instead of 5 and skip the 100 percent label condition without losing the core claims.
Dataset download stalls from PhysioNet S3	Low	Download script supports --only <dataset> for retries; CHB-MIT and EEGMMIDB are small enough to re-fetch quickly.