CLD-Trans · biosignals

Ankur

1 · PROBLEM

what and why

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 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 · APPROACH

model / method, and why this one

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

source, size, preprocessing

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 s3://physionet-open/mimic-iv- approx 90 approx 800k 12-lead, 10 s, 500 Hz MIMIC-IV-ECG v1.0 ecg/1.0/ GB ECGs EEG Motor Movement/Imagery s3://physionet- approx 3 GB 109 subjects, 64-ch, 160 Hz (EEGMMIDB) open/eegmmidb/1.0.0/ Downstream evaluation: Dataset Source Raw size Use approx 40 Zero-shot focal-lead localization (headline) + few- CHB-MIT PhysioNet GB 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. (My earlier 4 to 5 TB estimate was wrong; corrected here.)

4 · METRICS

how you’ll know it worked

Experiment Metric Zero-shot focal-lead localization (CHB-MIT, top-1 / top-3 / top-5 accuracy of argmin over i of (sum over j of tau_ij) vs. clinician headline) 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 Experiment Metric 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

`training time, memory, dependencies`

Estimates below are scaled from a measured single-GPU smoke run (python main.py mode=stage1 train.max_steps=1) on a development GPU. I have not run on MI300X yet, so I have padded the wall-clock by about 25 percent. Phase Hardware Estimated wall-clock Stage 1 — Motif VQ-VAE + LD-SEM pretraining (bf16, DDP across 8 GPUs, 8x MI300X approx 3 to 4 days eff. batch 256, approx 7.7k steps/epoch, approx 60 epochs) Stage 2 — Graph-ODE fine-tune 8x MI300X approx 1 day 1 to 2 Downstream eval (3 datasets x 3 budgets x 5 seeds) MI300X approx 1.5 days aggregate per run 1 to 8 Ablations + synthetic identifiability sweeps approx 2 days MI300X 8x MI300X approx 8 to 10 wall-clock days inside a 2- Total node week window, with buffer for re-runs 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 MI300X 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 . Bring-up is one command: bash scripts/server_master_run.sh --download public --smoke-only .

6 · DELIVERABLES

`model, paper, demo`

By the end of the two-week window I aim (not guarantee) to have: 1. A trained Stage-1 motif VQ-VAE checkpoint on EEGMMIDB + MIMIC-IV-ECG, plus the trained Graph-ODE on top. 2. Headline experiment table: zero-shot focal-lead localization on CHB-MIT vs. BIOT, BENDR, EEG-GCNN, DYNOTEARS, Rhino baselines. 3. Few-shot transfer numbers on CHB-MIT, PTB-XL, Sleep-EDF at 1 % / 10 % / 100 % label budgets. 4. Synthetic LD-SEM identifiability table (tau recovery error and edge-support F1 with and without the non-Gaussianity assumption). 5. 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/ . 6. The full repo remains MIT-licensed and public, with an MI300X / 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

`what breaks first, and plan B`

what breaks first, and plan B Risk Likelihood Mitigation Stage-1 pretraining doesn’t Early-stop on validation reconstruction; the headline experiments only need a converge in 60 epochs on the Medium converged motif codebook, not a fully converged loss curve. Can also subset MIMIC- larger MIMIC-IV-ECG corpus IV-ECG to approx 200k recordings without changing the story. ROCm-specific PyTorch op gap The lag op and ODE solver are isolated behind two modules with unit tests; CPU (e.g. torch.fft edge case, Medium fallback exists for the FFT operator, and I can swap dopri5 for rk4 if the adjoint torchdiffeq adjoint quirk) misbehaves on ROCm. Risk Likelihood Mitigation Fallback:

empirical lag-recovery on synthetic LD-SEM + zero-shot CHB-MIT result still Identifiability theorem doesn’t go Medium- stand on their own as an empirical paper. The theory becomes "supporting" rather than through cleanly in time high headline. Fall back to a self-supervised framing: pretrain + lag-only readout with light fine-tuning Zero-shot focal-lead result is weak Medium on a few seizure subjects. Still novel relative to BIOT/BENDR. Wall-clock overrun on the MI300X Low- Stage-1 schedule has a train.max_steps cap; can drop to 3 seeds instead of 5 and node medium skip the 100 percent label condition without losing the core claims. Dataset download stalls from Download script supports --only <dataset> for retries; CHB-MIT and EEGMMIDB are Low PhysioNet S3 small enough to re-fetch quickly. I will check in at the 1-week mark with intermediate Stage-1 numbers so we can decide together whether to push for the full 2-week plan or trim scope.