Semantic-Level Invariant Representation Learning for Cross-Hospital Clinical EEG Modeling

Kang Yin
Department of Artificial Intelligence
Korea University
charles_kang@korea.ac.kr

Hye-Bin Shin
Department of Brain and Cognitive Engineering
Korea University
hb_shin@korea.ac.kr

Seong-Whan Lee*
Department of Artificial Intelligence
Korea University
sw.lee@korea.ac.kr

Abstract

Robust deployment of clinical electroencephalography (EEG) models requires source-trained models to generalize to unseen hospitals without target-domain calibration, hospital identifiers, or clinical reports at inference. This is challenging because acquisition protocols, hardware, montages, and patient populations perturb low-level EEG statistics, while clinical interpretation is organized around higher-level semantic concepts. We argue that this abstraction mismatch limits conventional signal-level invariant learning: aligning feature distributions alone cannot distinguish clinically irrelevant variation from diagnostic content. We propose Text-Guided Invariant Learning (TG-IL), a deployment-oriented framework that uses clinician-authored reports only during training as semantic anchors for EEG representation learning. TG-IL combines stochastic recording-level aggregation, a variational semantic information bottleneck, and prototype-guided EEG–report alignment to preserve report-derived clinical semantics while mitigating site-specific nuisance variation. At inference, all text-side modules are discarded and the model operates solely on EEG. On a three-hospital clinical EEG benchmark, TG-IL improves zero-shot cross-hospital generalization and worst-case robustness while reducing hospital-specific information in learned representations.

1 Introduction

Robust generalization under distribution shift is a central barrier to deploying electroencephalography (EEG) models in clinical practice (Bang & Lee, 2022). A common deployment setting is to train on recordings from one or a few source hospitals and apply the model directly to an unseen hospital, without target-domain calibration, hospital identifiers, or clinical reports at inference time. This setting is difficult because acquisition protocols, hardware, montages, patient populations, and reporting conventions can vary substantially across institutions (Poziomska et al., 2025; Apicella et al., 2024). However, clinical EEG interpretation is not organized around stable low-level signal statistics alone, but around higher-level concepts such as background slowing, epileptiform activity, asymmetry, seizures, and normal variants. Cross-hospital robustness therefore requires reducing site-specific signal variation while preserving clinically meaningful semantic structure.

Most existing robustness methods enforce invariance at the signal or feature-distribution level, using data augmentation, distribution matching, or domain-adversarial training (Rommel et al., 2022; Rangwani et al., 2022; Ju et al., 2025; Yin et al., 2024). While effective in many settings, these methods leave the target of invariance under-specified: statistical alignment alone cannot determine which variations are clinically irrelevant and which encode diagnostic content. This abstraction

*Corresponding author.

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

Table 1: Deployment-driven design of TG-IL.

Clinical deployment constraint	**TG-IL response**
Reports summarize full recordings	Stochastic bag-level aggregation
Reports unavailable at inference	Train-time-only text modules
Site shifts perturb low-level EEG	Semantic information bottleneck
Clinical concepts are many-to-one	Prototype-level EEG–report alignment

mismatch motivates using semantic-level constraints for invariant representation learning, where invariance is anchored to the level at which EEG is interpreted in clinical practice rather than only to low-level signal statistics.

Clinician-written EEG reports provide a natural source of semantic supervision. Reports summarize long, heterogeneous recordings into high-level clinical impressions rather than dense temporal annotations (Tatum IV et al., 2016; Acharya & Acharya, 2022; Grant et al., 2014; Tveit et al., 2023). Although raw report wording may reflect site-specific templates or terminology, report-derived semantic representations can serve as approximate anchors for pathological interpretation. Our goal is not report generation or multimodal inference: reports are used only during training to shape EEG representations, while deployment remains strictly EEG-only.

Motivated by this use case, we propose *Text-Guided Invariant Learning* (TG-IL), a deployment-oriented framework for semantic-level invariant representation learning in clinical EEG. As summarized in Table 1, TG-IL translates clinical deployment constraints into four design choices: stochastic bag-level aggregation, training-time-only report supervision, a variational semantic information bottleneck, and prototype-level EEG–report alignment. Together, these components aim to preserve report-derived clinical semantics while mitigating site-specific nuisance variation.

We evaluate TG-IL on clinical EEG data from three hospitals under deployment-oriented cross-hospital protocols. Models are trained with source-hospital data, selected using source-only validation, and evaluated directly on unseen hospitals without target-domain adaptation. Our evaluation covers both predictive robustness and deployment constraints: zero-shot cross-hospital transfer, multi-source leave-one-hospital-out generalization, fine-grained subtype transfer, partial report availability, deployment-time efficiency, semantic-invariance diagnostics, domain leakage analysis, and controlled perturbations of EEG–report correspondence.

Our contributions are threefold: (i) we formulate robust clinical EEG modeling as a *semantic-level invariance* problem, identifying the abstraction mismatch between signal-level distribution shift and report-level clinical interpretation; (ii) we introduce TG-IL, a training-time report-guided framework that learns EEG-only representations through stochastic bag aggregation, a variational semantic information bottleneck, and prototype-guided semantic alignment; and (iii) we provide systematic cross-hospital evidence that semantic-level constraints improve EEG-only deployment robustness, tolerate incomplete report supervision, add negligible inference-time cost, reduce domain leakage, and provide evidence that the gains are linked to EEG–report semantic correspondence rather than generic compression alone.

2 Related Work

EEG domain generalization and signal-level invariance. Domain generalization (DG) aims to improve robustness to unseen domains by learning representations that suppress domain-specific variation (Musellim et al., 2022). In EEG modeling, this goal has been pursued through preprocessing, handcrafted features, pretraining, data augmentation (Tsoi et al., 2024), adversarial alignment, contrastive learning, prototype learning (Han et al., 2025), distribution matching, and invariant risk minimization (Rommel et al., 2022; Rangwani et al., 2022; Ju et al., 2025; Del Pup et al., 2025; Zhao et al., 2025; Huang et al., 2025; Xu et al., 2025; Lim et al., 2024). These methods are valuable for cross-subject, cross-session, and cross-site variability (Kim et al., 2025), but they typically define invariance at the level of signals or feature distributions. Clinical cross-hospital shifts are more structured: acquisition protocols, montages, hardware, patient populations, and reporting conventions may vary substantially across institutions (Wagh et al., 2022; Poziomska et al., 2025). Under such shifts, low-level statistical alignment does not determine which variations are clinically irrelevant and which encode pathological content. TG-IL addresses this abstraction mismatch by defining

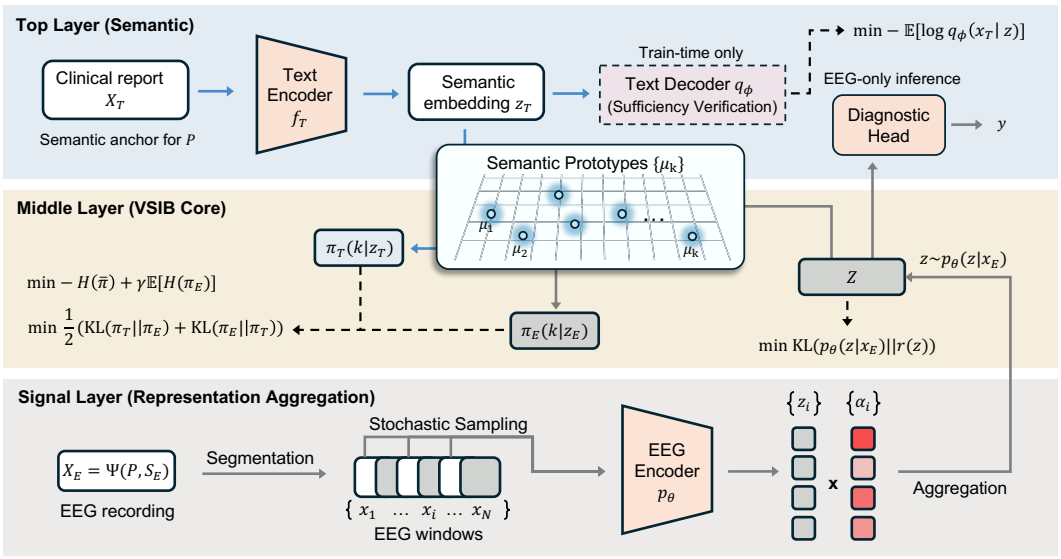


Figure 1: Overview of TG-IL. EEG windows are encoded and aggregated into a stochastic recording-level latent representation. During representation learning, clinician-authored reports provide training-time semantic constraints through a variational semantic bottleneck, prototype-guided EEG–text alignment, and report reconstruction for semantic sufficiency. Diagnostic labels and hospital identifiers are not used to learn the TG-IL representation. For downstream evaluation, all text-side modules are removed and a source-label diagnostic head is trained on EEG representations; inference on unseen hospitals is EEG-only.

the target of invariance at the level of report-derived clinical semantics rather than low-level EEG statistics.

Information bottleneck and invariant learning. The information bottleneck (IB) principle learns compact representations that preserve relevant information while discarding nuisance variability (Tishby et al., 2000; Alemi et al., 2016). IB and related invariant-learning methods have been used for robust prediction, domain generalization, and EEG representation learning (Liao et al., 2020; Wang et al., 2023; Yang et al., 2025; Gao et al., 2025; Liu et al., 2025). Classical formulations typically define relevance through task labels, environment annotations, or prediction objectives (Mahajan et al., 2021; Sheth et al., 2022; Lyu, 2025), including DG methods that combine invariance principles with IB-style compression using labels or environments as relevance signals (Ahuja et al., 2021; Li et al., 2022). These signals can be insufficient in clinical EEG, where diagnostic labels are often sparse, noisy, and too coarse to capture the semantic structure of a full recording. TG-IL instead uses clinician-authored report semantics as the relevance signal for learning EEG representations, turning IB-style compression into a training-time semantic regularizer for EEG-only cross-hospital deployment.

Report- and text-guided clinical representation learning. Training-time report supervision has precedents in medical image representation learning, including ConVIRT (Zhang et al., 2022) and GLoRIA (Huang et al., 2021); our use-inspired contribution is the clinical EEG problem framing and deployment-driven integration, rather than a new VIB formulation or the first use of training-time text. Recent multimodal and auxiliary-supervision approaches use text or other modalities to enrich neural representations (Jang et al., 2026), often by treating semantic outputs as prediction targets (Heo et al., 2026) or by requiring auxiliary modalities during inference (Chen et al., 2025; Yin et al., 2025). Clinical EEG reports are different from dense annotations: they summarize long, heterogeneous recordings into high-level impressions and reflect the abstraction process used by human EEG readers (Tatum IV et al., 2016; Acharya & Acharya, 2022; Grant et al., 2014; Tveit et al., 2023). However, raw report wording may also contain site-specific templates, terminology, and reporting preferences. TG-IL therefore uses report-derived semantic representations as approximate training-time anchors rather than raw lexical matching, and a lightweight report decoder is used only as a training-time sufficiency regularizer. The goal is not report generation or multimodal inference, but EEG-only representation learning that preserves clinically meaningful semantics under cross-hospital deployment shifts.

3 Method

We propose *Text-Guided Invariant Learning* (TG-IL), a training-time report-guided framework for learning EEG-only representations under cross-hospital deployment shifts. TG-IL uses clinician-authored reports as recording-level semantic anchors during representation learning, but removes all text-side modules at inference. The overall framework is shown in Fig. 1, and its design follows the deployment constraints summarized in Table 1.

3.1 Deployment Setting and Notation

We consider source hospital data $\mathcal{D}_{\mathrm{s}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ and an unseen target hospital $\mathcal{D}_{\mathrm{t}}$ used only for evaluation. The case $m = 1$ corresponds to single-source zero-shot transfer, while $m > 1$ corresponds to multi-source leave-one-hospital-out generalization. Each source recording is represented as (X_E, x_T, y), where $X_E = \{x_1, \ldots, x_N\}$ is a bag of EEG windows, x_T is the corresponding clinician-authored report, and y is a downstream diagnostic label.

The target deployment setting is strictly EEG-only: at inference, the model receives only X_E and does not use target-domain calibration data, target labels, hospital identifiers, or clinical reports. Accordingly, TG-IL learns representations from source-hospital EEG–report pairs, but does not use diagnostic labels or hospital identifiers during representation learning. Diagnostic labels are introduced only after representation learning to train a source-only diagnostic head. We view each recording as reflecting pathological factors P and EEG-side site-specific nuisance factors S_E, written informally as $X_E = \Psi(P, S_E)$. The goal is to preserve information about P while suppressing dependence on S_E.

Since clinical reports summarize findings over entire recordings, they provide weak recording-level semantic supervision rather than window-level annotations. We do not assume raw report text is domain-invariant; instead, TG-IL relies on report-derived semantic representations as approximate anchors, an assumption examined empirically in Sec. 4.5.1 and discussed further in Appendix B.1.

3.2 Recording-Level Semantic Bottleneck

Because reports describe full recordings rather than individual EEG windows, TG-IL first constructs a stochastic recording-level representation. At each training iteration, we sample a bag $B_M(X_E) = \{x_{i_1}, \ldots, x_{i_M}\}$ of M windows from recording X_E. Each sampled window is encoded independently as $h_i = f_\theta(x_i)$ and aggregated with attention pooling:

$$\alpha_i = \frac{\exp(g(h_i))}{\sum_{j=1}^{M} \exp(g(h_j))}, \qquad u_E = \sum_{i=1}^{M} \alpha_i h_i, \tag{1}$$

where $g(\cdot)$ is a lightweight scoring network. The aggregated representation parameterizes a stochastic bottleneck,

$$z \sim p_\theta(z \mid B_M(X_E)) = \mathcal{N}\big(\mu_\theta(u_E), \mathrm{diag}(\sigma^2_\theta(u_E))\big)\,. \tag{2}$$

For brevity, we write this induced posterior as $p_\theta(z \mid X_E)$ when the sampled bag is clear from context. Semantic constraints are imposed only on this recording-level latent variable z, matching the granularity of clinical reports.

TG-IL is motivated by a semantic information bottleneck objective:

$$\max_{\theta} \; I(Z; X_T) - \beta I(Z; X_E), \tag{3}$$

where the first term encourages sufficiency for report-level clinical semantics and the second term discourages retention of instance- and site-specific variability. In practice, we optimize a variational surrogate of Eq. (3): a text decoder $q_\phi(x_T \mid z)$ reconstructs the processed report as a sufficiency objective, while a prior $r(z) = \mathcal{N}(0, I)$ yields a KL compression penalty. This reconstruction serves only as a training-time check that the compressed EEG latent remains sufficient for report-level clinical semantics.

The full variational derivation is provided in Appendix A, with implementation details and design rationale in Appendix B.

3.3 Prototype-Guided EEG–Report Alignment

Clinical EEG semantics are many-to-one: recordings with different waveforms, montages, or patient characteristics may correspond to the same high-level interpretation. Rather than aligning individual EEG and report embeddings one-to-one, TG-IL aligns their distributions over shared semantic prototypes. Let $\{\mu_k\}_{k=1}^{K}$ denote learnable prototypes shared by EEG and report representations. For an EEG latent z and a report embedding z_T, we compute prototype assignment distributions

$$\pi_m(k|z_m) = \frac{\exp(\langle \hat{z}_m, \hat{\mu}_k \rangle / \tau)}{\sum_{j=1}^{K} \exp(\langle \hat{z}_m, \hat{\mu}_j \rangle / \tau)}, \qquad m \in \{E, T\}, \tag{4}$$

where $z_E = z$, $\langle \hat{z}_m, \hat{\mu}_k \rangle$ denote the dot product between ℓ_2-normalized vectors, and τ is a temperature parameter. The EEG and report assignments are aligned with a symmetric KL loss:

$$\mathcal{L}_{\text{proto}} = \frac{1}{2} \left[\mathrm{KL}(\pi_T \| \pi_E) + \mathrm{KL}(\pi_E \| \pi_T) \right]. \tag{5}$$

To avoid degenerate prototype usage, we use a diversity regularizer

$$\mathcal{L}_{\text{div}} = -H(\bar{\pi}_E) + \gamma \frac{1}{B} \sum_{b=1}^{B} H(\pi_E^{(b)}), \qquad \bar{\pi}_E = \frac{1}{B} \sum_{b=1}^{B} \pi_E^{(b)}, \tag{6}$$

which encourages batch-level prototype utilization while discouraging overly diffuse instance-level assignments.

3.4 Training Objective and EEG-Only Evaluation

The final TG-IL objective is

$$\mathcal{L}_{\text{TG-IL}} = -\mathbb{E}_{p_\theta(z|X_E)}[\log q_\phi(x_T \mid z)] + \beta \, \mathrm{KL}(p_\theta(z \mid X_E) \| r(z)) + \lambda \left(\mathcal{L}_{\text{proto}} + \mathcal{L}_{\text{div}} \right). \tag{7}$$

The first term preserves report-level clinical semantics, the second compresses recording- and site-specific variability, and the last two terms impose structured semantic alignment.

Representation learning optimizes Eq. (7) using only source-hospital EEG–report pairs. After representation learning, all text-side modules are discarded, and a lightweight diagnostic head is trained using source-hospital labels only. The resulting EEG-only model is then evaluated directly on unseen target hospitals without target adaptation.

4 Experiments

4.1 Experimental Setup

Deployment-oriented protocol. We evaluate TG-IL under source-only cross-hospital deployment: models are trained and selected using source-hospital data, then evaluated directly on unseen hospi-

Table 2: Single-source zero-shot cross-hospital generalization for normal vs. abnormal EEG classification. For each training hospital, models are trained on that hospital only and evaluated directly on unseen hospitals without adaptation. We report AUROC and balanced accuracy (Bal. Acc) (%) for each unseen evaluation hospital. Summary statistics (Avg/Worst/Std) are computed over unseen hospitals.

Training Hospital	Method	Eval (unseen): BIDMC		Eval (unseen): BCH		Eval (unseen): TUAB		Unseen Summary		
		AUROC	Bal. Acc	AUROC	Bal. Acc	AUROC	Bal. Acc	Avg	Worst	Std
BIDMC	ERM (supervised)	–	–	74.28±1.83	68.11±1.42	73.96±2.17	72.48±1.08	74.12	73.96	0.16
	Report-only aux	–	–	76.09±1.36	69.71±1.58	75.92±1.84	74.03±1.19	76.01	75.92	**0.09**
	CLIP-style	–	–	78.24±1.74	**72.91**±0.87	79.05±1.29	76.34±1.61	78.65	78.24	0.41
	VIB (supervised)	–	–	75.31±1.25	69.12±1.46	77.62±1.58	75.21±1.17	76.47	75.31	1.16
	TG-IL (ours)	–	–	**81.34**±0.72	70.88±1.15	**83.08**±0.64	**78.96**±0.82	**82.21**	**81.34**	0.87
BCH	ERM (supervised)	73.81±2.05	68.14±1.59	–	–	75.22±1.74	73.28±1.31	74.52	73.81	**0.71**
	Report-only aux	75.03±1.32	69.38±1.47	–	–	76.88±1.76	74.51±1.08	75.96	75.03	0.93
	CLIP-style	77.42±1.86	**73.74**±0.69	–	–	79.14±1.41	76.21±1.55	78.28	77.42	0.86
	VIB (supervised)	74.35±1.49	68.72±1.05	–	–	78.03±1.17	75.05±1.38	76.19	74.35	1.84
	TG-IL (ours)	**80.27**±0.83	72.05±1.23	–	–	**82.18**±0.62	**78.04**±0.88	**81.23**	**80.27**	0.96
TUAB	ERM (supervised)	71.18±1.97	66.24±1.51	71.64±1.84	67.06±1.42	–	–	71.41	71.18	0.23
	Report-only aux	73.57±1.69	68.11±1.28	73.26±1.93	69.02±1.17	–	–	73.42	73.26	0.16
	CLIP-style	75.48±1.43	69.68±1.57	75.11±1.32	70.27±1.54	–	–	75.30	75.11	0.19
	VIB (supervised)	72.36±1.55	67.34±1.14	74.21±1.19	69.61±1.29	–	–	73.29	72.36	0.93
	TG-IL (ours)	**78.92**±0.73	**71.18**±0.84	**78.65**±0.61	**72.31**±0.69	–	–	**78.79**	**78.65**	**0.14**

tals without target-domain calibration, target labels, hospital identifiers, or clinical reports. Reports are used only during source-hospital representation learning and all text-side modules are removed at inference.

Data and tasks. We use three clinical EEG domains: BIDMC and BCH from the Harvard Electroencephalography Database (Sun et al., 2025), and TUAB from the TUH Abnormal EEG Corpus (Harati et al., 2014). We evaluate two downstream tasks: recording-level normal/abnormal classification and low-resource abnormal subtype classification on abnormal recordings.

TUAB uses a fixed bipolar montage without raw reference or EOG channels, so the re-referencing and EOG regression used for BIDMC/BCH are not applied to TUAB. A common-bipolar control with both operations disabled across hospitals preserves the performance and leakage-reduction trends (Appendix I.3).

Evaluation protocols. We consider two cross-hospital settings. In single-source zero-shot transfer, models are trained on one hospital and evaluated on each remaining unseen hospital. In multi-source leave-one-hospital-out (LOHO) evaluation, models are trained on two hospitals and evaluated on the held-out hospital. All model selection uses source-only validation splits. Details of dataset construction, data preprocessing, and report processing are provided in Appendix C; cross-hospital protocols and baseline implementations are provided in Appendix D.

Metrics and implementation. For normal/abnormal classification, we report AUROC and balanced accuracy, together with average, worst-case, and standard deviation across unseen target hospitals. For subtype classification, we report macro one-vs-rest AUROC, macro-F1, and worst-case AUROC. Unless otherwise stated, main-text results use LaBraM-B (Jiang et al., 2024) as the EEG backbone and MedCPT (Jin et al., 2023) as the report encoder. Model configuration and optimization details are provided in Appendix F; results with alternative EEG backbones and text encoders are reported in Appendix G. Code is available at `https://github.com/Kang1121/TG-IL`.

4.2 EEG-only Cross-Hospital Generalization

We first ask whether semantic-level constraints improve the most stringent deployment setting: training from a single source hospital and direct transfer to unseen hospitals. Table 2 summarizes single-source zero-shot cross-hospital generalization results across all train→test directions. Across source hospitals, TG-IL generally improves average and worst-case AUROC across source hospitals, with smaller seed-level variation in most train→test AUROC estimates. These results suggest that report-guided semantic constraints are most beneficial under deployment shifts where source-trained models are vulnerable to hospital-specific signal variation.

In contrast, end-to-end ERM models trained directly on diagnostic labels exhibit pronounced performance degradation on unseen hospitals, highlighting their sensitivity to site-specific signal characteristics. Generally, TG-IL improves direct target-hospital performance relative to ERM. Standalone foundation-model and supervision-matched controls (Appendices I.1 and I.5) show contributions from both report supervision and the IB/prototype design; the full gain over EEG-only ERM should not be attributed to the invariance objective alone.

Table 3: LOHO evaluation under multi-source DG.

Method	TUAB	BCH	BIDMC	Avg	Worst
ERM	74.12±1.03	75.91±0.94	78.23±0.87	76.09	74.12
VIB	75.02±0.92	77.14±0.83	79.08±0.76	77.08	75.02
IRM	76.09±1.02	78.04±0.91	80.17±0.82	78.10	76.09
GroupDRO	77.08±0.88	79.22±0.81	**81.31**±0.69	79.20	77.08
CORAL	75.06±1.01	76.93±0.97	79.11±0.85	77.03	75.06
MMD	74.03±1.12	75.88±1.04	78.04±0.93	75.98	74.03
DANN	76.02±1.01	78.12±0.92	80.09±0.81	78.08	76.02
TG-IL	**78.41**±0.61	**80.27**±0.63	80.96±0.84	**79.88**	**78.41**

We further report results under the standard multi-source LOHO protocol in Table 3. TG-IL obtains the best average and worst-case AUROC, while being competitive on each held-out hospital.

4.3 Low-resource Transfer of Fine-grained Clinical Semantics

We next extend evaluation from binary abnormality detection to fine-grained abnormal subtype classification to assess semantic transfer under domain shift. To test whether TG-IL transfers *fine-grained* clinical semantics beyond coarse abnormality detection, we introduce an additional downstream task: abnormal subtype classification on abnormal recordings only. Subtype labels are extracted from impression text using a deterministic rule-based pipeline (Appendix E).

After representation learning with EEG–report supervision, we freeze the EEG encoder and train a lightweight linear probe using only 10% of labeled source recordings. The probe is then evaluated unchanged on unseen hospitals. We report macro OvR AUROC, macro-F1, and worst-case AUROC across target hospitals (see Appendix E.8 for details).

Table 4: Cross-hospital abnormal subtype classification under low-resource linear probing.

Method	Macro AUROC	Macro F1	Worst AUROC
ERM	61.92±0.87	49.83±1.14	57.64±1.06
Report aux	64.37±0.82	52.41±1.08	60.18±0.97
VIB	65.28±0.79	53.76±1.21	61.09±1.13
TG-IL	**68.04**±0.74	**56.91**±1.03	**64.87**±0.92

As shown in Table 4, TG-IL consistently improves cross-hospital subtype discrimination and worst-case robustness over label-supervised ERM and compression-based baselines, supporting transfer of report-derived clinical semantic structure. Because supervision and subtype labels derive from the same reports, this is not an independent clinical-annotation evaluation; full-label, explicit-label-removal, and disaggregated results are provided in Appendices I.4, I.6, and I.7.

4.4 Practical Deployment Constraints

A central motivation of TG-IL is to use clinician-authored reports as training-time semantic supervision without requiring them in deployment. We therefore evaluate two practical constraints that commonly arise in multi-center clinical EEG modeling: incomplete report availability during training and deployment-time efficiency.

Partial report availability. Although clinical EEG recordings are often accompanied by reports, fully paired EEG–report supervision may not always be available in practice. Some reports may be missing, delayed, incomplete, or inaccessible due to institutional data constraints. To test whether TG-IL requires exhaustive paired supervision, we randomly retain reports for only a fraction of source-hospital training recordings and remove report-based losses for the remaining recordings. The downstream EEG-only evaluation protocol is unchanged.

As shown in Table 5, average AUROC decreases smoothly as report coverage is reduced, while worst-case AUROC remains broadly stable under moderate report subsampling. Even with 25–50% report availability, TG-IL retains most of its cross-hospital robustness. The 0% row is included as an EEG-only lower-bound reference and corresponds to the ERM baseline. The gap between 10% and 0% suggests that even sparse report supervision provides useful semantic constraints beyond EEG-only training. These results indicate that TG-IL can be viewed as a training-time semantic regularizer that leverages available paired subsets, rather than a method that requires every training recording to have a clinical report.

Table 5: Robustness to partial report availability during training.

Fraction (%)	Avg AUROC ↑	Worst AUROC ↑
100	**80.74**±0.64	78.65±0.61
75	80.11±0.67	78.92±0.58
50	79.36±0.72	77.98±0.66
25	78.21±0.78	76.74±0.71
10	76.84±0.83	75.21±0.76
0 (ERM)	73.35±1.45	71.18±1.97

Deployment-time efficiency. We distinguish the training composite model from the exported deployment model: TG-IL uses report-side modules only during representation learning, whereas the deployed model is strictly EEG-only.

Table 6 shows that the additional cost of TG-IL is primarily a training-time cost. In the exported deployment checkpoint, TG-IL has the same number of inference-time parameters as the EEG-only ERM baseline (6.06M), does not require text at inference, and has nearly identical per-recording inference time. The larger training-time disk size includes the frozen MedCPT encoder, while the additional trainable parameters mainly come from the text decoder and lightweight semantic alignment modules, which are removed at deployment.

4.5 Evidence for Semantic-level Invariance

4.5.1 Diagnostic Evidence for Cross-Hospital Semantic Invariance

TG-IL relies on a weaker assumption than literal report invariance: raw report wording may vary across hospitals, but report-derived semantic representations should preserve clinical meaning while reducing institution-specific style. We therefore compare raw lexical features, EEG latent representations, and report semantic embeddings using cross-hospital descriptor distances, semantic retrieval,

Table 6: Training-time and deployment-time cost comparison. Text-side modules are used only during training and are removed from the exported deployment checkpoint. As a result, TG-IL preserves an EEG-only inference pipeline with the same number of deployed parameters as the ERM baseline.

Method	Deployment / inference-time			Training-time only			
	Disk (MB) ↓	Deployed params (M) ↓	Infer. time / rec. (s) ↓	Disk (MB) ↓	Trainable params (M) ↓	Peak GPU mem. (GB) ↓	Train time / epoch (min) ↓
ERM	92.14	6.06	6.22	92.14	5.96	13.06	54.62
TG-IL (MedCPT, FP16)	92.14	6.06	6.22	328.65	18.14	13.98	61.24

Table 7: Diagnostic cross-hospital semantic invariance analysis. All metrics are computed across different hospitals. Report semantic embeddings exhibit stronger semantic consistency and lower hospital identifiability than both raw lexical features and EEG latent representations.

Representation space	Same-desc. dist. ↓	Diff-desc. dist. ↑	Gap ↑	Sem. Recall@10 ↑	Sem. Purity@10 ↑	Hospital AUROC ↓
Raw lexical	0.43	0.56	0.13	0.41	0.37	0.74
EEG latent (ERM)	0.40	0.52	0.12	0.47	0.45	0.76
Report semantic	**0.31**	**0.58**	**0.27**	**0.63**	**0.58**	**0.52**

and hospital identifiability. The diagnostic uses canonical clinical descriptors extracted from report impressions, including normal, abnormal, slowing, epileptiform activity, seizure, asymmetry, and burst suppression (Appendix F.7).

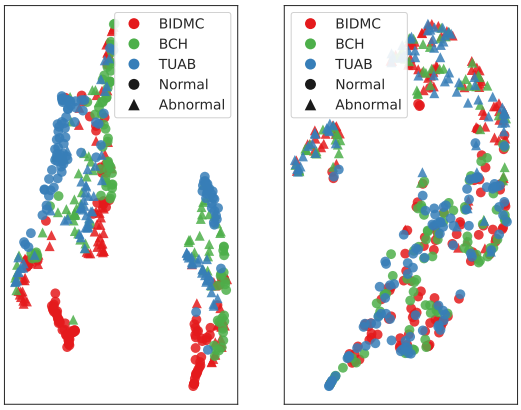
As shown in Table 7, report semantic embeddings achieve the lowest same-descriptor distance, the largest separation gap, the strongest semantic retrieval consistency, and near-chance hospital prediction. In contrast, raw lexical features and EEG latent representations retain substantially more hospital-specific information. These results support the use of report-derived semantic abstractions as relatively stable training signals in this benchmark.

This diagnostic does not imply that clinical reports are universally domain-invariant. Rather, it shows that, in this benchmark, dense report embeddings are less organized by hospital identity while remaining organized by clinically meaningful descriptors. Additional text-side diagnostics are provided in Appendix G.3. A content-preserving BWH template transformation also leaves performance essentially unchanged (Appendix I.2); this evidence is limited to the tested English template variation and does not establish universal report invariance.

4.5.2 Domain Leakage Analysis

We next test whether TG-IL reduces hospital-specific information in learned EEG representations. We freeze recording-level EEG representations and train a lightweight multinomial logistic regression classifier to predict hospital identity. Lower hospital-ID prediction indicates less domain leakage.

Figure 2 shows that ERM representations cluster strongly by hospital, whereas TG-IL representations are better mixed across hospitals within semantic classes. Quantitatively, Table 8 shows that TG-IL reduces domain AUROC from 0.83 for ERM to 0.56, substantially closer to chance-level hospital prediction and outperforming report auxiliary and VIB baselines.



(a) ERM (b) TG-IL

Figure 2: UMAP visualization of learned representations on the test set. ERM exhibits hospital-specific clustering, while TG-IL shows better cross-hospital mixing and semantic organization.

Prototype interpretability. We further analyze prototype semantics in Appendix G.4. Representative prototypes are associated with coherent clinical concepts, as illustrated by their report keywords and abnormality purity.

Table 8: Domain leakage analysis. Lower domain AUROC indicates less domain information.

Representation	ERM	Report Aux	VIB	**TG-IL (ours)**
Domain AUROC	0.83	0.68	0.68	**0.56**

Table 9: Ablation under zero-shot cross-hospital evaluation.

Method	**Avg AUROC** ↑	**Worst AUROC** ↑
ERM	73.35±1.45	71.18±1.97
+ Report recon.	75.13±1.06	73.26±1.93
+ VSIB w/o proto.	76.43±1.11	75.10±0.98
+ Proto. w/o div.	79.27±0.82	77.86±0.88
TG-IL	**80.74**±0.64	78.65±0.61

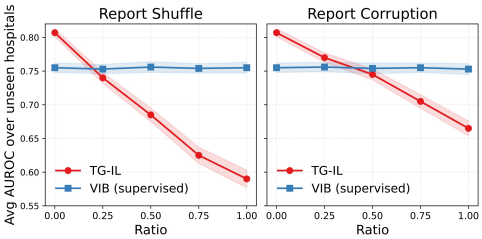


Figure 3: Robustness to disrupted EEG–report correspondence.

4.6 Mechanistic Analysis

4.6.1 Ablation Study

We ablate the main components of TG-IL under the same zero-shot cross-hospital protocol, using the same EEG backbone and downstream linear probe.

Table 9 shows that each component contributes to cross-hospital robustness. Report reconstruction provides the largest initial gain over ERM, indicating the value of report-level semantic supervision. The VSIB and prototype alignment further improve average and worst-case AUROC, while the full TG-IL model achieves the best overall and worst-case performance.

4.6.2 Counterexample Study: Semantic Bottleneck vs. Naive Compression

Finally, we test whether the gains come from genuine semantic correspondence rather than the generic regularization effect of stochastic compression. A potential explanation for the robustness gains of TG-IL is that they arise from generic stochastic compression induced by KL regularization, rather than from the proposed *semantic* information bottleneck principle. To disentangle these factors, we compare TG-IL against supervised VIB, an EEG-only label-supervised bottleneck baseline. This baseline tests whether a generic stochastic KL bottleneck, even when trained directly with source diagnostic labels, is sufficient to recover the robustness achieved by report-guided semantic constraints.

Controlled semantic perturbations. Beyond standard zero-shot cross-hospital evaluation, we introduce two targeted interventions that progressively destroy semantic correspondence between EEG recordings and reports: (i) *report shuffle*, which randomly permutes reports within each source hospital; and (ii) *report corruption*, which masks report tokens or replaces the impression section with generic templates.

Figure 3 shows average AUROC on unseen hospitals as a function of perturbation severity. Under clean training conditions (ratio $= 0$), TG-IL substantially outperforms the naive bottleneck baseline, supporting the benefit of semantic supervision. As semantic correspondence is progressively destroyed, TG-IL degrades sharply, whereas VIB (supervised) remains largely insensitive to perturbations. Under severe disruption, TG-IL can even underperform the naive bottleneck, consistent with forced alignment to incorrect or uninformative semantic signals.

4.7 Additional Analyses and Diagnostics

We conduct complementary analyses to further examine the semantic structure and robustness of the learned representations. These include text-side diagnostics and prototype semantic interpretability. Detailed results are provided in Appendix G; limitations are discussed in Appendix H.

5 Conclusion

We introduced TG-IL, a training-time report-guided framework for learning EEG-only representations under cross-hospital clinical deployment shifts. TG-IL uses clinician-authored reports as semantic anchors to align invariance with the abstraction level used in clinical EEG interpretation

rather than only with low-level signal statistics. Across three hospitals, TG-IL improves zero-shot generalization and worst-case robustness, while reducing domain leakage relative to signal-level invariant learning and generic bottleneck baselines. These results support semantic-level constraints as a useful principle for robust clinical EEG modeling under realistic institutional shifts.

Acknowledgments and Disclosure of Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University); No. RS-2024-00457882, AI Research Hub Project; No. RS-2025-02304828, Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents; No. RS-2026-25522672, Development of Unified Reasoning Technology Mimicking Human Cognition for Hierarchical Understanding and Unbounded Problem Solving; and No. RS-2024-00336673, AI Technology for Interactive Communication of Language Impaired Individuals).

The authors declare no competing interests.

References

Acharya, J. N. and Acharya, V. J. Eeg reporting: What's the score? *Clinical Neurophysiology Practice*, 7:264, 2022.

Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.-C., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., and Marchand, M. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Apicella, A., Arpaia, P., D'Errico, G., Marocco, D., Mastrati, G., Moccaldi, N., and Prevete, R. Toward cross-subject and cross-session generalization in eeg-based emotion recognition: Systematic review, taxonomy, and methods. *Neurocomputing*, 604:128354, 2024.

Bang, J.-S. and Lee, S.-W. Interpretable convolutional neural networks for subject-independent motor imagery classification. In *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*, pp. 1–5. IEEE, 2022. doi: 10.1109/BCI53720.2022.9734822.

Chen, S., Li, B., He, C., Li, D., Wu, M., Shen, X., Wang, S., Wei, X., Wang, X., Wu, H., and Liu, Q. Chineseeeg-2: An eeg dataset for multimodal semantic alignment and neural decoding during reading and listening, 2025. URL `https://arxiv.org/abs/2508.04240`.

Chen, X., Teng, X., Chen, H., Pan, Y., and Geyer, P. Toward reliable signals decoding for electroencephalogram: A benchmark study to eegnex. *Biomedical Signal Processing and Control*, 87: 105475, 2024.

Del Pup, F., Zanola, A., Tshimanga, L. F., Bertoldo, A., and Atzori, M. The more, the better? evaluating the role of eeg preprocessing for deep learning applications. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2025.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

Gao, X., Wang, D., Zhao, Y., and Wang, X. Emoeeg: A transferable generalist framework for eeg emotion recognition via information bottleneck theory. *Pattern Recognition*, pp. 112775, 2025.

Grant, A. C., Abdel-Baki, S. G., Weedon, J., Arnedo, V., Chari, G., Koziorynska, E., Lushbough, C., Maus, D., McSween, T., Mortati, K. A., et al. Eeg interpretation reliability and interpreter confidence: a large single-center study. *Epilepsy & Behavior*, 32:102–107, 2014.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Han, D.-K., Lee, M., and Lee, S.-W. Multi-layer prototype learning with dirichlet mixup for open-set eeg recognition. *Expert Systems with Applications*, 266:126047, 2025. doi: 10.1016/j.eswa.2024.126047.

Harati, A., Lopez, S., Obeid, I., Picone, J., Jacobson, M., and Tobochnik, S. The tuh eeg corpus: A big data resource for automated eeg interpretation. In *2014 IEEE signal processing in medicine and biology symposium (SPMB)*, pp. 1–5. IEEE, 2014.

Heo, J.-H., Wisniewska, A. J., Lee, S.-H., and Lee, S.-W. Cross-subject semantic decoding with shared-space alignment for generalized neural representation learning, 2026. URL `https://arxiv.org/abs/2607.19394`.

Huang, H., Si, X., Han, Y., and Ming, D. A novel conditional adversarial domain adaptation network for eeg cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 2025.

Huang, S.-C., Shen, L., Lungren, M. P., and Yeung, S. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021. URL `https://openaccess.thecvf.com/content/ICCV2021/html/Huang_GLoRIA_A_Multimodal_Global-Local_Representation_Learning_Framework_for_Label-Efficient_Medical_ICCV_2021_paper.html`.

Jang, H.-J., Shin, H.-B., and Yin, K. Cross-modal consistency-guided active learning for affective BCI systems. In *2026 14th International Conference on Brain-Computer Interface (BCI)*, pp. 1–4. IEEE, 2026. doi: 10.1109/BCI69045.2026.11435036.

Jiang, W.-B., Zhao, L.-M., and Lu, B.-L. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024.

Jin, Q., Kim, W., Chen, Q., Comeau, D. C., Yeganova, L., Wilbur, W. J., and Lu, Z. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Ju, X., Wu, X., Dai, S., Li, M., and Hu, D. Domain adversarial learning with multiple adversarial tasks for eeg emotion recognition. *Expert Systems with Applications*, 266:126028, 2025.

Kim, D.-Y., Han, D.-K., Jeong, J.-H., and Lee, S.-W. Calibration-free driver drowsiness classification with prototype-based multi-domain mixup. *IEEE Transactions on Intelligent Transportation Systems*, 26(3):2955–2966, 2025. doi: 10.1109/TITS.2024.3522308.

Li, B., Shen, Y., Wang, Y., Zhu, W., Li, D., Keutzer, K., and Zhao, H. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7399–7407, 2022.

Liao, H., Xu, J., and Yu, Z. Novel convolutional neural network with variational information bottleneck for p300 detection. *Entropy*, 23(1):39, 2020.

Lim, E. Y., Yin, K., Shin, H.-B., and Lee, S.-W. Baseline-guided representation learning for noise-robust EEG signal classification. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4. IEEE, 2024. doi: 10.1109/EMBC53108.2024.10781970.

Liu, P., Chen, C. L. P., He, Y., and Zhang, T. Cria: A cross-view interaction and instance-adapted pre-training framework for generalizable eeg representations, 2025. URL `https://arxiv.org/abs/2506.16056`.

Lyu, R. Deep learning approaches for eeg-based healthcare applications: A comprehensive review. *Frontiers in Human Neuroscience*, 19:1689073, 2025.

Mahajan, D., Tople, S., and Sharma, A. Domain generalization using causal matching, 2021. URL https://arxiv.org/abs/2006.07500.

Musellim, S., Han, D.-K., Jeong, J.-H., and Lee, S.-W. Prototype-based domain generalization framework for subject-independent brain-computer interfaces. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 711–714, 2022. doi: 10.1109/EMBC48229.2022.9871434. URL https://arxiv.org/abs/2204.07358.

Poziomska, M., Dovgialo, M., Olbratowski, P., Niedbalski, P., Ogniewski, P., Zych, J., Rogala, J., and Żygierewicz, J. Quantity versus diversity: Influence of data on detecting eeg pathology with advanced ml models. *Neural Networks*, pp. 108073, 2025.

Rangwani, H., Aithal, S. K., Mishra, M., Jain, A., and Radhakrishnan, V. B. A closer look at smoothness in domain adversarial training. In *International conference on machine learning*, pp. 18378–18399. PMLR, 2022.

Rommel, C., Paillard, J., Moreau, T., and Gramfort, A. Data augmentation for learning predictive models on eeg: a systematic comparison. *Journal of Neural Engineering*, 19(6):066020, 2022.

Rosenfeld, E., Ravikumar, P., and Risteski, A. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Sheth, P., Moraffah, R., Candan, K. S., Raglin, A., and Liu, H. Domain generalization – a causal perspective, 2022. URL https://arxiv.org/abs/2209.15177.

Sun, B., Feng, J., and Saenko, K. Correlation alignment for unsupervised domain adaptation. In *Domain adaptation in computer vision applications*, pp. 153–171. Springer, 2017.

Sun, C., Jing, J., Turley, N., Alcott, C., Kang, W.-Y., Cole, A. J., Goldenholz, D. M., Lam, A., Amorim, E., Chu, C., et al. Harvard electroencephalography database: A comprehensive clinical electroencephalographic resource from four boston hospitals. *Epilepsia*, 2025.

Tatum IV, W. O., Selioutski, O., Ochoa, J. G., Clary, H. M., Cheek, J., Drislane, F. W., and Tsuchida, T. N. American clinical neurophysiology society guideline 7: guidelines for eeg reporting. *The Neurodiagnostic Journal*, 56(4):285–293, 2016.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Tsoi, P., Kweon, Y.-S., and Lee, S.-W. Clustering-based augmentation for effective self-supervised learning in sleep staging. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2024. doi: 10.1109/IJCNN60899.2024.10650938.

Tveit, J., Aurlien, H., Plis, S., Calhoun, V. D., Tatum, W. O., Schomer, D. L., Arntsen, V., Cox, F., Fahoum, F., Gallentine, W. B., et al. Automated interpretation of clinical electroencephalograms using artificial intelligence. *JAMA neurology*, 80(8):805–812, 2023.

Wagh, N., Wei, J., Rawal, S., Berry, B. M., and Varatharajah, Y. Evaluating latent space robustness and uncertainty of eeg-ml models under realistic distribution shifts. *Advances in Neural Information Processing Systems*, 35:21142–21156, 2022.

Wang, G., Liu, W., He, Y., Xu, C., Ma, L., and Li, H. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

Wang, J., Ge, X., Shi, Y., Sun, M., Gong, Q., Wang, H., and Huang, W. Dual-modal information bottleneck network for seizure detection. *International journal of neural systems*, 33(01):2250061, 2023.

Xu, C., Song, Y., Zheng, Q., Wang, Q., and Heng, P.-A. Unsupervised multi-source domain adaptation via contrastive learning for eeg classification. *Expert Systems with Applications*, 261:125452, 2025.

Yang, Y., Duan, L., Hou, K., Kang, Z., Zhang, X., and Hu, B. Fddgnet: An information bottleneck-inspired feature disentanglement network for cross-subject eeg-based emotion recognition. *Neurocomputing*, pp. 132368, 2025.

Yin, K. and Shin, H.-B. NeuroLex: A lightweight domain language model for EEG report understanding and generation. In *2026 14th International Conference on Brain-Computer Interface (BCI)*, pp. 1–4. IEEE, 2026. doi: 10.1109/BCI69045.2026.11435041.

Yin, K., Lim, E. Y., and Lee, S.-W. Gitgan: Generative inter-subject transfer for eeg motor imagery analysis. *Pattern Recognition*, 146:110015, 2024. doi: 10.1016/j.patcog.2023.110015.

Yin, K., Shin, H.-B., Li, D., and Lee, S.-W. EEG-based multimodal representation learning for emotion recognition. In *2025 13th International Conference on Brain-Computer Interface (BCI)*, pp. 1–4. IEEE, 2025. doi: 10.1109/BCI65088.2025.10931743.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25. PMLR, 2022. URL `https://proceedings.mlr.press/v182/zhang22a.html`.

Zhao, Q., Jiang, H., Wu, Z., Zhang, L., Cui, K., Zheng, K., Liu, J., Cai, R., Zhao, M., Tian, F., et al. E-dann: An enhanced domain adaptation network for audio-eeg feature decoupling in explainable depression recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2025.

A Variational Semantic Information Bottleneck Derivation

This appendix provides a rigorous derivation of the variational bounds used to optimize the proposed *bag-level* Variational Semantic Information Bottleneck (VSIB) objective. The derivation follows standard arguments from the Variational Information Bottleneck (VIB) literature (Alemi et al., 2016), with adaptations to the multiple-instance (bag-level) setting.

A.1 Problem Setup

Let $X_E = \{x_1, \ldots, x_N\}$ denote a bag of EEG segments extracted from a single recording, and let x_T denote the corresponding clinical report that provides coarse-grained semantic supervision at the recording level. The latent representation Z is a recording-level variable produced by an EEG encoder $p_\theta(z \mid X_E)$.

We view EEG signals and reports as two imperfect observations of shared underlying clinical factors. Let P denote clinically meaningful physiological or pathological factors, S_E denote signal-side site-specific nuisance factors such as hardware, montage, referencing, and preprocessing conventions, and S_T denote report-side style or template variation. Informally,

$$X_E = \Psi(P, S_E), \qquad X_T = \Omega(P, S_T). \tag{8}$$

TG-IL aims to learn a representation that preserves information about P while reducing dependence on nuisance variation. Importantly, the inference encoder is constrained to be EEG-only:

$$p_\theta(z \mid X_E, X_T) = p_\theta(z \mid X_E), \tag{9}$$

or equivalently, under the inference model, Z is conditionally independent of the report given the EEG bag. Reports are therefore used only to define a training-time relevance signal for the representation; they are not inputs to the deployed model.

The semantic information bottleneck objective is written as the maximization problem

$$\max_\theta \ \mathcal{J}_{\mathrm{VSIB}} = I(Z; X_T) - \beta I(Z; X_E), \tag{10}$$

where $I(Z; X_T)$ encourages sufficiency for report-level clinical semantics and $I(Z; X_E)$ penalizes retention of instance-level, recording-specific, and site-specific nuisance information. Equivalently, the corresponding minimization loss is

$$\mathcal{L}_{\mathrm{VSIB}} = -I(Z; X_T) + \beta I(Z; X_E). \tag{11}$$

A.2 Variational Lower Bound on Semantic Sufficiency

We first derive a variational lower bound on the mutual information term $I(Z; X_T)$. By definition,

$$I(Z; X_T) = \mathbb{E}_{p(z, x_T)} \left[\log \frac{p(x_T \mid z)}{p(x_T)} \right] = H(X_T) - H(X_T \mid Z). \tag{12}$$

The true conditional distribution $p(x_T \mid z)$ is generally intractable. We therefore introduce a variational decoder $q_\phi(x_T \mid z)$ to approximate it. Using the non-negativity of the Kullback–Leibler divergence,

$$\mathbb{E}_{p(z)} \mathrm{KL}\big(p(x_T \mid z) \,\|\, q_\phi(x_T \mid z)\big) \geq 0, \tag{13}$$

which implies

$$\mathbb{E}_{p(z, x_T)}[\log p(x_T \mid z)] \geq \mathbb{E}_{p(z, x_T)}[\log q_\phi(x_T \mid z)]. \tag{14}$$

Substituting this inequality into the definition of mutual information yields the following variational lower bound:

$$I(Z; X_T) \geq \mathbb{E}_{p(z, x_T)}[\log q_\phi(x_T \mid z)] + H(X_T). \tag{15}$$

Since the entropy term $H(X_T)$ is constant with respect to the model parameters, maximizing $I(Z; X_T)$ motivates maximizing the expected report log-likelihood as a tractable surrogate.

In practice, the joint distribution $p(z, x_T)$ is induced by the data and the encoder, namely

$$p(z, x_T) = \int p_{\text{data}}(X_E, x_T)\, p_\theta(z \mid X_E)\, dX_E. \tag{16}$$

Therefore, the variational lower bound in Eq. (15) can be written as

$$I(Z; X_T) \;\geq\; \mathbb{E}_{p_{\text{data}}(X_E, x_T)} \mathbb{E}_{p_\theta(z|X_E)} \big[\log q_\phi(x_T \mid z) \big] + H(X_T), \tag{17}$$

which corresponds to the semantic sufficiency objective used in the main text. Since $H(X_T)$ is independent of the model parameters, we omit it in the optimization objective.

A.3 Variational Upper Bound on Invariant Compression

The compression term can be written as:

$$I(Z; X_E) = \mathbb{E}_{p_{\text{data}}(X_E)} \mathrm{KL} \left(p_\theta(z \mid X_E) \,\|\, \bar{p}_\theta(z) \right), \tag{18}$$

where

$$\bar{p}_\theta(z) = \int p_\theta(z \mid X_E) p_{\text{data}}(X_E) dX_E \tag{19}$$

is the aggregated posterior induced by the encoder and the data distribution.

Since the aggregated posterior is generally intractable, we introduce a tractable variational prior $r(z)$. For any choice of $r(z)$,

$$\begin{aligned} \mathbb{E}_{p_{\text{data}}(X_E)} \mathrm{KL} \left(p_\theta(z \mid X_E) \,\|\, r(z) \right) &= I(Z; X_E) + \mathrm{KL} \left(\bar{p}_\theta(z) \,\|\, r(z) \right) \\ &\geq I(Z; X_E). \end{aligned} \tag{20}$$

Thus,

$$I(Z; X_E) \leq \mathbb{E}_{p_{\text{data}}(X_E)} \mathrm{KL} \left(p_\theta(z \mid X_E) \,\|\, r(z) \right). \tag{21}$$

In all experiments, we use the standard Gaussian prior $r(z) = \mathcal{N}(0, I)$.

A.4 Resulting Variational Objective

Combining the lower bound on semantic sufficiency with the upper bound on invariant compression yields the tractable objective

$$\max_{\theta, \phi} \; \mathbb{E}_{p_{\text{data}}(X_E, x_T)} \mathbb{E}_{p_\theta(z|X_E)} [\log q_\phi(x_T \mid z)] - \beta \mathbb{E}_{p_{\text{data}}(X_E)} \mathrm{KL} \left(p_\theta(z \mid X_E) \,\|\, r(z) \right). \tag{22}$$

TG-IL augments this variational semantic bottleneck objective with prototype-guided EEG–report alignment and prototype diversity regularization, producing the final objective in Eq. (7).

A.5 Scope of the Information Bottleneck Formulation

The derivation above follows the standard information bottleneck and variational information bottleneck arguments. We do not claim a new information-theoretic bound in isolation. The role of the bottleneck in TG-IL is to instantiate a deployment-motivated semantic regularization principle for clinical EEG: the relevance variable is clinician-authored report semantics available during source-hospital representation learning, rather than a downstream task label or an environment identifier.

The methodological contribution lies in adapting this principle to the clinical EEG deployment setting through recording-level stochastic bags, training-time-only report supervision, and prototype-guided EEG–report semantic alignment. The variational bounds are not assumed to be tight; they serve as a tractable regularization objective whose components are empirically validated through ablations and representation diagnostics.

B Design Rationale and Method Implementation Details

B.1 Semantic Anchor Assumption

TG-IL relies on the assumption that clinician-authored reports provide useful recording-level semantic anchors for EEG representation learning. This assumption should not be interpreted as claiming that raw report text is domain-invariant or that reports provide complete descriptions of the underlying EEG. Instead, the assumption is weaker: after semantic encoding, report-derived representations are expected to be more closely associated with clinically meaningful pathological factors than with low-level acquisition nuisance factors.

Formally, we view an EEG recording as reflecting both pathological factors P and site-specific nuisance factors S_E:

$$X_E = \Psi(P, S_E),$$

where P denotes clinically meaningful factors such as slowing, epileptiform activity, seizures, asymmetry, or normal variants, and S_E denotes acquisition- or institution-specific factors such as hardware, montage, referencing scheme, preprocessing conventions, and local patient mix. The goal is not to remove all information unrelated to the downstream diagnostic label, but to learn a recording-level representation that preserves information about P while reducing dependence on S_E.

Clinical reports are suitable for this role because they summarize long, heterogeneous EEG recordings into high-level impressions rather than dense window-level annotations. This matches the granularity of the desired representation: semantic constraints should be imposed at the recording level, after aggregating information across EEG windows, rather than at the level of individual segments. At the same time, reports are imperfect supervision signals. They may contain site-specific templates, lexical preferences, reporting conventions, omissions, and variable levels of detail.

For this reason, TG-IL does not rely on raw lexical matching as the primary EEG–report alignment mechanism. Although a lightweight decoder reconstructs the processed report as a training-time sufficiency regularizer, the structured semantic alignment is imposed through report-derived dense embeddings and prototype-level assignment distributions. This design reduces reliance on site-specific wording while still using the paired report to verify that the EEG latent retains report-level clinical semantics.

This assumption motivates the main TG-IL design choices: semantic constraints are imposed after recording-level aggregation, text modules are used only during representation learning, and EEG–report alignment is performed through prototype assignment distributions rather than raw lexical or instance-level matching.

The assumption also defines an important limitation. If reports are systematically biased, weakly related to the EEG recording, or dominated by institution-specific templates, report-guided learning may align EEG representations to misleading semantic signals. For this reason, we include text-side sanity checks and report perturbation experiments to distinguish genuine EEG–report semantic correspondence from generic regularization effects.

B.2 Rationale for Recording-Level Stochastic Bag Sampling

Clinical EEG recordings often span long durations and contain heterogeneous temporal patterns, while corresponding clinical reports summarize pathological findings at a coarse semantic level. As a result, there is no deterministic or unique mapping between report semantics and any fixed subset of EEG windows. Treating all segmented windows as equally informative or enforcing deterministic window-to-report alignment may therefore encourage the model to overfit spurious or site-specific temporal patterns.

To address this issue, we adopt a stochastic bag-level sampling strategy during training. At each iteration, a fixed number of EEG windows are randomly sampled from the same recording, while the identical recording-level clinical report is used as the semantic target. Across training iterations, this procedure exposes the model to multiple partial observations of the same underlying pathology, effectively introducing stochasticity at the bag level.

This design can be interpreted as an inductive bias that encourages recording-level semantic consistency across incomplete observations. Rather than memorizing specific window-level patterns, the model is incentivized to extract features that remain stable across different subsets of the same recording. Such stability is particularly desirable under cross-site distribution shifts, where window-level characteristics may vary due to acquisition protocols, hardware differences, or referencing schemes.

Importantly, this strategy does not assume that any individual window is semantically sufficient. Instead, semantic supervision is applied exclusively at the aggregated recording level, aligning naturally with the weak temporal granularity of clinical reports.

B.3 Stochastic Bottleneck Implementation

Given a sampled EEG bag $B_M(X_E) = \{x_{i_1}, \ldots, x_{i_M}\}$ from recording X_E, the EEG encoder first extracts window-level features

$$h_i = f_\theta(x_i), \qquad i = 1, \ldots, M. \tag{23}$$

An attention pooling module produces a deterministic recording-level summary:

$$\alpha_i = \frac{\exp(g(h_i))}{\sum_{j=1}^{M} \exp(g(h_j))}, \qquad u_E = \sum_{i=1}^{M} \alpha_i h_i. \tag{24}$$

The stochastic bottleneck posterior is parameterized as:

$$p_\theta(z \mid B_M(X_E)) = \mathcal{N}\left(\mu_\theta(u_E), \mathrm{diag}(\sigma_\theta^2(u_E))\right). \tag{25}$$

Sampling is implemented with the reparameterization trick:

$$\epsilon \sim \mathcal{N}(0, I), \qquad z = \mu_\theta(u_E) + \sigma_\theta(u_E) \odot \epsilon. \tag{26}$$

For brevity, the main text writes this induced posterior as $p_\theta(z \mid X_E)$.

Following the variational semantic information bottleneck formulation, we approximate the marginal $p(z)$ with a tractable prior $r(z)$. In all experiments, we set $r(z) = \mathcal{N}(0, I)$, yielding a closed-form KL divergence:

$$\mathrm{KL}\big(\mathcal{N}(\mu_\theta, \sigma_\theta^2) \,\|\, \mathcal{N}(0, I)\big) = \frac{1}{2} \sum_{i=1}^{d} \big(\mu_{\theta,i}^2 + \sigma_{\theta,i}^2 - \log \sigma_{\theta,i}^2 - 1\big). \tag{27}$$

This term is averaged over EEG bags in the mini-batch and scaled by the hyperparameter β, producing the invariant compression loss:

$$\mathcal{L}_{\mathrm{IB}} = \beta\, \mathbb{E}_{X_E}\big[\mathrm{KL}(p_\theta(z \mid X_E) \,\|\, r(z))\big], \tag{28}$$

which corresponds to the variational upper bound on $I(Z; X_E)$ in Eq. (20). During optimization, gradients from $\mathcal{L}_{\mathrm{IB}}$ are propagated to the EEG encoder, encouraging it to discard instance-level variability that is not required for semantic sufficiency.

B.4 Prototype-Based Symmetric KL Alignment

The prototype-guided semantic structure alignment described in Sec. 3.3 is instantiated by a shared prototype layer and a symmetric KL divergence between prototype assignment distributions derived from bag-level EEG and text representations.

Given an EEG recording represented as a bag X_E and its associated report x_T, we obtain modality-specific representations

$$z_E = z, \qquad z_T = e_\psi(x_T), \tag{29}$$

where z is the EEG-side bottleneck latent and e_ψ denotes a pretrained biomedical text encoder followed by a lightweight projection head. Unless otherwise stated, we use MedCPT as the default report encoder; alternative biomedical text encoders are evaluated in Appendix G.2.

We maintain a learnable set of K semantic prototypes $\{\mu_k\}_{k=1}^{K}$, with $\mu_k \in \mathbb{R}^d$, shared across modalities. For each embedding, cosine-similarity logits are computed to derive a soft assignment:

$$s_k^E = \frac{\langle \hat{z}_E, \hat{\mu}_k \rangle}{\tau}, \quad s_k^T = \frac{\langle \hat{z}_T, \hat{\mu}_k \rangle}{\tau}, \tag{30}$$

$$\pi_E(k \mid z_E) = \frac{\exp(s_k^E)}{\sum_{j=1}^{K} \exp(s_j^E)}, \quad \pi_T(k \mid z_T) = \frac{\exp(s_k^T)}{\sum_{j=1}^{K} \exp(s_j^T)}, \tag{31}$$

where $\hat{z}$ and $\hat{\mu}_k$ denote ℓ_2-normalized vectors and τ is a temperature hyperparameter.

The prototype alignment loss is implemented as a symmetric KL divergence:

$$\mathcal{L}_{\text{proto}} = \frac{1}{2}\Big(\mathrm{KL}(\pi_T \,\|\, \pi_E) + \mathrm{KL}(\pi_E \,\|\, \pi_T)\Big), \tag{32}$$

as stated in Eq. (5). This loss is averaged over the batch and combined with the diversity regularization term:

$$\mathcal{L}_{\text{align}} = \lambda\big(\mathcal{L}_{\text{proto}} + \mathcal{L}_{\text{div}}\big). \tag{33}$$

Gradients from $\mathcal{L}_{\text{proto}}$ update both modality-specific encoders (except for the frozen text encoder) and the shared prototypes, enforcing semantic consistency at the bag level without requiring explicit clustering or domain labels.

B.5 Report Reconstruction as Semantic Sufficiency Verification

The text decoder $q_\phi(x_T \mid z)$ is used only during representation learning. Its role is to verify that the compressed EEG latent remains sufficient for report-level clinical semantics. This objective corresponds to maximizing the variational lower bound on $I(Z; X_T)$ derived in Appendix A.

We emphasize that the decoder is not intended to be a clinical report-generation system. Generation quality is not used as an evaluation target, and generated reports are never used for downstream diagnostic inference. The decoder provides a training-time semantic sufficiency constraint: if the bottleneck representation discards clinically relevant information, reconstructing the paired report becomes difficult.

At inference time, the report decoder is discarded. The deployed model receives only EEG inputs and does not require clinical reports, report embeddings, or generated text.

B.6 Two-Stage EEG-Only Diagnostic Evaluation

TG-IL separates report-guided representation learning from downstream diagnostic evaluation. In the first stage, the EEG encoder, stochastic bottleneck, report-side modules, and prototype layer are trained using only source-hospital EEG–report pairs by optimizing Eq. (7). Diagnostic labels, hospital identifiers, target-hospital EEG recordings, target reports, and target labels are not used to learn the representation.

In the second stage, all text-side modules are removed, and a lightweight diagnostic head is trained on source-hospital diagnostic labels using the frozen EEG representation. Hyperparameters and checkpoints are selected using source-only validation, and the final model is evaluated directly on unseen target hospitals without target-domain calibration or adaptation.

The exported deployment model contains only the EEG encoder, recording-level aggregation module, bottleneck module, and diagnostic head. Thus, inference is strictly EEG-only and does not require clinical reports, report embeddings, hospital identifiers, or target-domain information. Report-supervised baselines follow the same two-stage diagnostic-head protocol unless otherwise specified, while the ERM baseline is trained directly using source-hospital diagnostic labels under the same train/validation/test splits.

C Dataset Descriptions and Domain Characteristics

C.1 Overview of Multi-Center EEG Domains

This appendix provides detailed descriptions of the multi-center clinical EEG datasets used in our experiments. While Sec. 4.1 summarizes the experimental protocol, we elaborate on dataset-specific characteristics here to clarify the sources of domain shift and to support reproducibility.

Table 10 highlights substantial domain differences across the three hospitals, including patient populations, recording durations, acquisition protocols, and clinical focus. These systematic variations induce pronounced distribution shifts at the signal and representation levels. Despite these differences, all datasets provide supervision at the *recording level* via clinician-authored reports, which motivates our formulation of each EEG recording as a bag of segments with coarse-grained semantic supervision.

Table 10: Overview of multi-center EEG datasets and domain characteristics. Each hospital is treated as a distinct domain due to systematic differences in patient populations, acquisition protocols, and clinical focus.

	BIDMC	**BCH**	**TUAB (TUH)**
Institution type	Adult tertiary-care hospital	Pediatric specialty hospital	Large academic medical center
Patient population	Adult	Pediatric	Mixed (adult-dominant)
Primary clinical focus	Routine clinical EEG	Pediatric neurology	General abnormality screening
Sampling rate (Hz)	200 / 256 / 512	200 / 256	200 / 256 / 400 / 512 / 1,000
Electrode montage	10–20 (clinical variants)	10–20 (pediatric variants)	10–20 / extended variants
Report granularity	Recording-level summary	Recording-level summary	Recording-level summary
Number of recordings	57,552	61,255	3,065
Number of patients	17,578	25,045	2,391
Number of recordings used	4,903	4,844	3,065
Number of reports used	4,903	4,844	3,065

C.2 Beth Israel Deaconess Medical Center (BIDMC)

The BIDMC dataset consists of routine clinical EEG recordings collected at Beth Israel Deaconess Medical Center, a large adult tertiary-care hospital. Recordings are acquired during standard inpatient and outpatient EEG monitoring and are accompanied by clinician-authored reports that summarize salient findings at the recording level.

The dataset primarily reflects adult patient populations and standard clinical workflows. EEG recordings exhibit moderate duration and contain a mixture of normal activity and pathological patterns. Reports provide coarse-grained diagnostic impressions without precise temporal annotations, making the dataset well-suited for recording-level semantic learning.

C.3 Boston Children's Hospital (BCH)

The BCH dataset contains clinical EEG recordings acquired at Boston Children's Hospital, a specialized pediatric medical center. Compared to adult EEG data, pediatric EEG exhibits distinct signal characteristics due to neurodevelopmental factors, age-dependent background rhythms, and differences in clinical indications.

Recordings are paired with clinician-authored reports describing neurological findings at the recording level. As with BIDMC, reports do not provide precise temporal localization of events, but instead summarize clinically relevant observations across the entire recording.

The combination of pediatric patient populations and hospital-specific acquisition protocols introduces a substantial domain shift relative to adult EEG datasets.

C.4 Temple University Hospital Abnormal EEG Corpus (TUAB)

The Temple University Hospital Abnormal EEG Corpus (TUAB) is a large-scale public clinical EEG dataset collected at Temple University Hospital. TUAB was curated to support automated discrimination between normal and abnormal EEG recordings and represents one of the largest publicly available resources for this task.

EEG recordings in TUAB are accompanied by clinician-authored reports that include a high-level impression of whether the recording is within normal limits. While the broader TUH EEG corpus provides event-level annotations for seizure analysis, TUAB is specifically organized around recording-level normal/abnormal labels. In this work, we deliberately restrict supervision to recording-level reports and diagnostic labels, and do not use any event-level or temporal annotations, in order to maintain a consistent supervision setting across all hospitals.

TUAB recordings are typically longer and more heterogeneous than those from BIDMC and BCH, often containing sparse pathological events embedded within extended periods of normal background activity. This characteristic further amplifies the challenge of robust representation learning, as discriminative cues may be temporally diffuse and weakly expressed.

C.5 Clinical Report Processing and EEG–Report Matching

We preprocess raw clinician-authored EEG reports using a deterministic, rule-based pipeline and construct a curated paired EEG–report dataset through systematic matching and quality filtering.

Report preprocessing and selection. For each report, we locate the `FINDINGS` header using a case-insensitive pattern match and extract all subsequent text as the FINDINGS block, treating `IMPRESSION` / `INTERPRETATION` (different section names in each dataset) as a subsection when present. Subsection headers are parsed using the canonical uppercase-header-plus-colon pattern (e.g., `BACKGROUND:`, `SLEEP:`, `ABNORMALITY#1:`), with content assigned until the next header.

Subsection names are normalized by standardizing whitespace and punctuation, removing trailing colons, and collapsing redaction artifacts. Equivalent header variants are merged through a global canonical mapping and subsequently assigned to a fixed set of 13 core clinical categories via comprehensive pattern-based rules. Each report is then converted into a structured representation containing all core categories in a fixed order, with missing categories marked as `NA`. We further apply light text normalization to extracted report content, including lowercase conversion, whitespace canonicalization, and removal of non-informative punctuation, while explicitly preserving medical negation terms (e.g., `no`, `without`, `absent`) to support reliable semantic interpretation.

To ensure reliable semantic supervision, we retain only reports that contain a valid `IMPRESSION` section after preprocessing.

EEG recording filtering. EEG entries are first restricted to recordings explicitly labeled as EEG studies. To ensure sufficient signal duration for representation learning, we further retain only recordings with a minimum length of five minutes.

EEG–report pairing. Processed reports are linked to EEG recordings using de-identified recording-level identifiers provided in the metadata. We retain only the intersection between valid reports and filtered EEG recordings, yielding an initial set of usable EEG–report pairs.

Final dataset curation. From this paired set, we randomly sample up to 5,000 EEG–report pairs per hospital to ensure comparable dataset scale across domains. We further remove pairs whose EEG recordings lack required start/stop recording annotations or contain missing mandatory EEG channels. The remaining pairs constitute the final dataset used for training and evaluation.

C.6 EEG Preprocessing and Window Construction

All hospitals share a common preprocessing pipeline whenever raw channel data are available. We retain the intersection of 19 standard 10–20 EEG channels, together with ocular (LOC/ROC) and reference (A1/A2) channels when present. Signals are re-referenced to linked mastoids (A1/A2) if available, otherwise to average reference. EEG and EOG channels are high-pass filtered at 0.5 Hz, ocular artifacts are removed via linear regression using EOG signals, followed by low-pass filtering at 30 Hz and resampling to 200 Hz. Per-channel robust normalization is applied using median and interquartile range estimated from the source training set only, pooled across source hospitals in the LOHO setting, and then applied unchanged to validation and target recordings.

For TUAB, EEG recordings are provided in a fixed bipolar montage without raw reference or EOG channels. Consequently, re-referencing and regression-based ocular artifact removal cannot be applied. We therefore perform band-pass filtering, resampling, and normalization directly on the montage signals to ensure consistent frequency characteristics across datasets.

Each recording is segmented into windows of length $L{=}10$ s with stride 5 s. During training, we sample $M{=}32$ windows per recording to form a stochastic bag, and each recording is resampled 30 times per epoch. During evaluation, we use up to $M_{\text{test}}{=}64$ uniformly spaced windows and aggregate bag-level predictions.

D Baselines and Evaluation Protocols

This appendix provides additional details on the cross-hospital evaluation protocols and baseline implementations used in the main experiments. All protocols follow the source-only deployment setting described in Sec. 4.1: models are selected using source-hospital validation data and evaluated

directly on unseen hospitals without target-domain calibration, target labels, target reports, or target-domain model selection.

D.1 Single-Source Zero-Shot Evaluation

In the single-source zero-shot setting, each model is trained using recordings from one source hospital and evaluated directly on the remaining unseen hospitals. With three hospitals, this yields six train→test directions. Within each source hospital, we use patient-wise 80/10/10 splits for training, validation, and source testing. Hyperparameters and checkpoints are selected using the source validation split only.

This protocol represents the most stringent deployment scenario considered in this work: a model is developed at one institution and directly applied to a new hospital without target-hospital adaptation. For this reason, it is the primary protocol used for evaluating zero-shot cross-hospital robustness.

D.2 Single-Source Baseline Implementations

We include baselines designed to isolate the effects of diagnostic supervision, training-time report supervision, generic stochastic compression, and semantic EEG–report alignment.

ERM (supervised). End-to-end training from EEG to diagnostic labels without report supervision.

Report-only Auxiliary Supervision (Report-only aux). Uses the same EEG encoder, bag aggregation, and text decoder as TG-IL, but optimizes only report reconstruction (without IB/prototypes), isolating the effect of auxiliary report supervision.

CLIP-style. CLIP-style alignment between EEG and report embeddings using InfoNCE, without IB regularization, or prototype structure; all other components are identical for a controlled comparison.

VIB (supervised). Supervised VIB uses the same stochastic bag-level latent parameterization and KL regularization as TG-IL, but replaces report-based semantic objectives with a diagnostic cross-entropy loss and removes report reconstruction, text encoding, and prototype alignment.

D.3 Multi-Source Leave-One-Hospital-Out Evaluation

In the multi-source leave-one-hospital-out (LOHO) setting, models are trained on recordings from two hospitals and evaluated on the remaining held-out hospital. This setting provides multiple source domains during training and is widely used in the domain generalization literature because it allows methods to exploit environment structure.

For all methods, model selection is performed using source-hospital validation data only. The held-out target hospital is used exclusively for final evaluation. TG-IL does not use hospital identifiers during representation learning. For domain generalization baselines that require environment labels, we provide source-hospital identifiers during training.

D.4 Multi-Source DG Baseline Implementations

In addition to ERM and VIB, we compare against representative multi-source domain generalization methods: Invariant Risk Minimization (IRM) (Rosenfeld et al., 2020), Group Distributionally Robust Optimization (GroupDRO) (Sagawa et al., 2019), Correlation Alignment (CORAL) (Sun et al., 2017), Maximum Mean Discrepancy (MMD) (Dziugaite et al., 2015), and Domain-Adversarial Neural Networks (DANN) (Ajakan et al., 2014).

Discussion. Unlike the multi-source DG baselines above, TG-IL does not require explicit hospital identifiers or domain-adversarial objectives. Its robustness comes from training-time report-derived semantic constraints, which encourage invariance at the level of clinical interpretation rather than direct alignment of source-domain feature distributions.

E Abnormal Subtype Label Extraction and Evaluation Protocol

E.1 Motivation and Scope

Clinical EEG reports often describe abnormal findings at varying granularity and with institution-specific phrasing. To enable a standardized fine-grained downstream task across hospitals, we derive abnormal subtype labels from the impression section using a deterministic, keyword-driven extraction pipeline. This pipeline is used *only* to produce evaluation labels for the downstream probe task (Sec. 4.3); TG-IL does not require subtype labels for representation learning.

E.2 Text Preprocessing

Report preprocessing and normalization follow the standardized pipeline described in Appendix C.5. We directly operate on the processed impression/interpretation text to derive subtype labels for evaluation.

E.3 Subtype Taxonomy

We define a compact set of clinically common, cross-institutionally stable abnormal subtypes. The primary taxonomy used in our experiments contains five classes:

- **Epileptiform discharges (ED)**: spikes, sharp waves, spike-and-wave, periodic epileptiform patterns.
- **Generalized slowing (GS)**: diffuse/generalized slowing, generalized theta/delta, encephalopathy patterns.
- **Focal slowing (FS)**: focal/hemispheric slowing, lateralized delta/theta, regional dysfunction.
- **Abnormal background (ABG)**: abnormal background organization not primarily described as slowing, e.g., poorly organized background, low voltage, discontinuity (excluding burst suppression if too rare).
- **Artifact-dominant (ART)**: reports where artifact/muscle/movement is the dominant interpretation and pathological statements are absent or explicitly deferred due to artifact.

We intentionally avoid overly rare or highly specialized subtypes to reduce label noise and ensure sufficient per-class support in each hospital.

E.4 Keyword Rules

Each subtype is defined by a set of positive trigger phrases and a set of exclusion (or competing) phrases, which includes:

ED: `epileptiform`, `spike`, `sharp wave`, `spike-and-wave`, `polyspike`, `ictal`, `interictal`, `periodic discharges`, `lpds`, `gpds`. Exclusions: patterns explicitly negated (Sec. E.5).

GS: `generalized slowing`, `diffuse slowing`, `generalized theta`, `generalized delta`, `encephalopathy`, `diffuse cerebral dysfunction`.

FS: `focal slowing`, `lateralized slowing`, `left/right temporal slowing`, `hemispheric slowing`, `regional slowing`, `focal delta/theta`.

ABG: `abnormal background`, `poorly organized`, `disorganized background`, `low voltage`, `attenuation`, `discontinuity`, `asymmetry` (when discussed as background finding).

ART: `artifact`, `muscle`, `movement`, `electrode`, `technically limited`, `limited study`, `excessive artifact`. We additionally require that the report does not contain any non-negated ED/GS/FS triggers; otherwise ART is not assigned.

E.5 Negation Handling

To reduce false positives, we apply a simple window-based negation heuristic. A trigger match is suppressed if a negation word occurs within a fixed window (e.g., 0–5 tokens) preceding the

trigger, including phrases such as `no evidence of`, `without`, `absent`, `none`. For example, `no epileptiform discharges` does not activate ED. We also treat `rule out` / `cannot exclude` as *non-negated* (kept) since they indicate clinical suspicion.

E.6 Conflict Resolution and Label Assignment

Reports can mention multiple abnormalities. To construct a clean multi-class task, we enforce a single-label assignment using a fixed priority and ambiguity filter:

1. Apply all non-negated triggers to obtain a candidate set $\mathcal{C}$.
2. If $\mathcal{C} = \emptyset$, the recording is not used for subtype task.
3. If $\mathcal{C}$ contains multiple non-background subtypes (e.g., ED and FS), mark as **multi-finding** and exclude to reduce label noise.
4. Otherwise, assign the unique subtype in $\mathcal{C}$.
5. ART is assigned only if it is the sole active label after excluding other pathological triggers (Sec. E.4).

This design trades coverage for precision, yielding a more reliable evaluation of fine-grained semantic transfer across hospitals.

E.7 Inclusion Criteria and Class Balancing

We restrict the subtype task to recordings labeled abnormal by the original dataset label (when available) and additionally require that the extracted subtype is among the top-M most frequent categories in the *source training split* (we use M=5). We drop classes with fewer than a minimum number of source training examples to ensure stable probe estimation. All filtering thresholds are fixed prior to evaluation and applied identically across hospitals.

E.8 Evaluation Protocol for Subtype Classification

We evaluate abnormal subtype classification using a low-resource linear probing setup. After TG-IL representation learning, the EEG encoder is frozen and a multinomial logistic regression probe (or one-vs-rest logistic regression) is trained using 10% of labeled recordings from the source hospital only. The trained probe is evaluated unchanged on recordings from unseen hospitals.

We report: (i) **macro AUROC (OvR)** across classes, (ii) **macro-F1**, and (iii) **worst-case AUROC** over target hospitals. Hyperparameters are selected using source-only validation.

E.9 Label Quality Checks and Reproducibility Notes

To validate that subtype labels are not dominated by hospital-specific templates, we perform two checks: (1) hospital-ID prediction from the derived subtype labels alone (near chance is desired), and (2) manual spot-checking of a small random subset of extracted labels to estimate precision of the keyword rules. All keyword lists and priority rules are deterministic and will be released with the code for reproducibility.

F Additional Experimental Details

This appendix provides implementation details for EEG representation learning, text-side modules, optimization, and the semantic-invariance diagnostics used in the main experiments.

F.1 EEG Backbone Configuration

Unless otherwise stated, main-text results use LaBraM-B as the segment-level EEG feature extractor. To assess whether TG-IL depends on a particular EEG encoder, we further evaluate alternative backbones in Appendix G.1, including EEGPT, EEGNeX, and a lightweight spectral-feature MLP. All EEG encoders output fixed-dimensional segment-level representations and are paired with the same recording-level aggregation module, bottleneck parameterization, and TG-IL objective.

For each EEG recording, the preprocessed signal is segmented into 10-second windows with a stride of 5 seconds. During training, we randomly sample $M = 32$ windows from each recording to form a stochastic bag. Each recording is resampled 30 times per epoch, exposing the model to multiple partial observations of the same recording. During evaluation, we use up to $M_{\text{test}} = 64$ uniformly spaced windows and aggregate bag-level predictions.

F.2 Bag Aggregation Module

Attention pooling is implemented using a two-layer MLP scorer (256→128→1) with softmax temperature 1.0. This allows the model to weight informative EEG segments within each recording.

F.3 Variational Bottleneck Parameterization

The prior is a standard Gaussian $r(z) = \mathcal{N}(0, I)$, yielding a closed-form KL divergence. The bottleneck weight is set to $\beta = 5 \times 10^{-2}$ in all main experiments unless otherwise specified.

F.4 Text Module Configuration

For report modeling, we use MedCPT as the default pretrained biomedical text encoder. BioMedBERT and NeuroLex are evaluated as alternative text encoders in Appendix G.2. Each text encoder is followed by a lightweight projection head:

$$\text{Linear}(\text{d}_{\text{text}}, 256) \rightarrow \text{BN} \rightarrow \text{ReLU} \rightarrow \text{Linear}(256, d_z) \rightarrow \text{BN}.$$

The projected text representation is used for prototype-level EEG–report alignment during representation learning.

For report reconstruction, we use a lightweight transformer-based decoder trained with teacher forcing. The decoder is used only as a semantic sufficiency objective during representation learning. It is discarded after training and is not part of the deployed EEG-only model.

F.5 Prototype Layer and Alignment Configuration

TG-IL uses $K = 64$ learnable semantic prototypes shared across EEG and report representations. Unless otherwise stated, the prototype assignment temperature is $\tau = 0.1$, and the prototype alignment coefficient is $\lambda = 1.0$.

Given an EEG latent z_E and a projected report embedding z_T, both modalities are mapped to soft prototype assignment distributions using cosine similarity to the shared prototypes. The EEG and report assignment distributions are aligned using the symmetric KL objective described in Sec. 3. A diversity regularizer is applied to encourage batch-level prototype utilization while avoiding overly diffuse instance-level assignments.

F.6 Training Configuration

All models are trained with AdamW using cosine learning-rate decay. Unless otherwise specified, we train for 20 epochs with batch size 32 bags. Learning rates are set to 3×10^{-4} for EEG-side modules and 1×10^{-4} for text-related modules. We use weight decay 0.01 and mixed-precision training.

The same hyperparameters are used across train→test directions unless otherwise stated. All reported main results are averaged over three random seeds. Experiments are conducted on $4\times$ NVIDIA A6000 GPUs.

F.7 Semantic-Invariance Diagnostic Protocol

This section details the diagnostic protocol used in Sec. 4.5.1 to evaluate whether report-derived semantic representations are more stable across hospitals than raw lexical features or EEG latent representations.

Representation spaces. We compare three representation spaces: (i) **Raw lexical**, represented by TF-IDF features computed from report texts; (ii) **EEG latent**, represented by frozen recording-level EEG representations learned by the EEG encoder; and (iii) **Report semantic**, represented by dense embeddings from the pretrained biomedical report encoder used by TG-IL. All representations are ℓ_2-normalized before distance or retrieval computation.

Clinical semantic descriptors. For each clinical report, we extract a compact set of canonical semantic descriptors from the impression or summary section when available. The descriptor set includes common EEG phenotypes such as `normal`, `abnormal`, `slowing`, `epileptiform`, `seizure`, `asymmetry`, and `burst suppression`. Keyword variants are mapped to canonical tags; for example, "diffuse slowing" and "generalized slowing" are mapped to `slowing`. Simple negation handling is applied to exclude phrases such as "no epileptiform activity" or "without seizure activity".

These descriptors are used only for evaluation diagnostics and are never used to train TG-IL.

Cross-hospital descriptor distances. For samples drawn from different hospitals, we compute cosine distance between their representations. We report two quantities:

$$\mathrm{SameDescDist} = \mathbb{E}\big[1 - \cos(z_i, z_j) \,\big|\, d_i \neq d_j,\ \mathcal{A}_i \cap \mathcal{A}_j \neq \emptyset\big],$$

and

$$\mathrm{DiffDescDist} = \mathbb{E}\big[1 - \cos(z_i, z_j) \,\big|\, d_i \neq d_j,\ \mathcal{A}_i \cap \mathcal{A}_j = \emptyset\big],$$

where d_i denotes the hospital label and $\mathcal{A}_i$ denotes the semantic descriptor set of sample i. Lower same-descriptor distance indicates stronger cross-hospital semantic consistency, while higher different-descriptor distance indicates better semantic separation. We also report the separation gap:

$$\mathrm{Gap} = \mathrm{DiffDescDist} - \mathrm{SameDescDist}.$$

Cross-hospital semantic retrieval. For each query sample, we retrieve its top-10 nearest neighbors from other hospitals using cosine similarity. We report two retrieval metrics. Semantic Recall@10 measures whether at least one retrieved neighbor shares a descriptor with the query:

$$\mathrm{Recall@10}(q) = \mathbb{1}\left[\exists i \in \mathrm{NN}_{10}^{\mathrm{cross}}(q) : \mathcal{A}_q \cap \mathcal{A}_i \neq \emptyset\right].$$

Semantic Purity@10 measures the fraction of retrieved neighbors that share at least one descriptor with the query:

$$\mathrm{Purity@10}(q) = \frac{1}{10} \sum_{i \in \mathrm{NN}_{10}^{\mathrm{cross}}(q)} \mathbb{1}\left[\mathcal{A}_q \cap \mathcal{A}_i \neq \emptyset\right].$$

We report the average over all queries.

Hospital identifiability. To quantify domain-specific information, we train a lightweight multinomial logistic regression classifier to predict hospital identity from frozen representations. We report one-vs-rest hospital AUROC. Lower hospital AUROC indicates less hospital-specific information.

Additional text-side retrieval diagnostics. For the supplementary text-side diagnostics in Appendix G.3, we additionally compute Hospital Purity@k and Semantic Consistency@k over pooled report galleries. Hospital Purity@k measures the fraction of retrieved reports from the same hospital as the query, while Semantic Consistency@k measures the fraction of retrieved reports sharing at least one clinical descriptor with the query. These metrics are used only to analyze the geometry of frozen text representations.

G Additional Results

G.1 Results with Alternative EEG Backbones

To assess the robustness of TG-IL to the choice of EEG feature extractor, we evaluate the framework with alternative EEG backbones beyond the default LaBraM-B encoder. We first compare LaBraM-B with EEGPT (Wang et al., 2024) under the same TG-IL framework. All other components, including bag aggregation, semantic bottleneck, prototype alignment, and optimization settings, are kept identical.

Table 11 reports zero-shot cross-hospital generalization performance under the same protocol as the single-source zero-shot generalization setting in Sec. 4.2. TG-IL maintains strong robustness across pretrained EEG backbones, with only moderate variations in average performance, indicating that the proposed semantic-level invariance objective is not tied to a specific pretrained EEG encoder architecture.

Table 11: Zero-shot cross-hospital generalization with alternative pretrained EEG backbones. We report AUROC and balanced accuracy (%) for normal vs. abnormal classification.

Backbone	**BIDMC**		**BCH**		**TUAB**		**Summary**	
	AUROC	Bal. Acc	AUROC	Bal. Acc	AUROC	Bal. Acc	Avg	Worst
LaBraM-B	79.60±0.78	71.62±0.77	**80.00**±0.66	71.60±0.78	**82.63**±0.63	**78.50**±0.85	80.74	79.60
EEGPT	**80.21**±0.83	**75.34**±1.09	79.06±0.74	**73.58**±0.97	79.18±0.69	74.91±0.88	79.48	79.06

We further test whether the gains of TG-IL persist with weaker or non-pretrained EEG feature extractors. Specifically, we evaluate a spectral-feature MLP and EEGNeX (Chen et al., 2024) under the same zero-shot cross-hospital protocol, comparing ERM and TG-IL while keeping the downstream evaluation unchanged.

Table 12: Zero-shot cross-hospital generalization with weaker EEG feature extractors. TG-IL improves over ERM across both handcrafted spectral features and EEGNeX, suggesting that the benefit of semantic-level invariance is not restricted to large pretrained EEG backbones.

EEG feature encoder	**Method**	**Avg AUROC ↑**	**Worst AUROC ↑**	**Avg Bal. Acc ↑**	**Worst Bal. Acc ↑**
Spectral features + MLP	ERM	65.42±0.91	64.31±1.04	63.85±0.88	62.74±1.12
Spectral features + MLP	TG-IL	**69.08**±0.76	**68.92**±0.83	**68.47**±0.94	**66.30**±1.21
EEGNeX	ERM	72.21±0.84	71.08±0.96	70.86±0.89	69.72±1.03
EEGNeX	TG-IL	**76.94**±0.79	**74.81**±1.18	**76.47**±0.91	**73.33**±1.34

As shown in Table 12, TG-IL consistently improves over ERM for both weaker feature extractors. With spectral features, TG-IL improves average AUROC from 65.42 to 69.08 and worst-case AUROC from 64.31 to 68.92. With EEGNeX, TG-IL improves average AUROC from 72.21 to 76.94 and worst-case AUROC from 71.08 to 74.81. These results suggest that TG-IL is not merely leveraging the capacity of a particular pretrained backbone; rather, the report-guided semantic objective provides complementary robustness across different EEG feature extraction choices.

G.2 Results with Alternative Text Encoders

We further evaluate TG-IL with different pretrained biomedical text encoders to examine whether performance depends on a particular report representation. Specifically, we replace the default MedCPT encoder with BioMedBERT (Gu et al., 2021) and NeuroLex (Yin & Shin, 2026), while keeping all EEG-side components fixed.

Results are summarized in Table 13. Across encoders, TG-IL exhibits consistent cross-hospital robustness and competitive average performance. Although absolute AUROC varies slightly with text encoder capacity and pretraining corpus, the overall robustness trends remain stable, supporting the general applicability of semantic supervision across textual backbones.

Table 13: Zero-shot cross-hospital generalization with alternative text encoders. We report AUROC and balanced accuracy (%) for normal vs. abnormal classification.

Text Encoder	**BIDMC**		**BCH**		**TUAB**		**Summary**	
	AUROC	Bal. Acc	AUROC	Bal. Acc	AUROC	Bal. Acc	Avg	Worst
MedCPT	79.60±0.78	71.62±0.77	80.00±0.66	71.60±0.78	**82.63**±0.63	**78.50**±0.85	80.74	79.60
BioMedBERT	**79.86**±0.88	74.93±1.12	78.02±0.81	72.98±1.03	80.05±0.77	74.12±0.96	79.31	78.02
NeuroLex	79.06±0.62	**75.63**±0.94	**80.12**±0.61	**73.86**±0.82	79.24±0.53	74.91±0.71	79.47	79.06

G.3 Additional Text-Side Diagnostics with Alternative Encoders

The main text evaluates cross-hospital semantic invariance by comparing raw lexical features, EEG latent representations, and report semantic embeddings. Here, we provide an additional text-side diagnostic to examine whether the observed reduction of hospital-specific information is specific to the default MedCPT encoder or holds across alternative biomedical text encoders.

For each text representation, we embed clinical reports and evaluate two properties: (i) hospital identity prediction using multinomial logistic regression, and (ii) cross-hospital nearest-neighbor retrieval. Lower hospital identity predictability and lower hospital retrieval purity indicate reduced site-specific textual information, whereas higher semantic consistency indicates that nearby reports preserve shared clinical content. We include random embeddings as a chance-level reference and TF-IDF features as a raw lexical baseline.

Table 14: Additional text-side diagnostics with alternative text encoders. Lower hospital-ID predictability and retrieval purity indicate reduced domain-specific information, while higher semantic consistency reflects preserved clinical semantics.

Metric	Random	TF-IDF	MedCPT	BioMedBERT	NeuroLex
Hospital-ID AUROC (OvR) ↓	0.50	0.74	0.52	0.55	0.53
Hospital-ID Macro-F1 ↓	0.34	0.61	0.35	0.36	0.35
Hospital Purity@5 ↓	0.31	0.62	0.36	0.37	0.37
Hospital Purity@10 ↓	0.33	0.58	0.34	0.36	0.35
Semantic Consistency@5 ↑	0.19	0.46	0.71	0.71	0.73
Semantic Consistency@10 ↑	0.22	0.42	0.68	0.71	0.70

As shown in Table 14, across all pretrained encoders, hospital-ID prediction remains near chance and retrieval purity stays low, while semantic consistency is consistently high. This indicates that domain-invariant semantic structure is preserved across different text representations, and is not an artifact of a specific encoder.

G.4 Prototype Semantics

TG-IL aligns EEG and report representations through a shared set of semantic prototypes $\{\mu_k\}_{k=1}^{K}$, which serve as a structured bottleneck for organizing high-level pathological concepts. We further analyze whether the learned prototypes correspond to clinically meaningful semantics.

Prototype semantic interpretability. To examine the clinical meaning of each prototype, we collect the set of reports whose text-side assignment has the highest probability for prototype k:

$$\mathcal{R}_k = \{x_T : \arg\max_j \pi_T(j \mid z_T) = k\}.$$

For each prototype, we extract the top TF-IDF keywords from the impression section of associated reports and compute the proportion of samples labeled as abnormal. Table 15 presents representative prototypes. The correspondence between keyword sets and abnormality purity indicates that the learned prototypes capture coherent clinical concepts rather than arbitrary latent clusters.

Table 15: Semantic interpretability of learned prototypes. Top TF-IDF keywords and abnormal purity demonstrate coherent clinical concepts.

P. ID	Top-3 Keywords	Abn. Purity	Clinical Interpretation
P-19	*slowing, delta, diffuse*	98.2%	Generalized Slowing
P-25	*spike, wave, temporal*	95.5%	Epileptiform Discharges
P-32	*normal, limits, sleep*	2.1%	Normal Sleep/Background
P-44	*artifact, muscle, movement*	15.3%	Non-pathological Artifacts

H Limitations and Scope

We discuss limitations that are important for correctly interpreting the scope of TG-IL.

Report invariance is an assumption, not a guarantee. While clinician-authored reports capture high-level semantics, report style and terminology can vary across institutions, clinical services, and even individual clinicians. Our text-side sanity checks in Sec. G.3 provide evidence for (approximate) invariance in the studied benchmark, but this may not hold universally. In settings with strong site-specific report templates, TG-IL may require additional normalization (e.g., section-aware encoding, de-templating, or style-invariant text representations).

Granularity mismatch and rare events. Reports summarize findings at a coarse recording-level granularity and may omit rare or transient events. Our bag-level stochastic sampling mitigates weak temporal alignment, but extremely sparse pathologies may still be underrepresented in K sampled windows. This is a general limitation of recording-level supervision.

Prototype discretization. Prototypes provide a useful inductive bias for structured alignment, but discretizing semantics into K prototypes may underfit fine-grained clinical concepts when K is small or lead to redundancy when K is large. Although we include diversity regularization, exploring adaptive or hierarchical prototypes is an interesting future direction.

No claim of causal invariance. Our invariance objective is semantic-level and empirically validated under cross-hospital shifts. We do not claim the learned representations are causally invariant across arbitrary distribution shifts (e.g., new montages, unseen artifact types, or drastic demographic shifts). Robustness under such shifts remains an open problem.

Ethics and broader impacts. TG-IL is intended as a research method for improving robustness of clinical EEG representations under cross-hospital distribution shifts, not as a standalone diagnostic system. Potential benefits include more reliable cross-institutional EEG modeling and reduced dependence on site-specific acquisition artifacts. Potential risks include over-reliance on automated predictions, residual failures under untested demographic or acquisition shifts, and privacy concerns associated with clinical data. All experiments use de-identified secondary datasets under the applicable data-use constraints, and raw clinical data are not redistributed.

Report-derived labels and clinical scope. Subtype labels and representation-learning supervision originate from the same reports. Removing explicit subtype-trigger sentences reduces this overlap but does not make the remaining report context independent of diagnosis (Appendix I.6). The results support transfer of report-derived clinical semantic structure; independent clinical transfer requires separately collected annotations, diagnoses, or outcomes. The two evaluated tasks do not establish effectiveness for broader tasks such as seizure classification or sleep staging.

Institutional coverage, reporting bias, and fairness. The benchmark contains three hospital domains, so worst-case and LOHO results are specific to these domains. The BWH experiment changes only the source-report template; it adds no fourth EEG evaluation domain. All evaluated reports are English. Cross-language transfer, fundamentally different documentation systems, and diagnosis-correlated site-specific report phrases remain untested. This retrospective study does not establish clinical readiness or robustness to temporal shifts, changing prevalence, or new patient populations. Inconsistent demographic metadata prevent a subgroup fairness evaluation; improved cross-hospital performance is not evidence of fairness.

Unreported and temporally sparse information. Reports are incomplete semantic anchors, and attention pooling with stochastic resampling does not guarantee capture of extremely sparse events or provide event-level localization. Maximizing retention of all EEG-input information could also preserve acquisition artifacts and site-specific variation. Separating semantic and residual representations is a possible direction for preserving clinically relevant information omitted from reports.

I Additional Controlled Evaluations and Protocol Clarifications

We report additional controls for the source-only, EEG-only deployment setting. Unless stated otherwise, settings follow Appendices D and F. AUROC and F1 values are reported in percent, with mean and standard deviation across three random seeds. For the binary-task summaries below, Avg. averages the six individual source-to-unseen-target directions and Worst refers to the worst individual direction. These differ from taking the worst target-hospital mean after first averaging over source hospitals, as in Table 11.

I.1 Matched Comparison of Report- and Label-Derived Relevance

The supervised VIB baseline changes both the relevance signal and the alignment structure relative to TG-IL. To isolate these factors more closely, supervised prototype-VIB replaces report reconstruction with diagnostic-label cross-entropy and the report embedding with a learned label embedding. The stochastic bag encoder, bottleneck, KL weight, 64-prototype bank, symmetric-KL alignment, diversity regularization, optimization schedule, and downstream protocol are otherwise matched to TG-IL.

Table 16: Matched comparison of relevance signals and IB/prototype structure.

Method	Relevance signal	IB/prototypes	Avg. AUROC	Worst AUROC
ERM	Diagnostic labels	No	73.35 ± 1.45	71.18 ± 1.97
Supervised VIB	Diagnostic labels	IB only	75.31 ± 1.38	72.36 ± 1.55
Supervised prototype-VIB	Diagnostic labels	Yes	76.92 ± 1.21	75.23 ± 1.03
CLIP-style	Report semantics	No	77.41 ± 1.43	75.11 ± 1.32
TG-IL	Report semantics	Yes	80.74 ± 0.64	78.65 ± 0.61

Adding prototype alignment improves supervised VIB from 75.31/72.36 to 76.92/75.23 average/worst AUROC, demonstrating a contribution from the architecture under label supervision. With the matched IB/prototype structure, report-derived relevance further improves these metrics to 80.74/78.65. The CLIP-style comparison also shows that report access alone does not explain the full gain. These controls support contributions from both richer report supervision and the structured objective, rather than attributing all improvement to semantic-level invariance alone.

I.2 Content-Preserving Shift to an Unseen Report Template

Random report corruption disrupts EEG–report correspondence, whereas institutional template changes can preserve clinical meaning. We therefore reserialize source reports using the reporting format of Brigham and Women's Hospital (BWH), an additional institution in HEEDB outside the three-hospital EEG benchmark. Reserialization occurs after standard clinical-content extraction and canonicalization, and before text encoding. The transformed text is passed directly to the encoder without canonicalizing it back to the original section names or order.

The transformation changes section names and order while preserving clinical content, including negation, severity, laterality, and diagnostic polarity. Canonical clinical descriptors are checked to remain unchanged. Only the BWH reporting structure is used: no BWH EEG, diagnostic labels, or report-specific clinical content enter training or evaluation.

Table 17: Robustness to source-report reserialization using the BWH template.

Report condition during training	Avg. AUROC	Worst AUROC
EEG-only ERM	73.35 ± 1.45	71.18 ± 1.97
Current processed reports	80.74 ± 0.64	78.65 ± 0.61
BWH template shift	80.78 ± 0.71	78.73 ± 0.62

Performance is essentially unchanged. The small numerical increase does not establish superiority of the BWH template. This control supports robustness to the tested English section-organization variation during source representation learning, not to arbitrary terminology, other languages, funda-

mentally different reporting systems, or diagnosis-correlated institutional phrases. Reports remain unnecessary at target-hospital inference.

I.3 Harmonized EEG Preprocessing Control

The original preprocessing differs across hospitals because TUAB is supplied in a fixed bipolar montage without raw reference or EOG channels (Appendix C). BIDMC and BCH contain the same 19 standard 10–20 electrodes underlying that montage. We convert BIDMC and BCH to a common bipolar montage and disable re-referencing and EOG regression for all hospitals, leaving all other settings unchanged. Hospital-ID prediction and target-task performance are evaluated under both preprocessing conditions.

Table 18: Effect of harmonized preprocessing on hospital identifiability and target performance. Hospital-ID AUROC is lower-is-better; target AUROC is higher-is-better.

Preprocessing	Method	Hospital-ID AUROC	Avg. target AUROC	Worst target AUROC
Original	ERM	83.27 ± 0.46	73.35 ± 1.45	71.18 ± 1.97
Original	TG-IL	56.07 ± 0.29	80.74 ± 0.64	78.65 ± 0.61
Common bipolar	ERM	79.11 ± 0.42	74.12 ± 1.42	72.03 ± 1.20
Common bipolar	TG-IL	55.35 ± 0.31	80.90 ± 0.66	78.61 ± 0.64

The original preprocessing difference contributes to hospital identifiability, but does not explain the reduction achieved by TG-IL or its cross-hospital performance advantage. The common-bipolar rows disable both re-referencing and EOG regression for every hospital.

BIDMC and BCH are distinct domains despite belonging to the same Harvard resource: BIDMC covers adult tertiary care and BCH pediatric specialty care. Transfer is directional because each source trains a different model. Source coverage, patient population, case mix, and the target-task distribution can all affect AUROC. TUAB's curation for normal/abnormal screening also differs from the broader routine clinical populations at BIDMC and BCH. Consequently, transfer AUROC should not be interpreted as a symmetric distance between institutions.

I.4 Effect of Source-Label Availability on Subtype Probing

The original 10% source-label condition tests low-resource transfer with a frozen EEG encoder; it is not imposed by a lack of additional labels. We also train the subtype probe with 100% of available source labels, retaining source-only model selection and evaluation on unseen hospitals.

Table 19: Cross-hospital subtype classification with different source-label fractions.

Source labels	ERM macro AUROC	TG-IL macro AUROC
10%	61.92 ± 0.87	68.04 ± 0.74
100%	66.17 ± 0.69	72.08 ± 0.61

TG-IL retains its advantage with full source-label supervision: the mean gain is 6.12 AUROC points at 10% and 5.91 points at 100%. Both settings evaluate report-derived subtype structure, with the annotation-dependence limitation discussed below.

I.5 Standalone EEG Foundation Models and Matched Report Supervision

The ERM baseline in Table 2 is an EEG-only, end-to-end fine-tuned LaBraM-B model initialized from the same pretrained checkpoint as TG-IL. It uses source diagnostic labels but no reports, bottleneck, or prototype alignment. We additionally evaluate frozen LaBraM-B and EEGPT encoders, training only recording-level aggregation and the diagnostic head, and an EEG-only full fine-tuning baseline for EEGPT.

CLIP-style is the direct EEG–report InfoNCE baseline already included for LaBraM-B in Table 2. It uses the same paired source recordings and pretrained report embeddings as TG-IL, but omits the IB and prototype structure. We report this supervision-matched control for both backbones.

Table 20: Standalone foundation models and report-supervision-matched comparisons. Frozen-encoder rows train recording-level aggregation and the diagnostic head. Avg./Worst are computed over six individual transfer directions.

Backbone	Representation-learning method	Reports	Avg. AUROC	Worst AUROC
LaBraM-B	Frozen encoder	No	73.01 ± 1.51	71.23 ± 1.42
LaBraM-B	EEG-only full fine-tuning (ERM)	No	73.35 ± 1.45	71.18 ± 1.97
LaBraM-B	CLIP-style	Yes	77.41 ± 1.43	75.11 ± 1.32
EEGPT	Frozen encoder	No	73.18 ± 1.38	71.92 ± 1.41
EEGPT	EEG-only full fine-tuning	No	73.10 ± 1.23	71.77 ± 1.18
EEGPT	CLIP-style	Yes	76.89 ± 1.36	74.22 ± 1.46
LaBraM-B	TG-IL	Yes	80.74 ± 0.64	78.65 ± 0.61
EEGPT	TG-IL	Yes	79.48 ± 0.71	78.56 ± 0.63

Standalone pretrained EEG models remain near 73 average AUROC, while direct report alignment improves both backbones. For LaBraM-B, CLIP-style improves average AUROC from ERM's 73.35 to 77.41; TG-IL adds a further 3.33 average and 3.54 worst-direction AUROC points over CLIP-style. Thus, both access to clinical reports and the TG-IL objective contribute. The hospital-ID ERM results in Table 18 correspond to the fine-tuned LaBraM-B baseline; they are not a separate hospital-ID evaluation of every foundation-model variant.

I.6 Explicit-Label-Removal Controls and Annotation Dependence

For source recordings used to train the subtype probe, 100% of subtype labels are extracted from the IMPRESSION/INTERPRETATION subsection of the same reports used during representation learning. This overlap limits any interpretation as independent clinical transfer.

We rerun representation learning with two separate controls. For normal/abnormal classification, we remove sentences explicitly stating the overall interpretation, including "normal EEG," "abnormal EEG," and "within normal limits." For subtype classification, we remove the complete sentence containing the phrase used by the subtype-label extraction rule. Reapplying the extraction rules after sentence removal recovers the original subtype in 0% of affected reports. This is a statement about direct recoverability by those rules, not proof that all indirect diagnostic information has been removed.

Table 21: Explicit-label-removal controls. The two deletion conditions are separate representation-learning runs. Dashes denote task/condition combinations not reported.

Representation learning	Normal/abnormal Avg. AUROC	Subtype Macro AUROC	Subtype Macro-F1
ERM	73.35 ± 1.45	61.92 ± 0.87	49.83 ± 1.14
TG-IL, abnormality statements removed	80.07 ± 0.73	–	–
TG-IL, subtype-trigger sentences removed	–	66.91 ± 0.89	54.20 ± 1.18
TG-IL, full reports	80.74 ± 0.64	68.04 ± 0.74	56.91 ± 1.03

Removing explicit statements reduces performance, confirming their contribution to the original results. Nevertheless, both deletion variants remain above ERM. The remaining report context therefore provides useful supervision beyond the deleted label statements. This control does not resolve annotation dependence: independently collected clinician labels are unavailable in this benchmark. The binary task supports cross-hospital recording-level abnormality detection, and the subtype task supports transfer of report-derived clinical semantic structure.

I.7 Per-Hospital and Per-Class Subtype Results

We use the same frozen-encoder, 10%-labeled source-probe protocol as Table 4. Because the original evaluation retained aggregate metrics only, we independently reran the three-seed experiment while recording hospital- and class-level predictions. This rerun is distinct from the original aggregate experiment; its results should not be combined with the original run as if they shared predictions.

Table 22: Subtype performance by unseen target hospital and class in an independent rerun. Hospital rows report macro AUROC; class rows report one-vs-rest AUROC. Δ is the difference in reported means, in AUROC percentage points.

Breakdown	ERM	TG-IL	Δ
Unseen target hospital			
BIDMC	61.82 ± 1.04	67.49 ± 0.95	$+5.67$
BCH	62.21 ± 0.83	67.71 ± 0.80	$+5.50$
TUAB	62.87 ± 0.79	67.99 ± 0.84	$+5.12$
Subtype class			
Epileptiform discharges	64.53 ± 1.32	70.21 ± 1.17	$+5.68$
Generalized slowing	71.32 ± 1.06	76.31 ± 1.03	$+4.99$
Focal slowing	62.14 ± 1.31	70.08 ± 1.22	$+7.94$
Abnormal background	60.46 ± 0.91	67.35 ± 1.21	$+6.89$
Artifact-dominant	53.05 ± 0.73	54.70 ± 0.78	$+1.65$

TG-IL improves all three target-hospital means by 5.12–5.67 points and all five class means. Focal slowing and abnormal background show the largest gains; artifact-dominant recordings remain the most challenging and show the smallest improvement. The gains are therefore not concentrated in a single institution or subtype.

I.8 Report Examples and Remaining Protocol Clarifications

Report structure and preprocessing. Reports in all three datasets are clinician-authored narrative text, often organized by section headers, rather than checkbox-based or coded records. As described in Appendix C.5, extraction starts at FINDINGS, excluding preceding clinical history, indication, and administrative content. Equivalent section headers and formatting are standardized; explicit diagnostic terms and medical negations are retained. Hospital-specific wording within retained clinical text is not explicitly removed.

The following shortened, deidentified IMPRESSION excerpts illustrate information beyond a binary normal/abnormal label, such as recording state, localization, temporal pattern, and negated findings.

> **Example A (normal).** [IMPRESSION] This is a normal routine EEG in the awake and drowsy state. There are no epileptiform discharges or electrographic seizures. Patient did not reach sleep stage N2 in this recording. If there is high clinical suspicion, a sleep-deprived EEG may be helpful for further diagnosis.
>
> **Example B (abnormal).** [IMPRESSION] Abnormal EEG for a subject of this age due to:
>
> 1. Abundant epileptiform discharges in the left and right frontotemporal regions, indicative of underlying cortical hyperexcitability and increased risk for seizures.
> 2. Intermittent periods of lateralized rhythmic delta activity in the right temporal region, a finding commonly associated with temporal lobe epilepsy.
>
> There were no electrographic seizures.

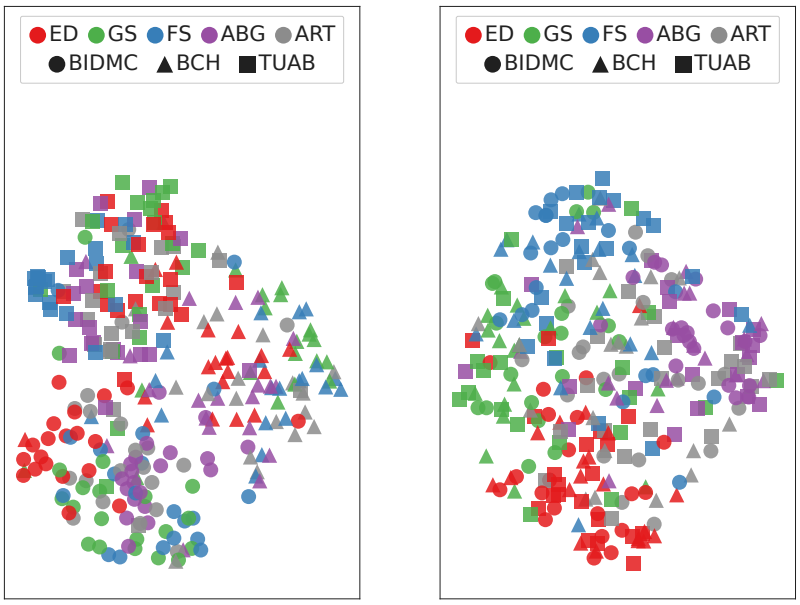
Recording-level aggregation. TG-IL uses learned attention pooling, not uniform averaging of sampled windows. Candidate windows densely tile each recording, and repeated stochastic sampling exposes training to complementary subsets. Clinical reports provide recording-level summaries, so semantic alignment is imposed after aggregation. Stochastic sampling does not guarantee coverage of extremely sparse events or establish event localization.

Source testing and prototype interpretation. The source-domain test split is retained only for post-hoc source diagnostics; it is not required for training, hyperparameter or checkpoint selection, or unseen-target AUROC evaluation. Prototype indices are permutation-identifiable across independently trained models, so encoder comparisons assess functional stability rather than one-to-one in-

dex correspondence. The alternative-encoder results in Appendix G and the template control above address complementary aspects of this stability.

I.9 Subtype-Colored Visualization Across Hospitals

Figure 4 complements the hospital-identity visualization in Figure 2 with the five abnormal subtypes evaluated in Table 4. Colors denote report-derived subtype labels, and marker shapes distinguish hospitals. The ERM projection shows pronounced grouping by hospital marker, whereas the TG-IL projection shows more mixing of hospital markers and visible regions enriched for particular subtypes. Subtype overlap remains in both projections. These two-dimensional visualizations provide qualitative context for the quantitative subtype results in Appendix I.7; they do not establish independent clinical transfer or quantify separation in the original representation space. As discussed in Appendix I.6, subtype annotations are derived from clinical reports.



(a) ERM (b) TG-IL

Figure 4: UMAP visualizations of EEG representations colored by abnormal subtype and marked by hospital. ED: epileptiform discharges; GS: generalized slowing; FS: focal slowing; ABG: abnormal background; ART: artifact-dominant. Circles, triangles, and squares denote BIDMC, BCH, and TUAB, respectively. The plots illustrate subtype-related organization and cross-hospital mixing; subtype labels are report-derived.

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 introduction state the main claims: TG-IL uses clinician-authored reports only during training to learn EEG-only representations for cross-hospital generalization. The claims are limited to the studied three-hospital clinical EEG benchmark and are supported by the experimental results in Sec. 4 .

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses limitations in Appendix H , including the approximate nature of report-derived semantic invariance, the need for paired reports during representation learning, recording-level granularity and rare events, prototype discretization, and the fact that the method does not claim causal invariance.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: [Yes]

   Justification: The paper does not claim new formal theorems, but provides the assumptions and variational derivations underlying the VSIB objective in Appendix A. The scope and limitations of the information bottleneck formulation are also discussed in Appendix A.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

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: Sec. 4.1 specifies the datasets, train/validation/test protocols, preprocessing, evaluation metrics, model components, and optimization settings. Additional dataset curation, baseline design, subtype-label extraction, and implementation details are provided in Appendices C–F, and anonymized source code is included in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: [Yes]

   Justification: Anonymized source code is provided in the supplementary material, and the existing datasets are cited with access and preprocessing details. However, the raw clinical EEG datasets are third-party clinical assets and are not redistributed with the submission; users must obtain them through the original data providers and comply with their terms of use.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: Sec. 4.1 describes the deployment-oriented evaluation protocol, source-only validation, cross-hospital splits, preprocessing, bag construction, metrics, model implementation, and optimization hyperparameters. Further experimental details are provided in Appendices C–F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: [Yes]

Justification: Main experimental results are reported as mean±standard deviation over three random seeds. Some diagnostic/ablation summaries report averages only.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: We report per-recording inference time, peak GPU memory, training time per epoch, and deployed/training-time parameter counts in Table 6. Appendix F.6 reports the hardware and optimization setup: experiments were conducted on 4Œ NVIDIA A6000 GPUs with mixed-precision training. We also report the number of epochs and batch size. Total GPU-hours for the full experimental suite are reported in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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 de-identified clinical EEG datasets under the applicable data-use constraints and does not involve re-identification, new patient contact, or autonomous clinical deployment. The work is intended for research on robustness and representation learning, and conforms to the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

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 clinically relevant limitations and deployment risks in Appendix H. Potential positive impacts include improving robustness of clinical EEG models across institutions, while potential negative impacts include over-reliance on imperfect automated predictions, residual domain shift, fairness concerns across patient populations, and privacy risks associated with clinical data.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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 high-risk generative models, scraped datasets, or patient-level raw clinical data. The released materials are limited to anonymized research code and documentation for reproducing experiments under the original datasets' access conditions.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: Existing datasets, EEG backbones, and biomedical text encoders used in the paper are cited in the main text and appendices. The work uses these assets according to their respective licenses, access restrictions, and data-use terms; the released code package includes corresponding attribution and license information.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The paper introduces anonymized source code and deterministic preprocessing/evaluation scripts as new research assets. These materials are documented in the supplementary package; no new raw clinical EEG dataset is released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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 study does not involve crowdsourcing, new participant recruitment, user studies, or direct interaction with human subjects. It uses retrospectively collected, de-identified clinical EEG datasets and does not involve participant instructions or compensation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: [Yes]

Justification: The study uses de-identified secondary clinical EEG datasets collected under the original institutional approvals, data-use agreements, or equivalent governance procedures associated with the source datasets. No new human-subject data were collected, and privacy risks are mitigated by de-identification and by not redistributing raw clinical data.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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: TG-IL uses pretrained biomedical text encoders, including MedCPT, BioMedBERT, and NeuroLex, as training-time report representation modules, and a lightweight transformer decoder for report reconstruction. These modules are described in Sec. 3.3 and Appendix F.4/G.2.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.