A Neurosymbolic Framework for Interpretable Skeleton-Based Seizure Detection via Concept-Driven Logical Reasoning

Talha Ilyas[1,2], Deval Mehta[2,3(✉)⋆], and Zongyuan Ge[2,3]

[1] Department of ECSE, Faculty of Engineering, Monash University, Australia
[2] AIM for Health Lab, Faculty of IT, Monash University, Australia
[3] Department of DSAI, Faculty of IT, Monash University, Australia
{talha.ilyas@monash.edu, deval.mehta1092@gmail.com}

Abstract. Video-based seizure detection is essential for the management of epilepsy patients, offering a non-invasive complement to electroencephalography. While several deep learning approaches have been developed for video-based seizure detection, none are inherently interpretable, limiting their adoption and translation into clinical practice. We present, to our knowledge, the first exploration of a neurosymbolic framework for video-based seizure detection that directly addresses this gap. Our approach (1) extracts patient-centric skeleton sequences from epilepsy monitoring units via a prompt-guided foundation model, (2) predicts binary spatio-temporal concept activations grounded in clinical motor semiology guidelines, and (3) composes them via differentiable logic into interpretable Boolean rules with auditable contributions. Furthermore, to mitigate false positives arising from the traditional binary formulation (seizure vs. non-seizure), we sub-classify non-seizure segments into clinically relevant normal activities, providing the model with fine-grained discriminative supervision. Evaluated on two public seizure video benchmarks, our framework achieves 89.78% sensitivity with 0.06 false detections per hour on SAHZU and 85.27% / 0.09 on IEEE, while producing complete three-level interpretability: every prediction decomposes into *which* motor primitives were detected, *how* they were logically composed, and *how much* each rule contributed to the clinical decision. We publicly release all annotations, extracted pose sequences, our data pipeline and code Link.

Keywords: Seizure detection · Neurosymbolic AI · Interpretability

1 Introduction

Epilepsy is a neurological disorder characterized by seizures, affecting approximately 50 million people worldwide [27]. Optimal diagnosis, treatment, and management of epilepsy patients depends on accurate seizure detection, including precise prediction of onset and offset times and reliable seizure counting.

⋆ now at School of Computing Technologies, RMIT University, Australia

Recently, video-based seizure detection has gained a significant interest, as it can be deployed in both hospital epilepsy monitoring units (EMUs) and out-of-hospital settings, saving time and resources for healthcare professionals and infrastructure [20,13,9].

In clinical practice, neurologists analyze video recordings to identify the observable motor manifestations of seizures, collectively termed *semiology*, following standardized guidelines published by the International League Against Epilepsy (ILAE) [3,24,2]. This vocabulary defines a taxonomy of motor signs, including *tonic* posturing (sustained muscular contraction), *clonic* jerking (repetitive rhythmic movements), *versive* deviation (forced head or eye turning), and *dystonic* posturing (sustained abnormal limb positions), among others. Crucially, clinicians reason over these observable primitives compositionally: a tonic-clonic seizure, for instance, is identified by the co-occurrence of bilateral limb stiffening followed by rhythmic jerking, not by a single holistic pattern. This structured clinical reasoning process establishes a standardized foundation for explainable decision-making aligned with established clinical guidelines.

A plethora of deep learning approaches have been proposed for video-based seizure detection, including spatiotemporal architectures [20,29,13] and skeleton-based pipelines [28,10,9]. Despite strong performance, these methods largely function as black boxes, offering limited interpretability. In clinical epileptology, where decisions rely on understanding seizure semiology and localization, such opacity constrains practical adoption [11,23]. Models that detect seizures without explicating the underlying motor signs provide limited clinical value. Additionally, most approaches adopt a binary formulation (*seizure* vs. *non-seizure*), collapsing heterogeneous non-seizure activity into a single negative class, which increases false positives during movement-intensive non-seizure activity [23].

In the general computer vision community, Concept Bottleneck Models (CBMs) [15,30] promote interpretability by routing predictions through human-defined concepts. However, primarily designed for image classification, standard CBMs treat concepts independently and do not model their logical composition, crucial for video understanding. Neurosymbolic AI addresses this limitation by integrating neural perception with symbolic reasoning, enabling differentiable Boolean combinations of concepts that mirror human inference [26,25]. Building on prior differentiable rule-learning frameworks [26,7], we lift this paradigm into the spatio-temporal video regime via ILAE-grounded spatial concepts, a shared temporal vocabulary, and a fine-grained multi-label normal-activity formulation tailored to the EMU setting. This compositional paradigm aligns naturally with epileptology, where the ILAE semiology vocabulary defines motor primitives over which clinicians reason logically. We therefore propose a neurosymbolic framework that learns Boolean rules over ILAE-grounded motion concepts via a concept bottleneck with differentiable logic layers, enabling clinically interpretable seizure detection. To further reduce false positives inherent in conventional binary (*seizure* vs. *non-seizure*) formulations, we introduce fine-grained normal activity subclasses that provide more discriminative supervision.

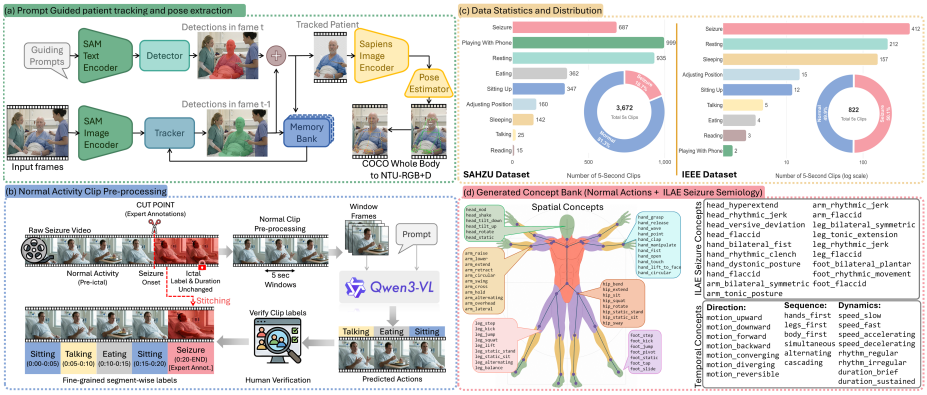

Fig. 1. (a) Prompt guided patient tracking and pose extraction, (b) VLM assisted normal activity sub-classification and verification, (c) Data Classes, statistics and distribution, (d) Generated concept bank for NTU-RGB+D 120, SAHZU and IEEE Dataset.

Our contributions are: **(1)** A neurosymbolic framework for seizure detection that learns differentiable Boolean rules over ILAE-grounded spatio-temporal motion concepts. **(2)** A fine-grained skeleton-based seizure dataset augmenting two public benchmarks [28,17] with eight normal activity subclasses via a patient-focused pose extraction pipeline (SAM 3 [4] + Sapiens [14]) and VLM-assisted annotation, publicly released. **(3)** Empirical evidence that fine-grained supervision combined with neurosymbolic reasoning reduces false positives on movement-intensive non-seizure segments while preserving interpretability.

2 Method

Our framework comprises three stages (Fig. 1): (1) patient-centric skeleton extraction and fine-grained activity labeling from raw EMU video (§2.1); (2) construction of a unified concept bank grounded in ILAE motor semiology (§2.2); and (3) neurosymbolic inference via differentiable logic layers (§2.3).

2.1 Data Pre-processing and Fine-grained Label Generation

We evaluate on two public seizure video datasets: the SAHZU dataset [28] and the IEEE Seizure Video dataset [17]. Public EMU video datasets remain scarce due to the highly sensitive nature of patient recordings, making these two among the only publicly accessible benchmarks. Both provide expert-labeled seizure onset/offset annotations but only binary (*seizure*/*non-seizure*) labels.

Patient Tracking and Pose Extraction EMU recordings differ substantially from standard human action recognition (HAR) settings: patients are typically supine, partially occluded by bedding, and share the frame with medical staff. To bridge this gap, we convert EMU videos into a standardized HAR representation using a two-stage foundation model pipeline (Fig. 1(a)). First, SAM 3 [4]

segments and tracks the patient across frames using text-prompt initialization and streaming memory to maintain identity under transient occlusions. The segmented regions are then processed by Sapiens-2B [14] to extract 133 COCO-WholeBody keypoints [12], which we map to the 25-joint NTU RGB+D format [18] via Ψ:$\mathbb{R}^{133\times2} \to \mathbb{R}^{25\times2}$. This alignment enables direct use of pretrained GCN backbones, crucial given the limited scale of seizure datasets.

Fine-Grained Normal Activity Annotation Conventional binary seizure/non-seizure annotation collapses heterogeneous normal activities into a single class, inflating false positives on movement-intensive segments (*e.g.*, patient adjusting position) that share superficial kinematic similarity with seizure activity [23]. To increase class specificity, we thus sub-classify non-seizure segments into fine-grained normal activities while preserving expert-labeled seizure boundaries untouched. Videos are clipped at the annotated seizure onset; only non-seizure segments are divided into 5-second clips and classified by Qwen3-VL [1] using structured prompts with clinical context. Predictions are then verified by human annotators and stitched back with the original seizure labels, yielding 8 non-seizure activity categories (Fig. 1(c)), 4 of which overlap with NTU RGB+D 120 [18] (Fig. 1(d)), facilitating knowledge transfer from large-scale HAR pretraining. At deployment, neither SAM 3, Sapiens-2B, nor Qwen3-VL is in the loop: they are invoked *offline* for one-time pose extraction and label refinement, while inference uses solely the compact GCN with logic layers introduced in §2.3

2.2 Seizure and Normal Activity Concept Bank Generation

Seizure detection requires a concept vocabulary spanning two distinct behavioral regimes: normal daily activities and pathological motor semiology. We construct a unified concept bank $\mathcal{C} = \mathcal{C}_\mathrm{s} \cup \mathcal{C}_\mathrm{t}$ of spatial and temporal concepts in two phases, following the concept bottleneck paradigm [15].

Phase 1: Normal activity concepts. Understanding normal activity primitives is essential for distinguishing seizure-specific motion patterns. To establish a comprehensive concept bank, we leverage the large-scale NTU RGB+D 120 dataset, which contains 120 diverse human action classes. Following [16], each action is decomposed into six body-part groups $\mathcal{P}$={`head`, `hand`, `arm`, `hip`, `leg`, `foot`}, and GPT-4o is prompted to generate part-wise motion descriptions. These descriptions are abstracted into canonical `verb-direction-modifier` patterns, embedded using a sentence encoder [22], and clustered via k-means to form spatial concepts $\mathcal{C}_\mathrm{s}^\mathrm{norm}$, and temporal concepts $\mathcal{C}_\mathrm{t}^\mathrm{norm}$ capturing motion dynamics (Fig. 1(d)). This process yields 74 concepts (53 spatial, 21 temporal).

Phase 2: Seizure-specific concepts from ILAE semiology. Seizure movements follow standardized clinical definitions that LLM-generated descriptions alone cannot capture. We treat seizure as a single holistic category and derive 19 spatial concepts from the ILAE glossary [3,24,2] using the same $\mathcal{P}$ partitioning, covering shared motor signatures across tonic, clonic, myoclonic, atonic, dystonic, and versive semiology (*e.g.*, `arm_tonic_posture`, `hand_bilateral_fist`, `head_versive_deviation`). The 21 temporal concepts from Phase 1 are shared

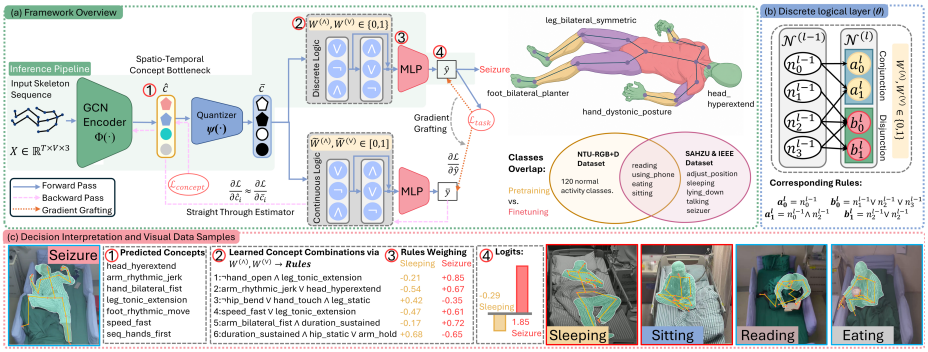

Fig. 2. (a) Model Framework overview along with body part specific concept of seizure dataset, (b) Discrete logic layer example along with it's corresponding rules, (c) Visual samples of normal and seizure activities along with rule interpretation guide.

across both regimes, as they already encode dynamics relevant to seizure discrimination (Fig. 1(d)). This yields a unified bank of $|\mathcal{C}|$=93 concepts (53 normal + 19 seizure spatial, 21 shared temporal).

Association matrix. Let $\mathcal{A}$ denote the set of action classes. We construct $\mathbf{M} \in \{0,1\}^{|\mathcal{A}|\times|\mathcal{C}|}$ where $\mathbf{M}_{a,c}$=1 indicates concept c is active for action a. Associations are generated via an LLM [8] and refined pairwise to ensure discriminability, *e.g.*, `adjusting_position` and `seizure` share `arm_extend` but differ on seizure-specific concepts such as `arm_tonic_posture`.

2.3 Neurosymbolic Seizure Detection Framework

Our framework (Fig. 2(a)) processes each skeleton sequence through four stages: a *spatio-temporal encoder* Φ extracts motion features, a *concept bottleneck* ψ maps them to binary concept activations, *differentiable logic layers* θ compose concepts into Boolean rules, and a *rule aggregator* produces class logits as a weighted sum of triggered rules.

Skeleton Encoding and Concept Bottleneck: Given an input sequence $\mathbf{X} \in \mathbb{R}^{T\times V\times 3}$ (T frames, V=25 joints), the encoder Φ models the skeleton as a spatio-temporal graph via the adaptive Hyper-GCN backbone [31], producing features $\mathbf{F} \in \mathbb{R}^{T\times V\times D}$. These are globally averaged into $\bar{\mathbf{F}} \in \mathbb{R}^{D}$ and projected to concept activations:

$$\hat{\mathbf{c}} = \sigma(\mathbf{W}_c\,\bar{\mathbf{F}} + \mathbf{b}_c) \in [0,1]^{|\mathcal{C}|}, \tag{1}$$

where each $\hat{c}_i$ is the predicted probability that concept c_i is active. With the target concept activations $\mathbf{c}^* = \mathbf{M}[a,:] \in \{0,1\}^{|\mathcal{C}|}$, derived from the action label a, the bottleneck is supervised as standard multi-label classification via binary cross-entropy ($\mathcal{L}_{\text{concept}}$), forcing the encoder to ground its representations in interpretable motion primitives before downstream reasoning. To mitigate concept leakage [7,25], we binarize activations via $\bar{c}_i = \mathbb{1}[\hat{c}_i > 0.5]$ and augment the

predicate space with negations: $\mathbf{p} = [\bar{\mathbf{c}};\, \neg\bar{\mathbf{c}}] \in \{0,1\}^{2|\mathcal{C}|}$, enabling rules over both presence and absence of motion concepts.

Differentiable Logic Layers: The logic layers θ compose binary predicates into rules via stacked conjunction-disjunction operations (Fig. 2(b)). Let $\mathbf{n}^{(0)} = \mathbf{p}$ denote the input predicate vector. Each layer l produces two sets of outputs: conjunction nodes $a_i^{(l)}$ that represent candidate AND-rules over predicates from the previous layer, and disjunction nodes $b_i^{(l)}$ that represent candidate OR-rules:

$$a_i^{(l)} = \bigwedge_{j:\, W_{ij}^{(l,\wedge)}=1} n_j^{(l-1)}, \quad b_i^{(l)} = \bigvee_{j:\, W_{ij}^{(l,\vee)}=1} n_j^{(l-1)}, \tag{2}$$

where $\mathbf{W}^{(l,\wedge)}, \mathbf{W}^{(l,\vee)} \in \{0,1\}^{N_l \times N_{l-1}}$ are learnable binary adjacency matrices that select which predicates participate in each rule. The concatenated output $\mathbf{n}^{(l)} = [\mathbf{a}^{(l)};\, \mathbf{b}^{(l)}]$ feeds subsequent layers, progressively building rules of increasing complexity in disjunctive normal form. After L layers, we collect the rule-activation vector $\mathbf{r} \in \{0,1\}^R$ over the $R{=}N_L$ learned boolean rules, with $r_k{=}1$ when the k-th rule fires for the input clip, e.g., when (`head_hyperextend` $\wedge$ `dynamics_rhythm_irregular`) holds, a typical seizure pattern.

Continuous relaxation and training: Since discrete Boolean operations yield zero gradients, we maintain continuous counterparts $\tilde{\mathbf{W}} \in [0,1]$ with differentiable logical activations [26]:

$$\tilde{a}_i^{(l)} = \mathcal{P}\Big(\prod\nolimits_j \big(1 - \tilde{W}_{ij}^{(l,\wedge)}(1 - \tilde{n}_j^{(l-1)})\big)\Big), \;\; \tilde{b}_i^{(l)} = 1 - \mathcal{P}\Big(\prod\nolimits_j \big(1 - \tilde{W}_{ij}^{(l,\vee)} \tilde{n}_j^{(l-1)}\big)\Big), \tag{3}$$

where $\mathcal{P}(x) = 1/(1 - \log x)$ prevents gradient vanishing; these reduce to standard AND/OR when inputs and weights are binary. The forward pass uses discrete weights $\mathbf{W} = \mathbb{1}[\tilde{\mathbf{W}} > 0.5]$ for interpretable inference, while backpropagation *grafts* gradients onto the continuous path to update $\tilde{\mathbf{W}}$. A Straight-Through Estimator [16] bridges the binarizer, enabling end-to-end gradient flow from the task loss through the logic layers to Φ. The total objective is $\mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{concept}} + \lambda \|\tilde{\mathbf{W}}\|_2$, where $\mathcal{L}_{\text{task}}$ is the multi-class cross-entropy between the predicted logits $\hat{\mathbf{y}}$ and the one-hot action label $a \in \mathcal{A}$, $\mathcal{L}_{\text{concept}}$ is the concept-bottleneck BCE defined above, and the sparsity term λ encourages parsimonious, interpretable rules.

Rule Aggregation and Decision Interpretation: A linear layer maps rule activations to class logits $\hat{\mathbf{y}} = \mathbf{V}\mathbf{r} + \mathbf{b}$, where each weight $V_{a,k}$ quantifies how rule r_k supports or opposes class a, yielding an additive decomposition $\hat{y}_a = \sum_k V_{a,k} \cdot r_k + b_a$. This produces a three-level audit trail (Fig. 2(c)): (1) *concept explanations*, which ILAE-aligned motor primitives are detected, (2) *rule explanations*, how concepts are composed into Boolean rules by tracing adjacency matrices, and (3) *rule contributions*, how much each rule influenced the prediction via weights $V_{a,k}$.

Table 1. Comparison with SOTA methods on SAHZU and IEEE benchmarks. Red = best; blue = second best; Interp.=interpretability; and OF=Optical Flow.

Method	Modality	SAHZU				IEEE				Interp.	Param.
		Sens.%	Spec.%	F1%	FDR/h	Sens.%	Spec.%	F1%	FDR/h		(M)
I3D+LSTM [13]	OF	71.23	68.45	69.81	0.87	67.84	65.12	66.43	1.04	NO	28.0
CNN+LSTM [29]	OF	68.36	65.92	67.08	1.13	64.51	62.87	63.62	1.31	NO	17.1
R3D+LSTM [21]	OF	72.68	69.37	70.94	0.79	69.15	66.48	67.74	0.95	NO	31.9
R3D+ViT [20]	OF	74.15	71.83	72.92	0.72	70.87	68.24	69.48	0.86	NO	32.6
JOLOGCN [28]	OF+Skel.	75.61	73.28	74.38	0.52	72.34	70.15	71.17	0.64	NO	6.9
RGBPoseConv3D [28]	RGB+Skel.	76.93	74.67	75.74	0.44	73.48	71.52	72.43	0.56	NO	3.2
AGCN+TCN [9]	RGB+Skel.	74.82	72.16	73.41	0.58	71.37	69.24	70.23	0.71	NO	37.1
RAMI [10]	OF+Skel.	78.54	76.83	79.77	0.61	75.16	73.42	76.28	0.74	NO	34.46
VSViG [28]	RGB+Patch	81.54	79.72	80.77	0.09	78.62	77.87	79.53	0.12	NO	5.4
STGCN [19]	Skeleton	76.47	74.31	75.32	0.38	73.82	71.65	72.67	0.46	NO	3.1
DG-STGCN [19]	Skeleton	78.24	76.18	77.15	0.31	75.63	73.71	74.59	0.38	NO	1.4
MSG3D [19]	Skeleton	79.86	77.54	78.63	0.27	76.91	74.82	75.79	0.34	NO	2.7
AAGCN [19]	Skeleton	78.53	76.42	77.41	0.33	75.28	73.36	74.24	0.41	NO	3.7
CTRGCN [5]	Skeleton	81.35	79.84	80.53	0.22	78.47	76.92	77.62	0.28	NO	1.4
HDGCN [28]	Skeleton	80.72	78.63	79.61	0.24	77.84	75.91	76.81	0.30	NO	1.7
InfoGCN [6]	Skeleton	80.18	78.27	79.16	0.26	77.23	75.48	76.28	0.32	NO	1.57
HyperGCN [31]	Skeleton	82.41	80.67	81.48	0.19	79.53	77.94	78.66	0.24	NO	2.4
Proposed HyperGCN+Logic Layers+Multi-label	Skeleton	89.78	87.64	86.93	0.06	85.27	83.28	83.59	0.09	**YES**	2.3

3 Experimental Setup

Datasets: We evaluate on two public seizure video datasets. **SAHZU** [28] contains EMU recordings of 14 patients with 33 focal or generalized tonic-clonic seizures (3.3 h total). **IEEE Seizure Video** [17] comprises recordings from 50 patients with 403 seizure segments (334 tonic-clonic, 69 absence) and an equal number of non-seizure segments. We use the official [28] 70/30 split for SAHZU and report 5-fold cross-validation for IEEE.

Preprocessing. We follow the standard skeleton-based HAR preprocessing protocol from CTR-GCN [5]: random rotation, spatial flip, and temporal cropping with uniform 64-frame sampling. Since our pose extraction pipeline (§2.1) outputs skeletons in the NTU-RGB+D 25-joint format, the same preprocessing applies directly to our seizure data without modification, ensuring reproducibility and compatibility with established baselines.

Pretraining on NTU-RGB+D 120: The limited scale of seizure datasets makes training from scratch impractical. We first pretrain the encoder Φ and concept bottleneck ψ on NTU-RGB+D 120 [18] using our 74 normal-activity concepts, achieving 86.7% on the cross-subject benchmark, on par with the state of the art [31]. This confirms the concept vocabulary captures sufficient discriminative information before extending it with seizure-specific concepts.

Fine-tuning and logic layer training: Starting from the pretrained encoder, we fine-tune the full framework on the seizure datasets for 400 epochs using AdamW with cosine learning rate decay on a single NVIDIA A6000 GPU with batch size 32. The concept bank is extended with the 19 seizure-specific concepts (§2.2), and the concept predictor head is expanded accordingly. Logic layers are activated after 15 warm-up epochs to ensure they receive converged concept predictions rather than random activations. The learning rate is set to 10^{-4} for the entire framework, with ℓ_1 regularization (λ=10^{-5}) inducing rule sparsity.

Table 2. Ablation studies on the SAHZU dataset.

(a) Model Configuration

Variant	Sens. %	F1 %	FDR/h
w/o NTU pretraining	78.62	76.34	1.34
w NTU pretraining	89.78	87.64	0.06
Binary model	83.31	81.26	0.21
Multilabel model	89.78	87.64	0.06

(b) Patient Tracking Pipeline

Variant	Sens. %	F1 %	FDR/h
Ilyas et al. [10]	85.84	83.97	0.65
Xu et al. [28]	80.27	81.48	0.11
Proposed	89.78	87.64	0.06

(c) Concept Configuration

Variant	Sens. %	F1 %	FDR/h
Spatial Only	80.61	81.43	0.25
Spatial + Temporal	89.78	87.64	0.06

(d) Latency (sec)	VSViG	IEEE
Xu et al. [28]	-6.8	-3.51
Proposed	-3.85	-1.47

Replacing Hyper-GCN's 121-class softmax head with our concept projection (Eq. 1) plus binary logic layers adds only ∼233K parameters, yielding a model (2.3M) marginally smaller than Hyper-GCN (2.4M).

4 Results

Comparison with state-of-the-art methods: Table 1 compares our framework against 18 baselines. Optical-flow and RGB-based methods suffer elevated false detection rates ($\geq$0.72/h) due to EMU noise, while multimodal approaches still lag behind skeleton-only models. Our framework surpasses the strongest skeleton baseline HyperGCN [31] by +**7.37** sensitivity and +**5.45** F1 on SAHZU with 68% FDR reduction (0.19→**0.06**), and by +**5.74** / +**4.93** on IEEE with 63% FDR reduction. Against VSViG [28], we gain +8.24% sensitivity with a 2.3× smaller model. Ours is the *only* method providing full interpretability.

Ablation studies: Table 2 isolates each design choice. *(a)* NTU-120 pretraining is essential (without it, sensitivity drops 11.16 points, FDR/h degrades 22×). The multi-label formulation adds +6.47 sensitivity with 3.5× FDR reduction, validating class specificity [23] in the video modality. *(b)* Our SAM 3 pipeline achieves +9.51 sensitivity over VSViG [28] pose extraction with zero manual annotation. Ilyas et al. [10] reach competitive sensitivity (85.84%) but with 10× higher FDR/h (0.65 vs. 0.06), as their optical-flow + lightweight pose front-end is sensitive to bedding motion under partial occlusion. *(c)* Temporal concepts boost sensitivity by 9.17 points, capturing rhythmicity and contraction dynamics missed by spatial concepts alone. *(d)* Our framework detects seizures 3.85 s before clinical onset on SAHZU (1.47 s on IEEE); VSViG achieves earlier detection (6.8 s / 3.51 s) via accumulative probability, but lacks interpretability.

Qualitative analysis: Fig. 3 visualizes our framework on a continuous SAHZU recording. The predicted timeline closely matches the ground truth, with seizure rules aligning with ILAE tonic semiology and resting rules suppressing other class activations. The t-SNE (Fig. 3(c)) confirms that multi-label training produces tighter clusters, explaining the 3.5× FDR reduction.

5 Conclusion

We introduced the first neurosymbolic framework for video-based seizure detection, grounding every prediction in ILAE motor semiology through a concept bottleneck with differentiable logic layers that produce fully auditable Boolean

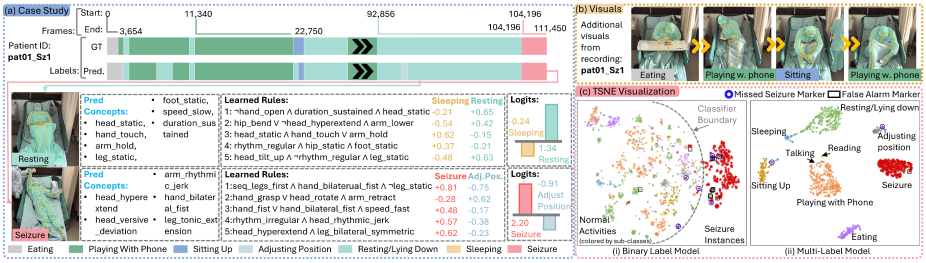

Fig. 3. Qualitative analysis on a SAHZU recording. (a) Ground-truth and predicted timelines with rule contributions for representative resting and seizure clips. (b) Additional frames from an earlier segment. (c) t-SNE plots for rule activations.

rules. Our fine-grained formulation for normal activity reduces false detections by 3.5× over binary classification, while achieving state-of-the-art sensitivity on two public benchmarks. Future work includes fine-grained seizure sub-classification across motor semiology types, expanding seizure-specific temporal concepts beyond the shared vocabulary, distilling the offline pose pipeline into a single lightweight extractor, and prospective multi-centre validation.

Acknowledgment and Disclosure of Interests. This work was supported in part by an NVIDIA Academic Grant which provided cloud access to 8xA100 GPUs. The authors declare no competing interests.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Beniczky, S., Tatum, W.O., Blumenfeld, H., Stefan, H., Mani, J., Maillard, L., Fahoum, F., Vinayan, K.P., Mayor, L.C., Vlachou, M., et al.: Seizure semiology: Ilae glossary of terms and their significance. Epileptic Disorders **24**(3), 447–495 (2022)
3. Beniczky, S., Trinka, E., Wirrell, E., Abdulla, F., Al Baradie, R., Alonso Vanegas, M., Auvin, S., Singh, M.B., Blumenfeld, H., Bogacz Fressola, A., et al.: Updated classification of epileptic seizures: Position paper of the international league against epilepsy. Epilepsia **66**(6), 1804–1823 (2025)
4. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
5. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13359–13368 (2021)
6. Chi, H.g., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: Proceedings of

the IEEE/CVF conference on computer vision and pattern recognition. pp. 20186–20196 (2022)
7. Gao, Y., Zhou, H., Gao, Z., Wang, B., Gao, S., Wang, S., Zhuang, X.: Learning concept-driven logical rules for interpretable and generalizable medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 291–300. Springer (2025)
8. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
9. Hou, J.C., Thonnat, M., Bartolomei, F., McGonigal, A.: Automated video analysis of emotion and dystonia in epileptic seizures. Epilepsy Research **184**, 106953 (2022)
10. Ilyas, T., Mehta, D., Sivathamboo, S., Wijaya, I., Steele, R., Simpson, H., Millist, L., O'Brien, T., Kwan, P., Ge, Z.: Privacy-centric seizure diagnosis via relation-aware fusion of minimally-invasive modalities. In: International Workshop on Applications of Medical AI. pp. 173–183. Springer (2025)
11. Jin, B., Wu, H., Xu, J., Yan, J., Ding, Y., Wang, Z.I., Guo, Y., Wang, Z., Shen, C., Chen, Z., et al.: Analyzing reliability of seizure diagnosis based on semiology. Epilepsy & Behavior **41**, 197–202 (2014)
12. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: European Conference on Computer Vision. pp. 196–214. Springer (2020)
13. Karácsony, T., Loesch-Biffar, A.M., Vollmar, C., Rémi, J., Noachtar, S., Cunha, J.P.S.: Novel 3d video action recognition deep learning approach for near real time epileptic seizure classification. Scientific Reports **12**(1), 19571 (2022)
14. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: European Conference on Computer Vision. pp. 206–228. Springer (2024)
15. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
16. Li, Y.L., Liu, X., Wu, X., Li, Y., Qiu, Z., Xu, L., Xu, Y., Fang, H.S., Lu, C.: Hake: A knowledge engine foundation for human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8494–8506 (2022)
17. Liang, J.: Seizure videos of epilepsy patients (2024). `https://doi.org/10.21227/nt6e-3x56`, `https://dx.doi.org/10.21227/nt6e-3x56`
18. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019)
19. Liu, M., Liu, H., Hu, Q., Ren, B., Yuan, J., Lin, J., Wen, J.: 3d skeleton-based action recognition: A review. arXiv preprint arXiv:2506.00915 (2025)
20. Mehta, D., Sivathamboo, S., Simpson, H., Kwan, P., O'Brien, T., Ge, Z.: Privacy-preserving early detection of epileptic seizures in videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 210–219. Springer (2023)
21. Pérez-García, F., Scott, C., Sparks, R., Diehl, B., Ourselin, S.: Transfer learning of deep spatiotemporal networks to model arbitrarily long videos of seizures. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 334–344. Springer (2021)
22. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural

Language Processing. Association for Computational Linguistics (11 2019), https://arxiv.org/abs/1908.10084
23. Saab, K., Tang, S., Taha, M., Lee-Messer, C., Re, C., Rubin, D.L.: Towards trustworthy seizure onset detection using workflow notes. npj Digital Medicine **7**(1), 42 (2024)
24. Turek, G., Skjei, K.: Seizure semiology, localization, and the 2017 ilae seizure classification. Epilepsy & Behavior **126**, 108455 (2022)
25. Vemuri, D.S., Bellamkonda, G., Pola, A., Balasubramanian, V.N.: Logiccbms: Logic-enhanced concept-based learning. arXiv preprint arXiv:2512.07383 (2025)
26. Wang, Z., Zhang, W., Liu, N., Wang, J.: Learning interpretable rules for scalable data representation and classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(2), 1121–1133 (2023)
27. World: Epilepsy (Feb 2024), https://www.who.int/news-room/fact-sheets/detail/epilepsy
28. Xu, Y., Wang, J., Chen, Y.H., Yang, J., Ming, W., Wang, S., Sawan, M.: Vsvig: Real-time video-based seizure detection via skeleton-based spatiotemporal vig. In: European Conference on Computer Vision. pp. 228–245. Springer (2024)
29. Yang, Y., Sarkis, R.A., El Atrache, R., Loddenkemper, T., Meisel, C.: Video-based detection of generalized tonic-clonic seizures using deep learning. IEEE Journal of Biomedical and Health Informatics **25**(8), 2997–3008 (2021)
30. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19187–19197 (2023)
31. Zhou, Y., Xu, T., Wu, C., Wu, X., Kittler, J.: Adaptive hyper-graph convolution network for skeleton-based human action recognition with virtual connections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12648–12658 (2025)