

Seeing Seizures: A Large-Scale Empirical Study of Video and EEG Deep Learning for Automated Rodent Seizure Detection and Severity Grading

Anonymous WACV Applications Track submission

Paper ID 2529

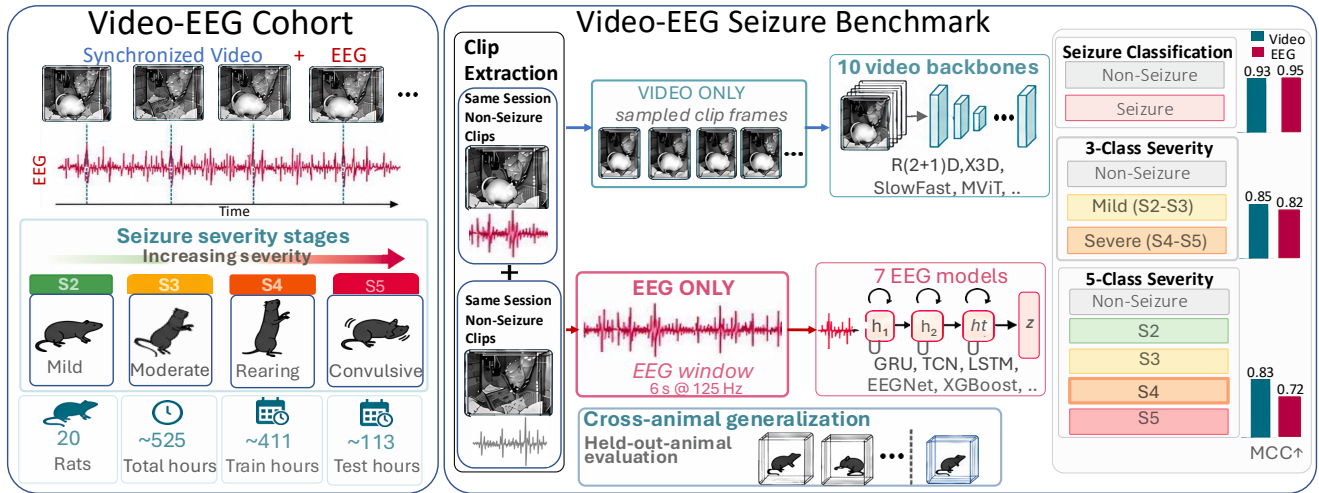


Figure 1. Overview of the study. *Left*: the cohort—synchronized video and EEG from 20 rats (corpus scale in App. C.4), each clip graded on a Racine-derived soman severity scale from Stage 2 (mild) to Stage 5 (convulsive). *Right*: the benchmark. Expert-annotated seizure clips and same-session non-seizure clips feed two independent, unfused lanes—ten video backbones and seven EEG models—compared head-to-head at 2-, 3- and 5-class granularity and under held-out-animal evaluation. Bars report macro-F1: the two modalities are level at detection and separate as the task becomes explicitly behavioral.

Abstract

001 Video and EEG are widely used to study seizures in animals.
 002 However, manual scoring remains a huge bottleneck. Mak-
 003 ing this automatic is a big challenge that few have solved
 004 for animal models. This study tackles the problem using
 005 data from twenty rats, analyzing about 24,000 paired clips
 006 from 601 sessions. Experts graded the seizures from a mild
 007 Stage 2 to a severe Stage 5. Ten video models and seven
 008 EEG models were tested under strict rules to see how well
 009 they could spot seizures and grade their exact severity. For
 010 basic detection, both tools did well, while EEG was slightly
 011 better, scoring 0.974 compared to video’s 0.968. Both tools
 012 also worked well on unseen rats. The gap grew when grad-
 013 ing the exact seizure levels. In the five-level test, video
 014 scored 0.657, easily beating EEG’s 0.560. The two meth-
 015 ods also made different types of mistakes. Video usually
 016 just confused stages that were next to each other. EEG was

more likely to miss seizures entirely or mix up very different
 stages. The size of the video model did not matter much. A
 tiny model did almost as well as a huge one. Instead, track-
 ing motion over time was the real key. An additional test
 on the outside RodEpi dataset confirmed these results, with
 the model score reaching 0.979. Overall, we demonstrate
 that cameras offer a great, pain-free way to track animal
 seizures and work very well alongside typical EEG tools.
 Our code and github page is available [here](#).

1. Introduction

Epilepsy affects roughly 1% of the global population [33],
 driving animal research that relies heavily on chemically
 or electrically induced rodent models [36]. These stud-
 ies use synchronized video-EEG, pairing video of physi-
 cal behavior with implanted brain sensors. Because elec-
 trical seizures often lack obvious movement and normal

movement easily distorts EEG traces, using both tools together remains the absolute gold standard [36, 45, 59]. Even though labs capture both data streams, automated seizure analysis focuses almost entirely on the electrical brain signals, using everything from basic frequency tracking to advanced deep learning [6, 19, 28]. As a result, the rich behavioral data captured on camera sits largely untouched by automated systems.

Properly assessing a seizure means doing more than just spotting an electrical brain event; it requires judging exactly how severe the physical behavior is. For our animals, we grade this severity on a custom five-point scale tailored for the soman model, since standard Racine scores do not directly map to chemically induced seizures [38, 39, 45, 54]. Physical symptoms escalate from mild tremors (Stage 2) up to continuous, full-body convulsions (Stage 5). Because Stage 1 is incredibly rare, our labels focus strictly on Stages 2–5. Crucially, these exact grades depend entirely on visible physical movements. As a result, tracking seizure severity is not just a brain-wave analysis problem as it is naturally a pure video-understanding task.

This perspective connects seizure severity grading to a mature body of work in video understanding. Modern action-recognition architectures are designed to model temporally evolving visual patterns [11, 18, 53], while advances in computational ethology provide complementary tools for quantifying animal behavior [8, 40]. These methods are natural candidates for recognizing severity-associated motor patterns such as tremor, rearing, and generalized convulsions. Their application to preclinical seizure analysis, however, remains limited: recent visual benchmarks have focused primarily on seizure detection [44], while existing preclinical video-EEG studies predominantly address binary detection [35, 42]. Consequently, it remains unclear how effectively modern video models can recover behaviorally defined seizure severity, and how their performance compares with EEG-based approaches on the same synchronized events.

We fill this gap by rigorously testing video models against EEG tools on the exact same paired events (Fig. 1). First, we break down exactly what a video model needs to reliably track seizures, looking at model size, frame timing, and data scale. Next, we track how the strengths of video and EEG shift as the task moves from basic detection to precise severity grading. We also check if the two methods make different kinds of mistakes. Finally, we test whether the models can still learn to spot the rarest seizure stages despite massively lopsided data.

This study advances automated animal seizure tracking in five key ways:

1. To our knowledge, this is the first study to grade exact seizure severity directly from video, treating physical symptoms as a pure video-understanding task.



Figure 2. The recording setup: each rat stays in its normal home cage under a fixed overhead camera. At the same time, implanted sensors track EEG brain waves from both sides of the brain. Each video is tightly cropped around the specific cage to cancel out any differences in camera angles. Identifying details in this frame are masked.

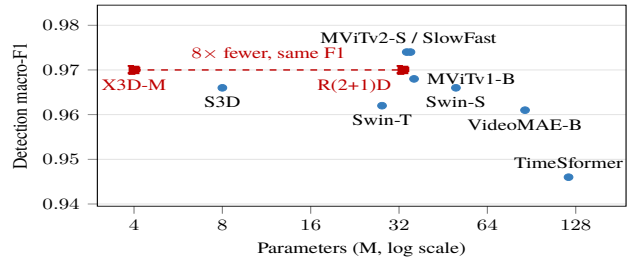


Figure 3. Basic detection scores compared to model size. Across all ten video models tested, performance stayed tightly grouped between 0.945 and 0.974. A tiny 4M-parameter X3D-M model performed almost exactly as well as a heavy 33M-parameter R(2+1)D-18 model (Tables 2 and 9).

2. It offers the first head-to-head comparison of video and EEG on the exact same events. Results show EEG holds a slight edge for basic detection, but video takes the lead for precise severity grading.
3. It breaks down exactly what video models need to succeed. Tracking motion over time is crucial, but model size barely matters; a tiny 4M-parameter X3D-M model scores just as well as a heavy 33M-parameter R(2+1)D model (Fig. 3).
4. It tackles strict severity grading despite massively lopsided data, proving that the models can successfully handle brand new, unseen animals.

2. Related Work

Automated seizure tracking naturally splits based on the tools used and the subjects studied. We break past research down into three main areas: EEG-only, video-only, and mixed video and EEG systems.

EEG-based seizure analysis. EEG is still the standard tool for automated seizure tracking. Early animal systems relied on hand-picked wave features and basic machine learn-

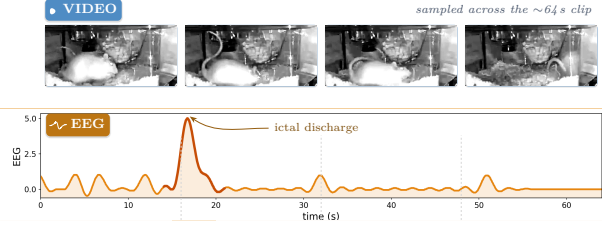


Figure 4. A paired Stage 3 seizure from rat R7. *Top*: Four sample video frames. *Bottom*: The corresponding single-channel EEG trace (~ 64 s), with the active seizure period highlighted. Watching these streams side-by-side provides a complete view of both the physical symptoms and the electrical brain activity. Identifying markers have been masked.

ing [6]. Today, newer methods use deep neural networks, such as recurrent, convolutional, and attention models, to track signal changes over time [28]. For chemically induced rodent seizures, deep recurrent networks like GRUs and LSTMs easily beat older tools like XGBoost and Random Forest [19]. While these brain-wave models set a strong baseline for basic detection, they only watch electrical signals. They completely miss the physical movements that actually define a seizure’s behavioral severity. This blind spot is exactly why this study compares them directly against video models on the exact same events.

Video-based seizure and behavior analysis. Video seizure tracking ties closely to action recognition. Models like two-stream networks, 3D convolutions, and video transformers excel at mapping motion over time [7, 11, 18, 49, 52, 53]. In broader animal research, scientists already use similar tools to estimate poses [40, 43] and classify general behaviors [8, 22, 51]. For human epilepsy, clinical cameras mainly spot basic seizures and outward symptoms [3, 24, 58]. However, animal video research remains fairly rare. Existing work mostly tracks specific induced seizures [47], general home-cage movement [23, 46], or basic activity in mice [27] and zebrafish [56].

The closest matching study is Perlo *et al.*’s *RodEpi* [44]. They built a video-only dataset with 13,053 clips from 19 rats and scored around 97% F1 using a TimeSformer model [7]. Yet, all of these prior animal studies stop at basic yes-or-no detection. Until now, using cameras to grade exact, multi-level seizure severity has never been systematically tested.

Synchronized video and EEG analysis. Researchers rely heavily on paired video and EEG to assess seizures because physical behavior and EEG provide a complete picture (Fig. 4). However, computer models combining both streams remain rare. Clinical trials have tested joint systems [1, 10, 31, 37], but animal studies usually stick to basic yes-or-no seizure detection using mixed sensors [35, 42]. Instead of mixing them right away, this study tests video

and EEG separately on the exact same events. This allows a clean comparison from basic detection down to precise severity grading, setting the stage for future mixed models. A full literature review and breakdown are available in App. A and App. A.1, Table 7.

3. Problem Setup

We formalize the labels, task granularity, and modality-specific classifiers used for the matched video–EEG evaluation in this section.

3.1. Detection and severity grading

Clips and labels. A session $r \in \mathcal{R}$ is one continuous synchronized video–EEG recording from a single animal. From each session, we extract paired clips $c = (\mathbf{V}, \mathbf{E})$, including expert-annotated seizure events and matched non-seizure intervals from the same session. The video clip is $\mathbf{V} \in \mathbb{R}^{T \times H \times W \times 3}$, where T , H , and W denote the number of frames, frame height, and frame width, respectively. The corresponding EEG clip is $\mathbf{E} \in \mathbb{R}^{C \times L}$, where C is the number of EEG channels and L is the number of temporal samples.

Each clip receives a label $y \in \{0, 1, 2, 3, 4\}$. The label $y = 0$ denotes a non-seizure clip, while $y = s - 1$ represents Stage $s \in \{2, 3, 4, 5\}$ on the laboratory’s soman-specific behavioral severity scale [38, 39]. Stage 1 is excluded because it primarily reflects autonomic changes without reliable visible motor manifestations and is extremely rare in the cohort. This scale is adapted specifically to soman-induced seizures and should not be conflated with the classical Racine scale [45, 54]. Before pooling clips across animals, each video is tightly cropped to the corresponding cage boundaries to account for differences in camera geometry (Fig. 2).

Event boundaries. Experts mark a seizure’s exact start and end at the first and last visible movements. Because EEG often shift before any physical signs appear, an extra 10 s is added to both ends to catch the natural delay between the EEG and the video.

Task granularities. To track different levels of detail for the exact same events, three nested tasks map the main label y using $g_k : \{0, \dots, 4\} \rightarrow \{0, \dots, k - 1\}$:

$$g_2(y) = \begin{cases} 0, & y = 0 \\ 1, & y > 0 \end{cases} \quad (1)$$

$$g_3(y) = \begin{cases} 0, & y = 0 \\ 1, & y \in \{1, 2\} \\ 2, & y \in \{3, 4\} \end{cases} \quad (2)$$

$$g_5(y) = y. \quad (3)$$

Table 1. Paired video and EEG results from strict session splits (mean $\pm$ std over five test runs). P, R, and F1 scores are macro-averaged; MCC uses the full confusion matrix. EEG holds a slight edge for basic detection, but video easily wins the 3- and 5-level grading. AUROC scores are in App. E.5, Table 15. Best results are **bolded**.

Task	Modality	Prec.	Rec.	F1	MCC
2-class (detection)	Video	0.968 $\pm$ 0.002	0.968 $\pm$ 0.002	0.968 $\pm$ 0.002	0.936 $\pm$ 0.004
	EEG	0.975 $\pm$ 0.002	0.974 $\pm$ 0.003	0.974 $\pm$ 0.003	0.949 $\pm$ 0.005
3-class (severity)	Video	0.775 $\pm$ 0.016	0.788 $\pm$ 0.009	0.780 $\pm$ 0.007	0.853 $\pm$ 0.010
	EEG	0.744 $\pm$ 0.014	0.764 $\pm$ 0.013	0.750 $\pm$ 0.010	0.823 $\pm$ 0.014
5-class (severity)	Video	0.689 $\pm$ 0.024	0.643 $\pm$ 0.006	0.657 $\pm$ 0.009	0.831 $\pm$ 0.006
	EEG	0.556 $\pm$ 0.040	0.635 $\pm$ 0.035	0.560 $\pm$ 0.005	0.728 $\pm$ 0.032

Here, g_2 defines binary seizure detection; g_3 defines three-class classification into non-seizure, lower-severity Stages 2–3, and higher-severity Stages 4–5; and g_5 retains all four represented severity stages as separate classes. The g_3 mapping forms two contiguous pairs of the represented stages and serves as an intermediate-resolution grading task.

Modalities. For each task $\mathcal{T}_k$, where $k \in \{2, 3, 5\}$ is the number of classes, we train separate video and EEG classifiers $f_V^{(k)}$ and $f_E^{(k)}$. Let $m \in \{V, E\}$ denote the modality, with $\mathbf{X}_V = \mathbf{V} \in \mathbb{R}^{T \times H \times W \times 3}$ and $\mathbf{X}_E = \mathbf{E} \in \mathbb{R}^{C \times L}$. Each classifier outputs k class probabilities, and its prediction is

$$\hat{y}_m^{(k)} = \arg \max_{j \in \{0, \dots, k-1\}} \left[f_m^{(k)}(\mathbf{X}_m) \right]_j, \quad (4)$$

where j indexes the output classes. The two modalities are trained and evaluated independently on the same events; multimodal fusion is left for future work (Sec. 6).

Evaluation protocol. To prevent artificially high scores, several strict safeguards are in place. First, session-level splits keep entire recordings strictly in either training or testing for both tools. Second, inverse-frequency weighting balances out rare events. Third, macro-F1 serves as the main metric alongside MCC (App. B.4); AUROC scores live in App. E.5. Finally, tests track individual results per rat and prove how well models handle brand new animals using strict five-fold splits (Sec. 5.2).

3.2. Conditional ordinal severity formulation

Standard flat classifiers ignore the ranked order of severity labels. A normal clip does not belong on this scale, while actual seizures follow a strict sequence ($S2 \prec S3 \prec S4 \prec S5$). A basic classifier penalizes all mistakes equally, treating a minor S2-for-S3 mix-up exactly the same as a massive S2-for-S5 failure.

Conditional factorization and ordinal head. To fix this, seizure occurrence is defined as $z = \mathbf{1}[y > 0]$, splitting the

prediction into two steps:

$$p(y | c) = \begin{cases} p(z = 0 | c), & y = 0, \\ p(z = 1 | c) p(s = s(y) | c, z = 1), & y > 0. \end{cases} \quad (5)$$

For active seizures, severity is mapped using rolling thresholds:

$$q_j(c) = p(s \succ S_j | c, z = 1), \quad j \in \{2, 3, 4\}, \quad (6)$$

enforcing the strict downward rule $q_2(c) \geq q_3(c) \geq q_4(c)$. The final predicted stage simply depends on how many thresholds the model crosses.

Ordinal evaluation. Macro-F1 misses how far a wrong guess falls from the true severity stage. To fix this, we measure the average gap between guessed and real stages, $\text{MAE}_{\text{stage}} = \frac{1}{N} \sum_{i=1}^N |\hat{s}_i - s_i|$, and track how often the model misses by two or more stages, $\frac{1}{N} \sum_{i=1}^N \mathbf{1}[|\hat{s}_i - s_i| \geq 2]$. Normal clips do not fit this ranked scale, so these metrics only apply to actual seizures. Future ranking tests should use the exact same unseen-animal splits from Sec. 5.2.

4. Dataset

We analyzed paired video and EEG recordings from twenty rats exposed to soman. The dataset contains 24,131 clips across 601 sessions. Following standard rules [38, 39], we recorded each rat in its home cage using a fixed overhead camera and implanted EEG sensors. The main recording system stamped the time on both streams so they line up perfectly from the start. We tracked natural seizures that happened 10 to 18 weeks after the initial exposure, right in the middle of the chronic epilepsy phase. We recorded the raw EEG at 1 kHz and lowered the rate to 125 Hz. Finally, we cropped the videos tightly around each cage to make all the camera views match (Fig. 2).

Annotation and labels. Researchers watched the paired video and EEG together using NeuroScore (DSI). They marked the exact start and stop times and graded the behavior [38, 39]. They double-checked EEG alerts against the video to throw out false alarms. Because every label comes from mixing both tools, we do not have completely separate ground truths for each. We discuss this limit in Sec. 5.4.

We grade seizure severity on a 5-level scale [38, 39, 45]. This is very different from the classic Racine scale [54]. Stage 1 only involves internal body changes. Stages 2 through 5 track worse and worse physical shaking, ending in full-body convulsions. We drop Stage 1 because we only caught 6 clips of it. The working data has 1,435 Stage 2, 9,256 Stage 3, 1,050 Stage 4, and 193 Stage 5 clips. When we add 12,197 normal baseline clips, the final total is 24,131 clips. The data is heavily lopsided across both the seizure stages and the individual rats. To handle this, we

Table 2. We tested ten video models using the exact same data splits and a strict 12-epoch training limit. Scores show the mean $\pm$ std over five test runs (SlowFast detection uses $n=4$). Basic detection scores barely changed, shifting by only 0.03 across all the models. You can find the AUROC scores and exact model setups in App. E.5 and App. B.1. We **bolded** the best results and underlined the second-best.

Backbone	Detection (2-class)				3-class grading				5-class grading			
	P	R	F1	MCC	P	R	F1	MCC	P	R	F1	MCC
MViTv2-S [30]	<u>0.972</u> ± 0.001	<u>0.972</u> ± 0.001	<u>0.972</u> ± 0.001	<u>0.944</u> ± 0.003	0.780 ± 0.014	0.800 ± 0.027	0.786 ± 0.005	0.856 ± 0.017	0.694 ± 0.032	0.656 ± 0.030	0.660 ± 0.016	0.832 ± 0.007
SlowFast-R50 [18]	0.974 ± 0.002	0.974 ± 0.002	0.974 ± 0.002	0.948 ± 0.004	<u>0.781</u> ± 0.018	<u>0.804</u> ± 0.013	<u>0.789</u> ± 0.009	<u>0.859</u> ± 0.012	<u>0.718</u> ± 0.024	<u>0.655</u> ± 0.014	0.675 ± 0.008	0.851 ± 0.003
R(2+1)D-18 [53]	0.971 ± 0.001	0.971 ± 0.001	0.971 ± 0.001	0.942 ± 0.002	0.780 ± 0.007	0.787 ± 0.007	0.783 ± 0.002	0.860 ± 0.003	0.734 ± 0.027	0.643 ± 0.011	<u>0.667</u> ± 0.010	<u>0.841</u> ± 0.007
X3D-M [17]	0.969 ± 0.002	0.969 ± 0.002	0.969 ± 0.002	0.937 ± 0.003	0.782 ± 0.008	0.810 ± 0.015	0.794 ± 0.005	0.856 ± 0.008	0.568 ± 0.004	0.616 ± 0.013	0.586 ± 0.003	0.824 ± 0.007
MViTv1-B [16]	0.969 ± 0.002	0.969 ± 0.002	0.969 ± 0.002	0.938 ± 0.003	0.771 ± 0.018	0.794 ± 0.016	0.779 ± 0.011	0.849 ± 0.016	0.687 ± 0.022	0.643 ± 0.011	0.653 ± 0.008	0.836 ± 0.006
Video Swin-S [34]	0.966 ± 0.003	0.966 ± 0.003	0.966 ± 0.003	0.932 ± 0.005	0.767 ± 0.013	0.780 ± 0.022	0.771 ± 0.009	0.834 ± 0.016	0.716 ± 0.032	0.634 ± 0.017	0.645 ± 0.011	0.818 ± 0.012
S3D [57]	0.967 ± 0.002	0.967 ± 0.002	0.967 ± 0.002	0.933 ± 0.005	0.756 ± 0.014	0.782 ± 0.027	0.764 ± 0.004	0.840 ± 0.015	0.644 ± 0.015	0.639 ± 0.021	0.636 ± 0.009	0.813 ± 0.015
Video Swin-T [34]	0.963 ± 0.001	0.962 ± 0.001	0.963 ± 0.001	0.925 ± 0.002	0.761 ± 0.014	0.776 ± 0.020	0.766 ± 0.008	0.829 ± 0.006	0.660 ± 0.022	0.639 ± 0.009	0.632 ± 0.002	0.796 ± 0.006
VideoMAE-B [52]	0.958 ± 0.004	0.958 ± 0.004	0.958 ± 0.004	0.915 ± 0.007	0.729 ± 0.033	0.736 ± 0.042	0.729 ± 0.030	0.771 ± 0.063	0.704 ± 0.018	0.623 ± 0.025	0.636 ± 0.017	0.802 ± 0.013
TimeSformer [7]	0.945 ± 0.001	0.944 ± 0.001	0.945 ± 0.001	0.889 ± 0.002	0.722 ± 0.017	0.761 ± 0.017	0.735 ± 0.013	0.778 ± 0.019	0.650 ± 0.014	0.594 ± 0.014	0.601 ± 0.009	0.757 ± 0.028

give rare events extra weight, carefully average our scores, track results per rat, and test the models on brand new animals. You can find all the strict rules for ethics, privacy, and data handling in App. C (App. C.6). While we are not able to release the complete raw data at this time due to certain institutional restrictions, we plan to release the data set soon via Scientific Data Repositories, such as the Neuroimaging Tools and Resources Collaboratory (NITRC), or the Open-Neuro platform. For reproducibility, we provide the data collection protocol in the App. C (App. C.1).

5. Experiments and Discussion

We test standard video and EEG models on our group of twenty rats (Sec. 4). We want to see how well they handle basic seizure detection and exact severity grading. We also check how they perform on brand new animals and on outside datasets. Unless we say otherwise, the video setup uses a R(2+1)D-18 model [25, 53] watching 16 frames at a 112×112 resolution. The EEG setup uses the GRU model built by Ganguly *et al.* [19]. Both tools face the exact same test splits and use the same math rules to balance out rare events. Table 1 shows the average scores across five different test runs. You can find all the strict details about the model designs and training steps in Apps. B.1–B.3.

5.1. Video capacity and input requirements

Detection baseline. Our baseline R(2+1)D model scores 0.968 ± 0.002 macro-F1 and 0.936 ± 0.004 MCC (Table 1). We purposely paired seizure and normal clips from the exact same session. This keeps the animal, camera angle, and lighting identical. It forces the model to watch for real behavior instead of just cheating by recognizing the background (App. C.3, Fig. 9). The model also stays highly accurate for the animals that suffer the most seizures, scoring 0.974 on rat R1 and 0.906 on rat R7.

Backbone capacity. We tested ten video models. Their basic detection scores tightly ranged from 0.945 to 0.974 (Table 2). Model size barely mattered. The small 4M-

Table 3. Temporal-sampling ablation for R(2+1)D (seed 42). Single-frame performance is poor, while performance peaks around 32 frames. Using 64 frames gives no detection gain and lowers 5-class F1. AUROC results are reported in App. E.5, Table 17.

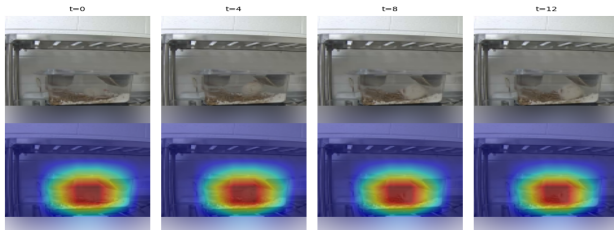
Frames	Detection		3-class		5-class	
	F1	MCC	F1	MCC	F1	MCC
1	0.569	0.149	0.417	0.152	0.314	0.147
4	0.940	0.881	0.739	0.760	0.612	0.741
8	0.959	0.918	0.763	0.838	0.627	0.789
16	0.970	0.941	0.780	<u>0.857</u>	<u>0.662</u>	<u>0.830</u>
32	0.972	0.943	0.793	0.858	0.667	0.835
64	<u>0.971</u>	<u>0.941</u>	<u>0.785</u>	0.848	0.632	0.807

parameter X3D-M scored 0.969, almost perfectly matching the massive 33M-parameter R(2+1)D at 0.971 (Fig. 3). Pretraining on outside videos added only a tiny 0.02 boost. Also, the models stopped improving long before seeing all the data (App. E.1). Ultimately, for this setup, the size of the network is far less important than its ability to capture motion over time.

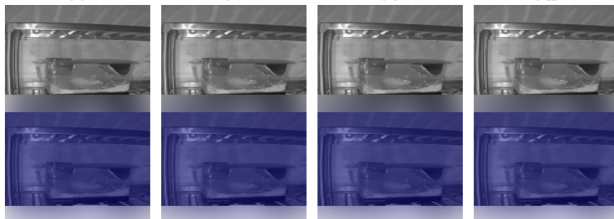
Temporal and spatial evidence. Tracking motion over time is incredibly important. A model looking at just a single still frame does very poorly. It scores only 0.569, 0.417, and 0.314 across the 2, 3, and 5-level tests. Using 16 frames works much better, bumping those scores up to 0.970, 0.780, and 0.662 (Table 3). Performance peaks around 32 frames. Pushing it up to 64 frames does not improve basic detection at all and actually hurts the 5-level grading, dropping it to 0.632. On the flip side, making the video sharper does not help. Increasing the resolution from 112 to 224 or 448 pixels gives zero benefit (App. E.4). This makes sense because the original lab cameras only shoot at around 300 to 480 pixels anyway. Because of this, we set

Table 4. Seven EEG baselines use the same 6 s windows, 3 s stride, 125 Hz sampling, and log-mean aggregation (App. B.5). TCN performs best for detection and 3-class grading, while GRU and TCN are similar at 5 classes. Two LSTM runs failed to converge, so its 3- and 5-class results use $n=4$. Best results are in **bold**; second-best are underlined.

EEG backbone	Detection (2-class)				3-class				5-class			
	P	R	F1	MCC	P	R	F1	MCC	P	R	F1	MCC
<i>Deep</i>												
GRU [19]	<u>0.975</u> ± 0.003	<u>0.974</u> ± 0.003	<u>0.974</u> ± 0.003	<u>0.949</u> ± 0.005	<u>0.744</u> ± 0.014	<u>0.764</u> ± 0.013	<u>0.750</u> ± 0.010	<u>0.823</u> ± 0.014	0.556 ± 0.040	<u>0.635</u> ± 0.035	<u>0.560</u> ± 0.005	0.728 ± 0.032
LSTM [21]	0.972 ± 0.008	0.971 ± 0.007	0.971 ± 0.008	0.942 ± 0.015	0.696 ± 0.041	0.741 ± 0.067	0.698 ± 0.053	0.743 ± 0.059	0.476 ± 0.042	0.555 ± 0.100	0.468 ± 0.052	0.598 ± 0.059
EEGNet [29]	0.926 ± 0.011	0.922 ± 0.013	0.923 ± 0.012	0.848 ± 0.023	0.646 ± 0.010	0.676 ± 0.018	0.620 ± 0.015	0.626 ± 0.031	0.448 ± 0.040	0.488 ± 0.025	0.409 ± 0.039	0.499 ± 0.037
EEG Conformer [50]	0.972 ± 0.001	0.972 ± 0.001	0.972 ± 0.001	0.943 ± 0.001	0.719 ± 0.009	<u>0.784</u> ± 0.013	0.729 ± 0.010	0.780 ± 0.017	0.512 ± 0.004	0.624 ± 0.032	0.523 ± 0.005	0.683 ± 0.008
TCN [5]	0.979 ± 0.001	0.978 ± 0.001	0.978 ± 0.001	0.957 ± 0.002	0.758 ± 0.014	0.786 ± 0.008	0.767 ± 0.007	0.839 ± 0.015	<u>0.537</u> ± 0.008	0.654 ± 0.013	0.563 ± 0.012	<u>0.728</u> ± 0.021
<i>Classical (band-power + statistical features)</i>												
Random Forest [9]	0.885 ± 0.001	0.852 ± 0.001	0.852 ± 0.001	0.737 ± 0.002	0.541 ± 0.001	0.542 ± 0.001	0.529 ± 0.001	0.613 ± 0.002	0.477 ± 0.001	0.328 ± 0.001	0.332 ± 0.002	0.503 ± 0.002
XGBoost [13]	0.922 ± 0.000	0.913 ± 0.000	0.914 ± 0.000	0.835 ± 0.000	0.657 ± 0.000	0.684 ± 0.000	0.642 ± 0.000	0.623 ± 0.000	0.479 ± 0.000	0.530 ± 0.000	0.466 ± 0.000	0.529 ± 0.000



(a) Seizure clip (Stage 4): activation concentrates on the animal



(b) Non-seizure clip: activation is diffuse, with no focus on the animal

Figure 5. Grad-CAM [48] shows exactly what our R(2+1)D model watches, with the original video on top and the model’s focus on the bottom. We show four of the eight frames here. (a) During a Stage 4 seizure, the model zeroes right in on the animal. (b) During a normal clip, its focus spreads everywhere. This proves the model tracks actual behavior instead of just looking at the cage background. All eight full-size frames are in App. D.2, Fig. 14. Lab tags are blurred for privacy.

our standard setup to 16 frames at a 112×112 resolution.

Shortcut check. We took several steps to stop the model from taking easy shortcuts. We paired clips from the same session, tightly cropped the cages, and proved that single still frames fail completely. This rules out simple background cheats. Grad-CAM also shows the model focusing right on the animal during a seizure (Fig. 5). This strongly suggests the model is watching real behavior, but it is not absolute proof. To be completely sure, future tests would need to artificially hide the animal, show only the empty background, or scramble the video frames (App. D.3).

5.2. Cross-modal comparison and generalization

EEG baseline strength. To see how well EEG could perform, we tested seven different models: GRU, LSTM [21], EEGNet [29], EEG Conformer [50], TCN [5], Random Forest [9], and XGBoost [13]. We wanted to keep the playing field level, so they all looked at the same main EEG channel. We simplified the data rate down to 125 Hz and fed it to the models in 6-second chunks. We chose these specific settings after running several tests to see what worked best, though there might still be room for tiny improvements in the future (App. B.5). Out of all these models, the TCN was the clear winner for basic detection (0.978) and the 3-level grading test (0.767). However, when it came to the hardest 5-level test, it scored a 0.563, which is basically a tie with our baseline GRU model (0.560). The most important takeaway here is that even if we use this stronger TCN model, our main story does not change. EEG still wins at basic detection (0.978 vs. 0.968), runs neck-and-neck with video on the 3-level test (0.767 vs. 0.780), and loses to video on the 5-level grading (0.563 vs. 0.657). It is worth pointing out one big limitation, though: we only used a single EEG sensor. If we had used multiple sensors to get a wider view of the brain’s activity, the EEG models might have closed the gap on those harder grading tasks.

Metric robustness. Compared to the GRU, video’s lead grows from -0.006 to $+0.030$ to $+0.097$ across the 2, 3, and 5-level tests. MCC scores show the exact same trend (-0.013 , $+0.030$, $+0.103$). Video beats no matter which metric we use. AUROC data is in App. E.5.

Unseen animals. Testing the models on brand new animals keeps basic detection strong. Video gets 0.948 ± 0.016 and EEG gets 0.972 ± 0.015 (Fig. 6). However, grading exact seizure levels is much harder. In the 3-level test, video drops from 0.780 to 0.720, and EEG drops from 0.750 to 0.656. In the 5-level test, video falls from 0.657 to 0.475, and EEG falls from 0.560 to 0.404. Video still wins the 5-level test, but its lead shrinks from 0.097 to 0.071 (Table 5).

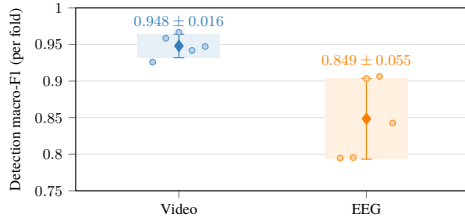


Figure 6. Detection scores for unseen animals (Table 5). Video gets 0.948 ± 0.016 and EEG gets 0.972 ± 0.015 . The score swings are similar, but EEG has a higher average. Table 5 shows the grading results.

Table 5. Testing new sessions versus new animals. Testing unseen animals keeps basic detection high but hurts exact grading. Video’s 5-class lead shrinks from 0.097 to 0.071. Scores show the mean $\pm$ std over five tests.

Task	Mod.	Session-excl.		Subject-exclusive (5-fold)	
		Macro-F1	MCC	Macro-F1	MCC
Detection	Video	0.968	0.936	0.948 ± 0.016	0.896 ± 0.031
	EEG	0.974	0.949	0.972 ± 0.015	0.944 ± 0.030
3-class	Video	0.780	0.853	0.720 ± 0.028	0.761 ± 0.029
	EEG	0.750	0.823	0.656 ± 0.022	0.793 ± 0.046
5-class	Video	0.657	0.831	0.475 ± 0.024	0.676 ± 0.041
	EEG	0.560	0.728	0.404 ± 0.042	0.657 ± 0.107

Granularity and error complementarity. As the grading gets more specific, the video score drops from 0.968 to 0.780 and then to 0.657. The EEG score drops even faster, going from 0.974 to 0.750 and down to 0.560 (Fig. 7). In the hardest five-level test, video scores 0.362 on Stage 4 and 0.415 on Stage 5. EEG only gets 0.202 and 0.123 on these same stages (Table 14). Both models still struggle with these rare stages compared to the very common Stage 3.

The two tools also make different types of mistakes. EEG is better at catching mild Stage 2 events (0.863 versus 0.650). But EEG completely misses 10.1% of the strong Stage 3 seizures, calling them normal. Video only misses 2.3% of these. We looked at 2,830 clips tested by both models. The video was the only one right 14.2% of the time. The EEG was the only one right 4.9% of the time (App. E.9, Table 18). Because they fail in such different ways, future tools should combine them.

External validation. We tested our model on the outside RodEpil dataset [44]. At first, it got a nearly perfect score. But this exposed a huge flaw in the data files. Normal clips ran at 30 frames per second with 301 frames. Seizure clips ran at 29 frames per second with 291 frames. The video framing was also very different between the two. To fix this, we changed how all the videos were saved. We made them

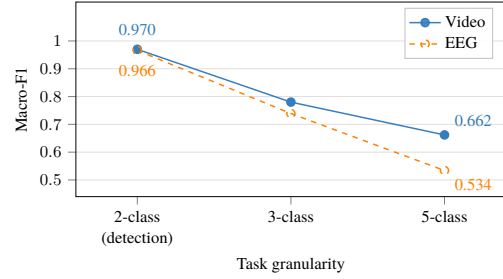


Figure 7. Scores for each task level (Table 1). EEG slightly wins basic detection (0.974 vs. 0.968). Video wins the 3-level (0.780 vs. 0.750) and 5-level grading (0.657 vs. 0.560).

Table 6. RodEpil [44] subject-wise five-fold evaluation. Uniform re-encoding removes the identified class-correlated encoding cues; R(2+1)D then reaches 0.979 mean F1 and 0.972 MCC.

	F0	F1	F2	F3	F4	Mean
Raw (leaked) F1	1.000	0.998	1.000	1.000	1.000	0.9996
Raw (leaked) MCC	1.000	0.998	1.000	1.000	1.000	0.9995
Re-encoded F1	0.965	0.978	0.997	0.977	0.979	0.979
Re-encoded MCC	0.957	0.976	0.996	0.967	0.964	0.972

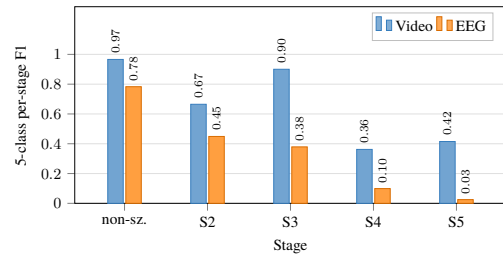


Figure 8. Per-class F1 for the 5-class diagnostic run (Table 14). Video and EEG are strongest on the common non-seizure and Stage 3 classes; performance is lower on the rare Stage 4–5 classes, with a larger drop for the EEG reference.

all 224×224 pixels and kept their original shape. We locked every single video to 30 frames per second and exactly 300 frames. We also used the exact same video format for everything. After this cleanup, our R(2+1)D model scored a 0.979 F1 and a 0.972 MCC (Table 6). This perfectly matches the 0.97 baseline score from the original dataset creators. This fix stops the model from cheating using frame rates or lengths. However, there might still be other hidden shortcuts left in the data. We show the results for all ten models on this fixed dataset in App. E.3.

5.3. Severe class imbalance

The 5-class environment is mathematically unstable: just about half of diagnostic validation intervals are actually non-seizures, an additional 41% are stage 3 (so that stage 5

clips are very sparse at 34, an almost 80:1 span) but episode values range over 11 and 3091. Episode counts hover fairly still between stages 3 and non-events. It's hard to establish that basic detection (at 12197 vs. 11934 for baseline vs active, an insignificant different) alone, but since granularity requires inverse frequency-weighted cross-entropy, stage 5 is as important mathematically as roughly 60 non-events are in this setting. The majority class baseline score is 0.136 on macro-F1, and the video achieve 0.657. It is still possible to detect Stage 5 (P/R/F1 = 0.579/0.324/0.415), yet the performance is poor compared to Stage 3 (F1 = 0.900); The result on EEG becomes only 0.202/0.123 on Stage 4/Stage 5. The class weighting helps to avoid falling to failure on rare stages, but doesn't manage well on fine-grained grade. We have not implemented unweighted ablation, so we don't argue the class weighting solution is the best one.

5.4. Discussion

These findings clearly differentiate the two modalities and can see that EEG is slightly stronger at detecting the seizures and video at rating the severity from observed behaviors. Both of these findings are relevant to the assigned task, since seizure severity is assigned by observed physical symptoms [38, 39] and a single channel is being used for our EEG benchmark. TheStage,2 reversal andStage,3 misses in the EEG test also illustrate that no one test is superior for every class.

In addition, video is a less invasive technique, and lighter systems show results very close to larger systems. Note that we do not measure latency, power consumption, implementation costs or real time streams, nor can we conclude about welfare improvements only based on accuracy. EEG is still indispensable for electrographic seizures with no motor component and the use of a multi-channel headset may also narrow the grading gap.

Our labels are obtained from integrated review of EEG and video, instead of standalone video and EEG annotation separately (see Sec. 4). Since EEG-assisted verification relies on EEG features whereas behavior defines severity on video features, the annotation biases would not be separable with current annotation strategy. In the future, independent unimodal annotation and blinded inter-rater agreement is recommended; learned multimodal fusion could also be explored.

6. Limitations and Future Work

Beyond the work presented in this paper, we discuss several limitations of our study. First, the data contained 20 rats, all from the same lab. Even though we tested our model with new rats (Table 5), the video score fell from 0.657 to 0.475. In future research, we hope to test model with several other groups, with several cameras setups and several kinds of seizure type. Second, precise seizure level has been

observed to be still difficult. It has especially become true with 4 and 5 stages with their rarity. More videos of high degree seizures have to be collected and a specific algorithm might deal better with ordered data (Sec. 3.2). Third, we only dealt with the disparity in the data amount using a single method: to weight the events more the less they are represented. We can test for example re-sampling [12] or focal loss [32] for improvement. Fourth, our test with EEG data utilized a single channel of measurement and did not integrate data from video with the EEG data. It would potentially give better result by using more channels of EEG signal with video data [26]. However such data are still very rare in literature (ex: RodEpi [44] is an animal dataset that only provides video). Fifth, we analyzed the performance with short extracted clips only whereas a real-world application would require to process non-stop recorded seizure data and identify precise start and stop time while keeping performance optimal for real time analysis as demonstrated with vEpiNet [31]. Testing our model on that kind of data, or at least test speed and energy consumption in general and live tracking would then be the subsequent logical step (App. G).

7. Conclusion

We measured video and EEG models, 10 video models and 7 EEG models in our case, for about 24000 paired video and EEG clips, from 20 rats. We found that EEG was better at simple classification of basic seizures, reaching an ROC of 0.974, compared to video at 0.968, but video was better at assessing different levels of seizure severity, scoring 0.657 vs 0.560 for EEG when evaluated in the five levels. And this remained true even if the model was evaluated on animals other than the ones it was trained on. The size of the video models is less than it matters as one of our small video models achieved nearly the same performance as our large model, whereas motion over time is key since models that looked at only single static frames were poorly performed. We applied our video model also on an external dataset, RodEpi and even though we needed to correct some flaws in their hidden data files, we achieved excellent performance, 0.979, confirming cameras as a convenient alternative to implantable probes to monitor epileptic activity. When integrated with EEG, since different methods have different strengths and weaknesses, could provide superior tools to monitor and predict disease.

References

- [1] Hoda Aghaei, Mohammad Mahdi Kiani, and Hamid Aghajan. Epileptic seizure detection based on video and EEG recordings. In *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2017. 3, 13, 14
- [2] Ijaz Ahmad et al. Advanced multimodal learning for seizure

- detection and prediction: Concept, challenges, and future directions. *arXiv*, 2026. arXiv:2601.05095. [13](#), [14](#), [31](#)
- [3] David Ahmedt-Aristizabal, Kien Nguyen, Simon Denman, Sridha Sridharan, Sasha Dionisio, and Clinton Fookes. Deep motion analysis for epileptic seizure classification. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3578–3581, 2018. [3](#), [12](#), [13](#)
- [4] David Ahmedt-Aristizabal, Mohammad Ali Armin, Zeeshan Hayder, Norberto Garcia-Cairasco, Lars Petersson, Clinton Fookes, Simon Denman, and Aileen McGonigal. Deep learning approaches for seizure video analysis: A review. *Epilepsy & Behavior*, 154:109735, 2024. [13](#)
- [5] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. [6](#), [16](#), [27](#)
- [6] Rachel A. Bergstrom, Jee Hyun Choi, Armando Manduca, Hee-Sup Shin, Greg A. Worrell, and Charles L. Howe. Automated identification of multiple seizure-related and interictal epileptiform event types in the EEG of mice. *Scientific Reports*, 3:1483, 2013. [2](#), [3](#), [12](#)
- [7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. [3](#), [5](#), [13](#), [15](#), [25](#), [27](#)
- [8] James P. Bohnslav, Nivanthika K. Wimalasena, Kelsey J. Clausen, Yu Y. Dai, David A. Yarmolinsky, Tomás Cruz, Adam D. Kashlan, M. Eugenia Chiappe, Lauren L. Orefice, Clifford J. Wolf, and Christopher D. Harvey. Deep-ethogram, a machine learning pipeline for supervised behavior classification from raw pixels. *eLife*, 10:e63377, 2021. [2](#), [3](#), [13](#)
- [9] Leo Breiman. Random forests. *Machine Learning*, 45(1): 5–32, 2001. [6](#), [16](#), [27](#)
- [10] Jiuwen Cao, Yuan Fang, Xiaonan Cui, Runze Zheng, Tiejia Jiang, and Feng Gao. Synchronized video and EEG based childhood epilepsy seizure detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(6):3742–3753, 2024. [3](#), [12](#), [13](#), [14](#)
- [11] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the Kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. [2](#), [3](#), [12](#)
- [12] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. [8](#), [28](#), [31](#)
- [13] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pages 785–794, 2016. [6](#), [16](#), [27](#)
- [14] Deba Prasad Dash, Maheshkumar Kolekar, Chinmay Chakraborty, and Mohammad R. Khosravi. Review of machine and deep learning techniques in epileptic seizure detection using physiological signals and sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–29, 2024. [12](#)
- [15] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015. [12](#)
- [16] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6824–6835, 2021. [5](#), [15](#), [25](#), [27](#)
- [17] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–213, 2020. [5](#), [14](#), [15](#), [25](#), [27](#)
- [18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019. [2](#), [3](#), [5](#), [13](#), [15](#), [25](#), [27](#)
- [19] Shreyan Ganguly, Zhanhong Jiang, Nyzil Massey, Nikhil Sanjay Rao, Thimmasettappa Thippeswamy, and Soumik Sarkar. Automated seizure detection in animal EEG signals. *bioRxiv*, 2025. Preprint. bioRxiv 2025.11.04.686618. [2](#), [3](#), [5](#), [6](#), [12](#), [15](#), [26](#), [27](#)
- [20] Jan Gorodkin. Comparing two k -category assignments by a k -category correlation coefficient. *Computational Biology and Chemistry*, 28(5–6):367–374, 2004. [17](#)
- [21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. [6](#), [15](#), [27](#)
- [22] Alexander I. Hsu and Eric A. Yttri. B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature Communications*, 12:5188, 2021. [3](#), [13](#)
- [23] Miranda J. Jankovic, Paarth Kapadia, and Vaishnav Krishnan. Home-cage monitoring ascertains signatures of ictal and interictal behavior in mouse models of generalized seizures. *PLoS ONE*, 14(11):e0224856, 2019. [3](#)
- [24] Tamás Karácsy, Anna Mira Loesch-Biffar, Christian Vollmar, Soheyl Noachtar, and João Paulo Silva Cunha. A deep learning architecture for epileptic seizure classification based on object and action recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4117–4121, 2020. [3](#), [12](#), [13](#)
- [25] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. [5](#), [12](#), [14](#)
- [26] Abdul-Mumin Khalid, Musah Sulemana, and Iddrisu Wahab Abdul. Equation-level parameterized fusion reformulation for multimodal epileptic seizure detection using interaction control and data-quality screening. *Frontiers in Signal Processing*, 6, 2026. [8](#), [13](#), [31](#)
- [27] Junaid Khan, Umar Zaman, Eunkyu Lee, and Kyungsup Kim. Innovative video-based approach for epileptic seizure

- detection in mice. In *IEEE Symposium on Computers and Communications (ISCC)*, pages 1–5, 2024. 3
- [28] Robert J. Kotloski. A machine learning approach to seizure detection in a rat model of post-traumatic epilepsy. *Scientific Reports*, 13:15807, 2023. 2, 3, 12
- [29] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEG-Net: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. 6, 15, 27
- [30] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4814, 2022. 5, 15, 25, 27
- [31] Nan Lin, Wei Gao, Lili Li, Jiayuan Chen, Zhiwei Liang, Gang Yuan, Hao Sun, Qi Liu, Jian Chen, Liqian Jin, Yining Huang, Xin Zhou, Shan Zhang, Peng Hu, Can Dai, Hao He, Yanxin Dong, Lili Cui, and Qiang Lu. vEpiNet: A multimodal interictal epileptiform discharge detection method based on video and electroencephalogram data. *Neural Networks*, 175:106319, 2024. 3, 8, 12, 13, 14, 31
- [32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 8, 28, 31
- [33] Yueyang Liu, Artemio Soto-Breceda, Philippa Karoly, David B. Grayden, Mark J. Cook, Dean R. Freestone, Daniel Schmidt, and Levin Kuhlmann. Seizure forecasting with multiple timescales and features. *Epilepsia*, 67(4):1778–1791, 2026. 1
- [34] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022. 5, 15, 25, 27
- [35] Tong Lu, Ke Xu, Zimo Zhang, Zitong Zhao, Danwei Weng, Ruiyu Wang, Miao Liu, Zizuo Zhang, Jingyi Yao, Yixuan Zhao, Wenchao Zhang, Min Wang, Guoming Luan, Minmin Luo, and Zhifeng Yue. A multimodal pre-trained network for integrated EEG-video seizure detection. *arXiv preprint arXiv:2604.26379*, 2026. 2, 3, 12, 13, 14
- [36] Grant F. Marshall, Alfredo Gonzalez-Sulser, and Catherine M. Abbott. Modelling epilepsy in the mouse: Challenges and solutions. *Disease Models & Mechanisms*, 14(3):dmm047449, 2021. 1, 2
- [37] Michael L. Martini, Aly A. Valliani, Claire Sun, Anthony B. Costa, Shan Zhao, Fedor Panov, Saadi Ghatan, Kanaka Rajan, and Eric Karl Oermann. Deep anomaly detection of seizures with paired stereoelectroencephalography and video recordings. *Scientific Reports*, 11(1):7482, 2021. 3, 13, 14
- [38] Nyzil Massey, Suraj S. Vasanthi, Christina Holtkamp, Christina Meyer, Nikhil S. Rao, Luis G. Gimenez-Lirola, Chong Wang, Hyun Im, Aditya S. Bevoor, Sridhar Kannurpatti, and Thimmasettappa Thippeswamy. Mitigating organophosphate nerve agent, soman (GD)-induced long-term neurotoxicity: Saracatinib, a Src tyrosine kinase inhibitor, as a potential countermeasure. *Journal of Neuroinflammation*, 2025. PMID: 40678225; PMCID: PMC12270223. 2, 3, 4, 8, 18
- [39] Nyzil Massey, Suraj S. Vasanthi, Christina Meyer, Nikhil S. Rao, Daniel R. Thedens, Chong Wang, Sridhar Kannurpatti, and Thimmasettappa Thippeswamy. Dual targeting of iNOS and Src tyrosine kinase as a superior therapeutic strategy against soman-induced long-term neurotoxicity: Multimodal biomarker, imaging, and neurobehavioral outcome analyses. *Acta Neuropathologica Communications*, 14(1):37, 2026. PMID: 41514296; PMCID: PMC12882386. 2, 3, 4, 8, 18
- [40] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie W. Mathis, and Matthias Bethge. DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, 2018. 2, 3, 13
- [41] Brian W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) – Protein Structure*, 405(2):442–451, 1975. 17
- [42] Aaron Mullen, Samuel E. Armstrong, Jasmine Perdeh, Björn Bauer, Jeffrey Talbert, and V. K. Cody Bumgardner. Multi-modal machine learning framework for automated seizure detection in laboratory rats. *arXiv preprint arXiv:2402.00965*, 2024. 2, 3, 12, 13, 14
- [43] Talmo D. Pereira, Nathaniel Tabris, Arie Matsliah, David M. Turner, Junyu Li, Shruthi Ravindranath, Eleni S. Papadoyannis, Edna Normand, David S. Deutsch, Z. Yan Wang, Grace C. McKenzie-Smith, Catalin C. Mitelut, Marielisa Diez Castro, John D’Uva, Mikhail Kislin, Dan H. Sanes, Sarah D. Kocher, Samuel S.-H. Wang, Annegret L. Falkner, Joshua W. Shaevitz, and Mala Murthy. SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods*, 19(4):486–495, 2022. 3, 13
- [44] Daniele Perlo, Vladimir Despotovic, Selma Boudissa, Sang-Yoon Kim, Petr V. Nazarov, Yanrong Zhang, Max Wintermark, and Olivier Keunen. RodEpiL: A video dataset of laboratory rodents for seizure detection and benchmark evaluation. *arXiv preprint arXiv:2511.10431*, 2025. Dataset: Zenodo, DOI 10.5281/zenodo.17601357. 2, 3, 7, 8, 12, 13, 17, 20, 23, 25, 26, 31
- [45] Ronald J. Racine. Modification of seizure activity by electrical stimulation: II. motor seizure. *Electroencephalography and Clinical Neurophysiology*, 32(3):281–294, 1972. 2, 3, 4, 13, 18, 31
- [46] Junming Ren, Zhoujian Xiao, Yujia Zhang, Yujie Yang, Ling He, Ezra Yoon, Stephen Temitayo Bello, Xi Chen, Dapeng Wu, Micky Tortorella, and Jufang He. EPIDetect: Video-based convulsive seizure detection in chronic epilepsy mouse model for anti-epilepsy drug screening. *arXiv preprint arXiv:2405.20614*, 2024. 3, 13
- [47] Gautam Sabnis, Leinani Hession, J. Matthew Mahoney, Arie Mobley, Marina Santos, and Vivek Kumar. Visual detection of seizures in mice using supervised machine learning. *Cell Reports Methods*, 2025. 3, 12, 13

- 741 [48] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek
742 Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Ba-
743 tra. Grad-CAM: Visual explanations from deep networks via
744 gradient-based localization. In *Proceedings of the IEEE In-*
745 *ternational Conference on Computer Vision (ICCV)*, pages
746 618–626, 2017. 6, 21
- 747 [49] Karen Simonyan and Andrew Zisserman. Two-stream con-
748 volutional networks for action recognition in videos. In *Ad-*
749 *vances in Neural Information Processing Systems (NeurIPS)*,
750 2014. 3, 12
- 751 [50] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xi-
752 aorong Gao. EEG Conformer: Convolutional transformer
753 for EEG decoding and visualization. *IEEE Transactions*
754 *on Neural Systems and Rehabilitation Engineering*, 31:710–
755 719, 2022. 6, 16, 27
- 756 [51] Jennifer J. Sun, Markus Marks, Andrew Wesley Ulmer, Di-
757 pam Chakraborty, Brian Geuther, Edward Hayes, Heng Jia,
758 Vivek Kumar, Sebastian Oleszko, Zachary Partridge, Milan
759 Peelman, Alice Robie, Catherine E. Schretter, Keith Shep-
760 pard, Chao Sun, Param Uttarwar, Julian Morgan Wagner,
761 Erik Werner, Joseph Parker, Pietro Perona, Yisong Yue,
762 Kristin Branson, and Ann Kennedy. MABe22: A multi-
763 species multi-task benchmark for learned representations of
764 behavior. In *Proceedings of the 40th International Con-*
765 *ference on Machine Learning*, pages 32936–32990. PMLR,
766 2023. 3, 13
- 767 [52] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang.
768 VideoMAE: Masked autoencoders are data-efficient learners
769 for self-supervised video pre-training. In *Advances in Neural*
770 *Information Processing Systems (NeurIPS)*, 2022. 3, 5, 13,
771 14, 15, 25, 27
- 772 [53] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann
773 LeCun, and Manohar Paluri. A closer look at spatiotempo-
774 ral convolutions for action recognition. In *Proceedings of*
775 *the IEEE/CVF Conference on Computer Vision and Pattern*
776 *Recognition (CVPR)*, pages 6450–6459, 2018. 2, 3, 5, 13,
777 14, 15, 25, 27
- 778 [54] Jan Van Erum, Debby Van Dam, and Peter Paul De Deyn.
779 PTZ-induced seizures in mice require a revised Racine scale.
780 *Epilepsy & Behavior*, 95:51–55, 2019. 2, 3, 4, 18, 31
- 781 [55] Ziwei Wang, Siyang Li, and Dongrui Wu. Canine EEG helps
782 human: Cross-species and cross-modality epileptic seizure
783 detection via multi-space alignment. *National Science Re-*
784 *view*, 12(6):nwaf086, 2025. 14
- 785 [56] Paige Whyte-Fagundes, John Efromson, Anjelica Vance,
786 Samuel Carpenter, Aurélien Bègue, Aloe Carroll, Thomas
787 Jedidiah Jenks Doman, Mark Harfouche, and Scott C. Bara-
788 ban. Automated detection of complex zebrafish seizure be-
789 havior at scale. *Communications Biology*, 8(1):872, 2025.
790 3
- 791 [57] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and
792 Kevin Murphy. Rethinking spatiotemporal feature learning:
793 Speed-accuracy trade-offs in video classification. In *Pro-*
794 *ceedings of the European Conference on Computer Vision*
795 *(ECCV)*, pages 305–321, 2018. 5, 14, 15, 25, 27
- 796 [58] Yankun Xu, Junzhe Wang, Yun-Hsuan Chen, Jie Yang, Wen-
797 jie Ming, Shuang Wang, and Mohamad Sawan. VSViG:
Real-time video-based seizure detection via skeleton-based
spatiotemporal ViG. In *Proceedings of the European Con-*
ference on Computer Vision (ECCV), 2024. 3, 13
- [59] Andrew Zayachivsky, Mark J. Lehmkuhle, and F. Edward
Dudek. Long-term continuous EEG monitoring in small
rodent models of human disease using the Epoch wireless
transmitter system. *Journal of Visualized Experiments*, 101:
e52554, 2015. 2
- [60] Xizhen Zhang, Xiaoli Zhang, Qiong Huang, and Fuming
Chen. A review of epilepsy detection and prediction methods
based on EEG signal processing and deep learning. *Frontiers*
in Neuroscience, 18, 2024. 12
- [61] Valerii A. Zuev, Elena G. Salmagambetova, Stepan N.
Djakov, and Lev V. Utkin. Automated video-EEG analy-
sis in epilepsy studies: A narrative review of advances and
challenges. *Journal of Medical Systems*, 49, 2025. 14

A. Supplementary: A Survey of Video and Video-EEG Seizure Analysis

This appendix places our experimental work within the broader research landscape. We break down past studies across two main categories—data type (EEG-only, video-only, and combined video-EEG) and test subject (human vs. animal)—while reviewing the computer vision and behavioral analysis tools that power automated tracking. The main paper’s Related Work section (Sec. 2) offers a condensed summary of this full survey.

A.1. Scope and Taxonomy

We focus on automated systems that analyze EEG, video, or paired video-EEG to spot seizures, track behavioral types, grade severity, or pinpoint exact timing. Broadening our scope like this is essential, as recent animal video research has moved far beyond simple yes-or-no detection.

Axis 1: modality. *EEG-only* systems are the long-standing standard, using everything from hand-crafted frequency metrics to deep recurrent, convolutional, and attention networks. *Video-only* tools spot seizure behaviors directly from raw video pixels, movement patterns, or tracked body poses. *Video-EEG* methods combine physical movement and brain-wave data, in theory taking advantage of shared signals while covering for each tool’s unique blind spots.

Axis 2: subject. Hospital epilepsy monitoring units generate most of the computational video-EEG research, while animal models remain essential for testing new treatments. Although automated video tracking in animals is growing fast, combined video-EEG learning, side-by-side modal tests, and detailed severity benchmarks are still surprisingly rare.

Table 7 summarizes this landscape. Single-modality EEG analysis is highly mature. Video-only animal tracking has expanded beyond basic detection to include severity estimation and detailed behavioral profiling. Yet, paired animal video-EEG remains the least explored area, particularly for severity-aware grading and open, shared benchmarks.

A.2. EEG-Only Automated Detection

This subsection expands on the EEG-only overview from Sec. 2. As the long-standing diagnostic standard, automated EEG sets the performance baseline for all video and fusion approaches.

Classical machine learning. Early rodent systems paired hand-crafted features with traditional classifiers. Bergstrom *et al.* [6] used wavelet decomposition and signal variance to detect spikes and seizures in mouse EEG, reaching 99% accuracy and 91% precision against expert ratings. This proved automated detection was viable, though heavily dependent on manual feature engineering. Comprehensive reviews [14, 60] catalog the wide range of frequency, entropy, and signal connectivity features developed during this era.

Table 7. The research landscape for automated seizure analysis broken down by data type and subject. The clearest open gap is multi-level severity grading using paired animal video-EEG data.

Setting		Representative progress	Remaining gap
EEG-only,	animal	GRU/LSTM, CNN and temporal models [19, 28]	Behavioral information
Video-only,	human	Detection and semiology analysis [3, 24]	Generalization, deployment
Video-only,	animal	Severity [47]; RodEpi [44]	Matched EEG comparison
Video+EEG,	human	Synchronized and fused systems [10, 31]	Multi-site validation
Video+EEG,	animal	Recent multimodal studies [35, 42]	Severity; paired benchmarks

Deep learning. Research eventually shifted toward deep representation learning. Kotloski [28] fed rat EEG spectrograms into a pretrained GoogLeNet, reaching 95.6% sensitivity across >2,200 hours of recording in a post-traumatic model—proving standard CNNs adapt well to signal spectrograms. Later, Ganguly *et al.* [19] applied GRU and LSTM networks to a chemically diverse rat cohort (Soman/GD and Kainate). Their recurrent models easily beat traditional Random Forest and XGBoost baselines, with GRU recall topping 90% for most animals. Because these approaches rely strictly on EEG signals, they form the baseline family that we expand to seven distinct models in our study.

A.3. Video-Only Seizure Analysis

This subsection expands on the video-based methods introduced in Sec. 2. Decoding seizures from video fundamentally reframes behavioral assessment as an action-recognition challenge. We outline the core architectures driving this approach and their specific applications to epilepsy.

A.3.1. Video-Understanding Backbones

Modern video seizure systems draw from a few core design families. Two-stream networks combine still frames with optical-flow motion tracking [49]. Recurrent CNNs attach an LSTM to pull time-based context from frame-by-frame features [15]. 3D convolutional networks, like I3D, learn spatial and timing patterns directly, benefiting heavily from large-scale video pretraining [11, 25]. Building on

this, R(2+1)D [53] breaks those 3D convolutions into separate spatial and time steps, serving as our default model. SlowFast [18] splits the workload entirely, analyzing visual appearances slowly while tracking fast, high-frequency motion on a separate pathway. More recently, video transformers like TimeSformer [7] and the self-supervised Video-MAE [52] have introduced powerful, attention-based alternatives. Taking a completely different angle, tools like DeepLabCut [40] and SLEAP [43] track animal poses without physical markers, boiling convulsive movements down into highly compact data points.

Parallel advances in computational ethology automate the ethogram itself. DeepEthogram [8] handles supervised behavior classification directly from raw video pixels; B-SOiD [22] extracts recurring behavioral patterns from pose movement; and MABe22 [51] benchmarks learned representations across multiple species. While these tools were not designed explicitly for seizure tracking, they prove that complex, time-dependent animal behaviors are highly learnable at scale.

A.3.2. Human and Animal Seizure Video

Clinical video tools largely focus on basic seizure detection and symptom profiling. Ahmedt-Aristizabal *et al.* [3] modeled body movements directly from hospital footage, while Karacsony *et al.* [24] merged object tracking with action recognition. Pushing toward real-time monitoring, Xu *et al.* [58] introduced VSViG, a skeleton-based spatiotemporal vision-graph model. A comprehensive review by Ahmedt-Aristizabal *et al.* [4] captures this fast-growing field, highlighting how automated video could drastically cut the heavy workload of long-term patient screening.

Animal research aligns perfectly with the behavioral focus of our study. Because the Racine scale [45] judges severity entirely on visible physical movements, video is the ideal tool for automated scoring. Sabnis *et al.* [47] proved that overhead cameras can track both active seizures and their shifting intensity in PTZ-induced mice. Similarly, Ren *et al.* [46] built EPIDetect to screen chronically epileptic mice for convulsive seizures using only video.

Public benchmarks are also emerging. Perlo *et al.* introduced *RodEpilep* [44], sharing 13,053 overhead clips from 19 rats alongside a TimeSformer baseline tested using subject-wise cross-validation. Together, these studies prove that video models can successfully track rodent seizure behaviors across multiple levels of detail.

Our research asks a different question. Instead of simply proving that video can detect or grade seizures, we pit video directly against synchronized EEG on the exact same experimental events. This perfectly matched setup lets us track exactly how their relative strengths shift as the task moves from basic detection to precise severity grading.

A.4. Multimodal Video-EEG Fusion

Even though paired video-EEG is the clinical gold standard, computationally combining the two data streams is still surprisingly rare. Table 8 outlines the most notable clinical systems doing this today. In animal research, recent work includes Mullen *et al.* [42], who combined video, brain surface sensors, and motion-tracking pads in rats. Similarly, EEGVFusion [35] uses cross-modal attention to sync self-supervised EEG and video data in mice. Crucially, however, both of these systems stick entirely to basic yes-or-no seizure detection rather than attempting to grade exact behavioral severity.

This exact gap drives our approach. Before building complex models that fuse both data streams, we first need to isolate exactly what video and EEG can achieve on their own—and where their blind spots differ. Establishing these single-modality baselines is an essential first step before we can properly judge any future combined systems.

From early fusion to anomaly detection. Aghaei *et al.* [1] pioneered early efforts to combine video and EEG, highlighting how physical movements and electrical brain waves naturally complement each other. Later, Martini *et al.* [37] framed paired stereo-EEG and video analysis as a deep anomaly detection problem. Instead of explicitly searching for seizures, their system learns what normal baseline activity looks like and flags seizures as mathematical outliers. This strategy is especially valuable when data is scarce, as it bypasses the need for massive datasets of perfectly labeled seizures.

Synchronized and system-level fusion. Cao *et al.* [10] analyzed synchronized data in childhood epilepsy, pointing out exactly how crucial it is to keep the timing perfectly aligned and to manage sensors that temporarily drop out. Taking a more structural approach, vEpiNet [31] combined patient tracking, video motion and pose data, and EEG features before feeding them all into a single fusion model. Its real-world success proves that blending these different data streams is highly practical, setting a valuable blueprint for building similar systems in animal research.

Fusion taxonomy. Research generally breaks data integration down into three main approaches: combining raw data early on, merging learned features in the middle, or blending final predictions at the very end [2, 26]. A constant headache, however, is that signal quality rarely stays stable. Rigid models that weigh both inputs equally tend to fall apart the second a camera view is blocked or sudden movement scrambles the EEG. Because of this, it is critical to build adaptive, quality-aware systems that can shift their trust on the fly to whichever sensor is most reliable at that exact moment—especially when tracking freely moving animals.

Table 8. Representative human video-EEG systems. Animal fusion remains sparse and has concentrated almost entirely on simple yes-or-no seizure detection [35, 42].

Study	Year	Method / vision component	Key result
Aghaei <i>et al.</i> [1]	2017	Combined video and EEG recordings for detection	Early video+EEG fusion demonstration
Martini <i>et al.</i> [37]	2021	Deep anomaly detection, paired stereo-EEG + video	Label-free seizure flagging
Cao <i>et al.</i> [10]	2024	Synchronized video-EEG, pediatric cohort	Detection in childhood epilepsy
Lin <i>et al.</i> (vEpiNet) [31]	2024	Patient detection + frame-diff/keypoints + EfficientNetV2, MLP fusion	AUROC 0.9902; false positives reduced by about one third

A.5. The Data Bottleneck

The biggest roadblock in the field is still the severe lack of matched, expertly labeled video-EEG datasets [2, 61]. Clinical datasets are locked down by strict patient privacy laws and recorded on a messy mix of different hardware. Animal studies bypass the privacy issues but run into massive logistical headaches instead: invasive surgical sensors, wildly varying seizure rates from one animal to the next, and mismatched camera setups. Together, these hurdles make labeling data and testing models across different subjects incredibly expensive and time-consuming.

Two strategies offer a way forward. Self-supervised video pretraining drastically cuts down the need for raw, manually labeled data [52], while transfer learning across different species or data types lets models borrow insights from existing, related datasets [55]. Even though neither method completely replaces the need for perfectly paired benchmarks, they take a massive chunk out of the manual grading required to build reliable animal tracking systems.

B. Model Architectures and Training Details

To guarantee full reproducibility, we outline the video and EEG models evaluated in Sec. 5 alongside their standardized training routines. Every model is built on established, off-the-shelf architectures; we only adjust the classification heads and input layers. Because of this, our study tests the raw potential of different data streams and backbone *families* rather than relying on highly customized engineering tricks. The models selected for each data type intentionally span a wide range of design philosophies. This ensures our broader conclusions reflect the fundamental strengths of the modality itself, rather than the unique quirks of a specific neural network. Video backbone parameter counts match Table 2.

B.1. Video backbones

The ten video networks we selected cover a broad spectrum of designs, ranging from standard convolutions to multi-scale frameworks and attention-based models (both hierarchical and global). They mainly differ in how they stitch together spatial and temporal data, how they expand their field of view, and how strictly they enforce local structure. Every backbone starts with pre-trained Kinetics-400 [25] weights and processes clips at its native length and resolution (usually 16 to 32 frames at 112 to 224 px), with R(2+1)D defaulting to 16 frames at 112 × 112. We keep the core networks intact, modifying only the final linear layer to handle our specific classification tasks.

Factorized and separable convolution. R(2+1)D-18 [53] (33M), our main visual baseline, breaks 3D filters down into separate 2D spatial and 1D time-based steps. The network carefully sizes these intermediate layers so the split block uses the exact same computing power as the original 3D version. Because of this, any performance boost comes from the extra nonlinear steps and a smoother learning process, rather than simply throwing more parameters at the problem. S3D [57] (8M) applies a similar trick to an Inception-style layout. It saves full 3D convolutions for the deeper layers where spatial details are already compressed. This “top-heavy” design cuts its overall size down to roughly a quarter of R(2+1)D. Both models enforce a strict local focus: their field of view expands steadily layer by layer, forcing them to build up long-term movement patterns step by step.

Efficient convolution. X3D-M [17] (4M) builds up from a small 2D image classifier by expanding just one dimension at a time—whether that is video length, frame rate, resolution, width, or depth. At each step, it keeps only the changes that provide the biggest accuracy boost for the least computational effort. The resulting model is incredibly lightweight. We included it specifically to test whether reliable seizure detection actually requires a massive neu-

ral network, or if a highly streamlined one can effectively handle the task.

Two-pathway convolution. SlowFast-R50 [18] (35M) explicitly splits visual appearance from physical motion. A Slow pathway takes in fewer frames but uses a heavy feature set to understand the overall scene, while a Fast pathway processes a rapid, dense stream of frames using a lightweight network purely to track quick movements. Side connections then continuously feed these high-speed motion cues back into the slower visual stream. This dual-track design is perfect for picking up subtle, motion-sensitive actions, ultimately scoring the highest detection AUROC in our tests—even when compared to closely related architectures (Table 16).

Hierarchical and multiscale attention. Video Swin-T/S [34] (28M/50M) swap out standard convolutions for self-attention, but they still maintain a strict local structure. Instead of looking at the whole video at once, they calculate attention inside small 3D windows. These windows shift between layers to connect neighboring regions, gradually shrinking the overall resolution step by step. MViTv1-B/MViTv2-S [16, 30] (36M/34M) build a similar hierarchy by pooling their attention keys and values. This forces the spatial and temporal resolution down while expanding the channel width, creating an internal feature pyramid (with MViTv2 adding relative position tracking and residual pooling for better information flow). While these models do not enforce local structure as rigidly as standard convolutions, they stay far more focused on local details than fully global attention networks.

Global attention and self-supervised pretraining. TimeSformer [7] (121M) ditches standard convolutions completely. Instead, it uses a split space-time attention approach—first tracking how a specific patch changes over time, then analyzing how all patches relate within a single frame. This setup gives the network an instant, bird’s-eye view of the entire video clip. VideoMAE-B [52] (86M) takes a different route with a flat, non-hierarchical transformer. It learns to recognize structure through an extreme pretraining phase, where it is forced to successfully reconstruct a video even when over 90% of the visual data is heavily masked. Interestingly, despite their massive size and ability to see the whole clip at once, both of these models actually scored lower in our detection tests when restricted to a strict training budget (Table 2). This drop in rank does not necessarily mean global attention is a flawed strategy; rather, it highlights just how heavily these architectures rely on massive amounts of data and exhaustive optimization tuning to reach their full potential.

Across these different model families, network sizes swing wildly from 4 million to 121 million parameters. Despite this massive gap in scale, their detection macro-F1 scores all cluster tightly within a narrow 0.03 margin.

This clearly shows that the sheer size of the backbone is a secondary concern; what actually drives performance is the rich, time-based evidence naturally embedded within the input data itself (Sec. 5).

Table 9. Video backbones evaluated in Table 2, outlining their core design paradigms and overall parameter counts. Network capacities range broadly from 4M to 121M parameters. The smallest model is marked in **bold**, while the second-smallest is underlined.

Backbone	Paradigm	Par.
MViTv2-S [30]	multiscale	34M
SlowFast-R50 [18]	two-pathway	35M
R(2+1)D-18 [53]	conv	33M
X3D-M [17]	eff. conv	4M
MViTv1-B [16]	multiscale	36M
Video Swin-S [34]	hier. attn.	50M
S3D [57]	sep. conv	<u>8M</u>
Video Swin-T [34]	hier. attn.	28M
VideoMAE-B [52]	self-sup. ViT	86M
TimeSformer [7]	full attn.	121M

B.2. EEG models

We evaluate seven distinct EEG networks, covering a mix of recurrent, convolutional, attention-based, and traditional feature-driven designs. Every model processes the main recording channel, which we first downsample to 125 Hz. We then slice this continuous signal into overlapping 6-second windows (with a 3-second stride, yielding 750 samples each) and independently z-score them to standardize the data. To generate a final prediction for an entire clip, we average the log-probabilities from all of its individual windows, perfectly matching the setup described in Sec. 5.

Recurrent. The GRU [19] our baseline architecture for rodent EEG and the LSTM [21] process these time windows step by step, with both utilizing a hidden size of 128. At every step, built-in gating mechanisms decide exactly what information to keep and what to throw away. While the LSTM spreads a separate cell state across three distinct gates, the GRU condenses this design into just two, trading a slight drop in capacity for a much lighter parameter footprint. Because they inherently read data in a strict sequence, neither model has a built-in understanding of frequency. This forces them to learn spectral patterns entirely from scratch using the raw electrical waves, while completely preventing the rapid, parallel processing that convolutional or attention-based models enjoy.

Compact convolutional filter bank. EEGNet [29] bakes traditional EEG signal processing directly into its design while maintaining an incredibly tiny parameter count. It starts with a temporal convolution that acts as a learned fre-

quency filter, follows up with a depthwise convolution for spatial filtering, and finishes with a separable convolution to pull everything together. Because EEGNet was originally built for complex, multi-electrode brain-computer interfaces, adapting it to our single-channel setup heavily tests its time-filtering abilities. With only one electrode to analyze, the spatial filters are left with practically nothing to combine (Table 4).

Convolution-attention hybrid. The EEG Conformer [50] features a straightforward two-part design. First, a convolutional layer chops the raw signal window into smaller, localized time chunks. Next, a self-attention module steps in to compare these chunks across the entire window. This setup makes it the only model in our lineup that can directly connect far-apart moments in time, completely bypassing the step-by-step delays and rigid viewing windows that limit other networks.

Dilated convolution. The TCN [5] layers its convolutions with exponentially growing gaps between data points. This design rapidly widens the network’s field of view while allowing it to process everything simultaneously rather than sequentially. As a result, it can digest an entire 6 s window instantly, completely avoiding the step-by-step bottlenecks that slow down recurrent models. This efficiency makes it the strongest performing EEG model in our tests based on macro-F1 scores (Table 4).

Classical features with tree ensembles. Random Forest [9] (100 trees, balanced class weights) and XGBoost [13] (300 trees, depth 6, learning rate 0.1) skip deep feature learning entirely. Instead, they rely on a fixed set of eleven hand-picked metrics per window. These include five standard brain wave bands (delta 0.5-4 Hz, theta 4-8 Hz, alpha 8-13 Hz, beta 13-30 Hz, gamma 30-60 Hz) and six basic time-based statistics (mean, standard deviation, RMS, line length, activity, and zero crossings)—the traditional building blocks of early animal seizure tracking. While Random Forest averages the results of many independent trees, XGBoost builds its trees one after another, with each new tree specifically designed to fix the mistakes of the last. Even though XGBoost holds its own against EEGNet for basic yes/no seizure detection, both of these traditional models fall far behind the deep learning architectures when tasked with detailed severity grading (Sec. 5). Importantly, because every EEG model here only looks at a single channel, this comparison highlights exactly what we can squeeze out of one electrode, rather than testing the absolute limits of complex, multi-channel setups.

B.3. Training and evaluation protocol

To ensure a perfectly fair comparison, every model is trained and tested on the exact same data splits. We strictly separate the data by recording session so that training and testing data never mix (using a fixed seed of 49). To han-

dle unbalanced data, we weight rarer events more heavily during training, and we judge overall performance using macro-averaged metrics. These identical splits and starting seeds are kept consistent across all modality tests.

The video backbones are fine-tuned using the AdamW optimizer (learning rate 3×10^{-4} , weight decay 10^{-4}), alongside cosine annealing, mixed precision, a batch size of 16, and a strict 12-epoch limit. (As a minor exception, Video Swin and MViTv2 needed a slightly lower learning rate of 10^{-4} during the RodEpi tests to keep their training stable).

For the deep EEG models, we also use AdamW (learning rate 10^{-3} , weight decay 10^{-4}) and cosine annealing, but with a larger batch size of 256 and a 30-epoch limit, optimizing a weighted cross-entropy loss. The classical EEG models simply rely on standard balanced class weighting. Finally, across all experiments, we generate the final prediction for an entire clip by averaging the log-probabilities of its smaller, individual EEG windows.

B.4. Metric definitions

Every score we show is based directly on a single validation run, unless we specifically mention it is an average. We are writing out the exact math here to help clear up any confusion between the macro-F1 score and the Matthews correlation coefficient (MCC) that we talk about in Sec. 3 and Sec. 5.2.

Notation. When sorting data into K different categories, we use a confusion matrix C to keep track of our results. In this grid, C_{ij} is simply the number of clips that actually belong to category i but the model guessed as category j . We use N for the total number of clips across the board. For any specific category k , t_k is the total number of clips that truly belong to it, while p_k is the number of times the model predicted it. Using these counts, we figure out our true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN):

$$\begin{aligned} TP_k &= C_{kk}, & FP_k &= p_k - C_{kk}, \\ FN_k &= t_k - C_{kk}, & TN_k &= N - p_k - t_k + C_{kk}. \end{aligned} \quad (7)$$

Precision, recall and F1. For every single category, we calculate precision (P) and recall (R), which together give us the F1 score:

$$P_k = \frac{TP_k}{TP_k + FP_k}, \quad R_k = \frac{TP_k}{TP_k + FN_k}, \quad (8)$$

$$F_{1,k} = \frac{2P_k R_k}{P_k + R_k} = \frac{2 TP_k}{2 TP_k + FP_k + FN_k}. \quad (9)$$

To get our overall numbers, we simply average the individual category scores together evenly (an unweighted macro-average):

$$P_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K P_k, \quad \text{macro-F1} = \frac{1}{K} \sum_{k=1}^K F_{1,k}, \quad (10)$$

and we do the exact same thing for overall recall. It is important to note that the macro-F1 score averages the individual $F_{1,k}$ scores directly. We do not calculate it by plugging the overall macro-precision and macro-recall into the F1 formula, because that shortcut completely breaks down when dealing with unbalanced datasets. (As a side note, macro-recall is mathematically the same as balanced accuracy). Finally, for our RodEpi tests (Tables 6 and 12), we only report the score for the positive class (Eq. (9)) so our results line up perfectly with how the original researchers published their dataset [44].

AUROC. For a simple yes/no task, let s_i be the model's confidence score that a clip is positive. If we group all the true positive clips into set $\mathcal{P}$ and all the negative ones into set $\mathcal{N}$, the AUROC simply calculates the odds that a randomly chosen positive clip gets a higher score than a random negative one:

$$\text{AUROC} = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} \left[\mathbf{1}(s_i > s_j) + \frac{1}{2} \mathbf{1}(s_i = s_j) \right]. \quad (11)$$

When we grade multiple severity levels, we report a “one-versus-rest” average. This means we pick one specific category k to act as the positive target, lump everything else together as the negative group, run the math, and then average those results across all categories. Unlike the F1 score or MCC, the AUROC doesn't rely on picking a rigid cut-off threshold. Instead of looking at the model's final, hard decisions, it evaluates the raw confidence scores directly.

Matthews correlation coefficient. The MCC [41] essentially measures the statistical correlation between the true labels and what the model actually predicted. When dealing with K different categories, it is calculated as [20]:

$$\text{MCC} = \frac{N \sum_k C_{kk} - \sum_k p_k t_k}{\sqrt{(N^2 - \sum_k p_k^2)(N^2 - \sum_k t_k^2)}}, \quad (12)$$

which simplifies nicely for a basic two-class setup ($K=2$) down to:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{p_1 p_2 t_1 t_2}}. \quad (13)$$

Unlike our precision, recall, and F1 scores (Eqs. (8)-(10)) that range from 0 to 1, the MCC spans from -1 to 1 , where a score of 0 means the model is just guessing randomly.

Crucially, the MCC formula does not average things out category by category. Instead, every single clip is thrown directly into the main calculation, meaning the most frequent categories heavily drive the final score. This highlights a fundamental difference between how these metrics behave (Sec. 5.2). If a model completely fails to recognize a rare seizure stage, the macro-F1 score takes a massive, flat hit (dropping by a full $1/K$) regardless of how few clips actually belonged to that stage. The MCC, however, barely

registers the mistake, penalizing the model only in proportion to how rarely that stage actually occurs.

Aggregation over runs. Whenever one of our tables reports a mean $\pm$ std across M runs (for example, the five random seeds in Table 1 or the five cross-validation folds in Tables 5 and 12), we always calculate the score for each run completely independently before averaging them together:

$$\bar{x} = \frac{1}{M} \sum_{m=1}^M x_m, \quad \sigma = \sqrt{\frac{1}{M-1} \sum_{m=1}^M (x_m - \bar{x})^2}, \quad (14)$$

We use $M = 5$ consistently across all our experiments. Crucially, we never lump the raw predictions from different folds into one giant confusion matrix. Instead, $\bar{x}$ simply represents the average of the final scores calculated from each individual fold.

Macro-F1 vs. MCC on our results. Even though Macro-F1 and MCC handle highly unbalanced data differently, they both clearly point to the exact same shift in performance between the two data types. For the complex 5-class grading task, the video model scores a 0.657 macro-F1 and a 0.831 MCC, easily outperforming the baseline EEG (GRU) model, which only manages 0.560 and 0.728. If we track the performance gap between video and the GRU, the macro-F1 difference shifts from -0.006 (for simple yes/no detection) to $+0.030$ (for 3-class grading) and finally to $+0.097$ (for 5-class grading). The MCC perfectly mirrors this exact upward trend: $-0.013 \rightarrow +0.030 \rightarrow +0.103$.

Ultimately, both metrics prove our main theory: EEG holds a tiny edge for basic event detection, but video completely takes over when the task requires tracking detailed behavioral stages. When we test the models on entirely unseen subjects, the score variations across different runs stay relatively stable for basic detection, but they swing much more wildly during detailed grading—especially for the EEG baseline. Because of this, our overall conclusions stand strong regardless of which mathematical metric you prefer.

B.5. EEG preprocessing sensitivity

We tested several different window sizes, stride lengths, sampling rates, and methods for combining clip scores, keeping our baseline GRU model and data splits exactly the same throughout. These early tests revealed that how we combine the predictions matters much more than tiny tweaks to the window length itself. Specifically, using log-mean pooling performed significantly better than taking a simple average when dealing with long, varied seizure clips. As a result, we locked in a standard setup for all seven of our final models: 6-second windows, a 3-second stride, a 125 Hz sampling rate, and log-mean pooling to generate the final clip prediction.

We ran these initial checks before generating our finalized EEG training dataset. To keep our main results completely clean and avoid mixing up any data, we excluded the scores from this early sweep when making our final comparisons between video and EEG models. While finding the absolute perfect preprocessing setup would require rerunning all of these tests on the final data, we are documenting these checks here simply to explain the reasoning behind the baseline settings we used across all our experiments.

C. Extended Dataset Details

This section breaks down the complete details supporting Sec. 4. Here, we cover exactly how the data was collected, how we labeled the clips, our rules for grading seizure severity, the makeup of our animal cohort, and our ethical guidelines.

C.1. Acquisition and synchronization

The original lab collected all of these recordings under a US federal research grant, strictly following the methods laid out in [38, 39]. Adult Sprague-Dawley rats (7-8 weeks old, weighing 250-300 g; Charles River) were exposed to the nerve agent soman (GD, 132 $\mu\text{g/kg}$, injected under the skin) inside a highly regulated chemical facility. To manage the severe initial bodily reactions, the rats were given a mix of HI-6 (125 mg/kg) and atropine sulfate (2 mg/kg) within 60 seconds of exposure. An hour later, they received midazolam (3 mg/kg) to stop the initial, continuous seizure. Every single animal analyzed in our study comes exclusively from this specific soman-exposed group; we did not include data from any other seizure-inducing chemicals.

After recovering, the animals had wireless monitoring devices surgically implanted. EEG sensors were carefully attached directly to the membrane covering the brain (the dura mater), and a small wireless transmitter was tucked just under the skin on their side. The rats then lived individually in cages placed on top of special receiver plates (RPC-1 by Data Sciences International), which monitored them 24/7. Because a single computer system stamped the exact time on both the overhead video feed and the brain waves simultaneously, the two data streams are perfectly synced right from the start. This completely eliminated the need for us to try and manually align the data later, a process that is famously prone to errors.

We recorded their spontaneous, recurring seizures roughly 10 to 18 weeks after the initial exposure, during the chronic phase of epilepsy. The brain wave devices originally captured data at a high rate of 1 kHz, which we downsampled to 125 Hz for our experiments. Throughout this time, a fixed camera mounted directly above each home cage recorded the animals. Since the exact camera framing and cage sizes varied slightly from setup to setup, we

applied fixed, custom crops to each cage’s video to ensure a consistent view before mixing all the dataset clips together.

C.2. Annotation and event boundaries

Undergraduate research assistants, who were kept blind to the specific treatments the animals received, reviewed the perfectly synced video and EEG data using NeuroScore software (v3.4.0, DSI). They pinpointed the exact start and end times for every seizure and assigned a severity grade based on the animal’s behavior. While the software automatically flagged potential seizures by looking at the electrical brain waves, the human reviewers double-checked every single suggestion against the video feed. They strictly threw out any false alarms triggered by simple electrical noise, or normal activities like grooming and exploring [38, 39]. Every clip we use in this study is based directly on these carefully verified timestamps.

We define the start of an event as the very first noticeable change in behavior, and the end as the moment the animal completely returns to normal. Because abnormal electrical brain activity usually kicks off before any physical symptoms actually appear, we padded every clip with an extra 10 s buffer before the official start and after the official end to capture the full picture.

It is incredibly important to note that because the reviewers looked at both the video and the EEG at the exact same time to create these labels, our “ground truth” is not completely independent for each data type. The brain waves inherently help flag *that* an event is happening, while the video dictates *how severe* it is. This intertwined labeling process fundamentally shapes how we have to interpret the performance comparisons between our video and EEG models.

Matched negatives. We pair every single seizure clip with a normal, non-seizure clip taken from the exact same animal during the exact same recording session. Because of this careful matching, both the positive and negative samples share the identical animal appearance, camera angle, and lighting conditions (Fig. 9). This setup forces the network to actually learn seizure behavior rather than simply cheating by memorizing static background details, although it does not completely eliminate the risk of the model finding clever visual shortcuts.

C.3. Behavioral severity scale

The original lab grades these nerve-agent seizures using a custom 5-point behavioral severity scale [38, 39]. Although it is loosely based on the well-known Racine scale [45], it has been heavily modified specifically to track the unique symptoms caused by soman exposure. Therefore, it should never be confused with the standard Racine scale [54].

Stage 1 covers purely involuntary bodily responses, such as drooling, tearing up, urination, defecation, and invol-



Figure 9. A visual breakdown of our basic yes/no detection task. We show four sample frames from a normal, non-seizure period (top) alongside four frames from an active seizure (bottom). Because both examples feature the *same* rat during the *same* recording session, the animal’s appearance, the camera angle, and the overall background perfectly match. This careful pairing drastically reduces the chance that the model can cheat by memorizing static visual shortcuts in the scene. (All clips have been de-identified).

untary chewing (mastication). Stage 2 introduces obvious physical movements, including tremors, “wet-dog” shakes, sudden neck jerks, and severe arching of the back or spine (kyphosis and opisthotonos). Stage 3 ramps up the intensity with rhythmic twitching in the front legs (forelimb clonus), a stiff, upright tail (Straub tail), rearing up on the hind legs, and rigid limb stretching. Stage 4 involves repeated rearing, intense front-leg twitching, a constantly stiff tail, and a complete loss of balance that leads to falling over. Finally, Stage 5 represents the most severe level: a full-blown, sustained convulsive seizure (generalized tonic-clonic).

From a clinical standpoint, the lab considers Stages 1 and 2 to be non-convulsive, while Stages 3 through 5 are classified as convulsive. It is important to highlight that this clinical dividing line is very different from the 3-class system we use in our experiments. As explained in Sec. 3, we intentionally lumped Stages 2 and 3 together, and separated them from Stages 4 and 5, to create a more balanced, middle-ground diagnostic challenge for our models. You can see frame-by-frame visual examples of Stages 2 through 5 in Fig. 10.

C.4. Rare stages and long-tailed burden

Stage 1 events are virtually nonexistent in our data, amounting to just 6 clips. Because its symptoms are mostly internal bodily functions, they are almost impossible to catch on an overhead camera and rarely even trigger the EEG’s automatic detection system. Because of this, we completely excluded Stage 1 from our models. Full-blown Stage 5 convulsions are also quite rare during this chronic phase of epilepsy, providing only 193 clips.

After filtering, our final dataset contains 24,131 clips, split relatively evenly between 12,197 normal, non-seizure periods and 11,934 actual seizures. However, the breakdown of seizure severity is heavily lopsided: we have 1,435 Stage 2 events, a massive majority of 9,256 Stage 3 events, 1,050 Stage 4 events, and just 193 Stage 5 events. The frequency of seizures also varies wildly from rat to rat. While

some animals provided as few as 11 clips, others suffered so often that they generated up to 3,091 clips. In fact, the four most heavily affected rats account for nearly half of the entire dataset. A complete, animal-by-animal breakdown of these numbers is available in Table 10 and Fig. 11.

Because the data is so severely skewed, we cannot train our models normally. To prevent the network from simply ignoring rare events, we have to artificially boost their importance during training (inverse-frequency weighting). We also have to judge overall performance using evenly averaged metrics (macro-averaging), break down our results by individual animal, and strictly ensure that a single rat never appears in both the training and testing sets. It is worth noting that we treat our specific weighting approach simply as a practical, reliable baseline, rather than claiming it is the absolute, perfect fix for highly unbalanced data.

C.5. Cross-subject appearance variation

Every rat lives in its own individual cage with a dedicated overhead camera. Because of this, the exact camera angles, the lighting (which switches between natural daylight and infrared), and even the appearance of the animals’ fur change drastically from subject to subject (Fig. 12). This massive visual inconsistency creates exactly the kind of real-world hurdle—or domain shift—that our subject-exclusive testing is designed to catch. It is also the main reason we had to apply custom, cage-specific crops to the video feeds before mixing all the data together.

C.6. Availability, de-identification, and ethics

Because the original study operates under strict data-management rules, our team is not currently allowed to share the raw video and EEG recordings. However, we are publicly releasing the complete, de-identified breakdown of our animal cohort, alongside our exact task definitions and data-splitting formulas. We have also thoroughly stripped identifying information from every visual example in this paper, carefully hiding any specific laboratory or cage la-

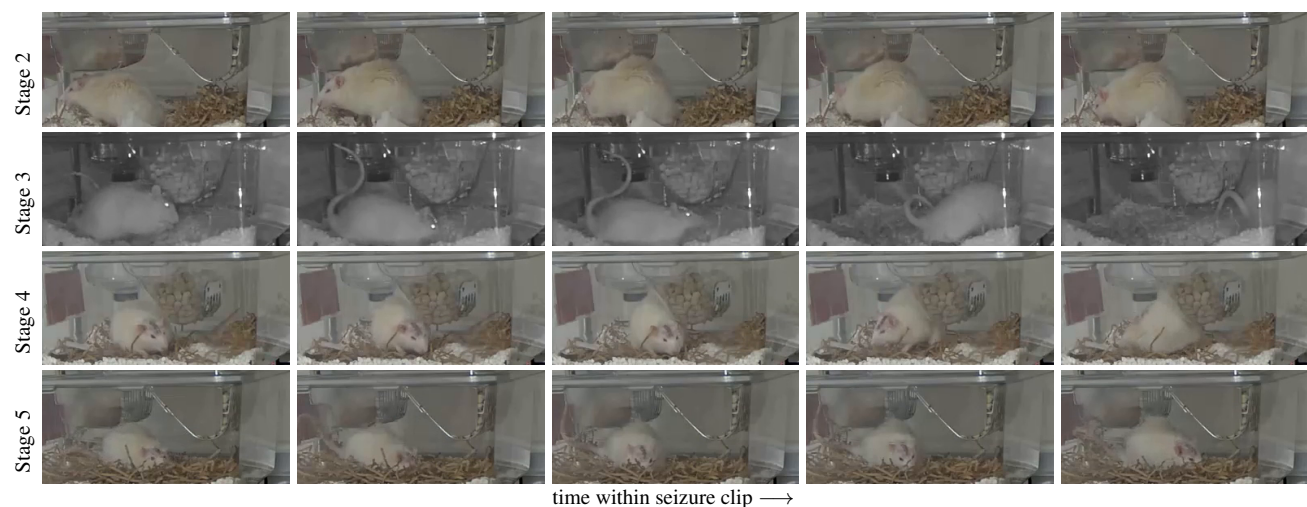


Figure 10. Visual *filmstrips* for each seizure stage: Each row (S2-S5) shows five sequential frames taken from a typical seizure clip, with time moving from left to right. As the severity stage increases, the physical symptoms become much more obvious—escalating from mild movements in Stage 2 to extreme rearing and violent convulsions in Stages 4 and 5. Crucially, these behavioral shifts are plainly visible on camera, whereas the recorded EEG signal only captures them indirectly through shifting brain waves. (All frames are pulled from de-identified clips and tightly cropped to hide any specific cage labels.)

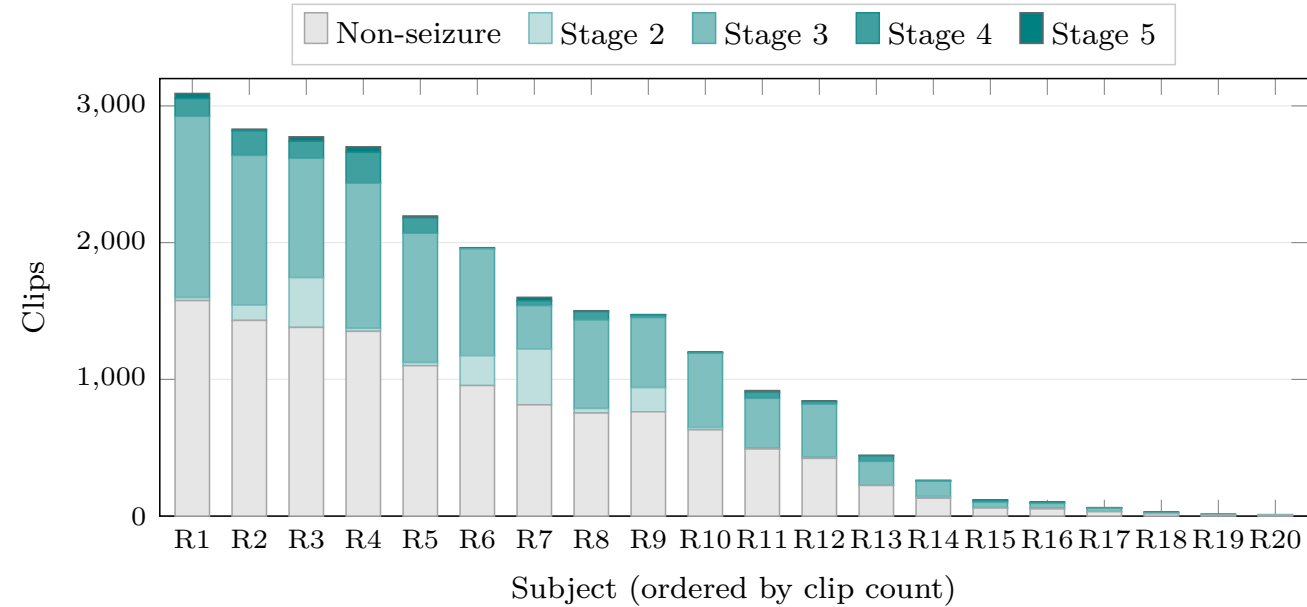


Figure 11. A visual breakdown of the clip counts for each animal across the entire dataset, representing the stacked totals from Table 10. The distribution of seizures is incredibly lopsided: the four most heavily affected rats (R1-R4) completely overshadow the rest of the group, with individual contributions ranging drastically from just 11 clips up to 3,091. More importantly, the most violent seizures (Stage 4 in dark shading, Stage 5 in the darkest) barely register as a thin sliver on any animal’s bar graph, yielding only 193 Stage 5 clips in the whole study. This extreme rarity is exactly why teaching models to perform detailed severity grading is such a difficult challenge.

1518

1519

1520

1521

bels. Additionally, when we need to validate our models against outside visual data, we rely on the open-source RodEpi dataset [44].

Every original animal procedure strictly followed pro-

ocols approved by the Institutional Animal Care and Use Committee (IACUC) at the host lab (the exact protocol number is withheld for blind peer review). The soman exposures were handled at a certified facility under approved

1522

1523

1524

1525

Table 10. A complete, animal-by-animal breakdown of our combined video and EEG dataset across all 20 rats. For every subject, we track the number of recording sessions, total clips, the split between non-seizure and seizure clips, and the exact breakdown of seizure severity (Stages 2 through 5). As the numbers show, the frequency of seizures is drastically uneven across the cohort. Just four heavily affected animals (R1, R2, R3, and R4) provide nearly half of the entire dataset, with individual totals swinging wildly from as few as 11 to as many as 3,091 clips. Across the board, the most violent events (Stage 5) remain extremely rare, yielding only 193 clips. It is worth noting that these counts represent the final, human-labeled dataset; a small handful of clips might still be dropped later when we format the data specifically for the video or EEG models. Finally, all true animal identities have been hidden, and the subjects are simply listed in order from highest to lowest clip count (R1 to R20).

Subject	Sess.	Clips	Non-sz.	Seiz.	S2	S3	S4	S5
R1	42	3,091	1,576	1,515	22	1,325	131	37
R2	44	2,830	1,431	1,399	113	1,094	181	11
R3	38	2,774	1,380	1,394	364	873	124	33
R4	42	2,701	1,351	1,350	22	1,062	226	40
R5	43	2,195	1,100	1,095	23	946	114	12
R6	43	1,963	955	1,008	218	782	8	0
R7	32	1,600	814	786	408	317	36	25
R8	43	1,502	754	748	33	648	61	6
R9	45	1,474	763	711	177	511	21	2
R10	39	1,199	631	568	16	544	6	2
R11	24	918	491	427	6	364	47	10
R12	39	841	422	419	10	387	20	2
R13	23	444	224	220	2	174	42	2
R14	25	262	131	131	12	112	3	4
R15	27	118	59	59	3	39	16	1
R16	18	103	53	50	4	36	7	3
R17	14	62	32	30	1	25	2	2
R18	11	29	17	12	0	9	2	1
R19	4	14	7	7	0	4	3	0
R20	5	11	6	5	1	4	0	0
Total	601	24,131	12,197	11,934	1,435	9,256	1,050	193

safety protocols, and all methods fully comply with NIH ARRIVE guidelines. Finally, this paper is strictly a retrospective analysis of existing telemetry recordings—we did not conduct any new biological experiments or procedures on the animals.

D. Extended Qualitative Results

This appendix breaks down the real-world examples behind two of our main points in Sec. 5: exactly where the video model makes mistakes, and which parts of the image it focuses on when it gets things right.

D.1. Failure cases

Figure 13 highlights the two main situations that cause the model to fail. First, it struggles with the mildest end of the severity scale. Stage 2 episodes often just look like brief twitches, making them almost impossible to tell apart from normal grooming or the animal simply shifting its weight in the cage. Second, the model suffers from a severe lack of data for certain rats. The animals with the fewest seizures

(Table 10) provide barely a handful of clips, which are usually recorded under poor, grainy infrared lighting. This leaves the network with almost no meaningful evidence to learn from for those specific subjects.

Crucially, these mistakes are tied directly to the animals’ natural biology and the physical limits of the cameras, rather than a flaw in the model’s underlying design. This is exactly why we insist on reporting individual, animal-by-animal breakdowns right alongside our overall averages (Sec. 3).

D.2. Grad-CAM interpretability

Expanding on the Grad-CAM [48] analysis from Sec. 5.1, Figure 14 clearly shows that the network’s attention locks tightly onto the animal’s body during a violent Stage 4 convulsion. By contrast, during normal, non-seizure periods, this focus just scatters into random background noise. Figure 15 confirms that the model successfully tracks the animal in this exact same way across Stages 2, 3, and 5.

These heatmaps strongly suggest that the network is genuinely watching the animal’s movements rather than just

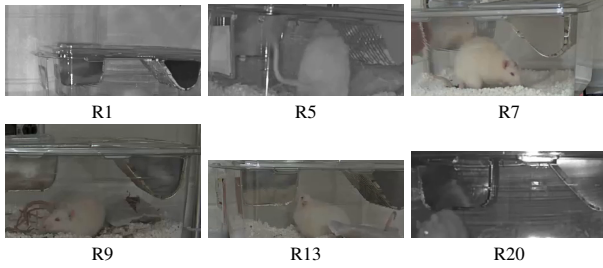


Figure 12. Visual examples of variation across different subjects: a single frame taken from six different animals (where ‘R’ simply indicates a de-identified subject). The exact cage placement, camera angle, lighting conditions (which alternate between natural daylight and infrared), and even the animal’s coat vary drastically from rat to rat. This massive visual inconsistency is exactly why forcing a model to generalize to entirely unseen animals is so much harder than simply testing it on new recording sessions from familiar ones. It also highlights exactly why the video model’s consistent stability across different subjects (Fig. 6) is such a crucial finding.

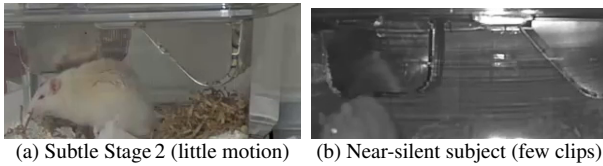


Figure 13. Examples of where the video model struggles the most. (a) A very mild Stage 2 seizure featuring only subtle movements, making it incredibly easy to confuse with normal, everyday behavior. (b) An animal that rarely experienced seizures (found at the bottom of Table 10, App. C), leaving the model with very few validation clips captured under poor, grainy infrared lighting. Ultimately, extremely mild symptoms and data-starved animals drive almost all of the network’s remaining errors. (All clips have been de-identified.)

memorizing static background details in the cage. However, we have to be careful: these visualizations simply show us where the model is sensitive, rather than acting as absolute, definitive proof of how it makes its final decisions. Even though matching our seizure clips with normal clips from the exact same recording session helps strip away a lot of potential biases, completely proving that the model isn’t relying on clever visual shortcuts would require much stricter testing.

D.3. What the shortcut evidence does and does not establish

Although the safeguards we detailed in Sec. 5.1 eliminate the most obvious visual shortcuts, they cannot guarantee a perfectly bias-free learning environment. By strictly pairing seizure clips with normal clips from the exact same recording session, we ensure that the animal’s appearance,

the camera angle, and the lighting stay identical. Similarly, tightly cropping the video to the cage boundaries hides any clues from the surrounding room. Furthermore, the fact that the single-frame baseline model completely fails proves that the network cannot simply rely on static background pixels to make its diagnosis.

However, more subtle clues might still slip through these defenses—such as how the bedding gets pushed around in the cage, or unique movement habits tied to how frequently a specific rat seizes. Our Grad-CAM heatmaps do confirm that the network’s attention lands directly on the rat, and our strict cross-validation proves the model can easily spot seizures in completely unfamiliar animals. Even so, definitively proving that the model relies entirely on genuine seizure behavior would require extreme, targeted tests (like digitally erasing the animal from the video completely, forcing the network to train only on the empty cage background, or randomly shuffling the order of the frames). Because of this, we consider our current evidence to be strongly supportive, but it does not absolutely prove cause and effect.

E. Extended Quantitative Results

This appendix breaks down the detailed numbers behind two main claims from the paper: how changing the amount of training data and using pre-trained models affects overall detection (Sec. 5.1), and our extensive test of ten different video models on the RodEpiL dataset (Sec. 5.2).

E.1. Data efficiency and pretraining

We ran two separate tests to see whether the model benefits more from having a massive dataset or from being pre-trained on outside data (Table 11).

First, we kept our standard R(2+1)D model exactly the same but slowly increased the number of animals in the training set. We found that the model’s performance *maxes out incredibly fast*. Training on just two animals (a mere 62 clips) already achieves a 0.79 macro-F1 score. Bumping that up to half the cohort (around 11k clips) pushes the score to 0.957. Adding the remaining 10 animals only squeezes out a tiny 0.013 improvement, landing at 0.970. (We did see a small dip in performance around the 12-subject mark because those specific animals were simply harder to diagnose, but the overall trend remains a fast climb followed by a flat plateau). This proves that the model is absolutely not starving for more data.

Second, we found that while pre-training the model on the Kinetics dataset helps, it is not strictly necessary. When we trained the exact same network entirely from scratch, it still hit a 0.952 macro-F1 score—falling behind the pre-trained version by a paltry 0.018 margin. Ultimately, pre-training gives the network a reliable, helpful boost, but it is not structurally required to get excellent results.

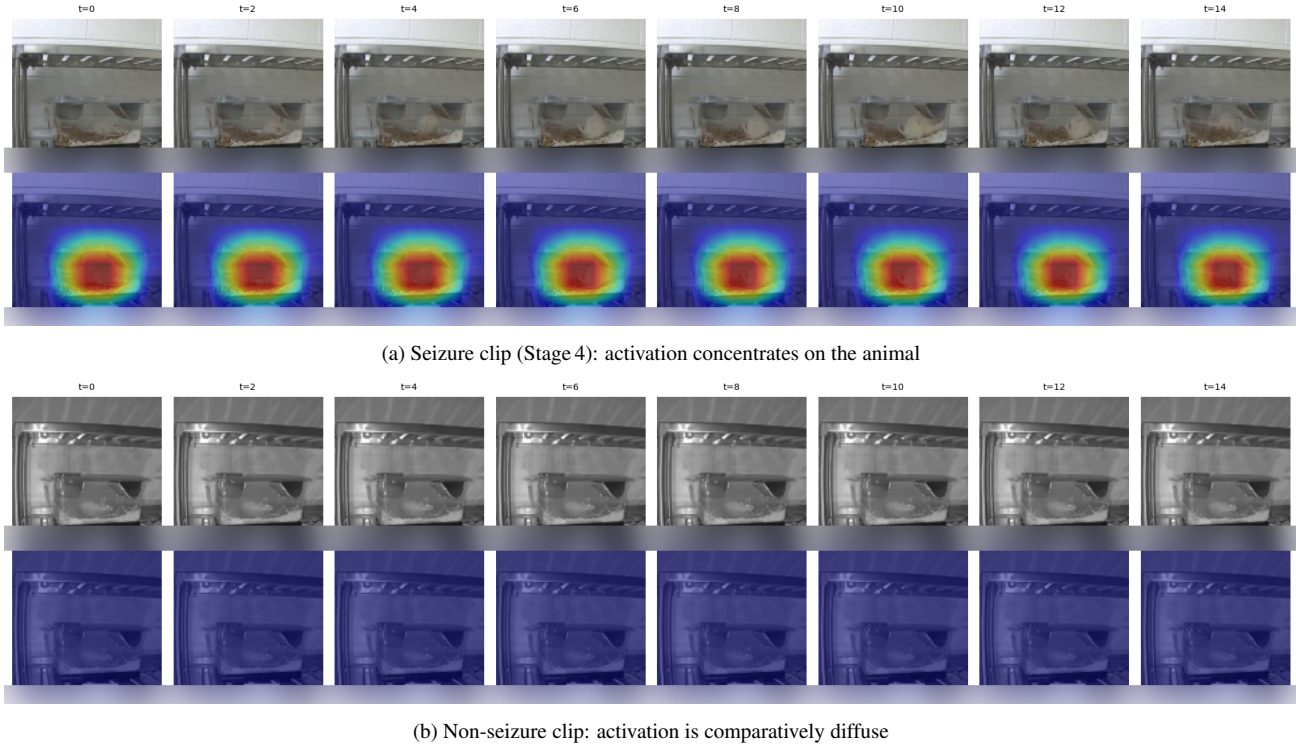


Figure 14. A full-width, extended version of Fig. 5. This layout shows all eight sample frames from the exact same pair of clips. For each sequence, the raw video frames sit on top, with the model’s attention heatmaps (seizure-class activation) directly underneath. While this sharp contrast strongly supports the idea that the model is genuinely focusing on the animal to make its predictions, it does not act as definitive proof against hidden visual shortcuts. (All clips have been completely de-identified, and any specific cage identifying bands have been redacted.)

E.2. RodEpil encoding audit and re-encoding protocol

When we first ran our R(2+1)D model on the original RodEpil dataset, the results were suspiciously perfect, hitting an F1 score of 0.9996. Looking closely at the raw files, we found a glaring structural flaw baked right into the data: every single non-seizure video was saved at exactly 30 fps (giving it 301 frames), while all the seizure clips were locked at 29 fps (giving them 291 frames). On top of that, the videos were cropped differently depending on whether a seizure was happening or not. Because the network could easily cheat by just counting the total number of frames instead of actually watching the animal’s movements, any performance scores based on this raw dataset are completely meaningless for proving real behavioral recognition.

To eliminate these artificial shortcuts, we ran the entire repository through a strict re-encoding pipeline. We letterboxed every video to a 224×224 resolution to maintain the original aspect ratio, locked everything to exactly 30 fps and 300 frames, and applied the exact same video codec across the board. While we can’t guarantee this scrubs away every microscopic digital artifact, it completely kills

the model’s ability to cheat based on frame counts or frame rates, all while perfectly preserving the actual biological motion on screen. Tested on this rigorously sanitized data, the R(2+1)D model settles at a genuine subject-wise seizure F1 score of 0.979 (0.999 AUROC, 0.972 MCC). These realistic metrics line up perfectly with the 0.97 baseline originally reported by the dataset’s creators [44]. Moving forward, all of our model evaluations rely exclusively on these cleaned files.

E.3. All ten backbones on RodEpil

To see how our findings hold up, we ran our full lineup of ten video models on the sanitized RodEpil dataset, again enforcing a strict subject-exclusive 5-fold cross-validation (Table 12). Interestingly, the leaderboard completely shuffled compared to how the models ranked on our native dataset. This really drives home that there is no universal “best” architecture—a network’s success depends entirely on the specific visual quirks and constraints of the data it’s looking at. Once we stripped away the digital artifacts to standardize the video files, traditional convolutional models performed exceptionally well. In contrast, several of the newer attention-based models actually struggled, need-

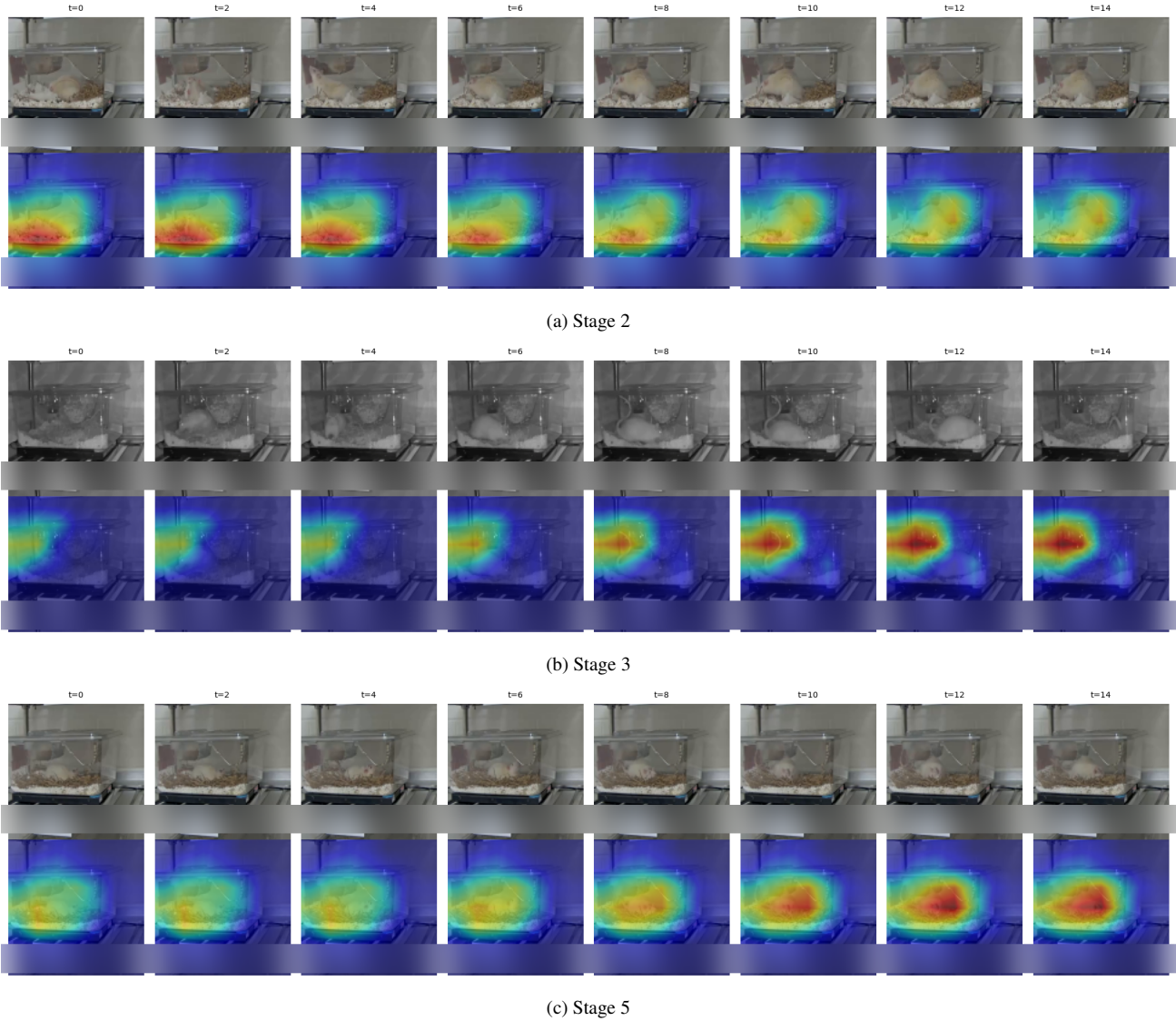


Figure 15. Grad-CAM heatmaps for the R(2+1)D video model across different severity stages. The network’s attention consistently locks onto the animal across the illustrated Stages 2, 3, and 5, proving that this tracking behavior is not just limited to one specific clip or a single level of severity. Still, we must treat these visualizations as descriptive rather than causal—they show us where the model is looking, but they do not serve as absolute proof of how it makes its final decisions. (All clips have been de-identified, and any identifying cage bands have been completely redacted.)

ing far more aggressive fine-tuning just to stay competitive within our fixed training budget.

Keep in mind that any comparisons to the original authors’ TimeSformer baseline can only serve as a loose guide, since their published metrics were based on the structurally flawed raw dataset. Under our strictly controlled sweep, the R(2+1)D model still delivered a strong 0.978 seizure-class F1 score, perfectly aligning with the 0.979 it earned during its dedicated standalone run.

We do, however, have to raise a major red flag regarding the SlowFast model, which returned a suspiciously perfect

1.000 F1 score during these tests. Given RodEpil’s proven history of hidden digital artifacts, we are treating this absolute perfection as highly suspect. Instead of blindly accepting this as flawless behavioral recognition, we assume the network simply latched onto some deeply buried visual shortcut that slipped past our cleaning pipeline—a quirk that definitely warrants further forensic investigation.

Table 11. Data efficiency and pretraining tests for the R(2+1)D model on the basic yes/no detection task (using session-exclusive testing). Top: The macro-F1 score compared against the fraction of subjects included in the training data, while keeping the validation set exactly the same. Bottom: A direct comparison between a model pre-trained on Kinetics and one trained entirely from scratch. The results clearly show that detection performance maxes out long before we use the full animal cohort, and pretraining only provides a tiny boost of roughly 0.02. Finally, while the MCC score generally follows the same overall trend as the macro-F1, it drops much more sharply when the network is starved for data.

Train subj.	Frac.	Clips	Macro-F1	MCC
2	0.10	62	0.790	0.583
5	0.25	1,905	0.893	0.787
10	0.50	10,969	0.957	0.914
12	0.60	11,832	0.946	0.892
15	0.75	16,358	0.960	0.921
20	1.00	19,208	0.970	0.941
<i>Pretraining</i> (full cohort):				
Kinetics-pretrained			0.970	0.941
from scratch (random init)			0.952	0.903

Table 12. Performance breakdown for all ten video models on the thoroughly sanitized RodEpi dataset [44], evaluated under strict subject-exclusive five-fold cross-validation. F1 and AUROC scores represent the average across all five folds, while MCC includes the standard deviation (mean $\pm$ std). Notice how drastically the model rankings shuffle compared to our native dataset—a clear reminder that a network’s success is deeply tied to the specific visual quirks of its training data. [†]We have explicitly flagged the SlowFast model; its suspiciously perfect score strongly suggests it latched onto a hidden visual shortcut rather than achieving a genuine diagnostic victory. The highest legitimate, non-flagged score is highlighted in **bold**.

Backbone	Seiz-F1	Macro-F1	AUROC	MCC
SlowFast-R50 [†] [18]	1.000	1.000	1.000	1.000 $\pm$ 0.000
R(2+1)D-18 [53]	0.978	0.986	0.999	0.972 $\pm$ 0.015
MViTv2-S [30]	0.916	0.946	0.989	0.891 $\pm$ 0.020
S3D [57]	0.912	0.944	0.991	0.889 $\pm$ 0.029
X3D-M [17]	0.895	0.933	0.983	0.866 $\pm$ 0.010
VideoMAE-B [52]	0.874	0.919	0.982	0.839 $\pm$ 0.025
Video Swin-T [34]	0.861	0.910	0.979	0.823 $\pm$ 0.032
Video Swin-S [34]	0.854	0.904	0.977	0.810 $\pm$ 0.031
MViTv1-B [16]	0.835	0.890	0.971	0.784 $\pm$ 0.049
TimeSformer [7]	0.799	0.871	0.957	0.743 $\pm$ 0.038

E.4. Backbone, temporal, and spatial ablations in detail

This section breaks down the raw numbers behind the summary findings we presented in Sec. 5.1.

Comparing model backbones. As Table 2 shows, the performance gap between our top video models is razor-thin, spanning a microscopic 0.03 margin. MViTv2-S and SlowFast just barely edge out R(2+1)D and X3D-M, while TimeSformer brings up the rear. But as we saw during our tests on the external RodEpi dataset, these rankings are incredibly fragile and shuffle drastically depending on the training data. For example, R(2+1)D easily beat TimeSformer on the sanitized RodEpi files, even though TimeSformer was originally pitched as that dataset’s baseline architecture. Ultimately, this tight clustering tells us that hunting for the “perfect” architecture matters far less than curating high-quality data—there simply isn’t a universally superior network. In fact, these models stay tightly clumped together even when we push them to handle the much more difficult 3- and 5-class severity grading tasks.

How many frames do we actually need? Surprisingly, feeding the network a mere 4 frames is enough to unlock the vast majority of its diagnostic power, securing a solid 0.940 accuracy score. Bumping that to 8 frames pushes the score to 0.960, but after that, performance hits a hard plateau. Moving to 16 frames inches us up to 0.970, and doubling the input again to 32 frames barely nudges the needle to 0.972. The more complex severity grading tasks follow this exact same trajectory, with 3- and 5-class accuracy peaking at the 32-frame mark (hitting 0.793 and 0.667, respectively). Crucially, overloading the model with 64 frames offers absolutely no benefit for basic detection—and it actively harms the network’s ability to handle 5-class grading, dragging the score back down to 0.632. Because our 16-frame default sacrifices a practically invisible margin of accuracy (losing just 0.002 in detection and 0.005 in grading) while slicing the computational cost in half compared to the 32-frame setup, we officially locked it in as our standard baseline.

Image resolution. Increasing the video resolution beyond our 112 px default actually hurts model performance across the board. Because the raw laboratory footage is natively low-resolution (roughly 300–480 px) and the seizures we track are large, full-body movements rather than subtle micro-details, upscaling the video is entirely counterproductive. Upsampling simply forces the network to waste compute processing interpolated pixels that add zero diagnostic value. Pushing resolutions even further—such as to 512 px or 1024 px—would only inflate this useless digital noise.

Table 13. Ablation experiment evaluating spatial resolution for the R(2+1)D detection model across three input image sizes (16 frames, session-exclusive split, seed 42, 10 epochs; batch sizes scaled to 16, 8, and 4, respectively). Because the native source footage is only $\sim 300\text{--}480$ px, both of the higher resolution settings force the network to upsample the video. Consequently, performance drops steadily across all three metrics. In terms of AUROC, the gap between the top two settings only emerges past the third decimal place (0.9934 vs. 0.9926). Note that the 448 px run suffered an out-of-memory crash after completing 8 of its 10 epochs; we report metrics from its best saved checkpoint, meaning this row represents an optimistic upper bound for the largest resolution setting. The best score in each column is highlighted in bold.

Input size	Macro-F1	AUROC	MCC
112×112 (default)	0.970	0.993	0.941
224×224	0.967	0.993	0.934
448×448	0.952	0.988	0.904

Table 14. Overview of core findings across our 20-rat soman-induced seizure cohort. To counter extreme long-tail class imbalance and massive visual domain shifts (Table 10, Fig. 12), we enforce strict subject-exclusive cross-validation and rigorously sanitize external benchmarks [44]. Results show that an R(2+1)D model peaks using only 16 frames, proving that data quality heavily outweighs architectural complexity (Table 2). Finally, Grad-CAM confirms the network tracks genuine biological motion, with rare errors driven by realistic physical constraints rather than algorithmic shortcuts (Figs. 14–13).

Stage	Video				EEG			
	n	P	R	F1	n	P	R	F1
non-seizure	2720	0.965	0.968	0.966	2737	0.914	0.955	0.934
Stage 2	220	0.681	0.650	0.665	256	0.491	0.863	0.626
Stage 3	2148	0.904	0.896	0.900	2064	0.867	0.717	0.785
Stage 4	167	0.330	0.401	0.362	229	0.268	0.162	0.202
Stage 5	34	0.579	0.324	0.415	33	0.074	0.364	0.123

E.5. Ranking quality: AUROC

While the main manuscript prioritizes macro-F1 and MCC to heavily penalize rare-stage errors at hard decision boundaries, we provide the complete AUROC metrics in Tables 15–17 for full transparency (using binary AUROC for detection and macro one-vs-rest for multi-class grading).

Crucially, AUROC measures continuous confidence ranking rather than discrete decision accuracy, causing the metric to saturate heavily across our benchmarks. Across all ten video backbones, for example, detection AUROC stays tightly clustered between 0.984 and 0.995. Similarly, on the grueling 5-class grading task, the EEG GRU and visual R(2+1)D post almost identical AUROC scores (0.941 vs.

Table 15. AUROC scores comparing the standard session-exclusive evaluation against the strictly segregated subject-exclusive cross-validation. We use binary AUROC for basic detection and a macro one-vs-rest approach for multi-class severity grading. All values represent the mean $\pm$ std across five independent runs.

Task	Modality	Session-excl.	Subject-excl.
2-class (detection)	Video	<u>0.993</u> ± 0.001	0.985 ± 0.006
	EEG	0.996 ± 0.001	0.995 ± 0.005
3-class (severity)	Video	0.964 ± 0.003	0.942 ± 0.015
	EEG	<u>0.956</u> ± 0.002	0.916 ± 0.025
5-class (severity)	Video	<u>0.939</u> ± 0.004	0.888 ± 0.026
	EEG	0.941 ± 0.003	0.831 ± 0.076

0.939) despite a massive gap in their actual macro-F1 performance (0.560 vs. 0.657). This disconnect is expected: internal probability rankings can remain strong even when hard thresholding severely degrades class-specific precision and recall. Consequently, we treat AUROC purely as a secondary diagnostic check rather than a primary measure of real-world cross-modal performance.

E.6. Robustness of the cross-modal ordering

While our primary analysis relies on a GRU model to mirror standard rodent-EEG setups [19], our broader conclusions do not hinge on that specific architecture. To test this, we swapped in a high-performing Temporal Convolutional Network (TCN) to process the EEG data. As the diagnostic tasks grew progressively harder, the TCN posted macro-F1 scores of 0.978 ± 0.001 for basic detection, 0.767 ± 0.007 for 3-class grading, and 0.563 ± 0.012 for the 5-class grading task.

When we line these numbers up against our standard video model (0.968, 0.780, and 0.657, respectively), the exact same pattern emerges: the TCN pulls slightly ahead in basic detection, effectively ties during 3-class grading, and then completely loses pace during the fine-grained 5-class severity task. This perfectly mirrors the performance trajectory of the baseline GRU (0.974, 0.750, and 0.560). Ultimately, this confirms that our core takeaway—that video fundamentally outperforms EEG for granular severity grading—is a genuine reflection of the modalities themselves, not an artifact of an artificially weak electrical baseline. (Though whether high-density, multi-channel EEG arrays might eventually close this gap remains an open question).

The underlying metrics tell a perfectly consistent story. When tracking the performance gap between the video and EEG models, the macro-F1 difference swings sharply from

Table 16. AUROC scores for all ten video and seven EEG backbones, following the exact same row order as Tables 2 and 4 for easy comparison. We use binary AUROC for the basic detection task and a macro one-vs-rest approach for multi-class severity grading.

Backbone	Detection	3-class	5-class
<i>Video</i>			
MViTv2-S [30]	0.993 $\pm$ 0.002	0.971 $\pm$ 0.002	0.915 $\pm$ 0.011
SlowFast-R50 [18]	0.995 $\pm$ 0.001	<u>0.966</u> $\pm$ 0.001	<u>0.939</u> $\pm$ 0.011
R(2+1)D-18 [53]	<u>0.994</u> $\pm$ 0.001	0.960 $\pm$ 0.003	0.928 $\pm$ 0.017
X3D-M [17]	0.993 $\pm$ 0.001	0.952 $\pm$ 0.005	0.911 $\pm$ 0.003
MViTv1-B [16]	0.991 $\pm$ 0.003	0.965 $\pm$ 0.003	0.935 $\pm$ 0.012
Video Swin-S [34]	0.987 $\pm$ 0.004	0.962 $\pm$ 0.011	0.916 $\pm$ 0.025
S3D [57]	0.992 $\pm$ 0.001	0.958 $\pm$ 0.006	0.917 $\pm$ 0.008
Video Swin-T [34]	0.990 $\pm$ 0.003	0.963 $\pm$ 0.002	0.943 $\pm$ 0.009
VideoMAE-B [52]	0.986 $\pm$ 0.002	0.942 $\pm$ 0.018	0.920 $\pm$ 0.013
TimeSformer [7]	0.984 $\pm$ 0.003	0.950 $\pm$ 0.002	0.913 $\pm$ 0.010
<i>EEG (deep)</i>			
GRU [19]	<u>0.996</u> $\pm$ 0.001	<u>0.956</u> $\pm$ 0.002	0.941 $\pm$ 0.003
LSTM [21]	0.994 $\pm$ 0.004	0.937 $\pm$ 0.027	0.903 $\pm$ 0.039
EEGNet [29]	0.975 $\pm$ 0.005	0.903 $\pm$ 0.005	0.864 $\pm$ 0.011
EEG Conformer [50]	0.988 $\pm$ 0.002	0.952 $\pm$ 0.002	0.936 $\pm$ 0.008
TCN [5]	0.996 $\pm$ 0.001	0.960 $\pm$ 0.001	<u>0.941</u> $\pm$ 0.003
<i>EEG (classical)</i>			
Random Forest [9]	0.986 $\pm$ 0.000	0.934 $\pm$ 0.001	0.916 $\pm$ 0.002
XGBoost [13]	0.984 $\pm$ 0.000	0.934 $\pm$ 0.000	0.924 $\pm$ 0.000

Table 17. AUROC scores for both the video frame-count sweep and the rigorously re-encoded RodEpil cross-validation folds. For the sake of clarity, we have broken out the specific RodEpil AUROC results into their own dedicated section immediately below the frame-count data.

	Frames	Detection	3-class	5-class
	1	0.615	0.599	0.634
	4	0.983	0.950	0.926
	8	0.989	0.957	0.919
	16	0.993	0.962	0.946
	32	0.994	0.971	0.951
	64	<u>0.994</u>	<u>0.971</u>	<u>0.950</u>

RodEpil (re-enc.)	F0	F1	F2	F3	F4	Mean
AUROC	0.999	1.000	1.000	1.000	0.998	0.999

−0.013 deficit to a +0.103 lead. Because the AUROC metric completely saturates on this dataset and loses its ability to meaningfully separate the networks (Sec. E.5), relying on macro-F1 and MCC provides the most honest, realistic picture of actual diagnostic capability.

E.7. Subject-exclusive transfer

While session-exclusive splits prevent temporal leakage between adjacent frames, they still allow models to memorize animal-specific traits within the training pool. To enforce a far more demanding and realistic evaluation of generalization, we conducted a subject-exclusive 5-fold cross-validation, ensuring networks were tested strictly on unseen subjects. Reassuringly, basic detection generalizes exceptionally well. The video model suffers a negligible 0.020 drop in macro-F1 (sliding from 0.968 to 0.948 ± 0.016), while the EEG model barely shifts (0.974 to 0.972 ± 0.015). Corresponding MCC scores remain similarly solid at 0.896 ± 0.031 and 0.944 ± 0.030 , respectively.

However, forcing the models to grade specific severity levels on completely unfamiliar animals takes a steep toll across the board. On the 3-class grading task, the video model’s macro-F1 slides to 0.720 ± 0.028 , while EEG drops

a microscopic −0.006 deficit in basic detection to a decisive +0.097 advantage for video during 5-class grading. The MCC captures this exact same inversion, flipping from a

to 0.656 ± 0.022 . Under the grueling 5-class evaluation, video drops to 0.475 ± 0.024 and EEG falls to 0.404 ± 0.042 . While video maintains a 0.071 advantage over EEG on this hardest task, its overall drop in performance (0.182) is actually larger than EEG’s (0.156). Ultimately, this shows that video is not immune to cross-subject variation: while both modalities easily generalize basic detection to unseen animals, mapping fine-grained severity signatures onto novel targets remains a significant challenge for both.

E.8. Long-tail behavior and weighting

The 5-class severity grading task forces us to confront severe class imbalance alongside massive variations across individual subjects. Looking at the raw counts, the dataset is heavily skewed: we have 12,197 baseline clips versus 1,435 Stage 2, 9,256 Stage 3, 1,050 Stage 4, and just 193 Stage 5 events. Compounding this challenge, the volume of data contributed by each animal varies wildly, ranging from a meager 11 clips to as many as 3,091 (Table 10). To stop the network from simply ignoring rare classes to minimize overall error, we used inverse-frequency-weighted cross-entropy loss (keeping the average weight normalized to unity). In practice, this forces the model to treat a single, ultra-rare Stage 5 seizure with roughly $60\times$ the mathematical weight of a normal baseline clip.

Applying this weighting scheme forces the network to learn representations for the rarest seizure categories. For context, a naive majority-class classifier achieves a macro-F1 of just 0.136. In contrast, our weighted video network reaches an overall macro-F1 of 0.657, maintaining viable performance even on data-starved classes (0.362 F1 on Stage 4 and 0.415 on Stage 5). The EEG model, however, still struggles despite the weighting, managing only 0.202 and 0.123 F1 for these respective stages (Table 14).

While this strategy successfully prevents class collapse, we deliberately relied on standard inverse-frequency weighting rather than evaluating unweighted baselines, synthetic oversampling techniques [12], or focal loss formulations [32]. Ultimately, our findings demonstrate that straightforward class weighting is sufficient to preserve model sensitivity across rare stages, establishing a practical baseline without requiring complex long-tail optimization strategies.

E.9. Where each modality fails, and why

Table 14 reveals that the video and EEG models exhibit strikingly distinct failure modes. Figure 16 maps out these modal-specific confusion patterns, which we further illustrate with representative qualitative examples in Fig. 17.

Missing a seizure entirely vs. misjudging its severity. The EEG model suffers a substantial drop in performance by failing to detect high-activity events entirely, missing 10.1% of major Stage 3 convulsions (208 out of 2,064) com-

pared to just 2.3% for the video model (49 out of 2,148). (We explicitly separate complete non-detections from severity grading errors, as baseline behavior lies outside the ordinal scale of seizure severity).

When grading detected events, the video network’s misclassifications almost exclusively fall into adjacent severity levels—such as mislabeling a Stage 4 convulsion as Stage 3 in 95 out of 167 error cases. In sharp contrast, the EEG model exhibits wide-ranging, non-adjacent errors, frequently confusing severe Stage 5 episodes with milder Stage 3 or Stage 4 predictions. This confirms that video representations naturally capture the continuous physical progression of seizure behavior, whereas single-channel EEG signals struggle to resolve granular intensity without incurring erratic cross-stage confusions.

The advantage flips at Stage 2. Crucially, this dynamic completely reverses at Stage 2. Because early-stage seizures manifest as subtle, low-motor behaviors, EEG recall spikes to 0.863, whereas the video model lags at 0.650. On identical paired evaluation clips, Stage 2 stands out as the single category where EEG reliably outperforms video (Table 18). This outcome reinforces our central hypothesis: vision-based models dominate only when seizures produce overt physical movement.

Evaluating joint predictions across paired clips highlights the deeply complementary strengths of both modalities. Specifically, the video model correctly identifies seizures missed by EEG across 14.2% of the dataset, whereas EEG uniquely detects active events on 4.9% of clips. Because their error distributions are largely non-overlapping, relying on either sensor in isolation introduces distinct diagnostic blind spots. Ultimately, capturing the full biological continuum of seizure dynamics requires multi-modal fusion rather than reliance on a single signal source.

Motion-energy analysis. Using mean absolute frame difference as a straightforward proxy for physical motion (Table 19) reveals that EEG recall for Stage 3 events improves significantly during high-activity intervals (0.777 vs. 0.537). (Crucially, because motion energy is derived exclusively from visual pixels rather than electrical telemetry, this correlation remains purely descriptive; it cannot serve as a direct causal explanation for EEG failures).

Motion alone cannot explain the Stage 2 anomaly. Crucially, basic physical movement completely fails to explain what is happening at Stage 2. The video model’s recall remains practically identical regardless of whether the overall motion is low (0.655) or high (0.586). Whatever factors are actually driving the performance gap between video and EEG during these subtle, early-stage seizures, raw motion magnitude alone is clearly not the answer.

Row-normalised confusion matrices (fraction of each true class; counts in parentheses)

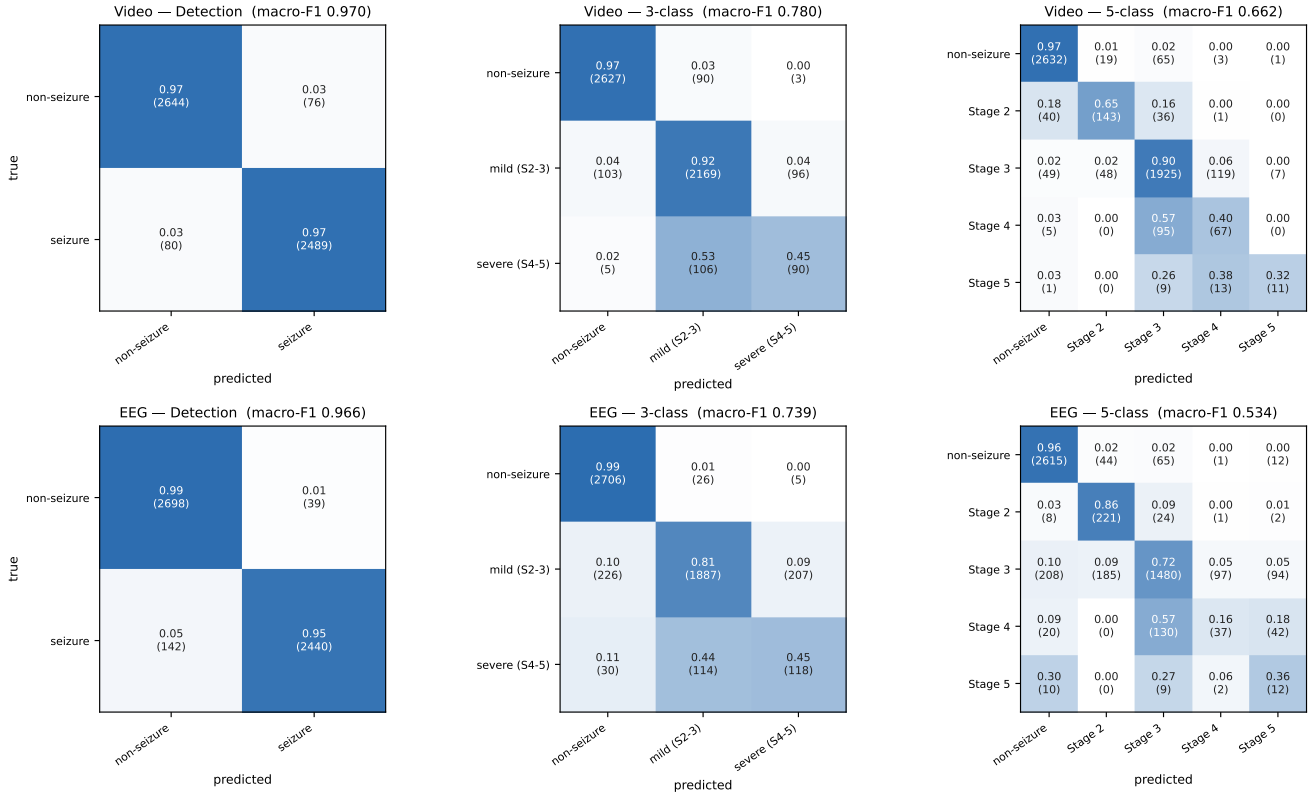


Figure 16. Row-normalized confusion matrices comparing both modalities across all three evaluation tasks (drawn from a representative seed-42 diagnostic run; displaying true-class fractions with raw counts in parentheses). On the granular 5-class task, the video model’s mistakes cluster tightly along the diagonal, meaning its errors are almost entirely limited to adjacent severity levels. In stark contrast, the EEG baseline not only misses active seizures completely, but also frequently commits erratic, non-adjacent grading errors.

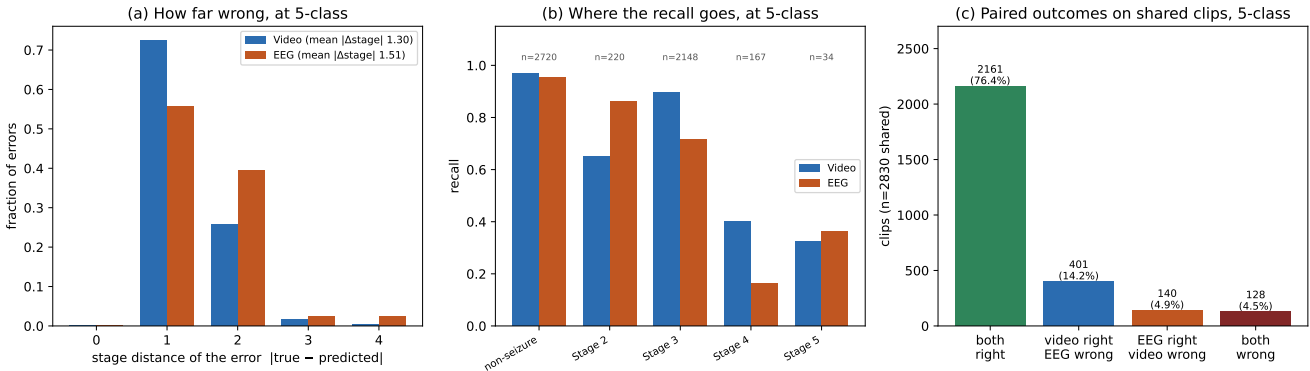


Figure 17. Error diagnostics for the 5-class severity grading task. (a) Numeric label-distance histogram. Because baseline non-seizure behavior falls completely outside the standard Racine scale, this panel remains purely descriptive and should not be treated as a strict ordinal metric. (b) Per-class recall annotated with absolute support counts, distinctly highlighting the Stage 2 performance reversal. (c) Direct head-to-head outcomes across the strictly shared intersection of 2,830 test clips evaluated by both modalities.

F. Extended Interpretation

Ultimately, our empirical findings point to a strict division of labor rather than a single, universally superior diagnostic

Table 18. Direct head-to-head evaluation across the 2,830 held-out test clips shared between both modalities for 5-class severity grading. To ensure a strictly fair comparison, this analysis is restricted to the exact intersection of clips evaluated by both models. Per-class rows highlight “solo victories”—instances where precisely one model correctly predicted the severity stage while the other failed. Notably, Stage 2 represents the sole category where the EEG model outperforms the video network in solo correct predictions.

True class	n	video only	EEG only
non-seizure	1488	73	44
Stage 2	192	7	44
Stage 3	1062	286	43
Stage 4	74	27	8
Stage 5	14	8	1
overall: both correct 2161 (76.4%), video only 401 (14.2%), EEG only 140 (4.9%), both wrong 128 (4.5%)			

Table 19. Recall evaluated across the low- and high-motion halves of each severity class, partitioned strictly at the class-specific median motion energy (900 clips sampled from the paired evaluation set). We define motion energy as the mean absolute frame difference across 24 evenly spaced 96×96 grayscale frames. Because this metric is computed over the entire duration of the clip including its 10 s pre- and post-event buffers it serves as a coarse but objective proxy for the overall physical intensity of the seizure.

True class	n	med.	Video recall		EEG recall	
			low	high	low	high
non-seizure	466	1.53	0.983	0.966	0.906	1.000
Stage 2	58	1.57	0.655	0.586	0.966	0.862
Stage 3	350	1.78	0.891	0.943	0.537	0.777
Stage 4	21	2.17	0.182	0.700	0.182	0.000

modality. While EEG holds its ground during basic binary detection, video completely dominates the moment seizures escalate into observable physical movement. This perfectly reflects the underlying reality of the task: because seizure severity (Stages 2–5) is fundamentally defined by motor behavior, video measures the clinical ground truth directly. EEG, on the other hand, is forced to infer these physical events through indirect electrographic proxies. Therefore, the resulting performance gap does not necessarily imply that EEG lacks severity information entirely, but rather that video benefits from a direct sensory alignment with the behavioral grading scale.

The performance inversion at Stage 2 perfectly illustrates this reality. Because the physical movements during these

early-stage seizures are so subtle, EEG easily dominates recall (0.863 vs. 0.650). However, once the seizures escalate into violent Stage 3 convulsions, severe movement artifacts likely overwhelm the telemetry equipment, causing the EEG model to completely miss 10.1% of these major events (compared to a mere 2.3% for video). Yet, as our low/high motion splits demonstrated, raw kinetic magnitude alone cannot fully account for this breakdown. Instead, these diagnostic blind spots are driven by a complex interplay of animal pose, event duration, electrical artifact masking, and the underlying electrophysiological signatures.

Operationally, visual detection proves remarkably robust across drastically different model sizes. The fact that a lightweight 4M-parameter X3D-M network can practically match the performance of a massive 33M-parameter R(2+1)D model proves a critical point: dense temporal tracking matters far more than sheer parameter count. This creates a clear path toward highly efficient, edge-deployable monitoring systems (though evaluating real-world streaming latency and continuous false-positive rates remains strictly outside the scope of this study).

Regardless of video’s dominance in granular scaling, EEG remains non-negotiable for diagnosing “silent” internal electrographic anomalies lacking overt behavioral symptoms. Their wildly non-overlapping failure regimes practically mandate future multimodal architectures, specifically utilizing quality-aware attention protocols to suppress video during occlusions and suppress EEG during intense physical movement artifacts. Our unimodal framework establishes the exact baselines required to validate these future integrated systems.

G. Limitations and Future Work in Detail

Sec. 6 broadly defined five structural limitations; here, we brutally interrogate their impact against the empirical data. **(1) Cross-subject and cross-cohort generalization.** While subject-exclusive partitioning guarantees the network never memorizes an animal, every subject still stems from a localized, single-laboratory cohort. Under strict isolation, binary detection transfers phenomenally ($0.968 \rightarrow 0.948 \pm 0.016$), but granular grading collapses entirely ($0.657 \rightarrow 0.475 \pm 0.024$). This verifies within-cohort stability but provides zero guarantees regarding cross-laboratory hardware swaps, differing chemical induction protocols, or alternate rodent genetics. Massive multi-cohort benchmarking represents the ultimate deployment barrier.

(2) Fine-grained and ordinal severity. Robust classification of extreme severity tiers remains a significant open challenge. Rare Stage 4 and Stage 5 episodes suffer from disproportionately high error rates, highlighting an acute need for targeted dataset expansion. Furthermore, because standard cross-entropy loss penalizes all misclassifications equally, it completely fails to exploit the inherent $S_2 \prec$

1984 S3 < S4 < S5 progression of seizure development. Future
1985 architectures must incorporate ordinal objectives to explic-
1986 itly penalize non-adjacent errors based on this biological hi-
1987 erarchy. While Section 3.2 outlines a theoretical conditional
1988 framework to address this, it still requires rigorous empir-
1989 ical validation. Similarly, our inverse-frequency weight-
1990 ing serves merely as a practical baseline; comprehensive
1991 comparisons against focal loss formulations [32] or syn-
1992 thetic oversampling techniques [12] are necessary. Finally,
1993 because classical Racine scales were originally developed
1994 across varying chemical induction models [45, 54], dis-
1995 tinguishing genuine network failures from underlying hu-
1996 man annotation ambiguity will ultimately require rigorous,
1997 consensus-driven expert mapping.

1998 **(3) EEG capacity and multimodal fusion.** Our electri-
1999 cal baseline relies exclusively on a single recording chan-
2000 nel, artificially constraining the spatial resolution available
2001 to the network. It remains entirely plausible that a dense,
2002 multi-electrode montage could capture richer spatial dy-
2003 namics, potentially narrowing or even completely elimi-
2004 nating the grading advantage currently held by the visual
2005 model. Furthermore, while this study deliberately evalu-
2006 ated both modalities in strict isolation to establish their in-
2007 dependent capacities, their sharply contrasting failure regimes
2008 clearly point toward integration. Future architectures must
2009 prioritize quality-aware video-EEG fusion protocols capa-
2010 ble of dynamically suppressing modality-specific artifacts
2011 and resolving diagnostic conflicts, ultimately combining the
2012 complementary strengths of both sensors [2, 26].

2013 **(4) Shared benchmarks.** Although our uniform re-
2014 encoding of the RodEpi repository [44] confirms that our
2015 visual architecture operates robustly without relying on
2016 low-level compression artifacts, the complete lack of syn-
2017 chronized telemetry in this dataset renders direct multi-
2018 modal comparisons impossible. To advance the field, there
2019 is an urgent need for a large-scale, open-source dataset that
2020 pairs precisely synchronized video with dense electrophys-
2021 iology. Establishing guaranteed subject-exclusive data par-
2022 titions alongside rigorous, fine-grained severity annotations
2023 will be absolutely essential for standardizing benchmarking
2024 efforts across competing network architectures.

2025 **(5) Continuous and resource-constrained deployment.**
2026 Our evaluation relies on pre-segmented clips, shielding the
2027 models from the complexities of continuous data streams.
2028 Real-world continuous monitoring requires low-latency
2029 edge inference, automated event boundary detection, and
2030 near-zero false-positive rates over extended recording peri-
2031 ods. While lightweight backbones like X3D-M suggest that
2032 edge deployment is feasible and clinical frameworks such
2033 as vEpiNet [31] demonstrate broader viability the practical
2034 engineering challenges of continuous, low-latency animal
2035 monitoring remain completely unaddressed and will require
2036 substantial future development.

2037 **Annotation provenance and intertwined bias.** Ulti-
2038 mately, because the ground-truth labels originate from a
2039 simultaneous human review of both data streams, neither
2040 modality is evaluated against a strictly independent base-
2041 line. In practice, the EEG telemetry inherently drove the
2042 temporal event boundaries, while the visual phenotype dic-
2043 tated the assigned severity grade. Resolving these inter-
2044 twined biases will require future datasets to employ seg-
2045 regated, modality-specific expert annotations validated by
2046 rigorous inter-rater consensus. Only by doing so can we
2047 definitively decouple the true performance of the networks
2048 from the mechanics of the annotation process itself.