CAN MULTIMODAL LARGE LANGUAGE MODELS UNDERSTAND PATHOLOGIC MOVEMENTS? A PILOT STUDY ON SEIZURE SEMIOLOGY

Lina Zhang
ECE UCLA
linazhang@g.ucla.edu

Tonmoy Monsoor
ECE UCLA
mtonmoy@g.ucla.edu

Prateik Sinha
Mathematics UCLA
prateiksinha@g.ucla.edu

Mehmet Efe Lorasdagi
ECE UCLA
lorasdagi@g.ucla.edu

Chong Han
Mathematics UCLA
chongh3814@g.ucla.edu

Peizheng Li
Mercedes-Benz AG
peizheng.li@mercedes-benz.com

Yuan Wang
ECE Zhejiang University
yuan2.24@intl.zju.edu.cn

Jessica Pasqua
Neurology UCLA
jessica.pasqua@gmail.com

Colin McCrimmon
Neurology UCLA
CMccrimmon@mednet.ucla.edu

Rajarshi Mazumder
Neurology UCLA
Rmazumder@mednet.ucla.edu

Vwani Roychowdhury
ECE UCLA
vwani@g.ucla.edu

***Abstract*—Multimodal Large Language Models (MLLMs) have demonstrated robust capabilities in recognizing everyday human activities, yet their potential for analyzing clinically significant involuntary movements in neurological disorders remains largely unexplored. This pilot study evaluates the capability of MLLMs for automated recognition of pathological movements in seizure videos. We assessed the zero-shot performance of state-of-the-art MLLMs on 20 ILAE-defined semiological features across 90 clinical seizure recordings. MLLMs outperformed fine-tuned Convolutional Neural Network(CNN) and Vision Transformer(ViT) baseline models on 13 of 18 features without task-specific training, demonstrating particular strength in recognizing salient postural and contextual features while struggling with subtle, high-frequency movements. Feature-targeted signal enhancement (facial cropping, pose estimation, audio denoising) improved performance on 10 of 20 features. Expert evaluation showed that 94.3% of MLLM-generated explanations for correctly predicted cases achieved $\geq$60% faithfulness scores, aligning with epileptologist reasoning. These findings demonstrate the potential of adapting general-purpose MLLMs for specialized clinical video analysis through targeted preprocessing strategies, offering a path toward interpretable, efficient diagnostic assistance.**

***Index Terms*—Multimodal Large Language Models, Vision Language Models, pathologic movements, seizure semiology, signal enhancement, explainable AI**

I. Introduction

Pathological movements are clinically fundamental, acting as direct diagnostic and prognostic indicators for a wide array of involuntary motor disorders [1]–[4]. Assessing these movements typically relies on subjective clinical observation, and manual video annotation review remains labor-intensive and time-consuming, creating bottlenecks in clinical workflows. Automated video analysis has made important strides in recent years, yet existing approaches face fundamental limitations. Most discriminative deep learning methods suffer from limited feature coverage, targeting only high-salience manifestations while neglecting subtle but diagnostically important cues. They lack interpretability, offering probability scores without explanation, a critical gap in high-stakes clinical environments [5]. Moreover, they exhibit fragility to real-world conditions such as occlusion, lighting variation, and background noise [6]. These limitations stem from a core constraint: supervised learning requires large annotated datasets for each feature, yet expert-labeled clinical videos are scarce, and the pathological movement vocabulary is too rich to exhaustively model with fixed-label classifiers [7].

Recent MLLMs offer a compelling alternative paradigm [8]–[12]. Pretrained on massive web-scale corpora, MLLMs demonstrate open-vocabulary reasoning, responding to flexible natural language queries rather than predicting from fixed label sets. They generate natural language explanations that align with clinical descriptors, providing much-needed transparency. Given these capabilities, a critical question arises: Can general-purpose MLLMs, trained primarily on everyday voluntary actions, recognize the subtle, involuntary, and often ambiguous manifestations of pathological movements in real-world clinical videos?

Epileptic seizure semiology serves as an ideal representative benchmark for this pilot investigation [13]. Seizures manifest through a diverse, temporally evolving spectrum of features, including motor, facial, autonomic, and vocal behaviors, constituting a comprehensive test for video understanding. Semiology spans from high-salience convulsions to subtle cues like oral automatisms and eye deviation. Demonstrating efficacy in seizure semiology would provide strong evidence for MLLMs'

potential in broader pathological movement analysis.

This pilot study presents the first systematic evaluation of MLLMs for comprehensive seizure semiology recognition. Our contributions are threefold:

Zero-shot MLLM Benchmarking: MLLMs outperformed fine-tuned CNN/ViViT baselines on 13/18 features across 20 ILAE-defined semiological features in 90 seizure videos, without task-specific training.

Feature-Targeted Signal Enhancement: Preprocessing strategies (facial cropping, pose overlays, audio denoising) improved performance on 10/20 features as a lightweight alternative to fine-tuning.

Clinical Explainability Analysis: MLLMs generated clinically interpretable justifications with 94.3% achieving $\geq$60% faithfulness scores, aligning with neurologist reasoning patterns.

II. Related Work

A. Discriminative models for pathological movement understanding

General-purpose video understanding models have made significant strides in action recognition. Spatiotemporal CNNs such as SlowFast [14] employ dual-pathway architectures to capture motion at multiple temporal resolutions, while self-supervised approaches like VideoMAE [15] learn robust representations through masked spatiotemporal prediction. VideoCLIP [16] further bridges vision and language by aligning video embeddings with text descriptions via contrastive learning. Despite their strong performance on everyday activities, these models are trained on general-domain action datasets (sports, daily routines) and fundamentally lack the clinical taxonomies needed to distinguish pathological movements from superficially similar voluntary behaviors.

Medical-specific vision models have emerged to address domain challenges in healthcare imaging. Models like MedViT [17] leverage vision transformers for tumor detection in CT scans or lesion classification in dermatology images. However, these architectures are designed for static pathology recognition and cannot capture the temporal motor dynamics that define neurological semiology. Pathological movements such as tonic-clonic seizures unfold over seconds, requiring joint modeling of spatial appearance and temporal evolution—capabilities absent in static imaging pipelines.

Seizure-specific detection systems represent the narrowest tier of existing work. Prior efforts employ specialized modules for isolated features: 3D CNNs for tonic-clonic detection [18], accelerometry-based classifiers [19], and optical flow segmentation [6]. While effective for their targeted symptoms, these approaches address single features in isolation, requiring separate models for each semiological component and yielding fragmented, non-scalable solutions.

These limitations reflect three fundamental gaps in discriminative approaches for pathological movement analysis. First, existing models suffer from taxonomic misalignment: general-purpose architectures cannot natively recognize the clinically-defined vocabulary of involuntary movements without extensive task-specific retraining, while medical models remain confined to static imaging tasks. Second, the data scarcity problem remains unresolved—supervised learning demands large expert-annotated corpora that do not exist for rare neurological conditions, and current self-supervised methods have not demonstrated efficacy on pathological movement tasks. Third, all these systems lack clinical explainability: they output probability scores or binary classifications without natural language justifications, a critical barrier in high-stakes medical environments where clinicians must understand and validate automated decisions before acting on them.

B. MLLMs for video understanding and medical applications

In contrast to discriminative models, MLLMs represent a fundamentally different generative paradigm. By unifying vision encoders with large language models, MLLMs such as GPT-4V [20], Gemini [21], and open-source alternatives like InternVL [22] and Qwen-VL [23] can process visual inputs with natural-language questions and generate free-form natural language descriptions. This generative capability has proven transformative for movement recognition: models can not only classify actions but also explain why a particular action was identified, describe contextual details, and respond to nuanced queries—capabilities unattainable with traditional classifiers.

Recent MLLMs have shown substantial promise primarily in the context of daily-life activity understanding, scene dynamics, and fine-grained motion reasoning [24]. The medical AI community has increasingly adopted MLLMs for clinical tasks, demonstrating strong performance in radiology report generation [25], dermatology diagnosis [26], and medical visual question answering [27]. However, these applications predominantly focus on static image analysis—interpreting chest X-rays, identifying skin lesions, or answering questions about anatomical structures. The temporal dimension remains underexplored: current medical MLLMs have not been systematically evaluated on pathological movement recognition, where clinical diagnoses depend on observing motor patterns evolving over time.

III. method

To systematically assess the potential of general-purpose MLLMs on pathological movement, we designed a comparative evaluation framework. Our methodology proceeds in two analytical stages. First, we benchmark the zero-shot performance of off-the-shelf MLLMs against task-specific, fine-tuned deep learning baselines. This comparison aims to determine whether generalist models can approximate the diagnostic accuracy of specialized supervised models without training cost. Second, we investigate if feature-targeted signal enhancement (e.g., facial cropping, pose estimation) can improve MLLMs zero-shot capabilities further (Fig. 1).

A. Clinical dataset curation

We collected 90 seizure videos from 29 consecutive adult patients undergoing video-EEG monitoring at UCLA Medical Center (2019–2023). Recordings were obtained using a

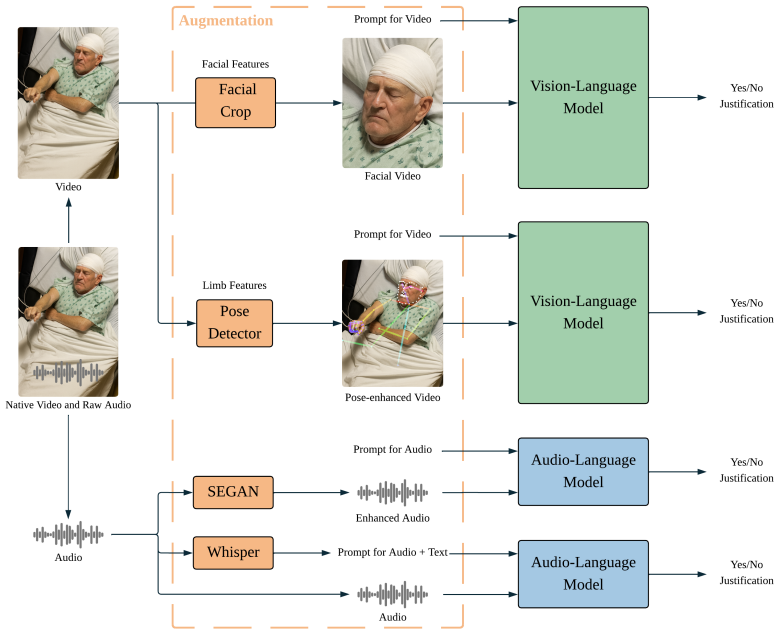

Fig. 1. MLLM-based medically induced involuntary movements recognition workflow with and without signal enhancement. (The person shown is an AI-generated virtual figure, not a real patient.)

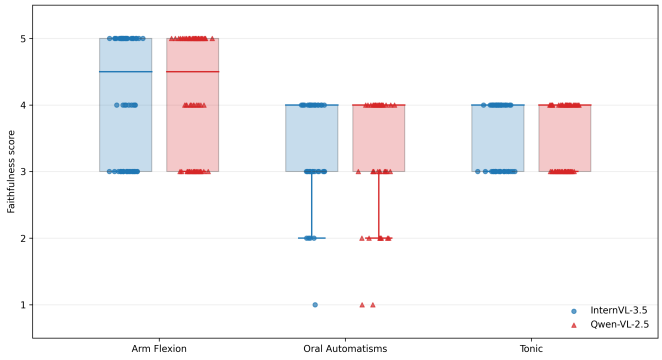

Fig. 2. MLLM justification faithfulness score distribution across 3 semiological features.

fixed overhead SONY EP 580 camera with a resolution of 1920 × 1080 pixels at 30 frames per second, with audio captured via unit-installed microphones at 44.1 kHz mono. To establish a robust ground truth, all videos were independently annotated by three epileptologists for the presence or absence of 20 standardized ILAE semiological motor features [28], [29].

B. Direct pathological movement recognition

Following established practices in video-based action recognition, we selected CNN [30] and Video Vision Transformer (ViViT) [31] architectures as comparative baselines. Both models were trained on the Kinetics dataset, which involves human actions [32]. They have high performance, so they provide a strong baseline for our task. Unlike MLLMs, which we evaluate in a zero-shot manner, these traditional baselines function as supervised classifiers and require task-specific fine-tuning; therefore, we initialized them from pretrained checkpoints and fine-tuned them separately for each feature using clinician-provided annotations. To ensure robust evaluation and prevent information leakage, we employed a patient-stratified approach rather than a simple video-level split. Specifically, we utilized three-fold cross-validation where patients were randomly partitioned into three distinct folds. Decision thresholds were calibrated independently for each feature by maximizing the F1 score on the training folds before being applied to the held-out test fold.

We then employed state-of-the-art MLLMs to extract pathological movement features from seizure videos without any task-specific fine-tuning. Given that subtle, clinically-defined involuntary movements can easily be confused with voluntary actions or normal behaviors, prompt design is critical for accurate feature extraction. We developed prompts collaboratively with three epileptologists, using clear, descriptive language rather than technical medical terminology to ensure the MLLMs could accurately interpret and identify the intended features. For instance, instead of simply asking about ”oral automatisms,” the prompt describes the specific observable behaviors: ”Does the patient exhibit repetitive, stereotyped mouth or tongue movements such as chewing, lip-smacking, or swallowing?” This approach translates clinical expertise into precise, accessible descriptions that leverage the MLLMs’ general understanding while guiding them toward clinically relevant observations. Representative examples of these expert-informed prompts are shown in Table I. Using these specialized prompts, we deployed InternVL-3.5 38B

TABLE I
SEIZURE SEMIOLOGICAL FEATURES AND MLLM PROMPTS SAMPLES.

Feature	MLLM Prompt
Oral Automatisms	Does the patient exhibit repetitive, stereotyped mouth or tongue movements such as chewing, lip-smacking, or swallowing?
"Figure4" Arms Posture	Does the patient's posture resemble a "figure-4" pattern, with one arm flexed and the other extended?
Pelvic Thrusting	Does the patient display repetitive, rhythmic, anteroposterior (forward-and-backward) movements of the hips?
Ictal Vocalization	Does the patient make any groaning, moaning, guttural sounds or do they utter stereotyped repetitive phrases?

[22] and Qwen-VL-2.5 32B [23] to extract visual semiology from video segments, while Audio Flamingo 3 (AF3) [33] was utilized to detect auditory features from the full audio recordings.

To facilitate a direct comparison, common preprocessing and inference protocols were applied. Videos were temporally downsampled to 2 fps to balance computational efficiency with the capture of dynamic movements. During inference, each seizure recording was divided into 30-second segments, with a 5-second overlap between consecutive segments, to ensure coverage of semiological events that may span segment boundaries. Finally, segment-level detection results for both baselines and MLLMs were aggregated across the full recording using an "any-yes" criterion: a given semiological feature was marked as present if it was detected in at least one segment of the video.

C. Feature-targeted signal enhancement

Pathological movements during seizures present inherent challenges for automated detection: they often involve rapid dynamics (e.g., clonic jerks), subtle manifestations (e.g., facial twitching), and are frequently masked by clinical noise such as background conversations or medical staff interventions appearing in the camera view. These characteristics make direct feature extraction from raw videos difficult for MLLMs. Inspired by clinical practice, where neurologists instinctively focus attention on specific body regions or filter out environmental noise to better identify subtle pathological movements, we hypothesized that targeted signal enhancement could similarly guide MLLMs toward clinically salient features. We grouped the 20 semiological features into 3 categories (i) facial features (ii) limb features, (iii) audio features and introduced category specific pre-processing enhancement procedures.

1) Facial Feature Enhancement: For facial features, such as eye closure and facial pulling, we hypothesized that restricting the field of view to the patient's face would help the MLLM focus on clinically relevant cues. We therefore applied face detection with temporal smoothing and cropped this region before passing the frames to the MLLM.

2) Limb Feature Enhancement: For features, such as arm flexion, tonic, and clonic, we used a pose detector (OpenPose [34]) to identify the coordinates of the patient's key limb joints in each frame. The resulting partial-skeleton was superimposed on each frame as additional information for the MLLM.

3) Audio Feature Enhancement: Key auditory features, such as ictal vocalizations and verbal responsiveness, are often masked by clinical noise (e.g., alarms, conversations). To reduce interference, we integrated a SEGAN-based speech enhancement module as a front-end preprocessing step [35]. To further explore the role of contextual information, we supplemented each audio clip with its corresponding transcript extracted using OpenAI's Whisper speech recognition model (large) to convert the WAV files to text format [36]. The audio was treated as the primary input and the transcript was provided as secondary evidence.

IV. RESULTS

A. MLLMs' zero-shot ability in detecting seizure semiological features

The best performing MLLMs outperformed the task-specific CNN and ViViT baselines (fine-tuned per feature) on 13/18 semiological features by F1 score (4/7 facial and 9/11 limb & body). On facial semiology, MLLMs achieved clear F1 gains for *closed eyes* (best baseline $F1 = 0.410 \rightarrow$ best MLLM $F1 = 0.524$), *blank stare* ($0.583 \rightarrow 0.632$), and *face pulling* ($0.463 \rightarrow 0.521$), and provided a smaller but consistent improvement on *face twitching* ($0.531 \rightarrow 0.548$). In limb & body semiology, MLLMs improved substantially on *arm straightening* ($0.447 \rightarrow 0.582$), *Figure 4* ($0.332 \rightarrow 0.462$), and *tonic* events ($0.506 \rightarrow 0.537$), while also exceeding baselines on broader contextual or salient motor patterns such as occur during sleep ($0.733 \rightarrow 0.771$) and arm flexion ($0.731 \rightarrow 0.800$). Notably, the MLLM advantage was not uniform: CNN/ViViT retained higher F1 on several fine-grained or high-frequency movement features, including *eye blinking* (best baseline $F1 = 0.388$ vs best MLLM $F1 = 0.250$), *head turning* (0.325 vs 0.320), *oral automatisms* (0.524 vs 0.479), *asynchronous movement* (0.690 vs 0.575), and *full body shaking* (0.513 vs 0.375).

MLLM performance patterns suggest that zero-shot generalization is strongest when the semiology is either contextual (scene-level) or visually unambiguous, and weakest when it depends on subtle, brief, or rapidly alternating motions. For example, occur during sleep reached $F1 = 0.771$ at 0.822 accuracy, consistent with reliable scene understanding. Similarly, distinct motor patterns such as arm flexion achieved $F1 = 0.800$, and the best MLLM variants performed competitively on several sustained motor phenomena (e.g., tonic $F1 = 0.537$). In contrast, performance degraded for semiologies dominated by small-amplitude facial dynamics or high-temporal-frequency movements: eye blinking remained low even for the best MLLM configuration ($F1 = 0.098$), and head turning suffered from low or unstable precision/recall trade-offs (best MLLM $F1 = 0.320$ despite high accuracies up to 0.811), indicating that these cues are often missed or confounded. Overall, these results are consistent with a zero-

TABLE II
FACIAL FEATURES PERFORMANCE. COMPARISON OF TRADITIONAL MODELS, VLMS, AND SIGNAL ENHANCED VLMS.

	Blank stare						**Closed eyes**						**Eye blinking**					
	CNN	*ViViT*	*Qwen*	*Intern*	*Crop+Qwen*	*Crop+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Crop+Qwen*	*Crop+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Crop+Qwen*	*Crop+Intern*
Accuracy	0.400	0.411	0.544	0.456	0.533	0.433	0.625	0.327	0.535	0.267	0.477	0.267	0.747	0.645	0.575	0.770	0.655	0.805
Precision	0.413	0.421	0.486	0.442	0.480	0.433	0.417	0.263	0.361	0.267	0.317	0.267	0.384	0.224	0.074	0.000	0.192	0.000
Recall	0.922	0.956	0.897	0.974	0.923	1.000	0.514	0.852	0.957	1.000	0.826	1.000	0.444	0.556	0.143	0.000	0.357	0.000
F1 Score	0.569	0.583	0.631	0.608	**0.632**	0.605	0.410	0.393	**0.524**	0.422	0.458	0.422	**0.388**	0.314	0.098	0.000	0.250	0.000
	Face pulling						**Face twitching**						**Oral automatisms**					
	CNN	*ViViT*	*Qwen*	*Intern*	*Crop+Qwen*	*Crop+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Crop+Qwen*	*Crop+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Crop+Qwen*	*Crop+Intern*
Accuracy	0.333	0.311	0.611	0.411	0.478	0.489	0.433	0.533	0.378	0.378	0.378	0.378	0.456	0.444	0.500	0.533	0.444	0.456
Precision	0.320	0.309	0.312	0.239	0.295	0.373	0.388	0.481	0.372	0.378	0.378	0.378	0.375	0.362	0.354	0.359	0.354	0.333
Recall	0.926	0.926	0.172	0.379	0.448	0.862	0.861	0.699	0.941	1.000	1.000	1.000	0.889	0.833	0.548	0.452	0.742	0.581
F1 Score	0.463	0.453	0.222	0.293	0.356	**0.521**	0.531	0.527	0.533	**0.548**	**0.548**	**0.548**	**0.524**	0.503	0.430	0.400	0.479	0.424
	Head turning																	
	CNN	*ViViT*	*Qwen*	*Intern*	*Crop+Qwen*	*Crop+Intern*												
Accuracy	0.667	0.611	0.811	0.800	0.767	0.800												
Precision	0.374	0.270	0.571	0.000	0.364	0.000												
Recall	0.417	0.486	0.222	0.000	0.222	0.000												
F1 Score	**0.325**	0.317	0.320	0.000	0.276	0.000												

TABLE III
LIMB & BODY FEATURES PERFORMANCE. COMPARISON OF TRADITIONAL MODELS, VLMS, AND SIGNAL ENHANCED VLMS.

	Occur during sleep						**Arm flexion**						**Arms move simultaneously**					
	CNN	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*
Accuracy	0.822	0.722	0.778	0.822	0.778	0.738	0.611	0.611	0.744	0.722	0.630	0.619	0.518	0.519	0.578	0.278	0.321	0.214
Precision	0.783	0.585	0.933	0.730	0.741	0.600	0.600	0.609	0.719	0.724	0.597	0.592	0.289	0.254	0.318	0.253	0.194	0.205
Recall	0.689	0.475	0.424	0.818	0.645	1.000	0.941	0.885	0.902	0.824	0.881	0.933	0.657	0.567	0.636	1.000	0.706	1.000
F1 Score	0.733	0.510	0.583	**0.771**	0.690	0.750	0.731	0.720	**0.800**	0.771	0.712	0.724	0.400	0.305	**0.424**	0.404	0.304	0.340
	Arm straightening						**Figure 4**						**Tonic**					
	CNN	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*
Accuracy	0.344	0.356	0.633	0.644	0.580	0.548	0.511	0.789	0.922	0.789	0.568	0.298	0.444	0.500	0.711	0.711	0.642	0.631
Precision	0.300	0.316	0.442	0.444	0.388	0.353	0.086	0.256	0.600	0.211	0.135	0.119	0.401	0.367	0.667	0.600	0.417	0.474
Recall	0.933	0.875	0.852	0.741	0.826	0.783	0.567	0.700	0.375	0.500	0.625	1.000	0.511	0.847	0.207	0.310	0.185	0.621
F1 Score	0.447	0.442	**0.582**	0.556	0.528	0.486	0.126	0.332	**0.462**	0.296	0.222	0.213	0.321	0.506	0.316	0.409	**0.537**	**0.537**
	Clonic						**Limb automatisms**						**Asynchronous movement**					
	CNN	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*
Accuracy	0.6	0.778	0.667	0.700	0.531	0.548	0.678	0.300	0.356	0.322	0.395	0.310	0.678	0.700	0.622	0.656	0.531	0.524
Precision	0.293	0.587	0.290	0.333	0.205	0.231	0.367	0.183	0.224	0.230	0.254	0.256	0.579	0.665	0.621	0.656	0.433	0.308
Recall	0.808	0.408	0.529	0.588	0.533	0.529	0.315	0.546	0.714	0.810	0.750	1.000	0.861	0.710	0.439	0.512	0.382	0.114
F1 Score	0.421	0.409	0.375	**0.426**	0.296	0.321	0.316	0.274	0.341	0.358	0.380	**0.408**	**0.690**	0.674	0.514	0.575	0.406	0.167
	Pelvic thrusting						**Full body shaking**											
	CNN	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*	*CNN*	*ViViT*	*Qwen*	*Intern*	*Pose+Qwen*	*Pose+Intern*						
Accuracy	0.644	0.589	0.778	0.756	0.432	0.607	0.598	0.528	0.644	0.556	0.395	0.405						
Precision	0.312	0.189	0.353	0.370	0.218	0.235	0.410	0.310	0.318	0.300	0.224	0.242						
Recall	0.733	0.467	0.400	0.667	0.800	0.533	0.783	0.733	0.292	0.500	0.765	0.833						
F1 Score	0.423	0.261	0.375	**0.476**	0.343	0.327	**0.513**	0.412	0.304	0.375	0.347	0.375						

shot MLLM regime that is effective for coarse contextual and sustained motor signatures.

B. Effect of feature-targeted signal enhancement

Feature-targeted pre-processing proved useful, though not universally effective, in boosting MLLMs zero-shot performance. As shown in Tables II, III, and IV, enhancements improved for 10 out of the 20 semiological features. The information added by preprocessing acts like a domain-specific attention mechanism, directing models toward clinically salient cues otherwise masked by distractors.

Facial semiological features showed most gains, and included enhanced recognition of blank stare, blinking, face pulling, twitching, oral automatisms, and sleep-related events, though sometimes with reduced precision (e.g., blinking). Pose estimation offered a strong abstraction layer: for tonic movements, InternVL-3.5 38B's F1 rose from 0.409 to 0.537, and limb automatisms improved consistently across VLMs.

For audio, SEGAN-based denoising alone provided little benefit, likely because generative filtering altered seizure-specific sounds. By contrast, pairing transcripts with audio

TABLE IV
AUDIO FEATURES PERFORMANCE. COMPARISON OF ALM(AF3) AND SIGNAL ENHANCEMENT WITH ALM.

	Verbal responsiveness			**Ictal vocalization**		
	AF3	*Segan + AF3*	*ASR + AF3*	*AF3*	*Segan + AF3*	*ASR + AF3*
Accuracy	0.434	0.321	0.245	0.765	0.581	0.744
Precision	0.468	0.375	0.431	0.850	0.654	0.759
Recall	0.361	0.291	0.327	0.708	0.500	**0.830**
F1 Score	0.380	0.286	0.193	0.773	0.567	**0.793**

TABLE V
MLLM GENERATED SEMIOLOGICAL FEATURE JUSTIFICATION SAMPLES WITH FAITHFULNESS SCORE.

Feature	**Justification**	**Score**
Oral Automatisms	The patient is observed holding and interacting with a tablet, conversing with the nurse, and raising arms to make gestures across multiple video segments; none of these actions involve repetitive, stereotyped mouth or tongue movements indicative of oral automatisms.	4
Oral Automatisms	The patient in the video doesn't exhibit oral automatisms.	2
Arm flexion	The patient does not flex their arms at the elbows throughout the video segments, as they are consistently holding and interacting with a laptop.	5

helped detect ictal vocalizations, slightly raising F1 (0.77 $\rightarrow$ 0.79) through improved recall. However, for verbal responsiveness, extra text reduced precision and F1, as the model sometimes misattributed background speech to the patient.

In summary, while targeted enhancements boosted recognition of several visual and auditory features, they also introduced risks, false positives, loss of context, or confusion, highlighting the need for further strengthening feature-specific preprocessing methodologies introduced in this paper.

C. MLLMs explainability on seizure semiology

Beyond predicting the presence or absence of semiological features, MLLMs can generate free-form natural language *justifications* describing the visual and auditory cues that support their decisions. This capability is particularly relevant in epilepsy care, where clinicians routinely rely on narrative interpretations of behaviors to characterize seizure type and infer likely seizure onset networks. Table V highlights representative explanations for multiple semiological categories, illustrating that modern MLLMs can produce descriptions that resemble clinical phrasing used in EMU reports.

To quantitatively assess explanation quality, we conducted a structured evaluation focusing on three representative features: arm flexion, oral automatisms, and tonic movements. For each feature, expert epileptologists evaluated MLLM-generated justifications from correctly predicted samples, including all true positive cases and an equal number of randomly sampled true negative cases, using a faithfulness score ranging from 1 to 5, where each score level represents incrementally higher correctness: 1 (20%), 2 (40%), 3 (60%), 4 (80%), and 5 (100%). Higher scores reflect justifications that are more specific, evidence-grounded, and clinically accurate in their description of observable features.

Figure 2 summarizes the faithfulness score distributions. Overall, MLLMs provided reliable and clinically interpretable explanations, with 94.3% justifications scoring $\geq$3 (60% correctness or higher). However, explanation quality varied by feature type. Salient motor behaviors (arm flexion) predominantly received scores of 4-5 (median 4.5), while oral automatisms (median 3.9) and tonic (median 4) showed more scores in the 3-4 range. This disparity reflects the inherent difficulty of describing subtle facial movements compared to clear postural patterns. These findings confirm that MLLMs can provide clinically valuable explanations that support clinician-in-the-loop review, particularly for prominent motor features.

V. DISCUSSION AND CONCLUDING REMARKS

This study demonstrates that general-purpose MLLMs can effectively recognize pathological movements in seizure semiology without task-specific training. Our key findings are threefold: First, zero-shot MLLMs outperformed fine-tuned CNN and ViViT baselines on 13 of 18 semiological features, achieving superior F1 scores despite requiring no domain-specific training. Second, Feature-targeted signal enhancement further improved performance on 10 of 20 features, offering a computationally efficient alternative to model fine-tuning. Techniques like pose estimation and facial cropping act as domain-specific attention mechanisms, guiding MLLMs toward clinically salient cues. Third, MLLMs provide clinically valuable explainability through natural language justifications that align with neurologist reasoning, with 94.3% explanations scoring $\geq$ 60% faithfulness. This interpretability is critical for high-stakes clinical deployment, addressing a fundamental limitation of traditional discriminative models that only output probability scores.

The limitations of our work remain apparent. Firstly, the zero-shot accuracy for many subtle or complex features is not yet at a clinical-grade level. Secondly, our dataset, while clinically authentic, is from a single center, which may limit the generalizability of our findings. Future work should focus on two key areas: (1) Domain-specific fine-tuning of MLLMs on larger, more diverse multi-center seizure video datasets to improve their grasp of nuanced clinical cues. (2) Exploring

more sophisticated methods for multimodal fusion methods, together with adaptive attention allocation mechanisms, to further enhance model accuracy and robustness.

MLLM-based systems could serve as intelligent screening assistants, automatically analyzing clinical videos and generating interpretable summaries of pathological movements. This approach significantly accelerates diagnostic workflows while maintaining transparency, establishing a practical pathway toward interpretable clinical AI.

REFERENCES

[1] R. Martínez-García-Peña, L. H. Koens, G. Azzopardi, and M. A. J. Tijssen, "Video-based data-driven models for diagnosing movement disorders: Review and future directions," *Movement Disorders*, vol. 40, no. 10, pp. 2046–2066, 2025.

[2] L. Timmermann, J. Gross *et al.*, "Mini-asterixis in hepatic encephalopathy induced by pathologic thalamo-motor-cortical coupling," *Neurology*, pp. 295–298, 2002.

[3] T. Uchihara and H. Tsukagoshi, "Limb myokymia with involuntary finger movement caused by peripheral neuropathy due to systemic lupus erythematosus: a case report," *Clinical Neurology and Neurosurgery*, pp. 321–324, 1994.

[4] T. Mainka, T. Büttner *et al.*, "The spectrum of involuntary vocalizations in humans: A video atlas," *Movement Disorders*, pp. 1036–1045, 2019.

[5] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 310, 2020.

[6] S. Kalitzin, G. Petkov, D. Velis, B. Vledder, and F. Lopes da Silva, "Automatic segmentation of episodes containing epileptic clonic seizures in video sequences," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 12, pp. 3379–3385, 2012.

[7] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. S. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 27, no. 2, pp. 166–176, 2021.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 28 492–28 518. [Online]. Available: https://arxiv.org/abs/2212.04356

[10] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc *et al.*, "Flamingo: a visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022. [Online]. Available: https://arxiv.org/abs/2204.14198

[11] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[12] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[13] R. S. Fisher, J. H. Cross, C. D'Souza, J. A. French, S. R. Haut, N. Higurashi, E. Hirsch, F. E. Jansen, L. Lagae, S. L. Moshé *et al.*, "Operational classification of seizure types by the international league against epilepsy: Position paper of the ilae commission for classification and terminology," *Epilepsia*, vol. 58, no. 4, pp. 522–530, 2017.

[14] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[15] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[16] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 6787–6800.

[17] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "Medvit: a robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 157, p. 106791, 2023.

[18] A. Boyne, H. J. Yeh, A. K. Allam, B. M. Brown, M. Tabaeizadeh, J. M. Stern, R. J. Cotton, and Z. Haneef, "Video-based detection of tonic-clonic seizures using a three-dimensional convolutional neural network," *Epilepsia*, vol. 66, no. 7, pp. 2495–2506, 2025.

[19] M.-Z. Poh, T. Loddenkemper, C. Reinsberger, N. C. Swenson, S. Goyal, M. C. Sabtala, J. R. Madsen, and R. W. Picard, "Convulsive seizure detection using a wrist-worn electrodermal activity and accelerometry biosensor," *Epilepsia*, vol. 53, no. 5, pp. e93–e97, 2012.

[20] J. Achiam, S. Adler, S. Agarwal *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[21] Gemini Team, Google, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023. [Online]. Available: https://arxiv.org/abs/2312.11805

[22] W. Wang, Z. Gao *et al.*, "Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv:2508.18265*, 2025.

[23] S. Bai, K. Chen *et al.*, "Qwen2.5-vl technical report," *arXiv:2502.13923*, 2025.

[24] W. Hong, Y. Cheng, Z. Yang, W. Wang, L. Wang, X. Gu, S. Huang, Y. Dong, and J. Tang, "Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 8450–8460.

[25] X. Zhang, Z. Meng, J. Lever, and E. S. L. Ho, "Gla-ai4biomed at rrg24: Visual instruction-tuned adaptation for radiology report generation," in *Proceedings of the BioNLP Workshop and Shared Task (ACL)*, 2024. [Online]. Available: https://aclanthology.org/2024.bionlp-1.54/

[26] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, S. Afvari, and X. Gao, "Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4," *Nature Communications*, vol. 15, p. 5606, 2024. [Online]. Available: https://www.nature.com/articles/s41467-024-50043-3

[27] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, and P. Rajpurkar, "Med-flamingo: a multimodal medical few-shot learner," in *Proceedings of Machine Learning Research (PMLR), 3rd Machine Learning for Health Symposium (ML4H)*, vol. 225, 2023, pp. 353–367.

[28] R. S. Fisher, J. H. Cross *et al.*, "Ilae classification of seizures: The 2017 revision and update," *Epilepsia*, pp. 522–530, 2017.

[29] ——, "Instruction manual for the ILAE 2017 operational classification of seizure types," *Epilepsia*, pp. 531–542, 2017.

[30] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[31] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[32] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[33] A. Goel, S. Ghosh, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle, and B. Catanzaro, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," *arXiv preprint arXiv:2507.08128*, 2025. [Online]. Available: https://arxiv.org/abs/2507.08128

[34] Z. Cao, T. Simon *et al.*, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, pp. 7291–7299.

[35] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," 2017.

[36] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356