ZERO-SHOT UNDERSTANDING OF MEDICALLY-INDUCED MOVEMENTS USING MULTIMODAL LARGE LANGUAGE MODELS

Lina Zhang[1], *Prateik Sinha*[3], *Tonmoy Monsoor*[1], *Chong Han*[3], *Peizheng Li*[3,4], *Yuan Wang*[1],
Jessica Pasqua[2], *Colin McCrimmon*[2], *Victor Morales*[2], *Hailey Marie Miranda*[2],
Rajarshi Mazumder[2], *Vwani Roychowdhury*[1]

[1]Department of Electrical and Computer Engineering, UCLA
[2]Department of Neurology, David Geffen School of Medicine , UCLA
[3]Department of Mathematics, UCLA
[4]Mercedes-Benz AG, [5]University of Tuebingen

ABSTRACT

Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated strong performance in recognizing voluntary human actions, yet their application to clinically significant involuntary movements, particularly in the context of neurological disorders, remains underexplored. This paper evaluates the zero-shot capabilities of MLLMs to identify and interpret 20 distinct behavioral features across 90 clinical seizure videos, and further examines the effect of feature targeted signal enhancement on model performance. The MLLM showed promising results in recognizing clear postural and gaze-related features and provided human-aligned natural language justifications for its predictions, enhancing model explainability, but its performance declined for complex motor events like tonic-clonic activity and subtle manifestations such as oral automatisms. Feature targeted signal enhancement led to performance improvements in 10 out of the 20 features. These findings highlight the potential of adapting general-purpose MLLMs for specialized clinical applications through targeted signal enhancement strategies.

Index Terms— Multimodal Large Language Models, Vision Language Models, involuntary movements, seizure semiology, signal enhancement

1. INTRODUCTION

Recent advances in Vision Language Models (VLMs) like CLIP [1], SwinBERT [2], VLTinT [3], ODMO [4], alongside Audio Language Models (ALMs) [5], have greatly improved machine learning (ML) systems' ability to recognize and describe voluntary human actions through spatiotemporal and semantic understanding. However, their application to involuntary movements, particularly those arising from medical conditions, remains limited.

While many involuntary movements arise from diverse physiological and pathological origins, like asterixis in hepatic encephalopathy [6], tremors in thyrotoxicosis and drug-induced syndromes [7], neuropsychiatric manifestations of autoimmune diseases [8], hiccups or stridor from psychiatric or infectious causes [9, 10]. Seizures, with their stereotyped, temporally evolving semiological features—like automatisms, ictal vocalizations, head turning, tonic and clonic movements provides critical insights into seizure localization and classification[11, 12]. Yet, most ML models address these features in isolation and lack holistic, interpretable frameworks. Prior works use specialized modules for specific semiological components (e.g., 3D CNNs for tonic-clonic detection [13], accelerometry-based models [14], optical flow segmentation [15]), resulting in fragmented and non-scalable solutions. Although explainable AI techniques have been applied to seizure electroencephalography(EEG) signal [16, 17], few models unify audio-visual temporal reasoning with interpretability.

To address these gaps, we propose a Multimodal Large Language Models (MLLMs) framework that processes video and audio inputs to identify and interpret 20 semiological features, without any targeted fine-tuning (zero-shot capabilities), from 90 seizure videos recorded in an Epilepsy Monitoring Unit (EMU). In addition, we also demonstrate that feature-targeted signal enhancement significantly boosts the zero-shot capabilities of MLLMs. Our approach illustrates the potential of MLLMs in understanding and explaining complex involuntary behaviors in context of seizures, laying the groundwork for future fine-tuning and signal optimization.

2. METHOD

Our methodology can be categorized into two sequential stages: in the first stage we evaluate zero-shot perfor-

Corresponding Author: vwani@g.ucla.edu

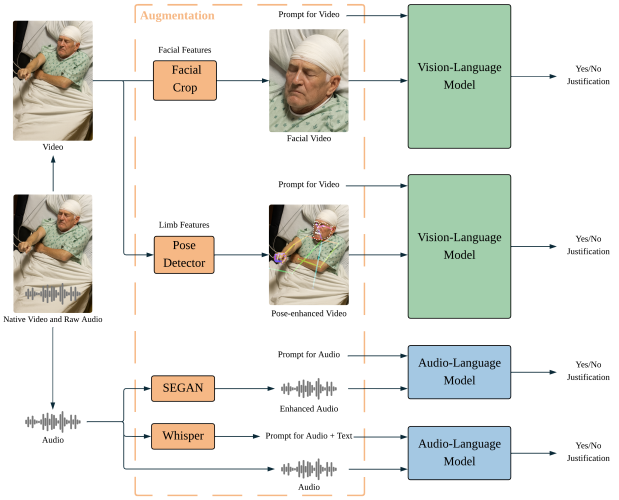

Fig. 1. MLLM-based feature extraction workflow with and without signal enhancement. (The person shown is an AI-generated virtual figure, not a real patient.)

mance of MLLMs on detecting and interpreting semiological features from raw seizure videos and in the second stage we determine if feature targeted signal processing can enhance MLLMs zero-shot capabilities.

2.1. Clinical dataset

We analyzed 90 video-recorded seizure events from 29 consecutive patients (aged >18 years) who underwent video-EEG monitoring at the UCLA Medical Center between 2019 and 2023. Recordings were obtained using a 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. Each recording included pre-ictal, ictal, and post-ictal phases of varying durations, with an average total length of approximately three minutes. All videos were independently annotated by three epileptologists for the presence or absence of 20 seizure semiological features, identified by the International League Against Epilepsy (ILAE) as essential for the systematic classification and clinical interpretation of seizures [18, 19].

2.2. Direct information extraction via prompting

The first phase of our experiment focused on direct information extraction from seizure recordings using state-of-the-art Multimodal Large Language Models (MLLMs), without any task-specific fine-tuning. This process relied on carefully designed prompts, developed collaboratively by three epileptologists based on their clinical expertise. Some representative prompts are shown in Table 1.

Table 1. Sample seizure semiological features and corresponding MLLM prompts.

Feature	**Prompt to MLLM**
Oral Automatisms	Does the patient exhibit repetitive, stereotyped mouth or tongue movements such as chewing, lip-smacking, or swallowing?
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?

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. For visual features extraction, we prompted state-of-the-art (SOTA) VLMs, specifically InternVL-3.5 38B [20] and Qwen-VL-2.5 [21] with the epileptologist designed prompts to extract visual semiological features from each video segment. Similarly for auditory feature extraction, we prompted Audio Flamingo 3 (AF3) to extract features like verbal responsiveness and ictal vocalization from the entire audio recording. The segment level extraction results of the VLMs were then aggregated across all segments of a recording using an any-yes criterion: a given semiological feature was marked as present if it was detected in at least one segment. This approach enabled robust detection of features while maintaining temporal sensitivity to localized seizure events.

2.3. Feature targeted signal enhancement

We grouped the 20 semiological features into three categories (i) facial features (ii) limb features and (iii) audio features and introduced category specific pre-processing enhancement procedures.

2.3.1. 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 (Fig. 1).

2.3.2. Limb Feature Enhancement

For features, such as arm flexion, tonic, and clonic, we used a pose detector (OpenPose [22]) 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 (Fig. 1).

2.3.3. 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 [23]. 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 [24]. The SEGAN-preprocessed audio was treated as the primary input and the transcript was provided as a secondary evidence (Fig. 1).

3. RESULTS

3.1. MLLMs zero-shot ability in detecting seizure semiological features

MLLMs performed well on high-level contextual features—for example, detecting whether a seizure occurred during sleep yielded high accuracy and F1 scores (Accuracy = 0.82,$F1$ = 0.77), reflecting strong scene understanding. They also showed promising performance on distinct motor and vocal events like arm flexion ($F1 = 0.77$) and ictal vocalization ($F1 = 0.77$), indicating an ability to recognize clear, unambiguous patterns. However, performance dropped sharply for features requiring nuanced interpretation of subtle or rapid muscle movements. Low F1 scores for eye blinking (0.1), head turning (0.32), and face pulling (0.29) highlight this gap (see Table 2). These features are often brief, small-scale posing challenges for models trained on general-domain data. Even with the aforementioned challenges *best performing MLLMs outperformed a naive bayes classifier on all three feature categories*: facial features (60% ↑), limb features (92% ↑), audio features (43% ↑).

3.2. MLLMs capabilities in explainable seizure semiology analysis

MLLMs not only demonstrate the capacity to identify semiological features in seizure videos but also provide accompanying natural language justifications for their predictions. Table 3 presents representative examples of model-generated explanations for various semiological features. These justifications often align with clinical descriptors used by epileptologists, suggesting that MLLMs can approximate human-like reasoning in describing complex visual and auditory phenomena.

Table 2. Performance metrics with and without signal enhancement. Each entry is reported as x|y (x = Qwen-VL-2.5, y = InternVL-3.5 38B). "Base" indicates no enhancement; "Crop / Pose / Segan / Text" are enhancement strategies. A ↑ marks features where at least one model improves with enhancement.

	Blank stare ↑		**Closed eyes**		**Eye blinking ↑**	
	Base	*Crop*	*Base*	*Crop*	*Base*	*Crop*
Accuracy	0.544 |0.456	0.533 |0.433	0.535 |0.267	0.477 |0.267	0.575 |0.770	**0.655** |**0.805**
Precision	0.486 |0.442	0.480 |0.433	0.361 |0.267	0.317 |0.267	0.074 |0.000	**0.192** |0.000
Recall	0.897 |0.974	**0.923** |**1.000**	0.957 |1.000	0.826 |1.000	0.143 |0.000	**0.357** |0.000
F1 Score	0.631 |0.608	**0.632** |0.605	0.524 |0.422	0.458 |0.422	0.098 |0.000	**0.250** |0.000
	Face pulling ↑		**Face twitching ↑**		**Oral automatisms ↑**	
	Base	*Crop*	*Base*	*Crop*	*Base*	*Crop*
Accuracy	0.611 |0.411	0.478 |**0.489**	0.378 |0.378	0.378 |0.378	0.500 |0.533	0.444 |0.456
Precision	0.312 |0.239	0.295 |**0.373**	0.372 |0.378	**0.378** |0.378	0.354 |0.359	0.354 |0.333
Recall	0.172 |0.379	**0.448** |**0.862**	0.941 |1.000	**1.000** |1.000	0.548 |0.452	**0.742** |**0.581**
F1 Score	0.222 |0.293	**0.356** |**0.521**	0.533 |0.548	**0.548** |0.548	0.430 |0.400	**0.479** |**0.424**
	Occur during sleep ↑		**Head turning**		**Arm flexion**	
	Base	*Pose*	*Base*	*Crop*	*Base*	*Pose*
Accuracy	0.778 |0.822	0.778 |0.738	0.811 |0.800	0.767 |0.800	0.744 |0.722	0.630 |0.619
Precision	0.933 |0.730	0.741 |0.600	0.571 |0.000	0.364 |0.000	0.719 |0.724	0.597 |0.592
Recall	0.424 |0.818	**0.645** |**1.000**	0.222 |0.000	0.222 |0.000	0.902 |0.824	0.881 |**0.933**
F1 Score	0.583 |0.771	**0.690** |0.750	0.320 |0.000	0.276 |0.000	0.800 |0.771	0.712 |0.724
	Arms move simultaneously		**Arm straightening**		**Figure 4**	
	Base	*Pose*	*Base*	*Pose*	*Base*	*Pose*
Accuracy	0.578 |0.278	0.321 |0.214	0.633 |0.644	0.580 |0.548	0.922 |0.789	0.568 |0.298
Precision	0.318 |0.253	0.194 |0.205	0.442 |0.444	0.388 |0.353	0.600 |0.211	0.135 |0.119
Recall	0.636 |1.000	**0.706** |1.000	0.852 |0.741	0.826 |**0.783**	0.375 |0.500	**0.625** |**1.000**
F1 Score	0.424 |0.404	0.304 |0.340	0.582 |0.556	0.528 |0.486	0.462 |0.296	0.222 |0.213
	Tonic ↑		**Clonic**		**Limb automatisms ↑**	
	Base	*Pose*	*Base*	*Pose*	*Base*	*Pose*
Accuracy	0.711 |0.711	0.642 |0.631	0.667 |0.700	0.531 |0.548	0.356 |0.322	**0.395** |0.310
Precision	0.667 |0.600	0.417 |0.474	0.290 |0.333	0.205 |0.231	0.224 |0.230	**0.254** |**0.256**
Recall	0.207 |0.310	0.185 |**0.621**	0.529 |0.588	**0.533** |0.529	0.714 |0.810	**0.750** |**1.000**
F1 Score	0.316 |0.409	0.256 |**0.537**	0.375 |0.426	0.296 |0.321	0.341 |0.358	**0.380** |**0.408**
	Asynchronous movement		**Pelvic thrusting**		**Full body shaking ↑**	
	Base	*Pose*	*Base*	*Pose*	*Base*	*Pose*
Accuracy	0.622 |0.656	0.531 |0.524	0.778 |0.756	0.432 |0.607	0.644 |0.556	0.395 |0.405
Precision	0.621 |0.656	0.433 |0.308	0.353 |0.370	0.218 |0.235	0.318 |0.300	0.224 |0.242
Recall	0.439 |0.512	0.382 |0.114	0.400 |0.667	**0.800** |0.533	0.292 |0.500	**0.765** |**0.833**
F1 Score	0.514 |0.575	0.406 |0.167	0.375 |0.476	0.343 |0.327	0.304 |0.375	**0.347** |0.375

	Verbal responsiveness			**Ictal vocalization ↑**		
	Base	*Segan*	*Text*	*Base*	*Segan*	*Text*
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**

3.3. Effect of feature targeted signal enhancement on MLLMs zero-shot ability

Feature-targeted pre-processing proved useful, though not universally effective, in boosting MLLMs zero-shot performance. As shown in Table 2, 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 in-

Table 3. Sample justifications provided by MLLM in support of semiological feature detection

Feature	Justification
Oral Automatisms	The patient in the video exhibits repetitive, stereotyped mouth movements, including chewing and lip-smacking actions.
Pelvic Thrusting	The patient is seen lying on their back on a hospital bed, wearing a white gown with black polka dots. The video shows the patient moving their hips in a repetitive, rhythmic, and anteroposterior manner, indicating forward-and-backward movements.
Ictal Vocalization	The patient makes groaning, moaning, and guttural sounds, which are indicative of physical discomfort or pain.

cluded 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 helped detect ictal vocalizations, slightly raising F1 (0.77 → 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.

4. DISCUSSION AND CONCLUDING REMARKS

This study demonstrates that modern MLLMs can be effectively repurposed for the complex task of seizure semiology recognition from clinical videos, even in a zero-shot setting. Our findings suggest a new paradigm where the vast world knowledge embedded in these general-purpose models serves as a powerful foundation for specialized medical analysis, potentially reducing the need for building bespoke models from scratch.

The principal finding is twofold. First, out-of-the-box MLLMs can identify a range of clinically relevant features and, crucially, explain their reasoning in human-readable language. This explainability is paramount in high-stakes clinical environments. Second, we have shown that targeted signal enhancement, framed as a proxy for domain-specific attention, can significantly boost performance. This approach offers a practical, computationally efficient alternative to full model fine-tuning, guiding a general model's focus without altering its internal weights. The success of techniques like pose estimation for tonic movements highlights the value of abstracting complex visual data into a more structured format for the MLLM to interpret.

While open-source MLLMs outperformed naive Bayes by large margins, showing average gains of 60%, 92%, and 43% for facial, limb, and audio features, it was infeasible to benchmark them against alternative non-MLLM video-understanding models. Existing efforts target only limited features (e.g., tonic, clonic, blinking) and to the best of our knowledge lack public code bases[13]. Moreover, building comprehensive baselines would require large-scale, expert-annotated seizure video datasets that are not currently available. We are collaborating with multiple institutions to create such a resource, though once available, fine-tuning and augmenting MLLMs remains a more powerful strategy, as they can jointly learn all features and provide explainable justifications.

The limitations of our work are quite 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.

The implications for clinical AI are significant. Such a system could function as an intelligent screening tool for neurologists, automatically analyzing lengthy video-EEG recordings, flagging segments of interest, and providing preliminary textual summaries of observed semiological features. This could dramatically expedite the diagnostic workflow in Epilepsy Monitoring Units. This study lays the groundwork for a new approach to medical video analysis, combining the power of large-scale pre-trained models with domain-specific signal processing to build transparent and effective clinical support tools.

5. REFERENCES

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, “Learning transferable visual models from natural language supervision,” *arXiv preprint arXiv:2103.00020*, 2021.

[2] Kevin Lin, Linjie Li, et al., “Swinbert: End-to-end transformers with sparse attention for video captioning,” *arXiv:2111.13196*, 2021.

[3] Koji Yamazaki, Kevin Vo, et al., “Vltint: Visual-linguistic transformer-in-transformer for coherent video paragraph captioning,” *arXiv:2211.15103*, 2022.

[4] Qiujing Lu, Yipeng Zhang, Mingjian Lu, and Vwani Roychowdhury, “Action-conditioned on-demand motion generation,” 2022.

[5] Xiulong Mei, Xinhao Liu, et al., “Diverse audio captioning via adversarial training,” in *ICASSP*, 2022, pp. 171–175.

[6] Juan Pablo Rissardo et al., “Flapping tremor: Unraveling asterixis—a narrative review,” *Medicina (Kaunas)*, p. 362, 2024.

[7] Lars Timmermann, Joachim Gross, et al., “Mini-asterixis in hepatic encephalopathy induced by pathologic thalamo-motor-cortical coupling,” *Neurology*, pp. 295–298, 2002.

[8] Toshiki Uchihara and Hiroshi 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.

[9] Thomas Mainka, Theresa Büttner, et al., “The spectrum of involuntary vocalizations in humans: A video atlas,” *Movement Disorders*, pp. 1036–1045, 2019.

[10] John L Faul, Kenneth C Wilson, et al., “Psychogenic cough, tic cough, and habit cough in adult and pediatric populations: Accp evidence-based clinical practice guidelines,” *Chest*, pp. 174S–179S, 2006.

[11] Linh T Nguyen et al., “Machine learning for automated seizure classification using semiological features from video-eeg,” in *EMBC*, 2022, pp. 3270–3273.

[12] Allan Krumholz, Richard G Wennberg, and John W Miller, “The epileptic seizure: Demystifying the ictal semiology,” *The Lancet Neurology*, pp. 305–317, 2023.

[13] James Boyne et al., “Video-based detection of tonic–clonic seizures using a three-dimensional convolutional neural network,” *Epilepsia*, pp. 2495–2506, 2025.

[14] Ivan Osorio, Mark G. Frei, et al., “Automated detection of tonic–clonic seizures using 3-d accelerometry and surface electromyography in pediatric patients,” *Epilepsy Behavior*, pp. 157–163, 2015.

[15] D. Downey et al., “Automatic segmentation of episodes containing epileptic clonic seizures in video sequences,” *Medical Engineering Physics*, pp. 763–771, 2012.

[16] Yong Yin, Jiawei Gong, et al., “Shap-driven feature analysis approach for epileptic seizure prediction,” *Biomedical Signal Processing and Control*, p. 104869, 2024.

[17] Nicolas Gaspard, Lawrence J Hirsch, and Aristea S Galanopoulou, “Semiology and seizure localization in the era of video-eeg and automated analysis,” *Epilepsy & Behavior*, p. 108963, 2023.

[18] Robert S. Fisher, J. Helen Cross, et al., “Ilae classification of seizures: The 2017 revision and update,” *Epilepsia*, pp. 522–530, 2017.

[19] Robert S. Fisher, J. Helen Cross, et al., “Instruction manual for the ILAE 2017 operational classification of seizure types,” *Epilepsia*, pp. 531–542, 2017.

[20] Weiyun Wang, Zhangwei Gao, et al., “Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency,” *arXiv:2508.18265*, 2025.

[21] Shuai Bai, Keqin Chen, et al., “Qwen2.5-vl technical report,” *arXiv:2502.13923*, 2025.

[22] Zhe Cao, Tomas Simon, et al., “Realtime multi-person 2d pose estimation using part affinity fields,” in *CVPR*, 2017, pp. 7291–7299.

[23] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, “Segan: Speech enhancement generative adversarial network,” 2017.

[24] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.