RatSeizure: A Benchmark and Saliency-Context Transformer for Rat Seizure Localization

Ting Yu Tsai[1], An Yu[1], Lucy Lee[1], Felix X.-F. Ye[1], Damian S. Shin[2], Tzu-Jen Kao[3], Xin Li[1], and Ming-Ching Chang[1⋆]

[1] University at Albany, State University of New York, Albany, NY, USA
{ttsai2,ayu,llee6,xye2,xli48,mchang2}@albany.edu
[2] University of Louisville, Louisville, KY, USA damian.shin@louisville.edu
[3] GE HealthCare Technology and Innovation Center, Niskayuna, NY, USA
kao@gehealthcare.com

Abstract. Animal models, particularly rats, play a critical role in seizure research for studying epileptogenesis and treatment response. However, progress is limited by the absence of datasets with precise temporal annotations and standardized evaluation protocols. Existing animal behavior datasets often provide limited accessibility, coarse labeling, and insufficient temporal localization of clinically meaningful events. To address these challenges, we introduce **RatSeizure**, the first publicly available benchmark designed for fine-grained seizure behavior analysis. The dataset is curated from long-form recordings and segmented into 10-second clips with dense annotations, including seizure-related action units and accurate temporal boundaries, enabling both behavior classification and temporal localization tasks. In addition, we propose **RaSeformer**, a saliency-context Transformer for temporal action localization that emphasizes behavior-relevant temporal context while suppressing redundant cues. Experimental results on RatSeizure demonstrate that RaSeformer achieves strong performance and provides a competitive reference model for this challenging setting. We further establish standardized dataset splits and evaluation protocols to promote reproducible benchmarking. Dataset is available at https://github.com/UA-CVML/RatSeizure.

Keywords: Temporal Action Localization · Seizure Detection· Animal Behavior Analysis · Transformer Models · Biomedical Video Analysis.

1 Introduction

Epilepsy is a prevalent neurological disorder affecting approximately $\sim 1\%$ of the global population [10]. Seizure episodes are characterized by stereotyped and stage-dependent behavioral manifestations, which provide critical signals for seizure classification, assessment of disease progression, and evaluation of therapeutic interventions [19,25]. In preclinical epilepsy research, rodent models, particularly rats, are extensively employed due to the controllability of seizure

⋆ Corresponding author

induction protocols and the ability to systematically observe and quantify behavioral phenotypes under standardized laboratory conditions.

Despite increasing interest in automated seizure behavior analysis, progress has been hindered by a fundamental infrastructure gap: the absence of publicly available benchmarks with dense temporal annotations and standardized evaluation protocols. Existing large-scale animal behavior datasets [5,7,11,18] primarily focus on generic behaviors and typically provide coarse video-level or weak temporal labels, limiting their utility for precise seizure event localization. Conversely, pathology-specific datasets [1,12] are often proprietary or lack sufficient temporal granularity, preventing reproducible benchmarking and fair comparison of algorithms. Consequently, there remains no dedicated, publicly accessible benchmark that supports fine-grained temporal localization of seizure-related behaviors, impeding methodological progress and translational impact in automated epilepsy research.

To address these limitations, we introduce **RatSeizure**, the first publicly available dataset designed for rat seizure behavior understanding. The dataset is curated from long-form recordings and segmented into 10-second clips with clinically action unit annotations and precise temporal boundaries, enabling unified evaluation of both behavior classification and temporal localization. We further establish standardized training and testing splits together with evaluation protocols. To characterize the challenges of this task, including fine-grained short-duration events and repetitive motion patterns, we benchmark representative temporal action localization methods and propose **RaSeformer**, a task-adapted reference model that achieves competitive performance against representative temporal action localization (TAL) baselines [8,14,23] on the RatSeizure benchmark. Our contributions are summarized below:

- **RatSeizure** is the first public video benchmark dataset for rat seizure behavior analysis with dense temporal annotations enabling clip-level classification and temporal localization evaluation.
- We define and annotate clinically motivated Action Units (AU) and introduce standardized data splits and evaluation protocols in RatSeizure to support reproducible comparison across methods.
- We introduce **RaSeformer**, a Transformer-based localization framework that explicitly models behavior-relevant temporal context and serves as a strong reference model on the benchmark.

2 Related Work

Animal behavior analysis datasets vary widely in annotation granularity, accessibility, and task focus. Many early clip-level datasets [1,6,7,11,17,21] are either limited in public availability or lack standardized benchmarks. Pose-centric resources [3,9,15,28] provide keypoint annotations but lack temporally localized behavioral events, while public video datasets such as CalMS21 [24], MARS [22] emphasize behavior classification without dense localization. AnimalKingdom [18] and MammalNet [5] introduce temporal annotations but focus

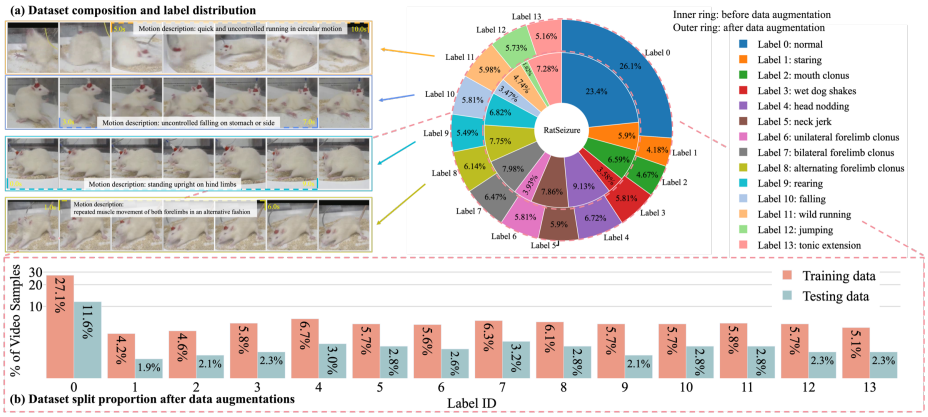

Fig. 1. RatSeizure dataset composition and split distribution: **(a)** Concentric donut chart showing behavioral class frequencies before (inner ring) and after (outer ring) data augmentation, with example frames from four action categories annotated with temporal boundaries. **(b)** Label distribution across training and test splits after augmentation, showing per-category relative frequencies (%). The $\sim 65:35$ ratio is consistent across all categories. Splitting is source-context constrained: augmented variants and their source clips are never separated across the train/test boundary.

on coarse multi-species behaviors, and clinically oriented datasets such as Pig Tail-biting [12] remain unavailable. RodEpil [20] focuses on binary seizure detection without timestamps or fine-grained action units, similar to VSViG [27], which adopts a skeleton-based representation centered on clip-level classification. To our knowledge, RatSeizure is the first publicly accessible benchmark with densely annotated seizure-specific behaviors and precise temporal boundaries, enabling systematic evaluation of fine-grained temporal localization. Table 1 summarizes representative datasets for animal behavior analysis.

Temporal action localization (TAL) has recently advanced through the adoption of end-to-end Transformer architectures that jointly predict action categories and temporal boundaries. TadTR [14] predicts action instances with query-based decoding, TE-TAD [8] adds time-aligned coordinate expressions and adaptive query selection, and TriDet [23] emphasizes boundary modeling and scalable granularity. More recent scaling efforts include AdaTAD [13] and efficiency-oriented improvements such as SNAG [16]. In contrast, RaSeformer enforces sparse temporal reasoning via local-window Top-K pruning, prioritizing behavior-relevant cues while preserving local temporal identity, which is well suited to fine-grained, short-duration events in repetitive seizure videos.

3 The RatSeizure Benchmark Dataset

The RatSeizure dataset (Fig. 1) comprise nine long-form rat seizure recordings, one per distinct rat. The recordings are segmented into 10-second clips, from

Table 1. Comparison of **RatSeizure** with representative animal behavior datasets. RatSeizure is the first publicly available dataset providing densely annotated, clip-based seizure behaviors with precise temporal boundaries and standardized evaluation.

Dataset	Public available?	Time tags?	# videos	Clinical?	Per video (h/m/s)	Species	Action localization?
Fish Action [21]	×	×	95	×	-	Fish	×
Dogs [1]	×	×	-	✓	-	Dog	×
Salmon Feeding [17]	×	×	76	×	∼83–833s	Salmon	×
Animal Pose [3]	✓	N/A	N/A	×	N/A	Multiple	N/A
Horse-30 [15]	✓	×	30	×	4–10s	Hourse	×
Pig Tail-biting [12]	×	✓	4,396	✓	1s	Pig	✓
Wildlife Action [11]	×	×	10,600	×	-	Multiple	×
Macaque Pose [9]	✓	N/A	N/A	×	N/A	Macaque	N/A
Broiler Chicken [6]	×	×	-	-	-	Chicken	×
Wild Felines [7]	×	×	2,700	×	<10s	Multiple	×
AP-10K [28]	✓	N/A	N/A	×	N/A	Multiple	N/A
CalMS21 [24]	✓	×	117	×	-	Mouse	✓
MARS [22]	✓	×	-	×	-	Mouse	✓
Animal Kingdom [18]	✓	✓	30,100	×	∼6s	Multiple	✓
MammalNet [5]	✓	✓	18,346	×	∼106s	Multiple	✓
RodEpil [20]	✓	×	13,053	✓	10s	Rodent	×
RatSeizure (Ours)	✓	✓	**794**	✓	**10s**	**Rat**	✓

which 794 clips with validated temporal boundaries of seizure-related behaviors are curated to support seizure action localization.

Experimental protocol: All animal use complied with the guidelines of the NIH and Albany Medical College (AMC) Institutional Animal Care and Use Committee (IACUC). Adult male Sprague-Dawley rats (Taconic, Germantown, NY) were used, and experiments were conducted during the light phase of the light–dark cycle (7am to 7pm). Female animals were excluded to avoid variability associated with estrous-cycle hormonal fluctuations that may affect neural excitability and seizure susceptibility. Seizures were pharmacologically induced via intraperitoneal pilocarpine (400 mg/kg) following scopolamine pretreatment (2 mg/kg, intraperitoneal) approximately 30 minutes prior. This controlled acquisition protocol supports reproducible benchmarking, though the use of adult male rats under a single experimental setting limits demographic and environmental generalizability.

Action labels: Seizure behaviors are described using 13 action units (AUs), together with a Normal category indicating absence of seizure activity; see Table 2. A seizure episode is defined as behavior persisting for at least 5 seconds. This 5-second threshold defines valid episodes for inclusion and clip extraction; individual AUs may be shorter and overlap temporally as multi-label segments.

Annotation and dataset curation: Nine recordings (each longer than 4 hours) were annotated with temporal AU labels. Valid seizure episodes guided clip extraction using 10-second windows with a 5-second stride. After verification, 601 clips were retained. Because substantial class imbalance was observed (for example, the Jumping AU appeared in only two recordings with 15 verified clips), lightweight **augmentation** including flipping and blurring was applied to reach a minimum target of 50 clips per AU, resulting in 794 total clips. Concur-

Table 2. Behavioral annotation covers 13 rat seizure Action Units (AUs) alongside a Normal category.

AU	Description	AU	Description
Normal	No seizure observed	Head Nodding	Repeated up/down "yes" head motion
Staring	Behavioral arrest without movement	Neck Jerk	Repeated, intense/quick "yes" motion
Mouth Clonus	Mouth and jaw twitching	Unilateral Forelimb Clonus	Repetitive forelimb movement on one side
Wet-Dog Shake	Whole body shaking	Bilateral Forelimb Clonus	Repetitive movement of both forelimbs
Rearing	Standing upright on hind limbs	Alternating Forelimb Clonus	Alternating forelimb movements
Jumping	Sudden upward jump	Falling	Loss of posture with uncontrolled fall
Tonic Extension	Limbs rigidly extended while lying	Wild Running	Rapid uncontrolled circular running

rent AUs were annotated as multi-label segments and evaluated independently. To avoid data leakage, the dataset was split before augmentation at the source-context level: clips from the same source recording, seizure episode, or overlapping temporal context were assigned to the same partition, and augmented variants inherited that partition. Thus, no original clip, augmented variant, overlapping adjacent clip, or source-related clip crosses the train/test boundary. The dataset was split into training (513 clips) and test (281 clips) sets. Per-class distributions across splits are shown in Fig. 1b.

Multi-stage expert review and verification: All long-form recordings were annotated using a sequential multi-expert protocol to ensure annotation stability. One expert created initial temporal segments, after which three additional experts independently verified and refined labels and temporal boundaries to reach consensus. Intermediate annotation snapshots were not retained for computing inter-rater agreement; instead, clip-level audit statistics are reported. Following clip extraction, Expert 1 revised approximately 20% of clips (labels and/or boundaries), while Expert 2 introduced additional corrections to 5%. The reduction in corrections suggests that most discrepancies were resolved in the first review stage, with only minor refinements required thereafter.

4 RaSeformer: Processing and Architecture

RaSeformer provides a strong reference model for temporal action localization in the RatSeizure benchmark, converting long-form video streams into temporally localized predictions through the three-phase architecture shown in Fig. 2.

Phase 1: Preprocessing and data curation: Raw rodent monitoring videos often contain spatial and temporal redundancy, along with irrelevant background regions. To focus on behavior-relevant areas and reduce noise, we apply a three-step preprocessing and curation procedure: (a) Temporal segmentation: Continuous recordings are split into overlapping 10-second clips using a 5-second stride. At 30 FPS, each clip contains 300 frames, providing consistent temporal coverage while enabling efficient downstream temporal localization. (b) Masking out mirror artifacts: To reduce false motion cues caused by reflections in glass and mirrors of the animal enclosures, we apply static binary masks to rule out

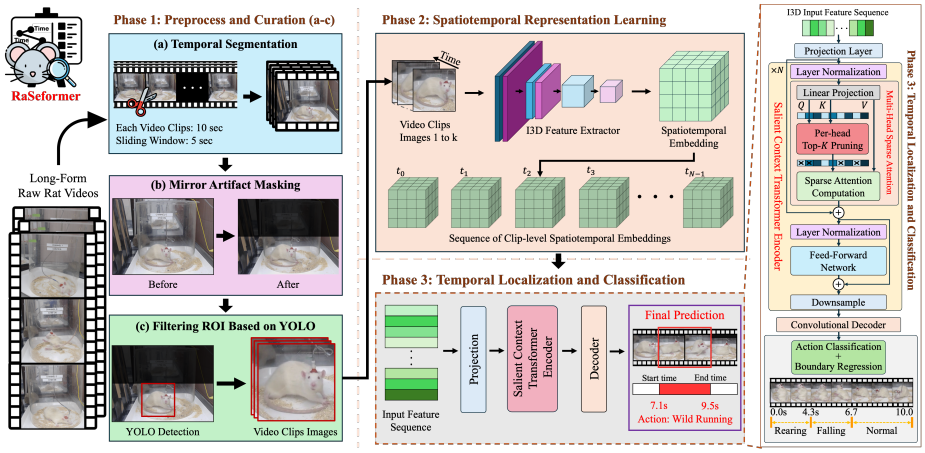

Fig. 2. The RaSeformer Pipeline: In phase 1 (a-c), raw rat videos are first processed through temporal segmentation, mirror artifact masking, and YOLO-based ROI cropping to produce standardized input tensors. These tensors are encoded by an I3D backbone into spatiotemporal feature sequences (phase 2). In the final phase, the sequences pass through the Salient Context Transformer Encoder, which uses per-head Top-K pruning to focus on salient behavioral cues. A decoder then fuses the encoded features, with parallel heads for temporal boundary regression and behavior classification.

reflection regions, eliminating spurious detections. (c) YOLO ROI filtering: A fine-tuned YOLOv12 [26] detector localizes the rat in each clip. To preserve contextual behavioral clues, each predicted bounding box is enlarged by 50 pixels on all sides, cropped into a square ROI and resized to 224×224. This removes background clutter, maintains spatial consistency across clips, and preserves fine-grained motion details for downstream tasks.

Phase 2: Spatiotemporal representation learning: Curated RGB clips are transformed into compact spatiotemporal embeddings using a pretrained Inflated 3D ConvNet (I3D) backbone [4]. Each 300-frame, 224×224 clip is processed to extract hierarchical features that capture both spatial posture and short-term temporal dynamics. Features are computed using a chunk size of 16 frames with a temporal stride of 4, producing sequences of approximately 71-72 embeddings per 10-second clip. These sequences serve as the input representation for downstream temporal localization and classification.

Phase 3: Temporal localization and classification: Spatiotemporal feature sequences are transformed into temporally localized behavior predictions (Fig. 2). This stage maps high-level behavioral embeddings to discrete action segments with precise temporal boundaries.

The sequence of I3D feature embeddings is first passed through a convolutional projection layer, which maps the features to internal embedding dimensions of the detection backbone. This standardizes the feature space and prepares the sequence for transformer-based temporal reasoning.

Table 3. Comparison of representative temporal action localization (TAL) baselines on RatSeizure. mAP is reported on the full test set and on non-augmented and augmented subsets separately. Parentheses indicate RaSeformer's average absolute mAP gain over all baselines.

Method	Full test mAP	Non-aug. video mAP	Aug. video mAP
ActionFormer [29]	54.51	55.13	40.60
TriDet [23]	55.23	55.65	42.02
TE-TAD [8]	45.70	50.27	34.30
TadTR [14]	42.75	43.82	31.14
RaSeformer (ours)	**57.08 (+7.53 avg.)**	**57.83 (+6.61 avg.)**	**46.67 (+9.65 avg.)**

In RaSeformer, the temporal modeling is the **Salient Context Transformer Encoder**, a multi-scale architecture comprising N stacked blocks at each level of a temporal feature pyramid. Each block uses pre-layer normalization to maintain training stability while capturing behavioral context across multiple temporal resolutions.

Within each block, features are processed by Multi-Head Sparse Attention. Queries Q, Keys K, and Values V are generated via depthwise convolutions followed by linear projections. For each query, a local temporal window size of 9 defines the candidate context. We implement a per-head Top-K pruning strategy based on Query-Key similarity: each attention head independently computes saliency scores and retains its own set of the most salient M neighboring tokens within the local window. This allows each head to specialize on distinct temporal regions and behavioral cues, while the center query token is always preserved to maintain local temporal identity. Sparse attention is computed over this reduced token set, substantially reduce the quadratic complexity of self-attention. The output is fused via a residual connection, followed by Layer Normalization and a feed-forward network, with a second residual connection applied to preserve information flow. At the end of each pyramid level, temporal downsampling propagates salient representations to the next pyramid level.

Convolutional Decoder and Final Prediction: The multi-scale encoded features are decoded using convolutional heads to produce final predictions. An action classification head outputs class probabilities for each candidate segment, while a boundary regression head predicts left and right temporal offsets. These offsets are converted into absolute start and end timestamps via a learnable scaling factor, yielding temporally localized behavior detections.

5 Experimental Validation and Results

All experiments were conducted on a single NVIDIA A100 GPU (80GB).

YOLO fine-tuning: We manually annotated 2,652 frames randomly sampled from the nine long-form videos to fine-tune a YOLOv12 [26] model using 2,113 training images and 539 test images. The resulting detector achieves an mAP@0.50 of 0.995 for rat detection.

Table 4. Ablation study of RaSeformer: Effects of saliency scoring and head specialization are evaluated using mAP on the full test set and two subsets.

Method	Full test mAP	Non-aug. video mAP	Aug. video mAP
ActionFormer [29]	54.51	55.13	40.60
w/ Static Key-Norm pruning	56.66	56.76	42.46
w/ Head-shared Top-K	56.80	56.94	43.29
RaSeformer (ours)	**57.08 (+1.09 avg.)**	**57.83 (+1.55 avg.)**	**46.67 (+4.55 avg.)**

Baseline and Comparison: On RatSeizure, we compare RaSeformer with representative temporal action localization (TAL) baselines, including TriDet [23], ActionFormer [29], TE-TAD [8], and TadTR [14]. Evaluation is performed on the test set using mAP averaged over temporal IoU thresholds 0.3-0.7, with one-to-one matching between predictions and ground truth at each threshold. During inference, multi-class Soft-NMS [2] (IoU threshold 0.1, σ=0.5) is applied, retaining up to 200 segments per video and filter predictions with scores below 0.001. Our encoder uses 4 attention heads and per-head Top-K sparse attention (keep ratio 0.5) and a local window sizes of W=9 across a six-level temporal pyramid, keeping the top $M=\lceil 0.5W \rceil$=5 tokens per head within each window.

Table 3 shows evaluation results where RaSeformer consistently outperforms all compared methods on the full test set as well as the non-augmented and augmented subsets. We report these subsets separately to distinguish in-distribution generalization from robustness under augmentation-induced distribution shift. On the full test set, RaSeformer achieves 57.08 mAP, corresponding to an average gain of +7.53 mAP over the compared baselines. This improvement remains on the non-augmented subset (57.83 mAP; +6.61 mAP on average), demonstrating strong in-distribution generalization. Although performance decreases for all methods on the augmented subset, RaSeformer remains the most robust (46.67 mAP; +9.65 mAP on average), as appearance perturbations such as blur and low lighting degrade feature quality and make precise temporal boundary localization more challenging.

Ablation analysis: Table 4 presents ablations on the full test set as well as the non-augmented and augmented subsets. RaSeformer consistently outperforms ActionFormer [29] and all ablated variants across all three evaluations. Replacing Query-Key similarity with Static Key-Norm pruning reduces mAP on every subset, highlighting the importance of query-dependent saliency for selecting relevant temporal context. Likewise, using head-shared Top-K selection instead of per-head Top-K consistently lowers performance, indicating that head specialization allows attention heads to capture complementary temporal cues. The largest improvement appears on the augmented subset (+4.55 mAP on average), suggesting that combining Query-Key pruning and head-specific sparsity improves robustness to appearance and motion perturbations introduced by augmentation. Together, these results confirm that both saliency modeling and head specialization are essential to RaSeformer's performance.

6 Conclusion

We introduce **RatSeizure**, a publicly available dataset with dense frame-level temporal annotations of rat seizure behaviors, designed to support reproducible benchmarking for behavior classification and fine-grained temporal localization. Along with the dataset, we provide standardized evaluation protocols and a strong reference model, **RaSeformer**, establishing competitive baselines for this challenging task. Together, these resources create a unified benchmark that enables fair comparison, promotes methodological development, and lowers the barrier to automated seizure behavior analysis.

Limitation of current work includes moderate dataset scale due to high cost and logistical complexity of collecting high quality rat seizure recordings with expert annotations. **Future work** will expand the dataset with additional animal types, seizure types, and recording conditions to improve diversity, scale, and translational relevance. Planned benchmark extensions include blinded re-annotation audits via temporal IoU analysis, broader comparisons with recent TAL methods, and additional metrics covering localization quality, class imbalance, and computational efficiency.

Acknowledgments. Felix X.-F. Ye is grateful for partial support from seed funding by the Center for Emerging Artificial Intelligence Systems at the University at Albany. This research was supported in part by the AI Computing Cluster at the UAlbany.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Barnard, S., Calderara, S., Pistocchi, S., Cucchiara, R., Podaliri-Vulpiani, M., Messori, S., Ferri, N.: Quick, accurate, smart: 3D computer vision technology helps assessing confined animals' behaviour. PloS ONE **11**(7), e0158748 (2016)
2. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-NMS — Improving Object Detection with One Line of Code. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 5562–5570 (2017)
3. Cao, J., Tang, H., Fang, H.S., Shen, X., Tai, Y.W., Lu, C.: Cross-domain adaptation for animal pose estimation. In: ICCV. pp. 9497–9506 (2019)
4. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: CVPR. pp. 4724–4733 (2017)
5. Chen, J., Hu, M., Coker, D.J., Berumen, M.L., Costelloe, B., Beery, S., Rohrbach, A., Elhoseiny, M.: Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In: CVPR. pp. 13052–13061 (2023)
6. Fang, C., Zhang, T., Zheng, H., Huang, J., Cuan, K.: Pose estimation and behavior classification of broiler chickens based on deep neural networks. Computers and Electronics in Agriculture **180**, 105863 (2021)
7. Feng, L., Zhao, Y., Sun, Y., Zhao, W., Tang, J.: Action recognition using a spatial-temporal network for wild felines. Animals **11**(2) (2021)
8. Kim, H.J., Hong, J.H., Kong, H., Lee, S.W.: Te-tad: Towards full end-to-end temporal action detection via time-aligned coordinate expression. In: CVPR. pp. 18837–18846 (2024)

9. Labuguen, R., Matsumoto, J., Negrete, S., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K.i., Shibata, T.: Macaquepose: A novel 'in the wild' macaque monkey pose dataset for markerless motion capture. bioRxiv (2020)
10. Leonardi, M., Ustun, T.B.: The global burden of epilepsy. Epilepsia **43**, 21–25 (2002)
11. Li, W., Swetha, S., Shah, M.: Wildlife action recognition using deep learning. In: TechRxiv (2020)
12. Liu, D., Oczak, M., Maschat, K., Baumgartner, J., Pletzer, B., He, D., Norton, T.: A computer vision-based method for spatial-temporal action recognition of tail-biting behaviour in group-housed pigs. Biosystems Engineering **195**, 27–41 (2020)
13. Liu, S., Zhang, C.L., Zhao, C., Ghanem, B.: End-to-end temporal action detection with 1b parameters across 1000 frames. In: CVPR. pp. 18591–18601 (2024)
14. Liu, X., Wang, Q., Hu, Y., Tang, X., Zhang, S., Bai, S., Bai, X.: End-to-end temporal action detection with transformer. TIP **31**, 5427–5441 (2022)
15. Mathis, A., Schneider, S., Yüksekgönül, M., Yuksekgonul, M., Bethge, M., Mathis, M.W.: Pretraining boosts out-of-domain robustness for pose estimation. WACV pp. 1858–1867 (2019)
16. Mu, F., Mo, S., Li, Y.: SnAG: Scalable and accurate video grounding. In: CVPR. pp. 18930–18940 (2024)
17. Måløy, H., Aamodt, A., Misimi, E.: A spatio-temporal recurrent network for salmon feeding action recognition from underwater videos in aquaculture. Computers and Electronics in Agriculture **167**, 105087 (2019)
18. Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: CVPR. pp. 19001–19012 (2022)
19. Ngugi, A.K., Bottomley, C., Kleinschmidt, I., Sander, J.W., Newton, C.R.: Estimation of the burden of active and life-time epilepsy: a meta-analytic approach. Epilepsia **51**(5), 883–890 (2010)
20. Perlo, D., Despotovic, V., Boudissa, S., Kim, S.Y., Nazarov, P.V., Zhang, Y., Wintermark, M., Keunen, O.: Rodepil: A video dataset of laboratory rodents for seizure detection and benchmark evaluation. arXiv preprint arXiv:2511.10431 (2025)
21. Rahman, S.A., Song, I., Leung, M., Lee, I., Lee, K.: Fast action recognition using negative space features. Expert Systems with Applications **41**(2), 574–587 (2014)
22. Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J.J., Perona, P., Anderson, D.J., Kennedy, A.: The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice. Elife **10**, e63720 (2021)
23. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: TriDet: Temporal action detection with relative boundary modeling. In: CVPR. pp. 18857–18866 (2023)
24. Sun, J.J., Karigo, T., Chakraborty, D., Mohanty, S.P., Wild, B., Sun, Q., Chen, C., Anderson, D.J., Perona, P., Yue, Y., et al.: The multi-agent behavior dataset: Mouse dyadic social interactions. NeurIPS **2021**(DB1), 1 (2021)
25. Thurman, D.J., Beghi, E., Begley, C.E., Berg, A.T., Buchhalter, J.R., Ding, D., Hesdorffer, D.C., Hauser, W.A., Kazis, L., Kobau, R., et al.: Standards for epidemiologic studies and surveillance of epilepsy. Epilepsia **52**, 2–26 (2011)
26. Tian, Y., Ye, Q., Doermann, D.: YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv preprint arXiv:2502.12524 (2025)
27. Xu, Y., Wang, J., Chen, Y.H., Yang, J., Ming, W., Wang, S., Sawan, M.: VSViG: Real-Time Video-Based Seizure Detection via Skeleton-Based Spatiotemporal ViG. In: ECCV. pp. 228–245 (2025)

28. Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: AP-10K: A benchmark for animal pose estimation in the wild. In: NeurIPS Datasets and Benchmarks Track (2021)
29. Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: ECCV. pp. 492–510. Springer (2022)