MooCap: A Multi-View Benchmark for Cow-Object-Human Interaction and Behavior Dynamics

Ian Noronha[1] Heather Neave[2,3] Upinder Kaur[1,*]
[1]Agricultural & Biological Engineering, [2]Animal Science, Purdue University
[3]Animal and Veterinary Sciences, Aarhus University, Denmark
[*]Corresponding author. {inoronha, hneave, kauru}@purdue.edu

Abstract

Understanding animal behavior requires modeling how bodies, objects, and other agents interact over time, not simply detecting isolated actions or estimating pose frame by frame. Existing animal video datasets target pose estimation or coarse, passively observed actions, and rarely provide the structured, multi-entity interaction annotations needed to study behavioral dynamics. We introduce MooCap, a multi-view video benchmark for animal-object-human interaction understanding under controlled experimental protocols. MooCap contains 42 hours of synchronized multi-camera video from 43 individually tested cows across seven standardized interaction scenarios, including novel environment, novel object, novel human, human approach, unfamiliar conspecifics (restricted and unrestricted) and maternal reunion (restricted and unrestricted). Recordings are densely annotated with 23 fine-grained behaviors, 39 body keypoints across 157 test sessions, 4 spatial zones, and 43 subjects, describing interactions among subjects, objects, humans, and other cattle. We establish three benchmarks on MooCap: (1) dense temporal action segmentation over 1200-1500-second sequences; (2) pose-based behavior and interaction recognition from keypoint trajectories; and (3) longitudinal behavioral classification linking adult behaviors with rearing conditions. Benchmarking results reveal that state-of-the-art temporal segmentation models achieve only 66.4% frame accuracy and 30.6% F1@0.5, with performance degrading further during interaction-heavy segments. Overall, MooCap bridges multi-view pose estimation, multi-entity tracking, and structured behavioral protocols to enable interaction-aware models for animal behavior analysis. Dataset available at https://github.com/IannoIITR/MooCap

1. Introduction

Understanding animal behavior is fundamental to advancing welfare science, conservation biology, and precision livestock management. Animals reveal critical information about their cognitive states, emotional well-being, and social dynamics through how they respond to environmental changes, novel stimuli, and conspecific encounters. Automated behavior recognition could enable real-time welfare monitoring and accelerate behavioral research, but developing such systems requires datasets that capture not just isolated actions but the contextual patterns of behavior across interaction paradigms.

Modern video sensors deployed across farms and research facilities now generate behavioral footage at unprecedented scales [10, 40]. However, processing this data to extract actionable insights remains a fundamental challenge. Manual annotation of behavioral sequences is labor intensive and does not scale to large datasets. Thus, automated animal behavior recognition systems are critical for capitalizing on these large unlabeled datasets and accelerating monitoring efforts at scale.

The first essential step toward building such systems is curating diverse, representative datasets that formalize these challenges as computer vision tasks. Recent large scale benchmarks such as Animal Kingdom [31] and MammalNet [10] have made important progress, providing thousands of hours of annotated footage across hundreds of species. However, these datasets share critical limitations: First they focus on coarse level behaviors (walking, standing) rather than fine scale actions (ruminating, threat displays), which indicate welfare and health status. Second, they rely almost entirely on passively collected footage of spontaneous behavior, systematically underrepresenting responses to novel or challenging situations. Third, they lack the controlled experimental structure necessary for semantic behavior profiling, making it difficult to disentangle individual differences from situational factors or to study consistent behavioral patterns across contexts [5].

To address these limitations, we propose MooCap, a multi camera video benchmark for detailed bovine behavior analysis. It uses structured experimental protocols and dense multimodal annotations. Unlike passively collected

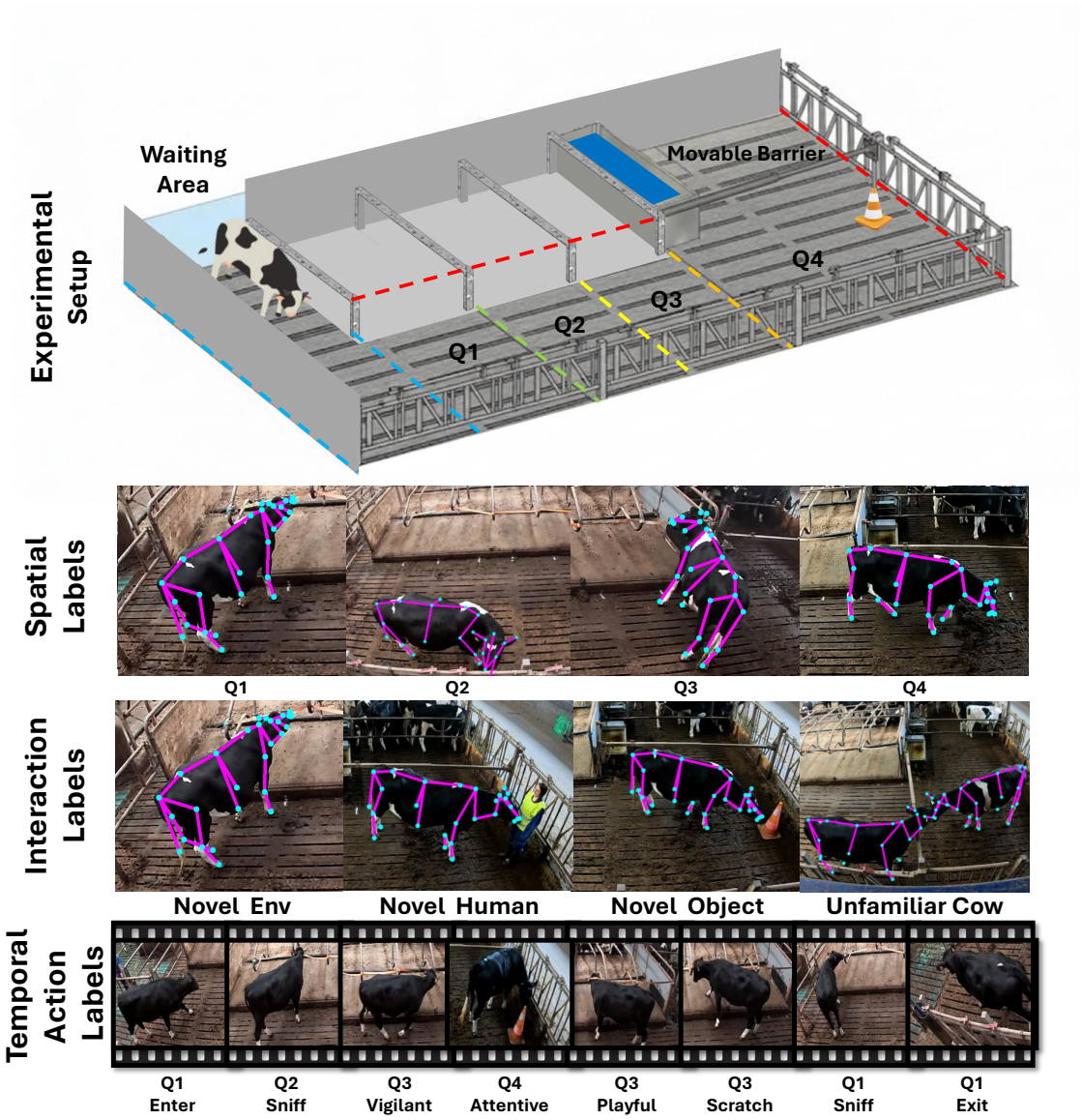

Figure 1. MooCap overview. The top graphics shows the layout of the data collection setup. The spatial, interaction, and temporal action label annotations compile a comprehensive set for behavior dynamics studies.

datasets, MooCap is built around ethologically validated behavioral assays in which each subject undergoes a standardized series of controlled stimuli: exposure to novel environments, unfamiliar objects, human approach, and social interactions with unfamiliar conspecifics or biological mothers. This experimental design systematically probes key dimensions of behavior, including exploratory motivation, neophobia, social competence, and the quality of human animal relationship. These dimensions are critical for welfare assessment yet are rarely visible in unstructured footage. The dataset comprises 42 hours of synchronized multi-view video from 43 dairy cattle subjects, each completing five to seven behavioral tests. Importantly, these animals represent three distinct early-life rearing treatments, enabling investigation of whether developmental history can be inferred from adult behavioral patterns.

MooCap provides three levels of annotation (Figure 1). First, frame-level action labels for 23 discrete behaviors spanning exploratory actions, attentional states, social interactions, and spatial preferences. Second, skeletal pose annotations with 39 anatomical keypoints per animal for joint modeling of kinematic and semantic patterns. Third, we provide longitudinal treatment labels that turn the dataset into a testbed for phenotypic inference. Models must learn

to predict latent traits from observable behavior, which is directly relevant for automated welfare monitoring.

We establish three core benchmarks on MooCap: temporal action segmentation, pose-based behavior recognition, and longitudinal phenotype classification. Our results demonstrate substantial room for improvement as state-of-the-art temporal segmentation models achieve only 66.39% frame accuracy, while skeleton-based recognition reaches just 0.39 mean F1 on action classification. These gaps highlight fundamental challenges in animal behavior understanding and provide opportunities for future research.

2. Related Work

2.1. From Controlled to In-the-Wild Benchmarks

The trajectory of animal behavior analysis in computer vision closely mirrors the evolution of Human Action Recognition (HAR).Early HAR work relied on small scale datasets such as *KTH* [34] and *HMDB51* [16] which contain simple actions in controlled settings. Over time the field moved toward large scale, "in-the-wild" benchmarks such as *ActivityNet* [8] and *Kinetics* [9], which capture the complexity and diversity of real-world human activities.

Animal behavior datasets followed a similar path. Initial works focused on single species in specific contexts, such as the *Cattle Visual Behaviours (CVB) dataset* [45] for livestock monitoring. This paradigm shifted with the introduction of massive, multi-species, "in-the-wild" benchmarks. *Animal Kingdom* [31] provided a large-scale dataset with 850 species, supporting tasks from action recognition to temporal grounding. *MammalNet* [10] further advanced this by providing 539 hours of taxonomy-guided video across 173 mammal categories. Concurrently, datasets like *IWildCam* [3] leveraged camera trap footage, establishing passive, large-scale observation as the dominant data collection methodology. **Table 1** further contextualizes MooCap relative to controlled single-species and large-scale passive observation datasets, demonstrating that MooCap uniquely combines dense pose and action annotations with longitudinal welfare labels enabling causality analysis.

2.2. Semantic Hierarchies & Fine-Grained Actions

A central challenge in action recognition is the semantic hierarchy of labels. In HAR, there is a distinction between atomic actions (e.g., *bend, lift*) and complex, temporally extended activities (e.g., *gardening*) [17, 36]. This semantic gap is even more pronounced in ethology. *MammalNet* [10] notes that simple atomic labels such as *running* are ambiguous and fail to capture high-level behavioral state (e.g., *hunting, escaping, or playing*). While most benchmarks focus on coarse-grained (CG) actions [46], the most critical welfare and health indicators lie in fine-grained (FG) behaviors [46]. However, capturing subtle FG actions is difficult due to occlusions, non-rigid bodies, and high intra-species variance [46], resulting in a scarcity of high-quality annotations for these crucial behaviors.

Table 1. Animal Behavior and Pose Estimation Benchmarks

Dataset	Dom.	Dur.	Subj.	Annotations
MooCap (Ours)	Farm	42h	43	Long.* Act, Pose, Spatial, ID
AcinoSet [14]	Wild	119kf	10	3D Pose, Kinematics
ChimpACT [26]	Zoo	16kf	23	Long. Act, Pose, ID
MARS [37]	Lab	14h	N/A	Action, Pose
Horse-30 [28]	Farm	8kf	30	Long. Pose
Stanford Dogs [15]	Multi	20kf	120	Pose, ID
AP-10K [44]	Wild	10kf	54s	Pose
MammalNet [10]	Wild	539h	N/A	Action, Detection
MBE-ARI [32]	Farm	12h	3	Long. Act, Pose, Int.
Animal Kingdom [31]	Wild	50h	N/A	Action, Grounding
LoTE-Animal [21]	Wild	35kf	11s	Long. Act, BBox

Long. Longitudinal study annotations.
*__Longitudinal Labels:__ Early-life dam contact history (full/half/none) is provided for causality analysis of behavior recorded 9 months post-treatment.

2.3. The Observational Bottleneck

While "in-the-wild" datasets [10, 31] provide ecological validity, they suffer from a fundamental observational bottleneck. This is a well-documented issue in computer vision, where passive data collection introduces significant dataset bias [38]. In animal benchmarks, this manifests as an over-representation of "attention-grabbing" behaviors (e.g., *fighting*) and a lack of data on crucial welfare indicators that occur less frequently or less dramatically [10].

Crucially, this passive paradigm restricts models to descriptive recognition (What is the animal doing?) without the systematic context needed for behavioral profiling (How does this individual respond to stimuli?). Current state-of-the-art research is thus split between two extremes: (1) large-scale, passive, "in-the-wild" datasets [10, 31] that lack context and control, and (2) small-scale, hypothesis-driven laboratory studies that use powerful pose-tracking tools [27, 33] but lack scalability and species diversity.

MooCap bridges this gap by embedding controlled experimental protocols within a scalable video framework. Rather than passively observing spontaneous behavior, we employ structured ethological assays, standardized stimuli designed to elicit specific, interpretable behavioral responses. This approach enables models to learn behavioral profiles under standardized conditions, supporting tasks such as phenotype classification and systematic behavioral characterization that require consistency across repeated testing protocols.

3. The MooCap Dataset

The MooCap dataset represents a departure from conventional animal video benchmarks by embedding structured

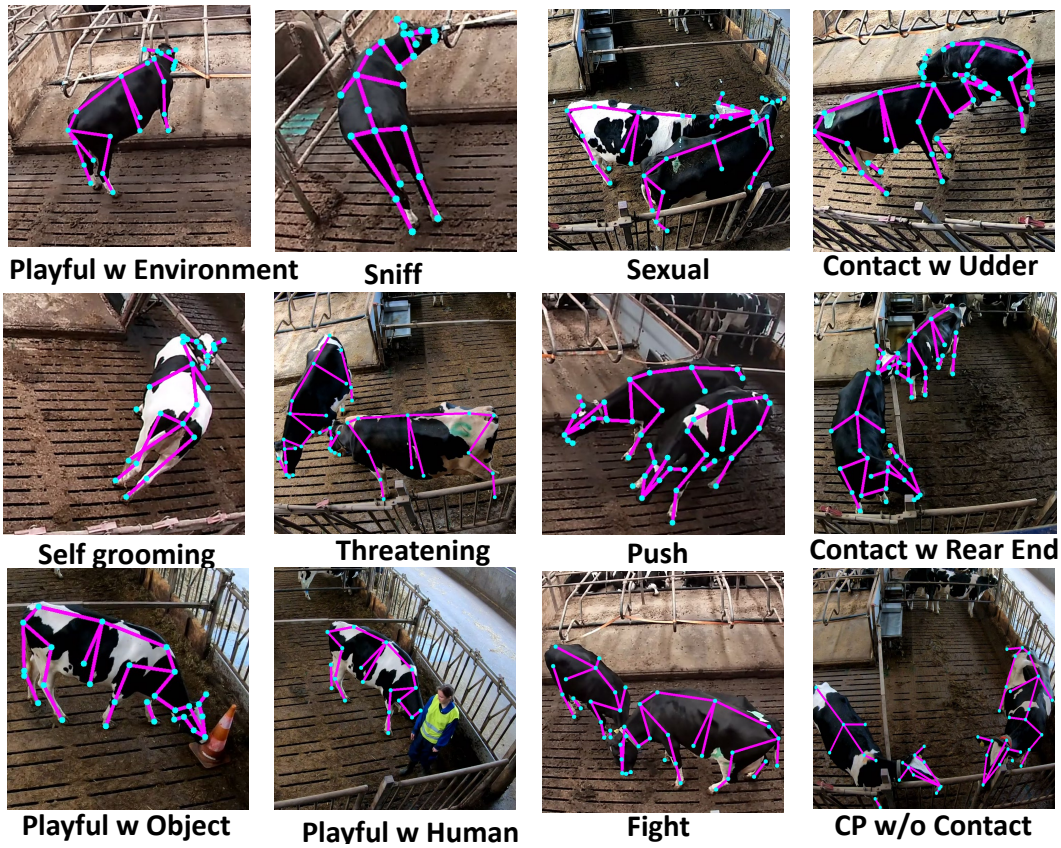

Figure 2. Examples of annotated behaviors across our seven experimental scenarios. Our dataset captures 23 distinct action categories, including social interactions (Contact with rear-end, Fight, Grooming), proximity behaviors (Close proximity, Touching), attention states (Attentive, Vigilant), and environmental responses (Sniff, Playful, Push). Videos include tests for: dam reunion, novel human, novel environment, novel object, and unfamiliar cow tests.

Table 2. Experiment Summary Statistics

Experiment	Avg Test Time (s)	# Tests	# Subj.	Total (s)
Novel Environment	180	32	28	5760
Novel Object	179	35	28	6299
Human Approach	180	28	28	5040
Unfamiliar Conspecific: Restricted	295	24	24	7080
Unfamiliar Conspecific: Unrestricted	300	24	24	7200
Dam: Restricted	300	7	7	2100
Dam: Unrestricted	300	7	7	2170

experimental design within a comprehensive video annotation framework. Rather than passively recording spontaneous behavior, we construct a controlled protocol in which each subject undergoes a standardized series of ethologically validated behavioral assays. This approach bridges the gap between small-scale laboratory pose-tracking studies [27, 33] and large-scale "in-the-wild" datasets [10, 31], enabling scalable behavioral interpretation.

3.1. Animal Taxonomy and Experimental Design

Our experimental paradigm is grounded in classical ethology, where behavioral responses to controlled stimuli reveal underlying traits and affective states. Each cow in our dataset is exposed to five core scenarios: introduction to a novel environment, presentation of a novel object, encounter with an unfamiliar human, approach by that human, and social interaction with an unfamiliar conspecific (Table 2). For a subset of subjects whose maternal dams remained in the herd, we added two additional scenarios: reunion with the biological mother with and without restriction. These scenarios are designed to probe specific dimensions of behavior that reveal how individuals respond to potentially stressful situations, and thus their temperament, affective state and welfare. Exposure to a novel environment typically assesses exploratory motivation or fear of isolation; novel object presentation evaluates neophilia versus neophobia [41]; human approach tests, where the animal voluntarily approaches a human or the animal is involuntar-

ily approached by the human, measure the human-animal relationship, a critical determinant of welfare in livestock settings [13, 39]; and social encounters, such as our two-stage protocol separating visual assessment from physical contact, reveal dominance hierarchies, affiliative bonding, and conspecific recognition capabilities.

Data was collected at the Danish Cattle Research Centre, Aarhus University, Tjele, Denmark. The dataset comprises 43 dairy heifers (approximately 19 months old at time of recording), with each subject completing the full five-scenario protocol for a total of approximately 20-25 minutes of video per animal. Critically, these animals were not randomly selected but instead represent three distinct longitudinal treatment cohorts, established during the calf rearing period as described in [30]. From 48 hours to 10 weeks of age, calves were assigned to one of three rearing conditions: full-time contact with their biological dam (23 hours per day), half-time contact (10 hours per day), or immediate separation at birth (control group). Following this 10-week treatment period, all calves were weaned and raised under identical standard management practices. The behavioral recordings used in MooCap were collected nine months post-weaning, creating a substantial temporal gap between treatment exposure and behavioral assessment. This longitudinal structure transforms the dataset from a simple action recognition benchmark into a unique testbed for inferring latent phenotypic traits from observable behavior, which remains a key challenge in behavioral genomics and welfare science. The experiment and data set details are summarized in Table 3.

3.2. Video Acquisition and Arena Setup

Beyond stimulus design, our data collection infrastructure ensures spatiotemporal precision and multi-view coverage. All tests were conducted in a single, standardized test pen to guarantee environmental consistency across subjects (Figure 1). Recording sessions occurred between 11:00 AM and 2:00 PM to minimize variation in circadian activity patterns. We deployed multiple synchronized GoPro cameras positioned on an elevated viewing platform, providing complementary perspectives that resolve occlusions and enable robust 3D pose reconstruction. The test arena measured 60 m² and was marked with one-meter grid intervals on the flooring, facilitating quantitative spatial analysis and trajectory tracking. Static environmental features, including a feed bunk and water trough, were maintained across all sessions, providing consistent reference points for calibration and spatial localization. Subjects were moved individually into the test arena, from a designated waiting area used between sessions to allow experimenters to reconfigure the pen for subsequent scenarios. Subjects always experienced the scenarios in the same consecutive order for the same amount of time (Table 2): (1) novel environment (3 min); (2) novel object (3 min); (3) human approach (3 min) followed by human approaching cow until retreat (approx. 1 min); (4) unfamiliar cow with restricted, visual only access (5 min); (5) followed by unrestricted, physical access (5 min). For those subjects with their maternal dam remaining in the herd, the restricted and unrestricted tests were repeated in a test of dam reunion (5 min each). This infrastructure supports not only video-level action recognition but also fine-grained pose estimation, spatial occupancy mapping, and multi-agent interaction modeling.

Table 3. MooCap Dataset Summary

Metric	**Value**
Total Subjects	43
Total Scenarios	5 per subject + 2 (cows with Dam)
Total Video Duration	42 hours
Avg. Video per Subject	22.5 minutes
Annotation Types	
Longitudinal Labels	3 rearing groups
Temporal Action Segments	23 behaviors
Spatial Segments	4 quadrants
Pose Keypoints	39 joints

3.3. Data Annotation and Quality Control

Our annotation protocol extends far beyond coarse-grained action labels. Using the BORIS behavioral observation software [12], we generated dense, per-frame annotations synchronized across multiple camera views with precise temporal alignment. To ensure observer annotation reliability, we implemented an inter-rater agreement protocol requiring Cohen's kappa values exceeding 0.8 for all behavioral categories. The resulting ethogram includes 23 discrete behaviors spanning multiple functional domains (examples shown in Figure 2). Exploratory behaviors capture investigative actions such as sniffing, licking, rubbing, and self-grooming. Spatial annotations divide the arena into four quadrants, providing ground-truth occupancy data that supports spatial preference analysis and approach-avoidance metrics. Attentional states, including vigilance, object-directed attention, and human-directed attention, enable modeling of perceptual focus and threat assessment.

Interactive behaviors encompass physical contact with objects or humans, as well as playful manipulation. Social behaviors are subdivided into affiliative categories (touching, allogrooming, head play) and agonistic categories (threats, pushing, fighting), as shown in Figure 2, allowing nuanced analysis of conspecific dynamics. This hierarchical structure mirrors the compositional action modeling in human activity benchmarks [17], where atomic actions compose into complex behavioral sequences. The frequency and durations of the major behaviors observed are

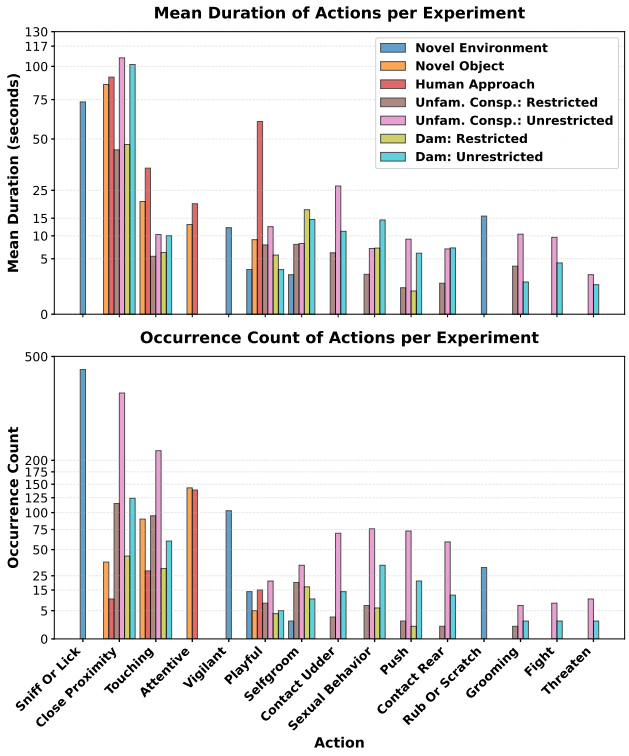

Figure 3. MooCap action statistics. Frequency and mean duration for the 15 most frequent actions.

summarized in Figure 3.

Complementing the temporal action annotations, we provide dense body pose data derived from manual keypoint labeling. We annotated approximately 3,000 frames distributed across the dataset, marking 39 anatomical landmarks per cow based on established protocols in animal pose estimation [18, 27, 29, 32]. These keypoints capture head orientation, limb positioning, tail carriage, and postural configuration, features that convey affective state, physical comfort, and motor intention. The integration of pose and action annotations enables joint modeling of kinematic signatures and semantic behavior labels, supporting tasks ranging from pose-conditioned action recognition to affective state inference from body posture.

4. Experiments and Results

The structured design and multi-modal annotations of MooCap enable a diverse suite of computational tasks that extend beyond conventional action recognition. We establish three core benchmarks that probe different aspects of behavioral understanding, from fine-grained temporal segmentation and pose-based recognition to long-term phenotype inference. These tasks address the dimensions of pose tracking, behavior mapping, and health and welfare, thereby establishing the unique contributions of this dataset in each. The training details for the models in the three benchmarks are described in the Supplementary A1.

4.1. Benchmark 1: Temporal Action Segmentation

Temporal action segmentation involves producing dense, per-frame action labels for untrimmed video, and remains one of the most challenging problems in video understanding. MooCap presents this task in a domain characterized by extreme temporal extent (20-25 minute sequences), high action diversity (23 classes), and severe class imbalance. Unlike human activity benchmarks where actions are often spatially localized and temporally sparse [17], animal behaviors frequently involve subtle, non-rigid motions distributed across the entire body with long, ambiguous transition regions between discrete states.

We evaluated three state-of-the-art supervised segmentation architectures on this task (Table 4). FACT [25], which employs hierarchical Transformer attention to model both frame-level features and segment-level context, achieved the strongest performance with 66.39% frame accuracy (MoF) and 30.57% F1@0.50. LTContext [2] reached 48.87% accuracy and 17.81% F1@0.50, demonstrating the importance of long-range temporal modeling but suggesting that additional architectural innovations are needed for sequences at this scale. DiffAct [22], a diffusion-based model, struggled with 35.65% accuracy and only 3.31% F1@0.50, indicating that diffusion approaches may require domain-specific adaptations for fine-grained animal behavior.

We also benchmarked TSA-ActionSeg [7], an unsupervised temporal segmentation method, to assess whether meaningful behavioral structure can be discovered without explicit labels (Table 4). The best configuration (FINCH clustering) achieved 14.73% MoF, revealing a substantial 51-point gap compared to supervised methods. This result highlights two key challenges: (1) the high semantic complexity of animal behavior, which spans multiple functional categories with subtle distinctions, makes unsupervised discovery difficult, and (2) the temporal ambiguity of behavior transitions, where boundaries between states are often gradual rather than discrete. The relatively low absolute performance of all methods underscores the difficulty of this benchmark and its potential to drive innovation in long-form temporal modeling.

4.2. Benchmark 2: Behavior Recognition

Inferring semantic actions from skeletal pose is critical for privacy-preserving monitoring, welfare monitoring, and tracking. We frame this as a skeleton-based action recognition task: given pose trajectories (time-series of 39 keypoint coordinates) [32], classify behaviors including threat displays, grooming, playful interaction, and close-proximity contact. This task isolates the question of whether kinematic patterns encode sufficient information for behavioral discrimination.

Table 4. **Benchmark 1: Temporal Action Segmentation**. Results for Supervised Models and Unsupervised methods. Supervised models are evaluated on frame-wise action prediction.

Supervised Models				
Model	**Acc (MoF)↑**	**F1@0.10↑**	**F1@0.25↑**	**F1@0.50↑**
FACT [25]	**66.39**	**40.76**	**36.94**	**30.57**
LTContext [2]	48.87	34.99	26.33	17.81
DiffAct [22]	35.65	19.83	11.57	3.31
ASFormer [43]	34.15	13.43	8.96	2.99
SSTDA [11]	32.27	13.41	7.32	1.22
MS-TCN++[20]	29.72	15.12	6.98	4.65
UVAST[4]	25.56	5.79	3.22	1.61
Unsupervised Methods [7]				
Clustering	**Acc (MoF)↑**	**IoU↑**	**F1↑**	**Edit↓**
kmeans	13.34	10.97	9.41	13.40
finch	**14.73**	11.37	9.28	14.75
spectral	13.23	11.07	9.29	**13.32**
twfinch	14.14	**11.95**	**9.75**	14.18

We benchmarked three graph convolutional network (GCN) architectures designed for skeleton-based recognition (Table 5). Each model was trained on pose trajectories extracted from annotated wireframes, with the skeleton graph structure defined by anatomical connectivity. MS-G3D [23] achieved the highest mean F1 of 0.39, with the strongest performance on grooming (0.51), playful behavior (0.52), and pushing (0.53). AMGCN [42] reached comparable overall performance (F1=0.36) but excelled on threat detection (0.50). The adaptive 2S-AGCN [35] showed strong performance on attentive behavior (0.62) but struggled with most other categories (mean F1=0.26).

These results reveal an important pattern: behaviors with stereotyped, repetitive motion signatures (grooming, head orientation changes) are more amenable to pose-based recognition, while ambiguous social interactions with subtle postural cues remain challenging. The modest absolute scores indicate significant room for improvement, particularly through attention mechanisms over specific body parts (e.g., head, tail) or multimodal fusion combining pose and appearance. This benchmark provides a testbed for advancing skeleton-based recognition beyond human-centric assumptions, where non-rigid bodies and species-specific kinematic priors must be incorporated.

4.3. Benchmark 3: Longitudinal Behavioral Classification

We introduce a novel video understanding task: predicting latent phenotypic traits from observable behavior. Given video from one or more experimental scenarios, the goal is to classify each subject into its early-life rearing group (full maternal contact, half-time contact, or permanent separation). Unlike standard action recognition, where labels correspond to visible actions, this task requires models to extract distributional signatures. This includes subtle patterns in exploration latency, vigilance duration, approach distances, and social engagement that correlate with developmental history nine months prior.

Table 5. **Benchmark 2: Behavior Recognition.** Comparison of GCN-based models on cow behavior recognition from pose data. All scores are reported as F1.

Action	**AMGCN** [42]	**MS-G3D** [23]	**2S-AGCN** [35]
Attentive	0.39	0.02	0.62
Threat	0.50	0.16	0.14
Close Proximity	0.09	0.50	0.13
Grooming	0.35	0.51	0.22
Playful	0.40	0.52	0.23
Push	0.45	0.53	0.24
Sexual Behavior	0.34	0.49	0.21
Mean F1 Score	0.36	0.39	0.26

Table 6. **Benchmark 3: Longitudinal Behavioral Classification.** Comparison of various video models on cow behavioral classification across multiple experimental setups. All accuracies are reported as percentages.

Experiment	**TimeS** [6]	**VSwin** [24]	**ViViT** [1]	**UniFormer** [19]
NE	25.18	18.00	30.00	88.10
NO	32.10	14.82	23.46	88.89
HA	22.22	22.22	20.99	83.95
UCR	24.00	17.00	25.00	85.00
UCU	25.00	16.00	28.39	87.00
DR	96.67	63.33	66.67	86.67
DU	93.33	56.67	76.67	70.00
Mean	45.50	29.72	38.74	84.23

Abbreviations: NE: Novel Environment, **NO**: Novel Object, **HA**: Human Approach, **UCR**: Unfamiliar Conspecific (Restricted), **UCU**: Unfamiliar Conspecific (Unrestricted), **DR**: Dam (Restricted), **DU**: Dam (Unrestricted).

We evaluated four video Transformer architectures across all seven experimental scenarios (Table 6). UniFormer [19] achieved the highest mean accuracy (84.23%), demonstrating particularly strong performance on novel object (88.89%) and novel environment (88.10%) tests. TimeSformer [6] reached 96.67% accuracy on dam reunion scenarios but struggled with human interaction (22.22%) and novel objects (32.10%), suggesting sensitivity to scenario-specific motion patterns rather than generalizable phenotypic features. Video Swin Transformer [24] (29.72%) and ViViT [1] (38.74%) underperformed, indicating that not all Transformer architectures effectively capture the temporal statistics relevant for this task.

The strong performance on dam reunion scenarios, where discriminative signal may be amplified by emotional response to the biological mother, contrasts with weaker performance on object and human interactions, where phenotypic differences manifest more subtly. This variability highlights the challenge of learning treatment-discriminative features that generalize across diverse behavioral contexts. Success on this benchmark would demonstrate a model's capacity to extract stable behavioral signa-

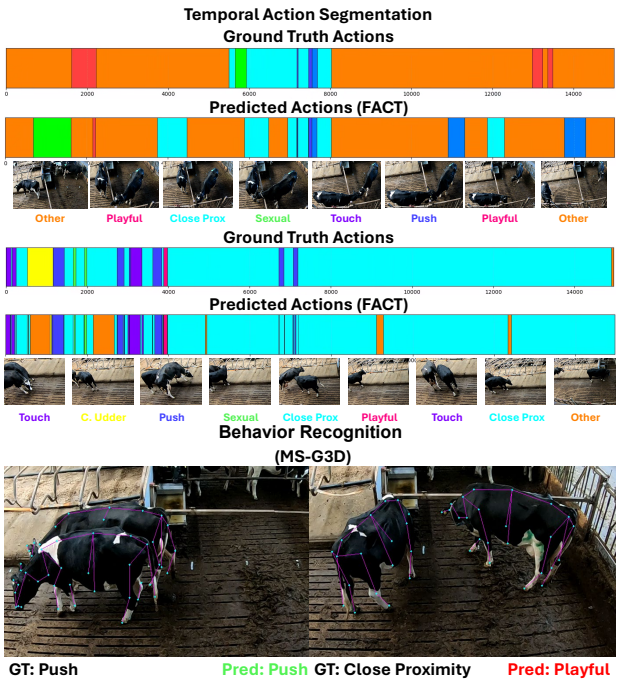

Figure 4. Qualitative analysis: For temporal action segmentation and behavior recognition, comparing ground truth and benchmarks.

tures: inferring latent causes (early-life experience) from observed effects (current behavior patterns), a capability with implications for behavioral genomics, personality assessment, and welfare diagnostics.

4.4. Qualitative Analysis

Figure 4 shows representative frames where models succeed and fail. FACT correctly segments extended grooming sequences, where repetitive licking motions provide strong temporal cues. However, it struggles with behavioral interpretation when subjects are at similar distances. For instance, when slight distance exists between cows, the model may ambiguously predict either ”close proximity” or ”sexual,” failing to grasp finer behavioral cues indicative of specific interactions such as push, touch, or contact with rear end. Conversely, when there is sufficient overlap between cows, the model tends to misclassify interaction behaviors as individual behaviors like ”other” or ”playful,” suggesting difficulty in distinguishing between social engagement and solitary actions based on spatial configuration alone.

Pose-based models exhibit complementary failure modes. MS-G3D correctly classifies ”push” actions where limb configurations are distinctive, but fails on ”close proximity” where spatial relationships matter more than joint angles. This suggests that pure skeleton-based approaches miss critical contextual information (inter-agent distance, environmental layout) that requires scene-level reasoning. Hybrid architectures combining pose with spatial scene graphs may address this limitation.

Dataset Limitations. We acknowledge limitations that motivate future work. Our dataset captures a single breed (Holstein dairy cattle) in a single facility, potentially limiting generalization to diverse farm environments and management practices. The fixed camera positions, while sufficient for our scenarios, constrain the viewpoint diversity compared to truly in-the-wild video capture. The relatively small subject count (N = 43), though typical for longitudinal behavioral studies, limits statistical power for certain phenotyping analyses. Future iterations could expand species diversity, incorporate true in-the-wild interaction scenarios (e.g., pasture-based mother-calf bonding), and scale to larger herds with automated tracking systems.

5. Conclusion

We presented MooCap, a multi-view video benchmark for fine-grained bovine behavior analysis featuring 42 hours of synchronized video across 43 subjects and seven structured interaction scenarios. Unlike existing animal datasets that rely on passive observation, MooCap embeds controlled ethological stimuli, including novel environments, objects, humans, and conspecifics, enabling models to learn relationships between visual patterns and behavioral responses. Our dense annotations span three modalities: frame-level action labels (23 behaviors), skeletal pose (39 keypoints), and spatial occupancy, supporting diverse tasks from temporal segmentation to pose-based recognition.

Baseline results across our three core benchmarks reveal substantial room for improvement. State-of-the-art temporal segmentation models achieve only 66.39% frame accuracy on long-form sequences, while skeleton-based approaches reach just 0.39 mean F1 on action classification. The longitudinal phenotyping task demonstrates that Transformer models can extract some phenotypic signatures, but struggle to generalize across interaction contexts. These gaps highlight fundamental challenges in animal behavior analysis, including extreme temporal extent, subtle non-rigid motions, multi-entity interactions, and the need to infer stable individual differences from distributional patterns rather than isolated action instances. Overall, MooCap bridges computer vision and behavioral science, providing a testbed for algorithms that move beyond frame-level action recognition toward analysis of behavioral dynamics.

Dataset, code, and evaluation tools is made publicly available to support reproducible research.

6. Acknowledgments

We thank the barn staff at the Danish Cattle Research Centre, Aarhus University (Tjele, Denmark) for management

and care of the animals; Emma Hvidtfeldt Jensen (Aarhus University) for assistance with animal management and data collection; and Julie Tekieli (ISARA, France) for video observation and behavioural coding. We also thank Prof. Margit Bak Jensen (Aarhus University) for overall leadership of the original experimental project from which these data were derived. The behavioral data collection at Aarhus University was supported by Independent Research Fund Denmark as part of the project "Can dairy cows have the best of both worlds?" (2020–2024).

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 7

[2] Emad Bahrami, Gianpiero Francesca, and Juergen Gall. How much temporal long-term context is needed for action segmentation? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6, 7

[3] Sara Beery, Arush Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 3

[4] Nadine Behrmann, S Alireza Golestaneh, Zico Kolter, Juergen Gall, and Mehdi Noroozi. Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In *European Conference on Computer Vision (ECCV)*, pages 52–68. Springer, 2022. 7

[5] Daniel Berckmans. Precision livestock farming technologies for welfare management in intensive livestock systems. *Animal Frontiers*, 4(1):14–18, 2014. 1

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 7

[7] Elena Bueno-Benito, Biel Tura Vecino, and Mariella Dimiccoli. Leveraging triplet loss for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4921–4929, 2023. 6, 7

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 3

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 3, 11

[10] Jun Chen, Ming Hu, Blair Costelloe, Darren J Coker, Michael L Berumen, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13052–13061, 2023. 1, 3, 4

[11] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9454–9463, 2020. 7

[12] Olivier Friard and Marco Gamba. Boris: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, 7 (11):1325–1330, 2016. 5

[13] P.H Hemsworth. Human–animal interactions in livestock production. *Applied Animal Behaviour Science*, 81(3):185–198, 2003. International Society for Applied Ethology Special Issue: A selection of papers from the ISAE international congresses, 1999-2001. 5

[14] Daniel Joska, Liam Clark, Naoya Murrone, et al. Acinoset: A 3d pose estimation dataset and baseline models for cheetahs in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[15] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011. 3

[16] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2556–2563. IEEE, 2011. 3

[17] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 780–787, 2014. 3, 5, 6

[18] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Steffen Schneider, T Nath, M M Rahman, V Di Santo, D Soberanes, G Feng, and et al. Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods*, 19 (4):496–504, 2022. 6

[19] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning, 2022. 7

[20] Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 7

[21] Dan Liu, Jin Hou, Shaoli Huang, Jing Liu, Yuxin He, Bochuan Zheng, Jifeng Ning, and Jingdong Zhang. Loteanimal: A long time-span dataset for endangered animal behavior understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20064–20075, 2023. 3

[22] Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6, 7

[23] Ziyu Liu, Hongsheng Zhang, Zhenghao Wang, Wanli Ouyang, and Guan Wang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 143–152, 2020. 7

[24] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 7

[25] Zijia Lu and Ehsan Elhamifar. FACT: Frame-action cross-attention temporal modeling for efficient supervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18175–18185, 2024. 6, 7

[26] Xiaoxuan Ma, Stephan Kaufhold, Jiajun Su, Wentao Zhu, Jack Terwilliger, Andres Meza, Yixin Zhu, Federico Rossano, and Yizhou Wang. Chimpact: A longitudinal dataset for understanding chimpanzee behaviors. *Advances in Neural Information Processing Systems*, 36:27501–27531, 2023. 3

[27] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie W Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018. 3, 4, 6

[28] Alexander Mathis, Thomas Biasi, Steffen Schneider, et al. Pretraining boosts out-of-distribution robustness for pose estimation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 3

[29] Mackenzie W Mathis and Alexander Mathis. Deep learning tools for the measurement of animal behavior in neuroscience. *Current opinion in neurobiology*, 60:1–11, 2020. 6

[30] Heather W. Neave, Emma Hvidtfeldt Jensen, Marine Durrenwachter, and Margit Bak Jensen. Behavioral responses of dairy cows and their calves to gradual or abrupt weaning and separation when managed in full- or part-time cow-calf contact systems. *Journal of Dairy Science*, 107(4):2297–2320, 2024. 5

[31] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19023–19034, 2022. 1, 3, 4

[32] Ian Noronha, Advait Prasad Jawaji, Juan Camilo Soto, Jiajun An, Yan Gu, and Upinder Kaur. MBE-ARI: A multimodal dataset mapping bi-directional engagement in animal-robot interaction. In *Accepted to ICRA 2025*, 2025. 3, 6

[33] Talmo D. Pereira, Nathaniel Tabris, Arie Matsliah, David M. Turner, Junyu Li, Shruthi Ravindranath, Eleni S. Papadoyannis, Edna Normand, David S. Deutsch, Z. Yan Wang, Grace C. McKenzie-Smith, Catalin C. Mitelut, Marielisa Diez Castro, John D'Uva, Mikhail Kislin, Dan H. Sanes, Sarah D. Kocher, Samuel S.-H. Wang, Annegret L. Falkner, Joshua W. Shaevitz, and Mala Murthy. SLEAP: A deep learning system for multi-animal pose tracking. *Nature Methods*, 19(4):486–495, 2022. 3, 4

[34] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, pages 32–36. IEEE, 2004. 3

[35] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12026–12035, 2019. 7

[36] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding, 2016. 3

[37] Jennifer J. Sun, Tomomi Karigo, Dipam Chakraborty, Sharada P. Mohanty, Benjamin Wild, Quan Sun, Chen Chen, David J. Anderson, Pietro Perona, Yisong Yue, and Ann Kennedy. The multi-agent behavior dataset: Mouse dyadic social interactions, 2021. 3

[38] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528. IEEE, 2011. 3

[39] Susanne Waiblinger, Xavier Boivin, Vibeke Pedersen, Maria-Vittoria Tosi, Andrew M. Janczak, E. Kathalijne A. Visser, and Robert B. Jones. Assessing the human–animal relationship in farmed species: a critical review. *Applied Animal Behaviour Science*, 101(3–4):185–242, 2006. 5

[40] Aharon Weissbrod, Alexander Shapiro, Genadiy Vasserman, Liat Edry, Molly Dayan, Assif Yitzhaky, Libi Hertzberg, Ofer Feinerman, and Tali Kimchi. Automated long-term tracking and social behavioural phenotyping of animal colonies within a semi-natural environment. *Nature Communications*, 4, 2013. 1

[41] {Megan M.} {Woodrum Setser}, {Heather W.} Neave, and {Joao H.C.} Costa. The history, implementation, and application of personality tests in livestock animals and their links to performance. *Applied Animal Behaviour Science*, 268, 2023. 4

[42] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 7

[43] F. Yi, H. Wen, and T. Jiang. ASFormer: Transformer for action segmentation. In *Proc. BMVC*, 2021. 7

[44] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild, 2021. 3

[45] Ali Zia, Renuka Sharma, Reza Arablouei, Greg Bishop-Hurley, Jody McNally, Neil Bagnall, Vivien Rolland, Brano Kusy, Lars Petersson, and Aaron Ingham. Cvb: A video dataset of cattle visual behaviors, 2023. 3

[46] Ali Zia, Renuka Sharma, Abdelwahed Khamis, Xuesong Li, Muhammad Husnain, Numan Shafi, Saeed Anwar, Sabine Schmoelzl, Eric Stone, Lars Petersson, and Vivien Rolland. A review on coarse to fine-grained animal action recognition. *arXiv preprint arXiv:2506.01214*, 2025. 3