A. Appendices

A.1. Ethical Considerations and Animal Welfare

Ensuring high ethical standards was paramount to this study. All experimental procedures were conducted at the Danish Cattle Research Centre, Aarhus University (Tjele, Denmark). All animal procedures were approved by the Danish Animal Experiments Inspectorate (Permit No. 2021–15–0201–00989) in accordance with the Danish Ministry of Environment and Food Act No. 474 (May 15, 2014). Trained animal handlers supervised all sessions and could terminate tests if animals showed signs of acute distress (although early terminations were not necessary). All treatments met or exceeded standard industry welfare guidelines, and animals received comprehensive veterinary care throughout.

This research is ethically justified by its potential to advance automated welfare monitoring systems that could improve the quality of life for millions of livestock animals globally. By developing computer vision tools for early detection of distress, illness, or suboptimal housing conditions, this work contributes to the welfare science goal of continuous, non-invasive monitoring at scale. Detailed protocols and ethical guidelines will be released with the dataset to support reproducible research.

A.2. Benchmark 1: Temporal Action Segmentation

A.2.1. Input Feature Extraction

To ensure consistent inputs across all temporal segmentation models, we construct a shared visual representation using a 3D convolutional backbone. We fine-tune an I3D network [9] on the MooCap dataset using 64-frame clips sampled at 25 fps, optimizing for classification over all 23 annotated behaviors. After fine-tuning, the I3D model is frozen and applied to the full-length untrimmed videos. From the penultimate layer we extract per-frame descriptors and form a sequence

$$F_0 \in \mathbb{R}^{T\times D}, \tag{1}$$

where T denotes the number of frames in each scenario and $D = 1024$ is the feature dimensionality. All vectors are ℓ_2-normalized. These features serve as compact, motion-aware descriptors suitable for long-range temporal reasoning and ensure that all benchmarked models operate under the same visual input distribution.

A.2.2. Evaluation Metrics

We follow the standard temporal action segmentation protocol used in prior work on long-form activity datasets. Let $\hat{y}_t$ and y_t be the predicted and ground-truth action labels for frame t. Framewise accuracy is computed as

$$\text{Acc} = \frac{1}{T}\sum_{t=1}^{T}[\hat{y}_t = y_t]. \tag{2}$$

To evaluate temporal localization, we report segmental F1 scores at overlap thresholds $\theta \in \{0.10, 0.25, 0.50\}$. A prediction segment p matches a ground-truth segment g when

$$\text{IoU}(p, g) = \frac{|p \cap g|}{|p \cup g|} \geq \theta, \tag{3}$$

Finally, we report the Edit score, which measures sequence-level ordering agreement using normalized Levenshtein distance:

$$\text{Edit} = 1 - \frac{\text{Lev}(\hat{Y}, Y)}{\max(|\hat{Y}|, |Y|)}. \tag{4}$$

These metrics capture both per-frame correctness and the structural quality of predicted action sequences across long, fine-grained behavioral recordings.

A.2.3. Training Configuration

All supervised segmentation models are trained under a unified optimization setup for comparability. We use AdamW with an initial learning rate of 1×10^{-4}, weight decay of 0.05, and a batch size of one full video. Each model receives the same I3D feature sequence F_0, ensuring that architectural differences drive performance variance. We use the Adam optimizer with an initial learning rate of 0.001 and a StepLR scheduler that decays the learning rate every 10 epochs by a multiplicative factor of 0.1, with a batch size of 32. Models are trained with early stopping based on validation performance to prevent overfitting. All evaluations strictly follow the official training/validation/test splits, with no subject identity overlap across sets.

A.3. Benchmark 2: Pose-Based Behavior Recognition

A.3.1. Pose Sequence Construction

For pose-driven behavior recognition, we leverage the dense skeletal annotations available in MooCap. Each frame contains 39 anatomical keypoints, yielding a pose vector $\mathbf{p}_t \in \mathbb{R}^{39\times 2}$. Missing keypoints due to occlusion are filled via per-joint temporal interpolation. Each annotated video produces a trajectory

$$P \in \mathbb{R}^{T\times 39\times 2}, \tag{5}$$

which is normalized by subtracting per-joint means and dividing by per-joint standard deviations estimated over the training set. These pose sequences constitute the sole input modality for all GCN-based models.

A.3.2. Evaluation Metrics

Following standard skeleton-based action recognition protocols, we report per-class F1 scores and mean F1 across

the 23 behavior categories. For each behavior class c, we compute

$$\mathrm{F1}_c = \frac{2 \cdot \mathrm{P}_c \cdot \mathrm{R}_c}{\mathrm{P}_c + \mathrm{R}_c}, \tag{6}$$

and report the average

$$\mathrm{mF1} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{F1}_c, \tag{7}$$

where $C = 23$ is the number of annotated behaviors. This metric emphasizes balanced performance across highly imbalanced behavior classes, reflecting the distributional challenges of fine-grained ethological actions.

A.3.3. Training Configuration

All pose-based models in this benchmark are trained using a shared optimization protocol derived from the MSG3D training setup to ensure fairness across architectures. We use the Adam optimizer with an initial learning rate of 0.001 and a StepLR scheduler that reduces the learning rate every 10 epochs by a factor of 0.1, using a batch size of 32. Training proceeds with early stopping based on validation mF1 to ensure stable convergence across models. Data augmentation includes random temporal cropping, sequence jittering, and Gaussian perturbation of joint coordinates, which helps models remain robust to annotation noise and subtle variations in pose extraction. All evaluations follow the official subject-disjoint split, ensuring that no cow identity appears across both training and testing sets.

A.4. Benchmark 3: Longitudinal Behavioral Classification

A.4.1. Input Processing

For the longitudinal phenotype classification task, we evaluate whether video models can infer early-life rearing treatment namely full contact, half contact, or separation at birth from observable behavior across entire testing scenarios. For this benchmark, we operate directly on RGB video frames. Each model receives uniformly sampled clips from each scenario, with clip lengths of 8–16 frames depending on model requirements. Frames are resized to 224×224 and normalized following standard ImageNet-style preprocessing.

To account for the extreme length of MooCap recordings, we sample non-overlapping clips from each video, producing a sequence of clip-level embeddings which are temporally averaged before classification. This strategy preserves global behavioral statistics while enabling tractable training on long sequences.

A.4.2. Evaluation Metrics

We report classification accuracy for each experimental scenario as well as overall mean accuracy. Given subject-level predictions $\hat{z}$ and ground-truth treatment labels z, accuracy is defined as

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} [\hat{z}_i = z_i], \tag{8}$$

where N is the number of test subjects. Because each subject undergoes multiple scenarios, we compute accuracy separately for each scenario to assess the degree to which behavioral signatures generalize across contexts.

A.4.3. Training Configuration

We fine-tune four Transformer-based video models TimeSformer, ViViT, Video Swin, and UniFormer using AdamW with a base learning rate of 2×10^{-4} and a warm-up of 20 epochs, followed by a StepLR scheduler that reduces the learning rate every 10 epochs by a factor of 0.1. All models are trained with early stopping based on validation accuracy. Training is performed using the official subject-disjoint splits to prevent leakage of individual motion signatures. Models are trained independently per scenario and evaluated both per-scenario and across-scenario to quantify the stability of behavioral indicators. Data augmentation includes random horizontal flips, color jitter, and random temporal resampling to increase robustness to lighting and motion variability.