KABR-TOOLS: AUTOMATED FRAMEWORK FOR MULTI-SPECIES BEHAVIORAL MONITORING

Jenna Kline[1], Maksim Kholiavchenko[2], Samuel Stevens[1], Nina van Tiel[3], Alison Zhong[1], Namrata Banerji[1], Alec Sheets[1], Sowbaranika Balasubramaniam[1], Isla Duporge[4], Matthew Thompson[1], Elizabeth Campolongo[1], Jackson Miliko[5], Neil Rosser[6], Tanya Berger-Wolf[1], Charles V. Stewart[2], and Daniel I. Rubenstein*[4,5]

[1]The Ohio State University
[2]Rensselaer Polytechnic Institute
[3]École Polytechnique Fédérale de Lausanne
[4]Princeton University
[5]Mpala Research Centre
[6]University of Miami
*corresponding author

ABSTRACT

A comprehensive understanding of animal behavior ecology depends on scalable approaches to quantify and interpret complex, multidimensional behavioral patterns. Traditional field observations are often limited in scope, time-consuming, and labor-intensive, hindering the assessment of behavioral responses across landscapes. To address this, we present kabr-tools (**K**enyan **A**nimal **B**ehavior **R**ecognition **Tools**), an open-source package for automated multi-species behavioral monitoring. This framework integrates drone-based video with machine learning systems to extract behavioral, social, and spatial metrics from wildlife footage. Our pipeline leverages object detection, tracking, and behavioral classification systems to generate key metrics, including time budgets, behavioral transitions, social interactions, habitat associations, and group composition dynamics. Compared to ground-based methods, drone-based observations significantly improved behavioral granularity, reducing visibility loss by 15% and capturing more transitions with higher accuracy and continuity. We validate kabr-tools through three case studies, analyzing 969 behavioral sequences, surpassing the capacity of traditional methods for data capture and annotation. We found that, like Plains zebras, vigilance in Grevy's zebras decreases with herd size, but, unlike Plains zebras, habitat has a negligible impact. Plains and Grevy's zebras exhibit strong behavioral inertia, with rare transitions to alert

behaviors and observed spatial segregation between Grevy's zebras, Plains zebras, and giraffes in mixed-species herds. By enabling automated behavioral monitoring at scale, `kabr-tools` offers a powerful tool for ecosystem-wide studies, advancing conservation, biodiversity research, and ecological monitoring.
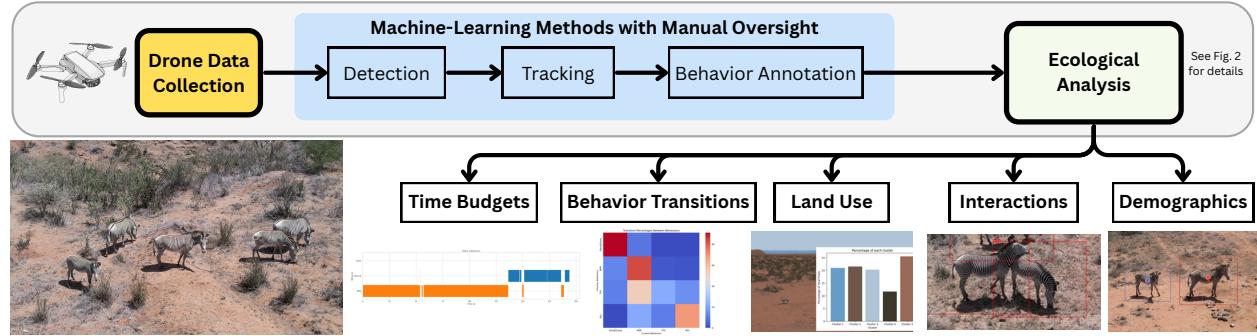
1 Introduction

Figure 1: `kabr-tools` computational framework for automated wildlife behavioral monitoring. The modular pipeline processes drone video through object detection, individual tracking, and machine learning-based behavioral classification to generate ecological metrics including time budgets, behavioral transitions, land and habitat use, social interactions, and demographic data. Framework design enables integration of novel ML models and adaptation across species and study systems. See Fig. 2 for technical details on ML methods with manual oversight.

Large-scale wildlife monitoring across diverse ecosystems increasingly requires computational frameworks capable of processing multi-dimensional behavioral datasets to understand behavioral patterns [Koger et al., 2023, Duporge et al., 2025]. The data-intensive nature of modern ecology, intensified by climate change impacts, habitat fragmentation, and biodiversity loss, demands automated approaches that can scale beyond the temporal and spatial limitations of traditional observation methods [Anderson and Perona, 2014, Besson et al., 2022]. While field studies of animal behavior have successfully employed systematic observation methods [Altmann, 1974], these approaches constrain researchers to small-scale, labor-intensive studies that cannot efficiently address landscape-level ecological questions or multi-species comparative analyses.

Traditional behavioral sampling methods, such as focal and scan sampling, have provided the foundation for understanding species ecology and behavioral adaptations through the construction of time budgets [Bateson and Martin, 2021]. However, these approaches face fundamental scalability limitations when applied to contemporary ecological challenges [Smith and Pinter-Wollman, 2021]. Focal animal sampling, while providing detailed behavioral sequences, restricts observations to single individuals and cannot efficiently capture the simultaneous responses of multiple group members to environmental pressures [Amato et al., 2013]. Moreover, by the time the final individual is observed, its state and needs may differ from those observed earlier in the sequence. Scan sampling addresses temporal alignment by recording instantaneous group behavior, but it frequently misses brief but ecologically critical events such as vigilance responses or predator detection behaviors [Amato et al., 2013]. Both methods require extensive human observation

effort that constrains the spatial scale and temporal duration of studies, limiting our ability to understand behavioral responses to environmental change across communities and landscapes.

Recent advances in remote sensing and computational analysis offer transformative opportunities for scaling behavioral monitoring to meet ecological research needs [Hughey et al., 2018, Corcoran et al., 2021, Besson et al., 2022, Koger et al., 2023, Kline et al., 2025d]. Drone technology provides aerial observation platforms that can simultaneously monitor multiple individuals across groups while minimizing observer effects and visual obstructions [Afridi et al., 2024, Pedrazzi et al., 2025, Schad and Fischer, 2023]. More critically, the integration of computer vision and machine learning enables automated processing of the large video datasets generated by drone surveys, transforming behavioral ecology from observation-limited to data-intensive science [Petso et al., 2021, Zhang et al., 2024, Saoud et al., 2024, Axford et al., 2024, Chan et al., 2024, Elhorst et al., 2025, Kline et al., 2025a]. These computational approaches can process behavioral information at scales previously impossible with manual methods, enabling more efficient landscape-level monitoring, multi-species comparative studies, and long-term assessment of behavioral responses to environmental change.

Building on this foundation, the Kenyan Animal Behavior Recognition (KABR) project initially developed a dataset of drone-based behavioral observations to train machine learning models for automated behavior classification [Kholiavchenko et al., 2024b]. This foundational work established methodological approaches for drone-based behavioral data collection and demonstrated the feasibility of automated behavior recognition across multiple African savanna species, including Plains zebras (*Equus quagga*), Grevy's zebras (*Equus grevyii*) and reticulated giraffes (*Giraffa reticulata*). However, the original KABR framework was primarily designed for machine learning dataset preparation rather than comprehensive ecological analysis. Here, we present an expanded computational framework that extends the KABR project beyond dataset creation to provide comprehensive tools for ecological behavioral analysis, following the application-driven machine learning framework [Rolnick et al., 2024]. Our `kabr-tools` framework transforms the original machine learning training pipeline into a complete software suite for automated wildlife behavioral monitoring. We build upon the validated methodological foundation established by the original KABR dataset while significantly expanding both analytical capabilities and taxonomic scope. This expanded framework addresses the scalability challenges of traditional methods by providing tools for processing large-scale behavioral datasets across multiple species and ecological contexts.

We demonstrate the effectiveness of drones in collecting data, paired with machine learning for analysis, by benchmarking the KABR dataset against traditional, ground-based expert observation methods [Kline et al., 2025b]. We validate our expanded framework using five key metrics and three case studies drawn from the behavioral data in the original KABR video clips and 50 minutes of newly released full-length videos [Kline et al., 2025c]. We provide worked examples for fourteen new videos, including annotated detections, tracks, behavior labels, and telemetry data [Kline et al., 2025e]. Our enhanced framework provides automated extraction of time budgets, behavioral transitions, social interactions, and habitat associations across these diverse study systems. The modular design enables integration of emerging machine learning models and adaptation to new species and study contexts. We demonstrate the ecological

3

applications of this approach through three case studies, which include the computational analysis of anti-predator behavior in Grevy's zebras, transitions in zebra behavior, and inter- and intra-specific social interactions. Our analysis reveals species-specific risk management strategies across habitat gradients. We provide comprehensive documentation [1], open-source code [2]. All the data is available on HuggingFace, including methodological comparisons[3], and worked examples for three sessions, including full length videos full-length videos[4], detections, tracks, behavior annotations, and associated telemetry[5]. We provide a glossary bridging behavioral ecology and computer vision terminology to facilitate interdisciplinary adoption (Tab. 1). This evolution from dataset creation to a comprehensive computational ecology framework advances automated wildlife monitoring capabilities while providing scalable tools for addressing contemporary conservation biology and ecosystem management challenges.

Computer Science Term	Definition
Annotation	Manual data labeling process – used to train ML models or classify behaviors in datasets
CVAT	Computer Vision Annotation Tool – open-source platform for annotating videos and manually labeling animal behaviors [CVAT.ai, 2023]
Head Classes	Common behaviors with many training examples (e.g., Walk, Graze, Head Up)
KABR	Kenya Animal Behavior Recognition project – a computational framework for automated animal behavior analysis from drone footage [Kholiavchenko et al., 2024b]
Machine Learning (ML)	A type of AI where algorithms learn patterns from data and improve over time without explicit programming. Used in ecology to automate behavior analysis.
Mini-scene	Cropped video sequence focused on a single animal, extracted from drone footage and centered on the individual
Tail Classes	Rare behaviors with few training samples (e.g., Fighting), harder to classify
Ecological Term	**Definition**
Animal Behaviour Pro	Smartphone app for logging field observations with timestamps and behavior categories [Newton-Fisher, 2021]
Drone-based Sampling	Behavior is extracted from drone footage, either manually or via AI. Enables focal sampling of all individuals simultaneously.
Ethogram	A catalog of species-specific behaviors with standardized descriptions (See example in Tab. 13)
Focal Sampling	Continuous observation of a single individual recording all behaviors for a predetermined time [Altmann, 1974]
Ground-based Sampling	Observer records behavior manually (e.g., via binoculars) using focal or scan sampling [Altmann, 1974]
Instantaneous Sampling	Behavior is recorded only at specific observation instants, often using scan sampling
Scan Sampling	Recording behavior of all visible group members at regular intervals [Altmann, 1974]
Time Budget	Proportion of time animals spend on various activities like feeding or moving

Table 1: Glossary of Computer Science and Ecological Terms and Tools

---

[1]https://imageomics.github.io/kabr-tools/
[2]https://github.com/Imageomics/kabr-tools
[3]https://huggingface.co/datasets/imageomics/kabr-methodology
[4]https://huggingface.co/datasets/imageomics/kabr-full-video
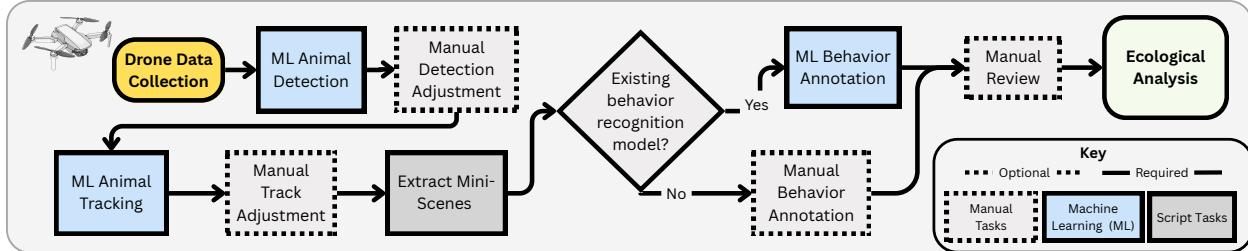[5]https://huggingface.co/datasets/imageomics/kabr-worked-examples

Figure 2: `kabr-tools` video processing pipeline for drone-based behavioral analysis. Raw drone footage undergoes automated animal detection and tracking, followed by mini-scene extraction to create individual-focused video segments. Manual behavioral annotation using standardized ethograms produces continuous behavioral records, with quality control measures ensuring dataset accuracy and consistency.

2 Methodology

We present an end-to-end framework for drone-based behavioral ecology that integrates scalable video processing, expert annotation, machine learning classification, and ecological analysis. Our methodological design prioritizes automation, ecological validity, and minimal disturbance to wildlife. The `kabr-tools` pipeline, illustrated in Fig. 2, transforms raw drone footage into behavior-annotated videos. We validate our proposed drone-based behavioral analysis pipeline by collecting data on Grevy's zebras, Plains zebras, and reticulated giraffes at Kenya's Mpala Research Centre, using standardized protocols to minimize animal disturbance (Section 2.1). We describe the pipeline in detail (Section 2.2), including mini-scene extraction (Section 2.2.1) and the video annotation protocol (Section 2.2.2). We describe our comprehensive multi-tiered validation strategy (Section 2.3), which includes machine learning evaluation, cross-method comparisons, and application-driven ecological tasks [Rolnick et al., 2024]. We demonstrate that our modular framework enables reproducible, non-invasive behavioral monitoring at scale and provides a template for extending AI-driven analysis to other ecological systems.

2.1 Data Collection for Evaluation

We collected the data at the Mpala Research Center in Laikipia, Kenya in January 2023, summarized in Tab. 2. First, we collected ground-based scan and focal sampling to familiarize the field observers with behavior observations and refine the ethograms (see Appendix 7.4). Next, we manually flew a series of drone missions to collect video footage of animal behaviors, while also conducting simultaneous ground-based focal sampling for selected missions. Acoustic guidelines were used to mitigate disturbance during flight operations [Duporge et al., 2021]. We collected behavior data of three ungulate species: Grevy's zebras, plains zebras, and reticulated giraffes across diverse habitat types, including open savanna, mixed woodland, and acacia woodland environments. When possible, demographic information, including age and sex classification, was recorded for the herds. We conducted 19 drone missions using a DJI Air 2S drone, collecting 10.5 hours of video (54 GB) in 4K and 5.4K resolution at 30 frames per second. All research was conducted under Institutional Animal Care and Use Committee approval (Princeton University IACUC 1835F) and Kenya's National Commission for Science, Technology & Innovation research license (NACOSTI/P/22/18214).

Category	Metric	Value
Dataset Overview	Study Period	January 11-21, 2023
	Location	Mpala Research Centre, Laikipia, Kenya
	Animals observed	97+ individuals
	Mini-scene count	969 (807 manually annotated, 162 automatically annotated with ML)
Species Distribution	Plains Zebra	378 mini-scenes (39.0%)
	Grevy's Zebra	275 mini-scenes (28.4%)
	Zebra (species unspecified)	198 mini-scenes (20.4%)
	Reticulated Giraffes	118 mini-scenes (12.2%)
Group Demographics	Group Size Range	2-16 individuals
	Mean Group Size	6.1 individuals
Drone Sessions	Total Sessions	17
	Session Duration Range	5 - 52 minutes
	Typical Session Length	24 minutes
	Recording Time Range	9:30 AM - 4:30 PM
Data Files	Detections	Bounding boxes locating and tracking animals through the video
	Mini-scenes	Cropped video clips created from detections focused on a single individual animal
	Behavior	Behavior label associated with each frame in the mini-scene
	Drone telemetry	GPS, altitude, heading, and speed
	Drone Session Summary	Session date-time, species, herd size, field notes
	Concurrent ground-based scan and focal data	8 sessions (2 giraffe, 4 Grevy's, 2 Plains)

Table 2: Summary of dataset used to validate `kabr-tools` pipeline and perform case studies analysis. Contains behavior data from eight concurrent scan and focal sampling sessions, 969 annotated mini-scenes obtained from drone footage, and drone session summary.
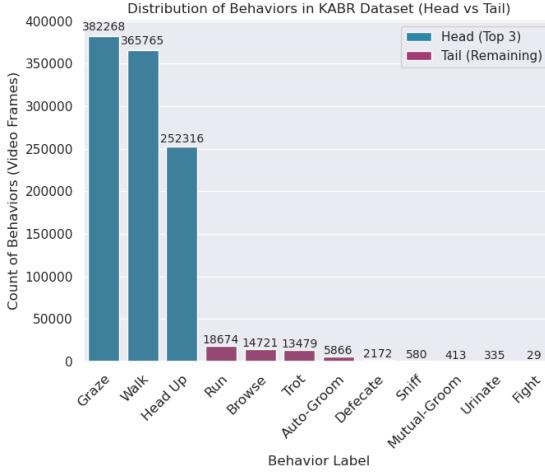

Figure 3: Distribution of behaviors in the dataset. Head classes constitute the majority of the dataset, which include *Graze*, *Walk*, and *Head Up*.

The dataset includes ground-based sampling data and drone videos capturing ungulate behaviors (Tab. 2). Ground-based scan and focal behavior data was collected for eight sessions to validate the automated classification approaches. We use 969 mini-scenes from 17 drone recording sessions annotated with our pipeline to validate our methodology and perform the case study analysis. The dataset contains twelve distinct behavior categories, shown in Fig. 3). Detailed definitions of the behavioral categories are summarized Appendix 7.4. The top seven categories occurring with the greatest frequency (*Graze, Walk, Head Up, Run, Browse, Trot, Auto-Groom*, as well as *Out of Sight*) are used for statistical analysis, including fine-tuning the KABR X3D model [Kholiavchenko et al., 2024c]. Primary behaviors common to all species *Walk* (forward locomotion with at least three consecutive steps), *Head Up* (head raised above shoulder height with ears oriented forward, indicating vigilance or scanning), and *Auto-groom* (self-directed grooming using mouth or hind limb). Secondary behaviors include species-specific activities such as *Browse* for giraffes, describing animals with their heads up eating leaves, and *Graze* for zebras, describing animals with their heads down eating grass. Social interactions, including *Mutual-Groom* and *Fight*, were recorded when observed but occurred too infrequently for statistical analysis.

2.2 Video Processing Pipeline

Our video processing methodology transforms raw drone footage into annotated behavioral datasets suitable for machine learning analysis. This approach enables simultaneous focal sampling of multiple individuals while maintaining the detailed behavioral records traditionally achieved through single-animal focal sampling. The pipeline consists of two main stages. First, the raw videos are processed into mini-scenes (Section 2.2.1). Next, the mini-scenes are labeled with behaviors (Section 2.2.2). Quality control measures ensure annotation consistency and accuracy through inter-annotator reliability testing and expert review. Ultimately, the resulting dataset is utilized for ecological analyses.

2.2.1 Mini-scene Extraction

The initial processing stage begins with object detection to identify all animals in the video frame, followed by continuous tracking to maintain individual identity across frames. From these tracked individuals, we extract *mini-scenes*: cropped video segments that focus on single animals while preserving behavioral context (Fig. 4). Mini-scenes simulate focal sampling conditions for multiple individuals simultaneously by maintaining each animal at the center of its respective video frame throughout the observation period. Our processing pipeline employs a YOLO-based convolutional neural network for object detection [Jocher et al., 2023], a state-of-the-art model pre-trained on the COCO dataset [Lin et al., 2015]. Detected animals are tracked continuously across frames using spatial proximity and visual feature matching, maintaining consistent individual identification even during temporary occlusions or when animals move
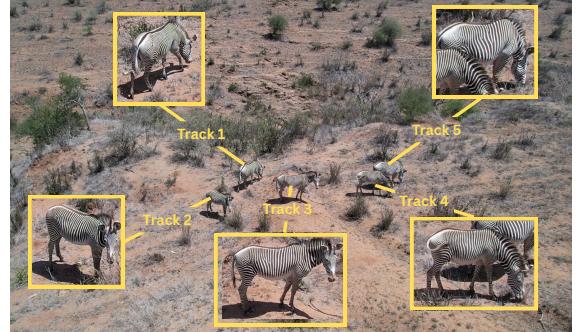

Figure 4: A mini-scene is a sub-image cropped from the drone video footage centered on and surrounding a single animal. Mini-scenes simulate the camera as well-aligned with each individual animal in the frame, compensating for the movement of the drone and ignoring everything in the large field of view but the animals' immediate surroundings. The KABR dataset comprises mini-scenes and their frame-by-frame behavior annotations [Kholiavchenko et al., 2024b].

close together. The system dynamically crops the full video frame around each tracked individual while preserving sufficient surrounding context for behavioral classification. This approach compensates for drone movement and camera angle changes, ensuring consistent framing regardless of the animal's position within the larger field of view. Quality control measures included manual checks identified evident tracking errors where the algorithm switched between individuals, which were excluded. We also filter out mini-scenes consisting of less than 90 frames (i.e. 3 seconds at 30 fps) to ensure a sufficient duration to reliably classify behavior for the focal animal. This duration threshold of three seconds was empirically determined to provide sufficient temporal context for accurate behavioral classification while maximizing data retention. Our manual review and adjustment capabilities corrected tracking errors and refine detection boundaries when necessary.

2.2.2 Video Annotation Protocol

Once the mini-scenes are extracted, they are annotated with behavior labels frame-by-frame using standardized behavioral labels. This creates continuous behavioral records for each tracked individual. These behavioral labels may be produced manually using an annotation tool such as CVAT, or automatically if a suitable behavior recognition model is available, such as KABR X3D [Kholiavchenko et al., 2024c] or YOLO-Behaviour [Chan et al., 2024]. To build our dataset, we conducted manual behavior annotations using the Computer Vision Annotation Tool (CVAT). CVAT is an open-source platform enabling precise frame-by-frame labeling of video data [CVAT.ai, 2023]. This annotation produces continuous behavioral records analogous to traditional focal sampling, with every frame labeled according to standardized species-specific ethograms (Appendix 7.4). The annotation protocol requires labeling every frame with the focal individual's current behavioral state, creating temporally precise behavioral sequences. We provide a CVAT user guide developed through this process in our Github repository.

Our annotation team consists of nine trained individuals: four field observers who collected the original ground-based data and five additional assistants who underwent comprehensive training to ensure consistency. Rigorous quality control measures were used to ensure annotation accuracy and consistency. Our standardized training program includes intensive review of species-specific behavioral definitions with video examples, technical instruction on the CVAT interface, and practice annotation sessions until achieving greater than 90% agreement with expert annotations. We employed weekly calibration sessions throughout the annotation period to address interpretation drift and maintain consistency across all annotators. These included random double-annotation of 10% of mini-scenes to monitor inter-annotator reliability (achieving $\kappa = 0.88$ for primary behavioral categories), weekly calibration sessions to address any annotation drift, and final expert review by field-experienced team members for all completed annotations. This multi-layered approach ensured that the resulting behavioral dataset met the precision standards required for both traditional ethological analysis and machine learning model training.

2.3 Evaluation Methodology

Our evaluation framework integrates cross-method sampling comparisons and application-driven ecological tasks to assess the effectiveness of our drone-based behavioral analysis pipeline. In Section 2.3.1, we validate ML-enabled drone-based observations against established ground-based methods by comparing behavior distributions and detection sensitivity across three approaches. (1) scan versus focal sampling (ground-based), (2) focal sampling (ground versus drone), and (3) manual versus ML annotations. In Section 2.3.2, we demonstrate the utility of our pipeline through three case studies, each aligned with key ecological tasks: time budget analysis, behavioral transitions, social interactions, habitat use, and group composition. These case studies illustrate how automated behavioral data extracted from drone footage can be applied to answer ecological questions at individual, group, and community levels. Together, this multi-level evaluation framework validates the technical reliability and ecological relevance of drone-based behavior monitoring systems, enabling scalable, non-invasive observation of wildlife in natural habitats.
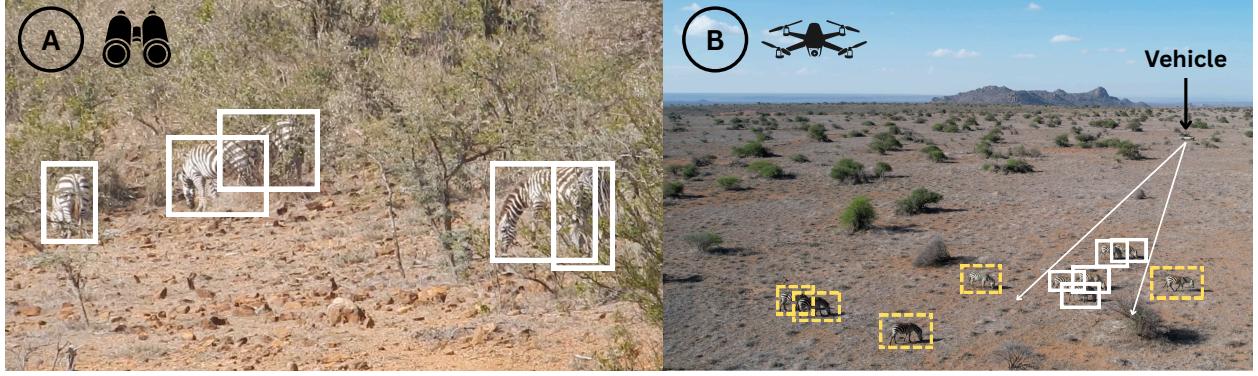
Figure 5: Comparison of ground-based (left) and drone-based (right) perspectives of the same group of plains zebras. While ground-based observations captured only five visible individuals (white solid outline) due to occlusion from vegetation, the drone view clearly revealed all ten (yellow dotted outline), illustrating the advantage of aerial monitoring for behavioral data collection.

2.3.1 Drone-based Observation Validation

We employed multiple analytical approaches to validate ML-enabled drone-based behavior observations against traditional ground-based scan and focal sampling methods. We aim to capture differences in out-of-field annotation protocol with the in-field data collection, since replaying videos typically allows annotators to capture behavioral transitions with greater precision. First, we compared the results from simultaneous ground-based scanning and focal sampling sessions that observed the same herd. Next, we compared the results from simultaneous ground-based focal sampling and drone-based focal sampling, observing the same individual during the same time. Finally, we compared the results of manually annotated drone footage with those of ML-annotated drone footage. To measure the agreement between the methods, we use confusion matrices to quantify the agreement between simultaneous observations, revealing method-specific detection capabilities for different behaviors.

To conduct ground-based observations in the field, observers worked in pairs. One partner observed the behaviors using a pair of binoculars and called out their observations. The other partner recorded the behaviors using the AnimalBehaviorPro app [Newton-Fisher, 2021]. Scan sampling captured data on the whole herd every two minutes, while focal sampling recorded the behavior of a single individual continuously. Animals were monitored simultaneously using both scan and focal sampling methods (Tab. 8). We used timestamp data captured with the AnimalBehaviorPro app to match the behavior observations temporally. To allow for comparison to the continuous focal sampling data, the discrete scan sampling observations were converted to continuous observations propagated across the subsequent two-minute period.

During select drone missions, we recorded zebra behaviors using focal sampling, allowing for a comparison of the behavior data gleaned from the drone footage with data observed from the ground (Tab. 9). We identified and matched the individual zebras between ground observations and drone footage using distinctive physical characteristics, including unique stripe patterns, scars, body size, and sex-specific features through standardized ground-level photography from multiple angles (Fig. 5). Digital photo catalogs were created for each observation session with GPS coordinates and

9

Data Harmonization Step	Description
Visibility Filtering	Excluded time periods when individuals were out of sight (ground focal) or out of frame/occluded (drone focal) to ensure valid time budget comparisons.
Behavioral Category Harmonization	Applied consistent behavioral definitions across both ground and drone datasets to allow for direct comparison.
Individual Matching Validation	Limited comparative analyses to individuals with clear, unambiguous identification across both methods.
Temporal Resolution Matching	Aggregated drone annotations to match the coarser temporal resolution of ground observations when required.

Table 3: Data harmonization steps used to align drone and ground-based behavioral datasets.

timestamps to facilitate matching across data collection platforms. Ground-based focal sampling and drone video collection were synchronized using GPS timestamps and detailed field notes recording drone flight initiation and termination times. Video annotations were temporally aligned with ground observations using metadata from both data streams. Only periods with complete temporal overlap between ground focal sampling and corresponding drone mini-scenes were included in comparative analyses. To ensure valid comparisons between drone and ground-based behavioral datasets, we applied four harmonization steps (Tab. 3). We first filtered out periods where individuals were not visible in either method. Then, we harmonized behavior categories, validated individual identity matches, and matched temporal resolutions by aggregating drone data. These steps minimized bias and improved alignment between datasets for robust comparative analysis.

In our previous work, we evaluated four ML models to automatically classify behavior from drone videos: I3D [Carreira and Zisserman, 2017], SlowFast [Feichtenhofer et al., 2019], X3D [Feichtenhofer, 2020], and UniformerV2 [Li et al., 2022] [Kholiavchenko et al., 2024a]. The X3D model consistently performed the best, so it was used here to evaluate the performance of automatic behavior recognition compared to manually annotated mini-scenes. The X3D model achieves a 91.32% F-1 score for the three most common behaviors, which include *Graze*, *Walk*, *Head Up*, but significantly lower performance on long-tailed tail classes, consisting of rarer behaviors (50.31% F-1). We used the X3D model to automatically annotate the behaviors used for Case Study 3 (Section 3.2.1).

2.3.2 Ecological Tasks and Case Studies

We focus on five broad ecological tasks: time budget analysis, behavioral transitions, social interactions, land use and movement patterns, and social group composition analysis [Hughey et al., 2018, Smith and Pinter-Wollman, 2021, Pedrazzi et al., 2025], summarized in Tab. 4. These tasks provide a multifaceted understanding of animal behavior by capturing not just how much time animals spend in different activities, but also how behaviors transition over time, vary across habitats, and are shaped by social context. Time budgets and transition matrices offer standardized metrics to quantify behavioral patterns and flexibility, while drone-enabled detection of interactions and movement pathways reveals dynamics often missed by ground observers. Together, these methods help ecologists link behavior to environmental and social drivers, which is essential for interpreting animal responses to habitat change, human disturbance, and conservation interventions.

Ecological Task	Description	Key Metrics / Comparisons
Time Budget Analysis	Calculates proportion of time spent in each behavioral state. Excludes periods when the animal was not visible.	• Time Budget$_i = \frac{\sum t_i}{\sum t_{\text{visible}}} \times 100\%$ • Comparison across drone vs. ground observations
Behavioral Transitions	Examines sequential behavior using transition matrices.	• $P(j\|i) = \frac{n_{ij}}{\sum_k n_{ik}}$ • Drone vs. ground transition patterns
Social Interactions	Detects social interactions based on spatial overlap and behavioral synchrony in drone footage.	• $> 50\%$ bounding box overlap for > 3 frames • (Optional) Behavioral categorized interactions (e.g., mutual-grooming)
Land Use and Movement Patterns	Uses GPS and visual classification to link behavior to habitat type.	• Habitat: open grassland, dense woodlands, mixed • Movement paths in annotated landscapes
Social Group Composition Analysis	Analyzes group size, demographics, and spatial structure based on high-resolution video.	• Age-sex categories (adult male/female, subadult, juvenile, infant) • Group size and spatial configuration

Table 4: Summary of behavioral and ecological analyses enabled by drone footage

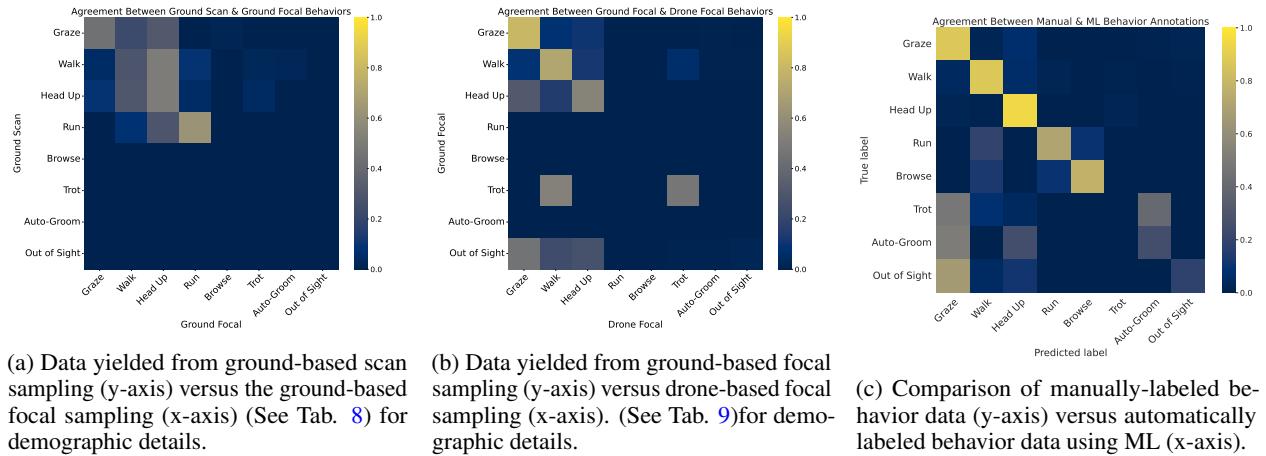| Case Study	Description	Ecological Tasks
Grevy's Zebra Landscape of Fear	Analyzed 368 mini-scenes of Grevy's zebra groups (2–8 individuals) across open grassland and bush habitats to assess anti-predator behavior. Behavioral labels were generated using a fine-tuned X3D model [Kholiavchenko et al., 2024c].	• Time budgets • Land use • Social group composition
Zebra State Shifts	Modeled behavior sequences from 748 mini-scenes to estimate transition probabilities between grazing, walking, trotting, and running.	• Behavioral transitions • Time budgets
Cross-Species Encounters	Detected 5,035 inter- and inter-species interactions among 16 individuals using the bounding box overlap method across 157 annotated mini-scenes.	• Social interaction detections • Social group composition

Table 5: Case studies illustrating use of analytical tasks for drone-based behavioral ecology

To demonstrate the analytical capabilities enabled by drone-based observation, we conducted three case studies leveraging the tasks (Tab. 5). Herbivores must balance competing needs to feed and avoid predation. These competing needs cause animals to avoid areas where food is abundant but predation risk is high, known as the landscape of fear phenomenon [Riginos, 2015]. We examine this phenomenon in our first case-study, *Grevy's Zebra Landscape of Fear*, which combines time budgets, habitat use, and social group composition to explore how landscape features and group structure influence perceived risk and anti-predator behavior. We model the likelihood of transitions between locomotor states in the *Zebra State Shift* case study to assess behavioral flexibility and alertness. Finally, the *Cross-Species Encounters* case study uses spatial proximity and behavioral synchrony to detect and classify social interactions, providing insight into mixed-herd dynamics that are difficult to observe from the ground.

3 Results

We evaluate the performance of our end-to-end system for automated wildlife behavior analysis through a combination of machine learning benchmarking, method comparison, and ecological case studies. In Section 3.1, we report the results of drone-based observations compared to traditional ground-based focal and scan sampling methods, and evaluate the performance of deep learning models for automatic behavior labeling. In Section 3.2, we report the findings of our three case studies. The results demonstrate the potential of drone-based, ML-enabled behavioral analysis to match and surpass traditional methods in accuracy, scale, and ecological insight.

3.1 Comparison of Behavior Data Collection Methods

(a) Data yielded from ground-based scan sampling (y-axis) versus the ground-based focal sampling (x-axis) (See Tab. 8) for demographic details.

(b) Data yielded from ground-based focal sampling (y-axis) versus drone-based focal sampling (x-axis). (See Tab. 9)for demographic details.

(c) Comparison of manually-labeled behavior data (y-axis) versus automatically labeled behavior data using ML (x-axis).

Figure 6: Comparison of behavior data collected using different methodologies applied to the same individual animal during the same time. Drone-based methods enable concurrent focal sampling of all individuals in a herd, and ML models can automatically annotate behaviors with high accuracy–particularly for common behaviors–closely matching manual and ground-based observations while capturing events often missed from the ground. Behavior categories are arranged in descending order by frequency of occurrence. (a) Ground-based scan and focal sampling show stronger agreement for common behaviors, but scan sampling fails to capture rare behaviors. (b) Ground-based focal sampling matches drone-based focal sampling, but misses behaviors when the animals are out of sight from the ground-based observers. (c) Automatic ML behavior annotation shows strong performance compared to manual annotations, especially for common behaviors.

To evaluate the effectiveness of drone-based behavioral sampling, we perform three key comparisons: (1) ground-based scan versus ground-based focal sampling (Section 3.1.1), (2) ground-based focal versus drone-based focal sampling (Section 3.1.2), and (3) manual versus automated behavior annotations of drone-based observations (Section 3.1.3) [Kline et al., 2025b]. We use confusion matrices to compare concordance between these methodologies (Fig. 6). Confusion matrices are tabular summaries that quantify agreement between two classification methods by reporting the frequency with which each behavior label from one method matches or differs from the corresponding label in the other. Our comparison confirmed that scan sampling captures concurrent behavior of the entire herd, but captures fewer behavior categories overall compared to individual focal sampling. Drone-based focal sampling reduced out-of-sight time by 14.7% and captured more behavioral transitions compared to ground-based focal sampling. A comparison of manual versus automated drone annotations showed high agreement for dominant behaviors, confirming the utility

of automated pipelines while revealing systematic errors in rare or subtle behaviors. Together, these comparisons demonstrate that drone-based sampling with machine learning provides a powerful complement to traditional methods, combining the comprehensive coverage of scan sampling with the fine granularity of focal sampling, while enabling scalable, automated behavioral analysis at high resolution.
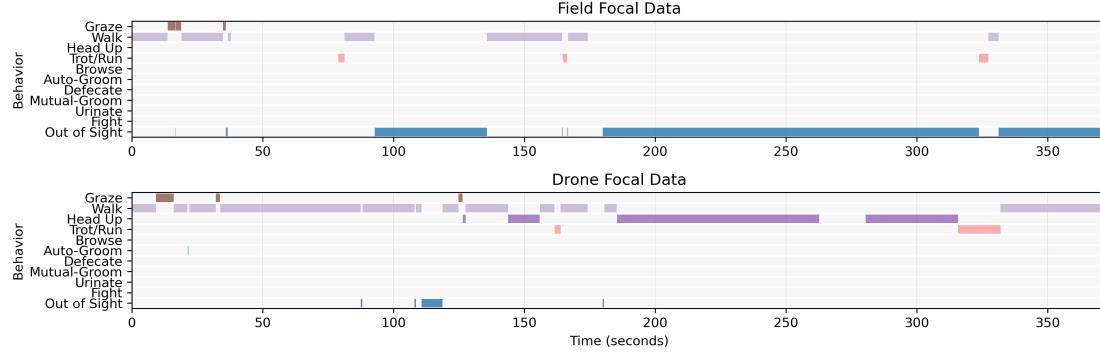
3.1.1 Comparison of ground-based focal and scan sampling

Simultaneous scan and focal sampling in the field reveal key differences in behavioral granularity, illustrated in Fig. 6a. Focal sampling captured significantly more behavioral categories than scan sampling (focal: mean $= 6.2 \pm 1.4$ categories; scan: mean $= 3.1 \pm 0.8$ categories; paired t-test: $t_5 = 4.73$, $p = 0.005$). Focal sampling captured rare and transient behaviors, including *Trot*, *Auto-Groom*, and *Fight*, which were often missed by the coarser scan sampling approach. Scan sampling captures *Graze*, *Walk*, *Head Up* and *Browse*. When scan sampling captures *Graze*, focal sampling revealed this included *Graze* (44%), *Walk* (25%), *Head Up* (28%), and *Auto-Groom* (2%).
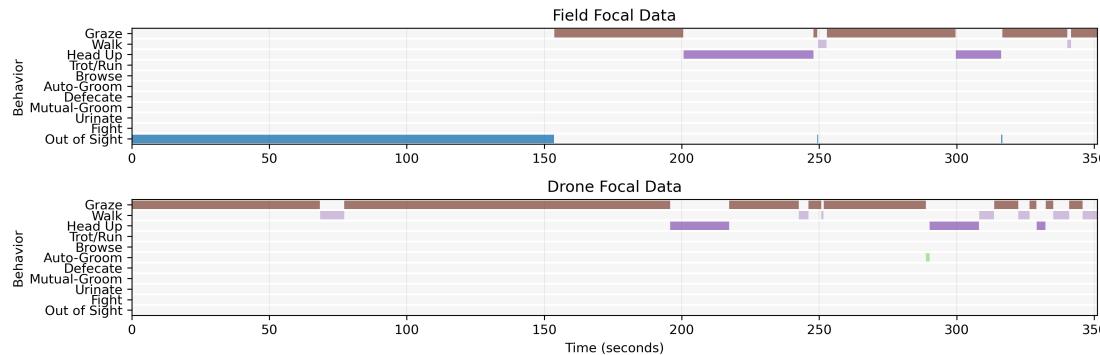
3.1.2 Comparison of ground-based focal and drone focal sampling

We compare ground-based focal sampling with drone-based focal sampling using temporally aligned video segments from four uniquely identifiable zebras (Tab. 9). Temporal overlap for each sampling session ranged from 193.9 to 405.8 seconds for a total of 22 minutes. Drone-based sampling was found to offer substantial advantages in visibility and coverage: ground observers lost sight of focal animals for an average of 23.4% of the observation time, whereas drones experienced only 8.7% of occlusion. Ground-based and drone-based focal sampling were largely in agreement for common behaviors, *Graze, Walk, Head Up* (Fig. 6b). Behavior instances labeled *Trot* in ground-based sampling was frequently labeled as *Walk* in the drone focal sampling. Ground sampling suffered from extended *Out of Sight* periods, while drone sampling experienced shorter *Out of Sight* periods. When the animals were out of sight of the ground observers, the drone captured *Graze* (46%), *Walk* (24%), and *Head Up* (26%) behaviors. When both methods successfully capture behavior, agreement was highest for common behaviors: *Graze* (80% concordance), *Walk* (71% concordance), but lower for brief behaviors: *Head Up* (54% concordance), *Trot/Run* (47% concordance).
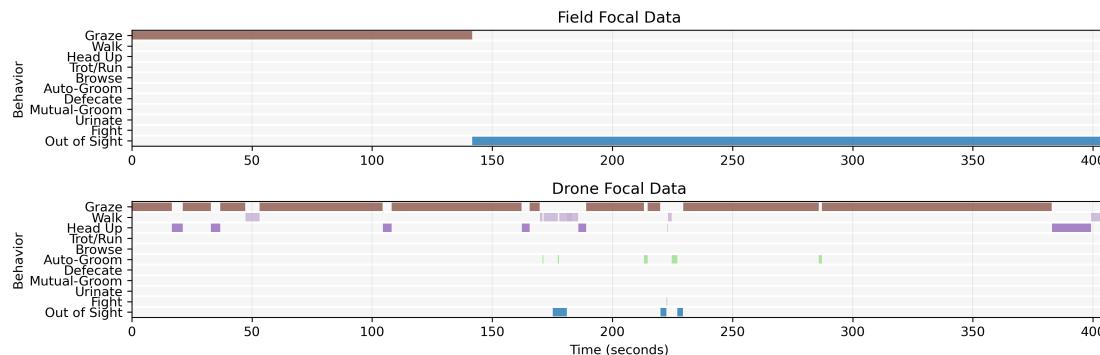
Gantt chart comparisons further demonstrated that drones captured more behavioral transitions and continuous observation windows, providing richer behavioral sequences, as shown in the Gantt charts in Fig. 7. Drone footage consistently captures more behavior transitions than ground observations. For Zebra A, the drone-based focal sampling largely matches the ground-based focal sampling for the first 100 seconds, until the zebra is out of sight of the ground observers. For Zebra B, the individual was out of sight of the ground-based observers for the first 150 seconds, while the drone captured instances of *Graze* and *Walk*. Once Zebra B was in view of the ground observers, the behaviors captured were the same as those of the drone-based observers. Notably, the drone-based observations captured more behavioral transitions and a brief instance of *Auto-Groom*. Similarly, for Zebra C, the drone-based focal observations captured instances of *Graze*, *Walk*, and *Head Up*, with rapid behavioral changes, while ground-based observations only captured *Graze* during the same time. After 150 seconds, Zebra C was out of sight of the ground-based observers,

(a) Zebra A Gantt Chart

(b) Zebra B Gantt Chart

(c) Zebra C Gantt Chart

(d) Zebra D Gantt Chart

Figure 7: Temporal comparison of behavior captured from ground-based focal sampling versus drone-based focal sampling of four zebras (Tab. 9). Drone observations yielded finer granularity and have far fewer instances of *Out of Sight* compared to ground-based observations. Behavior categories are organized in descending order of occurrence.

14

while the animal stayed in view of the drone for 400 seconds, with a few brief instances of occlusion. For Zebra D, the drone-based focal sampling agreed with the ground-based focal observations, although the drone captured the behavior transition from *Head Up* to *Walk* sooner than the ground observers. After 30 seconds, Zebra D was out of sight of the ground-based observers, while the animal stayed in view of the drone for 200 seconds. Note, this small sample size ($n = 4$) precludes statistical inference and serves only for qualitative method comparison. Future studies require larger sample sizes for robust statistical analysis.

3.1.3 Comparison of manually-annotated and automatically-annotated behaviors

We assess the performance of automated behavior recognition using the best-performing X3D model against expert manually annotated drone footage. The X3D model is evaluated using 25% of the mini-scenes for testing, while the remaining 75% were used for training [Kholiavchenko et al., 2024b]. The confusion matrix (Fig. 6c) compares manual behavior annotations (rows) to predictions generated by the X3D model (columns) across the eight behavioral categories. The X3D model demonstrates high accuracy for core locomotor and foraging behaviors, correctly classifying *Graze*, *Walk*, and *Head Up* with accuracies of 87%, 87%, and 94%, respectively. Moderate performance is observed for *Trot* (71%) and *Run* (77%), although some misclassification occurs between these and *Walk*, likely due to overlapping kinematic features. Less frequent or more visually subtle behaviors such as *Browse* and *Auto-Groom* exhibit higher rates of confusion, particularly with grazing and scanning postures. Misclassifications were most common between visually similar categories and in occluded frames, which were often predicted as *Graze*. For example, *Browse* is frequently predicted as *Graze* (48%) and *Auto-Groom* is often misclassified as *Graze* (50%) or *Head Up* (25%). The *Out of Sight* category is partially misclassified as *Graze* (66%), suggesting that occluded frames may share visual characteristics with feeding behaviors.

3.2 Ecological Case Studies

We demonstrate the capabilities of our behavioral analysis pipeline through three ecological case studies. For Case Study 1, *Grevy's Zebra Landscape of Fear* (Section 3.2.1), we found that vigilance declined with increasing herd size, consistent with the group-vigilance trade-off, but habitat type did not explain variation in behavior, unlike Plains zebras. For Case Study 2, *Zebra State Shifts* (Section 3.2.2), zebras showed strong behavioral inertia, with infrequent but detectable transitions to alert behaviors (e.g., *Walk* to *Run*). These patterns offer a foundation for predictive, behavior-adaptive remote sensing systems. For Case Study 3, *Cross-Species Encounters* (Section 3.2.3), we analyzed the spatial overlap among Grevy's zebras, Plains zebras, and reticulated giraffes in a mixed-species herd. Intra-species proximity was highest for zebras, while interactions between zebras and giraffes were not detected, indicating spatial segregation between species despite sharing a grazing area.

Behavior	R^2	Habitat Effect (Open vs Closed)	Herd Size Effect (Small vs Large)
Walk	0.402 ***	0.087†	0.307 ***
Head Up	0.250 ***	-0.008	0.321 ***
Graze	0.562 ***	-0.076	-0.641 ***
Other	0.002	-0.003	0.013

Table 6: Summary of linear regression results examining habitat and herd size effects on Grevy's zebra behaviors. The reported coefficients represent a change in the proportion of time spent on behavior. Small herds significantly increased walking, vigilance (head up), and decreased grazing compared to large herds (all p < 0.001), while habitat type showed minimal effects on behavior, with herd size explaining 25-56% of behavioral variation depending on the activity. Significance levels: $***p < 0.001; ** \ p < 0.01; *p < 0.05; \ p < 0.10$. Reference categories: Closed habitat, Large herds.

3.2.1 Case Study 1: Grevy's Zebras Landscape of Fear

We analyze anti-predator strategies in Grevy's zebras using 83 mini-scenes [Kline et al., 2024b, 2025e]. We selected mini-scenes from six representative sessions, containing a mix of herd sizes and habitat types (Tab. 10). To ensure sufficiently long behavior sequences for analysis, we selected mini-scenes ≥ 60 seconds. We classify the mini-scenes by habitat type based on vegetation density by visually examining the mini-scene footage to classify habitat type as open (little to no vegetation) or closed (dense vegetation). Examples of habitat categories are illustrated in Fig. 8. The median herd size is three individuals; herds ≤ 3 are categorized as small, herds > 3 are large. For each mini-scene, the time budget was calculated as a proportion of *Walk*, *Head Up*, *Graze*, and *Other*. We use *Head Up* behaviors to indicate instances of vigilance, where animals survey their habitat for potential risks.

We conducted linear regression analyses to examine the effects of habitat type (closed vs. open) and herd size (large vs. small) on the four behavioral categories, summarized in Tab. 6. Categorical predictor variables were dummy coded with closed habitat and large herds serving as reference categories. For each behavior, we fitted separate multiple linear regression models using ordinary least squares estimation. Model assumptions were verified through residual analysis, and statistical significance was assessed using t-tests with $\alpha = 0.05$. We calculated 95% confidence intervals for all coefficients and quantified effect sizes using Cohen's f^2, visualized in Fig. 9. We also tested for potential interaction effects between habitat and herd size but found no significant interactions, confirming that the effects of these variables are additive rather than synergistic. The linear regression models revealed substantial variation in how social and environmental context influence different behaviors, with model performance (R^2) ranging from negligible for *Other* behaviors ($R^2 = 0.002$) to strong for *Graze* ($R^2 = 0.562$). Herd size emerged as the dominant predictor across all behaviors, showing significant effects for *Walk* ($\beta = 0.307, p < 0.001$), *Head Up* ($\beta = 0.321, p < 0.001$), and *Graze* ($\beta = -0.641, p < 0.001$), with small herds consistently showing higher activity levels except for *Graze*. Habitat effects were generally weaker and non-significant, with only marginal significance for walking behavior ($\beta = 0.087, p = 0.052$). The strongest and most predictable response was *Graze*, which decreased substantially in small compared to large herds, while vigilance and locomotory behaviors increased.

(a) Example of closed habitat with dense vegetation (Session 4 from Tab. 10), [Kline et al., 2024a].

(b) Example of open habitat with with little vegetation (Session 5 from Tab. 10), [Kline et al., 2025e].

Figure 8: Example of Closed versus Open habitats analyzed for Case Study 1: *Grevy's Zebras Landscape of Fear*.
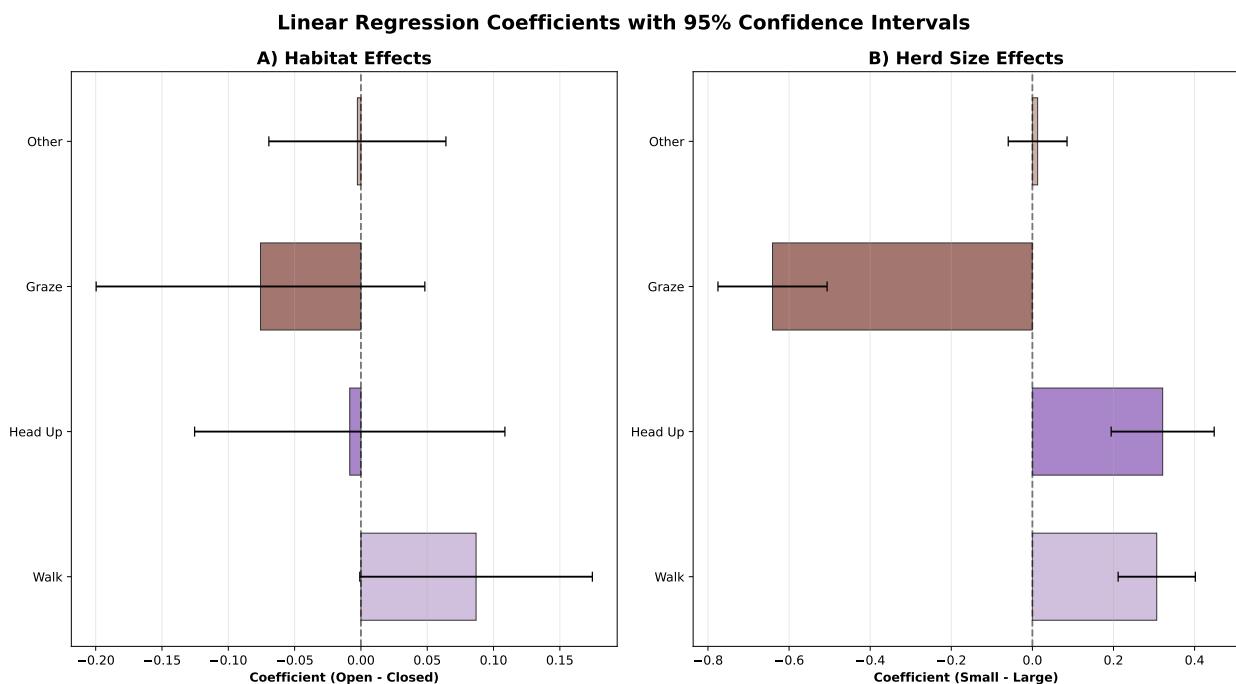

Figure 9: Linear regression coefficients with 95% confidence intervals showing the effect of habitat (Open vs Closed) and herd size (Small vs Large) on four behaviors, analyzed for Case Study 1: *Grevy's Zebras Landscape of Fear*. Larger herds allow Grevy's zebras to spend more time grazing and less time scanning or walking, confirming the classic group – vigilance trade-off. The habitat panels show that any raw differences among closed, mixed, and open bush are largely explained by the bigger groups that congregate in dense cover – once herd size is accounted for, habitat adds little additional effect.

3.2.2 Case Study 2: Zebra State Shifts

We analyzed 748 mini-scenes [Kline et al., 2024b]. The mini-scenes total 2 hours 30 minutes, containing zebra locomotion behaviors of *Graze*, *Walk*, *Trot*, and *Run*. We quantify the likelihood of zebras transitioning between behaviors (Fig. 10) by normalizing the transitions by the count of occurrences, down-sampled every 10 seconds. Individuals are most likely to remain in the same behavioral state, with high self-transition probabilities for *Graze* (91.1%), *Walk* (83.0%), and *Run* (68.4%). Previous instances of *Trot* were most likely to transition to *Walk* (53.4%). Transitions from relaxed states to more alert behaviors are also evident: for instance, individuals in a *Walk* state transition to *Trot* 2.6% of the time and to *Run* 2.0% of the time, while *Trot* transitions to *Run* with a relatively high probability of 14.3%. These transition dynamics provide insight into how animals escalate movement in response to internal or external stimuli and

Figure 10: Transition matrix for zebra locomotion behaviors, ordered by least to greatest velocity, for Case Study 2: *Zebra State Shifts*.

offer a foundation for modeling behavioral state changes. Understanding these patterns is critical for developing behavior-adaptive autonomous drones that can anticipate transitions, such as from grazing to fleeing, and adjust drone flight paths accordingly to minimize disturbance while maximizing data capture.
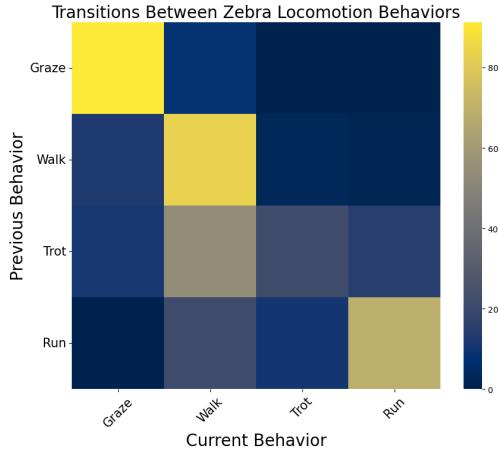
3.2.3 Case Study 3: Cross-Species Encounters

A mixed species herd of Grevy's zebras, Plains zebras, and giraffes was observed grazing together, illustrated in Fig. 11) [Kline et al., 2025e]. From the drone videos, 157 mini-scenes were generated and automatically labeled with behaviors using the KABR X3D model [Kholiavchenko et al., 2024c]. To assess patterns of spatial proximity, we computed normalized bounding box overlap counts between individuals of the same and different species (Tab. 7). Normalization was performed by dividing the total number of overlaps by the number of possible pairs for each species combination. Both zebra species showed higher overlap compared to giraffes. Grevy's zebras exhibited a high normalized overlap rate (87.93 overlaps per pair), suggesting frequent and sustained close proximity among group members. Plains zebras showed the highest normalized value (93.00), though this reflects a single dyad and should be interpreted with caution due to the small sample size. Giraffes had a moderate normalized overlap rate (26.00), consistent with a more spatially dispersed grouping structure. Cross-species overlaps were rare. The normalized overlap between Grevy's and plains zebras was low (1.27), and no overlaps were observed between giraffes and zebras (0.00), indicating strong spatial segregation between species in the observed dataset. Overall, these patterns suggest that intra-species proximity is most prominent in Grevy's zebras, with limited interaction between species.

Species Pair	Overlap Count	Possible Pairs	Normalized Overlap
Grevy's Zebra – Grevy's Zebra	4836	55	87.93
Plains Zebra – Plains Zebra	93	1	93.00
Giraffe – Giraffe	78	3	26.00
Grevy's Zebra – Plains Zebra	28	22	1.27
Giraffe – Plains Zebra	0	6	0.00
Giraffe– Grevy's Zebra	0	33	0.00

Table 7: Bounding box overlap counts and normalized values for intra- and inter-species interactions for Case Study 3: *Cross-Species Encounters.*. Normalized values represent average overlaps per possible pair. Same species: $n(n-1)/2$ possible pairs. Cross-species: $n_1 \times n_2$. 11 Grevy's zebras, 3 reticulated giraffes, and 2 Plains zebras.
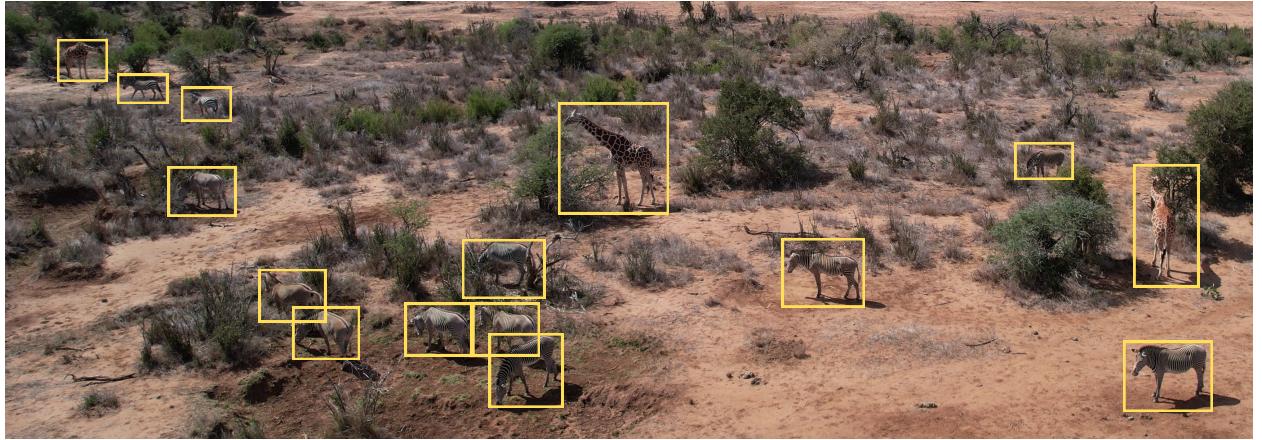

Figure 11: Image from mini-scene of mixed species herd consisting of 11 Grevy's zebras (center), 2 Plains zebras (left), and 3 reticulated giraffes (left, center, right) analyzed for Case Study 3: *Cross-Species Encounters* [Kline et al., 2025e].

4 Discussion

Our results demonstrate that drone-based behavioral observation, augmented by machine learning, offers a scalable, non-invasive, and ecologically valid method for studying animal behavior in natural environments. In Section 4.1, we discuss our proposed hybrid semi-automated approach for video processing to balance scalability with accuracy. In Section 4.2, we discuss how drone-based observation overcomes the fundamental limitations of traditional ground-based methods by providing simultaneous focal sampling of multiple individuals with superior temporal resolution and fewer observation gaps, bridging the trade-off between behavioral detail and group-level context. In Section 4.3, we discuss the ecological insights derived from the case studies, including species-specific vigilance patterns in Grevy's zebras, structured behavioral state transitions, and spatial segregation in mixed-species herds. Together, these findings suggest that drone-based automated behavioral analysis provides a valuable complement to traditional methods, offering researchers new opportunities to study animal behavior at larger scales while acknowledging the ongoing need for methodological refinement and human oversight.

4.1 Video Processing Pipeline

Our pipeline processes drone video to enable scalable, semi-automated analysis of animal behavior in ecological field studies. It integrates automated machine learning with targeted manual annotation, balancing scalability with accuracy. While fully manual annotation yields highly detailed behavioral records, the time investment becomes prohibitive at scale. Instead, we propose a hybrid workflow in which machine learning models handle dominant behaviors, while human annotators validate edge cases and review rare or ambiguous behaviors.

We designed our pipeline to be modular, allowing users to easily integrate emerging machine learning models. Our X3D model performed well in classifying dominant behaviors such as *Walk, Graze*, and *Head* Up, achieving over 95% mean average precision (mAP) and a 91.4% F1 score for these classes. However, its performance declined for rare or visually ambiguous behaviors, such as *Run* and *Auto-groom*. This outcome reflects the long-tail problem in machine learning, where rare behaviors suffer from limited training examples and, as a result, lower predictive accuracy. Despite this limitation, the X3D model is sufficient for identifying broad behavioral trends with a per-instance (micro) average accuracy of 86.7% [Kholiavchenko et al., 2024b]. It is also sufficient for automatically estimating coarse-grained time budgets, such as those used for Case Study 1 (Section 3.2.1). Automated methods struggle with high-resolution or rare-behavior analyses, motivating the need for human-in-the-loop processing pipelines to capture these instances. When integrating new models into the pipeline, we recommend setting model confidence thresholds (e.g., > 0.8) to enhance precision, with the understanding that this may reduce recall. For studies requiring high accuracy or detailed behavioral timing, manual review remains essential.

Manual annotation provides fine-grained, accurate behavioral labels but is labor-intensive. Based on our annotation rate model ($T_{\text{annotation}} = 1.5nt$ for n individuals over t seconds), a 10-minute video with three individuals requires approximately 15 minutes of annotation time per animal, for a total of 45 minutes to process the entire scene. This annotation burden scales linearly with group size and video length, making fully manual approaches impractical for large datasets. Our project included 969 annotated mini-scenes, which would have required hundreds of hours of effort without automation. The semi-automated workflow we developed offers a practical compromise between scale and precision. Automated classification handles the bulk of labeling for frequent behaviors, while human annotators focus on reviewing edge cases, validating model outputs, and labeling rare behaviors. To ensure reliability across annotators and datasets, we emphasize the importance of standardized annotation protocols and regular calibration sessions to maintain inter-annotator consistency.

The effectiveness of the pipeline depends not only on model accuracy but also on data quality and experimental design. Common technical challenges include object detection failures when animals are obscured by vegetation or overlapping conspecifics, leading to *Out of Sight* segments. In addition, video resolution limits can constrain the identification of individuals and subtle behaviors, particularly at greater distances or flight altitudes. Careful flight planning is essential to optimize resolution and field of view for behavioral monitoring. Finally, behavior definitions must be operationalized clearly to reduce inconsistencies, especially for transitional or compound behaviors.

Our results demonstrate that semi-automated analysis offers a powerful path forward for behavioral ecology, enabling analysis at scales that would otherwise be infeasible. While current model performance imposes constraints–particularly for rare behaviors–hybrid approaches that combine machine learning with human oversight provide a scalable and accurate framework. Future improvements in model architecture, training data, and multimodal integration will continue to expand the ecological questions that automated pipelines like `kabr-tools` can address.

4.2 Drone-based Behavioral Observations

Our systematic comparison of behavioral sampling methods reveals distinct advantages and limitations of traditional ground-based scan and focal sampling. Focal sampling consistently captures more detailed behavioral information than scan sampling, detecting rare behaviors (*Auto-groom, Fight*) that are missed by instantaneous scan observations. However, scan sampling provides the critical advantage of capturing behavioral data from multiple individuals under identical environmental and social contexts, eliminating the temporal confounds inherent in sequential focal observations. Drone-based observations bridge these methodological gaps by fine-grained focal sampling of multiple individuals under identical conditions. Compared to ground-based methods, drones experienced fewer observation gaps, captured more behavioral transitions, and provided richer behavioral sequences, especially for brief behaviors like vigilance. We did not conduct simultaneous drone missions and scan sampling sessions. However, this would be a valuable contribution for future work, particularly in comparing tradeoffs in cost and labor for each methodology and granularity of behavior data captured. The superior temporal resolution of drone annotations stems from two key factors: (1) the aerial perspective minimizes visual obstructions, and (2) post-hoc video analysis allows annotators to pause and rewind footage, ensuring precise alignment of behavioral transitions with video frames. This precision advantage is particularly evident for brief behaviors like vigilance responses, where ground observers may introduce recording delays that artificially inflate behavior durations.

4.3 Ecological Case Studies

To demonstrate the capabilities of our `kabr-tools` pipeline, we applied it to three ecological case studies: anti-predator behavior, behavioral state transitions, and group structure in mixed-species herds (Section 4.3.1). These case studies underscore the types of ecological insights that drone-based automated analysis enables at scale (Section 4.3.2).

4.3.1 Case Studies Findings

Case Study 1: Grevy's Zebra Landscape of Fear. We analyzed 86 mini-scenes of Grevy's herds ranging from two to seven individuals across open and closed habitats. We found that individual vigilance decreased significantly with increasing group size, consistent with the group-vigilance trade-off [Rubenstein, 1986]. Notably, habitat type had no additional effect on vigilance after controlling for group size–unlike in plains zebras, where habitat remains a significant factor [Chen et al., 2021]. This novel finding reveals a fundamental difference in risk assessment by zebra species. The tradeoff between survival and nutrition significantly impacts future reproduction [Rubenstein, 1986]. The difference in

vigilance trade-offs between Grevy's and plains zebras likely stems from social structure, specifically the differences in the fluidity of each species' social system [Rubenstein, 2010]. Plains zebras exhibit more rigid groupings, while Grevy's zebras form flexible, fission-fusion groups, enabling fine-tuned responses to risk through dynamic group size adjustment. While plains zebras live in fixed closed membership groups that can join and leave herds, Grevy's zebra individuals are freer to move in and out of herds as conditions change, so responding to changes in risk can happen more quickly, making individuals less risk-averse than individual plains zebras, their close evolutionary kin. However, it should be noted that these conclusions are drawn from five sampling sessions representing 20 individuals. Further data collection should be conducted to explore this relationship further. Grevy's zebras are endangered, with less than two thousand individuals remaining in the wild [Rubenstein et al., 2016], so understanding factors driving their future reproduction is vital for protecting and preserving these at-risk populations.

Case Study 2: Zebra State Shifts. We utilized 748 mini-scenes of behavioral sequences to construct a transition matrix between locomotor states (*Graze*, *Walk*, *Trot*, *Run*). These revealed strong behavioral inertia and structured escalation patterns, where transitions from relaxed states (e.g., *Graze*) to alert states (e.g., *Run*) occurred in measurable steps. This structure provides predictive value for behavior-adaptive autonomous sensing, offering a potential feedback loop between real-time drone behavior and animal response.

Case Study 3: Cross-Species Encounters. We examined spatial interactions in a mixed-species herd using bounding box overlap metrics. Intra-species proximity was strongest in both Grevy's and Plains zebras, while inter-species interactions were rare despite shared space. These findings indicate spatial segregation across species and demonstrate the utility of drone imagery for quantifying social cohesion and species-level group structure in dynamic field settings.

4.3.2   Implications of Case Studies

Across the three case studies, our framework demonstrated three key strengths: scale, simultaneity, and standardization. We analyzed 969 mini-scenes – far beyond what traditional methods could feasibly capture or annotate. This scale enables robust statistical inference on behavior and interactions across ecological gradients. Drone-based observation captures all individuals in a group under identical environmental and social conditions, providing superior simultaneity. This eliminates temporal confounds common in sequential focal sampling and was essential for detecting group-level behavioral patterns. Finally, automated labeling ensures consistent behavioral definitions across observations, minimizing observer bias and facilitating standardized comparisons across species and sites.

While our results reveal important ecological insights, they also highlight the current limitations of automated behavioral classification. With a classification accuracy of approximately 91% for the three behavior classes used for *Grevy's Landscape of Fear* case study, we can be reasonably certain the automatically generated time budgets are accurate. However, the automatic recognition of rare behaviors and precise timing remains challenging. Despite this, the strength of statistical patterns emerging from large sample sizes suggests that automated annotation is already a powerful tool for broad ecological inference. Future work will focus on improving behavior classification through model fine-tuning, incorporating multimodal data, and human-in-the-loop workflows to expand the range of ecological questions that can

be addressed. Ultimately, this work demonstrates that drone-based automated analysis is not merely a faster alternative to manual methods, but a transformative tool enabling new modes of behavioral and ecological inquiry.

5   Conclusion

Our work demonstrates that the integration of drone technology, computer vision, and machine learning creates new possibilities for ecological research while highlighting the importance of rigorous validation and performance assessment. The success of this approach depends on continued collaboration between ecological and computational disciplines, with particular attention to developing tools that serve genuine research needs while maintaining the highest standards of scientific rigor and animal welfare. The kabr-tools framework and associated dataset provide a foundation for advancing automated wildlife behavioral analysis. Realizing the full potential of these approaches requires sustained effort in model development, standardization, and building collaborative frameworks, rooted in application-driven machine learning [Rolnick et al., 2024]. The future of ecological informatics lies in creating seamless integration between field observations, automated data collection, and predictive modeling to support evidence-based conservation and management decisions.

Acknowledgments

6   Data Availability Statement

All code, annotated datasets, and documentation are publicly available at https://github.com/Imageomics/kabr-tools. Documentation in Github includes annotation guidelines and best practices for managing large-scale ecological video datasets. The data used for analysis is available on HuggingFace and includes full-length videos [Kline et al., 2025c], worked examples [Kline et al., 2025e], and methodology comparison [Kline et al., 2025b].

The data was collected at the Mpala Research Centre in Kenya, in accordance with Research License No. NA-COSTI/P/22/18214. The data collection protocol adhered strictly to the guidelines set forth by the Institutional Animal Care and Use Committee under permission No. IACUC 1835F.

References

Saadia Afridi, Kasper Hlebowicz, Dylan Cawthorne, and Ulrik Lundquist. Unveiling the impact of drone noise on wildlife: A crucial research imperative. pages 1409–1416, 06 2024. doi: 10.1109/ICUAS60882.2024.10557094.

Jeanne Altmann. Observational study of behavior: Sampling methods. *Behaviour*, 49(3–4):227–266, January 1974. ISSN 0005-7959, 1568-539X. doi: 10.1163/156853974X00534.

Katherine R. Amato, Sarie Van Belle, and Brianna Wilkinson. A comparison of scan and focal sampling for the description of wild primate activity, diet and intragroup spatial relationships. *Folia Primatologica; International Journal of Primatology*, 84(2):87–101, 2013. ISSN 1421-9980. doi: 10.1159/000348305.

David J. Anderson and Pietro Perona. Toward a science of computational ethology. *Neuron*, 84(1):18–31, October 2014. ISSN 0896-6273. doi: 10.1016/j.neuron.2014.09.005.

Daniel Axford, Ferdous Sohel, Mathew Vanderklift, and Amanda Hodgson. Collectively advancing deep learning for animal detection in drone imagery: Successes, challenges, and research gaps. *Ecological informatics*, page 102842, 2024.

M. Bateson and P. Martin. *Measuring behaviour: an introductory guide*. Cambridge university press, 2021.

Marc Besson, Jamie Alison, Kim Bjerge, Thomas E Gorochowski, Toke T Høye, Tommaso Jucker, Hjalte MR Mann, and Christopher F Clements. Towards the fully automated monitoring of ecological communities. *Ecology Letters*, 25(12):2753–2775, 2022.

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*, pages 6299–6308, 2017.

Alex Hoi Hang Chan, Prasetia Putra, Harald Schupp, Johanna Köchling, Jana Straßheim, Britta Renner, Julia Schroeder, William D Pearse, Shinichi Nakagawa, Terry Burke, et al. Yolo-behaviour: A simple, flexible framework to automatically quantify animal behaviours from videos. *bioRxiv*, pages 2024–08, 2024.

Anping Chen, Leslie Reperant, Ilya R. Fischhoff, and Daniel I. Rubenstein. Increased vigilance of plains zebras (equus quagga) in response to more bush coverage in a kenyan savanna. *Climate Change Ecology*, 1:100001, July 2021. ISSN 2666-9005. doi: 10.1016/j.ecochg.2021.100001.

Evangeline Corcoran, Megan Winsen, Ashlee Sudholz, and Grant Hamilton. Automated detection of wildlife using drones: Synthesis, opportunities and constraints. *Methods in Ecology and Evolution*, 12(6):1103–1114, 2021. ISSN 2041-210X. doi: 10.1111/2041-210X.13581.

CVAT.ai. Computer Vision Annotation Tool (CVAT), November 2023. URL https://github.com/cvat-ai/cvat.

Isla Duporge, Marcus P Spiegel, Eleanor R Thomson, Tatiana Chapman, Curt Lamberth, Caroline Pond, David W Macdonald, Tiejun Wang, and Holger Klinck. Determination of optimal flight altitude to minimise acoustic drone disturbance to wildlife using species audiograms. *Methods in Ecology and Evolution*, 12(11):2196–2207, 2021.

Isla Duporge, Maksim Kholiavchenko, Roi Harel, Scott Wolf, Daniel I Rubenstein, Margaret C Crofoot, Tanya Berger-Wolf, Stephen J Lee, Julie Barreau, Jenna Kline, et al. Baboonland dataset: Tracking primates in the wild and automating behaviour recognition from drone videos: I. duporge et al. *International Journal of Computer Vision*, pages 1–12, 2025.

Reinoud Elhorst, Martyna Syposz, and Katarzyna Wojczulanis-Jakubas. Behave-facilitating behaviour coding from videos with ai-detected animals. *Ecological Informatics*, 87:103106, 2025.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 6201–6210, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi: 10.1109/ICCV.2019.00630. URL https://ieeexplore.ieee.org/document/9008780/.

Lacey F Hughey, Andrew M Hein, Ariana Strandburg-Peshkin, and Frants H Jensen. Challenges and solutions for studying collective animal behaviour in the wild. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1746):20170005, 2018.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Yolo by ultralytics, January 2023. URL https://github.com/ultralytics/ultralytics.

Maksim Kholiavchenko, Jenna Kline, Maksim Kukushkin, Otto Brookes, Sam Stevens, Isla Duporge, Alec Sheets, Reshma R. Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, Jackson Miliko, Eduardo Bessa, Majid Mirmehdi, Thomas Schmid, Tanya Berger-Wolf, Daniel I. Rubenstein, Tilo Burghardt, and Charles V. Stewart. Deep dive into kabr: a dataset for understanding ungulate behavior from in-situ drone video. *Multimedia Tools and Applications*, December 2024a. ISSN 1573-7721. doi: 10.1007/s11042-024-20512-4. URL https://doi.org/10.1007/s11042-024-20512-4.

Maksim Kholiavchenko, Jenna Kline, Michelle Ramirez, Sam Stevens, Alec Sheets, Reshma Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, Jackson Miliko, Eduardo Bessa, Isla Duporge, Tanya Berger-Wolf, Daniel Rubenstein, and Charles Stewart. Kabr: In-situ dataset for kenyan animal behavior recognition from drone videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 31–40, January 2024b.

Maksim Kholiavchenko, Maksim Kukushkin, Otto Brookes, Jenna Kline, Sam Stevens, Isla Duporge, Alec Sheets, Reshma R. Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, Jackson Miliko, Eduardo Bessa Mirmehdi, Thomas Schmid, Tanya Berger-Wolf, Daniel I. Rubenstein, Tilo Burghardt, and Charles V. Stewart. Kabr model, 2024c. URL https://huggingface.co/imageomics/x3d-kabr-kinetics.

Jenna Kline, Maksim Kholiavchenko, Tanya Berger-Wolf, Charles V. Stewart, and Christopher Stewart. Integrating biological data into autonomous remote sensing systems for in situ imageomics: A case study for kenyan animal

behavior sensing with unmanned aerial vehicles (uavs). In *Proceedings of the First Workshop on Imageomics: Discovering Biological Knowledge from Images using AI, held as part of AAAI 24*, 2024a.

Jenna Kline, Maksim Kholiavchenko, Otto Brookes, Tanya Berger-Wolf, Charles V. Stewart, and Christopher Stewart. Kabr behavior and telemetry dataset, 2024b. URL https://imageomics.github.io/KABR/.

Jenna Kline, Saadia Afridi, Edouard GA Rolland, Guy Maalouf, Lucie Laporte-Devylder, Christopher Stewart, Margaret Crofoot, Charles V Stewart, Daniel I Rubenstein, and Tanya Berger-Wolf. Studying collective animal behaviour with drones and computer vision. *Methods in Ecology and Evolution*, 2025a. URL https://doi.org/10.1111/2041-210X.70128.

Jenna Kline, Maksim Kholiavchenko, Michelle Ramirez, Sam Stevens, Reshma Ramesh Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, Jackson Miliko, Isla Duporge, Neil Rosser, Chuck Stewart, Tanya Berger-Wolf, and Daniel Rubenstein. kabr-tools methodology dataset, 2025b. URL https://huggingface.co/datasets/imageomics/kabr-methodology.

Jenna Kline, Maksim Kholiavchenko, Michelle Ramirez, Samuel Stevens, Alec Sheets, Reshma Ramesh Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, Jackson Miliko, Neil Rosser, Isla Duporge, Charles Stewart, Tanya Berger-Wolf, and Daniel Rubenstein. Kabr raw videos: Unprocessed drone footage for kenyan animal behavior analysis (revision fda789c), 2025c. URL https://huggingface.co/datasets/imageomics/KABR-raw-videos.

Jenna Kline, Alison Zhong, Kevyn Irizarry, Charles V Stewart, Christopher Stewart, Daniel I Rubenstein, and Tanya Berger-Wolf. Wildwing: An open-source, autonomous and affordable uas for animal behaviour video monitoring. *Methods in Ecology and Evolution*, 2025d. URL https://doi.org/10.1111/2041-210X.70018.

Jenna Kline, Alison Zhong, Maksim Kholiavchenko, Sam Stevens, Alec Sheets, Reshma Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, Jackson Miliko, Neil Rosser, Charles Stewart, Tanya Berger-Wolf, and Daniel Rubenstein. Kabr worked example: Manually annotated detections and behavioral analysis for kenyan wildlife pipeline demonstration, 2025e. URL https://huggingface.co/datasets/imageomics/kabr-worked-examples.

Benjamin Koger, Adwait Deshpande, Jeffrey T. Kerby, Jacob M. Graving, Blair R. Costelloe, and Iain D. Couzin. Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision. *Journal of Animal Ecology*, 92(7):1357–1371, 2023. ISSN 1365-2656. doi: 10.1111/1365-2656.13904.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Nicholas E. Newton-Fisher. Animal behaviour pro: 1.6, September 2021. URL https://kar.kent.ac.uk/82465/.

Lucia Pedrazzi, Hemal Naik, Chris Sandbrook, Miguel Lurgi, Ines Fürtbauer, and Andrew J King. Advancing animal behaviour research using drone technology. *Animal Behaviour*, 222:123147, 2025.

Tinao Petso, Rodrigo S Jamisola Jr, Dimane Mpoeleng, Emily Bennitt, and Wazha Mmereki. Automatic animal identification from drone camera based on point pattern analysis of herd behaviour. *Ecological informatics*, 66: 101485, 2021.

Corinna Riginos. Climate and the landscape of fear in an african savanna. *Journal of Animal Ecology*, 84(1):124–133, 2015.

David Rolnick, Alan Aspuru-Guzik, Sara Beery, Bistra Dilkina, Priya L Donti, Marzyeh Ghassemi, Hannah Kerner, Claire Monteleoni, Esther Rolf, Milind Tambe, et al. Application-driven innovation in machine learning. *arXiv preprint arXiv:2403.17381*, 2024.

D. Rubenstein, B. Low Mackey, ZD Davidson, F. Kebede, and S. R. B. King. Equus grevyi. The IUCN Red List of Threatened Species 2016: e.T7950A89624491, 2016. https://dx.doi.org/10.2305/IUCN.UK.2016-3.RLTS.T7950A89624491.en, Accessed on 06 August 2025.

D. I. Rubenstein. Ecology and sociality in horses and zebras. In D. I. Rubenstein and R. W. Wrangham, editors, *Ecological Aspects of Social Evolution*, pages 282–302. Princeton University Press, Princeton, NJ, 1986.

Daniel I Rubenstein. Ecology, social behavior, and conservation in zebras. In *Advances in the Study of Behavior*, volume 42, pages 231–258. Elsevier, 2010.

Lyes Saad Saoud, Atif Sultan, Mahmoud Elmezain, Mohamed Heshmat, Lakmal Seneviratne, and Irfan Hussain. Beyond observation: Deep learning for animal behavior and ecological conservation. *Ecological Informatics*, page 102893, 2024.

Lukas Schad and Julia Fischer. Opportunities and risks in the use of drones for studying animal behaviour. *Methods in Ecology and Evolution*, 14(8):1864–1872, 2023.

Jennifer E Smith and Noa Pinter-Wollman. Observing the unwatchable: Integrating automated sensing, naturalistic observations and animal social network analysis in the age of big data. *Journal of Animal Ecology*, 90(1):62–75, 2021.

Guoqing Zhang, Yongxiang Zhao, Ping Fu, Wei Luo, Quanqin Shao, Tongzuo Zhang, and Zhongde Yu. A reliable unmanned aerial vehicle multi-target tracking system with global motion compensation for monitoring procapra przewalskii. *Ecological Informatics*, 81:102556, 2024.

7 Appendix

7.1 Field protocols

We manually recorded ground-based behavioral data of zebras and giraffes at the Mpala Research Center to assess the consistency between drone-based and traditional behavioral sampling methods. We visually located the herds on the ground from the field vehicle, launched the drone, and began manual ground-based behavioral observations once the animals were visible from both the drone and the field vehicle.

7.1.1 Ground-based observation protocols

Field observations were conducted by trained observers who underwent rigorous calibration to ensure data quality and consistency. Observers participated in field calibration exercises, practicing simultaneous observations overseen by an expert field ecologist. Throughout data collection, teams conducted daily 15-minute synchronization sessions to maintain consistency and address any emerging discrepancies in behavioral classification. Each observation team utilized 10 X 42 binoculars for detailed behavioral observation. We used the iOS version of the Animal Behaviour Pro [Newton-Fisher, 2021] to record timestamped behavior observations data synchronized with GPS coordinates on our iPhone mobile devices.

We used two complementary sampling methods to capture different aspects of behavioral data: scan sampling and focal sampling. Scan sampling sessions targeted groups of 3-12 individuals observed for an average duration of 10.2 ± 2.3 minutes. During these sessions, instantaneous scans were conducted every 2 minutes, with observers systematically recording the behavior of all visible individuals from left to right across the group. This approach provided population-level behavioral data while controlling for environmental and social context. Focal sampling was conducted simultaneously by a second observer team, with continuous observation of single individuals for 10-minute periods. Focal animals were selected based on clear visibility and distinctive identifying features such as unique stripe patterns for zebras or spot configurations for giraffes. All behavioral transitions were recorded with 1-second precision, providing fine-scale temporal resolution of behavioral sequences. We completed six simultaneous scan and focal sessions across different groups and species, providing paired datasets for methodological comparison.

7.1.2 Drone protocol

To minimize behavioral disturbance, we employed standardized approach protocols. We visually located the herds and slowly approached in out field vehicle until we were positioned approximately 100 meters away with the takeoff location positioned with the vehicle between the pilot and the animals. After observing the animals for five to ten minutes to gauge vigilance, the drone pilot exited the vehicle on the opposite side from the animals to position the drone for takeoff. After takeoff, the drone's altitude was set immediately to 40-50 meters. The drone hovered in place for 5 minutes to gauge whether the animals reacted. If the animals appeared undisturbed, the pilot gradually decreased altitude to 20-30 meters. Note, this is lower than typical wildlife drone surveys (70-100m) to maintain sufficient video resolution for

individual behavioral identification, and hovered briefly before slowly approaching the group. We immediately paused drone operations whenever animals exhibited signs of agitation or increased vigilance behavior. This conservative approach prioritized animal welfare while maintaining data quality for behavioral analysis. The drone's camera was angled approximately 45 degrees with respect to the terrain to better capture the side-view of the animals, with flight speeds kept below 5 m/s when tracking moving groups to maintain smooth footage. Our battery management protocols ensured maximum 20-minute flights with 20% power reserve for safe return. All videos included embedded GPS coordinates and timestamps enabling precise synchronization with ground observations. We uploaded full-length drone videos and corresponding mission telemetry data on the Ohio Supercomputer Center (OSC) for storage and processing.

7.2 Demographic Information

Session	Species	Male	Female
1	Plains Zebra	1	2
2	Plains Zebra	1	4
3	Grevy's Zebra	3	0
4	Grevy's Zebra	3	0
5	Reticulated giraffe	1	1
6	Reticulated giraffe	2	0

Table 8: Sessions capturing behavior with simultaneous scan and focal sampling, [Kline et al., 2025b]. Two simultaneous scan and focal sampling sessions were captured for each of the three ungulate species with a mix of male and females in each group. Simultaneous scan and focal sampling were collected to compare the relative strengths and weaknesses of each methodology.

Zebra ID	Demographic information	Temporal Overlap
A	Adult female lactating Grevy's zebra travelling with her foal and two female plains zebras	96 sec
B	Adult female Grevy's zebra, travelling with Grevy's herd	141 sec
C	Female foal Grevy's zebra, travelling with Grevy's herd	197 sec
D	Plains male zebra, travelling with plains harem	72 sec

Table 9: Demographic details of individual zebras' behaviors analyzed to compare ground focal with drone focal annotations, Kline et al. [2025b]. The temporal overlap quantifies the time when ground and drone focal sampling coincide.

Session	Habitat	Herd Size Category (Count)	Date
1	Closed	Small (3)	11/01/23
2	Open	Small (3)	11/01/23
3	Open –> Closed	Small (3)	12/01/23
4	Closed	Large (7)	16/01/23
5	Open	Small (2)	18/01/23
6	Open	Large (5)	20/01/23

Table 10: Summary of data used for Case Study 1: Grevy's Zebras Landscape of Fear, by habitat and herd size, [Kline et al., 2024a, 2025e]. Herds larger than the median (4.2) are considered 'Large', all others are considered 'Small'. For session 3, the habitat type transitioned from Open to Closed as the animals moved to a more dense habitat during the video duration.

7.3 Machine learning performance

Model	Mean Average Precision (mAP) (%)			Precision	Recall	F1
	Overall	Head Classes	Tail Classes			
I3D	65.06	96.81	54.48	**67.17**	62.94	64.52
SlowFast	66.10	96.72	55.90	67.05	**65.28**	**65.82**
X3D-L	**66.36**	**96.96**	**56.16**	66.44	63.65	64.70
UniformerV2-B	61.78	95.38	50.58	64.37	54.82	57.41

Table 11: Behavior recognition machine learning model performance, summarized from [Kholiavchenko et al., 2024a]. The mean average precision (mAP %) is reported overall (all 8 classes), Head classes (*Walk*, *Graze*), and Tail classes (*Head Up, Browse, Auto-Groom, Trot, Run, Out of Sight/Occluded*). The precision (P), recall (R), and F1 score (F1) are reported for each model. The highest scores across all metrics are shown in **bold** for each evaluation setting. Refer to our previously published work for additional model training and performance details [Kholiavchenko et al., 2024a].

Metric	Formula
Precision	$\dfrac{TP}{TP + FP}$
Recall	$\dfrac{TP}{TP + FN}$
F1 Score	$2 \times \dfrac{Precision \times Recall}{Precision + Recall}$
Average Precision (AP)	Area under the Precision-Recall curve
Mean Average Precision (mAP)	$\dfrac{1}{N} \sum_{i=1}^{N} AP_i$

Table 12: Summary of evaluation metrics used in object detection and classification. TP = True Positives, FP = False Positives, FN = False Negatives, N = Number of classes, and AP_i = Average Precision for class i.

7.4 Ethogram

Behaviour	Code	Species	Notes
Auto-groom	AG/SG	Both	Self grooming, rubbing nose or teeth on the neck, back or other body parts, or rubbing body on a landscape feature, such as a tree or branch
Browsing	B	G	Eating leaves off trees, chewing and looking around
Chasing	C	Z	Focal animal chasing another animal of same species
Defecating	DF/D	Both	Elimination of solid waste
Drinking	D	Z	Ingesting water
Dusting	DU	Z	Rolling on back to covered body in dust
Fighting	F	Both	Aggression towards another animal, involves biting, rearing and kicking with fore or hind legs
Graze	G	Z	Head down, biting and chewing grass
Head Up	HU	Both	Vigilance, standing, sleeping, or scanning
Herding	H	Z	Head down with neck stretched out, directing another animals' movements
Lying down	L	Z	Body lying on ground
Mounting/Mating	M	Both	Engaging in mating behaviors, such as mounting
Mutual grooming	MG	Both	Rubbing with nose or teeth each others neck, back or other body parts
Running	R	Both	A high-speed gait where both diagonal pairs of legs move together, but there is a moment when all four hooves are off the ground.
Sniff	S	Z	Sniffing another animal or nuzzling the ground
Trotting	TR	Z	A faster, more energetic gait where diagonal pairs of legs (front left with back right, and vice versa) move together, resulting in a brief period when all four hooves are off the ground.
Urinating	U	Both	Elimination of liquid waste
Walking	W	Both	A slow, steady gait where one leg moves at a time, with at least one hoof in contact with the ground at all times
Technical annotations			
Occluded*	OCL	Both	Animal hidden behind bush/another animal
Out of Focus*	OOC	Both	Object detector switches to wrong animal
Out of Frame*	OOF	Both	Animal not visible in video frame

Table 13: Combined ethogram for zebras and giraffes. Species column indicates applicability: Z = zebras only, G = giraffes only, Both = applicable to both species. *Technical categories can overlap (e.g., detector error may precede animal going out of frame).