MammAlps: A multi-view video dataset of wild mammals behavior monitoring in the Swiss Alps

Supplementary Materials

Contents

A. Data acquisition

A.1. Site descriptions and period of acquisition

The Swiss National Park is located in Eastern Switzerland and has a substantially higher density of ungulates compared to neighboring regions. Additionally, the park is a strictly protected nature reserve, and thus human activities are restricted to be minimal [6]. This makes the region particularly interesting to acquire data on the naturalistic behaviors of ungulates from camera trap videos over a relatively short period of time.

We identified three sites for habitat monitoring. The three sites used for the study are located between 1840 m and 1890 m of altitude, at which elevation mostly red deers and roe deers are found, chamois foraging generally higher at this period of the year. For privacy reasons, we do not disclose the exact location.

Site 1 is a clearing within an alpine forest composed of larch, cembra pine, mountain pine and spruce facing South-West. Site 2 is located at the intersection of multiple game paths, in a similar forest type facing North. Site 3 is located by a water stream where the terrain creates two small water pounds, and is facing towards South. The three sites were chosen by purpose to acquire a behavioral dataset as diverse as possible since observing different behavioral expressions is of a high chance in these sites. Cameras acquired video and audio data for 6 weeks between August and October 2023. This period corresponds to the rutting season of red deer, and thus many events represent rutting-related behaviors.

A.2. Camera settings

Camera traps (Browning's Spec Ops Elite HP5) acquired videos of fixed duration (either for 1 or 2 minutes at daytime, and 20 seconds when with the IR flash). Cameras were set to fast trigger mode with a delay of 1 second between subsequent videos, with long-range motion detection enabled. Cameras were fixed either on wooden poles or on trees, around 60 cm above ground. Cameras were positioned on the sites with varying levels of field-of-view overlap, while Site 3 had the most con-focal setup, and site 1 had the least.

We report video acquisition statistics per camera and per site (Fig. S1a-b). When cameras began to run out of battery, the recordings at night were automatically shortened by the hardware, leading to many nighttime clips with short durations (below 20 seconds). Among these clips, We kept only the ones containing at least 30 frames (1 second).

B. Details on data processing and annotation

B.1. From events to tracklets

We used MegaDetector [1, 7] v5a at a sampling period of five frames to detect recordings with animals among the raw videos ($N = 3794$). Videos which did not have at least two animal detections above a permissive animal detection confidence threshold of 0.3, were considered as false positives. The videos with detections ($N = 1961$) were then trimmed to the segment between the first and last MegaDetector detection. We ran MegaDetector v5a again on every frame of the trimmed videos to obtain dense animal detection predictions.

To obtain animal tracks we adapted the matching algorithm from ByteTrack [12]. Indeed, ByteTrack performance depends on the performance of the object detector and the frame rate (the more frequent the better). However, as MegaDetector was not fine-tuned on our data, we observe a high rate of missing detections either because of long-term occlusions (*e.g.* an animal passing behind a tree), low frame

quality (*e.g.* at night), and relatively low frame rate (for tracking, i.e., 30 FPS). To improve tracking performance, we used the generalized intersection-over-union matching cost (GIoU), instead of the (IoU) originally proposed in ByteTrack to allow the matching of bounding boxes even when they do not overlap. We added an area difference matching cost to avoid matching animals with small false detections from MegaDetector (*e.g.* rain drops). We also gave maximum certainty to the measurements (MegaDetector bounding boxes) during the Kalman Filter integration process to avoid long-term interpolations and bounding boxes that would lag behind the animal after long occlusions. Specifically, we used a detection threshold of 0.2, a track activation threshold of 0.5, a lost track buffer of 300 frames, and a minimum matching threshold for high confidence pairs of 0.75. The cost C between two bounding boxes B_i and B_j is defined as follows:

$$C(B_i, B_j) = 1 - (GIoU(B_i, B_j) - 2 * A(B_i, B_j)) + 3)/4$$
$$A(B_i, B_j) = \frac{|Area(B1) - Area(B2)|}{area(B1) + Area(B2)} \tag{1}$$

After dense prediction and tracking, resulting tracks were all visually examined and corrected in CVAT [5] when necessary. Specifically, tracks were corrected for identity switches and duplicated or lost tracks. We also removed any false positive tracks (*e.g.* a rock), yielding a total 2139 animal tracks.

A video tracklet of dimension 380×380 was created for each individual track by cropping the original video and padding it with the background to preserve the 1:1 aspect ratio. In crowded scenes, it is common that multiple animals expressing different behaviors are visible on the same tracklet, which may ultimately impact model performance.

The curated tracks include five species: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), fox (*Vulpes vulpes*), wolf (*Canis lupus*) and mountain hare (*Lepus timidus*). Other species were not included, either because too few events were captured or because individuals were too small.

B.2. Behavior annotations

We report the list of behaviors used in the study, along with their definitions and their associated actions, which were automatically gathered from the annotations (Tab. S1). We used a mixed approach to select relevant behaviors. First we sourced behaviors from ethogram studies of related deer species. Then, we adjusted the list based on what was interpretable from video-data, and the behavior observed in our data. The *ruminating* behavior was discarded since it was difficult to detect, especially at nighttime, and was hence merged with *standing head up*. The *exploring* behavior was also difficult to differentiate from others, and thus merged with *foraging*. Some social behaviors such as *parenting* or other non-agonistic behaviors between individuals were not included as they are relatively difficult to define in space and time in a consistent manner.

B.3. Reference scene segmentation maps

Before dismounting cameras, a reference picture of the scene was collected for each of them by manually triggering the camera trap (Fig. S2).

The reference scenes (Fig. S2) were annotated in CVAT [5] for 10 classes: *bush*, *pole*, *rock*, *grass*, *soil/path*, *log*, *tree trunk*, *foliage*, *water*, and *background* (Fig. S3).

B.4. Quantification of cameras temporal drift

We quantified the temporal drift between pairs of cameras for each site, as shown in Fig. S4. This was achieved by manually selecting frames that depicted the same animal pose from at least two camera views, and reporting the date and time of the respective frames. Site 1 shows the biggest drift, while cameras in Site 2 seems less prone to temporal drift. Site 3 contains limited data as Camera 1 battery ran off early, which limits the quantification of the drift.

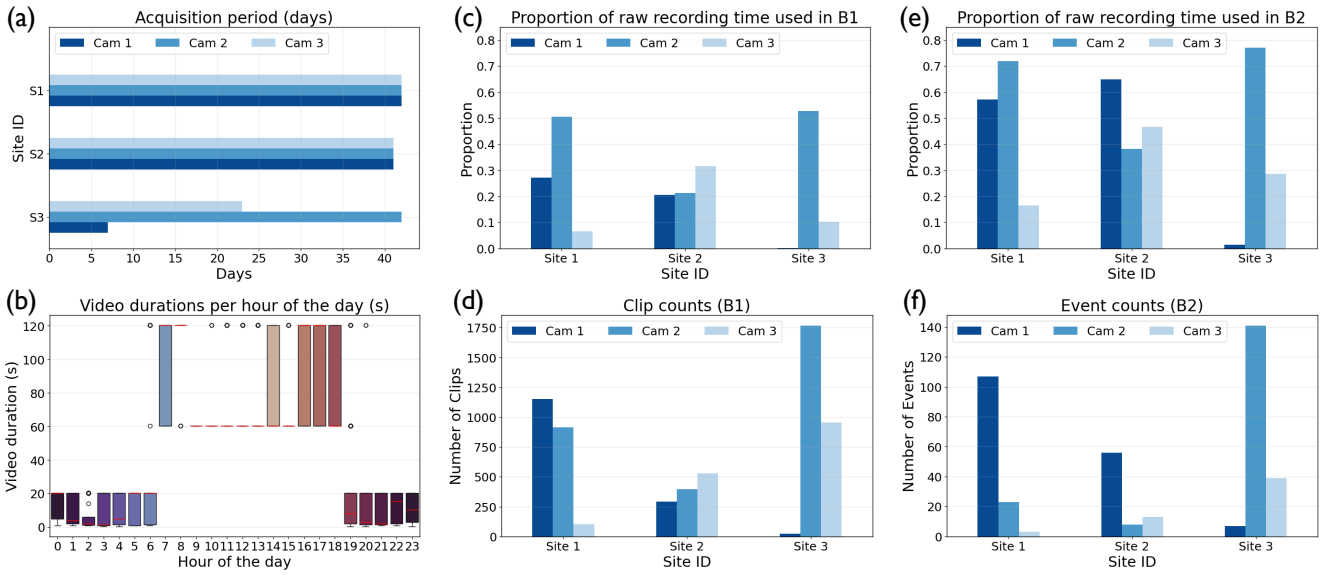

Figure S1. **Dataset statistics on the acquired data per camera, and on the data used in Sec. 3.3 and 3.4.** (a) summarizes the number of recording days before curation. Note that the batteries for two of the cameras at site S3 ran out earlier. Video durations per hour of the day (b) were computed on the subset of raw videos belonging to either benchmark B1 or B2.

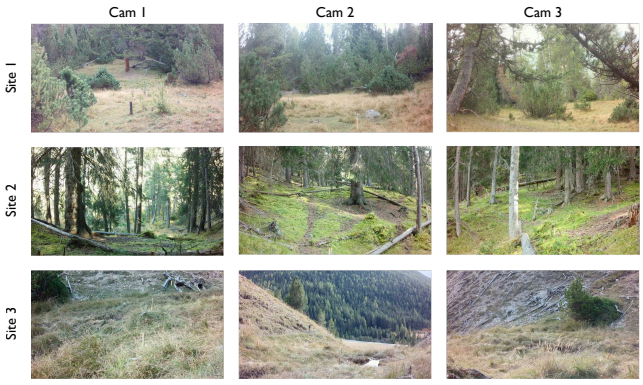

Figure S2. **Reference picture of the scene for each camera.**

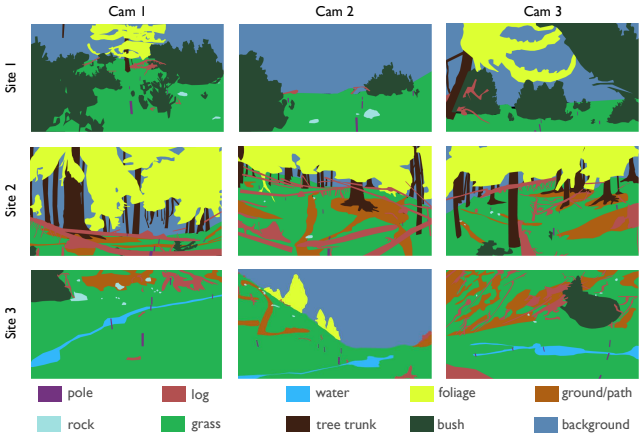

Figure S3. **Reference scene segmentation maps.**

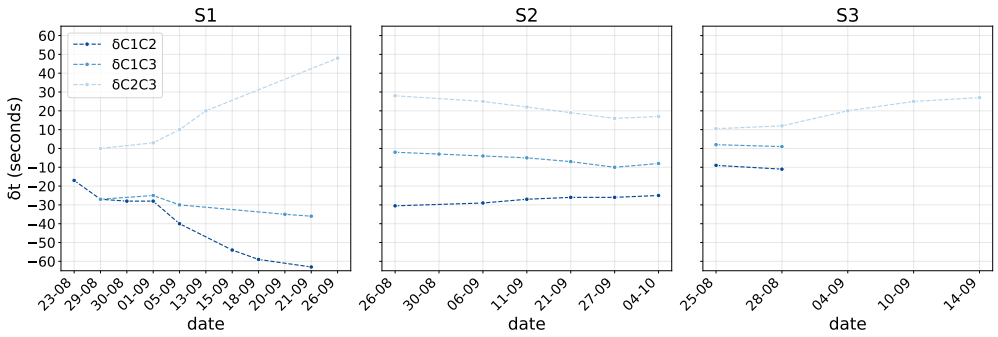

Figure S4. **Temporal drift between pairs of cameras over time.**

Activity	Associated actions	Definition
Camera Reaction	standing head up, looking at the camera, running, sniffing, jumping, walking	Any type of behavior that involves reacting to a camera.
Chasing	running, walking	Whenever a predator chases a prey.
Courtship	standing head up, running, vocalizing, bathing, scratching antlers, laying, walking	Behaviors related to breeding, uniquely for red deer at this period of the year. It can involve a single stag (*e.g*. vocalizing) or multiple individuals (*e.g*. running after a hind).
Escaping	running, vocalizing, walking, jumping	Escaping from a predator, or running away from another individual from the same species.
Foraging	standing head up, laying, unknown, running, drinking, sniffing, vocalizing, standing head down, bathing, defecating, grazing, walking, urinating, scratching body	Large family of behaviors related to energy acquisition, from environment sensing (*e.g*. sniffing) to actual consumption (*e.g*. grazing).
Grooming	standing head up, shaking fur, bathing, standing head down, scratching antlers, defecating, scratching hoof, laying, walking, urinating, scratching body	Behaviors involving a single individual that cleans its body and fur, either by scratching in multiple ways or while bathing.
Marking	standing head up, defecating, bathing, scratching antlers, standing head down, jumping, scratching hoof, walking, urinating	Behaviors related to a single stag that marks specific features from the environment.
Playing	standing head up, running, sniffing, standing head down, jumping, scratching hoof, walking	Behaviors involving one or multiple individuals, often young ones, and characterized by running or jumping in the absence of negative stimuli.
Resting	standing head up, bathing, scratching antlers, standing head down, laying	Whenever an animal stays in place for a long time and does not appear to be in vigilance or foraging.
Unknown	standing head up, unknown, running, sniffing, standing head down, jumping, scratching hoof, walking	Sometimes the behavior cannot be deduced from the current context, for example, because of occlusion or some decisive parts of the body being out-of-frame.
Vigilance	standing head up, looking at the camera, running, sniffing, standing head down, defecating, grazing, walking	Any behavior where an animal or a group of animals are actively sensing the environment either to detect potential predators or other sources of threat, or in reaction to another individual's vocalization.

Table S1. **Definition of the activities present in the dataset and their associated actions.**

C. Benchmark 1: Multimodal Species and Behavior recognition

C.1. Multimodal VideoMAE Implementation details

We adopted a condensed version of VideoMAE [10] from InternVideo [11], for which we used the pre-trained weights on Kinetics 700 dataset [3]. We replaced the original classification head with three classification heads to predict species (Spe), activities (ActY) and actions (ActN) simultaneously, while using the loss weights of 1, 2.5 and 2. Meanwhile, we implemented a balanced sampling strategy to deal with the unbalanced number of samples across different classes. For all the models with different modality inputs, we trained them with 150 epochs with the learning rate decreasing from 10^{-5} to 10^{-7}.

An overview of the model trained for B1 was created (Fig. S5). We made several modifications so that the VideoMAE [10] model can take different modalities as input (video, audio and segmentation masks). First, the video modality is naturally trivial – we sampled 16 frames similar to the original VideoMAE [10] and then transformed them to $16 \times 14 \times 14$ patches. It needs to be noted that we only sampled frames within 5 seconds of randomly selected windows since some behaviors span long times; this captured evidence more compactly. For the audio inputs, we first found the audio clip simultaneous to the video clip and then transformed the original audio signal to a spectrogram, similar to AudioMAE [8]. We adopted a smaller audio sample length (10 in comparison to the original 25) so that the spectrogram can be generated with fewer audio samples. We applied masking across temporal and frequency domains during training for data augmentation. The spectrogram was interpolated to 256 tokens to obtain the same input length across different samples. Finally, for the segmentation inputs, we sampled 16 frames simultaneous to the sampled video frames. Segmentation inputs were represented as one-hot encoded matrices for every frame so that the model did not rely on spurious linear dependencies between the class indices.

We optimized model parameters by back-propagating the three task-specific cross-entropy (CE) losses. After the quantitative comparison between binary cross-entropy and CE loss for ActN recognition, CE ultimately increased optimization speed, most likely since there are at most two actions and often only one. For both B1 and B2 we used balancing sampling. To account for multiple labels, we computed a sampling weight proportional to the *sum* of their inverse class frequencies.

C.2. Baseline performances.

To contextualize the difficulty of B1, we ran additional experiments on the ActY recognition task for videos (Tab. S2).

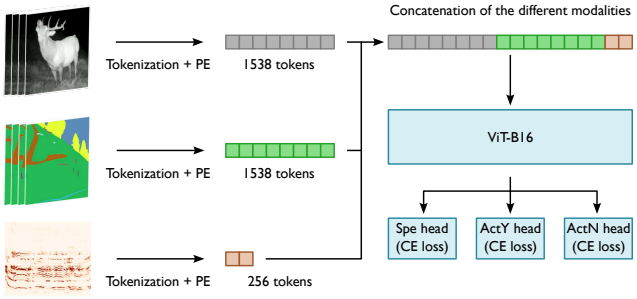

Figure S5. **Multimodal Video Transformer implementation for B1.** Note that the transformer backbone is similar to both B1 and B2. In B2, the backbone is followed by four classification heads instead of the three depicted here, one for each of the classification tasks.

Note that the model evaluated on KABR [9] and MammalNet [4] show behavior recognition scores of 0.66 (mAP on X3D-L) and 0.378 (top-1 balanced acc. on mViTv2), respectively, indicating that the difficulty is in the range of related datasets for this single unimodal task.

Baseline	mAP	top-1 balanced accuracy
SlowFast-8x8†	0.203	0.197
X3D-M†	0.251	0.256
mViT-v2†	0.259	0.156
VideoMAE† (ours)	0.410	0.274
VideoMAE (ours)	0.414	0.403

Table S2. Additional baseline performances on the ActY recognition task from videos. †: uniform sampling

C.3. Models performance per class

We report model performances (F1-scores and average precisions) per class (Tab. S3, Tab. S4, Tab. S5 and S6). The advantage of reporting the mAP (or AP when considering single classes) is that the metric better represents the area under the curve as it computes the precision over multiple thresholds, and it can be equally applied to multi-class and multi-label problems. To compute the F1-score, we used a threshold of 0.5 on the softmax and sigmoid outputs for multi-class and multi-label tasks, respectively.

Activity	Support	F1-score									
Trained on Modality		ActY. V	ActY.+ActN. V	ActY.+Spe. V	All V	All A	All S	All A+S	All V+S	All V+A	All V+A+S
Cam. reaction	7	0.167	0.182	0.000	0.000	0.080	0.000	0.000	0.111	0.000	**0.190**
Chasing	3	**1.000**	**1.000**	0.857	**1.000**	0.000	0.462	0.250	**1.000**	0.857	0.750
Courtship	56	0.565	0.532	0.429	0.442	0.589	0.143	0.512	0.330	0.574	**0.617**
Escaping	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Foraging	688	0.782	0.795	0.760	0.801	0.677	0.651	0.709	0.783	**0.822**	0.789
Grooming	24	0.350	**0.359**	0.264	0.293	0.014	0.108	0.150	0.230	0.356	0.310
Marking	76	0.667	0.583	0.504	0.569	0.509	0.230	0.382	0.516	0.775	**0.787**
Playing	21	0.000	0.000	0.000	0.000	**0.067**	0.049	0.000	0.000	0.000	0.000
Resting	48	**0.250**	0.185	**0.250**	0.189	0.000	0.039	0.000	0.207	0.154	0.185
Unknown	92	0.426	0.398	0.394	0.378	0.030	0.275	0.229	0.441	**0.508**	0.393
Vigilance	228	0.625	**0.664**	0.619	0.637	0.025	0.183	0.338	0.589	0.640	0.621
Macro	1244	**0.439**	0.427	0.371	0.392	0.181	0.194	0.234	0.382	0.426	0.422

Table S3. **F1-scores per activity for the behavior recognition benchmark (B1).** V: video clips; A: audio spectrograms; S: segmentation map clips; ActY.: Activities; ActN.: Actions; Spe.: Species.

Activity	Support	AP									
Trained on Modality		ActY. V	ActY.+ActN. V	ActY.+Spe. V	All V	All A	All S	All A+S	All V+S	All V+A	All V+A+S
Cam. reaction	7	0.089	0.114	0.169	0.119	0.018	0.042	0.073	0.104	0.114	**0.194**
Chasing	3	**1.000**	**1.000**	**1.000**	**1.000**	0.017	0.362	0.423	**1.000**	**1.000**	0.917
Courtship	56	0.540	0.552	0.425	0.419	0.638	0.113	0.569	0.369	0.633	0.**651**
Escaping	1	0.059	0.034	**0.333**	0.023	0.006	0.004	0.015	0.077	0.017	0.038
Foraging	688	0.850	0.870	0.857	0.867	0.613	0.703	0.735	0.840	**0.873**	0.870
Grooming	24	0.280	0.291	0.216	**0.308**	0.020	0.101	0.116	0.152	0.307	0.222
Marking	76	0.739	0.619	0.572	0.654	0.534	0.155	0.321	0.556	**0.794**	0.788
Playing	21	0.017	0.022	0.026	0.024	0.042	**0.071**	0.050	0.055	0.036	0.030
Resting	48	0.275	**0.289**	0.286	0.267	0.070	0.051	0.068	0.205	0.280	0.218
Unknown	92	0.342	0.367	0.344	0.357	0.104	0.227	0.208	0.421	**0.456**	0.395
Vigilance	228	0.651	0.706	0.646	0.672	0.218	0.241	0.310	0.608	**0.713**	0.651
Macro	1244	0.440	0.442	0.443	0.428	0.207	0.188	0.262	0.399	**0.475**	0.452

Table S4. **Average precisions (AP) per activity for the behavior recognition benchmark (B1).** V: video clips; A: audio spectrograms; S: segmentation map clips; ActY.: Activities; ActN.: Actions; Spe.: Species.

Action	Support	F1-score									
Trained on Modality		ActN. V	ActY.+ActN. V	ActN.+Spe. V	All V	All A	All S	All A+S	All V+S	All V+A	All V+A+S
Bathing	2	**0.400**	0.286	**0.400**	**0.400**	0.013	0.028	0.071	0.133	0.286	**0.400**
Defecating	6	0.000	0.000	0.000	0.000	0.026	0.022	**0.040**	0.000	0.000	0.013
Drinking	6	**0.500**	0.444	0.400	0.444	0.033	0.062	0.156	0.267	0.345	0.316
Grazing	184	**0.684**	0.613	0.650	0.616	0.425	0.510	0.508	0.564	0.592	0.564
Jumping	7	0.000	0.000	**0.222**	0.000	0.044	0.108	0.000	0.000	0.143	0.000
Laying	53	0.312	**0.435**	0.394	0.317	0.102	0.062	0.051	0.314	0.344	0.303
Look. at cam.	2	0.000	0.000	0.000	**0.333**	0.000	0.041	0.118	0.074	0.000	0.000
Running	36	0.466	0.416	0.376	**0.471**	0.162	0.305	0.325	0.313	0.455	0.330
Scratch. antlers	55	0.638	0.645	0.626	0.680	0.280	0.188	0.258	0.508	**0.745**	0.686
Scratch. body	10	**0.250**	0.187	0.211	0.000	0.000	0.015	0.030	0.083	0.091	0.139
Scratch. hoof	24	0.294	0.321	0.373	0.280	0.236	0.127	0.286	0.200	0.429	**0.430**
Shaking fur	11	0.545	**0.571**	0.400	0.538	0.020	0.101	0.161	0.359	0.273	0.350
Sniffing	38	**0.479**	0.143	0.232	0.193	0.064	0.081	0.116	0.120	0.218	0.118
Stand. head down	180	0.464	0.375	**0.467**	0.400	0.279	0.300	0.298	0.385	0.400	0.381
Stand. head up	265	0.689	**0.712**	0.648	0.677	0.359	0.390	0.397	0.585	0.702	0.629
Unknown	75	**0.578**	0.551	0.507	0.497	0.147	0.283	0.217	0.357	0.502	0.401
Urinating	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Vocalizing	37	0.323	0.500	0.328	0.505	**0.604**	0.188	0.481	0.306	0.598	0.511
Walking	300	**0.786**	0.746	0.780	0.714	0.400	0.458	0.491	0.548	0.730	0.658
Macro	1292*	**0.390**	0.366	0.369	0.372	0.168	0.172	0.211	0.269	0.361	0.328

Table S5. **F1-scores per action for the behavior recognition benchmark (B1).** *Note that since there can be up to two actions per sample, this increases the total number of samples since each label is considered independently. V: video clips; A: audio spectrograms; S: segmentation map clips; ActY.: Activities; ActN.: Actions; Spe.: Species.

Action	Support	AP									
Trained on Modality		ActY. V	ActY.+ActN. V	ActY.+Spe. V	All V	All A	All S	All A+S	All V+S	All V+A	All V+A+S
Bathing	2	0.507	0.509	0.528	**0.550**	0.011	0.254	0.508	0.503	0.520	0.507
Defecating	6	0.008	0.012	0.008	0.006	0.014	0.007	**0.061**	0.005	0.005	0.007
Drinking	6	0.633	0.555	0.714	0.513	0.051	0.029	0.166	**0.800**	0.621	0.502
Grazing	184	**0.857**	0.746	0.848	0.847	0.388	0.493	0.544	0.792	0.834	0.812
Jumping	7	0.023	0.023	**0.207**	0.024	0.032	0.042	0.043	0.014	0.105	0.024
Laying	53	0.315	0.368	0.381	0.369	0.054	0.085	0.093	0.244	**0.382**	0.321
Look. at cam.	2	0.008	0.013	0.012	**0.238**	0.002	0.126	0.035	0.026	0.047	0.030
Running	36	0.669	**0.684**	0.586	0.634	0.172	0.315	0.457	0.489	0.646	0.521
Scratch. antlers	55	0.674	0.672	0.654	0.716	0.184	0.160	0.192	0.558	0.742	**0.760**
Scratch. body	10	**0.164**	0.091	0.152	0.067	0.009	0.014	0.017	0.044	0.054	0.097
Scratch. hoof	24	0.294	0.212	0.198	0.292	0.289	0.069	0.299	0.166	0.470	**0.519**
Shaking fur	11	**0.559**	0.400	0.323	0.516	0.020	0.134	0.126	0.248	0.345	0.272
Sniffing	38	**0.517**	0.320	0.399	0.456	0.037	0.101	0.122	0.246	0.407	0.167
Stand. head down	180	0.575	0.477	**0.578**	0.535	0.234	0.197	0.181	0.313	0.521	0.346
Stand. head up	265	0.778	**0.853**	0.806	0.851	0.035	0.362	0.462	0.772	0.830	0.806
Unknown	75	0.610	0.607	**0.619**	0.572	0.093	0.280	0.278	0.507	0.576	0.519
Urinating	1	**0.007**	0.002	**0.007**	0.003	0.002	0.005	0.002	0.004	0.001	0.003
Vocalizing	37	0.415	0.688	0.500	0.606	0.835	0.100	0.724	0.561	0.787	**0.836**
Walking	300	0.890	0.881	0.876	**0.901**	0.284	0.477	0.569	0.841	0.895	0.878
Macro	1292*	0.447	0.427	0.442	0.458	0.161	0.171	0.257	0.375	**0.463**	0.417

Table S6. **Average precisions (AP) per action for the behavior recognition benchmark (B1).** *Note that since there can be up to two actions per sample, this increases the total number of samples since each label is considered independently. V: video clips; A: audio spectrograms; S: segmentation map clips; ActY.: Activities; ActN.: Actions; Spe.: Species.

D. Benchmark 2: Multi-view Long-term Event Understanding

Here we detail our simple baseline method for B2. In particular, we illustrate how we performed token merging, how we trained the model and additional results.

D.1. Selecting false positive events

The raw video dataset contains 43 h of raw data, where the majority comes from false positive samples in Camera 1 of site 3 (Fig. S1a-b). While having these false positive events is important for B2 as they represent true data and are common in camera trap surveys, a disproportionate number of them leads to unnecessarily high computational costs. To construct the dataset for B2, we therefore discarded any event that was longer than 15 minutes (cumulative recording time among all points of view) which eliminated 10 false positive events and three true positive ones, and effectively reducing the dataset size to 14 hours with 3 hours of false positive events.

D.2. Offline Token Merging strategy

We describe our offline token merging strategy over time in Algorithm 1, and illustrate the process (Fig. S6). After spatial merging with ToME [2], we select the tokens of every second frame and merge them with any other tokens from all the other frames, following the same soft-bipartite graph matching algorithm used in the original method [2]. The process is repeated iteratively the final number of video tokens is equal or inferior to the original number of tokens in a single frame. Note that the final number of video tokens increases with the video duration since we perform the algorithm in chunks. The embedding dimension is 768, and the chunk size is 615 frames.

Algorithm 1 Offline Token Merging

Require: Video frames $\mathcal{F}$,
Pretrained Vision-MAE with token merging [2] ToME,
ToME reduction factor r,
Chunk size c
Ensure: Condensed video tokens $\mathcal{T}_{\text{final}}$

for each chunk $\mathcal{C}_i \subset \mathcal{F}$ of size c **do** ▷ Process in chunks
  $\mathcal{T} \leftarrow \text{ToME}(f_j, r), \forall f_j \in \mathcal{C}_i$ ▷ Spatial Merging
  $N_f \leftarrow |\mathcal{T}_j|$ for any j ▷ Tokens in each frame
  $N_v \leftarrow N_f \times |\mathcal{C}_i|$ ▷ Tokens in chunk
  while $N_v > N_f$ **do** ▷ Temporal Merging
    $\mathcal{T}_{\text{selected}} \leftarrow \{\mathcal{T}_j \mid j \text{ is even}\}$
    $\mathcal{T}_{\text{other}} \leftarrow \{\mathcal{T}_j \mid j \text{ is odd}\}$
    $\mathcal{T} \leftarrow \text{Merge}(\mathcal{T}_{\text{selected}}, \mathcal{T}_{\text{other}})$
    $\mathcal{C}_i \leftarrow \{f_j \mid j \text{ is even}, \forall f \in \mathcal{C}_i\}$
    $N_v \leftarrow N_f \times |\mathcal{C}_i|$
  end while
end for
$\mathcal{T}_{\text{final}} \leftarrow \bigcup_i \mathcal{T}_{\mathcal{C}_i}$ ▷ Concatenate tokens across chunks

D.3. Transformer encoder implementation details

We used the same code base as for B1 for the long-term event understanding task. Instead of giving video frames to a video tokenizer as input to a transformer encoder, we concatenated all video tokens corresponding to a given event, while adding spatial (Cam_{ID}: camera id) and positional encodings (ΔT_{event}: elapsed time w.r.t event start), and input them to the transformer encoder. We also added the source frame and patches from the offline token merging process to each individual video token as positional embedding ($Source$). We used the same encoder as a ViT-base model, without using pretraining weights (i.e. trained from scratch).

Models were trained for 300 epochs with a learning rate decreasing from 10^{-5} to 10^{-7} using the Adam-weighted optimizer. We applied the same sampling balancing strategy as in B1. We trained the activity recognition task with binary cross-entropy loss, and the other three tasks with categorical cross-entropy loss. We did not apply loss weighting to any of the four classification heads.

D.4. Camera-views ablation

We ablated camera-views: C_1, C_2 (Table S7). Models are tested on the same multi-view subset of events $E_{C1} \cup E_{C2}$, which are seen by either one or both views. Experiments demonstrate the advantage of using multiple views for complex tasks such as ActY recognition and number of individuals recognition.

Train events	ActY mAP	Ind. mAP	Avg. mAP
E_{C1}	0.379	0.474	0.407
E_{C2}	0.464	0.445	0.446
$E_{C1} \cup E_{C2} \setminus E_{C1} \cap E_{C2}$	0.480	0.445	0.456
$E_{C1} \cup E_{C2}$	0.522	0.510	0.501

Table S7. **Camera-view ablations for B2.** Models are trained with all positional embeddings and $r = 14$ on the joint recognition task. ActY: Activity; Ind. Number of individuals; Avg. Overall per-class

D.5. Models performance per class

We report F1-scores and average precisions per class computed similarly as for B1 (Tab. S9 and S8). We show the results when using a ToME [2] reduction factor of $r = 14$ and $r = 11$, and all types of positional encodings (Cam_{ID}, ΔT_{event}, $Source$).

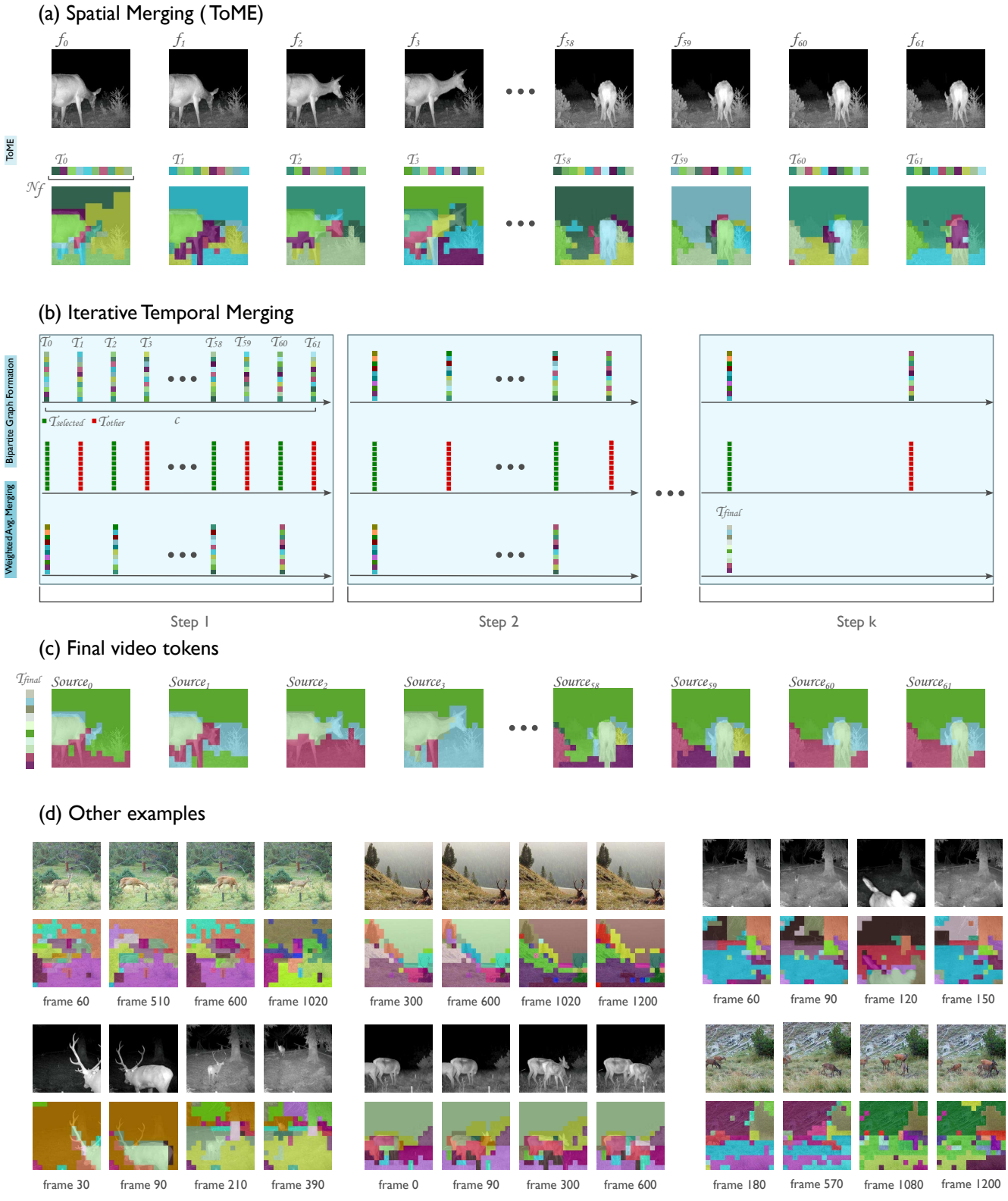

Figure S6. **Offline token merging strategy.** (a) We apply ToME [2] first spatially and then (b-c) temporally. (d) Multiple examples show initial frames and the source patch and frames for each final video token, corresponding to a unique color. For (a-c), refer to Algorithm 1 for variable names. For (a-b) we use a ToME reduction factor of 16, for (d) we use a ToME reduction factor of 14.

Class	Support	$r = 14$	$r = 11$
Activities AP			
Cam. reaction	5	**0.300**	0.080
Chasing	2	0.022	**0.175**
Courtship	5	0.385	**0.396**
Escaping	2	0.021	**0.038**
Foraging	49	0.863	**0.890**
Grooming	5	**0.648**	0.499
Marking	5	**0.557**	0.465
None	28	**0.959**	0.924
Playing	1	**0.042**	0.020
Resting	3	**0.459**	0.411
Unknown	30	**0.786**	0.755
Vigilance	35	**0.759**	0.751
Macro	170*	**0.483**	0.450
Species AP			
Fox	1	0.030	**0.500**
Hare	1	**0.020**	0.019
None	28	0.919	**0.978**
Red deer	53	0.938	**0.973**
Roe deer	3	0.118	**0.148**
Wolf	1	**0.033**	0.018
Macro	87*	0.343	**0.439**
Meteorological Conditions AP			
Clear	30	**0.803**	0.798
Overcast	15	0.466	**0.533**
Rainy	9	**0.416**	0.348
Sunny	32	**0.927**	0.858
Macro	86	**0.653**	0.634
Counting Individuals AP			
0	28	0.917	**0.985**
1	42	0.684	**0.798**
2	10	0.170	**0.348**
3+	6	0.014	**0.239**
Macro	86	0.478	**0.593**

Table S8. **Average precisions (AP) per class for the long-term event understanding benchmark (B2).** *Note that since there can be multiple species and activities per sample, this increases the total support since each label is considered independently.

Class	$r = 14$	$r = 11$	
Activities F1-scores			
Cam. reaction	5	0.222	0.000
Chasing	2	0.000	0.000
Courtship	5	**0.333**	**0.333**
Escaping	2	0.000	0.000
Foraging	49	**0.889**	0.871
Grooming	5	**0.400**	0.250
Marking	5	0.286	**0.333**
None	28	0.926	**0.964**
Playing	1	0.000	0.000
Resting	3	**0.500**	0.000
Unknown	30	**0.812**	0.704
Vigilance	35	0.658	**0.667**
Macro	170*	**0.419**	0.344
Species F1-scores			
Fox	1	0.000	0.000
Hare	1	0.000	0.000
None	28	**0.926**	**0.926**
Red deer	53	**0.909**	0.907
Roe deer	3	**0.222**	0.182
Wolf	1	0.000	0.000
Macro	87*	**0.343**	0.336
Meteorological Conditions F1-scores			
Clear	30	**0.778**	**0.778**
Overcast	15	**0.545**	0.300
Rainy	9	**0.333**	**0.333**
Sunny	32	0.939	**0.941**
Macro	86	**0.649**	0.588
Counting Individuals F1-scores			
0	28	0.926	**0.964**
1	42	0.690	**0.833**
2	10	**0.174**	0.000
3+	6	0.000	0.000
Macro	86	0.448	**0.449**

Table S9. **F1-scores per class for the long-term event understanding benchmark (B2).** *Note that since there can be multiple species and activities per sample, this increases the total support since each label is considered independently.

References

[1] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019. 1

[2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 8, 9

[3] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 5

[4] Jun Chen, Ming Hu, Darren J Coker, Michael L Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13052–13061, 2023. 5

[5] CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT), 2023. 2

[6] Jeannine Fluri, Pia Anderwald, Fränzi Korner-Nievergelt, Sonja Wipf, and Valentin Amrhein. The influence of wild ungulates on forest regeneration in an alpine national park. *Forests*, 14(6):1272, 2023. 1

[7] Andres Hernandez, Zhongqi Miao, Luisa Vargas, Rahul Dodhia, and Juan Lavista. Pytorch-wildlife: A collaborative deep learning framework for conservation, 2024. 1

[8] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35: 28708–28720, 2022. 5

[9] Maksim Kholiavchenko, Jenna Kline, Michelle Ramirez, Sam Stevens, Alec Sheets, Reshma Babu, Namrata Banerji, Elizabeth Campolongo, Matthew Thompson, Nina Van Tiel, et al. 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, 2024. 5

[10] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 5

[11] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 5

[12] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 1