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

Valentin Gabeff[1] Haozhe Qi[1] Brendan Flaherty[1] Gencer Sumbül[1]

Alexander Mathis[1] alexander.mathis@epfl.ch

Devis Tuia[1] devis.tuia@epfl.ch

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

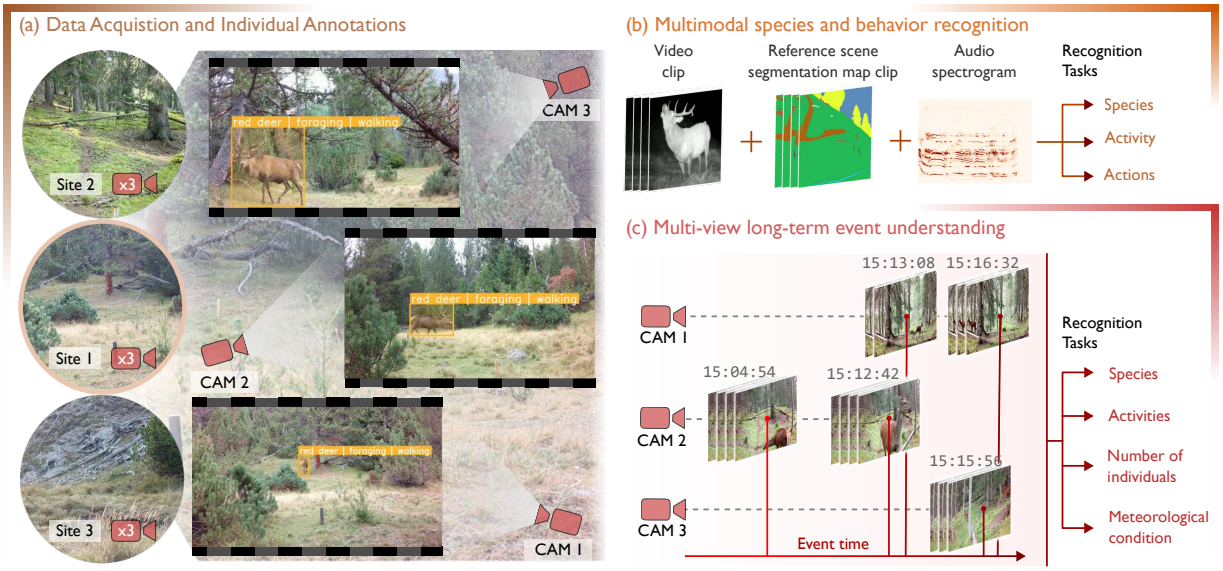

Figure 1. **MammalAlps: Overview of the data and proposed benchmarks.** (a) Nine camera traps were installed at three different sites in the Swiss National Park and recorded video and audio of animal activity for six weeks. (b) We propose a multimodal species and hierarchical behavior recognition benchmark for wildlife based on video, audio and segmentation maps. (c) We propose the first multi-view, long-term event understanding benchmark that aims at summarizing long-term ecological events into meaningful information for behavioral ecology.

Abstract

Monitoring wildlife is essential for ecology and ethology, especially in light of the increasing human impact on ecosystems. Camera traps have emerged as habitat-centric sensors enabling the study of wildlife populations at scale with minimal disturbance. However, the lack of annotated video datasets limits the development of powerful video understanding models needed to process the vast amount of fieldwork data collected. To advance research in wild animal behavior monitoring we present MammAlps, a multimodal and multi-view dataset of wildlife behavior monitoring from 9 camera-traps in the Swiss National Park. MammAlps contains over 14 hours of video with audio, 2D segmentation maps and 8.5 hours of individual tracks densely labeled for species and behavior. Based on 6‘135 *single animal clips, we propose the first hierarchical and multimodal animal behavior recognition benchmark using audio, video and reference scene segmentation maps as inputs. Furthermore, we also propose a second ecology-oriented benchmark aiming at identifying activities, species, number of individuals and meteorological conditions from 397 multi-view and long-term ecological events, including false positive triggers. We advocate that both tasks are complementary and contribute to bridging the gap between machine learning and ecology. Code and data are available at https://github.com/eceo-epfl/MammAlps.*

1. Introduction

Due to unprecedented rates of biodiversity loss, monitoring wild animals behavior has become a crucial task in conservation ecology and wildlife management [6, 44]. More broadly, understanding animal behavior is important across many fields [15, 32, 46]. Wild animal behavior can be monitored with a variety of sensors. Animal-centric sensors such as bio-loggers are traditionally used to obtain broad behavioral information over large spatio-temporal extents [14, 15, 26, 46]. Conversely, habitat-centric imagery acquired from camera traps [11, 15, 17, 46] provides more fine-grained information on wildlife-environment interactions. With the most recent camera trap setups achieving enhanced battery life and storage, it is now becoming possible to study animal behavior at scale in the wild with video traps [10, 29, 30].

However, these advances in camera traps hardware also drastically increased dataset sizes, along with the complexity of the behavioral traits observed and to be quantified. To address this challenge, deep learning (DL) models were developed to support the analysis of wild animal videos for behavior recognition, segmentation and detection [5, 8, 9, 20, 28, 38, 39, 52].

Simultaneously, wild animal datasets are being curated to support the training of DL models to effectively classify a wide range of behaviors across many species and geographical regions. Existing datasets annotated for wild animal behavior can generally be categorized in either fieldwork data, or internet scrapped data. Fieldwork data is generally constrained to a small geographical location, focuses on one or few species and mostly contains common behaviors [46]. They have the advantage of representing "real world" data. In contrast, large scale datasets scrapped from the internet such as MammalNet [13] contain a rich set of behaviors and species, potentially with an over-representation of rare behaviors that are challenging to acquire in field surveys. Yet, they still suffer from an important domain gap between the videos scrapped (*e.g*. scenes from documentaries) and the type of data used by experts (*e.g*. camera trap imagery). Both sources of data are complementary, but the field still lacks publicly available and curated fieldwork datasets to unify them. Additionally, insights from ethology and neuroscience can improve animal behavior recognition models by better representing behaviors in these wild animal datasets [2, 43]. Indeed, currently available datasets all categorize behaviors as independent classes, often without any kind of behavioral structure.

To address these shortcomings and advance research at the interface between computer vision and behavioral ecology, we collected and annotated MammAlps, a unique camera-trap video dataset consisting of footage acquired at three different sites in the Northern European Alps, at the Swiss National Park (SNP). MammAlps contains 8.5 hours of curated mammals behavior recordings. Three cameras with varying level of field-of-view overlap were deployed at each site to provide multi-view information (Fig. 1a). Additionally, cameras built-in microphones were used to acquire audio and a segmentation map was created for each camera reference scene. To better represent the hierarchical nature of animal behavior, individual tracklets were densely annotated at two levels of complexity, *i.e*. high-level activities and low-level actions.

Along with the dataset, we propose the first multimodal species and behavior recognition benchmark from the camera trap video clips, the associated audio recordings and the reference scene segmentation map clips (Fig. 1b). We also provide a second benchmark consisting of summarized annotations at the event level (*e.g*. a set of multiple videos capturing the same ecological scene) for long-term scene understanding task (Fig. 1c). This task consists of multiple predictive objectives at the event level from multiple views: Listing all detected species along with their activities, classifying the number of individuals into group sizes, and classifying meteorological conditions. In this second task, spatio-temporal precision is traded for larger spatio-temporal context which suits different needs in behavioral ecology.

Our contributions are:

- A unique multimodal and multi-view camera-trap video dataset containing 8.5 hours of densely annotated wild mammals behavior acquired in the Swiss Alps (Fig. 1a).
- A multimodal species and behavior recognition benchmark to foster method development for wildlife monitoring (Fig. 1b).
- A unique multi-view and long-term event understanding benchmark designed to meet key unaddressed needs of ecologists, along with an offline method to condense long events into few visual tokens. (Fig. 1c).

2. Related Works

Wild animal behavioral datasets. Thanks to advances in sensor design and availability [15, 46], a number of fieldwork-based datasets for wildlife behavior monitoring from videos became available recently (Tab. 1). LoTE offers a collection of camera trap datasets (images and videos) from South East Asia [29]. While a subset of the images are labeled with bounding boxes, the behavior annotations for the video dataset are not at the individual level. Brookes et al. share a camera trap video dataset of great apes in Africa [10]. Each video is associated with a set of behavior labels that occur within the video, and a subset of the dataset also comprises individual tracks. A larger part of the dataset contains richer behavior descriptions, yet without individual tracks. The meerkat behavior dataset contains rich behavioral annotations at the individual level [37].

Dataset	Video hours (processed)	Source	# Videos	# Species	# Behav.	Annot. level	Hierarch. Behav.	Multi-Modal	Multi-View
Meerkats [37]	4	Zoo	35	1	15	individual	✗	✗	✗
ChimpACT [31]	2	Zoo	163	1	23	individual	✗	✓*	✗
KABR [27]	10	Drone	13k	3	8	individual	✗	✗	✗
BaboonLand [19]	20	Drone	30k	1	12	individual	✗	✗	✗
PanAf20k [10]	80	CT	20k	2	18	video	✗	✗	✗
PanAf500 [10]	2	CT	500	2	9	individual	✗	✗	✗
LoTE [29]	N/A	CT	10k	11	21	video	✗	✓*	✗
PandaFormer [30]	2	CT	1431	1	5	video	✗	✗	✗
AnimalKingdom [33]	50	Youtube	30k	850	140	video	✗	✓*	✗
MammalNet [13]	394	Youtube	20k	173	12	video	✗	✗	✗
MammAlps (clips)	8.5	CT	6k	5	11+19	individual	✓	✓	✗
MammAlps (events)	14.5	CT	2384	5	11	event	✓	✓*	✓

Table 1. **Prominent and publicly available video datasets of wild animals behavior monitoring.** *Multimodal data is available but it is not used for an action recognition benchmark. MammAlps is available at 10.5281/zenodo.15040901.

Similarly, ChimpACT contains individual level annotations, along with animal body pose annotations [31]. However, both datasets are recorded in zoos. KABR and BaboonLand use drone footage and provide dense behavior labels for four African species at the individual level [19, 27]. PandaFormer [30] contains almost two hours of wild pandas recordings spanning five behaviors. Recently, a 1-h long dataset with recordings of 17 bird species and seven behavioral classes became available[36].

Scraping the web can also yield relevant datasets. Animal Kingdom [33] contains 50 hours of behavioral videos spanning 850 species and 140 behavioral descriptions. MammalNet [13] is the largest dataset of wild animal videos, containing around 400 hours of footage from different sources (*e.g*. documentaries, zoos) depicting 173 mammal species and around 20 behaviors shared across mammals. While some of these works propose exclusively low-level behavior recognition [27, 30] (*e.g*. actions like walking, grazing), others annotate more high-level behaviors [10, 13, 19] (*e.g*. chasing, hunting).

Multi-modal action recognition. With the development of the transformer architecture [47] and expanding computational power, leveraging multimodal data for action understanding was increasingly feasible[12, 40, 41, 49, 50, 54]. LaViLa [54] learns video representations from pre-trained large language models. TIM [12] designs time interval encodings to incorporate visual and audio events. In the domain of wildlife behavior understanding, researchers sometimes use multiple sensors (*i.e*. modalities) conjointly to monitor animal behavior [1, 3, 23]. In [3], the authors make a first attempt at using audio-visual inputs from camera traps to classify two specific wild primate behaviors.

Overall, our work is most similar to [3, 10, 19, 27]. On top of the dense behavioral annotations at the individual level, our dataset brings additional value over all previous datasets as (1) we follow a hierarchical representation of behavior [2, 43], and propose separate tasks for low-level action and high-level activity recognition; (2) we provide audio recordings and segmentation maps from the fixed camera reference scenes to further guide models via multiple modalities; (3) events are being recorded from up to three points-of-view, which provides detailed information for long-term event understanding (Tab. 1); (4) MammAlps is the only camera trap video dataset focusing on species from the European Alps, which is a region particularly vulnerable to climate change [22, 48].

3. MammAlps dataset and proposed benchmarks

In this section, we detail the dataset collection and preprocessing (Sec. 3.1) of MammAlps, as well as the annotation protocol (Sec. 3.2) and the two benchmarks proposed (Sec. 3.3 and 3.4). For clarity, we defined a list of terms used throughout the study in Tab. 2.

3.1. Data collection and pre-processing

Data collection. Nine camera traps (Browning's Spec Ops Elite HP5) were installed in the Swiss National at three sampling sites representing different ecological habitats. The project was approved by the Research Commission of the National Park. For each site, three cameras were positioned with different perspectives, in order to capture the scene from multiple angles and to provide more context for interpreting behavior (Fig. 1a). Triggered by motion, videos were collected for six weeks (between June and August 2023) during daytime and nighttime. At nighttime, videos are recorded with an IR flash invisible to the species of interest. Videos are captured at high resolution (1920×1080) with a frame rate of 30 FPS. Cameras recorded 43 hours of raw footage, with varying levels of false positive triggers. Data acquisition details and sampling site descriptions can be found in the Supplementary Materials.

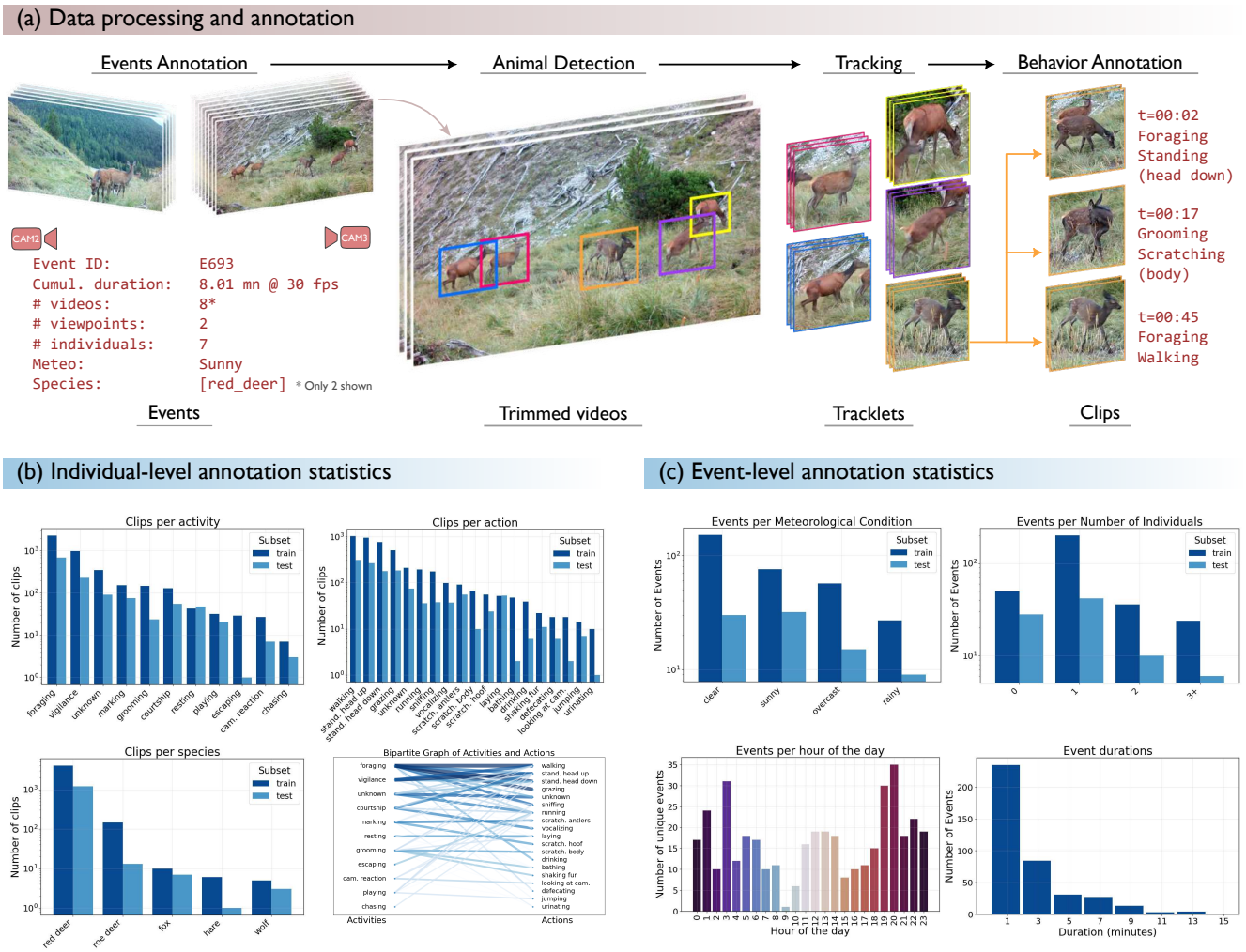

Figure 2. **Data processing pipeline and analysis.** (a) Raw videos were first aggregated into events. We then applied MegaDetector [4, 24] and ByteTrack [53] to generate animal tracklets, which were manually corrected. We annotated these tracklets for species and behavior at two levels of complexity. (b-c) Various statistics of the dataset.

Data pre-processing. The data processing pipeline is as follows (Fig. 2a): *raw videos* were first grouped into *events*, corresponding to periods without more than five minutes of inactivity at the corresponding site. We then removed false positive videos and trimmed the true positive ones by running them through MegaDetector [4, 24]. The dense animal detection predictions of the *trimmed video* were used as inputs to an adapted version of ByteTrack [53] yielding individual *tracks*. The *tracks* were then manually corrected in CVAT [16] to remove identity switches, lost tracks, and any remaining false positive segment. We did not correct localization errors (*e.g.* body parts outside of bounding boxes) since our proposed benchmarks do not require this level of spatial precision. Each animal track was converted into a video *tracklet* (380×380) padded with background to avoid distortions. We further partition the tracklets into short video *clips* displaying a single behavioral expression (Sec. 3.2). Data processing and model parameters are detailed in the Supplementary Materials.

Cameras synchronization and temporal drift. Cameras have a built-in accuracy of one minute and are subject to drift over time (see Supplementary Materials). Temporal drift between camera pairs extended up to one minute in Site 1. This drift further increases the difficulty of the second benchmark, while reflecting fieldwork conditions. Auditory data could be used for syncing.

3.2. Data annotation

Individual counts and meteorological conditions were annotated at the event level, while behaviors and species were annotated at the individual level (Fig. 2b) and then aggregated at the event level when necessary (Fig. 2c).

Raw video	Raw camera trap recording of fixed duration.
Event	Collection of raw videos corresponding to an ecological event. Events are separated by a period of inactivity of at least 5 minutes. The events are used as input for the long-term scene understanding task (Sec. 3.4).
Trimmed video	Segment within a raw video contained between the first and the last MegaDetector [4, 24] detections.
Track	Sequence of bounding boxes with associated individual identifier, built automatically from ByteTrack [53] and MegaDetector predictions [4, 24] and manually adjusted in CVAT [16].
Tracklet	Animal-centered video of aspect ratio 1:1 cropped from an animal track labeled for species and densely annotated for behavior.
Clip	Segment within a tracklet with a single behavioral expression. The clips are used as input for the behavior recognition benchmark (Sec. 3.3).

Table 2. **Terminology used at the different stages of the data processing and annotation pipeline.**

Species and behavior annotations. Animal *tracklets* were densely labeled in CVAT [16] for species and behaviors. We focused on five species: red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), fox (*Vulpes vulpes*), wolf (*Canis lupus*) and mountain hare (*Lepus timidus*). Behaviors were annotated by two annotators at two levels of complexity [2, 43]: 1) *Actions* (*e.g.* walking, grazing), are stereotypical combinations of a few basic movements and can usually be identified from a few frames; 2) *Activities*, which generally require longer spatio-temporal context and may be the composition of multiple *actions* (*e.g.* foraging) or the interaction between different individuals of the same species (*e.g.* courtship) or between different species (*e.g.* chasing). Each frame is labeled with one activity and either one or two non-mutually exclusive actions. For both levels, we included an *unknown* class, which indicates a behavior that could not be identified, either because of occlusion or by lack of information.

Individual counts. The total number of individuals in an event is determined by visual examination of all the videos from all viewpoints recording it. Automatic aggregation from the track annotations was not possible, since camera traps were not perfectly temporally synchronized nor spatially referenced in a 3D model. Individual counts per species were summed and grouped into four categories (0, 1, 2, 3+). Thus, the counting task assesses the group size (none, individual, pair or group).

Meteorological conditions. During this process, meteorological conditions were visually determined and categorized into four general conditions: *clear weather* (including day and night), *sunny*, *overcast* and *rainy* (including day and night).

Reference scenes segmentation. Since camera traps are placed at a fixed position, a single segmentation map was annotated for each of the scene's viewpoints. A reference picture (without animals) was taken with each camera after the video acquisition. We annotated the segmentation masks for ten classes using CVAT [16]. Some classes are unique to a site (*e.g.* water pound only occurs at Site 3), while others are shared across the three sites (*e.g.* grass). The segmentation maps are then processed into video clips by generating a tracklet based on the animal tracks for every video clip. Hence, these segmentation map clips represent the background classes surrounding (and behind) the animal, synchronized in location and time to the animal video clips. Examples are shown in the Supplementary Materials.

3.3. Multimodal Species and Behavior Recognition Benchmark: B1

Action recognition is a common challenge across multiple wildlife monitoring datasets [10, 13, 19, 27, 30, 33]. While all of them are limited to RGB visual inputs, we enrich the video modality with audio and reference scene segmentation maps. We hypothesized that audio can help identify some specific actions such as vocalization and walking, while segmentation maps of the reference scenes may guide classification for behaviors involving specific environmental features (*e.g.* drinking from a water source). The dataset for this task (B1) consists of 6135 short video clips spanning 11 activities, 19 actions and 5 species and a total of 8.5 hours of recordings. Because a sample can be annotated with up to two actions, action recognition is a multi-label classification task, while species and activity recognitions are multi-class ones. We refer to *behavior* recognition as the recognition task that encompasses both action and activity recognition. The data was randomly split in a train, validation and test set at the day level, while matching label distributions across splits. Clips that contained occlusions were labeled as unknown activity and actions since we considered that a model cannot provide a reliable behavioral estimate with such limited context.

3.4. Multi-view long-term event understanding Benchmark: B2

Benchmark B1 is a computer science-oriented benchmark focused on a single sensor (with multiple modalities). However, to reliably identify events all the available sensors

Training task	Spe.(↑)	ActY.(↑)	ActN.(↑)
Single Task Prediction			
Spe.	0.537	-	-
ActY.	-	0.440	-
ActN.	-	-	0.447
Joint Task Prediction			
Spe. + ActY.	0.437	**0.443**	-
Spe. + ActN.	**0.539**	-	0.442
ActY. + ActN.	-	0.442	0.427
All.	0.487	0.428	**0.458**

Table 3. **Comparison of single vs. joint task prediction (B1).** mAP for single and joint task predictions from video clips. In all cases, VideoMAE is used as the base model [45]. ActY.: Activities; ActN.: Actions; Spe.: Species.

should be used. Additionally, understanding events requires long-term context understanding (more than 16 frames), especially when expressed activities are temporally related to other individuals (*e.g.* prey-predator relationships) or are composed of multiple actions (*e.g.* foraging). Being able to efficiently summarize events into broad categories is also necessary to facilitate the annotation process of very large camera trap datasets. To this end, we propose a second, long-term event understanding benchmark (B2) that takes as input the raw multi-view videos of a given event with the objective of predicting high-level behaviors (activities), the species detected, the number of individuals (in the grouped categories defined in Sec. 3.2) and the meteorological conditions. Activity and species recognition are multi-label classification tasks, while meteorological condition and number of individuals are multi-class classification ones. This benchmark is particularly challenging as the event duration varies greatly (from 1 second to 12 minutes), activities and species are highly imbalanced, and counting individuals requires to intelligently integrate information across camera views and over time. The dataset for task B2 is composed of 397 events, 2384 videos, totaling 14.2 hours of recordings. Similarly as for Sec. 3.3, the events were randomly split (at the day level) in a train and test set. Data spans 11 activities and 5 species (the same as for Sec. 3.3), 4 group size categories and 4 meteorological conditions.

For both benchmarks B1 and B2, we report the mean average precision (mAP) averaged over the label categories of each task, which is a convenient metric to compare tasks that are either multi-label or multi-class. When applicable, for joint predictions on all tasks, we report the mAP averaged over all label categories of all tasks in column 'Avg.'. Models for benchmarks B1 and B2 were trained with four and eight A100 GPUs, respectively.

Modalities	Spe.(↑)	ActY.(↑)	ActN.(↑)	Avg.(↑)
V	0.487	0.428	0.458	0.453
S	0.414	0.188	0.171	0.211
A	0.223	0.207	0.161	0.184
V+S	0.457	0.399	0.375	0.394
A+S	0.334	0.262	0.257	0.270
V+A	**0.503**	**0.475**	**0.463**	**0.472**
V+A+S	0.482	0.452	0.417	0.438

Table 4. **Hierarchical action recognition from multimodal data (B1).** mAP for joint task prediction from multimodal data using VideoMAE as the base model [45]. V: video clips; A: audio spectrograms; S: segmentation map clips; ActY.: Activities; ActN.: Actions; Spe.: Species; Avg.: overall per-class average. Note: the 'V' row, corresponds to the last row of Tab. 3.

4. Experiments

4.1. B1: Multi-modal species and behavior recognition

In order to utilize multi-modal data for action recognition, we adapted the VideoMAE model [45] so that it could take video, audio and segmentation maps as inputs simultaneously. Specifically, we sampled 16 frames within 5 seconds of randomly selected windows for both video and one-hot encoded segmentation map clips. For the audio inputs, we first found the audio clip simultaneous to the video clip and then transformed and tokenized the original audio signal to a spectrogram, similarly to AudioMAE [25]. To compensate for the label imbalance, clips were sampled with a probability proportional to the sum of the inverse label frequencies for each class. Because test clips greatly vary in their duration, we aggregated predictions over ten random samples of 16 frames for every test clip.

When considering only the video modality, VideoMAE leads to improved results for all tasks when considering the joint task prediction (Tab. 3). Multi-modal results indicate that combining the audio and video modalities improves the performance over the video-only model ($+0.019$ mAP), with an overall class-average mAP of 0.472 (Tab. 4). However, in our baseline model, the reference segmentation map clips did not improve over their video-only or video-audio counterparts, but they did increase the performance of the audio-only model ($+0.059$ mAP) suggesting that this modality contains distinct information relevant to the tasks. More details, baselines and results per class can be found in the Supplementary Materials.

4.2. B2: Multi-view long-term event understanding

To the best of our knowledge, due to the size no existing video model can process multi-view and long-term (ecological) data for our task of interest, so we propose a simple

Training task	r	Cont. Len.	Spe.(↑)	ActY.(↑)	Met. Cond.(↑)	Indiv.(↑)	Avg. (↑)
Single Task Prediction							
Spe.	14	4096	**0.481**	-	-	-	
ActY.	14	4096	-	0.478	-	-	
Met. Cond.	14	4096	-	-	**0.681**	-	
Indiv.	14	4096	-	-	-	0.592	
Joint Task Prediction							
All	14	4096	0.343	**0.483**	0.653	0.478	0.476
All	11	8192	0.439	0.450	0.634	**0.593**	**0.498**

Table 5. **mAP for long-term event understanding from the multi-view events (B2).** All models use the transformer encoder from ViT-Base. "r": ToME [7] reduction factor. A larger reduction factor leads to more patches being merged at the frame level and fewer video tokens; "Cont. Len.": context length: number of tokens per sample; ActY.: Activities; Spe.: Species.; Met. Cond.: Meteorological Conditions; Indiv.: Number of individuals categories.; Avg.: overall per-class average.

method as baseline. Taking inspiration from token merging in vision transformers (ToME) [7] and follow-up works focusing on merging tokens online over time [35, 42], we propose a fully offline method to merge the frame patch tokens from a pretrained vision-MAE transformer first in the spatial and then in the temporal dimensions (see Supplementary Materials). To account for the large range of video durations, we perform token merging over time in blocks of fixed duration and concatenate the resulting tokens, so that long videos ultimately yield more tokens than short ones. We add three cosine positional embeddings [18] to every video token: 1) The information from the camera identity for the given site (Cam_{ID}); 2) the elapsed time with respect to the event start (ΔT_{event}); and 3) the frame and patch identities of the source frame tokens composing each individual video token (see Supplementary Materials for details). We input these condensed video tokens to a transformer backbone with four output heads, each corresponding to one of the predictive tasks. We set a maximum input context length based on the longest event and pad shorter ones with masked tokens.

The best joint recognition performance (average per-class mAP of 0.498) was achieved with a ToME [7] reduction factor (r) at the frame level of 11, yielding between 65 and 390 tokens per video depending on their duration (Tab. 5). When $r = 11$, the overall mAP is slightly higher ($+0.022$) than when $r = 14$ (yielding between 29 and 174 tokens per video) but not on all tasks.

We evaluated the model performance when ablating r and the different positional embeddings (Tab. 6). We focused on the task where these embeddings are thought to contribute the most: number of individuals classification. Here, the model with all positional embeddings lead to the highest scores independent of the value of r. While with $r = 14$, the highest increase is observed for the single task ($+0.078$ mAP), this is the opposite when $r = 11$ (increase in joint task mAP of $+0.109$). More details, ablations and results per class can be found in the Supplementary Materials.

5. Discussion

Contributions of the audio and segmentation map modalities. Adding the audio modality improves the overall performance over a video-only model (Tab. 4). When looking at specific classes (Supplementary Materials), classes with distinct sounds such as marking or vocalizing improved for the audio-video model over the video-only model ($+0.20$ and $+0.09$ F1-scores, respectively). Conversely, the resting activity which is mostly silent remains with a low F1-score (from 0.19 with video to 0.15 with the audio and video). While the reference segmentation map modality did not improve performance when combined with videos, it did improve over the audio-only model especially on classes involving specific scene features such as grazing ($+0.08$) or walking ($+0.09$) despite that these actions already emit some sound.

Impact of token merging on classifying the number of individuals. In B2 (Sec. 3.4), classifying the number of individuals is particularly challenging as the model needs to integrate information from multiple views and videos. Hence the model needs to extract individual identities. Yet, it is common that tokens representing different animals become merged by our offline approach. This is expected as the algorithm merges tokens based on similarity and two different individuals might show little visual differences when they are from the same species. We address this issue by both increasing the number of tokens per video and by adding a positional embedding to the video tokens that contains summarized information about their source frames and patches. With the former, we aim that different individuals are represented by different tokens, while with the

ToME parameters		Positional embeddings			mAP	
r	Cont. Len./BS	Cam_{ID}	ΔT_{event}	Source	Indiv.(↑) (Single|Joint)	Indiv. 2+ (↑) (Single|Joint)
14	4096/32				0.514|0.505	0.222|0.120
14	4096/32	✓	✓		0.562|0.461	0.192|0.112
14	4096/32	✓	✓	✓	**0.592**|0.478	**0.329**|0.156
11	8192/8				0.502|0.566	0.145|0.223
11	8192/8	✓	✓		0.527|0.484	0.200|0.136
11	8192/8	✓	✓	✓	0.527|**0.593**	0.184|**0.294**

Table 6. **Ablation study on the effect of the number of video tokens, and the addition of the different positional embeddings on the number of individuals recognition task (B2).** All models are the transformer encoder from ViT-Base. "r": ToME [7] reduction factor."Cont. Len.": context length: maximum number of tokens per sample; BS: Batch Size; ActY.: Activities; Spe.: Species; Met. Cond.: Meteorological Conditions; Indiv.: Number of individuals in categories; 'Indiv. 2+': Predictions for groups containing more than a single individual. Results for the 'Indiv.' and 'Indiv. 2+' tasks are provided for both the single and joint task prediction.

latter, we indicate if a single video token comes from one or multiple discontinuous spatio-temporal segments. The ablation realized suggests that this design successfully increases the performance for test events with more than two individuals (Tab. 6).

Hierarchical description of behaviors. The decision to represent behaviors as a combination of one activity and one or two actions seems to facilitate learning, as suggested by the higher performance of joint recognition models over single ones (Tab. 3). However, the hierarchical relationship between activities and actions could be further exploited in both benchmarks. For example in B2 (Sec. 3.4), predicted actions could influence higher-level activity prediction, *e.g.* chasing is a high-level activity composed of the running action in a prey-predator context.

Dataset bias and limitations. First, annotating animal behavior is a complex task, as behavior categorization remains a subjective process, prone to annotators' biases. This concerns particularly social behaviors such as those related to courtship. These uncertainties lead to varying level of label noise per class. To mitigate these biases, uncertain samples were tagged and discussed among annotators to ensure annotation consistency. Additionally, the set of species in the dataset remains limited, as all three sites were located in the same National Park, at the same elevation and over a short period of time (*i.e.* until the camera battery exhaustion). This limits the possibility to learn common behavioral expression across species as done in [13]. Other mammal species that are common in the European Alps are absent from the dataset in its current form. Likewise, despite containing 80GB of raw video data, the dataset of the long-term video understanding benchmark only contains 86 test events, which may be insufficient to properly assess the performance of the model on rare classes. However, this is the first dataset considering events-level information in the wildlife domain and which defines a new task for the field. Future surveys (by the authors themselves and desirably by the wider and very active 'AI for ecology' community) will progressively increase the quantity of the data for this task.

6. Conclusion and outlook

We develop MammalAlps, a novel multimodal, multi-view camera-trap video dataset of annotated hierarchal, mammalian behavior in the Swiss Alps. We propose two benchmarks to motivate the development of behavior understanding methods for ecology, based on event and clip annotations. In particular, we propose the first long-term event understanding task that aims to summarize long-term ecological events into meaningful information for the ecologists. We believe this task is particularly interesting to spur research on efficient architectures that can flexibly integrate multiple sources of information over diverse temporal ranges to reach better conclusions.

MammAlps can be extended in a multitude of ways, for instance by adding new modalities such as animal body pose [52], segmentation masks [34], depth [51], and language [9, 21], as all these modalities introduce complimentary behavior information.

By publicly sharing MammAlps, we aim to provide rich behavioral annotations that can fuel the development of holistic animal behavior understanding models. These models have the potential to identify and quantify observable behavioral traits of wild individuals, opening the doors to AI-assisted data processing and scientific discoveries.

Acknowledgments: We thank members of the Mathis Group for Computational Neuroscience & AI (EPFL) and of the Environmental Computational Science and Earth Observation Laboratory (EPFL) for their feedback and fieldwork efforts. We also thank members of the Swiss National Park monitoring team for their support and feedback. This project was partially funded by EPFL's SV-ENAC I-PhD

program (G.V.), Boehringer Ingelheim Fonds PhD stipend (H.Q.) and Swiss SNF grant (320030-227871).

References

[1] Daniel Alempijevic, Ephrem M Boliabo, Kathryn F Coates, Terese B Hart, John A Hart, and Kate M Detwiler. A natural history of chlorocebus dryas from camera traps in lomami national park and its buffer zone, democratic republic of the congo, with notes on the species status of cercopithecus salongo. *American Journal of Primatology*, 83(6): e23261, 2021. 3

[2] David J Anderson and Pietro Perona. Toward a science of computational ethology. *Neuron*, 84(1):18–31, 2014. 2, 3, 5

[3] Max Bain, Arsha Nagrani, Daniel Schofield, Sophie Berdugo, Joana Bessa, Jake Owen, Kimberley J Hockings, Tetsuro Matsuzawa, Misato Hayashi, Dora Biro, et al. Automated audiovisual behavior recognition in wild primates. *Science advances*, 7(46):eabi4883, 2021. 3

[4] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019. 4, 5

[5] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context r-cnn: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13075–13085, 2020. 2

[6] Oded Berger-Tal, Tal Polak, Aya Oron, Yael Lubin, Burt P Kotler, and David Saltz. Integrating animal behavior and conservation biology: a conceptual framework. *Behavioral Ecology*, 22(2):236–239, 2011. 2

[7] 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. 7, 8

[8] Otto Brookes, Majid Mirmehdi, Hjalmar Kühl, and Tilo Burghardt. Triple-stream deep metric learning of great ape behavioural actions. *arXiv preprint arXiv:2301.02642*, 2023. 2

[9] Otto Brookes, Majid Mirmehdi, Hjalmar Kuhl, and Tilo Burghardt. Chimpvlm: Ethogram-enhanced chimpanzee behaviour recognition. *arXiv preprint arXiv:2404.08937*, 2024. 2, 8

[10] Otto Brookes, Majid Mirmehdi, Colleen Stephens, Samuel Angedakin, Katherine Corogenes, Dervla Dowd, Paula Dieguez, Thurston C Hicks, Sorrel Jones, Kevin Lee, et al. Panaf20k: a large video dataset for wild ape detection and behaviour recognition. *International Journal of Computer Vision*, pages 1–17, 2024. 2, 3, 5

[11] Anthony Caravaggi, Peter B Banks, A Cole Burton, Caroline MV Finlay, Peter M Haswell, Matt W Hayward, Marcus J Rowcliffe, and Mike D Wood. A review of camera trapping for conservation behaviour research. *Remote Sensing in Ecology and Conservation*, 3(3):109–122, 2017. 2

[12] Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, and Dima Damen. Tim: A time interval machine for audio-visual action recognition. *arXiv preprint arXiv:2404.05559*, 2024. 3

[13] 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. 2, 3, 5, 8

[14] Steven J Cooke. Biotelemetry and biologging in endangered species research and animal conservation: relevance to regional, national, and iucn red list threat assessments. *Endangered species research*, 4(1-2):165–185, 2008. 2

[15] Iain D Couzin and Conor Heins. Emerging technologies for behavioral research in changing environments. *Trends in Ecology & Evolution*, 38(4):346–354, 2023. 2

[16] CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT), 2023. 4, 5

[17] Zackary J Delisle, Elizabeth A Flaherty, Mackenzie R Nobbe, Cole M Wzientek, and Robert K Swihart. Next-generation camera trapping: systematic review of historic trends suggests keys to expanded research applications in ecology and conservation. *Frontiers in Ecology and Evolution*, 9:617996, 2021. 2

[18] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[19] Isla Duporge, Maksim Kholiavchenko, Roi Harel, Dan Rubenstein, Meg Crofoot, Tanya Berger-Wolf, Stephen Lee, Scott Wolf, Julie Barreau, Jenna Kline, et al. Baboonland dataset: Tracking primates in the wild and automating behaviour recognition from drone videos. *arXiv preprint arXiv:2405.17698*, 2024. 3, 5

[20] Michael Fuchs, Emilie Genty, Klaus Zuberbühler, and Paul Cotofrei. Asbar: an animal skeleton-based action recognition framework. recognizing great ape behaviors in the wild using pose estimation with domain adaptation. *bioRxiv*, pages 2023–09, 2023. 2

[21] Valentin Gabeff, Marc Rußwurm, Devis Tuia, and Alexander Mathis. Wildclip: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. *International Journal of Computer Vision*, pages 1–17, 2024. 8

[22] Andreas Gobiet, Sven Kotlarski, Martin Beniston, Georg Heinrich, Jan Rajczak, and Markus Stoffel. 21st century climate change in the european alps—a review. *Science of the total environment*, 493:1138–1151, 2014. 3

[23] Jonathan M Handley, Andréa Thiebault, Andrew Stanworth, David Schutt, and Pierre Pistorius. Behaviourally mediated predation avoidance in penguin prey: in situ evidence from animal-borne camera loggers. *Royal Society open science*, 5 (8):171449, 2018. 3

[24] Andres Hernandez, Zhongqi Miao, Luisa Vargas, Rahul Dodhia, and Juan Lavista. Pytorch-wildlife: A collaborative deep learning framework for conservation, 2024. 4, 5

[25] 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. 6

[26] Roland Kays, Margaret C Crofoot, Walter Jetz, and Martin Wikelski. Terrestrial animal tracking as an eye on life and planet. *Science*, 348(6240):aaa2478, 2015. 2

[27] 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. 3, 5

[28] 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. 2

[29] Dan Liu, Jin Hou, Shaoli Huang, Jing Liu, Yuxin He, Bochuan Zheng, Jifeng Ning, and Jingdong Zhang. Loteanimal: A long time-span dataset for endangered animal behavior understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20064–20075, 2023. 2, 3

[30] Jing Liu, Jin Hou, Dan Liu, Qijun Zhao, Rui Chen, Xiaoyuan Chen, Vanessa Hull, Jindong Zhang, and Jifeng Ning. A joint time and spatial attention-based transformer approach for recognizing the behaviors of wild giant pandas. *Ecological Informatics*, 83:102797, 2024. 2, 3, 5

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

[32] Mackenzie Weygandt Mathis and Alexander Mathis. Deep learning tools for the measurement of animal behavior in neuroscience. *Current opinion in neurobiology*, 60:1–11, 2020. 2

[33] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19023–19034, 2022. 3, 5

[34] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 8

[35] Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou. Testa: Temporal-spatial token aggregation for long-form video-language understanding. *arXiv preprint arXiv:2310.19060*, 2023. 7

[36] Javier Rodriguez-Juan, David Ortiz-Perez, Manuel Benavent-Lledo, David Mulero-Pérez, Pablo Ruiz-Ponce, Adrian Orihuela-Torres, Jose Garcia-Rodriguez, and Esther Sebastián-González. Visual wetlandbirds dataset: Bird species identification and behavior recognition in videos. *arXiv preprint arXiv:2501.08931*, 2025. 3

[37] Mitchell Rogers, Gaël Gendron, David Arturo Soriano Valdez, Mihailo Azhar, Yang Chen, Shahrokh Heidari, Caleb Perelini, Padriac O'Leary, Kobe Knowles, Izak Tait, et al. Meerkat behaviour recognition dataset. *arXiv preprint arXiv:2306.11326*, 2023. 2, 3

[38] Frank Schindler and Volker Steinhage. Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. *Ecological Informatics*, 61: 101215, 2021. 2

[39] Frank Schindler, Volker Steinhage, Suzanne TS van Beeck Calkoen, and Marco Heurich. Action detection for wildlife monitoring with camera traps based on segmentation with filtering of tracklets (swift) and mask-guided action recognition (maroon). *Applied Sciences*, 14(2):514, 2024. 2

[40] Ketul Shah, Anshul Shah, Chun Pong Lau, Celso M de Melo, and Rama Chellappa. Multi-view action recognition using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3381–3391, 2023. 3

[41] Md Salman Shamil, Dibyadip Chatterjee, Fadime Sener, Shugao Ma, and Angela Yao. On the utility of 3d hand poses for action recognition. *arXiv preprint arXiv:2403.09805*, 2024. 3

[42] Leqi Shen, Tianxiang Hao, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. Tempme: Video temporal token merging for efficient text-video retrieval. *arXiv preprint arXiv:2409.01156*, 2024. 7

[43] Lucas Stoffl, Andy Bonnetto, Stéphane d'Ascoli, and Alexander Mathis. Elucidating the hierarchical nature of behavior with masked autoencoders. In *European Conference on Computer Vision*, pages 106–125. Springer, 2025. 2, 3, 5

[44] Joseph A Tobias and Alex L Pigot. Integrating behaviour and ecology into global biodiversity conservation strategies. *Philosophical Transactions of the Royal Society B*, 374 (1781):20190012, 2019. 2

[45] 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. 6

[46] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022. 2

[47] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[48] Yann Vitasse, Sylvain Ursenbacher, Geoffrey Klein, Thierry Bohnenstengel, Yannick Chittaro, Anne Delestrade, Christian Monnerat, Martine Rebetez, Christian Rixen, Nicolas Strebel, et al. Phenological and elevational shifts of plants, animals and fungi under climate change in the european alps. *Biological Reviews*, 96(5):1816–1835, 2021. 3

[49] Lichen Wang, Zhengming Ding, Zhiqiang Tao, Yunyu Liu, and Yun Fu. Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6212–6221, 2019. 3

[50] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks

for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 3

[51] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 8

[52] Shaokai Ye, Anastasiia Filippova, Jessy Lauer, Steffen Schneider, Maxime Vidal, Tian Qiu, Alexander Mathis, and Mackenzie Weygandt Mathis. Superanimal pretrained pose estimation models for behavioral analysis. *Nature Communications*, 15(1):5165, 2024. 2, 8

[53] 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. 4, 5

[54] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 3