ChimpACT: A Longitudinal Dataset for Understanding Chimpanzee Behaviors

Xiaoxuan Ma [1,⋆] maxiaoxuan@pku.edu.cn

Stephan P. Kaufhold [2,⋆] spkaufho@ucsd.edu

Jiajun Su [1,⋆] sujiajun@pku.edu.cn

Wentao Zhu [1] wtzhu@pku.edu.cn

Jack Terwilliger [2] jterwilliger@ucsd.edu

Andres Meza [2] anmeza@ucsd.edu

Yixin Zhu [3,5,✉] yixin.zhu@pku.edu.cn

Federico Rossano [2,✉] frossano@ucsd.edu

Yizhou Wang [1,3,4] yizhou.wang@pku.edu.cn

⋆ X. Ma, S. Kaufhold, and J. Su contributed equally. ✉ Corresponding authors

[1] CFCS, School of Computer Science, Peking University, China

[2] Department of Cognitive Science, University of California, San Diego, USA

[3] Institute for Artificial Intelligence, Peking University, China

[4] Nat'l Eng. Research Center of Visual Technology, China

[5] PKU-WUHAN Institute for Artificial Intelligence, China

https://shirleymaxx.github.io/ChimpACT/

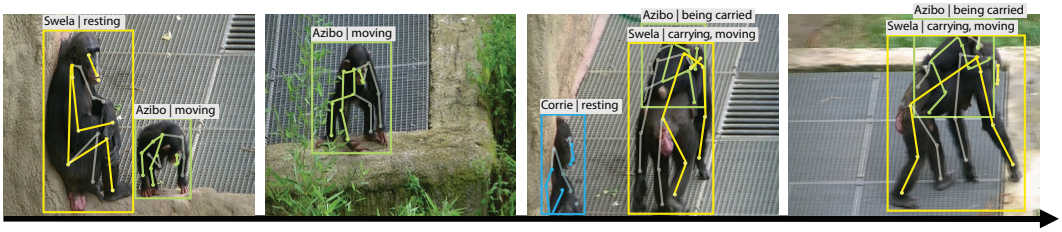


Figure 1: **Sample frames and annotations from a ChimpACT clip.** While we also annotate visibility for both the bounding box and the keypoint, these are omitted here for clarity.

Abstract

Understanding the behavior of non-human primates is crucial for improving animal welfare, modeling social behavior, and gaining insights into distinctively human and phylogenetically shared behaviors. However, the lack of datasets on non-human primate behavior hinders in-depth exploration of primate social interactions, posing challenges to research on our closest living relatives. To address these limitations, we present ChimpACT, a comprehensive dataset for quantifying the longitudinal behavior and social relations of chimpanzees within a social group. Spanning from 2015 to 2018, ChimpACT features videos of a group of over 20 chimpanzees residing at the Leipzig Zoo, Germany, with a particular focus on documenting the developmental trajectory of one young male, Azibo. ChimpACT is both comprehensive and challenging, consisting of 163 videos with a cumulative 160,500 frames, each richly annotated with detection, identification, pose estimation, and fine-grained spatiotemporal behavior labels. We benchmark representative methods of three tracks on ChimpACT: (i) tracking and identification, (ii) pose estimation, and (iii) spatiotemporal action detection of the chimpanzees. Our experiments reveal that ChimpACT offers ample opportunities for both devising new methods and adapting existing ones to solve fundamental computer vision tasks applied to chimpanzee groups, such as detection, pose estimation, and behavior analysis, ultimately deepening our comprehension of communication and sociality in non-human primates.

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

1 Introduction

Studying the behavior of non-human primates is essential for gaining evolutionary insights (Langergraber et al., 2012), conducting biomedical research (Schapiro et al., 2005), and improving animal welfare (Dawkins, 2003; Gonyou, 1994). Furthermore, given the close phylogenetic proximity between humans and non-human primates, it provides an ethically sound and effective avenue to probe the roots of human sociality (The Chimpanzee Sequencing and Analysis Consortium, 2005). Traditional field research typically requires researchers to enter wildlife conservation areas for extended durations, sometimes spanning multiple years. This involves habituating primate groups to human presence, capturing video footage, and laboriously manually coding these videos for subsequent statistical analysis (Hobaiter et al., 2017; Fröhlich et al., 2020; Surbeck et al., 2017; Luncz et al., 2018; Sirianni et al., 2015). While video coding is heralded as the gold standard for distilling rich, nuanced behavioral patterns (Wiltshire et al., 2023), its practical utility hinges on the efficiency of the coding process. This not only demands researchers with specialized expertise but is also prone to attentional biases.

Recent strides in computer vision offer promise for the automated analyses of non-human primate behaviors, especially those of chimpanzees. Nevertheless, the scarcity of high-quality longitudinal datasets remains a bottleneck. Assembling chimpanzee behavioral data is a formidable endeavor, necessitating substantial resources and expertise. This process entails continuous video recording and meticulous manual annotation, with a keen emphasis on annotation accuracy and consistency. While some datasets (Marks et al., 2022; Bala et al., 2020) confine subjects to indoor enclosures, resulting in atypical and constrained environments, others resort to sourcing and labeling chimpanzee images online (Labuguen et al., 2021; Desai et al., 2022; Ng et al., 2022; Yao et al., 2023). Unfortunately, these often overlook the intricate social dynamics inherent to chimpanzee groups, hindering a comprehensive study of their social behaviors and social relationships.

Addressing the existing limitations, we introduce `ChimpACT`, a comprehensive longitudinal dataset tailored for the in-depth study of chimpanzee social behavior in a semi-naturalistic setting, replete with annotations of instance bounding boxes, body poses, and spatial-temporal action labels. A comparison with other datasets is provided in Tab. 1. `ChimpACT` encompasses footage of a specific chimpanzee group residing at Leipzig Zoo, Germany, with a particular focus on a juvenile male named Azibo (refer to Fig. 1). The data, gathered between 2015 and 2018, employs *focal sampling* (Altmann, 1974). Born in April 2015, Azibo[1] has been living in the group since birth, providing a unique perspective on the development of an individual within a chimpanzee group characterized by

Table 1: **Comparison of `ChimpACT` with existing primate behavioral datasets.** Square-bracketed numbers denote label counts for the chimpanzee category. ⊘ denotes undocumented. For the "Species" row, G represents general, P for primates, M for macaque, and C for chimpanzee. In the "Source" row, I stands for Internet, Z for zoo, C for cage, W for wild, and CP for captive.

Dataset	Species	Track 1 detection, tracking, ReID				Track 2 pose estimation				Track 3 action recognition		Source
		ID #	frame #	box #	track	frame #	pose #	track	dim.	class #	label #	
AP-10K (Yu et al., 2021)	G	✗	✗	✗	✗	10,015	13,028 [<500]	✗	2D	✗	✗	I
AnimalKingdom (Ng et al., 2022)	G	✗	✗	✗	✗	33,099	99,297 [576]	✗	2D	140	30,100 [⊘]	I
OpenApePose (Desai et al., 2022)	P	✗	✗	✗	✗	71,868	71,868 [18,010]	✗	2D	✗	✗	I
OpenMonkeyChallenge (Yao et al., 2023)	P	✗	✗	✗	✗	111,529	111,529 [<10,000]	✗	2D	✗	✗	I & Z
OpenMonkeyStudio (Bala et al., 2020)	M	✗	✗	✗	✗	194,518	33,192 [0]	✓	3D	✗	✗	C ($6.7m^2$)
MacaquePose (Labuguen et al., 2021)	M	✗	✗	✗	✗	13,083	16,393 [0]	✗	2D	✗	✗	I & Z
SIPEC (Marks et al., 2022)	M	4	191	2,200 [0]	✓	✗	✗	✗	✗	4	⊘	C ($15m^2$)
CCR (Bain et al., 2019)	C	13	936,914	1,937,585	✓	✗	✗	✗	✗	✗	✗	W
`ChimpACT` (Ours)	C	23	160,500	56,324	✓	16,028	56,324	✓	2D	23	64,289	CP ($4400m^2$)

[1]Details about Azibo can be found at https://tinyurl.com/azibo-chimp/.

well-defined kin relationships. (also depicted in Fig. 2a). The footage covers the daily lives of over 20 chimpanzees in a group, aggregating to 163 video recordings, approximately 160,500 frames, and spanning around 2 hours.

Our annotations on ChimpACT are extensive, marking each individual's detection, tracking, identification, pose estimation, and spatiotemporal action detection. Sample frames with their corresponding annotations are illustrated in Fig. 1. Each chimpanzee's identity is confirmed by a seasoned behavioral researcher familiar with the Leipzig chimpanzees, ensuring data precision and trustworthiness. Crucially, we employ an ethogram (detailed in Fig. 2b) devised by the same expert for fine-grained action labels. To our knowledge, ChimpACT is the first to furnish ethogram annotations for the machine learning and computer vision community. This bespoke ethogram delineates behaviors into four categories: locomotion, object interaction, social interaction, and others, with each encompassing several detailed actions we diligently annotate.

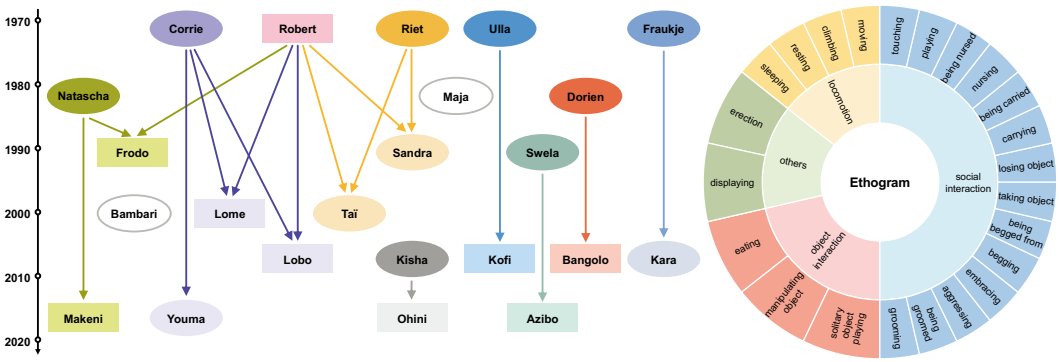
While advancements in computer vision have notably addressed human-centric tasks, such as human pose estimation (Sun et al., 2019; Xiao et al., 2018), the dearth of chimpanzee datasets has curtailed progress on chimpanzee-specific challenges. Despite their genetic closeness to humans (The Chimpanzee Sequencing and Analysis Consortium, 2005), deciphering chimpanzee behaviors is intricate due to their unique morphology, appearance, and keypoint articulation. Highlighting the importance of crafting sophisticated chimpanzee perception models, we evaluate prominent human perception methods on three tracks: (i) detection, tracking, and identification (ReID), (ii) pose estimation, and (iii) spatiotemporal action detection. Our findings underscore ChimpACT's potential as a platform for the community to pioneer advanced techniques for better perception of the chimpanzees and ultimately contribute to a deeper understanding of non-human primates.

2 Related work

Computer vision for animals A myriad of datasets and benchmarks have emerged, harnessing computer vision techniques to advance animal research. For instance, 3D-ZeF20 (Pedersen et al., 2020) introduces 3D tracking of zebrafish to the MOT benchmarks. AnimalTrack (Zhang et al., 2023) emphasizes multi-animal tracking across a spectrum of species. AP-10K (Yu et al., 2021) and APT-36K (Yang et al., 2022) venture into animal pose estimation for diverse species. AnimalKingdom (Ng et al., 2022) extends its focus to fine-grained multi-label action recognition. Moreover, several studies have delved into multi-agent behavior understanding from a social interaction perspective (Sun et al., 2021, 2023). Distinctively, ChimpACT stands out as a holistic benchmark, encompassing three varied downstream tasks and boasting rich annotations of social interactions.

Human video datasets In contrast to animal-centric video datasets, a more substantial collection is tailored to human subjects, addressing diverse human-centric video understanding tasks. For instance, the MOT Challenge (Milan et al., 2016) is curated for multi-person tracking. Other benchmarks like COCO (Lin et al., 2014) and MPII (Andriluka et al., 2014) cater to human pose estimation. Meanwhile, datasets such as Kinetics (Kay et al., 2017), ActivityNet (Fabian Caba Heilbron and Niebles, 2015), and AVA (Gu et al., 2018) are dedicated to human action recognition. With ChimpACT, we encompass analogous tasks but introduce challenges specific to chimpanzee behavior.

Datasets on primate behavioral understanding Most existing primate datasets are tailored towards individual primate detection and pose estimation. These either stem from confined indoor settings (Bala et al., 2020; Marks et al., 2022) or are amassed and labeled from online sources (Labuguen et al., 2021; Desai et al., 2022; Ng et al., 2022; Yao et al., 2023). The former can induce atypical behavioral patterns, while the latter often omits longitudinal interactions, rendering them suboptimal for analyzing chimpanzee social dynamics. A notable exception is the CCR dataset (Bain et al., 2019), chronicling 13 chimpanzees in the Bossou forest over two years. Yet, it primarily focuses on individual detection and recognition, lacking behavioral annotations, which limits its efficacy for probing the social nuances of wild primates. Tab. 1 offers a comprehensive comparison. The narrow focus of most primate datasets on singular tasks restricts their breadth and adaptability to diverse research inquiries. Contrarily, ChimpACT presents a multifaceted approach, encompassing identities, kinship, detection labels, pose annotations, ethograms, and fine-grained action labels. This richness positions it as an indispensable tool for devising advanced chimpanzee behavior analysis methods and enriching the overarching comprehension of primate behavior.



(a) (b)

Figure 2: **(a) Kinship of the observed chimpanzee group.** Rectangles and ellipses represent males and females, respectively, with arrows flowing from the parents to the child. Their vertical position relative to the time axis indicates the year of birth. **(b) Ethogram with annotated behaviors.**

Methods for primate behavioral analysis Deciphering primate behavior is instrumental in understanding their social dynamics and cognitive abilities. Behavioral analysis often encompasses subtasks like individual detection, tracking, and identification (Bain et al., 2019; Marks et al., 2022), pose estimation (Labuguen et al., 2021; Desai et al., 2022; Mathis et al., 2018; Wiltshire et al., 2023), and behavior recognition (Ng et al., 2022; Bain et al., 2021). While each task has specialized techniques, many are rooted in human behavioral research. Numerous algorithms exist for human tracking (Bewley et al., 2016; Pang et al., 2021), pose estimation (Sun et al., 2019; Xiao et al., 2018), and behavior recognition (Feichtenhofer et al., 2019). However, due to the dearth of primate datasets, primate behavioral analysis often repurposes algorithms designed for humans, including:

- **Detection, tracking, and ReID** identify individual primates in videos, often leveraging established object or human detection algorithms like Mask-RCNN (He et al., 2017). For instance, SIPEC (Marks et al., 2022) employs Mask-RCNN with a ResNet backbone (He et al., 2016) to track and segment macaque. Bain et al. (2019) utilize CNNs to crop and identify individual chimpanzees.
- **Pose estimation** discerns primate poses, frequently adapting human pose estimation methods like SimpleBaseline (Xiao et al., 2018). DeepLabCut (Mathis et al., 2018; Lauer et al., 2022), for instance, employs ResNet-50 with ImageNet pre-trained weights for 2D animal pose estimation. SIPEC (Marks et al., 2022) modifies SimpleBaseline for 2D macaque poses.
- **Behavior recognition** identifies primate actions and interactions. Contemporary methods (Bain et al., 2021; Bohnslav et al., 2021) often derive from human action recognition algorithms like SlowFast (Schindler and Steinhage, 2021). Notably, Bain et al. (2021) integrates audio cues for classifying two simple non-interactive behaviors: nut cracking and buttress drumming. In contrast, `ChimpACT` encompasses over 20 daily behaviors under an ethogram hierarchy, capturing both solitary actions and intricate social interactions.

In essence, primate behavioral analysis is a multifaceted endeavor, intertwining computer vision, machine learning, and primatology. The advent of `ChimpACT` marks a significant stride towards unraveling the intricate social tapestry of our primate kin.

3 ChimpACT

3.1 Dataset description

`ChimpACT` comprises about 2-hour video footage of chimpanzees recorded at the Leipzig Zoo in Germany between 2015 and 2018. The videos focus on one male chimpanzee, Azibo, who was born in April 2015 to Swela and has lived with the A-chimpanzee group[2] at the Leipzig Zoo ever since. The longitudinal observation of Azibo offers a rare lens into his behavioral evolution, social dynamics,

[2]The A-chimpanzee group is among the most extensively studied zoo-residing chimpanzee cohorts. Its members have been subjects of both behavioral and cognitive studies, spanning observational and experimental designs, conducted by researchers affiliated with the MPI for Evolutionary Anthropology (Baker, 2022; McEwen et al., 2022).

and intra-group relationships. With over 20 individuals in the group, ChimpACT serves as a treasure trove of insights into chimpanzee behavior and social intricacies. Key attributes of ChimpACT are delineated below.

Longitudinal data Spanning four years, ChimpACT chronicles the life of a stable zoo-residing chimpanzee group, offering a rare glimpse into the nuances of chimpanzee social behavior development. Tracking the growth and interactions of a young chimpanzee within this group sheds light on chimpanzee socialization, the evolution of social skills (Matsuzawa, 2013), the formation of social bonds and integration into the dominance hierarchy (Matsuzawa et al., 2006), and the acquisition of group-specific cultural behaviors (Van Leeuwen, 2021; Musgrave et al., 2021).

Semi-naturalistic and social environment The videos in ChimpACT capture chimpanzees in their semi-naturalistic habitats at Leipzig Zoo, split between indoor (96 videos) and outdoor (67 videos) enclosures. The indoor space, spanning roughly 400 m^2, features a plethora of environmental enrichments, ranging from wooden climbing structures and hammocks to vegetation and foraging boxes. When weather permits, the chimpanzees have access to a 4000 m^2 outdoor area, replete with vegetation, surrounded by an artificial river, and complemented by enrichments similar to the indoor space. This blend of environments ensures the dataset's relevance for both naturalistic and artificial environments. The multifaceted physical and social surroundings of the chimpanzees further imbue the dataset with intricate behaviors and social dynamics.

Ethogram with solitary and social behaviors ChimpACT captures the daily life of group-living chimpanzees, offering invaluable insights into the evolution and sustenance of their social behaviors and relationships (Nishida et al., 2010). By focusing on a juvenile chimpanzee, ChimpACT illuminates facets of social learning, communication, bonding, and more, all pivotal in the social and ecological life of chimpanzees (Bard et al., 2014). To systematically represent these behaviors, we composed an ethogram—a detailed catalog of behavioral categories, depicted in Fig. 2b (further details in Appx. A). This ethogram organizes behaviors into four primary categories, like locomotion and social interaction, each further subdivided into several fine-grained actions, meticulously annotated and validated with expert oversight. By delving into these behaviors, ChimpACT elucidates not only the social dynamics shaping social relationships but also the cognitive and ecological influences on juvenile chimpanzee behaviors.

3.2 Dataset collection

The focal video data were collected with the Chimpanzee-A group housed at Leipzig Zoo, Germany, using focal sampling (Altmann, 1974). Videographers were instructed to focus on Azibo and his mother, Swela, but also on capturing the environmental context and his interactions with other chimpanzees. Videos from ChimpACT were sampled from a larger set of around 405 hours of longitudinal focal video recordings of the dyad between 2015 and 2018. These videos were recorded by several research assistants during the daytime (7am–4pm) using tripod-mounted RGB cameras. Two JVC Everio camera models were utilized across the years, filming with a framerate of 25 (Codec H.264) and with resolutions of 720×578 and 1280×720, respectively. The mother-infant dyad was filmed for about five hours each week during the observation period. The footage contains both optical zoom and camera movements.

3.3 Dataset tasks and annotations

ChimpACT supports three tracks: (i) chimpanzee detection, tracking, and ReID, (ii) chimpanzee pose estimation, and (iii) spatiotemporal action detection. We provide fine-grained annotations for each track. From the extensive footage, we curated 163 video clips, each approximately 1000 frames in length. Fifteen adept annotators were then tasked with annotating bounding boxes, body keypoints, and fine-grained behavioral classes for each chimpanzee at intervals of every 10 frames. To ensure accuracy and consistency, a behavioral researcher familiar with the chimpanzee group meticulously reviewed and refined the identity and behavioral class annotations. For a deeper dive into the annotation process and its quality, please refer to Appx. A and our dedicated website.

Detection, tracking, and ReID This task encompasses the detection and tracking of individual chimpanzees across video sequences, subsequently coupled with their re-identification. ChimpACT features over 23 distinct chimpanzee individuals, each identified by a primate expert familiar with the Leipzig A-group chimpanzees. Initially, annotators were instructed to delineate the bounding

box of each chimpanzee, ensuring consistent box IDs for the same individual throughout a video clip. Subsequently, the expert matched these box IDs with the corresponding true names of the chimpanzees, resulting in the identification of 23 unique individuals. Additionally, every annotated bounding box is attached with a visibility attribute, indicating if the chimpanzee is fully visible, truncated, or occluded in a given frame. Such visibility annotations can support the reasoning of the chimpanzee behavior, potentially bolstering tracking robustness. Fig. 3a illustrates the occurrence frequency (on a *log* scale) of each individual, revealing a long-tail distribution. This pattern aligns with the focal sampling strategy, where Azibo is the primary subject. Notably, Swela, Azibo's mother, also exhibits a high occurrence frequency, resonating with prior studies (Boesch, 1996).

Pose estimation Pose estimation aims to predict the locations of the chimpanzee joints that have semantic meaning, such as the knee and shoulder, from an input image. There are four keypoints on the chimpanzee's face (*i.e.*, two for the eyes, and one each for the upper and lower lips), for a total of 16 chimpanzee keypoints (refer to Sec. 3.3 and Fig. 4). Annotators are tasked with marking the 2D joint coordinates and the visibility status of each joint. We adopt the visibility protocol from the COCO 2D human keypoint annotations (Lin et al., 2014), where a value of 0 indicates a joint outside the image frame, 1 signifies an obscured joint within the image, and 2 designates a clearly visible joint. Such an annotation protocol affords reason about chimpanzee's orientation and action based on facial joint visibility. For instance, the chimpanzee might be eating something if the two lips are apart. Sample frames showcasing pose annotations are depicted in Fig. 1. Notably, `ChimpACT` holds the potential for future expansion to encompass pose tracking tasks, analogous to the PoseTrack (Andriluka et al., 2018) for humans.

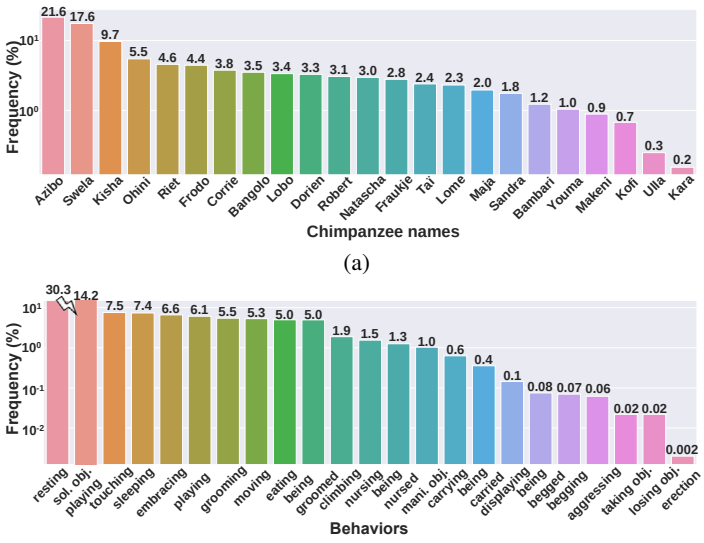


Figure 3: **(a) Distribution (in log scale) of annotations for each individual. (b) Distribution (in log scale) of annotations for each behavior.** Vector graphics; zoom for details.

Table 2: **Keypoint definitions for chimpanzee.**

No.	Definition	No.	Definition
0	Root of hip	8	Right eye
1	Right knee	9	Left eye
2	Right ankle	10	Right shoulder
3	Left knee	11	Right elbow
4	Left ankle	12	Right wrist
5	Neck	13	Left shoulder
6	Upper lip	14	Left elbow
7	Lower lip	15	Left wrist

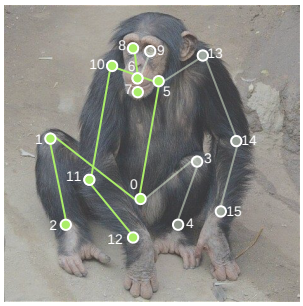
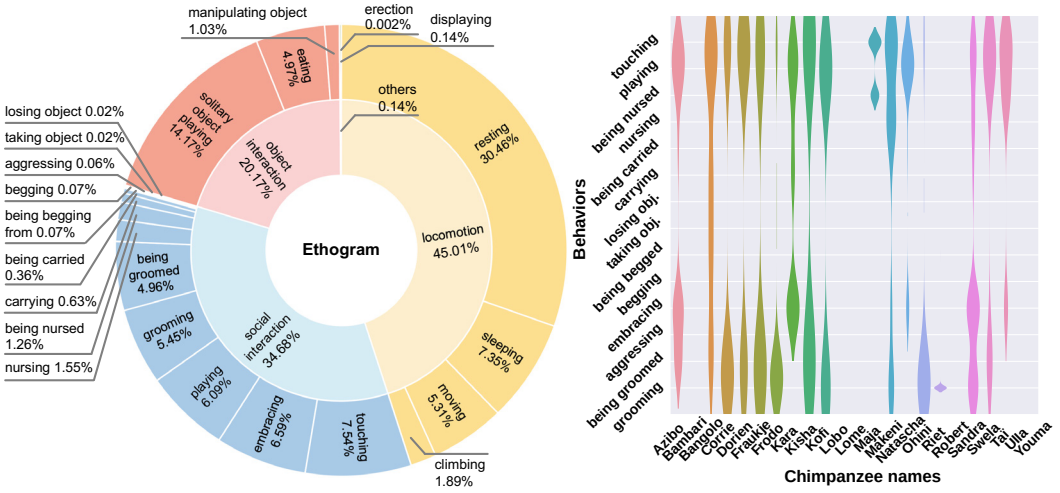


Figure 4: **Keypoint definitions for chimpanzee.**

Spatiotemporal action detection Spatiotemporal action detection seeks to attribute one or multiple behavioral labels to each bounding box containing a chimpanzee, leveraging the spatiotemporal context within a video clip. Our ethogram, detailed in Fig. 2b, delineates 23 nuanced subcategories of behaviors and guides the fine-grained annotations of chimpanzee behavior, such as "climbing" within the "locomotion" category. Notably, within the realm of social interactions, we meticulously differentiate between the action performer and receiver. For instance, the grooming behavior is bifurcated into "grooming" and "being groomed." Every chimpanzee in a frame has its subcategory behavior annotated. It is not uncommon for an individual to simultaneously exhibit multiple behaviors,

exemplified by Swela's "carrying" and "moving" actions in Fig. 1. The distribution of these behavioral annotations, visualized in Fig. 3b on a *log* scale, reveals a long-tail distribution, mirroring the authentic behavioral tendencies of chimpanzees in their natural habitats.



(a) (b)

Figure 5: **(a) Distribution of the annotated behavior categories. (b) Distribution showcasing individuals alongside their respective social behaviors.** Vector graphics; zoom for details.

Fig. 5a showcases the distribution of the annotated behaviors, with social interactions constituting approximately 35% of the total annotations. Furthermore, Fig. 5b delineates the distribution of labeled social behaviors across distinct individuals, highlighting grooming, playing, and touching as predominant activities within the social dynamics of the group-living chimpanzees.

In essence, ChimpACT emerges as an invaluable resource for researchers spanning the domains of primatology, comparative psychology, computer vision, and machine learning. It furnishes a comprehensive and varied array of annotations, paving the way for in-depth analysis of multifaceted chimpanzee behaviors and catalyzing the development of advanced machine learning algorithms. The inherent long-tail distribution not only presents a formidable challenge for chimpanzee identification and behavior recognition but also beckons explorations into few-shot learning in future endeavors.

4 Experiments

To rigorously assess ChimpACT, we benchmark a suite of representative methods across the aforementioned three tracks: (i) detection, tracking, and ReID, (ii) pose estimation, and (iii) spatiotemporal action detection. Our computational framework leverages four NVIDIA GeForce RTX 3090 GPUs (24GB) for both training and evaluation across all tracks. In the subsequent sections, we delve into the implementation details, baseline methods, and evaluation metrics for each track.

4.1 Detection, tracking, and ReID

Setting We evaluate several prominent Multiple Object Tracking (MOT) algorithms on ChimpACT, including both classical methods such as SORT (Bewley et al., 2016), DeepSORT (Wojke et al., 2017), and Tracktor (Bergmann et al., 2019), as well as the state-of-the-art methods such as ByteTrack (Zhang et al., 2022), and OC-SORT (Cao et al., 2023). All implementations are based on the MMTracking (Contributors, 2020a) codebase. For those methods supporting flexible detection backbones, we trial two typical detectors, including the two-stage detector Faster R-CNN (Ren et al., 2015) and the one-stage detector YOLOX (Ge et al., 2021). Each method undergoes training for 10 epochs, adhering to the official configurations, which encompass optimizer settings, batch size, data augmentation techniques, and pre-trained models. Given that the three classical methods lack inherent ReID modules, we supplement with a dedicated ReID network built on ResNet-50 (He et al., 2016). The training curves of select methods (refer to Fig. A2a) affirm convergence within the training epochs.

We split the video clips in ChimpACT into 80% train, 10% validation, and 10% test. Both the train set and test set cover all the individuals. Models are trained on the training set, with performance metrics

Table 3: **Results of the detection, tracking, and ReID track on the `ChimpACT` test set.** The row highlighted in light blue is the performance reference on the human tracking dataset MOT-17 (Milan et al., 2016). – denotes not applicable. ⊘ denotes unreported.

Method	Detector	ReID	HOTA ↑	MOTA ↑	MOTP ↑	IDF1 ↑	mAP ↑	nFP ↓	nFN ↓	nIDs ↓
SORT (Bewley et al., 2016)	Faster R-CNN	ResNet-50	$42.6_{\pm 1.0}$	$47.4_{\pm 0.6}$	$22.9_{\pm 1.3}$	$42.7_{\pm 1.2}$	$70.7_{\pm 1.6}$	$19.1_{\pm 0.3}$	$31.4_{\pm 0.5}$	$2.1_{\pm 0.0}$
	YOLOX		$39.8_{\pm 0.8}$	$43.2_{\pm 1.3}$	$20.3_{\pm 0.5}$	$37.7_{\pm 1.7}$	$71.4_{\pm 1.6}$	$16.1_{\pm 3.1}$	$37.8_{\pm 1.7}$	$2.8_{\pm 0.5}$
DeepSORT (Wojke et al., 2017)	Faster R-CNN	ResNet-50	$47.6_{\pm 0.4}$	$46.7_{\pm 0.5}$	$23.0_{\pm 1.2}$	$52.8_{\pm 1.5}$	$70.7_{\pm 1.6}$	$19.0_{\pm 0.3}$	$31.4_{\pm 0.5}$	$2.9_{\pm 0.1}$
	YOLOX		$40.2_{\pm 1.0}$	$43.2_{\pm 1.2}$	$20.3_{\pm 0.5}$	$38.4_{\pm 1.9}$	$71.4_{\pm 1.6}$	$16.1_{\pm 3.1}$	$37.8_{\pm 1.7}$	$2.9_{\pm 0.6}$
Tracktor (Bergmann et al., 2019)	Faster R-CNN	ResNet-50	$49.5_{\pm 0.7}$	$50.5_{\pm 1.1}$	$22.6_{\pm 1.1}$	$55.6_{\pm 1.2}$	$70.7_{\pm 1.6}$	$13.8_{\pm 0.5}$	$35.2_{\pm 0.7}$	$0.5_{\pm 0.0}$
QDTrack (Pang et al., 2021)	Faster R-CNN	–	$50.3_{\pm 3.2}$	$54.2_{\pm 4.6}$	$22.2_{\pm 1.4}$	$55.8_{\pm 3.6}$	$77.8_{\pm 2.0}$	$19.7_{\pm 3.6}$	$24.6_{\pm 0.8}$	$1.4_{\pm 0.2}$
ByteTrack (Zhang et al., 2022)	Faster R-CNN	–	$43.7_{\pm 0.3}$	$36.9_{\pm 2.2}$	$24.6_{\pm 0.3}$	$48.8_{\pm 1.3}$	$68.2_{\pm 1.1}$	$27.7_{\pm 1.1}$	$34.2_{\pm 1.0}$	$1.2_{\pm 0.2}$
	YOLOX	–	$49.2_{\pm 0.8}$	$43.9_{\pm 1.3}$	$20.3_{\pm 1.0}$	$55.2_{\pm 1.1}$	$70.3_{\pm 1.0}$	$18.0_{\pm 7.4}$	$37.4_{\pm 6.1}$	$0.7_{\pm 0.0}$
OC-SORT (Cao et al., 2023)	Faster R-CNN	–	$43.4_{\pm 1.0}$	$38.2_{\pm 1.9}$	$24.3_{\pm 0.2}$	$48.7_{\pm 2.2}$	$68.7_{\pm 0.8}$	$25.0_{\pm 1.6}$	$35.6_{\pm 1.5}$	$1.2_{\pm 0.1}$
	YOLOX	–	$47.9_{\pm 0.4}$	$42.1_{\pm 2.6}$	$20.5_{\pm 0.8}$	$53.3_{\pm 0.8}$	$70.5_{\pm 0.8}$	$20.3_{\pm 1.3}$	$36.6_{\pm 2.1}$	$1.1_{\pm 0.3}$
OC-SORT (Cao et al., 2023)	YOLOX	–	63.2	78.0	⊘	77.5	⊘	2.7	19.0	0.3

reported on the test set. We employ widely-accepted evaluation metrics, drawing from convention in human/object detection, tracking, and ReID (Bewley et al., 2016; Pang et al., 2021; Zhang et al., 2022). Specifically, we utilize (i) mean Average Precision (mAP) Lin et al. (2014) to gauge the detection accuracy, and (ii) the CLEAR metrics (Bernardin and Stiefelhagen, 2008) (MOTA, MOTP, FP, FN, IDs), IDF1 (Ristani et al., 2016), and HOTA (Luiten et al., 2021) to evaluate various facets of the tracking performance. It is worth noting that for FP, FN, and IDs, we report normalized values and denote these metrics as nFP, nFN, and nIDs, respectively.

Results Tab. 3 summarizes these tracking algorithms' performances on the `ChimpACT` test set. We conducted three runs for each method and reported the average and variance of these metrics. Notably, the variance across multiple runs is minimal, underscoring the robust reproducibility of our benchmarking. A holistic view of the results reveals that QDTrack (Pang et al., 2021) emerges as the top performer. However, it does suffer from a higher count of identity switches compared to other methods. In terms of detection performance, the YOLOX algorithm (Ge et al., 2021) stands toe-to-toe with Faster R-CNN (Ren et al., 2015). A discernible trend is evident among contemporary tracking methods, which seem to excel in identity association capabilities over their classical counterparts. This is corroborated by marked improvements in tracking metrics like IDF1 and IDs. Such a trend intimates that the latest tracking methods might be adept at maintaining consistent object identities, a pivotal aspect when tracking and analyzing individual trajectories within chimpanzee cohorts.

While the results garnered by the array of tracking algorithms are commendable, they still lag behind the benchmarks set on human-centric datasets (Zhang et al., 2022; Pang et al., 2021; Cao et al., 2023). This disparity can be attributed to challenges like the low contrast and low color variation of the body fur of chimpanzees, compounded by intricate self-occlusions. Nonetheless, this very observation accentuates the significance of `ChimpACT`. It not only offers a challenging arena for tracking algorithms but also stands as an ideal platform for pioneering and refining tracking methods tailored for chimpanzees and other non-human primates.

4.2 Pose estimation

Setting We benchmark several state-of-the-art human pose estimation methods on `ChimpACT`, including CPM (Wei et al., 2016), SimpleBaseline (Xiao et al., 2018), HRNet (Sun et al., 2019), DarkPose (Zhang et al., 2020). Broadly, human pose estimation methods can be bifurcated into two primary paradigms: heatmap-based and regression-based. We harness the MMPose (Contributors, 2020b) framework for implementing these methods. Please refer to Appx. C for more implementation details. All the models undergo training for 210 epochs, maintaining the official configurations for optimizers, batch sizes, and learning rates. To gauge any potential model overfitting, we present the validation curve on the AP metric in Fig. A2b, reassuringly suggesting an absence of overfitting.

The train/test partitioning mirrors that of the first track. We use mAP with various thresholds, adhering to the conventions of human pose estimation (Lin et al., 2014). Additionally, we report the Percentage of Correctly estimated Keypoints (PCK) metric (Andriluka et al., 2014; Ng et al., 2022). PCK@α quantifies the fraction of accurately predicted keypoints within a distance threshold defined as

Table 4: **Results of the pose estimation track on `ChimpACT` test set.** The row highlighted in light blue is the performance reference on the human pose estimation dataset COCO (Lin et al., 2014). ⊘ denotes unreported.

	Method	Backbone	PCK@0.05	PCK@0.1	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Regression	SimpleBaseline (Xiao et al., 2018)	ResNet-50	$25.3_{\pm 0.5}$	$46.2_{\pm 0.5}$	$8.6_{\pm 0.4}$	$27.4_{\pm 1.3}$	$3.9_{\pm 0.4}$	$0.3_{\pm 0.1}$	$12.5_{\pm 0.5}$	$17.3_{\pm 0.7}$
		ResNet-101	$26.2_{\pm 1.0}$	$46.4_{\pm 1.1}$	$8.7_{\pm 0.4}$	$27.5_{\pm 0.6}$	$4.2_{\pm 0.5}$	$0.3_{\pm 0.0}$	$12.9_{\pm 0.2}$	$17.7_{\pm 0.4}$
		ResNet-152	$26.3_{\pm 0.4}$	$47.3_{\pm 0.8}$	$9.3_{\pm 0.1}$	$29.2_{\pm 1.1}$	$4.7_{\pm 0.3}$	$0.5_{\pm 0.0}$	$13.4_{\pm 0.2}$	$18.6_{\pm 0.0}$
	RLE (Li et al., 2021)	MobileNetV2	$27.5_{\pm 1.4}$	$48.1_{\pm 1.7}$	$16.7_{\pm 0.8}$	$43.1_{\pm 2.7}$	$11.1_{\pm 0.8}$	$2.0_{\pm 0.7}$	$17.7_{\pm 0.8}$	$19.5_{\pm 0.9}$
		ResNet-50	$28.2_{\pm 1.7}$	$47.1_{\pm 3.1}$	$16.3_{\pm 2.5}$	$41.2_{\pm 6.9}$	$11.4_{\pm 1.4}$	$1.3_{\pm 0.8}$	$17.4_{\pm 2.8}$	$20.0_{\pm 1.6}$
		ResNet-101	$28.2_{\pm 3.5}$	$46.5_{\pm 4.3}$	$16.2_{\pm 2.6}$	$41.1_{\pm 5.7}$	$10.8_{\pm 2.4}$	$2.1_{\pm 0.1}$	$17.3_{\pm 2.8}$	$20.1_{\pm 2.1}$
		ResNet-152	$30.0_{\pm 1.3}$	$48.4_{\pm 2.2}$	$18.1_{\pm 2.8}$	$43.0_{\pm 7.9}$	$13.5_{\pm 0.6}$	$1.4_{\pm 0.3}$	$19.2_{\pm 3.2}$	$22.3_{\pm 1.1}$
Heatmap-based	CPM (Wei et al., 2016)	CPM	$40.7_{\pm 0.2}$	$60.4_{\pm 0.0}$	$21.6_{\pm 0.1}$	$51.0_{\pm 0.4}$	$17.1_{\pm 0.1}$	$9.5_{\pm 0.6}$	$22.4_{\pm 0.1}$	$25.4_{\pm 0.1}$
	Hourglass (Newell et al., 2016)	Hourglass-4	$44.6_{\pm 0.5}$	$60.8_{\pm 0.1}$	$20.6_{\pm 0.3}$	$48.9_{\pm 0.1}$	$16.0_{\pm 0.4}$	$4.6_{\pm 0.1}$	$23.7_{\pm 0.6}$	$28.2_{\pm 0.2}$
	MobileNetV2 (Sandler et al., 2018)	MobileNetV2	$39.8_{\pm 0.4}$	$59.4_{\pm 0.4}$	$19.4_{\pm 0.1}$	$48.5_{\pm 0.6}$	$14.3_{\pm 0.8}$	$2.3_{\pm 0.1}$	$20.6_{\pm 0.1}$	$23.2_{\pm 0.1}$
	SimpleBaseline (Xiao et al., 2018)	ResNet-50	$43.3_{\pm 0.2}$	$61.7_{\pm 1.2}$	$22.1_{\pm 0.2}$	$51.5_{\pm 0.4}$	$17.7_{\pm 0.2}$	$3.7_{\pm 0.4}$	$23.4_{\pm 0.2}$	$26.3_{\pm 0.1}$
		ResNet-101	$42.8_{\pm 0.3}$	$60.7_{\pm 0.2}$	$21.7_{\pm 0.1}$	$52.5_{\pm 0.4}$	$16.7_{\pm 0.0}$	$4.3_{\pm 0.2}$	$23.0_{\pm 0.1}$	$26.2_{\pm 0.2}$
		ResNet-152	$43.9_{\pm 0.4}$	$61.6_{\pm 0.1}$	$22.7_{\pm 0.4}$	$53.4_{\pm 0.6}$	$18.3_{\pm 0.4}$	$5.3_{\pm 0.5}$	$23.9_{\pm 0.4}$	$27.1_{\pm 0.1}$
	HRNet (Sun et al., 2019)	HRNet-W32	$48.6_{\pm 0.9}$	$65.6_{\pm 0.6}$	$25.9_{\pm 0.4}$	$58.2_{\pm 0.8}$	$22.1_{\pm 0.4}$	$6.1_{\pm 0.4}$	$27.0_{\pm 0.6}$	$30.3_{\pm 0.5}$
		HRNet-W48	$47.3_{\pm 0.2}$	$64.5_{\pm 0.2}$	$25.1_{\pm 0.1}$	$57.2_{\pm 0.6}$	$21.0_{\pm 0.1}$	$6.9_{\pm 0.9}$	$26.2_{\pm 0.3}$	$29.6_{\pm 0.1}$
	DarkPose (Zhang et al., 2020)	ResNet-50	$43.7_{\pm 0.0}$	$62.1_{\pm 0.6}$	$22.8_{\pm 0.1}$	$53.8_{\pm 0.8}$	$18.8_{\pm 0.6}$	$3.4_{\pm 0.2}$	$24.1_{\pm 0.0}$	$27.1_{\pm 0.1}$
		ResNet-101	$43.1_{\pm 0.9}$	$61.2_{\pm 1.4}$	$22.1_{\pm 0.3}$	$52.6_{\pm 0.6}$	$17.6_{\pm 0.7}$	$4.0_{\pm 0.4}$	$23.4_{\pm 0.3}$	$26.5_{\pm 0.3}$
		ResNet-152	$43.5_{\pm 0.3}$	$61.2_{\pm 0.2}$	$22.4_{\pm 0.1}$	$53.2_{\pm 0.1}$	$17.4_{\pm 0.3}$	$4.6_{\pm 0.0}$	$23.7_{\pm 0.1}$	$26.7_{\pm 0.1}$
		HRNet-W32	$48.7_{\pm 0.5}$	$65.6_{\pm 0.9}$	$25.7_{\pm 0.4}$	$58.4_{\pm 0.8}$	$21.3_{\pm 0.8}$	$5.6_{\pm 0.4}$	$26.9_{\pm 0.2}$	$30.1_{\pm 0.2}$
		HRNet-W48	$47.6_{\pm 0.7}$	$64.5_{\pm 1.0}$	$25.8_{\pm 0.4}$	$58.0_{\pm 1.7}$	$21.5_{\pm 0.3}$	$6.6_{\pm 0.5}$	$27.0_{\pm 0.4}$	$30.2_{\pm 0.5}$
	HRFormer (Yuan et al., 2021)	HRFormer-S	$45.1_{\pm 0.4}$	$61.4_{\pm 0.4}$	$23.0_{\pm 0.0}$	$53.1_{\pm 0.4}$	$19.7_{\pm 0.2}$	$5.5_{\pm 1.6}$	$24.1_{\pm 0.2}$	$27.1_{\pm 0.1}$
		HRFormer-B	$46.4_{\pm 0.3}$	$63.0_{\pm 0.1}$	$24.1_{\pm 0.6}$	$55.3_{\pm 1.0}$	$20.1_{\pm 0.1}$	$5.2_{\pm 0.4}$	$25.4_{\pm 0.5}$	$28.2_{\pm 0.4}$
	HRNet (Sun et al., 2019)	HRNet-W32	⊘	⊘	74.4	90.5	81.9	70.8	81.0	79.8

$\alpha \times max(height, width)$, derived from the bounding box of the chimpanzee. This metric is widely recognized for its accuracy in body joint localization in both human and animal pose estimation.

Results Tab. 4 consolidates these pose estimators' performances on the `ChimpACT` test set. Notably, the heatmap-based DarkPose (Zhang et al., 2020) with an HRNet (Sun et al., 2019) backbone emerges as the top-performing model. This trend aligns with observations in human pose estimation, where heatmap-centric methods (Wei et al., 2016; Xiao et al., 2018; Newell et al., 2016; Sun et al., 2019) predominantly lead the pack, attributed to their robustness against pose and appearance variations. However, the heatmap representation may be less accurate in scenarios where multiple joints are occluded or closely spaced, and it demands heftier computational and memory resources. Conversely, the newer regression-based methods (Li et al., 2021) are computationally leaner but tend to be more susceptible to overfitting and generally lag in performance.

These results underscore that the task of chimpanzee pose estimation is distinct and nuanced, and cannot be seamlessly addressed by merely repurposing human-centric pose estimation methods. We believe there are two primary reasons for this: (i) chimpanzees exhibit unique joint flexibility and a broader range of motion, and (ii) the visual texture and appearance of chimpanzee fur diverge significantly from human skin. These insights emphasize the need for chimpanzee specific pose estimation strategies.

4.3 Spatiotemporal action detection

Setting We benchmark four representative human action detection baselines on `ChimpACT` using the MMAction2 (Contributors, 2020c) codebase, including ARCN (Sun et al., 2018), LFB (Wu et al., 2019), and SlowFast with its variant SlowOnly (Feichtenhofer et al., 2019). All models undergo training for 20 epochs with a batch size of 32. Convergence is evident from the training curves in Fig. A2c. We maintain consistent optimizers and learning rates as in official implementations. Ground-truth bounding boxes for each chimpanzee are provided during both training and testing, as per Tang et al. (2020). Please refer to Appx. C for further details on ablative modules.

Table 5: **Results of spatiotemporal action detection track on `ChimpACT` test set.** The row highlighted in light blue is the performance reference on the human action dataset AVA (Gu et al., 2018). − denotes not applicable. "*w.* NL/Max/Avg LFB" denotes using non-local, max, or average LFB module. "*w.* Ctx" indicates using both the RoI feature and the global pooled feature for classification. "mAP," "mAP$_L$," "mAP$_O$," "mAP$_S$," and "mAP$_o$" represent the overall mAP and mAP for Locomotion, Object interaction, Social interaction, and others.

Method	Frame sampling	Module	mAP	mAP$_L$	mAP$_O$	mAP$_S$	mAP$_o$
ACRN (Sun et al., 2018)	$8 \times 8 \times 1$		$24.4_{\pm 0.5}$	$58.7_{\pm 0.7}$	$33.8_{\pm 1.7}$	$14.7_{\pm 0.4}$	$0.0_{\pm 0.0}$
	$4 \times 16 \times 1$		$23.9_{\pm 1.3}$	$57.8_{\pm 0.4}$	$35.0_{\pm 4.0}$	$13.8_{\pm 1.6}$	$0.0_{\pm 0.0}$
LFB (Wu et al., 2019)	$4 \times 16 \times 1$	*w.* NL LFB	$22.0_{\pm 0.9}$	$50.1_{\pm 0.8}$	$32.3_{\pm 0.9}$	$13.5_{\pm 1.6}$	$0.6_{\pm 0.1}$
	$4 \times 16 \times 1$	*w.* Max LFB	$23.2_{\pm 0.7}$	$45.0_{\pm 1.5}$	$31.2_{\pm 0.8}$	$17.7_{\pm 1.4}$	$0.5_{\pm 0.0}$
	$4 \times 16 \times 1$	*w.* Avg LFB	$21.3_{\pm 1.6}$	$45.0_{\pm 3.6}$	$29.8_{\pm 1.1}$	$14.7_{\pm 2.6}$	$0.5_{\pm 0.0}$
SlowOnly (Feichtenhofer et al., 2019)	$8 \times 8 \times 1$		$20.9_{\pm 1.9}$	$48.1_{\pm 7.0}$	$36.2_{\pm 2.8}$	$11.5_{\pm 1.0}$	$0.0_{\pm 0.1}$
	$4 \times 16 \times 1$		$19.2_{\pm 1.1}$	$47.0_{\pm 2.5}$	$28.3_{\pm 2.5}$	$11.0_{\pm 1.2}$	$0.0_{\pm 0.1}$
	$8 \times 8 \times 1$	*w.* Ctx	$22.3_{\pm 1.9}$	$52.3_{\pm 3.2}$	$31.2_{\pm 1.3}$	$13.8_{\pm 2.4}$	$0.1_{\pm 0.1}$
	$4 \times 16 \times 1$	*w.* Ctx	$21.4_{\pm 0.9}$	$47.6_{\pm 2.0}$	$33.0_{\pm 1.2}$	$13.2_{\pm 2.2}$	$0.2_{\pm 0.1}$
SlowFast (Feichtenhofer et al., 2019)	$8 \times 8 \times 1$		$21.9_{\pm 1.0}$	$53.0_{\pm 0.7}$	$30.6_{\pm 2.2}$	$12.9_{\pm 1.2}$	$0.0_{\pm 0.1}$
	$4 \times 16 \times 1$		$22.0_{\pm 0.8}$	$52.9_{\pm 2.3}$	$33.1_{\pm 2.3}$	$12.6_{\pm 0.9}$	$0.0_{\pm 0.0}$
	$8 \times 8 \times 1$	*w.* Ctx	$24.3_{\pm 0.6}$	$56.8_{\pm 1.6}$	$31.5_{\pm 2.0}$	$15.6_{\pm 0.8}$	$0.1_{\pm 0.1}$
	$4 \times 16 \times 1$	*w.* Ctx	$24.1_{\pm 0.9}$	$56.6_{\pm 2.0}$	$34.7_{\pm 2.7}$	$14.6_{\pm 0.4}$	$0.1_{\pm 0.1}$
SlowFast (Feichtenhofer et al., 2019)	$8 \times 8 \times 1$		25.8	−	−	−	−

We adopt the same train-test split as previous tracks. Performance is gauged using mAP across 23 action classes, as per standard (Feichtenhofer et al., 2019; Tang et al., 2020). Additionally, we evaluate the mAP within the four behavioral types separately.

Results Tab. 5 summarizes the action detection algorithms' performances on the `ChimpACT` test set. The overall mAP aligns with results on human action datasets, underscoring the feasibility of automated action detection for video coding and further analyses. Locomotion behaviors achieve a notably higher mAP, likely due to their solitary nature and distinct patterns. Conversely, Conversely, the "others" category registers the lowest mAP, attributed to its limited data—comprising just 0.14% of action instances across two fine-grained classes. This imbalance suggests the potential benefit of few-shot learning methods in the future. The results highlight both the promise and areas for improvement in the dataset, positioning it as a valuable platform for advancing spatiotemporal action detection algorithms. We anticipate that `ChimpACT` will further studies into the social dynamics of non-human primates in semi-naturalistic environments.

5 Conclusion

In this work, we introduced `ChimpACT`, a novel longitudinal video dataset capturing the intricate behaviors of group-living chimpanzees, focusing on the juvenile chimpanzee, Azibo. Our meticulous annotations and diverse social interactions within the dataset offer a unique view into the world of our closest evolutionary relatives. Through comprehensive experiments, we underscored the challenges and nuances of applying human-centric computer vision algorithms to the distinct behaviors and interactions of chimpanzees. The dataset's depth, combined with its long-tail distribution, not only emphasizes its significance but also paves the way for interdisciplinary research bridging primatology, comparative psychology, computer vision, and machine learning. By making this resource available, our aspiration is to catalyze advancements in video understanding, inspire the research community to craft specialized techniques for non-human primates and deepen our collective insights into their intricate social fabric and dynamics.

Limitation and future work `ChimpACT` is based on captive chimpanzees living in a semi-natural environment, limiting the observable range of behaviors. Natural foraging, responses to predators, and intergroup encounters are absent. Focusing on Azibo overrepresents certain individuals and underrepresents others, limiting the assessment of the full social network. Nevertheless, we plan to contribute more data and labels to create a larger and more comprehensive chimpanzee dataset.

Acknowledgement

The authors would like to thank the Wolfgang Köhler Primate Research Center, BasicFinder CO., Ltd., and Keyue Zhang for annotations and quality check, Zihao Yin for discussions and preliminary experiments on the chimpanzee detection models, Guangyuan Jiang and Yuyang Li for their technical support on the GPU cluster, and NVIDIA for their generous support of GPUs and hardware. X. Ma, J. Su, W. Zhu, Y. Zhu, and Y. Wang are supported in part by the National Key R&D Program of China (2022ZD0114900), and Y. Zhu is supported in part by the Beijing Nova Program and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

References

Altmann, J. (1974). Observational study of behavior: sampling methods. *Behaviour*, 49(3-4):227–266. 2, 5

Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., and Schiele, B. (2018). Posetrack: A benchmark for human pose estimation and tracking. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 6

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3, 8

Bain, M., Nagrani, A., Schofield, D., Berdugo, S., Bessa, J., Owen, J., Hockings, K. J., Matsuzawa, T., Hayashi, M., Biro, D., et al. (2021). Automated audiovisual behavior recognition in wild primates. *Science Advances*, 7(46):eabi4883. 4

Bain, M., Nagrani, A., Schofield, D., and Zisserman, A. (2019). Count, crop and recognise: Fine-grained recognition in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2, 3, 4

Baker, T. A. (2022). Wolfgang köhler primate research center. *Encyclopedia of Animal Cognition and Behavior*, page 7310. 4

Bala, P. C., Eisenreich, B. R., Yoo, S. B. M., Hayden, B. Y., Park, H. S., and Zimmermann, J. (2020). Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature Communications*, 11(1):4560. 2, 3

Bard, K. A., Dunbar, S., Maguire-Herring, V., Veira, Y., Hayes, K. G., and McDonald, K. (2014). Gestures and social-emotional communicative development in chimpanzee infants. *American Journal of Primatology*, 76(1):14–29. 5

Bergmann, P., Meinhardt, T., and Leal-Taixe, L. (2019). Tracking without bells and whistles. In *International Conference on Computer Vision (ICCV)*. 7, 8, A5

Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10. 8

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. In *IEEE International Conference on Image Processing (ICIP)*. 4, 7, 8, A5

Boesch, C. (1996). The emergence of cultures among wild chimpanzees. In *Proceedings-British Academy*. 6

Bohnslav, J. P., Wimalasena, N. K., Clausing, K. J., Dai, Y. Y., Yarmolinsky, D. A., Cruz, T., Kashlan, A. D., Chiappe, M. E., Orefice, L. L., Woolf, C. J., et al. (2021). Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels. *Elife*, 10:e63377. 4

Cao, J., Pang, J., Weng, X., Khirodkar, R., and Kitani, K. (2023). Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 7, 8, A5

Contributors, M. (2020a). MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking. 7

Contributors, M. (2020b). Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose. 8

Contributors, M. (2020c). Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2. 9

Dawkins, M. S. (2003). Behaviour as a tool in the assessment of animal welfare. *Zoology*, 106(4):383–387. 2

Desai, N., Bala, P., Richardson, R., Raper, J., Zimmermann, J., and Hayden, B. (2022). Openapepose: a database of annotated ape photographs for pose estimation. *arXiv preprint arXiv:2212.00741*. 2, 3, 4

Fabian Caba Heilbron, Victor Escorcia, B. G. and Niebles, J. C. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *International Conference on Computer Vision (ICCV)*. 4, 9, 10, A9, A12

Fröhlich, M., Müller, G., Zeiträg, C., Wittig, R. M., and Pika, S. (2020). Begging and social tolerance: Food solicitation tactics in young chimpanzees (pan troglodytes) in the wild. *Evolution and Human Behavior*, 41(2):126–135. 2

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*. 7, 8

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12):86–92. A14

Gonyou, H. W. (1994). Why the study of animal behavior is associated with the animal welfare issue. *Journal of Animal Science*, 72(8):2171–2177. 2

Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al. (2018). Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3, 10

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *International Conference on Computer Vision (ICCV)*. 4

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4, 7

Hobaiter, C., Samuni, L., Mullins, C., Akankwasa, W. J., and Zuberbühler, K. (2017). Variation in hunting behaviour in neighbouring chimpanzee communities in the budongo forest, uganda. *PloS One*, 12(6):e0178065. 2

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*. 3

Labuguen, R., Matsumoto, J., Negrete, S. B., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K.-i., and Shibata, T. (2021). Macaquepose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture. *Frontiers in Behavioral Neuroscience*, 14:581154. 2, 3, 4

Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., Inoue, E., Inoue-Muruyama, M., Mitani, J. C., Muller, M. N., et al. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences (PNAS)*, 109(39):15716–15721. 2

Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M. M., Di Santo, V., Soberanes, D., Feng, G., et al. (2022). Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods*, 19(4):496–504. 4

Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., and Lu, C. (2021). Human pose regression with residual log-likelihood estimation. In *International Conference on Computer Vision (ICCV)*. 9, A8, A9

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*. 3, 6, 8, 9

Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., and Leibe, B. (2021). Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision (IJCV)*, 129:548–578. 8

Luncz, L. V., Sirianni, G., Mundry, R., and Boesch, C. (2018). Costly culture: differences in nut-cracking efficiency between wild chimpanzee groups. *Animal Behaviour*, 137:63–73. 2

Marks, M., Jin, Q., Sturman, O., von Ziegler, L., Kollmorgen, S., von der Behrens, W., Mante, V., Bohacek, J., and Yanik, M. F. (2022). Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. *Nature Machine Intelligence*, 4(4):331–340. 2, 3, 4

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289. 4, A16

Matsuzawa, T. (2013). Evolution of the brain and social behavior in chimpanzees. *Current Opinion in Neurobiology*, 23(3):443–449. 5

Matsuzawa, T., Tomonaga, M., and Tanaka, M. (2006). *Development in chimpanzees*. Springer. 5

McEwen, E. S., Warren, E., Tenpas, S., Jones, B., Durdevic, K., Rapport Munro, E., and Call, J. (2022). Primate cognition in zoos: Reviewing the impact of zoo-based research over 15 years. *American Journal of Primatology*, 84(10):e23369. 4

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. (2016). Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*. 3, 8, A5

Musgrave, S., Lonsdorf, E., Morgan, D., and Sanz, C. (2021). The ontogeny of termite gathering among chimpanzees in the goualougo triangle, republic of congo. *American journal of physical anthropology*, 174(2):187–200. 5

Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*. 9, A8, A9

Ng, X. L., Ong, K. E., Zheng, Q., Ni, Y., Yeo, S. Y., and Liu, J. (2022). Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 3, 4, 8

Nishida, T., Zamma, K., Matsusaka, T., Inaba, A., and McGrew, W. C. (2010). *Chimpanzee behavior in the wild: an audio-visual encyclopedia*. Springer Science & Business Media. 5

Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., and Yu, F. (2021). Quasi-dense similarity learning for multiple object tracking. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4, 8, A5

Pedersen, M., Haurum, J. B., Bengtson, S. H., and Moeslund, T. B. (2020). 3d-zef: A 3d zebrafish tracking benchmark dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 7, 8

Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*. 8

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 9, A8, A9

Schapiro, S. J., Perlman, J. E., Thiele, E., and Lambeth, S. (2005). Training nonhuman primates to perform behaviors useful in biomedical research. *Lab Animal*, 34(5):37–42. 2

Schindler, F. and Steinhage, V. (2021). Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. *Ecological Informatics*, 61:101215. 4

Sirianni, G., Mundry, R., and Boesch, C. (2015). When to choose which tool: multidimensional and conditional selection of nut-cracking hammers in wild chimpanzees. *Animal Behaviour*, 100:152–165. 2

Sun, C., Shrivastava, A., Vondrick, C., Murphy, K., Sukthankar, R., and Schmid, C. (2018). Actor-centric relation network. In *European Conference on Computer Vision (ECCV)*. 9, 10, A12

Sun, J. J., Karigo, T., Chakraborty, D., Mohanty, S. P., Wild, B., Sun, Q., Chen, C., Anderson, D. J., Perona, P., Yue, Y., et al. (2021). The multi-agent behavior dataset: Mouse dyadic social interactions. *arXiv preprint arXiv:2104.02710*. 3

Sun, J. J., Marks, M., Ulmer, A. W., Chakraborty, D., Geuther, B., Hayes, E., Jia, H., Kumar, V., Oleszko, S., Partridge, Z., et al. (2023). Mabe22: A multi-species multi-task benchmark for learned representations of behavior. In *International Conference on Machine Learning (ICML)*. 3

Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3, 4, 8, 9, A8, A9

Surbeck, M., Boesch, C., Girard-Buttoz, C., Crockford, C., Hohmann, G., and Wittig, R. M. (2017). Comparison of male conflict behavior in chimpanzees (pan troglodytes) and bonobos (pan paniscus), with specific regard to coalition and post-conflict behavior. *American Journal of Primatology*, 79(6):e22641. 2

Tang, J., Xia, J., Mu, X., Pang, B., and Lu, C. (2020). Asynchronous interaction aggregation for action detection. In *European Conference on Computer Vision (ECCV)*. 9, 10

The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87. 2, 3

Van Leeuwen, E. J. (2021). Temporal stability of chimpanzee social culture. *Biology Letters*, 17(5):20210031. 5

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. A9

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 8, 9, A8, A9

Wiltshire, C., Lewis-Cheetham, J., Komedová, V., Matsuzawa, T., Graham, K. E., and Hobaiter, C. (2023). Deepwild: Application of the pose estimation tool deeplabcut for behaviour tracking in wild chimpanzees and bonobos. *Journal of Animal Ecology*. 2, 4

Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing (ICIP)*. 7, 8, A5

Wu, C.-Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., and Girshick, R. (2019). Long-term feature banks for detailed video understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 9, 10, A9, A12

Xiao, B., Wu, H., and Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*. 3, 4, 8, 9, A8, A9

Yang, Y., Yang, J., Xu, Y., Zhang, J., Lan, L., and Tao, D. (2022). Apt-36k: A large-scale benchmark for animal pose estimation and tracking. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3

Yao, Y., Bala, P., Mohan, A., Bliss-Moreau, E., Coleman, K., Freeman, S. M., Machado, C. J., Raper, J., Zimmermann, J., Hayden, B. Y., et al. (2023). Openmonkeychallenge: Dataset and benchmark challenges for pose estimation of non-human primates. *International Journal of Computer Vision (IJCV)*, 131(1):243–258. 2, 3

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. A3

Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., and Tao, D. (2021). Ap-10k: A benchmark for animal pose estimation in the wild. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*. 2, 3

Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., and Wang, J. (2021). Hrformer: High-resolution vision transformer for dense predict. In *Advances in Neural Information Processing Systems (NeurIPS)*. 9, A8, A9

Zhang, F., Zhu, X., Dai, H., Ye, M., and Zhu, C. (2020). Distribution-aware coordinate representation for human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 8, 9, A8, A9

Zhang, L., Gao, J., Xiao, Z., and Fan, H. (2023). Animaltrack: A benchmark for multi-animal tracking in the wild. *International Journal of Computer Vision (IJCV)*, 131(2):496–513. 3

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., and Wang, X. (2022). Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision (ECCV)*. 7, 8, A5