# AlphaChimp: Tracking and Behavior Recognition of Chimpanzees

Xiaoxuan Ma[1†], Yutang Lin[2,5,6,7,8,9†], Yuan Xu[1], Stephan P. Kaufhold[3], Jack Terwilliger[3], Andres Meza[4], Yixin Zhu[2,5,7,8,9*], Federico Rossano[3*], Yizhou Wang[1,7]

[1]Center on Frontiers of Computing Studies, School of Computer Science, Peking University.
[2]School of Psychological and Cognitive Sciences, Peking University.
[3]Department of Cognitive Science, University of California San Diego.
[4]Department of Computer Science & Engineering, University of California San Diego.
[5]Institute for Artificial Intelligence, Peking University.
[6]Yuanpei College, Peking University.
[7]State Key Laboratory of General Artificial Intelligence, Peking University.
[8]Beijing Key Laboratory of Behavior and Mental Health, Peking University.
[9]Embodied Intelligence Lab, PKU-Wuhan Institute for Artificial Intelligence.

*Corresponding author(s). E-mail(s): yixin.zhu@pku.edu.cn; frossano@ucsd.edu;
Contributing authors: maxiaoxuan@pku.edu.cn; yutang.lin@stu.pku.edu.cn; xuyuan@stu.pku.edu.cn; spkaufho@ucsd.edu; jterwilliger@ucsd.edu; anmeza@ucsd.edu; yizhou.wang@pku.edu.cn;
†Equal contribution.

**Abstract**

Understanding non-human primate behavior is essential for advancing animal welfare and uncovering the roots of human sociality. However, automated analysis remains limited. Existing methods, many of which are human-centric, are typically task-specific, handling detection, tracking, or behavior recognition in isolation. We present AlphaChimp, an end-to-end and unified framework for chimpanzee detection, tracking, and spatiotemporal behavior recognition. To address the unique challenges of chimpanzee video analysis, such as frequent occlusions and social interactions, we build upon a DETR-based architecture with crucial modifications. Our model integrates multi-resolution temporal features to capture long-term contextual cues and employs attention mechanisms to model spatial relationships between individuals. This unique design allows AlphaChimp to jointly capture individual and social interactions. Evaluated on the `ChimpACT` dataset, the only benchmark with fine-grained spatiotemporal annotations of chimpanzee behavior, our method achieves state-of-the-art performance, with a 10% gain in tracking and a 20% improvement in behavior recognition, particularly for social behaviors. It also surpasses recent video foundation models. Our approach bridges computer vision and primatology, enabling scalable, objective analysis of primate social behavior to facilitate future research. We release our code and models at project page and hope this will facilitate future research in animal social dynamics.

**Keywords:** Computer Vision, CV for Animals, Primatology

# 1 Introduction

Studying non-human primate behavior is essential for understanding human evolution (Langergraber et al., 2012) and advancing animal welfare (Dawkins, 2003; Gonyou, 1994). As our closest relatives, primates offer a uniquely ethical and effective model for exploring the origins of complex social behavior (The Chimpanzee Sequencing and Analysis Consortium, 2005). However, traditional fieldwork is labor-intensive, requiring years of habituation, video collection, and manual behavior coding (Hobaiter et al., 2017; Fröhlich et al., 2020; Surbeck et al., 2017; Luncz et al., 2018; Sirianni et al., 2015). Although manual annotation remains the gold standard for capturing subtle, context-rich behaviors (Wiltshire et al., 2023), it is time-consuming, expertise-dependent, and susceptible to subjective bias. These challenges underscore the need for scalable, objective, and efficient methods to assist discovery in primate research.

Recent advances in computer vision have enabled promising directions for the automated analysis of non-human primate behavior, particularly in chimpanzees (He et al., 2017; Liu et al., 2022). However, most existing methods are human-centric, designed for human detection, tracking (Cao et al., 2023; Zhang et al., 2022; Gritsenko et al., 2024), and action recognition (Feichtenhofer et al., 2019; Yu et al., 2022), and transfer poorly to primate contexts. A major reason for this poor transferability is the paucity of high-quality chimpanzee datasets, hindering progress in this field. Assembling comprehensive chimpanzee behavioral data is an arduous task that demands substantial resources and expertise. This process involves continuous video recording coupled with meticulous manual annotation, with an emphasis on annotation accuracy.

Existing primate datasets present various limitations as summarized in Tab. 1. Some studies (Marks et al., 2022; Bala et al., 2020; Fuchs et al., 2024) restrict subjects to indoor enclosures, which may result in unnatural behaviors. Other works (Labuguen et al., 2021; Desai et al., 2023; Ng et al., 2022; Yao et al., 2023) rely on labeling primate images from online resources. While offering scaling-up advantages, they fail to capture complex social dynamics inherent to group-living primates such as chimpanzees. Other works (Bain et al., 2019; Brookes et al., 2024b, 2025) collect data from the wild, but they lack clear social bonds or behavioral ethogram. This oversight significantly limits comprehensive studies of chimpanzees' social behaviors and relationships, which are crucial aspects of their natural behavior patterns. Therefore, this work is based on our `ChimpACT` dataset (Ma et al., 2023), which was previously published at NeurIPS 2023. `ChimpACT` is a comprehensive longitudinal dataset for in-depth study of chimpanzee social behavior in a semi-naturalistic setting, with detailed annotations across multiple tasks. See Fig. 1 for a dataset example.

Yet, to date, no method, including recent general-purpose video models (Gritsenko et al., 2024; Zhao et al., 2024), offers an end-to-end framework that jointly performs detection, tracking, and recognition of both *fine-grained* solitary actions (*e.g.*, resting) and complex *social* interactions (*e.g.*, grooming, being groomed) in general videos or chimpanzee videos. Transferring human-centric models can often only handle one specific task and leads to notable performance drops, particularly in recognizing social behaviors (as validated in Tab. 3). Moreover, such task-specific, multi-stage pipelines may suffer from error propagation (Tab. 5), for instance, inaccuracies in bounding box detection can cascade into downstream action recognition errors. This underscores the need for chimpanzee-adapted joint tracking, detection, and behavior recognition perception models.

To this end, we propose the first end-to-end framework, AlphaChimp, for simultaneous chimpanzee detection, tracking, and behavior recognition in videos (Fig. 4). While our model builds upon a general-purpose DETR-based architecture (Carion et al., 2020; Zhu et al., 2021; Zhang et al., 2023a), we introduce key adaptations tailored to the unique challenges of primate video analysis, such as frequent occlusions, and fine-grained social interactions. Specifically, we integrate multi-resolution temporal features to capture contextual cues over time and employ attention mechanisms (Vaswani et al., 2017) to model spatial relationships between individuals. By focusing on both contextual and spatiotemporal dynamics, AlphaChimp can capture the nuances of chimpanzee behavior more effectively,


**Fig. 1**: **Sample frames and annotations from the `ChimpACT` dataset.** We present three video sequences where an infant chimpanzee, named Azibo, is focused. While we also annotate visibility for both the bounding box and the keypoint, these are omitted here for clarity.

which enhances the accuracy of recognizing complex *social* behaviors among chimpanzees.

We evaluate our method on the `ChimpACT` dataset (Ma et al., 2023), the only benchmark with both chimpanzee tracking and fine-grained spatiotemporal action labels. Experiments show that our model not only streamlines primate video understanding but also achieves State-of-the-Art (SOTA) performance across tasks, with a 10% improvement in tracking accuracy and a 20% accuracy gain in behavior recognition, particularly for complex social interactions. Notably, it even outperforms recent large video foundation models such as VideoPrism (Zhao et al., 2024). We hope this work could underscore the potential of specialized, integrated models in advancing primate behavioral research through enhanced computational perception, offering a powerful new tool for analyzing complex chimpanzee behaviors and social dynamics.

In summary, our contributions are threefold:

- We introduce AlphaChimp, the first end-to-end and unified framework designed for automated detection, tracking, and fine-grained behavioral recognition of chimpanzees in video footage.
- AlphaChimp achieves notable improvement over all existing SOTA models on the `ChimpACT` benchmark across diverse tasks, with a 10% improvement in tracking and a 20% improvement in behavior recognition, despite those models being specifically tailored for each task.
- This unified framework AlphaChimp and the `ChimpACT` dataset collectively offer a new resource and platform for the community for advanced techniques for better perception of chimpanzees, ultimately contributing to a deeper understanding of non-human primates.

## 2 Related work

### 2.1 Computer vision for animals

In recent years, several new datasets and benchmarks leveraging computer vision techniques to

**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" column, G represents general, P for primates, M for macaques, B for baboons, C for chimpanzees, and C+g for chimpanzees and gorillas. In the "Source" column, I stands for Internet, Z for zoo, G 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., 2023) | 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 | ✗ | ✗ | G ($6.7m^2$) |
| 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 | ⊘ | G ($15m^2$) |
| BaboonLand (Duporge et al., 2025) | B | ✗ | 30,000 | ⊘ | ✓ | ✗ | ✗ | ✗ | ✗ | 12 | ⊘ | W |
| PanAf20K (Brookes et al., 2024b) | C+g | ✗ | 179,956 [⊘] | ⊘ | ✗ | ✗ | ✗ | ✗ | ✗ | 9 | 201,516 [⊘] | W |
| CCR (Bain et al., 2019) | C | 13 | 936,914 | 1,937,585 | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | W |
| ChimpBehave (Fuchs et al., 2024) | C | ✗ | 12,000 | ⊘ | ✗ | ✗ | ✗ | ✗ | ✗ | 7 | ⊘ | Z |
| PanAf-FGBG (Brookes et al., 2025) | C | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | 14 | ⊘ | W |
| `ChimpACT` (Ours) (Ma et al., 2023) | C | 23 | 160,500 | 56,324 | ✓ | 16,028 | 56,324 | ✓ | 2D | 23 | 64,289 | ($4400m^2$) CP |

advance animal research have been introduced. These efforts span a wide range of species and tasks. For example, 3D-ZeF20 (Pedersen et al., 2020) introduces 3D tracking of zebrafish to the Multi-Object Tracking (MOT) benchmarks, while AnimalTrack (Zhang et al., 2023b) focuses on multi-animal tracking across various species. In the realm of pose estimation, AP-10K (Yu et al., 2021) and APT-36K (Yang et al., 2022) address this task for diverse species. AnimalKingdom (Ng et al., 2022) extends the scope to fine-grained multi-label action recognition. Several studies explore multi-agent behavior understanding from a social interaction perspective (Sun et al., 2021, 2023). KABR (Kholiavchenko et al., 2024) contributes by collecting videos from drones flown over the Mpala Research Centre in Kenya. Recently, PanAf20K (Brookes et al., 2024b) and PanAf-FGBG (Brookes et al., 2025) curate datasets for chimpanzee behavior recognition, though they still lack clear social bonds or fine-grained ethogram. Distinctively, `ChimpACT` (Ma et al., 2023) stands out as a comprehensive benchmark, encompassing three varied downstream tasks and featuring rich annotations of social interactions within the same chimpanzee group. This approach allows for a more nuanced and longitudinal analysis of animal behavior, particularly in the context of chimpanzee social dynamics.

## 2.2 Human video datasets

In contrast to animal-centric video datasets, a more substantial collection exists for human subjects, addressing diverse human-centric video understanding tasks. These datasets cover a wide range of applications in computer vision. For multi-person tracking, the MOT Challenge (Milan et al., 2016) serves as a primary benchmark. Human pose estimation is well-served by datasets such as COCO (Lin et al., 2014) and MPII (Andriluka et al., 2014), which provide extensive annotations for body keypoints. In the domain of action recognition, datasets like Kinetics (Kay

et al., 2017), ActivityNet (Fabian Caba Heilbron and Niebles, 2015), and AVA (Gu et al., 2018) offer large-scale video collections with diverse human activities. While `ChimpACT` (Ma et al., 2023) encompasses analogous tasks to these human-centric datasets, it introduces unique challenges specific to chimpanzee behavior. This approach allows for the adaptation and advancement of human-centric computer vision techniques to the study of non-human primates, bridging the gap between human and animal behavior analysis.

### 2.3 Datasets on primate understanding

Most existing primate datasets focus primarily on individual primate detection and pose estimation, limiting their utility for contextual behavioral understanding. Lab-based datasets (Bala et al., 2020; Marks et al., 2022) may induce atypical behaviors, while online-sourced data (Labuguen et al., 2021; Desai et al., 2023; Ng et al., 2022; Yao et al., 2023) often lack longitudinal interactions essential for studying social dynamics. Although the CCR dataset (Bain et al., 2019) tracks wild chimpanzees over two years, it lacks behavioral annotations. BaboonLand (Duporge et al., 2025) introduces a drone-captured dataset for baboon tracking and behavior analysis. Among available resources, `ChimpACT` (Ma et al., 2023) stands out by providing comprehensive annotations, including detection, tracking, and fine-grained spatiotemporal action labels, covering both solitary and social behaviors. In contrast, recent datasets like PanAf20K (also PanAf500, a subset of PanAf20K) (Brookes et al., 2024b), ChimpBehave (Fuchs et al., 2024), and PanAf-FGBG (Brookes et al., 2025) offer only coarse or short-term annotations, with limited coverage of social interactions. Given its richness and level of detail, we adopt `ChimpACT` (Ma et al., 2023) as the benchmark dataset for training and evaluating our method.

### 2.4 Methods on primate understanding

Understanding primate behavior requires first detecting and tracking individuals, followed by recognizing their actions. However, current methods rarely support this full pipeline. Most focus on either individual localization (Bain et al., 2019; Marks et al., 2022) or behavior classification (Bain et al., 2021; Brookes et al., 2024a; Zhao et al., 2024), but not both. Detection and tracking (Marks et al., 2022; Bain et al., 2019) often rely on general object detectors like Mask R-CNN (He et al., 2017). SIPEC (Marks et al., 2022) applies it with a ResNet backbone (He et al., 2016) for macaques, while Bain *et al.* (Bain et al., 2019) use a two-stage CNN for chimpanzees. Behavior recognition often builds on human action models (Bain et al., 2021; Brookes et al., 2024a; Zhao et al., 2024). Bain *et al.* (Bain et al., 2019) use SlowFast (Feichtenhofer et al., 2019) for classifying simple behaviors like nut cracking. Recent large-scale video foundation models such as VideoPrism (Zhao et al., 2024) can recognize spatiotemporal actions but lack support for primate detection and tracking.

In contrast, we propose the first unified framework AlphaChimp that jointly detects chimpanzees and recognizes over 20 fine-grained behaviors, significantly outperforming state-of-the-art methods across both detection and recognition tasks.

## 3 Preliminary work: `ChimpACT`

To facilitate understanding of this work, we first provide a brief overview of our `ChimpACT` (Ma et al., 2023) dataset. It is a comprehensive collection of high-resolution video footage documenting chimpanzees at the Leipzig Zoo in Germany from 2015 to 2018, which is semi-naturalistic habitats comprising both indoor and outdoor enclosures; see also Fig. 2. The indoor enclosure (Fig. 2b), spanning approximately 400 $m^2$, is equipped with a diverse array of environmental enrichments. These include 15 wooden climbing structures with heights ranging from 2 to 5 meters, 8 hammocks, vegetation consisting of over 20 plant species, and 12 foraging boxes designed to simulate natural food-finding behaviors. Weather permitting, the chimpanzees have access to a 4000 $m^2$ outdoor area (Fig. 2c). This expansive space features abundant vegetation, including 30 trees of various species, and is bordered by a 3-meter wide artificial river. The outdoor environment is further enhanced with enrichments similar to those


(a) aerial view of Leipzig Zoo

(b) indoor enclosure

(c) outdoor enclosure

**Fig. 2**: **Semi-naturalistic habitats at Leipzig Zoo.** (a) The aerial view of Leipzig Zoo (Earth, 2024), including both indoor and outdoor enclosure. (b) An example scene inside the indoor enclosure (Lehmann, 2018). Photo used under CC BY-SA 4.0. (c) An example scene of the outdoor enclosure.

found in the indoor space. Although the semi-naturalistic setting aims to approximate wild conditions, it still differs from fully wild habitats, where collecting high-quality longitudinal data remains an open challenge.

The dataset comprises approximately 2 hours of recordings, focusing primarily on Azibo, a male chimpanzee born in April 2015 to Swela. Azibo has been a member of the A-chimpanzee group since birth, providing a unique opportunity for longitudinal observation of his behavioral development and social interactions. The A-chimpanzee group, consisting of over 20 individuals, is one of the most extensively studied zoo-residing chimpanzee cohorts. Its members have been subjects of numerous behavioral and cognitive studies, including both observational and experimental designs, conducted by researchers affiliated with the Max Planck Institute for Evolutionary Anthropology (Baker, 2022; McEwen et al., 2022).

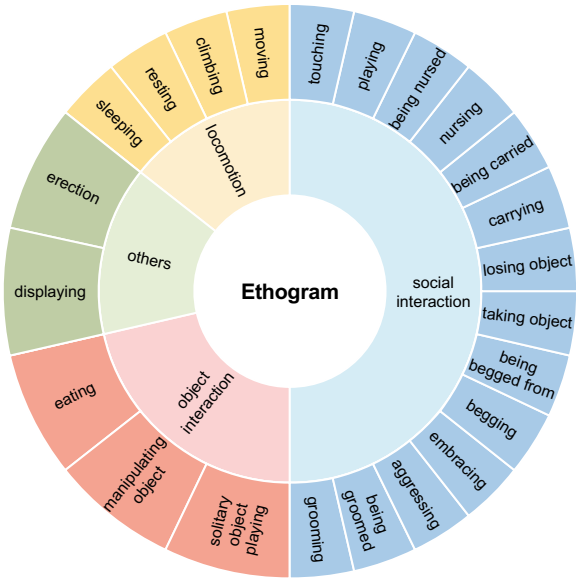


**Fig. 3**: **Ethogram with annotated behaviors.** The ethogram is structured into four categories: locomotion, object interaction, social interaction, and others. Each category includes annotated behaviors that belong to it.

ChimpACT's longitudinal nature offers an unprecedented window into Azibo's growth, social interactions, and intra-group relationships within this complex social environment. To systematically categorize and analyze these behaviors, we have developed a detailed ethogram, which serves as a comprehensive catalog of behavioral categories. This ethogram, visually represented in Fig. 3, is structured into four primary categories, including locomotion and social interaction. Each of these categories is further delineated into several fine-grained actions, allowing for a nuanced analysis of chimpanzee behavior. The dataset captures a wide range of behaviors and social dynamics, and these are the behavioral categories our method aims to predict.

For more details of the dataset, including data collection process, annotations, and statistics, readers are directed to Ma et al. (2023).

# 4 AlphaChimp

In this section, we present AlphaChimp, a unified framework for video-based chimpanzee detection

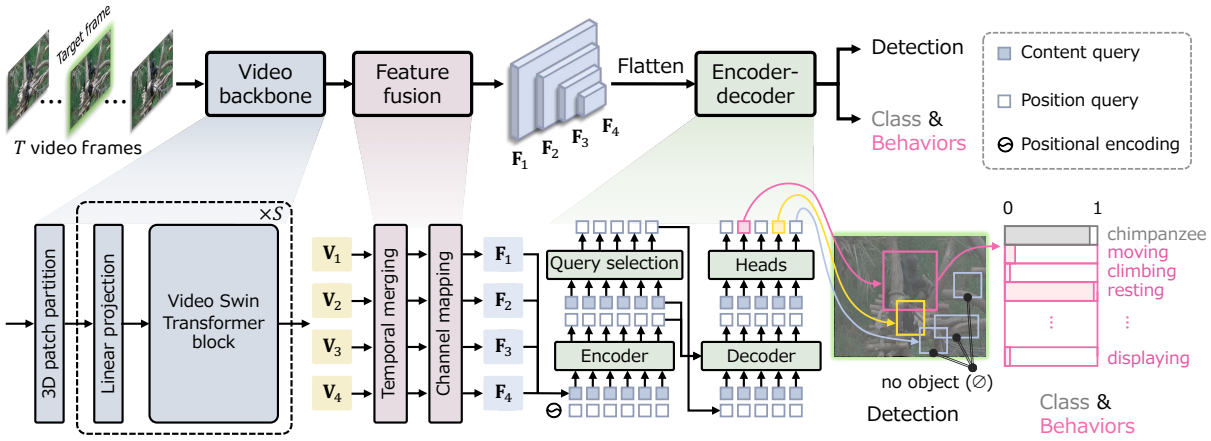


**Fig. 4**: **Overview of our framework for tracking and behavior recognition of primates.** Given a video clip, our method predicts a set of detection boxes for the target frame and simultaneously identifies the class label and behaviors within each bounding box.

and fine-grained spatiotemporal behavior recognition. As shown in Fig. 4, our method integrates spatial and temporal information from video input to jointly localize individuals and classify their behaviors in a fully end-to-end fashion. The framework consists of three main components: a multi-scale temporal feature extractor, *i.e.* video backbone (Sec. 4.1), a temporal feature fusion module (Sec. 4.2), and a Transformer-based encoder-decoder for joint detection and behavior classification (Sec. 4.3).

## 4.1 Multi-scale temporal feature extraction

To extract multi-resolution temporal features, we employ the Video Swin Transformer (Liu et al., 2022) as our backbone network. This architecture processes a video sequence of $T$ frames, where each frame is composed of $H \times W \times 3$ pixels. The backbone treats each 3D patch of size $2 \times 4 \times 4 \times 3$ as a token, enabling it to capture both spatial and temporal information effectively.

The initial 3D patch partition layer transforms the input into $\frac{T}{2} \times \frac{H}{4} \times \frac{W}{4}$ 3D tokens, with each token represented by a $C_{in}$-dimensional feature vector. This transformation sets the stage for subsequent processing through a series of $S$ Video Swin Transformer blocks (Liu et al., 2022), each generating temporal features at different resolutions.

Each block incorporates a linear projection layer, adhering to the standard practice outlined in Liu et al. (2022). These projection layers perform patch merging along the spatial dimensions, progressively reducing spatial resolution while increasing feature dimensionality. It's worth noting that the linear projection in the first block is an exception, as it maintains the original spatial dimensions.

The output of this process is a set of $S$ multi-scale temporal features, denoted as $\{\mathbf{V}_i\}_{i=1}^{S}$, where each feature $\mathbf{V}_i$ is extracted from its corresponding block. This multi-scale approach allows our model to capture both fine-grained details and broader contextual information, crucial for accurate chimpanzee detection and behavior recognition.

For a more detailed description of the architecture, including specific layer configurations and feature dimensions, we refer readers to Sec. A1.

## 4.2 Temporal feature fusion

The multi-scale temporal features extracted by the backbone network provide essential contextual information for predicting the target frame. This section describes the process of fusing these features to create a unified representation that integrates temporal context effectively.

Initially, each feature $\mathbf{V}_i$ in the set $\{\mathbf{V}_i\}_{i=1}^{S}$ has a temporal dimension of size $\frac{T}{2}$, reflecting the temporal extent of the input video sequence. To

condense this temporal information, we employ a temporal merging layer. This layer utilizes convolution operations to reduce the temporal dimension of each feature to 1, effectively aggregating information across the time axis.

Following temporal merging, a channel mapping layer is applied. This layer, also implemented using convolution, serves to standardize the feature channels across all scales to a common dimension. The result of this process is a set of $S$ multi-scale features, denoted as $\{\mathbf{F}_i\}_{i=1}^{S}$, where each feature $\mathbf{F}_i$ has $C$ feature dimensions. For example, $\mathbf{F}_1 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ represents the feature at the first scale, maintaining the spatial dimensions of $\frac{H}{4} \times \frac{W}{4}$ with $C$ channels.

These fused features effectively integrate the temporal context from the input video sequence, providing a robust foundation for the subsequent tasks of chimpanzee detection and behavior recognition in the target frame. By preserving information across multiple scales while condensing temporal information, this fusion process enables our model to capture both fine-grained details and broader contextual cues necessary for accurate analysis of chimpanzee behavior.

## 4.3 Joint detection and behavior classification

Building upon advanced DETR-series models (Carion et al., 2020; Zhu et al., 2021; Zhang et al., 2023a), particularly DINO (Zhang et al., 2023a), we develop a comprehensive model for chimpanzee detection and behavior classification. Our approach simultaneously determines the category, location, and behaviors of chimpanzees in the target frame, leveraging a combination of position and content queries to enhance accuracy.

**Overview.** The process begins with cross-scale feature fusion using the Transformer encoder's self-attention mechanism. This integrates information from the multi-level features $\{\mathbf{F}_i\}_{i=1}^{S}$, capturing both fine-grained details and high-level spatial context. We flatten and concatenate features of different scales to form initial content queries for the encoder, with corresponding positional encodings serving as position queries.

A query selection mechanism then identifies $Q$ encoder features as initial position queries for the decoder. The decoder, augmented with auxiliary heads, transforms these position queries into bounding boxes while optimizing content queries for class and behavior determination. This dual-query approach enables precise detection and simultaneous classification of chimpanzees and their behaviors, including fine-grained multi-label prediction for complex actions.

For training, we adopt loss functions common to the DETR series (Carion et al., 2020; Zhang et al., 2023a) for box regression and classification, employing one-to-one bipartite matching. To address multi-label behavior classification, we implement focal loss (Lin et al., 2017), which effectively handles class imbalance in such scenarios.

**Query selection.** To optimize position queries efficiently, we adopt a query selection scheme inspired by works (Zhang et al., 2023a; Zhu et al., 2021). This process involves appending a classification head network after the encoder to compute confidence scores for each query, representing the likelihood of containing a chimpanzee. The top $Q$ queries with the highest confidence are selected as initial position queries for the decoder.

This selection mechanism serves to focus the model's attention on the most relevant spatial locations, potentially improving both detection accuracy and computational efficiency. By prioritizing high-confidence regions, the decoder can more effectively refine its predictions. The architecture of the classification head network used here is shared with other head networks in our model, details of which are elaborated in the subsequent section.

**Prediction heads.** Our model employs two specialized prediction heads: one for bounding box regression (detection) and another for behavior classification. Both are implemented as Multilayer Perceptrons (MLPs) and utilize refined query features from the last decoder layer's output.

The box regression head transforms position queries into $\mathbf{B} = \{\mathbf{b}_q\}_{q=1}^{Q} \in \mathbb{R}^{Q \times 4}$, where each $\mathbf{b}_q \in [0, 1]^4$ represents the predicted box position for the $q^{th}$ query. Box positions are defined using center coordinates, height, and width, all relative to the image size. Once the bounding boxes for

each frame are obtained, we adopt the same association strategy proposed in ByteTrack (Zhang et al., 2022) to associate bounding boxes across frames and obtain tracking results.

The behavior head network converts content queries into two outputs: $\mathbf{C} = \{c_q\}_{q=1}^{Q} \in \mathbb{R}^{Q\times 1}$, representing class probabilities, and $\mathbf{A} = \{\mathbf{a}_q\}_{q=1}^{Q} \in \mathbb{R}^{Q\times K}$, representing behavior probabilities, where $K = 23$ is the total number of behavior classes. Here, $c_q \in [0, 1]$ provides the confidence score for the $q^{th}$ query, indicating the probability of a chimpanzee in the bounding box. $\mathbf{a}_q \in [0, 1]^K$ is a multi-label probability vector obtained using the Sigmoid function, where each dimension represents the probability of a corresponding behavior class.

For query selection, we integrate an additional behavior head following the encoder. This head generates class probabilities used as confidence scores for selecting queries, as discussed in the previous section.

**Training losses.** Our training process follows established practices in object detection (Carion et al., 2020; Zhang et al., 2023a). We begin with bipartite matching based on bounding box positions and class labels to establish a one-to-one correspondence between predicted and ground-truth (GT) sets. The matching result could be represented in pairs of $(\sigma(i), i)$, denoting the $i$-th GT box is matched with the $\sigma(i)$-th predicted box. This matching ensures that each prediction is uniquely associated with a GT object, facilitating more effective learning.

Following the matching, we apply set prediction losses for both box regression and classification. For box regression, we employ a combination of $L_1$ loss and Generalized Intersection over Union (GIoU) loss (Rezatofighi et al., 2019). The $L_1$ loss addresses direct positional errors:

$$\mathcal{L}_{\text{L1}}(\hat{b}_i, b_{\sigma(i)}) = \|\hat{b}_i - b_{\sigma(i)}\|_1, \tag{1}$$

where $\hat{b}_i$ denotes the GT box coordinates and $b_{\sigma(i)}$ represents the predicted box at index $\sigma(i)$. The GIoU loss captures the overall geometric similarity:

$$\mathcal{L}_{\text{GIoU}}(\hat{b}_i, b_{\sigma(i)}) = 1 - \left( \frac{|\hat{b}_i \cap b_{\sigma(i)}|}{|\hat{b}_i \cup b_{\sigma(i)}|} - \frac{|H \setminus (\hat{b}_i \cup b_{\sigma(i)})|}{|H|} \right), \tag{2}$$

where $|\cdot|$ denotes the area of the box, $\setminus$ denotes relative complement, while $H$ represents the smallest convex hull that encloses both $\hat{b}_i$ and $b_{\sigma(i)}$.

For classification, including both object class and behavior classes, we utilize focal loss (Lin et al., 2017):

$$\mathcal{L}_{\text{FL}}(p) = -(1-p)^{\gamma} \log(p). \tag{3}$$

Here, $p$ is the model's estimated probability for the target class, $\gamma \geq 0$ is a tunable focusing parameter that reduces the loss contribution from easy examples. This choice is particularly effective for handling class imbalance, which is common in multi-label classification scenarios like behavior recognition. The overall loss is the weighted sum of the box regression losses and the clasification loss, formulate as:

$$\mathcal{L} = \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{GIoU}}\mathcal{L}_{\text{GIoU}} + \lambda_{\text{FL}}\mathcal{L}_{\text{FL}}. \tag{4}$$

This comprehensive loss formulation ensures that our model learns to accurately localize chimpanzees while simultaneously classifying their behaviors, addressing the multi-faceted nature of our task.

# 5 Experiments

## 5.1 Implementation details

Our framework processes video sequences of $T = 8$ frames. We employ Swin-L (Liu et al., 2022) as the video backbone, comprising $S = 4$ Swin Transformer blocks. The Transformer architecture consists of 12 encoder and 12 decoder layers, incorporating 4 reference points (Zhu et al., 2021) in the deformable attention module. Detailed architectural specifications are available in Sec. A1. Based on our dataset analysis revealing an average of 3 chimpanzees per image, with a maximum of 9, we set the decoder's fixed query number to $Q = 10$. For tracking, we use a resolution of $576\times576$, while for spatiotemporal action detection, we employ a $256 \times 256$ resolution. These resolutions align with existing methods in their respective tracks,

**Table 2: Results of the detection and tracking on the `ChimpACT` test set (Ma et al., 2023).** The row highlighted in light blue is the performance reference on the human tracking dataset MOT-17 (Milan et al., 2016). – denotes unreported. To ensure a fair comparison with prior work, results are reported at an input resolution of $576 \times 576$, which differs from the original `ChimpACT` (Ma et al., 2023) setting.

| Method | HOTA ↑ | MOTA ↑ | MOTP ↑ | IDF1 ↑ | mAP ↑ | nFP ↓ | nFN ↓ | nIDs ↓ |
|---|---|---|---|---|---|---|---|---|
| OC-SORT (Cao et al., 2023) | 63.2 | 78.0 | – | 77.5 | – | 2.7 | 19.0 | 0.3 |
| SORT (Bewley et al., 2016) | 27.2 | 34.4 | 23.1 | 24.0 | 63.7 | 12.4 | 48.8 | 4.1 |
| DeepSORT (Wojke et al., 2017) | 30.8 | 31.8 | 23.0 | 32.4 | 63.7 | 12.1 | 48.6 | 7.1 |
| QDTrack (Pang et al., 2021) | 47.9 | 46.5 | 24.1 | 52.6 | 73.9 | 24.8 | 27.2 | 1.4 |
| ByteTrack (Zhang et al., 2022) | 38.0 | 35.4 | **25.7** | 45.2 | 58.5 | **9.3** | 49.4 | 0.7 |
| OC-SORT (Cao et al., 2023) | 41.5 | 39.1 | 25.1 | 47.2 | 64.4 | 14.2 | 45.3 | 1.2 |
| YOLOv8 (Jocher et al., 2023) | 41.1 | 39.6 | 20.8 | 43.8 | 60.5 | 21.6 | 37.9 | 4.9 |
| YOLO11 (Jocher et al., 2024) | 48.2 | 53.4 | 19.0 | 51.6 | 67.7 | 14.6 | 27.3 | 4.7 |
| **AlphaChimp** | **56.3** | **60.0** | 21.6 | **65.6** | **75.2** | 14.2 | **25.1** | **0.5** |

facilitating fair comparisons. We set $\lambda_{\text{L1}} = 5.0$, $\lambda_{\text{GIoU}} = 2.0$ and $\lambda_{\text{FL}} = 2.0$ to balance between box regression and behavior classification. We also set $\gamma = 2$ for focal loss in behavior classification. Our model is optimized using Adam with a learning rate of $1e-4$ for 20K iterations. We train our model on the `ChimpACT` (Ma et al., 2023) training set with a batch size of 64 end-to-end and report the results on its test set. 8 HGX-A800-SXM GPUs are used in the training.

## 5.2 Comparison with SOTAs

To the best of our knowledge, there is currently no existing baseline that jointly performs tracking and spatiotemporal action detection for chimpanzees. Therefore, we design our evaluation by separating the task into two tracks: (1) detection and tracking, and (2) spatiotemporal action detection. For both tracks, we benchmark our method against a combination of classic human-centric algorithms adapted to chimpanzee data, as well as recent SOTA methods. Notably, for action detection, we include VideoPrism (Zhao et al., 2024), which is a foundation model with extensive pretraining.

### 5.2.1 Detection and tracking

**Setting.** We evaluate several prominent tracking algorithms on `ChimpACT`, such as SORT (Bewley et al., 2016), DeepSORT (Wojke et al., 2017), and OC-SORT (Cao et al., 2023). All implementations are based on the MMTracking (MMTracking, 2020) codebase. For the detector backbone, we follow the default configurations in MMTracking (MMTracking, 2020) and adopt YOLOX (Ge et al., 2021) as the detection backbone for SORT, DeepSORT, ByteTrack, and OC-SORT, while Faster R-CNN is used for QDTrack as required by its official implementation. For DeepSORT, we additionally adopt a ResNet-50-based (He et al., 2016) ReID network following the default MMTracking setting. We additionally benchmarked against YOLOv8 (Jocher et al., 2023) and YOLO11 (Jocher et al., 2024), more powerful YOLO-based detection backbones. 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 to keep a fair comparison. We strive to ensure a fair comparison, but we adopt certain model-specific default settings to guarantee the convergence of each method, which we acknowledge as a potential limitation.

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 reported on the test set. We report the common metrics (Milan et al., 2016) used in MOT task, including mean Average Precision (mAP) (Lin et al., 2014) for detection accuracy, as well as a range of tracking-specific metrics. These include the CLEAR metrics (Bernardin and Stiefelhagen, 2008), that is, Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), False Positives (FP), False Negatives

(FN), Identity Switch (IDs). MOTA summarizes overall tracking performance by accounting for FP, FN, and IDs, and MOTP measures the localization precision of predicted bounding boxes. Additionally, we report the Identity F1 Score (IDF1) (Ristani et al., 2016) and Higher Order Tracking Accuracy (HOTA) (Luiten et al., 2021) to evaluate various facets of the tracking performance. IDF1 evaluates the consistency of identity assignments over time, and HOTA jointly reflects detection, association, and localization quality in a single unified score. Note that for FP, FN, and IDs, we report normalized values and denote these metrics as nFP, nFN, and nIDs, respectively. Note that our framework detects all chimpanzees in each frame, and adopts the same association strategy as proposed in ByteTrack (Zhang et al., 2022).

**Results.** As shown in Tab. 2, our method achieves SOTA performance on this chimpanzee tracking benchmark, and outperforms prior approaches with approximately 10% higher HOTA scores, indicating stronger overall tracking quality across both detection and association components. We also observe a 16% accuracy improvement in MOTA, reflecting fewer missed detections and identity switches, and a 13% accuracy gain in IDF1, demonstrating significantly improved identity consistency across frames. MOTP remains comparable across methods, suggesting that localization precision is already saturated, while our high mAP confirms accurate per-frame detection. These consistent improvements across multiple metrics highlight the strength of our joint detection and tracking framework in modeling complex multi-agent interactions in chimpanzee videos.

### 5.2.2 Spatiotemporal action detection

**Setting.** We benchmark several representative human action detection baselines on `ChimpACT` using the MMAction2 (MMAction2, 2020) codebase. The evaluated methods include ARCN (Sun et al., 2018), LFB (Wu et al., 2019), and SlowFast (Feichtenhofer et al., 2019). These methods represent a range of approaches in spatiotemporal action detection, from recurrent networks to long-term feature banks and multi-pathway architectures. All models undergo training for 20 epochs with a batch size of 32 and are observed to converge, maintaining consistent optimizers and learning rates as in their official implementations. These methods are provided with either GT (*with GT box*) or detected (*with Det. box*) bounding boxes for each chimpanzee during both training and testing.

In addition, we report results from recent SOTA methods (Yu et al., 2022; Wang et al., 2022; Li et al., 2023; Zhao et al., 2024), some of which utilize foundation models. For instance, VideoPrism (Zhao et al., 2024) represents a state-of-the-art approach in video understanding. The reported numbers for these methods are sourced from VideoPrism (Zhao et al., 2024), which adheres to the same training and evaluation protocols as our benchmark, and all use GT bounding boxes.

We maintain consistency with previous tracks by adopting the same train-test split. We report the overall mean Average Precision (mAP) (Lin et al., 2014) across 23 behavior categories defined in `ChimpACT` (Ma et al., 2023), as well as the category-wise mAP for three key behavior groups: locomotion ($\text{mAP}_L$), object interaction ($\text{mAP}_O$), and social interaction ($\text{mAP}_S$). Since our model is designed to jointly predict bounding boxes and behavior categories in an end-to-end manner, it does not rely on GT boxes. For multi-stage methods, we re-implement open-source baselines such as SlowFast (Feichtenhofer et al., 2019) using detection results from our trained model to ensure fair comparison.

**Results.** Tab. 3 compares AlphaChimp's action detection performance with existing SOTAs. The bottom block of Tab. 3 (*with Det. box*) reveals AlphaChimp's substantial improvement over existing algorithms, achieving a 20% increase in overall mAP. This advancement is particularly evident in the challenging "social" category ($\text{mAP}_S$), where AlphaChimp shows a 20% enhancement. These improvements stem primarily from the advantages of end-to-end joint modeling of detection, tracking, and behavior recognition, which enables holistic optimization across tasks.

Tab. 3 compares AlphaChimp's action detection performance with existing SOTAs. The bottom block of Tab. 3 (*with Det. box*) demonstrates AlphaChimp's substantial improvement over prior methods, achieving a 20% increase in overall mAP. This advancement is particularly pronounced in

**Table 3**: **Results of spatiotemporal action detection on `ChimpACT` test set.** The row highlighted in light blue is the performance reference on the human action dataset AVA (Gu et al., 2018). "*with GT box*" and "*with Det. box*" mean using GT bounding boxes or detected boxes, respectively. "*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," "$\text{mAP}_L$," "$\text{mAP}_O$," and "$\text{mAP}_S$ represent the overall mAP and mAP for Locomotion, Object interaction, and Social interaction. ⊘ denotes not applicable. – denotes unreported.

| Method | Module | mAP | $\text{mAP}_L$ | $\text{mAP}_O$ | $\text{mAP}_S$ |
|---|---|---|---|---|---|
| SlowFast (Feichtenhofer et al., 2019) | | 25.8 | ⊘ | ⊘ | ⊘ |
| VideoPrism-B (Zhao et al., 2024) | | 30.6 | ⊘ | ⊘ | ⊘ |
| VideoPrism-g (Zhao et al., 2024) | | 36.2 | ⊘ | ⊘ | ⊘ |
| *with GT box* | | | | | |
| ACRN (Sun et al., 2018) | | 24.4 | 58.7 | 33.8 | 14.7 |
| LFB (Wu et al., 2019) | *w.* NL LFB | 22.0 | 50.1 | 32.3 | 13.5 |
| | *w.* Max LFB | 23.2 | 45.0 | 31.2 | 17.7 |
| | *w.* Avg LFB | 21.3 | 45.0 | 29.8 | 14.7 |
| SlowOnly (Feichtenhofer et al., 2019) | | 20.9 | 48.1 | 36.2 | 11.5 |
| | *w.* Ctx | 22.3 | 52.3 | 31.2 | 13.8 |
| SlowFast (Feichtenhofer et al., 2019) | | 21.9 | 53.0 | 30.6 | 12.9 |
| | *w.* Ctx | 24.3 | 56.8 | 31.5 | 15.6 |
| CoCa-B (Yu et al., 2022) | | 12.6 | – | – | – |
| InternVideo-B (Wang et al., 2022) | | 24.0 | – | – | – |
| InternVideo-L (Wang et al., 2022) | | 25.7 | – | – | – |
| UMT-B (Li et al., 2023) | | 25.0 | – | – | – |
| UMT-L (Li et al., 2023) | | 24.7 | – | – | – |
| VideoPrism-B (Zhao et al., 2024) | | 28.8 | – | – | – |
| VideoPrism-g (Zhao et al., 2024) | | 31.5 | – | – | – |
| *with Det. box* | | | | | |
| ACRN (Sun et al., 2018) | | 13.4 | 26.8 | 14.4 | 7.1 |
| SlowOnly (Feichtenhofer et al., 2019) | | 11.8 | 25.8 | 13.1 | 5.2 |
| | *w.* Ctx | 13.9 | 27.4 | 14.4 | 7.7 |
| SlowFast (Feichtenhofer et al., 2019) | | 13.5 | 27.2 | 13.7 | 7.3 |
| | *w.* Ctx | 16.2 | 27.5 | 14.3 | 11.9 |
| **AlphaChimp** | | **34.3** | **50.3** | **31.3** | **29.3** |

the challenging "social" category ($\text{mAP}_S$), where AlphaChimp attains a 20% improvement. These gains arise from our architecture's ability to effectively fuse temporal features via Transformer-based self-attention and to jointly optimize detection, tracking, and behavior recognition in an end-to-end manner, which helps mitigate error propagation inherent to multi-stage pipelines.

Notably, our method even surpasses recent foundation models such as VideoPrism (Zhao et al., 2024), which are also fine-tuned on the `ChimpACT` (Ma et al., 2023) dataset. As VideoPrism does not support end-to-end simultaneous detection and fine-grained action recognition, they rely on GT chimpanzee bounding boxes during inference. Despite this advantage, our approach achieves superior performance, highlighting the effectiveness of our architecture for modeling socially interactive animals such as chimpanzees, beyond what can be achieved through fine-tuning general-purpose video models alone.

We observe in Tab. 3 that the performance gap between the GT-box and detected-box settings indicates that detection quality plays a crucial role in downstream behavior recognition. This is because failures in detection inevitably lead to missing action predictions, causing error accumulation across the pipeline. To mitigate this issue, our proposed multi-task framework is designed to avoid treating detection merely as a pre-processing step; instead, it enables detection and behavior recognition to be jointly optimized in an end-to-end manner.

## 5.3 Ablation study

**Query Number $Q$.** Fig. 5 illustrates the impact

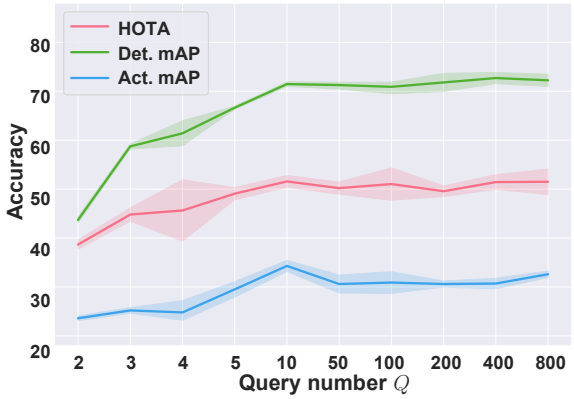


**Fig. 5: The effect of different query number $Q$ on tracking and spatiotemporal action detection.** We plot the results on HOTA, detection mAP (Det. mAP), and the overall action detection mAP (Act. mAP). Performance improves consistently up to $Q = 10$, where it stabilizes, providing a balance between efficiency and stability.

**Table 4: Ablative results of varying the frame length $T$ of the video input on the `ChimpACT` test set for behavior classification.**

| $T$ | mAP | mAP$_L$ | mAP$_O$ | mAP$_S$ |
|---|---|---|---|---|
| 4 | 32.3 | 44.5 | 31.8 | 28.0 |
| 8 | **34.3** | **50.3** | 31.3 | **29.3** |
| 16 | 34.0 | 49.1 | **36.5** | 27.8 |

of different query numbers $Q$ in query selection. We plot the mean and variance for HOTA, detection mAP (Det. mAP), and action detection mAP (Act. mAP). Based on `ChimpACT` statistics, there is an average of 3 chimpanzees per frame, with a maximum of 9. As expected, we observe rapid improvement in model performance as $Q$ increases up to 10, after which performance stabilizes. To balance computation and model stability, we choose $Q = 10$ for our experiments. This value results in low variance with stable and superior performance across all metrics. It effectively accommodates the typical range of chimpanzees in scenes while providing headroom for more crowded scenarios.

**Video input frames $T$.** Tab. 4 presents the ablation results for varying frame length $T$ of the input video. Using only 4 video frames yields lower performance compared to 8 frames, as a larger temporal window provides more temporal information. However, increasing the number of input frames beyond 8 leads to a slight decline in performance. This may be due to additional frames introducing more complex behavioral changes, making accurate estimation more challenging. The results suggest that 8 frames strike an optimal balance between temporal context and performance.

To further understand this behavior, we also computed the action recognition mAP on the training set for different temporal window sizes $T$, which are 0.6434, 0.6455, and 0.6966 for $T = 4, 8$, and 16, respectively. We conjecture that the performance at $T = 16$ may start to overfit on the current training set due to the limited data scale. Nevertheless, we anticipate that larger temporal windows could become more beneficial as more data become available in the future. However, collecting large-scale chimpanzee video data remains challenging, as discussed in Sec. 3, and we hope this will motivate continued collective efforts from the broader research community. Scaling up the dataset and systematically revisiting the choice of $T$ is left for future work.

**Table 5: Ablation on training paradigm for spatiotemporal action detection on ChimpACT test set.**

| Paradigm | mAP | mAP$_L$ | mAP$_O$ | mAP$_S$ |
|---|---|---|---|---|
| Two-stage | 20.6 | 36.2 | 30.7 | 12.2 |
| **Ours** | **34.3** | **50.3** | **31.3** | **29.3** |

**Training paradigm.** We conduct a controlled two-stage training experiment (reported in Tab. 5). Specifically, we first trained the detection branch, then froze the detection branch and trained only the action prediction head, while keeping all other training settings identical to our end-to-end model. We emphasize that this two-stage setting serves only as a proof-of-concept comparison rather than a strictly ablation pipeline, because our detection and action branches share a common backbone and thus cannot be fully decoupled in practice. To prevent changes in detection outputs, the only way is to freeze the backbone as well, which inevitably limits the backbone's ability to adapt its feature representations toward action-relevant cues during feature fusion. Consequently, we observe that this

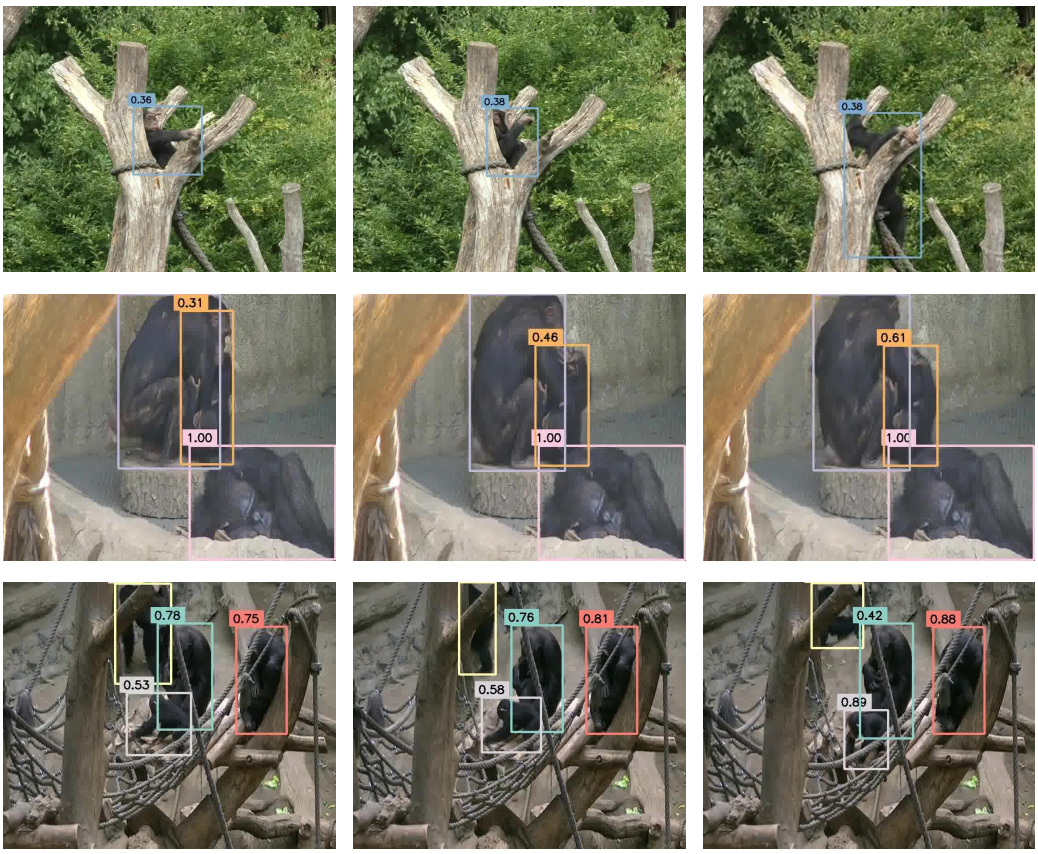


**Fig. 6**: **Visualization of AlphaChimp's detection and tracking results in three different scenarios from `ChimpACT` test set.** Consistent colored boxes indicate successful tracking of the same chimpanzee across frames. The numbers represent the confidence scores of chimpanzee classification.

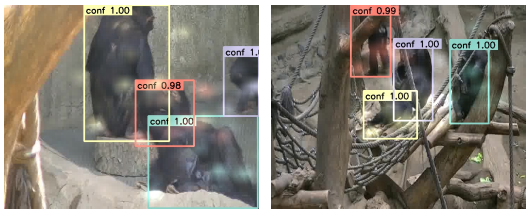


**Fig. 7**: **Visualization of the reference points in the deformable attention module, with a blurring effect applied based on the attention weights.** We visualize the reference points for each query (represented by each box in the image) using the same color as the box's outline. Each reference point is blurred proportionally to its attention weight, with brighter points indicating greater significance.

two-stage pipeline performs worse than end-to-end training, supporting the hypothesis that fixing detection outputs can lead to error accumulation and propagation to action prediction. This also explains the performance gap in behavior recognition when using GT versus detected bounding boxes. In contrast, our end-to-end approach enables joint optimization of detection and action prediction, which helps mitigate such error propagation.

## 5.4 Qualitative results

Fig. 6 visualizes our model's detection and tracking results. Each row represents a sequence from a `ChimpACT` test set video clip. The numbers above each bounding box indicate categorization probability scores, while consistently colored boxes denote successful tracking of individual chimpanzees across frames. Our model demonstrates robustness in managing occlusions, as evidenced in the second row where an infant chimpanzee


**Fig. 8**: **Visualization of AlphaChimp's tracking and behavior detection results in different scenarios from `ChimpACT` test set.** Consistent colored boxes indicate successful tracking of the same chimpanzee across frames. The numbers represent the confidence scores of chimpanzee classification.

(orange box) is momentarily blocked by an adult, and in the third row where a chimpanzee (yellow box) is largely obscured by a tree. Despite these challenges, our method maintains accurate detection and consistent tracking.

Fig. 7 illustrates the reference points within the deformable attention module, with each point blurred according to its attention weights for enhanced visualization. Notably, these reference points predominantly focus on keypoint areas of each chimpanzee. This observation suggests that joint regions may exhibit distinct patterns and unique features crucial for chimpanzee identification and differentiation from other objects.

In Fig. 8, we present three examples of AlphaChimp's tracking and behavior recognition results on the `ChimpACT` test set. The method simultaneously predicts detection boxes around chimpanzees, maintains consistent tracking (indicated by boxes of the same color across frames), and recognizes behaviors. The number above each box represents the categorization probability score. These examples highlight AlphaChimp's ability to handle the significant challenges presented by `ChimpACT`, including accurate detection of occluded chimpanzees and precise behavior determination. A notable instance is in the first row, where the method successfully identifies a young chimpanzee heavily obscured by adults.

We further visualize the gradient $\frac{\partial \mathbf{a}}{\partial \mathbf{I}}$ of our behavior prediction probabilities in Fig. 10. This visualization highlights that, in predicting behaviors, our model focuses not only on the individual chimpanzee but also on other interacting entities within the scene. This broader attention is crucial for accurately recognizing social behaviors that involve multiple participants. For instance, in the first column, where a young chimpanzee is seen 'playing' with an adult, our AlphaChimp effectively highlights the relevant body parts of the adult chimpanzee. This indicates that our model comprehensively considers the dynamics

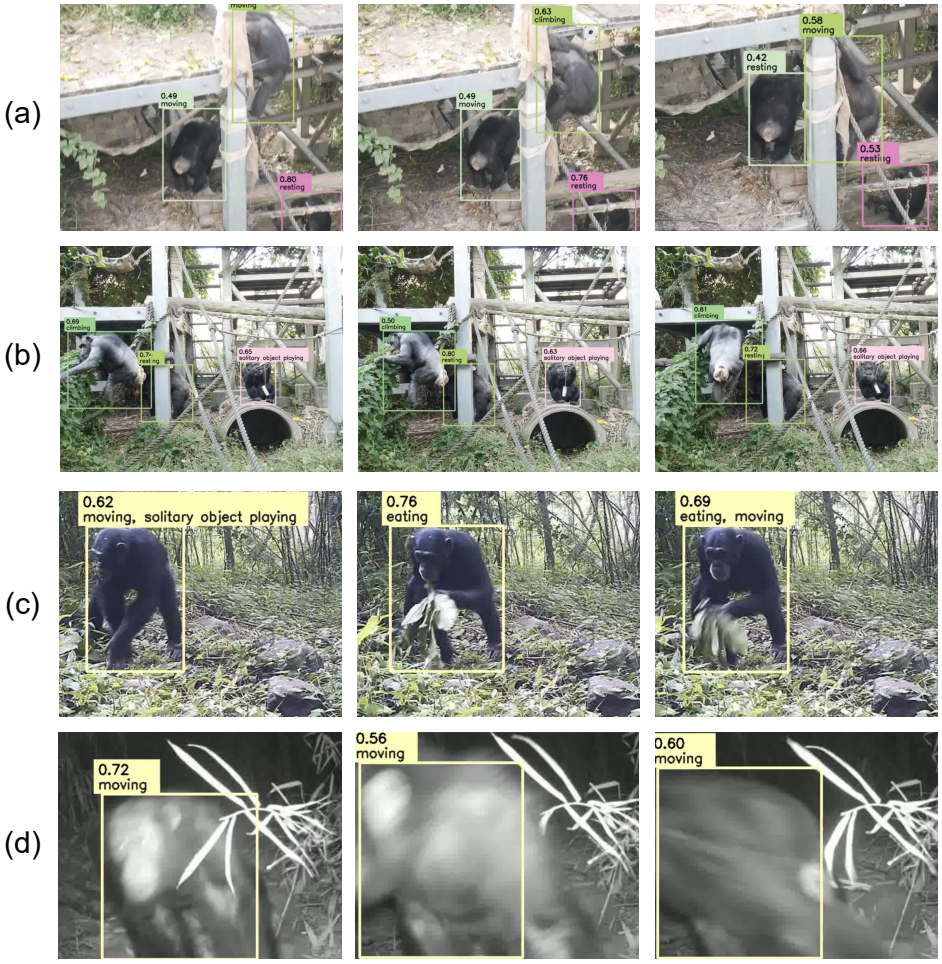


**Fig. 9**: **Qualitative results on (a-b) unseen Internet video data and (c-d) the PanAf-FGBG (Brookes et al., 2025) dataset.** We directly apply our model for inference on this data.


**Fig. 10**: **Visualization of the behavior recognition probability gradient.** We visualize $\frac{\partial \mathbf{a}}{\partial \mathbf{I}}$, where $\mathbf{a}$ represents the behavior recognition probability for the specific chimpanzee boxed in the image $\mathbf{I}$.

between the individuals, allowing for a nuanced understanding of social interactions within the group.


**Fig. 11**: **Typical failure case of AlphaChimp.** There is a missing detection of the occluded chimpanzee.

In addition, Fig. 9 presents qualitative results of our model on (a-b) unseen Internet videos and the (c-d) PanAf-FGBG (Brookes et al., 2025) dataset. Although our model was not trained on either source, it can robustly detect, track, and

recognize chimpanzee behaviors across a variety of challenging conditions. The examples illustrate scenarios with different camera types, viewpoints, lighting conditions (e.g., nighttime in (d)), and (c) wild scenarios, as well as diverse occlusion patterns and background complexities. Despite these variations, our model produces coherent detections and behavior predictions, indicating reasonable generalization capability to out-of-distribution data.

These qualitative results underscore AlphaChimp's effectiveness in handling complex scenarios involving multiple chimpanzees, occlusions, and diverse behaviors. The model's ability to maintain consistent tracking and accurate behavior recognition, even in challenging conditions, demonstrates its potential for advancing automated chimpanzee behavior analysis in real-world settings.

Fig. 11 illustrates one typical failure case encountered by our model, where a missed behavior estimation due to a detection error occurs when the chimpanzee is occluded. This failure case highlights the inherent challenges in chimpanzee perception and behavior analysis. Factors such as the intrinsic appearance similarity among chimpanzees, frequent occlusions in their natural environment, and the complexity of social behaviors involving multiple individuals contribute to these difficulties. These examples underscore the need for continued refinement of our model to better handle such challenging scenarios in chimpanzee behavior analysis.

Please refer to Sec. A2 and supplementary video for more qualitative results.

# 6 Conclusion

This work introduces AlphaChimp, an end-to-end approach that detects, tracks, and recognizes chimpanzee fine-grained behaviors in a unified framework. This framework enhances tracking and behavior recognition by integrating temporal context and spatial relationships. Experiments show that it outperforms existing methods across tasks, highlighting its capabilities.

Despite these advances, several **limitations** remain. First, our framework does not incorporate pose estimation, which may provide richer kinematic cues for fine-grained behavior understanding. Second, while our model recognizes individual behaviors effectively, it does not explicitly model behavioral transitions, which we identify as an important direction for future work. Third, precise detection and tracking under heavy occlusion or ambiguity remain challenging, reflecting difficulties that even human observers face in densely interacting chimpanzee groups.

From a data perspective, our `ChimpACT` dataset represents a unique longitudinal, in-the-wild observation of the same chimpanzee social group over four years, enabling the study of development and social dynamics. However, focal sampling may introduce biases, and manual data collection is costly. Future efforts could explore complementary, non-intrusive data acquisition methods such as camera traps with lightweight on-device preprocessing to support scalable and sustainable data collection.

More broadly, this work takes an initial step toward automated understanding of non-human primate behavior. We hope that bridging primatology and computer vision will inspire the development of primate-specific perception and behavior models, ultimately contributing to deeper insights into chimpanzee social dynamics, animal welfare, and comparative studies of social intelligence.

**Acknowledgements.** The authors would like to thank the Wolfgang Köhler Primate Research Center for assisting in data collection, BasicFinder CO., Ltd. and Keyue Zhang for annotations and quality check, Jiajun Su, Wentao Zhu, and Zihao Yin for discussions and preliminary experiments, 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, Y. Xu, and Y. Wang are supported in part by the National Natural Science Foundation of China (6247070125). Y. Lin and Y. Zhu are supported in part by the National Natural Science Foundation of China (32595491, 62376009), the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

**Data Availability Statement.** The dataset used in this work is `ChimpACT` (Ma et al., 2023), which has been made publicly available on our project website.

**Ethical Approval Statement.** The `ChimpACT` dataset raises no ethical concerns regarding the privacy information of human subjects, as it solely focuses on chimpanzees. Studying the social behavior of chimpanzees provides an ethical and efficient means to explore aspects of human sociality due to our phylogenetic proximity. By analyzing their behaviors, we can gain insights into the evolution of human social behavior and potentially contribute to both the scientific and ethical understanding of the human condition. The ethics committee of the Wolfgang Köhler Primate Research Center approved the observational data collection for this project.

**Conflict of Interest Statement.** The authors declare that they have no competing interests.

**Funding Declaration Statement.** The funding agencies had no role in study design, data collection, analysis, or manuscript preparation.

**Consent for Publication Statement.** Not applicable. The manuscript does not contain any individual person's data.

**Author Contribution Statement.** Xiaoxuan Ma led the project, including project conception, method design, dataset collection and processing, experiments, and manuscript writing. Yutang Lin and Yuan Xu contributed to method design, experiments, and manuscript writing. Stephan P. Kaufhold, Jack Terwilliger, and Andres Meza contributed to dataset collection and processing. Federico Rossano contributed to the original data collection. Yixin Zhu, Federico Rossano, and Yizhou Wang provided overall project supervision, research guidance, and manuscript revision.

**Consent to Participate Statement.** Not applicable. The study does not involve human participants. All data were collected from non-human primates under approved observational protocols.

# References

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

Bain M, Nagrani A, Schofield D, et al (2019) Count, crop and recognise: Fine-grained recognition in the wild. In: Proceedings of International Conference on Computer Vision Workshops (ICCV Workshops)

Bain M, Nagrani A, Schofield D, et al (2021) Automated audiovisual behavior recognition in wild primates. In: Science Advances, vol 7. American Association for the Advancement of Science, p eabi4883

Baker TA (2022) Wolfgang köhler primate research center. In: Encyclopedia of Animal Cognition and Behavior, p 7310

Bala PC, Eisenreich BR, Yoo SBM, et al (2020) Automated markerless pose estimation in freely moving macaques with openmonkeystudio. In: Nature Communications, p 4560

Bernardin K, Stiefelhagen R (2008) Evaluating multiple object tracking performance: the clear mot metrics. In: EURASIP booktitle on Image and Video Processing, vol 2008. Springer, pp 1–10

Bewley A, Ge Z, Ott L, et al (2016) Simple online and realtime tracking. In: IEEE International Conference on Image Processing (ICIP)

Brookes O, Mirmehdi M, Kuhl H, et al (2024a) Chimpvlm: Ethogram-enhanced chimpanzee behaviour recognition. In: Proceedings

of Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)

Brookes O, Mirmehdi M, Stephens C, et al (2024b) Panaf20k: A large video dataset for wild ape detection & behaviour analysis. In: International Journal of Computer Vision (IJCV)

Brookes O, Kukushkin M, Mirmehdi M, et al (2025) The panaf-fgbg dataset: Understanding the impact of backgrounds in wildlife behaviour recognition. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), pp 5433–5443

Cao J, Pang J, Weng X, et al (2023) Observation-centric sort: Rethinking sort for robust multi-object tracking. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)

Carion N, Massa F, Synnaeve G, et al (2020) End-to-end object detection with transformers. In: Proceedings of European Conference on Computer Vision (ECCV)

Dawkins MS (2003) Behaviour as a tool in the assessment of animal welfare. In: Zoology, vol 106. Elsevier, pp 383–387

Desai N, Bala P, Richardson R, et al (2023) Openapepose, a database of annotated ape photographs for pose estimation. In: Elife, p RP86873

Duporge I, Kholiavchenko M, Harel R, et al (2025) Baboonland dataset: Tracking primates in the wild and automating behaviour recognition from drone videos: I. duporge et al. International Journal of Computer Vision (IJCV) pp 1–12

Earth G (2024) Google earth. https://www.google.com/earth/about/

Fabian Caba Heilbron BGVictor Escorcia, Niebles JC (2015) Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)

Feichtenhofer C, Fan H, Malik J, et al (2019) Slowfast networks for video recognition. In: Proceedings of International Conference on Computer Vision (ICCV)

Fröhlich M, Müller G, Zeiträg C, et al (2020) Begging and social tolerance: Food solicitation tactics in young chimpanzees (pan troglodytes) in the wild. In: Evolution and Human Behavior, vol 41. Elsevier, pp 126–135

Fuchs M, Genty E, Bangerter A, et al (2024) From forest to zoo: Great ape behavior recognition with chimpbehave. In: Proceedings of Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)

Ge Z, Liu S, Wang F, et al (2021) Yolox: Exceeding yolo series in 2021. In: arXiv preprint arXiv:2107.08430

Gonyou HW (1994) Why the study of animal behavior is associated with the animal welfare issue. In: booktitle of Animal Science, vol 72. Oxford University Press, pp 2171–2177

Gritsenko AA, Xiong X, Djolonga J, et al (2024) End-to-end spatio-temporal action localisation with video transformers. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), pp 18373–18383

Gu C, Sun C, Ross DA, et al (2018) Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)

He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)

He K, Gkioxari G, Dollár P, et al (2017) Mask r-cnn. In: Proceedings of International Conference on Computer Vision (ICCV)

Hobaiter C, Samuni L, Mullins C, et al (2017) Variation in hunting behaviour in neighbouring chimpanzee communities in the budongo forest, uganda. In: PloS One, vol 12. Public Library of Science San Francisco, CA USA, p e0178065

Jocher G, Qiu J, Chaurasia A (2023) Ultralytics YOLO. URL https://github.com/ultralytics/ultralytics

Jocher G, Qiu J, Chaurasia A (2024) Ultralytics YOLO. URL https://github.com/ultralytics/ultralytics

Kay W, Carreira J, Simonyan K, et al (2017) The kinetics human action video dataset. In: arXiv preprint arXiv:1705.06950

Kholiavchenko M, Kline J, Ramirez M, et al (2024) Kabr: In-situ dataset for kenyan animal behavior recognition from drone videos. In: Proceedings of Winter Conference on Applications of Computer Vision Workshops (WACV Workshops)

Labuguen R, Matsumoto J, Negrete SB, et al (2021) Macaquepose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture. In: Frontiers in Behavioral Neuroscience, vol 14. Frontiers Media SA, p 581154

Langergraber KE, Prüfer K, Rowney C, et al (2012) Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. In: Proceedings of the National Academy of Sciences (PNAS), vol 109. National Acad Sciences, pp 15716–15721

Lehmann F (2018) Wikimedia commons. https://commons.wikimedia.org/wiki/File:Zoo_Leipzig_April_2018_(41).jpg

Li K, Wang Y, Li Y, et al (2023) Unmasked teacher: Towards training-efficient video foundation models. In: Proceedings of International Conference on Computer Vision (ICCV)

Lin TY, Maire M, Belongie S, et al (2014) Microsoft coco: Common objects in context. In: Proceedings of European Conference on Computer Vision (ECCV)

Lin TY, Goyal P, Girshick R, et al (2017) Focal loss for dense object detection. In: Proceedings of International Conference on Computer Vision (ICCV)

Liu Z, Ning J, Cao Y, et al (2022) Video swin transformer. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)

Luiten J, Osep A, Dendorfer P, et al (2021) Hota: A higher order metric for evaluating multi-object tracking. In: International Journal of Computer Vision (IJCV), vol 129. Springer, pp 548–578

Luncz LV, Sirianni G, Mundry R, et al (2018) Costly culture: differences in nut-cracking efficiency between wild chimpanzee groups. In: Animal Behaviour, vol 137. Elsevier, pp 63–73

Ma X, Kaufhold S, Su J, et al (2023) Chimpact: A longitudinal dataset for understanding chimpanzee behaviors. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Marks M, Jin Q, Sturman O, et al (2022) Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. In: Nature Machine Intelligence, vol 4. Nature Publishing Group UK London, pp 331–340

McEwen ES, Warren E, Tenpas S, et al (2022) Primate cognition in zoos: Reviewing the impact of zoo-based research over 15 years. In: American booktitle of Primatology, vol 84. Wiley Online Library, p e23369

Milan A, Leal-Taixé L, Reid I, et al (2016) Mot16: A benchmark for multi-object tracking. In: arXiv preprint arXiv:1603.00831

MMAction2 (2020) Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2

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

Ng XL, Ong KE, Zheng Q, et al (2022) Animal kingdom: A large and diverse dataset for animal behavior understanding. In: Proceedings of Conference on Computer Vision and Pattern

Recognition (CVPR)

Pang J, Qiu L, Li X, et al (2021) Quasi-dense similarity learning for multiple object tracking. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)

Pedersen M, Haurum JB, Bengtson SH, et al (2020) 3d-zef: A 3d zebrafish tracking benchmark dataset. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)

Rezatofighi H, Tsoi N, Gwak J, et al (2019) Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)

Ristani E, Solera F, Zou R, et al (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of European Conference on Computer Vision Workshops (ECCV Workshops)

Shao S, Li Z, Zhang T, et al (2019) Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of International Conference on Computer Vision (ICCV)

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

Sun C, Shrivastava A, Vondrick C, et al (2018) Actor-centric relation network. In: Proceedings of European Conference on Computer Vision (ECCV)

Sun JJ, Karigo T, Chakraborty D, et al (2021) The multi-agent behavior dataset: Mouse dyadic social interactions. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Sun JJ, Marks M, Ulmer AW, et al (2023) Mabe22: A multi-species multi-task benchmark for learned representations of behavior. In: Proceedings of International Conference on Machine Learning (ICML)

Surbeck M, Boesch C, Girard-Buttoz C, et al (2017) Comparison of male conflict behavior in chimpanzees (pan troglodytes) and bonobos (pan paniscus), with specific regard to coalition and post-conflict behavior. In: American booktitle of Primatology, vol 79. Wiley Online Library, p e22641

The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. In: Nature, pp 69–87

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Wang Y, Li K, Li Y, et al (2022) Internvideo: General video foundation models via generative and discriminative learning. In: arXiv preprint arXiv:2212.03191

Wiltshire C, Lewis-Cheetham J, Komedová V, et al (2023) Deepwild: Application of the pose estimation tool deeplabcut for behaviour tracking in wild chimpanzees and bonobos. In: booktitle of Animal Ecology. Wiley Online Library

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

Wu CY, Feichtenhofer C, Fan H, et al (2019) Long-term feature banks for detailed video understanding. In: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)

Yang Y, Yang J, Xu Y, et al (2022) Apt-36k: A large-scale benchmark for animal pose estimation and tracking. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Yao Y, Bala P, Mohan A, et al (2023) Openmonkeychallenge: Dataset and benchmark challenges for pose estimation of non-human primates. In: International Journal of Computer Vision (IJCV), vol 131. Springer, pp 243–258

Yu H, Xu Y, Zhang J, et al (2021) Ap-10k: A benchmark for animal pose estimation in the wild. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS)

Yu J, Wang Z, Vasudevan V, et al (2022) Coca: Contrastive captioners are image-text foundation models. In: Transactions on Machine Learning Research

Zhang H, Li F, Liu S, et al (2023a) DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: Proceedings of International Conference on Learning Representations (ICLR)

Zhang L, Gao J, Xiao Z, et al (2023b) Animaltrack: A benchmark for multi-animal tracking in the wild. In: International Journal of Computer Vision (IJCV), vol 131. Springer, pp 496–513

Zhang Y, Sun P, Jiang Y, et al (2022) Bytetrack: Multi-object tracking by associating every detection box. In: Proceedings of European Conference on Computer Vision (ECCV)

Zhao L, Gundavarapu NB, Yuan L, et al (2024) Videoprism: A foundational visual encoder for video understanding. In: Proceedings of International Conference on Machine Learning (ICML)

Zhu X, Su W, Lu L, et al (2021) Deformable DETR: Deformable transformers for end-to-end object detection. In: Proceedings of International Conference on Learning Representations (ICLR)

# A1 Additional method details

## A1.1 Architecture details

Fig. A2 illustrates the detailed architectural configuration of the video backbone. In line with Liu et al. (2022), the 3D patch partition layer obtains $\frac{T}{2} \times \frac{H}{4} \times \frac{W}{4}$ 3D tokens, with each patch (*i.e.* token) consisting of a $C_{in}$-dimensional feature. Subsequently, four successive stages transform these video tokens into multi-resolution features: specifically, $\mathbf{V}_1 \in \mathbb{R}^{\frac{T}{2} \times \frac{H}{4} \times \frac{W}{4} \times M}$, $\mathbf{V}_2 \in \mathbb{R}^{\frac{T}{2} \times \frac{H}{8} \times \frac{W}{8} \times 2M}$, $\mathbf{V}_3 \in \mathbb{R}^{\frac{T}{2} \times \frac{H}{16} \times \frac{W}{16} \times 4M}$, and $\mathbf{V}_4 \in \mathbb{R}^{\frac{T}{2} \times \frac{H}{32} \times \frac{W}{32} \times 8M}$.

The feature fusion module then transforms these multi-scale temporal features into $\mathbf{F}_1 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$, ..., $\mathbf{F}_4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C}$. This module operates in two key steps: first, temporal merging compresses the features along the temporal dimension using 3D convolutional layers. Next, a channel mapping layer adjusts the feature channels to ensure uniformity, standardizing them to $C$ using 2D convolutional layers.

Before inputting the multi-scale features $\mathbf{F}$ into the Transformer encoder-decoder, we flatten and concatenate them along the spatial dimension. This process results in a feature input of size $(\frac{H}{4} \times \frac{W}{4} + \frac{H}{8} \times \frac{W}{8} + \frac{H}{16} \times \frac{W}{16} + \frac{H}{32} \times \frac{W}{32}) \times C$.

## A1.2 Implementation details

In practice, we set $C_{in} = 192$, $M = 192$, and $C = 512$. Following the successful pertaining paradigm in foundation models, we first train our model on the Object365 dataset (Shao et al., 2019) for $40K$ iterations. We then fine-tune the model on our `ChimpACT` dataset for $20K$ iterations. We set the thresholds for category and behavior classification at 0.3 and 0.3, respectively.

# A2 Additional experimental results

We report the accuracy of several subcategory behaviors in Tab. A1. It is evident that when using the same detected boxes as input, our method AlphaChimp shows significant improvements over the baselines, especially in social behaviors. For example, previous methods almost completely failed in categories like playing or being nursed, whereas we achieved substantial improvements. Even when compared to baselines using GT boxes as input, our method AlphaChimp still maintains impressive accuracy.

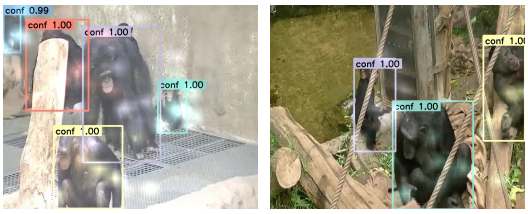


**Fig. A1**: **Additional visualization of the reference points in the deformable attention module, with a blurring effect applied based on the attention weights.** We visualize the reference points for each query (represented by each box in the image) using the same color as the box. Each reference point is blurred proportionally to its attention weight, with brighter points indicating greater significance.

We present additional detection and tracking results by our AlphaChimp in Fig. A3. These examples demonstrate the robustness and effectiveness of our approach, particularly in challenging scenarios. Even when significant occlusion occurs, such as in the third row where a chimpanzee is partially hidden from view, our method successfully detects and tracks the occluded chimpanzee.

Fig. A1 visualizes more examples of the reference points within the deformable attention module, with each point blurred according to its attention weights for clarity. These reference points mainly target keypoint areas on chimpanzees, indicating that joint regions may have unique features essential for distinguishing chimpanzees from other objects.

We present additional qualitative results in Fig. A5, showing detection, tracking and spatiotemporal behavior predictions made by our AlphaChimp on the test set of the `ChimpACT` dataset. These results illustrate the comprehensive capabilities of our model, which not only predicts detection bounding boxes but also performs simultaneous classification of both the class and behaviors in an end-to-end manner. This integrated approach allows AlphaChimp to efficiently process complex visual scenes, identifying individual chimpanzees and their corresponding actions. By effectively combining detection and classification tasks, AlphaChimp provides a holistic view of the observed scenes, making it a powerful tool for analyzing and understanding primate behavior in

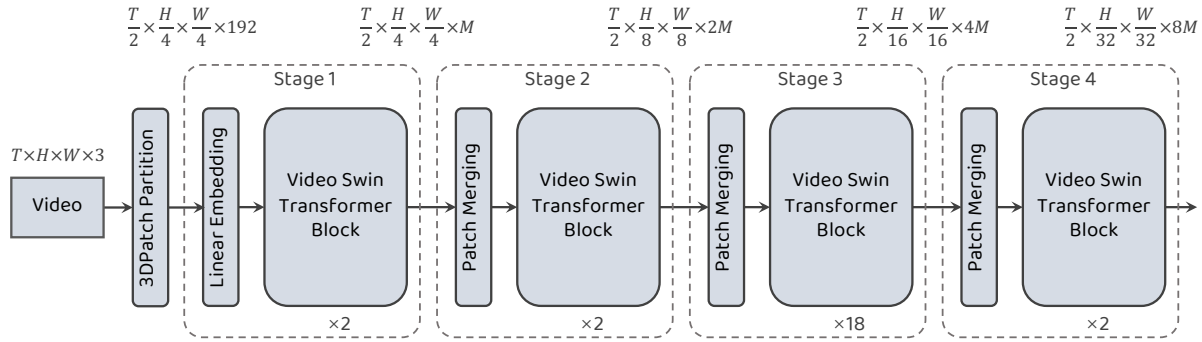


**Fig. A2**: **Detailed architecture of the video backbone in our AlphaChimp**. We follow the Swin-L (Liu et al., 2022) architecture design.

**Table A1**: **Results of spatiotemporal action detection track on `ChimpACT` test set.** "*with GT box*" and "*with Det. box*" mean using GT bounding boxes or detected boxes, respectively.

| Method | mAP | moving | climbing | sol. obj. | playing | eating | grooming | playing | being begged from | aggressing | being nursed |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *with GT box* | | | | | | | | | | | |
| ACRN (Sun et al., 2018) | 24.4 | 60.2 | 23.2 | 38.2 | 54.3 | 7.7 | 42.9 | 0.0 | 0.0 | 4.4 |
| LFB (Wu et al., 2019) | 22.4 | 45.3 | 10.0 | 34.4 | 56.3 | 8.7 | 51.0 | 0.4 | 0.0 | 32.1 |
| SlowOnly (Feichtenhofer et al., 2019) | 24.5 | 56.1 | 31.6 | 41.0 | 45.4 | 10.4 | 43.0 | 0.0 | 0.0 | 7.5 |
| SlowFast (Feichtenhofer et al., 2019) | 24.5 | 60.9 | 37.2 | 47.3 | 35.3 | 10.4 | 49.2 | 0.0 | 0.0 | 7.5 |
| *with Det. box* | | | | | | | | | | |
| SlowOnly (Feichtenhofer et al., 2019) | 11.8 | 13.4 | 3.5 | 19.4 | 19.9 | 0.3 | 9.4 | 0.0 | 0.0 | 0.0 |
| SlowOnly *w.* Ctx (Feichtenhofer et al., 2019) | 13.9 | 16.3 | 6.1 | 16.3 | 19.7 | 1.3 | 12.7 | 0.0 | 0.0 | 0.0 |
| SlowFast (Feichtenhofer et al., 2019) | 13.5 | 18.4 | 3.5 | 19.8 | 18.4 | 0.1 | 5.5 | 0.0 | 0.0 | 0.0 |
| SlowFast *w.* Ctx (Feichtenhofer et al., 2019) | 16.2 | 16.8 | 6.4 | 20.4 | 19.4 | 0.2 | 1.8 | 0.0 | 0.0 | 0.0 |
| **AlphaChimp** (Ours) | 34.3 | 33.3 | 33.4 | 37.1 | 55.0 | 1.9 | 48.9 | 0.3 | 0.0 | 74.0 |

natural settings. Please refer to our project page for more video results.

Fig. A4 illustrates typical failure cases encountered by our model. Panels (a-c) demonstrate tracking failures, while (d) highlights behavior recognition issues. Tracking failures primarily involve detection errors and ID mismatches. In (a), two closely positioned chimpanzees are mistakenly identified as one (yellow box) due to their small appearance in the image. Panel (b) shows an inaccurate bounding box for a young chimpanzee, excluding its left hand, a challenging scenario even for human observers. In (c), an ID change occurs after occlusion, resulting in a bounding box color change upon the chimpanzee's reappearance.

Behavior recognition failures are exemplified in panel (d), which shows an incorrect behavior recognition despite successful detection. Specifically, an adult chimpanzee carrying a young one is not correctly identified, and the young chimpanzee's state of "being carried" is unrecognized in subsequent frames.

These failure cases highlight the inherent challenges in chimpanzee perception and behavior analysis. Factors such as the intrinsic appearance similarity among chimpanzees, frequent occlusions in their natural environment, and the complexity of social behaviors involving multiple individuals contribute to these difficulties. These examples underscore the need for continued refinement of our model to better handle such challenging scenarios in chimpanzee behavior analysis.

In addition, Fig. A6 presents more qualitative results of our model on (a-c) unseen Internet videos and the (d-f) PanAf-FGBG (Brookes et al., 2025) dataset. Although our model was not trained on these data, it is able to reliably detect, track, and recognize chimpanzee behaviors under diverse and challenging conditions. The examples cover different camera types, viewpoints (b), lighting conditions (d & f), and wild environments in (d-f), along with various occlusion patterns and background complexities. Across these settings, our model yields consistent detections

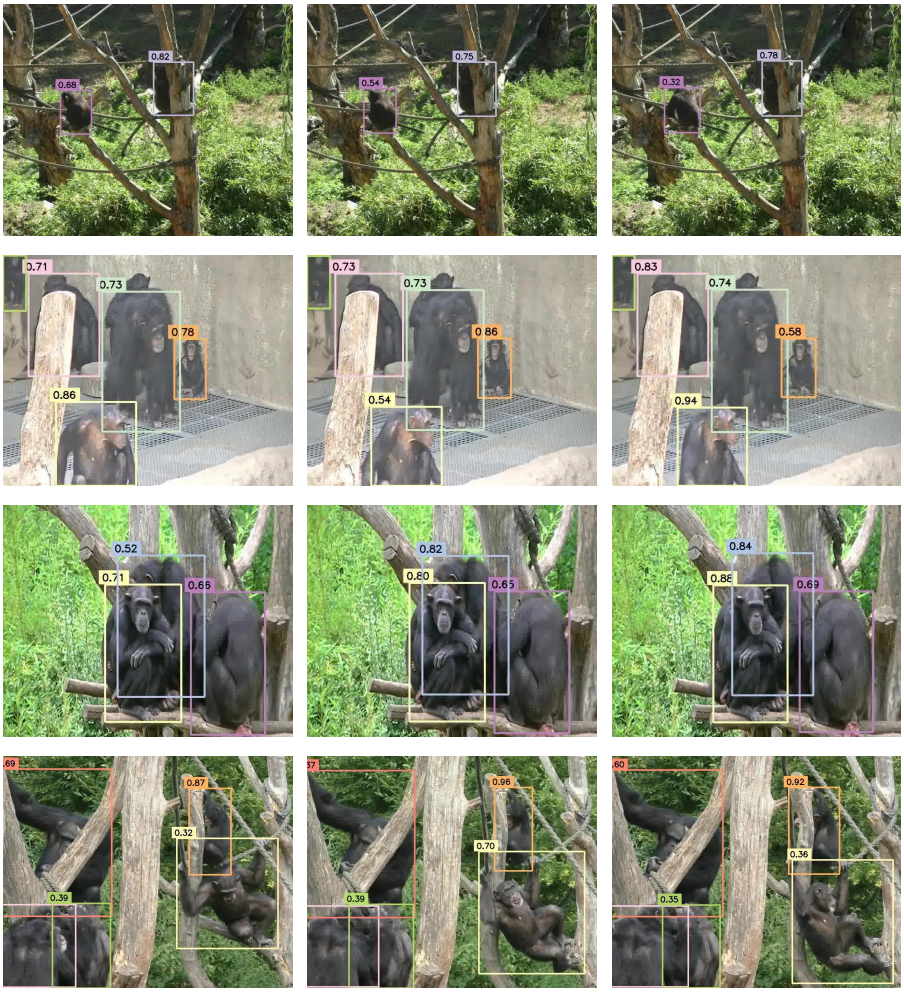

**Fig. A3**: **Additional visualization of AlphaChimp's tracking results in different scenarios from `ChimpACT` test set.** Consistent colored boxes indicate successful tracking of the same chimpanzee across frames. The numbers represent the confidence scores of chimpanzee classification.

and behavior predictions, suggesting reasonable generalization to out-of-distribution data.

Furthermore, we conducted inference on unseen bonobo data (Fig. A7). Bonobos are a species of great apes closely related to chimpanzees in appearance and behavior, and the results show that our model can successfully recognize bonobo behaviors as well. We note that our framework is in principle applicable to other animal species beyond chimpanzees, which we leave for future work.

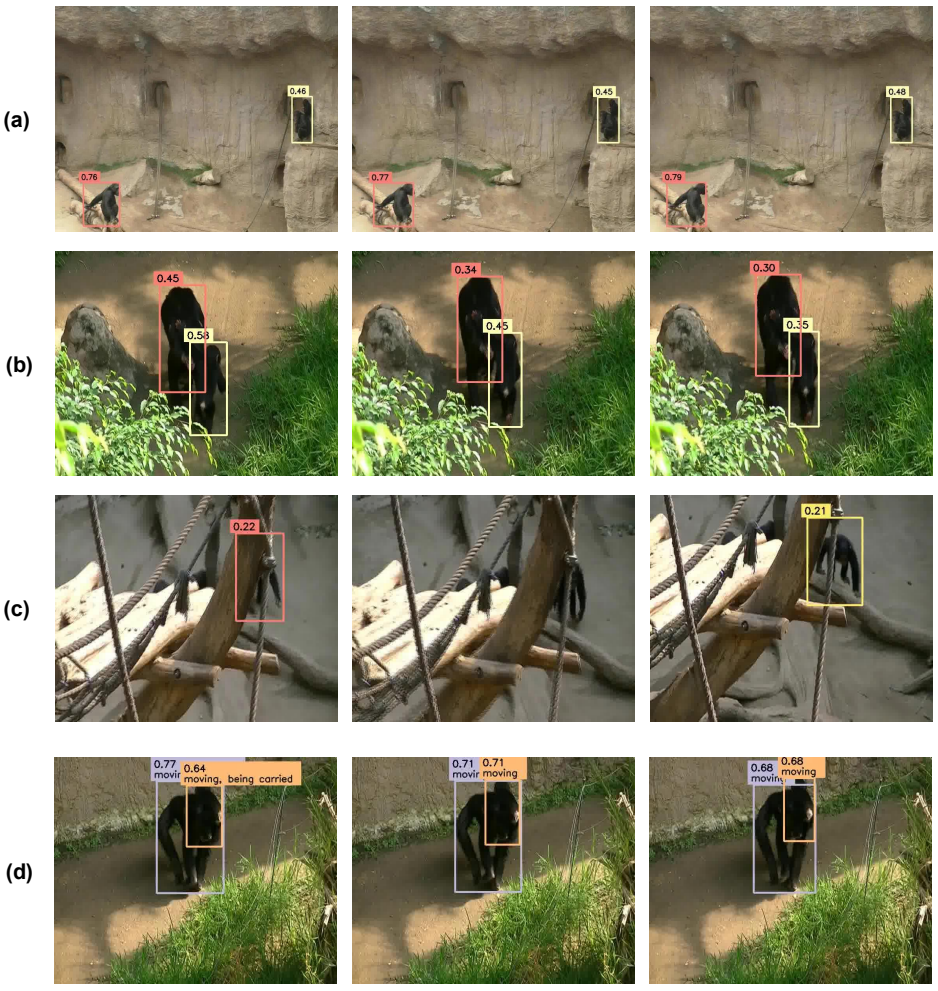


**Fig. A4**: **Typical failure cases of AlphaChimp's on `ChimpACT` test set.** (a-c) show the failure cases in chimpanzee detection and tracking. (d) shows the failure cases in chimpanzee behavior recognition.


**Fig. A5**: **Additional visualization of AlphaChimp's detection, tracking, and behavior detection results in different scenarios from `ChimpACT` test set.** Consistent colored boxes indicate successful tracking of the same chimpanzee across frames. The numbers represent the confidence scores of chimpanzee classification.

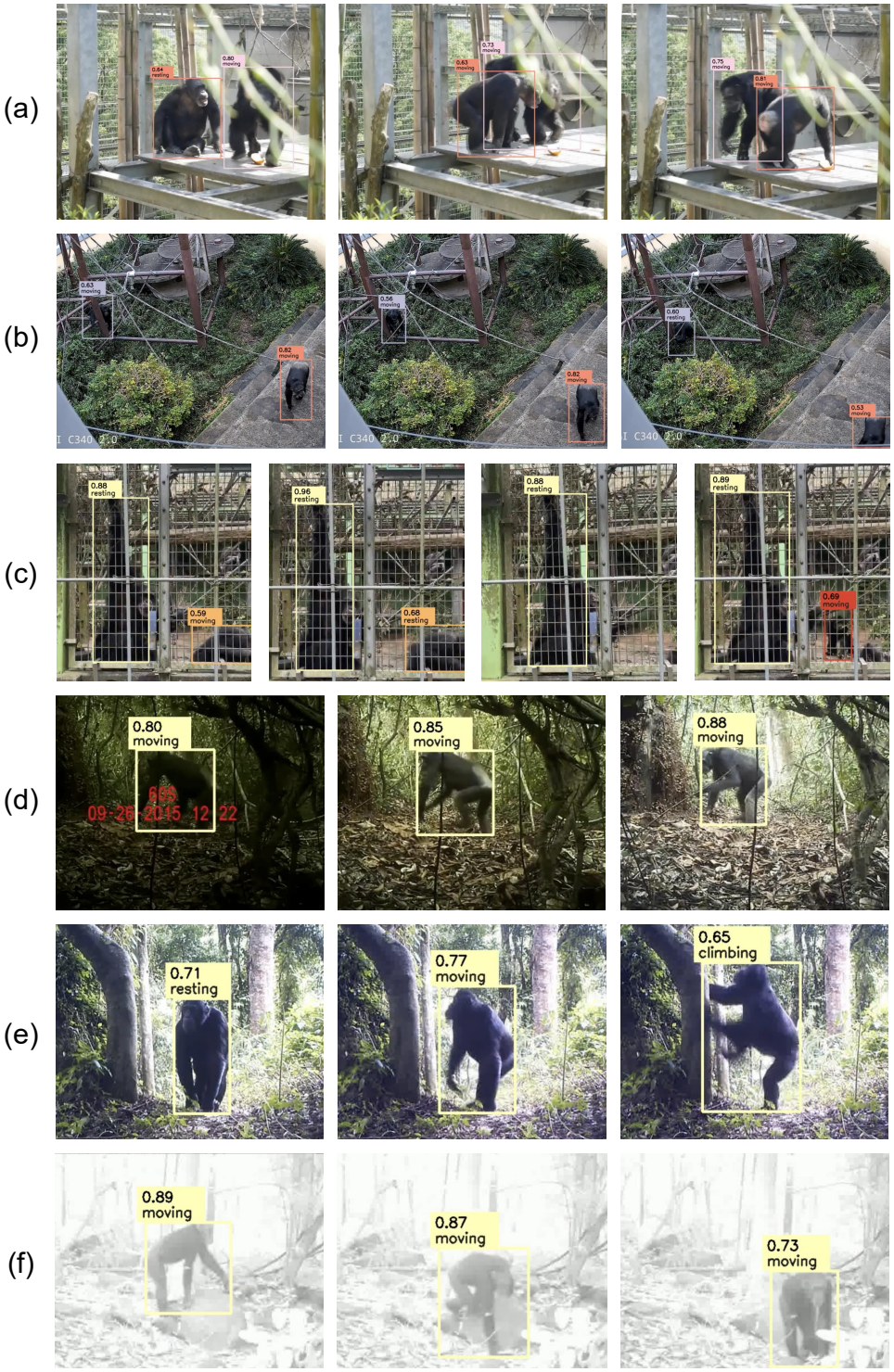


**Fig. A6**: **Additional qualitative results on (a-c) unseen Internet video data and (d-f) the PanAf-FGBG (Brookes et al., 2025) dataset.** We directly apply our model for inference on this data.


**Fig. A7**: **Qualitative results on unseen Internet video data of bonobos.** We directly apply our model trained on chimpanzees for inference on videos captured for bonobos.