Zhai et al. *Zool. Res.* 2025, 46(2): 355−369
https://doi.org/10.24272/j.issn.2095-8137.2024.268
https://cstr.cn/32248.14.zr.2024.268

ZR | Zoological Research

Article Open Access

InteBOMB: Integrating generic object tracking and segmentation with pose estimation for animal behavior analysis

Hao Zhai[1,2,#], Hai-Yang Yan[1,2,#], Jing-Yuan Zhou[3], Jing Liu[1], Qi-Wei Xie[4], Li-Jun Shen[1], Xi Chen[1,*], Hua Han[1,2,*]

[1] *Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology*, *Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*

[2] *School of Future Technology*, *School of Artificial Intelligence*, *University of Chinese Academy of Sciences*, *Beijing* 101408, *China*

[3] *School of Electronic Information and Electrical Engineering*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*

[4] *Research Base of Beijing Modern Manufacturing Development*, *Beijing University of Technology*, *Beijing* 100124, *China*

ABSTRACT

Advancements in animal behavior quantification methods have driven the development of computational ethology, enabling fully automated behavior analysis. Existing multi-animal pose estimation workflows rely on tracking-by-detection frameworks for either bottom-up or top-down approaches, requiring retraining to accommodate diverse animal appearances. This study introduces InteBOMB, an integrated workflow that enhances top-down approaches by incorporating generic object tracking, eliminating the need for prior knowledge of target animals while maintaining broad generalizability. InteBOMB includes two key strategies for tracking and segmentation in laboratory environments and two techniques for pose estimation in natural settings. The “background enhancement” strategy optimizes foreground-background contrastive loss, generating more discriminative correlation maps. The “online proofreading” strategy stores human-in-the-loop long-term memory and dynamic short-term memory, enabling adaptive updates to object visual features. The “automated labeling suggestion” technique reuses the visual features saved during tracking to identify representative frames for training set labeling. Additionally, the “joint behavior analysis” technique integrates these features with multimodal data, expanding the latent space for behavior classification and clustering. To evaluate the framework, six datasets of mice and six datasets of non-human primates were compiled, covering laboratory and natural scenes. Benchmarking results demonstrated a 24% improvement in zero-shot generic tracking and a 21% enhancement in joint latent space performance across datasets, highlighting the effectiveness of this approach in robust, generalizable behavior analysis.

Keywords: Generic object tracking; Pose estimation; Behavior analysis; Background subtraction; Online learning; Selective labeling; Joint latent space

INTRODUCTION

Quantifying and analyzing animal behavior from video recordings is fundamental to research in zoology, neuroscience, ethology, medicine, and computer science (Anderson & Perona, 2014; Luxem et al., 2023). The recent emergence of computational ethology has driven the development of automated methods designed to capture various aspects of animal motion (Pereira et al., 2020). Behavioral descriptors span from coarse representations, such as body centroids, to detailed body part poses (Mathis et al., 2018). Analytical approaches have evolved from classic handcrafted feature extraction to advanced deep neural networks (Romero-Ferrero et al., 2019). Furthermore, the focus has expanded from tracking single animals to multiple individuals interacting within the same scene, often leading to occlusion and contact challenges (Walter & Couzin, 2021). These advancements have imposed increasingly complex demands on multi-task behavioral recognition frameworks, necessitating robust methodologies for animal tracking, segmentation, and pose estimation.

Existing state-of-the-art multi-animal pose estimation methods (Lauer et al., 2022; Pereira et al., 2022) primarily adopt either bottom-up or top-down approaches. Bottom-up frameworks detect individual body parts before grouping them into distinct animals, whereas top-down strategies first identify

Received: 11 November 2024; Accepted: 11 December 2024; Online: 12 December 2024

Foundation items: This work was supported by the STI 2030-Major Projects (2022ZD0211900, 2022ZD0211902), STI 2030-Major Projects (2021ZD0204500, 2021ZD0204503), National Natural Science Foundation of China (32171461), and National Key Research and Development Program of China (2023YFC3208303)

[#]Authors contributed equally to this work

*Corresponding authors, E-mail: xi.chen@ia.ac.cn; hua.han@ia.ac.cn

individuals and subsequently estimate body parts. To the best of our knowledge, all emerging top-down deep-learning-based approaches rely on a hierarchical tracking process, detecting entire bodies before refining the analysis to body parts (Luxem et al., 2023; Pereira et al., 2020). Although several approaches incorporate spatiotemporal constraints (Biderman et al., 2024) or employ reidentification or tracklet (i.e., fragments of trajectories) stitching to improve continuity across frames (Lauer et al., 2022), few track animals directly by matching. Unlike tracking-by-detection, tracking-by-matching establishes correspondences between objects across frames, enabling more robust generic object tracking (Javed et al., 2023). This framework does not require *a priori* knowledge of the tracked animal and is not species-constrained. To exploit these advantages, this study integrated Siamese networks (Chen et al., 2020) and discriminative correlation filters (Bhat et al., 2020b) into top-down pose estimation. These two widely used paradigms in generic object tracking and segmentation formed the basis of InteBOMB, our advanced successor to SiamBOMB (Chen et al., 2020). InteBOMB functions as a zero-shot tracker, markedly reducing manual annotations while maintaining comparable performance.

To develop a robust and adaptive framework for behavioral quantification, InteBOMB was designed with two key strategies tailored for laboratory scenes. The “background enhancement” strategy acquires extra background information from the relatively stable environment of home cages via few-shot learning. The “online proofreading” strategy corrects and stitches multi-animal tracks automatically or manually for long-term video recordings via active learning. These strategies are integrated into a newly developed plug-and-play PyTorch-based software, which enhances interoperability, interactivity, and user accessibility.

Recent advances have rapidly transformed computational ethology, making it more automated, scalable, and reproducible (Luxem et al., 2023). From a computational perspective, emerging vision foundation models learn general visual embeddings from natural scenes (Kirillov et al., 2023; Oquab et al., 2024). From an ethological perspective, end-to-end behavior analysis approaches extract behavioral patterns directly from raw frames or other signals (Bohnslav et al., 2021; Schneider et al., 2023). Integrating these approaches, this study explored the potential for general visual features to enhance behavior quantification. To further improve generalization, the visual features extracted from the encoder of generic object trackers for two additional techniques were reused in the InteBOMB workflow. The “automated labeling suggestion” technique uses visual features as meta-knowledge (Evans & Foster, 2011) of data for k-means or k-nearest neighbors (kNN)-based selective labeling. The “joint behavior analysis” technique uses visual features as additional data to extend the joint latent space produced by behavioral and neural data (Schneider et al., 2023).

To benchmark our InteBOMB workflow, which integrates two tracking-by-matching strategies for laboratory scenes and two data-efficient techniques for natural scenes, analyses were conducted on publicly available datasets of mice and non-human primates (Wilkinson et al., 2016). The mouse datasets encompass six experimental settings, including the open field test (with fisheye lenses), elevated plus maze (EPM), and forced swim test (FST), among others. Since mice are fundamental models for behavioral research and numerous commercial or open-source tools have been designed for laboratory-based tracking (Panadeiro et al., 2021), these datasets were used to assess plug-and-play tracking and segmentation. The non-human primate datasets also include six benchmarking conditions featuring monkeys and apes. Given the greater complexity of primate quadrupedal postures and social interactions (Marks et al., 2022), three of the six datasets were used to evaluate cross-dataset generalization for pose estimation in natural scenes. Importantly, these datasets do not overlap with those used in previous studies (Li et al., 2023; Liu et al., 2022; Yang et al., 2023), which primarily focused on single, in-house, home-cage datasets of specific primate species.

The contributions of this study are as follows: (i) integration of generic object tracking into top-down approaches with background enhancement and online proofreading strategies; (ii) reuse of visual features for automated labeling suggestion and joint behavior analysis; and (iii) benchmarking of the workflow using diverse datasets of mice and non-human primates. The following sections describe the study in detail. Section 1 of the Materials and Methods presents the collected datasets. Section 2 outlines the generic object tracking framework and provides an overview of the modular components of InteBOMB. Sections 3 and 4 detail the two proposed strategies, corresponding to contribution (i). Sections 5 and 6 focus on visual feature reuse, corresponding to contribution (ii). The Results section presents experimental evaluations, including a case study on the tree shrew. The Discussion section concludes with a discussion of findings and future research directions.

MATERIALS AND METHODS

Benchmarking datasets for mice and non-human primates

Quantitative analysis of animal behavior has been revolutionized by deep-learning-based frameworks, such as DeepLabCut (Lauer et al., 2022; Ye et al., 2024), SLEAP (Pereira et al., 2022), A-SOiD (Tillmann et al., 2024), and Lightning Pose (Biderman et al., 2024), dedicated to mouse behavior recognition via video recordings. Recently, emerging methods have been designed for non-human primates, including DeepLabCut, SIPEC (Marks et al., 2022), A-SOiD, and species-specific approaches such as MonkeyTrail (Liu et al., 2022), MonKit (Li et al., 2023), and BARN (Yang et al., 2023). A major challenge in behavioral quantification is achieving cross-species generalization, which requires models to maintain performance across diverse taxa. Some methods, including DeepLabCut, SLEAP, SIPEC, LabGym (Hu et al., 2023b), and SBeA (Han et al., 2024), have demonstrated generalization across multiple datasets spanning different species. To systematically assess the generalizability of InteBOMB, a series of open-access mouse and non-human primate datasets (Figure 1) were compiled to evaluate its performance across species, experimental procedures, camera parameters, and other related properties.

In constructing a benchmark for generic mouse tracking, dataset diversity was prioritized to capture variations in environmental setup, imaging conditions, subject numbers, and morphological differences. Table 1 summarizes six laboratory mouse datasets integrated from the research community, including “fisheye” (Salem et al., 2019), WFCI (Musall et al., 2019), EPM, FST (Sturman et al., 2020), “headset” (Liu et al., 2020), and “mice_hc” (Pereira et al., 2022). The EPM and FST datasets represent two classic behavioral paradigms using overhead camera views, while WFCI captures “reward-seeking” behavior (Luxem et al., 2023). The “fisheye” dataset uses an ultra-wide-angle lens that produces strong visual distortion intended to create a

panoramic image. The "headset" dataset, which presents challenges due to interference from electroencephalographic (EEG) and electromyographic (EMG) cables attached to the mice, monitors brain activity and motor behaviors. Finally, the “mice_hc” dataset consists of multiple interacting mice housed within a shared enclosure.

Similarly, to construct a comprehensive benchmark for non-human primate pose estimation, side-view datasets sourced from six publicly available sources were compiled, with three selected for evaluation in this study, including AP-10K,

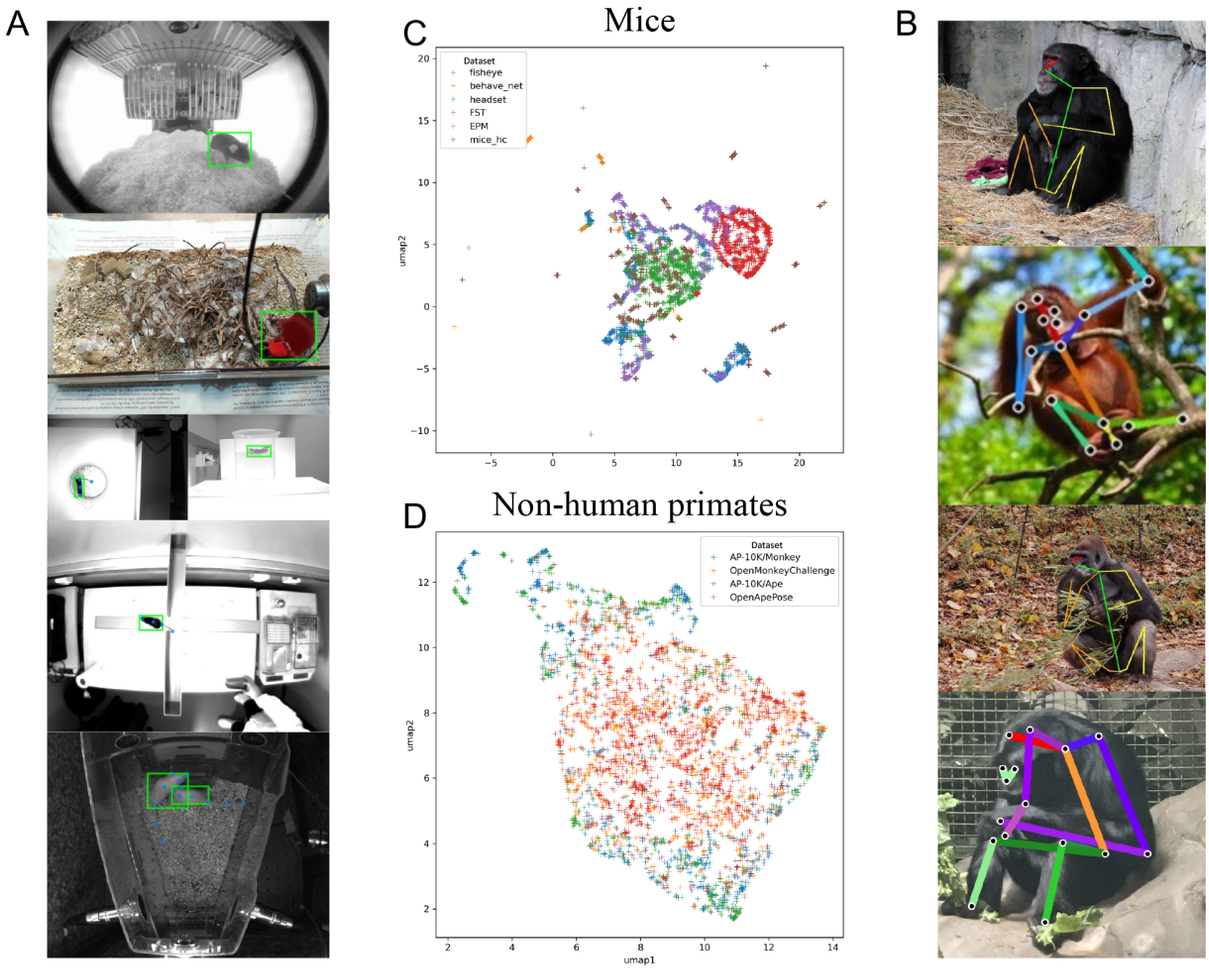

Figure 1 Overview of benchmarking datasets

A, B: Example images with manual annotations for the six mouse datasets in laboratory scenes (A) and three non-human primate datasets in natural scenes (B). C, D: UMAP visualization of manual annotations with normalized bounding boxes for mouse datasets (C) and position points for non-human primate datasets (D).

Table 1 Characteristics of benchmarking datasets

Name	Species	Label	Type	Annotated frames	Total duration
“fisheye”	Mouse (multi-view)	B-box, keypoint, behavior	Sparse	1 004	~5 min×15 trials
WFCI	Mouse (two-view)	Calcium imaging, keypoint, behavior	Dense	189×549 trials	6.3 s×549 trials
EPM	Mouse	B-box, keypoint, behavior	Sparse; dense	140; ~15 000×3 trials	~10 min×3 trials
FST	Mouse (two-view)			580; ~9 000×3 trials	~6 min×3 trials
“headset”	Mouse	B-box, mask, EEG, EMG, behavior	Sparse; dense	3 000; 5 000	2 min 47 s
“mice_hc”	Mouse (two-id)	B-box, keypoint	Sparse; dense	1 474; 2 560	1 min 42 s
“flies13”	*Drosophila* (two-id)			2 000; 2 560	1 min 42 s
AP-10K	3 apes and 5 monkeys (*Alouatta*, monkey, noisy night monkey, spider monkey, uakari)	Keypoint	Sparse	10 015 images collected and filtered from 54 species	
OpenMonkey Challenge	6 apes, 6 New World monkeys and 14 Old World monkeys	Keypoint	Sparse	111 529 images of 26 species	
OpenApePose	5 apes and non-siamang gibbons	Keypoint	Sparse	71 868 images of 6 species	

Species, label formats, number of annotated frames, and total duration for mouse (top) and non-human primate datasets (bottom). “Two-view” means two cameras with different viewpoints, “two-id” means two interacting individuals, “sparse” means annotated frames are randomly selected from video clips, and “dense” means continuous frame-by-frame annotations in the same video clip.

featuring three ape species and five monkey species (Yu et al., 2021); OpenMonkeyChallenge (OMC), encompassing six ape species and 20 monkey species (Yao et al., 2022); and OpenApePose (OAP), containing six ape species (Desai et al., 2023). AP-10K contains a greater number of animals per image compared to OMC and OAP and also features more extensive natural backgrounds. Additional datasets, including MacaquePose, Animal Kingdom, and APT-36K, were also considered but excluded from evaluation for the following reasons: (i) MacaquePose includes only two macaque species, making it unrepresentative of the distribution for either monkeys or apes. (ii) Animal Kingdom, despite spanning 850 species, only contains 2 975 images for non-human primates, a considerably smaller dataset than OMC (111 529) and OAP (71 868). (iii) APT-36K, as an extension of AP-10K, exhibits similar species composition, background characteristics, and data distribution (see Supplementary Materials).

To further illustrate dataset diversity, sample images were visualized alongside latent space distributions derived from labeled positions using uniform manifold approximation and projection (UMAP). As shown in Figure 1, these normalized landmark coordinates based on image size provide standardized and interpretable spatial representations (Desai et al., 2023).

Adhering to the FAIR principles —findability, accessibility, interoperability, and reusability (Wilkinson et al., 2016) —this study aimed to establish benchmark datasets that facilitate generalizable mouse tracking and non-human primate pose estimation. These benchmarks were designed for easy adoption by the research community, supporting evaluations of both commercial products (e.g., EthoVision by Noldus) and open-source frameworks. The full dataset collection is available in the Supplementary Materials and at: https://github.com/JackieZhai/awesome-behavior-datasets.

InteBOMB integrates tracking-by-matching methods

InteBOMB is a comprehensive workflow designed to support multi-modal behavioral analysis through a modular pipeline that extends from data acquisition to multi-purpose behavioral quantification. As shown in Figures 2 and 3, the framework consists of four core modules, including the model manager, data adapter, exception handler, and behavior analyzer.

To ensure broad applicability across different animal datasets, the model manager module allows both top-down and bottom-up approaches, as well as tracking-by-detection and tracking-by-matching strategies. Notably, the framework integrates alternative segmentation techniques for tracking models, enabling outputs that include both bounding boxes and masks. As pose estimation using masks enhances performance in multi-animal scenarios (Han et al., 2024), this flexibility is particularly beneficial. Two state-of-the-art pose estimation methods, SLEAP (Pereira et al., 2022) and multi-animal DeepLabCut (Lauer et al., 2022; Ye et al., 2024), are also incorporated via high-level application programming interfaces (APIs). To accommodate models from diverse sources while maintaining flexibility, the model manager offers both a streamlined model selection interface for ease of use and an advanced configuration panel for detailed customization. InteBOMB is model-agnostic, enabling users to select the best-suited model for their dataset without constraints imposed by predefined settings.

A key advancement in InteBOMB is the integration of generic object tracking models to serve as tracking-by-matching approaches. These models include Siamese networks and discriminative correlation filters (DCFs),

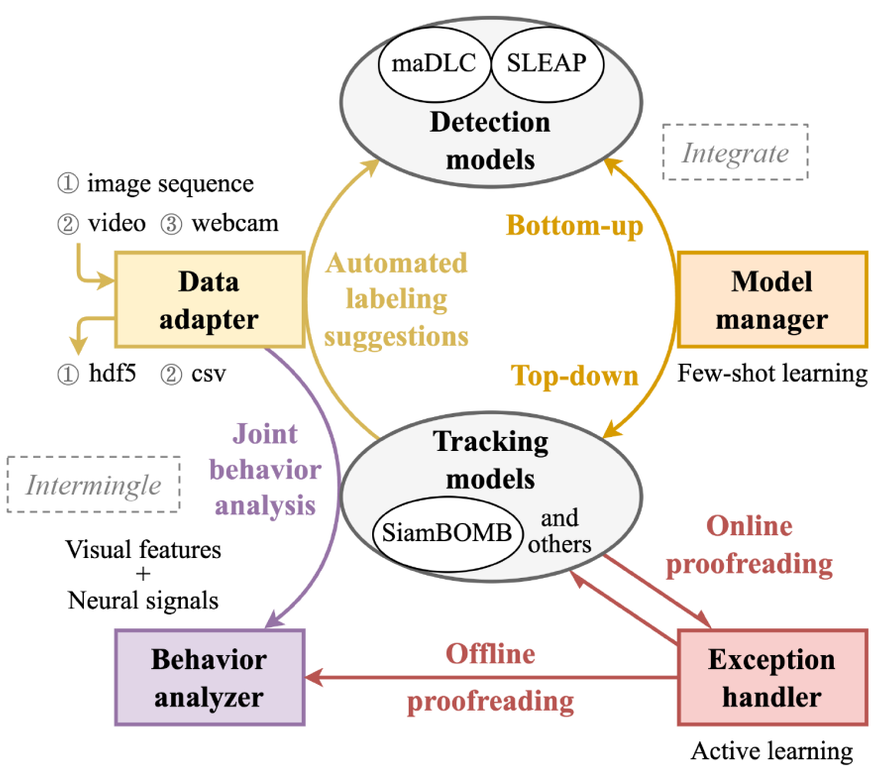

Figure 2 Overview of InteBOMB pipeline

Four main modules of InteBOMB include the data adapter, model manager, exception handler, and behavior analyzer. Model manager maintains top-down tracking models (SiamBOMB) and bottom-up detection models (maDLC and SLEAP). The data adapter (automated labeling suggestions) and behavior analyzer reuse visual features from trackers for automated labeling suggestions and joint behavior analysis techniques.

incorporating convolutional neural networks (CNNs) and transformer-based architectures (Cui et al., 2024; Hu et al., 2023a), as well as Vision Transformers (ViTs), such as SAM (Kirillov et al., 2023) and DINOv2 (Oquab et al., 2024). For laboratory-based video recordings, InteBOMB introduces a background enhancement strategy to optimize generic object tracking via few-shot learning (Figure 3). Details are described in the following first section.

To support efficient data management across labeling, training, and inference, InteBOMB includes a dual-track data adapter to facilitate both rapid inspection and compact storage. This adapter enables seamless interoperability by implementing CSV and HDF5 reader-writer formats in parallel, with CSV files saving the metadata and settings and HDF5 files organizing corresponding data and models in a key-value style. A highly versatile input-output API was also designed to convert well-arranged 1D–4D signals and datasets, such as video recordings, calcium imaging, and EEG and EMG recordings. To ensure broad compatibility with existing video analysis tools, InteBOMB adheres to best practices in computational ethology (Luxem et al., 2023), integrating widely used frameworks such as PyTorch (employed in DeepLabCut, Lightning Pose, BARN, and our generic object tracking), HDF5 (used in DeepLabCut and SLEAP), and CSV (used in most behavioral analysis tools).

Designed with researchers in neuroscience and ethology in mind, InteBOMB provides an intuitive graphical user interface (GUI) that eliminates the need for programming expertise. The GUI incorporates an exception handler, which enables automatic or manual intervention through real-time pop-ups, facilitating a pause-modify-resume operation. This module also utilizes a back-end online proofreading strategy driven by active learning (Figure 4). For users requiring high-performance inference, an alternative command-line interface (CLI) is available, supported by low-level APIs from graphics processing unit (GPU) acceleration and multi-processing pools.

As the successor to SiamBOMB (Chen et al., 2020), InteBOMB introduces several key innovations. A newly integrated automated labeling suggestion technique enhances annotation efficiency, while the built-in behavior analyzer enables classification and clustering across multiple modalities

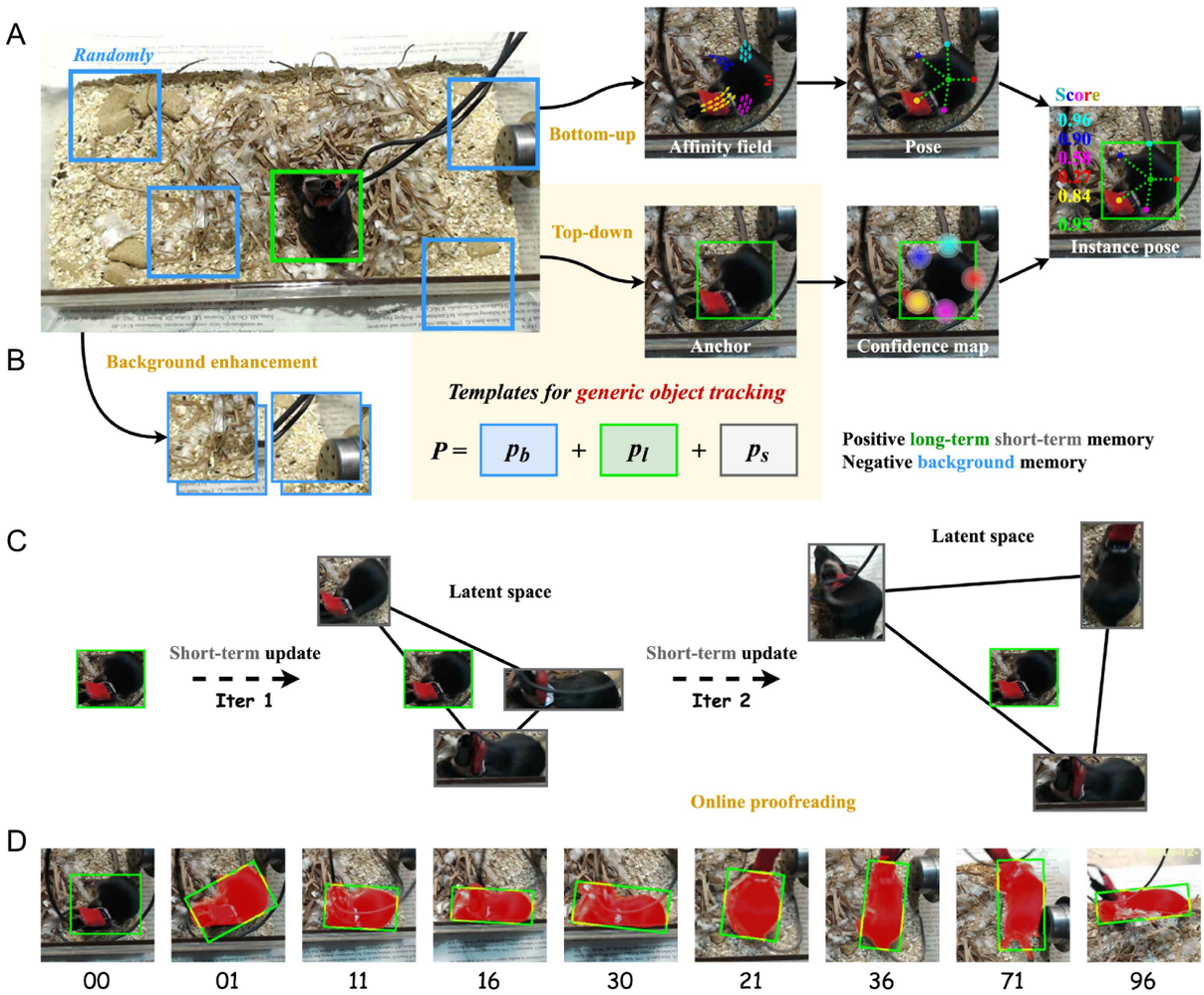

Figure 3 Top-down tracking-by-matching methods with background enhancement

A: InteBOMB implements both bottom-up and top-down approaches for animal tracking, segmentation, and pose estimation. Bottom-up approaches first detect all body parts with affinity fields and then group body parts into instances. Top-down approaches first locate instance anchors and then predict confidence maps of body parts within anchors. B: Background-enhanced generic object tracking methods use P templates (i.e., anchors) for the first step of top-down approaches. Online dynamic memory contains positive p_l long-term (green), p_s short-term (gray), and negative p_b background (blue) templates. C: Short-term memory iteratively updates its latent space as bounding box and mask predictions for animal bodies. D: Frame numbers are shown at the bottom.

within a joint latent space (Figures 2, 3). These techniques are detailed in Sections 3 and 4. To ensure reproducibility and accessibility, InteBOMB employs standardized tools for version control, packaging, and documentation, allowing users to deploy and utilize the workflow with minimal setup.

Background enhancement strategy via few-shot learning

A core challenge in computational ethology is accurately detecting all animal candidates and tracking individual targets to enable detailed behavior analysis. While tracking-by-detection approaches have advanced significantly, several interdisciplinary challenges remain, particularly in generalization to unseen animals (Kennedy, 2022). To address this, a shift in perspective was adopted, introducing tracking-by-matching strategies through Siamese networks and DCFs (Javed et al., 2023).

Traditional background subtraction considers the animal as the foreground while assuming a relatively stable background, a principle widely applied in laboratory-based tracking (Liu et al., 2022). Building on this foundation, foreground-background pairs or target-match-mismatch triplets were incorporated into stable generic animal tracking models, enabling training through few-shot learning algorithms. In this framework, video sequences with ground-truth annotations provide support sets from the first frame, while subsequent frames within the same clip act as query sets. For example, using contrastive loss, Siamese networks predict whether two samples are a positive (matching) or negative (mismatching) pair. The loss function minimizes the distance between visual embeddings of positive pairs and maximizes the distance between embeddings of negative pairs. Triplet loss extends this approach by introducing a target sample alongside two additional samples (one matching, one not), forcing the model to refine its embeddings for more precise differentiation between positive and negative matches.

In the first stage of top-down tracking, animal bodies are localized by matching against manually specified bounding boxes in the first frame of a video. According to the principles of generic object tracking, the deep network is trained offline on a large dataset of paired animal patches to learn a matching function $g(f)$, after which this network as a function $h[g(f)]$ is evaluated online for real-time tracking. The function f represents the shared backbone, acting as an encoder for both the target animal of interest x in the first frame and the broader search region z in the next frame. The function g applies a cross-correlation operation $\star$,

$$g(x,z) = f(x) \star f(z) + b \tag{1}$$

where b denotes an offset scalar and $g(x,z)$ represents a response map that quantifies the similarity between z and x. Furthermore, the model incorporates multiple decoder heads h to facilitate multi-task learning. To track animal bodies, an anchor-based bounding-box regression method was implemented using a regional proposal network (RPN), consisting of a classification head $h_c[g(x,z)]$ and regression head $h_r[g(x,z)]$. For enhanced localization, a parallel segmentation head $h_s[g(x,z), f(x)]$ with multi-scale skip-connections was integrated. Simple cross-correlation $\star$ is often replaced with depth-wise cross-correlation to produce multi-channel response maps (Hu et al., 2023a). By leveraging a Siamese network architecture, the tracker learns the generic relationship between animal motion and appearance, enabling it to generalize to previously unseen species, even those absent from the training dataset.

The InteBOMB model manager incorporates SiamBOMB (Chen et al., 2020) with small modifications. As described in the original SiamBOMB paper, the recordings generated from fixed cameras contain relatively stable backgrounds, with only

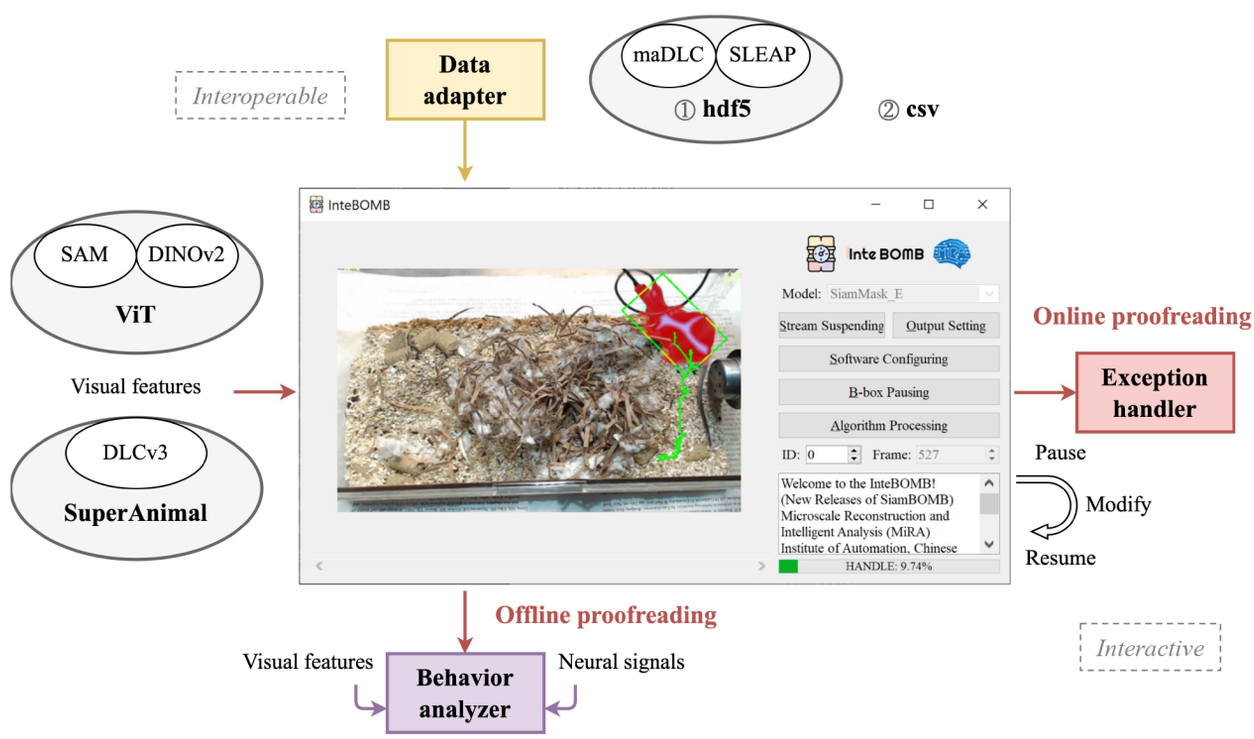

Figure 4 User-friendly software interface with online proofreading and offline joint behavior analysis

Screenshot of main InteBOMB graphical user interface (GUI) with interactive modules in InteBOMB. The workflow saves ViT features and runs SuperAnimal models via a command-line interface (CLI). The exception handler actualizes online pause-modify-resume proofread operations via InteBOMB pop-ups.

small perturbations (e.g., illumination variation). Contrastive information was simultaneously collected from both few-shot foreground and background patches (Figure 3). Notably, whenever all animals in the frame are manually labeled by the bounding box, the software randomly samples some patches from the rest of the same frame. Both margin contrastive loss $\mathcal{L}_{\text{con}}$ and triplet loss $\mathcal{L}_{\text{tri}}$ were considered (Hu et al., 2023a):

$$\mathcal{L}_{\text{con}}(\theta)=\sum_{x\in\mathcal{D}}|f_\theta(z)-f_\theta(x)|_2^2+(1-y_{xz})\max\left(0,\varepsilon_1-|f_\theta(z)-f_\theta(x)|_2^2\right) \tag{2}$$

$$\mathcal{L}_{\text{tri}}(\theta)=\sum_{(x^p,x^n)\in\mathcal{D}}\max\left(0,|f_\theta(z)-f_\theta(x^p)|_2^2-|f_\theta(z)-f_\theta(x^n)|_2^2+\varepsilon_2\right) \tag{3}$$

where $|\bullet|_2^2$ is the Euclidean distance of L_2-normalized features, $\mathcal{D}$ is the support set, $y_{xz}\in\{0,1\}$ indicates whether x and z are matched or not, θ is the network parameters, and ε_1 and ε_2 are the minimum distance margin that pairs or triplets distinguishing mismatched objects should satisfy. These losses exploit the pairwise and triplet relationships and structural connections between the positive and negative templates (x^p,x^n) of the animal target (Javed et al., 2023). Cross-entropy loss $\mathcal{L}_{\text{cls}}$ is for the classification head, smooth L_1 loss $\mathcal{L}_{\text{reg}}$ is for the bounding box regression head, and binary logistic regression loss $\mathcal{L}_{\text{mask}}$ is for the mask head. These losses compose the multi-task loss:

$$\mathcal{L}_{\text{mul}}=\text{”}\lambda_1\mathcal{L}_{\text{cls}}+\text{”}\lambda_2\mathcal{L}_{\text{reg}}+\text{”}\lambda_3\mathcal{L}_{\text{mask}} \tag{4}$$

with $\text{”}\lambda_1=\text{”}\lambda_2=1$ and $\text{”}\lambda_3=32$ (Hu et al., 2023a).

Online proofreading strategy via active learning

To ensure accurate identity tracking across frames, InteBOMB incorporates an online proofreading strategy based on active learning. Once tracklets are produced, users must verify and correct identity assignments of animal individuals to maintain tracking consistency. Both online and offline proofreading methods have been implemented within the exception handler. Offline proofreading, similar to DeepLabCut and SLEAP, provides users with an interface to review and correct results after the tracker has processed a batch of frames. In contrast, online proofreading functions as a back-end process within the GUI, sequentially processing frames from video recordings as they become available, with newly proofread frames continuously incorporated to refine the best tracker for future frames at each proofreading iteration. In this context, online and offline refer to the concept of machine learning rather than internet connectivity.

To enhance proofreading efficiency, dynamic memory (Yang & Chan, 2021) is employed to store foreground-background templates derived from the generic object tracker. This dynamic memory structure includes positive long-term and short-term memory and negative memory components (Figure 3). Long-term memory contains p_l templates, including first-frame and online-proofread annotations. These templates are annotated through pause-modify-resume operations derived from InteBOMB pop-ups, which select score-based (i.e., confidence-based) unreliable templates (Figure 4). Short-term memory contains p_s dynamic templates, writing and reading previous templates to deal with target appearance variations. Negative memory contains p_b templates for distractors, identified and extracted through the background enhancement strategy. This process aligns with the active learning paradigm, allowing the algorithm to selectively identify data it wants to learn from and achieve a higher degree of accuracy with fewer training annotations.

Siamese networks can benefit from spatiotemporal regularization to compute the correlation response map within cross-correlation. Accordingly, classification head h_c was applied to predict scores in CNN-based Siamese trackers (e.g., SiamMask), and an extra score token with attention layers was added to predict scores in Transformer-based Siamese trackers (e.g., MixFormer).

Additionally, DCF-based trackers inherently support online appearance model updates. The DCF aims to learn a filter ω, such that $x*\omega\approx y$ for all samples (x,y) in the dataset. Filter ω can be learned by minimizing a generalized objective function (Bhat et al., 2020b),

$$L(\omega)=\sum_{(x,y)\in\mathcal{D}}|W_\theta(y)[x*\omega-B_\theta(y)]|^2+\lambda|\omega|^2 \tag{5}$$

Using the steepest descent iteration, the approximate solution $\omega^{i+1}=\omega^i-\alpha^i\nabla^i$ is obtained, where step-length α^i minimizes loss along the gradient direction $\alpha^i=\text{argmin}_\alpha L\left(\omega^i-\alpha\nabla^i\right)$, and $\nabla^i=\nabla L\left(\omega^i\right)$ represents the gradient at ω^i (Bhat et al., 2020b). The gradient of the updating learnable filter ω is then utilized as a scoring mechanism. These predicted scores (i.e., classification scores from Siamese networks and gradient-based scores from DCFs), form the basis for uncertainty and diversity sampling within the active learning framework (Tillmann et al., 2024). Finally, given $P=p_l+p_s+p_b$ template patches from dynamic memory, the trackers establish correspondences between the support set $P\times H_x\times W_x\times 3$ and query search region $H_z\times W_z\times 3$ in subsequent frames (Cui et al., 2024), where H and W denote the height and width of patches in three-channel (RGB) frames.

Metadata-driven labeling suggestions for pose estimation

In behavior classification from raw video frames, methods that capture general image features may improve accuracy over pose-based features (Pereira et al., 2020). With the development of vision foundation models, end-to-end approaches offer strong potential for generalization, enabling direct behavior classification from video recordings. Both unsupervised frameworks, such as BehaveNet (Batty et al., 2019), and supervised methods, such as DeepEthogram (Bohnslav et al., 2021), have demonstrated the feasibility of this approach. This study proposes an alternative strategy, leveraging general features as an initial perceptual layer to automate labeling suggestions for pose estimation.

Existing pose estimation tools, such as DeepLabCut and SLEAP, employ distinct techniques for automated labeling. DeepLabCut directly downsamples videos and applies k-means clustering, treating each frame as a feature vector. SLEAP, in contrast, uses handcrafted image features and utilizes principal component analysis (PCA) to reduce feature dimensionality for k-means clustering. For instance, SLEAP generates a bag of feature vectors using the histogram of oriented gradients (HOG) descriptors, enabling the selection of visually distinctive frames for annotation.

In contrast to these methods, the InteBOMB workflow reuses the general features extracted from generic object tracking models. Empirical evaluations indicated that Transformer-based trackers exhibited greater robustness and superior performance (see Results). Consequently, ViT embeddings from foundation models (e.g., semi-supervised SAM or self-supervised DINOv2) were selected as visual features for each frame. Although UMAP and t-distributed stochastic neighbor embedding (t-SNE) are excellent

nonlinear methods, they are not as directly interpretable as PCA. To maintain interpretability and consistency, PCA was applied for dimensionality reduction, projecting 2D features into a 1D vector. Inspired by meta-knowledge concepts (Evans & Foster, 2011), these PCA-vectorized visual features were defined as "automated" metadata, generalized representations of dataset structure. Two clustering strategies were implemented for prototype selection, k-means clustering for prototype selection and iterative random sampling combined with kNN to filter neighbor vectors at each step. Through this process, representative frames were automatically identified for selective labeling.

InteBOMB natively generates and saves visual features during tracking and segmentation of animal bodies, allowing users to choose between automated metadata-based selective labeling or prediction-assisted labeling for the next training procedure of pose estimation. The latter method implements pretrained SuperAnimal (Ye et al., 2024) models to generate initial results, which users can modify to produce faster annotations.

Behavioral analyses intermingle with existing visual features

A fundamental goal of neuroscience is to map behavioral actions to neural activity. For example, Musall et al. (2019) used widefield calcium imaging (WFCI) alongside video recordings to study neural activity during auditory and visual decision-making. Similarly, Liu et al. (2020) characterized natural brain states and motor behaviors in freely moving mice by synchronizing EEG, EMG, and video recordings. Further advancements, such as MoseVenue3D (Han et al., 2022), have combined high-resolution miniature two-photon microscopy and behavioral tracking, enabling synchronous neural recording in mice.

Recently, deep neural networks, including foundation models, have been introduced to provide precise behavior quantification and joint behavior analysis. For example, CEBRA (Schneider et al., 2023) integrates behavioral and neural data using both supervised hypothesis-driven and self-supervised discovery-driven approaches to produce consistent latent spaces. This framework has been instrumental in producing consistent latent spaces, aligning data across two-photon imaging and neuropixel recordings, and achieving high-accuracy decoding of neural signals from the visual cortex.

Existing visual features were reused to integrate with neural signals, such as calcium imaging. The InteBOMB PCA-vectorized metadata and neural signals are combined to create a joint latent space. This procedure extends the dimensionality of feature spaces, increasing the likelihood of matching diverse behaviors. Following CEBRA, DINO models were applied to embed video frames into latent space, and SAM models were also implemented for the same purpose. A built-in behavior analyzer was developed to utilize joint latent spaces for both unsupervised clustering and supervised classification (see Supplementary Materials for details). For clustering, UMAP visualization and k-means partitioning were used to uncover behavioral structures. For classification, several ensemble methods were employed, including boosting, bagging, and combined approaches (Fauvel et al., 2019).

RESULTS

The InteBOMB pipeline (Figure 2) represents the implementation of generic object tracking for animal behavioral quantification. To enhance tracking accuracy in laboratory-based video recordings, two strategies were designed: "background enhancement" (Figure 3) and "online proofreading" (Figure 4). Moreover, two techniques were introduced to reuse visual features generated and saved during object tracking for downstream behavior analysis. The "automated labeling suggestion" technique identifies representative frames for training set annotation in pose estimation, while the "joint behavior analysis" technique extends the latent space for behavior classification and clustering.

To quantify the performance of InteBOMB, evaluations were conducted across bounding boxes, masks, and specific keypoints for tracking, segmentation, and pose estimation in mouse and non-human primate datasets. The benchmarking datasets comprised video recordings of multiple species (mice, flies, monkeys, and apes) captured in varied environments (laboratory and natural settings), across various viewing angles (including multiple view setups) and diverse animal numbers.

General-purpose methods reach comparable performance in laboratory scenes

A comprehensive evaluation of animal tracking requires a wide range of models and datasets for comparison. To ensure robustness, six distinct laboratory home-cage datasets (Table 1) were selected, representing standard behavioral experiments (WFCI, EPM, and FST), social interaction studies ("mice_hc" and "files13"), and synchronous neural-behavior recordings ("headset"). Additionally, a broad collection of open-source generic object tracking models was compiled to assess their adaptability to animal tracking. The evaluated tracking models were categorized based on their underlying architectures: (i) Siamese network-based trackers: SiamMask (Hu et al., 2023a), MixFormer (Cui et al., 2024); (ii) DCF-based trackers: DiMP (Danelljan et al., 2020), KYS (Bhat et al., 2020a), ToMP (Mayer et al., 2022); (iii) Trackers with alternative segmentation: SiamMask, LWL (Bhat et al., 2020b), RTS (Paul et al., 2022); (iv) CNN-based trackers: SiamMask, DiMP, KYS, LWL, RTS; (v) Transformer-based trackers: MixFormer, ToMP; and (vi) Graph neural network (GNN)-based trackers: KeepTrack (Mayer et al., 2021). Please see Supplementary Materials for the full list.

Each tracker was initialized using the bounding box of the first frame for each video clip. To simulate human-in-the-loop correction, reinitialization with online proofreading was applied when tracking failures occurred. In datasets where the first frame contained no animals (e.g., EPM and FST), the initial tracking frame was defined as the moment the first animal appeared. Tracking accuracy was evaluated using root-mean-square error (RMSE) and intersection over union (IoU) between ground-truth and predicted bounding boxes.

As shown in Table 2, performance was compared against SuperAnimal-TopViewMouse (Ye et al., 2024) and supervised methods such as SLEAP and SIPEC. The background enhancement (+B) and online proofreading (+O) strategies improved performance in both CNN-based trackers (SiamMask) and Transformer-based trackers (MixFormer). The integrated "SM+B+O" and "MF+B+O" models surpassed SuperAnimal and reached performance levels comparable to SLEAP and SIPEC. Transformer-based trackers demonstrated higher accuracy for size-invariant animals (EPM, FST, and "headset"), while CNN-based trackers performed better in datasets with high variability in animal shapes and sizes ("fisheye" and "mice_hc"). Our trackers

demonstrated “zero-shot” performance, effectively tracking unseen datasets and species without additional fine-tuning.

One noted limitation of SuperAnimal models was their sensitivity to background interference (e.g., laboratory assistants in EPM and FST). This issue was mitigated by cropping frames to the region of animal movements, an approach further enhanced by the background enhancement strategy. For pose estimation, both maDLC (TensorFlow-based v.2.3.10) and DLCv3 (PyTorch-based v.3.0.0rc2) top-down HRNet-w32 models were tested. These models were pretrained on the TopViewMouse-5K dataset (https://doi.org/10.5281/zenodo.10618947) using default settings. In contrast, SLEAP and SIPEC were trained on in-distribution datasets using official protocols.

Several tracking models, including SiamMask, LWL, RTS, and SIPEC, were capable of generating mask outputs. Among these, the first three generic object trackers predicted masks automatically based on the input bounding boxes without requiring additional mask annotations. In contrast, SIPEC, a Mask R-CNN-based tracker, required explicit mask annotations for training. To facilitate direct comparison, 3 000 frames of mice from the “headset” dataset were manually annotated for SIPEC training, enabling mask-based segmentation evaluation. Tracking and segmentation performance was quantified using precision and recall for ground-truth and predicted masks, Jaccard index (IoU) for masks, and the F1 score for bipartite matching between boundary pixels of masks (Hu et al., 2023a).

As shown in Table 3, integrating background enhancement and online proofreading significantly improved SiamMask performance, allowing it to match the accuracy of SIPEC with extra in-distribution annotations. Results demonstrated that generic object segmentation models can be effectively adapted for fixed-camera video recordings, with performance reaching human-level accuracy in top-view mouse laboratory datasets. The best-performing approach exhibited higher recall, improved robustness, and enhanced tracking accuracy.

Selective labeling improves cross-dataset performance in natural scenes

To evaluate pose estimation, the PyTorch-based toolbox MMPose was reproduced, incorporating the HRNet network architecture (Wang et al., 2021). DeepLabCut (SuperAnimal) and SLEAP also employ this kind of keypoint detection network. The evaluation followed the standard top-down settings detailed in the official implementation documents: https://mmpose.readthedocs.io/en/latest/model_zoo/animal_2d_keypoint.html.

Two non-human primate datasets recorded in natural environments were selected for testing: AP-10K and OMC. AP-10K (Yu et al., 2021) comprises 10 015 annotated images, spanning 54 species across 23 families. A subset of 400 images from *Hominidae* (three species) and 674 images from *Cercopithecidae* (five species) was used. OMC (Yao et al., 2022) includes 111 529 annotated images representing 26 species, including six New World monkeys, 14 Old World monkeys, and six apes. The validation set of 22 306 images was used for analysis. The imbalance between these datasets, with OMC containing approximately 20 times more

Table 2 Zero-shot tracking methods surpass general-purpose detection methods and show comparable performance to supervised detection methods

Model	“fisheye”		EPM		FST	
Metric	RMSE	IoU	RMSE	IoU	RMSE	IoU
SiamMask	41.8±80.4	0.638±0.419	22.5±52.6	0.638±0.187	5.71±1.64	0.659±0.081
DiMP	44.9±81.7	0.626±0.320	19.2±64.4	0.671±0.209	3.75±1.72	0.735±0.082
KYS	67.0±79.4	0.444±0.380	67.7±165	0.559±0.251	33.6±23.4	0.318±0.335
LWL	83.0±86.2	0.347±0.338	10.2±18.2	0.561±0.144	126±137	0.184±0.234
KeepTrack	66.6±114	0.591±0.348	8.75±20.1	0.724±0.175	3.62±1.97	0.731±0.094
ToMP	38.2±80.1	0.658±0.282	8.99±19.9	0.642±0.140	4.00±3.10	0.678±0.102
MixFormer	113±129	0.409±0.404	4.47±17.9	0.816±0.103	2.39±1.31	0.789±0.078
SM+B+O	**35.5±69.8**	**0.661±0.211**	9.02±18.4	0.719±0.139	4.14±2.40	0.667±0.108
MF+B+O	81.5±76.3	0.425±0.346	**3.79±13.0**	**0.841±0.084**	**2.31±1.23**	**0.795±0.070**
DLC (SA)	139±96.3	0.151±0.191	90.6±120	0.513±0.353	7.22±5.01	0.536±0.201
DLC (SA) w/ crop	–	–	7.66±23.2	0.750±0.111	4.57±2.59	0.683±0.114
Model	**“headset”**		**“mice_hc”**		**“flies13”**	
Metric	RMSE	IoU	RMSE	IoU	RMSE	IoU
SiamMask	21.5±16.2	0.678±0.173	89.6±88.7	0.358±0.201	11.0±18.3	0.728±0.218
DiMP	20.1±14.7	0.680±0.155	91.0±111	0.335±0.268	13.8±16.9	0.714±0.206
KYS	23.2±17.0	0.658±0.157	157±174	0.232±0.243	27.2±33.0	0.598±0.309
LWL	36.8±26.8	0.424±0.215	93.4±63.2	0.205±0.150	8.49±6.16	0.659±0.097
KeepTrack	25.3±16.4	0.625±0.154	80.3±96.7	0.324±0.250	6.53±8.78	0.800±0.110
ToMP	27.5±15.5	0.616±0.146	85.2±90.5	0.290±0.234	4.24±2.42	0.818±0.072
MixFormer	17.1±13.6	0.729±0.148	94.7±109	0.289±0.254	4.11±4.73	0.850±0.083
SM+B+O	7.52±6.27	0.854±0.081	**52.1±45.2**	**0.609±0.188**	6.70±9.33	0.795±0.108
MF+B+O	**5.05±4.29**	**0.873±0.062**	53.6±64.7	0.583±0.173	4.05±5.01	0.851±0.074
SLEAP	–	–	93.9±70.7	0.344±0.295	**1.46±1.03**	**0.944±0.028**
SIPEC	5.48±4.34	0.863±0.066	–	–	–	–

Generic object tracking methods, generic object tracking methods (SM: SiamMask; MF: MixFormer) with background enhancement (+B) and online proofreading (+O) strategies, and general-purpose methods (SuperAnimal from DLC) trained by TopViewMouse-5K dataset or supervised methods (SLEAP and SIPEC) trained by in-distribution datasets. The numbers in bold indicate the highest-ranking results. –: Not available.

images than AP-10K, provided a challenging scenario for assessing data selection methods under uneven sampling conditions.

The HRNet models were trained separately on each dataset, followed by in-distribution evaluations and cross-dataset generalization tests. RMSE was used to quantify keypoint prediction accuracy by comparing corresponding keypoints with identical semantics across datasets. Mutual keypoints were aligned across AP-10K, OMC, and SuperAnimal-Quadruped datasets using conversion mapping, vocabulary projection, and dataset merging (Ye et al., 2024). Please see Supplementary Materials for details on the body part mapping table.

As shown in Figure 5, the first column represents in-distribution testing, while subsequent columns correspond to out-of-distribution testing. The final column illustrates a fine-tuning experiment using only 1% (227 images) of selective labels to refine SuperAnimal models. For each out-of-distribution test, the training set excluded the out-of-distribution testing set. RMSE heatmap of all 15 keypoints indicated that automated labeling suggestions improved performance of the general-purpose SuperAnimal models, surpassing direct cross-dataset testing. The RMSE distributions in Figure 6 highlight that these improvements were particularly pronounced for looming hard-to-detect body parts, such as shoulders, elbows, paws, and knees.

A qualitative comparison was conducted against DeepLabCut and SLEAP on both mouse and non-human primate datasets to further evaluate the effectiveness of automated labeling selection. DeepLabCut uses k-means clustering in the feature space of downsampled images, while SLEAP uses k-means clustering in the feature space of HOG (or other classic operators) and CEBRA uses kNN in DINO feature space, pretrained by ImageNet (Schneider et al., 2023). Given the limited capacity of the training set, clear k-means centers can facilitate the identification of diverse frames for general pose estimation. As shown in Figure 7, UMAP visualization revealed that the automated labeling suggestion technique produced more compact intra-class clusters while maintaining greater separation between clusters. In particular, cross-dataset data points were more evenly distributed. The degree of clustering separation was further quantified using Silhouette coefficients (Rousseeuw, 1987), which measure cluster compactness and separation. As shown in the bottom-left corner of each plot in Figure 7, our “embedding_vit_pca” clusters achieved the highest Silhouette score for both mice and non-human primates.

Joint analysis assists in behavior clustering and classification

To evaluate animal behavior analysis, a series of widely used classification ensemble methods were implemented. The

Table 3 Zero-shot tracking and segmentation performance

Model	“headset”			
Metric	Precision	Recall	IoU	F1
Human (two annotators)*	–	–	0.879	0.896
LWL	0.629±0.295	0.654±0.293	0.502±0.231	0.490±0.229
SiamMask	0.822±0.060	0.887±0.051	0.773±0.094	0.768±0.178
RTS	0.810±0.078	0.953±0.035	0.778±0.071	0.783±0.090
SM+B (SiamBOMB)*	–	–	0.844	0.852
SM+B+O	0.892±0.044	**0.963±0.030**	**0.849±0.057**	**0.858±0.076**
SIPEC	**0.961±0.051**	0.867±0.057	0.836±0.054	0.853±0.112

Generic object segmentation methods (SM: SiamMask) with background enhancement (+B) and online proofreading (+O) strategies, supervised methods (SIPEC) trained by in-distribution datasets. Human performance refers to variation between two annotators. “*” indicates results from original SiamBOMB paper (Chen et al., 2020). –: Not available. The numbers in bold indicate the highest-ranking results.

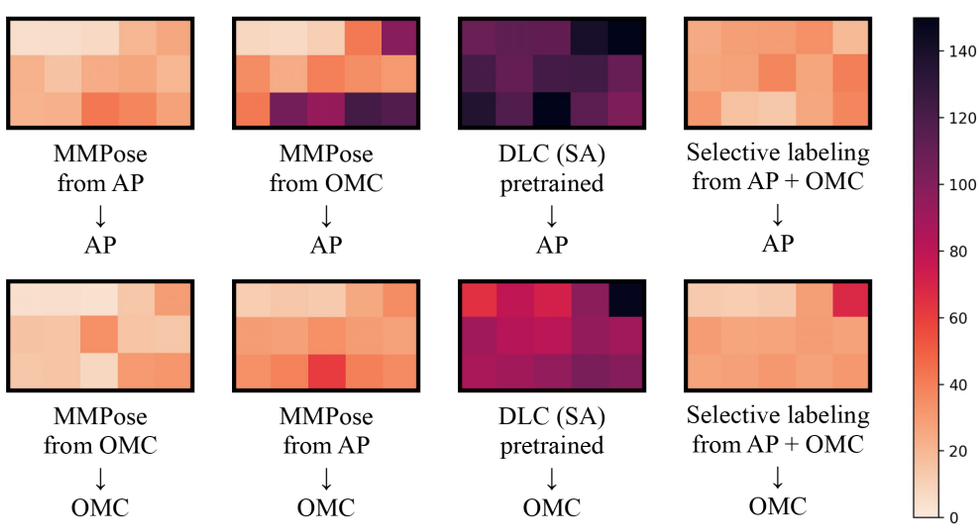

Figure 5 Cross-dataset generalization of pose estimation

Heatmap of each corresponding keypoint prediction result from cross-dataset generalized models. The first column represents in-distribution supervised learning, while the other three columns represent out-of-distribution transfer learning. The third column represents zero-shot learning DeepLabCut (SuperAnimal-Quadruped), while the final column represents fine-tuned SuperAnimal-Quadruped with selective labeling.

construction of an ensemble method involves combining multiple accurate and diverse individual predictors. Two well-established approaches were selected to modify the distribution of the original training data with complementary effects on bias-variance (underfitting-overfitting) trade-off: bagging (variance reduction) and boosting (bias reduction). Random Forest was deployed as a bagging method, while XGBoost served as a boosting method. Additionally, LCE, which combines bagging and boosting through a divide-and-conquer approach, was incorporated to individualize predictor errors on different parts of the training data (Fauvel et al., 2019). All ensemble methods were integrated into the behavior analyzer module (Figure 4) within the GUI and CLI of InteBOMB, following the official implementation guidelines: http://lce.readthedocs.io/en/latest/tutorial.html.

To assess the effectiveness of these classifiers, the WFCI dataset was selected to test common behavioral-neural recordings. This dataset features a head-fixed mouse performing a visual decision-making task, with concurrent neural activity recorded across the dorsal cortex using WFCI. Behavioral data were captured by two grayscale cameras (one side view and one bottom view) at 30 Hz, synchronized with neural activity recordings at the same frame rate. The data consisted of 549 trials, each spanning 189 frames (Batty et al., 2019).

Classification accuracy was quantitatively assessed by comparing the precision, recall, and F1 scores of four different classifiers, with and without the inclusion of PCA-vectorized visual features (metadata). The evaluation focused on two prediction tasks: an eight-class target, called “stimTime”, and a seven-class target, called “spoutTime”. All 549 trials were used, with a training-to-testing set ratio of 7:3. As shown in Table 4, classification performance was improved by the additional visual information from general tracking models, forming a higher dimensional joint latent space to be more easily mapped by classifiers.

To further investigate the impact of visual features, a qualitative analysis was conducted on latent space properties with and without visual feature integration. The clustering structure was examined using UMAP visualization on trial 15, focusing on the discrete behavioral variable “R_Spout” (three categorical choices) and the continuous variable “R_paw” (ranging from −4 to 3). As shown in Figure 8, the clustering results in the right column exhibit greater intra-class cohesion and increased inter-class separation compared to the left column. Additionally, the incorporation of automated metadata disrupted the elongated, string-like structure of neural signals in the latent space, resulting in a more compact and interpretable representation of behavioral patterns.

Case study of a never-before-seen species (tree shrew)

To further evaluate the generalization capability of the InteBOMB workflow, its performance was evaluated on a previously unseen species. The tree shrew has long been proposed as an alternative laboratory animal model due to its close evolutionary relationship with primates (Yao et al., 2024). To simulate a real-world scenario, tracking, segmentation, and selective labeling were conducted on a demo video recording of a tree shrew, using default settings and without prior species-specific training.

Only the first frame was manually annotated with a bounding box, after which generic object models were employed to automatically track and segment the tree shrew across subsequent frames (Figure 9). MixFormer was used for predicting bounding boxes and SiamMask for masks, both enhanced with background enhancement and online updating strategies. As these models have different real-time efficiency,

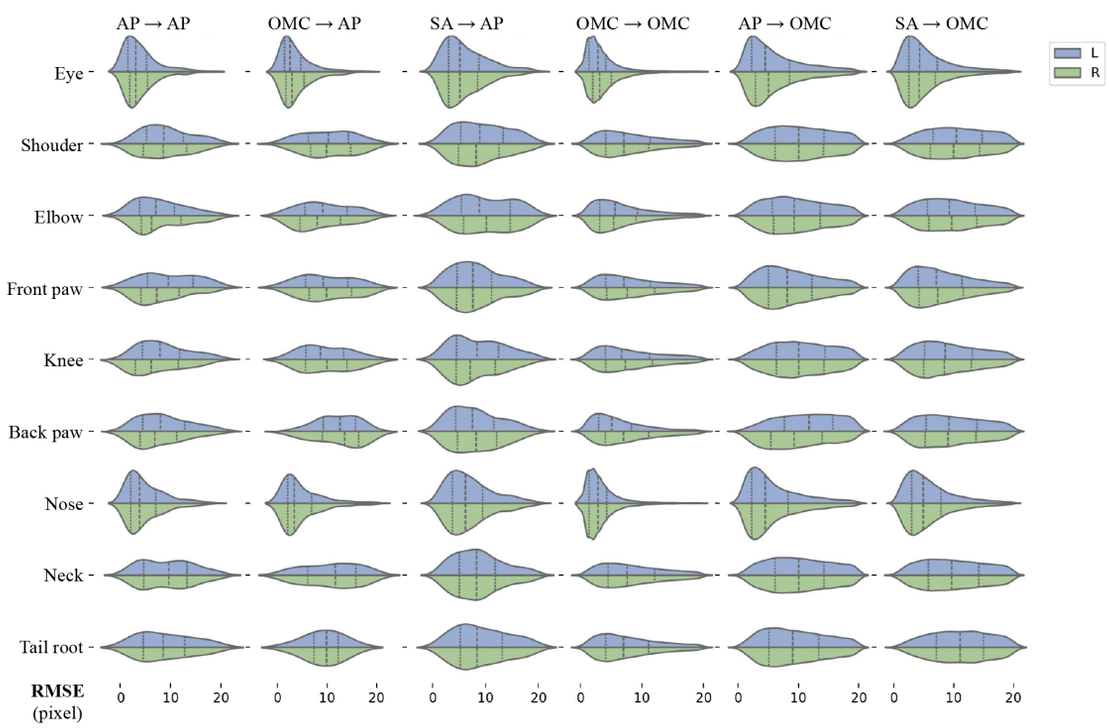

Figure 6 Distribution of keypoint prediction errors for pose estimation

Violin plots display keypoint RMSE for three scenarios: Training and testing on the same dataset, training and testing on different datasets, and training with selective labeling followed by testing on AP-10K or OMC (from left to right). Upper halves represent left-side body parts, while bottom halves represent right-side body parts. Vertical dotted lines indicate first, second, and third quartiles.

bounding boxes were computed at the original 25 frames per second (FPS), while masks and additional embeddings were extracted at a sparser 6 FPS (see Supplementary Videos).

As shown in Figure 9, PCA-vectorized ViT embeddings were generated and saved during tracking and segmentation. These embeddings were projected using UMAP and clustered into 10 groups via k-means or spectral clustering. All cluster centers (red stars) were adopted for next-step selective labeling to train pose estimation models. Notably, two outlier clusters (dashed boxes) exhibited pronounced stretching and deformation, suggesting the need for more concentrated labeling. Overall, these results demonstrated that the automated labeling suggestion technique effectively captures the diverse postural variations of tree shrews, enabling efficient model adaptation to novel species.

DISCUSSION

This study introduced InteBOMB, a general-purpose workflow for animal tracking, segmentation, and pose estimation. To the best of our knowledge, this is the first framework to integrate generic object tracking into top-down approaches, bridging advancements in Siamese networks and discriminative correlation filters. Extensive investigations and experiments were conducted to optimize these trackers using background enhancement and online proofreading, using few-shot learning and active learning procedures to improve tracking and segmentation in laboratory environments. Additionally, existing visual features extracted by generic object trackers were repurposed for pose estimation in natural scenes, introducing PCA-vectorized visual features, also called automated metadata. These embeddings facilitated further applications,

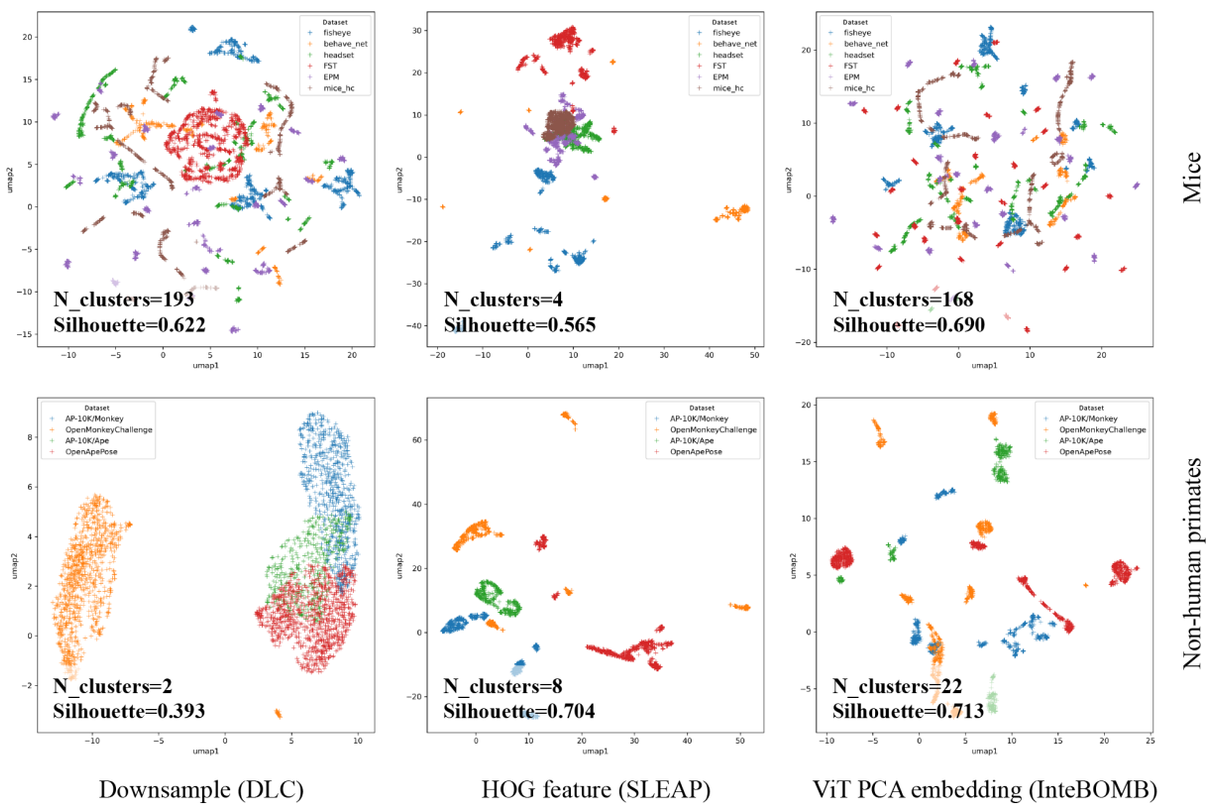

Figure 7 Comparison of data selection methods with UMAP visualization

Different data selection methods —DeepLabCut (downsampled images), SLEAP (HOG features), and InteBOMB (PCA-vectorized visual features)—are shown for mouse (top row) and non-human primate datasets (bottom row). Best number of k-means clusters (N_clusters) and Silhouette coefficients are shown in the left-bottom corner of each plot.

Table 4 Comparison of behavior classification with and without visual features

Target		"stimTime"			"spoutTime"		
Metric		P	R	F1	P	R	F1
Neural data only	LCE	0.510	0.570	0.535	0.405	0.479	0.426
	XGBoost	0.418	0.509	0.444	0.512	0.545	0.494
	DecisionTree	0.366	0.358	0.357	0.355	0.364	0.356
	RandomForest	0.437	0.582	0.493	0.376	0.539	0.441
Neural PCA data+ViT PCA embedding (InteBOMB)	LCE	0.495	**0.600**	**0.536**	**0.491**	**0.648**	**0.554**
	XGBoost	**0.448**	**0.552**	**0.466**	**0.662**	**0.709**	**0.636**
	DecisionTree	0.359	**0.382**	**0.367**	**0.575**	**0.612**	**0.588**
	RandomForest	**0.449**	**0.600**	**0.508**	**0.545**	**0.685**	**0.572**

WFCI dataset trials 0–548, with a 7:3 train-test split. "stimTime" contains eight classes and "spoutTime" contains seven classes. P, precision; R, recall; F1, F1 score. The numbers in bold indicate improved results.

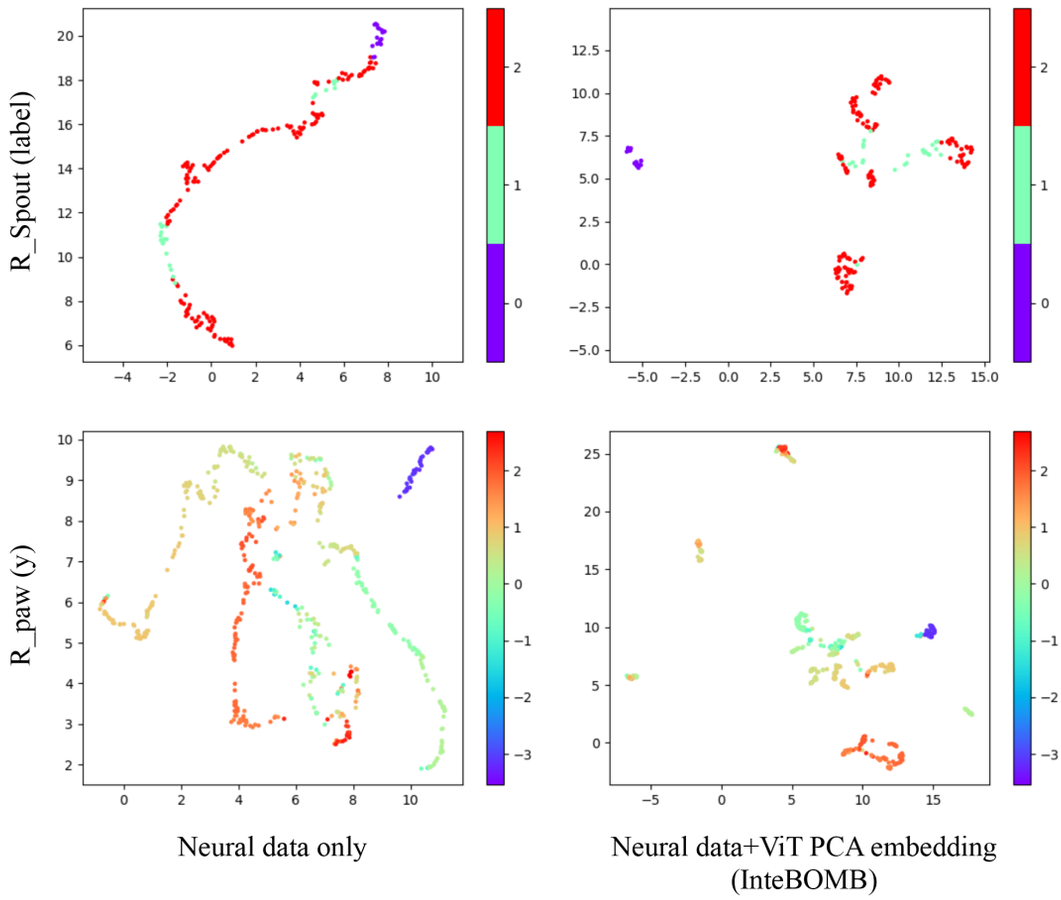

Figure 8 Comparison of behavior clustering with and without visual features

R_Spout (label) contains three classes, R_paw (y) is continuous and ranges from −4 to 3.

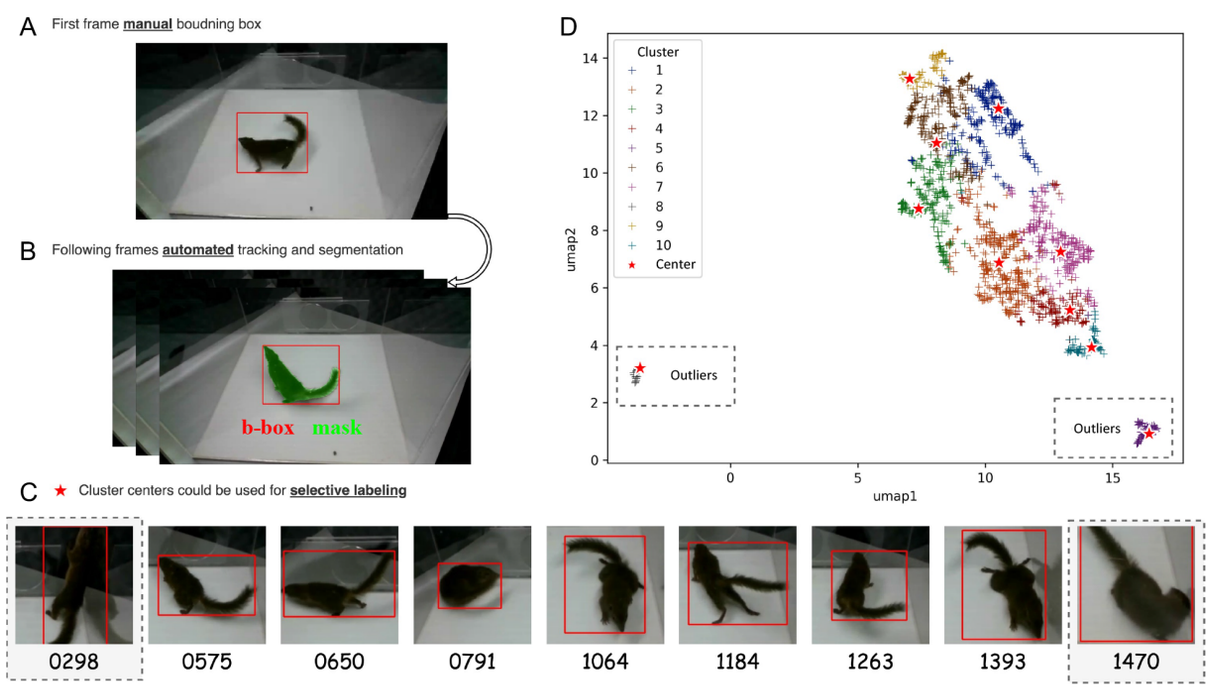

Figure 9 Automated tracking and selective labeling for a newly introduced species, the tree shrew, in a laboratory setting

A–C: General trackers only require first frame manual annotation to generate bounding boxes (A) and masks (B) across subsequent frames (C, similar to Figure 3). D: UMAP visualization of vectorized embeddings (similar to Figure 7), where cluster centers can be used for selective labeling to train further models. Note: Two outlier clusters (dashed boxes) may require more focused labeling. Frame numbers are shown at 6 FPS, as segmentation and embedding extraction were performed at this rate.

namely, automated labeling suggestion and joint behavior analysis. Based on these back-end algorithms, InteBOMB offers a user-friendly interface with four key functional modules —data adapter, model manager, exception handler and behavior analyzer —ensuring interoperability and accessibility for both researchers and the broader open-source community.

Quantifying multi-animal video recordings presents substantial challenges, particularly due to frequent interactions that introduce occlusions and ambiguous tracklet associations, especially among visually similar animals (Lauer et al., 2022). The ongoing debate regarding the suitability of top-down versus bottom-up approaches for behavior quantification remains unresolved. InteBOMB was designed to accommodate both paradigms, allowing users to choose and compare their relative efficacy. Notably, the latest DeepLabCut single and multi-animal code framework (DLCv3) supports top-down, bottom-up, and a state-of-the-art "hybrid" approach, called bottom-up conditional top-down (BUCTD) (Lauer et al., 2022; Zhou et al., 2023). While top-down approaches have been empirically validated as more efficient and accurate in common experiment settings, where animals are moderate in number and sufficiently large within the video frame (Pereira et al., 2022), our evaluations on mouse datasets further substantiated this point (Figure 1; Tables 2, 3). Beyond initial object tracking, multi-object tracking (MOT) and reidentification models remain critical for assigning consistent identities to individual animals. Previous studies indicate that top-down tracking models outperform bottom-up ID models, achieving 3–16× faster processing speeds without compromising accuracy (Pereira et al., 2022). Therefore, we believe that reformative generic object tracking models in a top-down fashion, combined with appropriate ID models, will have a wide range of applications for multi-species, multi-animal behavior analysis in the near future.

InteBOMB provides universal API support for HDF5 and CSV file formats, and it is accessible through both GUI and CLI, facilitating seamless integration into existing computational ethology workflows. Several potential future directions include: (i) Expanding large-scale FAIR datasets to support diverse species and experiment settings, fostering broader applications in computational ethology; (ii) Developing an end-to-end unified framework that integrates tracking, segmentation, and pose estimation within a single foundation model for multi-animal analysis; (iii) Extending research to emerging 3D animal pose estimation, using multi-view calibrated cameras (Han et al., 2024); (iv) Enhancing robust real-time performance to benefit closed-loop behavioral experiments (Lauer et al., 2022), with optimizations for CPU-based multi-processing and multi-threading (Abdulhussain et al., 2022; Flayyih et al., 2024; Mahmmod et al., 2024) and GPU-accelerated implementation (Biderman et al., 2024; Pereira et al., 2022); (v) Optimizing joint analysis and modeling methods to accommodate heterogeneous visual recording and neural signal modalities, such as radio frequency identification devices (RFID), depth, or infrared cameras (Pereira et al., 2020). By integrating advanced tracking methodologies, adaptive learning strategies, and multi-modal behavior quantification, InteBOMB provides an extensible foundation for future advancements in computational ethology and neuroscience research.

DATA AVAILABILITY

Our open-source InteBOMB workflow is maintained on a GitHub repository: https://github.com/JackieZhai/InteBOMB. The six mouse datasets and six non-human primate datasets can also be found at GitHub: https://github.com/ JackieZhai/awesome-behavior-datasets. The tree shrew demo video used as an example is available at: https://www.zoores.ac.cn/en/supplement/bb843abe-3823-434a-b7e4-eabf2fba0a1d.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The SiamBOMB (CN111582214A) patent was invented by X.C., H.Z., L.J.S., Q.W.X., and H.H., and was also submitted by the Institute of Automation, Chinese Academy of Sciences.

AUTHORS' CONTRIBUTIONS

H.H. and X.C. conceptualized the research. H.Z. designed the InteBOMB workflow. H.Z., H.Y.Y., and J.Y.Z. developed the software and performed the experiments. Q.W.X. and L.J.S. provided technical support for the software and datasets. H.Z., H.Y.Y., and J.L. wrote the manuscript. All authors read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We thank the Transdisciplinary Platform of Functional Connectome and Brain-inspired Intelligence, Huairou Science City, Beijing and the Microscopic Technology and Analysis Center, Institute of Automation, Chinese Academy of Sciences, for providing technical support and equipment resources.

REFERENCES

Abdulhussain SH, Mahmmod BM, Baker T, et al. 2022. Fast and accurate computation of high-order Tchebichef polynomials. *Concurrency and Computation: Practice and Experience*, **34**(27): e7311.

Anderson DJ, Perona P. 2014. Toward a science of computational ethology. *Neuron*, **84**(1): 18–31.

Batty E, Whiteway MR, Saxena S, et al. 2019. BehaveNet: nonlinear embedding and Bayesian neural decoding of behavioral videos. *In*: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc. , 1407.

Bhat G, Danelljan M, Van Gool L, et al. 2020a. Know your surroundings: exploiting scene information for object tracking. *In*. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 205–221.

Bhat G, Lawin FJ, Danelljan M, et al. 2020b. Learning what to learn for video object segmentation. *In*. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 777–794.

Biderman D, Whiteway MR, Hurwitz C, et al. 2024. Lightning Pose: improved animal pose estimation via semi-supervised learning, Bayesian ensembling and cloud-native open-source tools. *Nature Methods*, **21**(7): 1316–1328.

Bohnslav JP, Wimalasena NK, Clausing KJ, et al. 2021. DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels. *eLife*, **10**: e63377.

Chen X, Zhai H, Liu DQ, et al. 2020. SiamBOMB: a real-time ai-based system for home-cage animal tracking, segmentation and behavioral analysis. *In*: Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama: ijcai. org, 5300–5302.

Cui YT, Jiang C, Wu GS, et al. 2024. MixFormer: end-to-end tracking with iterative mixed attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **46**(6): 4129–4146.

Danelljan M, Van Gool L, Timofte R. 2020. Probabilistic regression for visual tracking. *In*: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 7181–7190.

Desai N, Bala P, Richardson R, et al. 2023. OpenApePose, a database of annotated ape photographs for pose estimation. *eLife*, **12**: RP86873.

Evans JA, Foster JG. 2011. Metaknowledge. *Science*, **331**(6018): 721–725.

Fauvel K, Masson V, Fromont É, et al. 2019. Towards sustainable dairy

management - a machine learning enhanced method for estrus detection. *In*: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage: ACM, 3051–3059.

Flayyih WN, Al-Sudani AH, Mahmmod BM, et al. 2024. High-performance krawtchouk polynomials of high order based on multithreading. *Computation*, **12**(6): 115.

Han Y, Chen K, Wang Y, et al. 2024. Multi-animal 3D social pose estimation, identification and behaviour embedding with a few-shot learning framework. *Nature Machine Intelligence*, **6**(1): 48–61.

Han YN, Huang K, Chen K, et al. 2022. MouseVenue3D: a markerless three-dimension behavioral tracking system for matching two-photon brain imaging in free-moving mice. *Neuroscience Bulletin*, **38**(3): 303–317.

Hu WM, Wang Q, Zhang L, et al. 2023a. SiamMask: a framework for fast online object tracking and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**(3): 3072–3089.

Hu YJ, Ferrario CR, Maitland AD, et al. 2023b. *LabGym*: quantification of user-defined animal behaviors using learning-based holistic assessment. *Cell Reports Methods*, **3**(3): 100415.

Javed S, Danelljan M, Khan FS, et al. 2023. Visual object tracking with discriminative filters and siamese networks: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**(5): 6552–6574.

Kennedy A. 2022. The what, how, and why of naturalistic behavior. *Current Opinion in Neurobiology*, **74**: 102549.

Kirillov A, Mintun E, Ravi N, et al. 2023. Segment anything. *In*: Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 3992–4003.

Lauer J, Zhou M, Ye SK, et al. 2022. Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods*, **19**(4): 496–504.

Li CX, Xiao ZF, Li YR, et al. 2023. Deep learning-based activity recognition and fine motor identification using 2D skeletons of cynomolgus monkeys. *Zoological Research*, **44**(5): 967–980.

Liu DQ, Li WF, Ma CY, et al. 2020. A common hub for sleep and motor control in the substantia nigra. *Science*, **367**(6476): 440–445.

Liu MS, Gao JQ, Hu GY, et al. 2022. MonkeyTrail: a scalable video-based method for tracking macaque movement trajectory in daily living cages. *Zoological Research*, **43**(3): 343–351.

Luxem K, Sun JJ, Bradley SP, et al. 2023. Open-source tools for behavioral video analysis: Setup, methods, and best practices. *eLife*, **12**: e79305.

Mahmmod BM, Flayyih WN, Abdulhussain SH, et al. 2024. Performance enhancement of high degree Charlier polynomials using multithreaded algorithm. *Ain Shams Engineering Journal*, **15**(5): 102657.

Marks M, Jin QH, Sturman O, et al. 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.

Mathis A, Mamidanna P, Cury KM, et al. 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, **21**(9): 1281–1289.

Mayer C, Danelljan M, Bhat G, et al. 2022. Transforming model prediction for tracking. *In*: Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 8721–8730.

Mayer C, Danelljan M, Pani Paudel D, et al. 2021. Learning target candidate association to keep track of what not to track. *In*: Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 13424–13434.

Musall S, Kaufman MT, Juavinett AL, et al. 2019. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, **22**(10): 1677–1686.

Oquab M, Darcet T, Moutakanni T, et al. 2024. DINOv2: learning robust visual features without supervision. *Transactions on Machine Learning Research*. **2024**.

Panadeiro V, Rodriguez A, Henry J, et al. 2021. A review of 28 free animal-tracking software applications: current features and limitations. *Lab Animal*, **50**(9): 246–254.

Paul M, Danelljan M, Mayer C, et al. 2022. Robust visual tracking by segmentation. *In*: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 571–588.

Pereira TD, Shaevitz JW, Murthy M, et al. 2020. Quantifying behavior to understand the brain. *Nature Neuroscience*, **23**(12): 1537–1549.

Pereira TD, Tabris N, Matsliah A, et al. 2022. SLEAP: a deep learning system for multi-animal pose tracking. *Nature Methods*, **19**(4): 486–495.

Romero-Ferrero F, Bergomi MG, Hinz RC, et al. 2019. idtracker. ai: tracking all individuals in small or large collectives of unmarked animals. *Nature Methods*, **16**(2): 179–182.

Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**: 53–65.

Salem G, Krynitsky J, Hayes M, et al. 2019. Three-dimensional pose estimation for laboratory mouse from monocular images. *IEEE Transactions on Image Processing*, **28**(9): 4273–4287.

Schneider S, Lee JH, Mathis MW. 2023. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, **617**(7960): 360–368.

Sturman O, Von Ziegler L, Schläppi C, et al. 2020. Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. *Neuropsychopharmacology*, **45**(11): 1942–1952.

Tillmann JF, Hsu AI, Schwarz MK, et al. 2024. A-SOiD, an active-learning platform for expert-guided, data-efficient discovery of behavior. *Nature Methods*, **21**(4): 703–711.

Walter T, Couzin ID. 2021. TRex, a fast multi-animal tracking system with markerless identification, and 2D estimation of posture and visual fields. *eLife*, **10**: e64000.

Wang JD, Sun K, Cheng TH, et al. 2021. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**(10): 3349–3364.

Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**: 160018.

Yang S, Chen ZY, Liang KW, et al. 2023. BARN: Behavior-Aware Relation Network for multi-label behavior detection in socially housed macaques. *Zoological Research*, **44**(6): 1026–1038.

Yang TY, Chan AB. 2021. Visual tracking via dynamic memory networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**(1): 360–374.

Yao Y, Bala P, Mohan A, et al. 2022. OpenMonkeyChallenge: dataset and benchmark challenges for pose estimation of non-human primates. *International Journal of Computer Vision*, **131**(1): 243–258.

Yao YG, Lu L, Ni RJ, et al. 2024. Study of tree shrew biology and models: A booming and prosperous field for biomedical research. *Zoological Research*, **45**(4): 877–909.

Ye SK, Filippova A, Lauer J, et al. 2024. SuperAnimal pretrained pose estimation models for behavioral analysis. *Nature Communications*, **15**(1): 5165.

Yu H, Xu YF, Zhang J, et al. 2021. AP-10K: a benchmark for animal pose estimation in the wild. *In*: Proceedings of the 35th Conference on Neural Information Processing Systems. NeurIPS.

Zhou M, Stoffl L, Mathis MW, et al. 2023. Rethinking pose estimation in crowds: overcoming the detection information bottleneck and ambiguity. *In*: Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 14643–14653.