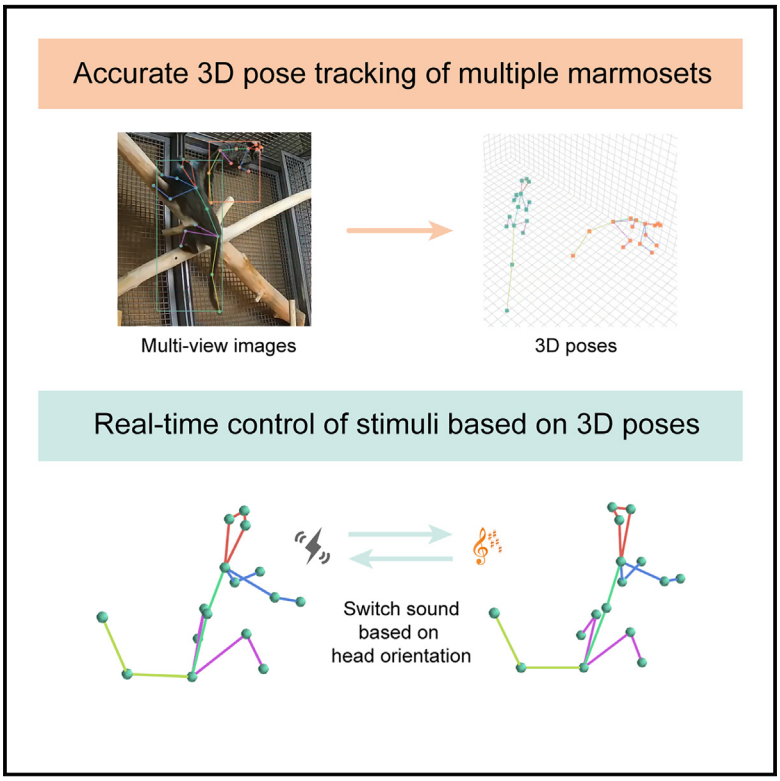
Cell Reports
Methods

Article

A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates

Graphical abstract



Authors

Chaoqun Cheng, Zijian Huang, Ruiming Zhang, Guozheng Huang, Han Wang, Likai Tang, Xiaoqin Wang

Correspondence

xiaoqin.wang@jhu.edu

In brief

Cheng et al. present MarmoPose, a deep-learning-based system for the real-time 3D pose tracking of multiple freely moving marmosets. This user-friendly system requires minimal hardware and supports real-time closed-loop experimental control, enabling advanced behavioral studies.

Highlights

- Accurately tracks 3D poses of multiple freely moving marmosets
- Enables real-time closed-loop experimental control based on 3D poses and positions
- Optimizes 3D poses and estimates invisible body locations using a marmoset skeleton model
- Adapts to new environments with minimal modifications

Cheng et al., 2025, Cell Reports Methods *5*, 100986
February 24, 2025
https://doi.org/10.1016/j.crmeth.2025.100986

CellPress

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

Cell Reports Methods

CellPress
OPEN ACCESS

Article

A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates

Chaoqun Cheng,[1,2] Zijian Huang,[1,2] Ruiming Zhang,[1] Guozheng Huang,[1,2] Han Wang,[1] Likai Tang,[1,2] and Xiaoqin Wang[1,2,3,4,*]

[1]Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China
[2]School of Biomedical Engineering, Tsinghua University, Beijing, China
[3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
[4]Lead contact
*Correspondence: xiaoqin.wang@jhu.edu
https://doi.org/10.1016/j.crmeth.2025.100986

MOTIVATION Tracking the three-dimensional (3D) poses of multiple non-human primates (NHPs) is essential for quantifying their behaviors. However, most existing methods are limited to offline processing and lack the capability for closed-loop experiments. The ability to deliver sensory or optogenetic stimuli based on events detected from the 3D positions and poses of NHPs is a powerful tool for studying their behaviors. To address this limitation, we developed MarmoPose, a real-time 3D pose tracking system for multiple marmosets. MarmoPose is capable of accurately tracking the 3D poses of multiple freely moving marmosets in their home cage with minimal hardware requirements and supports real-time closed-loop experimental control.

SUMMARY

The ability to track the positions and poses of multiple animals in three-dimensional (3D) space in real time is highly desired by non-human primate (NHP) researchers in behavioral and systems neuroscience. This capability enables the analysis of social behaviors involving multiple NHPs and supports closed-loop experiments. Although several animal 3D pose tracking systems have been developed, most are difficult to deploy in new environments and lack real-time analysis capabilities. To address these limitations, we developed MarmoPose, a deep-learning-based, real-time 3D pose tracking system for multiple common marmosets, an increasingly critical NHP model in neuroscience research. This system can accurately track the 3D poses of multiple marmosets freely moving in their home cage with minimal hardware requirements. By employing a marmoset skeleton model, MarmoPose can further optimize 3D poses and estimate invisible body locations. Additionally, MarmoPose achieves high inference speeds and enables real-time closed-loop experimental control based on events detected from 3D poses.

INTRODUCTION

The common marmoset (*Callithrix jacchus*) has emerged in recent years as a promising non-human primate model in neuroscience research, offering unique advantages over other animal models. Compared to rodents, marmosets have more complex brain architecture and exhibit closer cognitive ability to humans. Unlike larger primates like macaques, marmosets are easier to breed in captivity and have a shorter developmental stage and faster reproductive cycle.[1,2] Marmosets have been widely used in various fields of scientific research, including vocal and auditory studies,[3–8] visual neuroscience,[9–11] and transgenic studies for disease modeling.[12–16]

Marmosets are particularly suitable for a wide range of behavioral experiments due to their small body size and social behaviors.[17] However, most behavioral experiments on marmosets have been conducted based on manual recordings or with movement constraints.[18–20] Therefore, the ability to automatically capture and quantify behaviors of marmosets in natural environments and social scenarios is highly desired by the marmoset research community. Such a system could be integrated with other experimental methodologies to advance marmoset research. For instance, using quantified behaviors to identify the differences in behavioral phenotypes between normal and genetically modified marmosets could shed light on the relationships between genes and behaviors.

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

In addition, synchronizing behavioral quantification with neural activity recording technologies has the potential to reveal the neural mechanisms underlying behaviors.

In recent years, there has been a rapid development of automated pose tracking systems for animal behavioral studies. DeepLabCut offers multi-animal two-dimensional (2D) pose estimation capabilities[21,22] and can be extended to provide low-latency 3D pose estimation for a single animal.[23–25] SLEAP provides versatile support for multi-animal 2D pose tracking across a variety of network architectures.[26,27] DANNCE enables direct 3D pose estimation for single rodent and can be transferred to other species.[28] MAMMAL provides the capability to capture 3D surface motions of pigs and dogs.[29] Additionally, species-specific systems have been developed, such as OpenMonkeyStudio for macaques,[30] DeepFly3D for *Drosophila*,[31] and FreiPose for rats.[32]

DeepBhvTracking,[33] MarmoDector,[34] and FulMAI[35] are offline tracking systems specifically designed for marmosets. However, the functions of these systems are confined to tracking the positional trajectories of marmosets in 2D or 3D spaces, lacking crucial pose information for behavioral analyses. While DeepLabCut offers 2D pose tracking capability for marmosets, and DANNCE has the potential to be adapted for estimating the 3D poses of single marmoset, a significant amount of new training data is required for new experimental setups, such as adding more marmosets with new identities, which considerably limits their applications in marmoset behavioral experiments.

Due to the highly social nature and rapid movements of marmosets in 3D space, a system capable of tracking 3D poses of multiple marmosets is highly desired. In addition, the ability to control experimental stimuli in real time based on the marmoset's positions or actions[25] would give researchers power to conduct a wider range of behavioral and physiological experiments in freely roaming marmosets. Existing systems are unable to fulfill these specific requirements. In this study, we have developed an efficient and user-friendly real-time 3D pose tracking system for multiple marmosets, which can be flexibly adapted by a wide range of researchers to study the marmoset's natural behaviors.

The MarmoPose described in this report is a deep-learning-based 3D pose tracking system, with minimal hardware requirements, specifically designed for reconstructing the 3D poses of single or multiple marmosets freely moving in their home cage environment. In MarmoPose, multi-view images captured by four (or more) cameras are first processed by deep neural networks to predict the 2D coordinates of 16 body locations of each marmoset. Subsequently, visible 3D body locations are reconstructed using triangulation, while invisible 3D body locations are estimated through a denoising autoencoder (DAE) incorporating a marmoset skeleton model. MarmoPose offers several advantages over existing systems: (1) this is the first system to enable comprehensive 3D pose tracking for multiple marmosets; (2) this system supports real-time closed-loop experimental control based on the 3D poses and positions of marmosets, which could be integrated with other experimental functions, including stimulus playback and neural recording; (3) this system employs a marmoset skeleton model for 3D coordinate optimization, thereby improving the precision of the reconstructed 3D poses and rendering it possible to estimate invisible body locations in the cameras' blind spots; (4) this system is flexible, as each module can be independently modified to accommodate new experimental setups, such as varying the number of marmosets in the cage or different obstacle configurations; and (5) this system is designed for user-friendly deployment in a typical marmoset family cage (0.7 × 1 × 0.8 m) without additional modifications and can, therefore, be easily adapted to other housing or experimental environments.

RESULTS

Overview of the MarmoPose system

MarmoPose is a 3D pose tracking system specifically designed for both single and multiple marmosets. It processes video streams from multiple cameras and outputs the estimated body locations of marmosets in a 3D space. The system was developed for a typical marmoset family cage (1 × 0.7 × 0.8 m) with four video cameras mounted on the upper corners, as shown in Figure 1A. Marmosets can freely move around in the cage, which is equipped with wooden perches, wire mesh platforms, and small sticks to encourage naturalistic behaviors. Notably, the cameras are fixed on the four top corners inside the home cage so that such a setup does not require any modifications of the cage and thus can be easily deployed in other housing cages.

We selected 16 locations of the body to capture the posture of a marmoset (head, left/right ear, neck, spinemid, left/right elbow/hand/knee/foot, tailbase, tailmid, and tailend). As annotated in Figure 1B, each dot represents one of the body locations, and the lines represent the marmoset skeleton. Invisible body locations from this view are indicated by transparent dots with a white outline. In scenarios involving multiple marmosets, maintaining consistent identification of each individual across different camera views is crucial for accurate 3D triangulation. To ensure reliable identification of individuals, we applied a harmless dye to their ears to clearly distinguish them (for n marmosets, $n - 1$ are marked by different colors, while one is unmarked). For instance, the marmoset marked with blue dye is correspondingly annotated with blue in Figure 1B.

Figure 1C illustrates the workflow of MarmoPose, consisting of two main stages. In the first stage, 2D predictions were generated for each video. For images containing multiple marmosets, we first trained a detection model, adapted from RTMDet,[36] to detect the bounding box as well as the identity of each marmoset (Figure 1D). Subsequently, we trained a pose estimation model, adapted from RTMPose,[37] to predict 16 body locations based on the images cropped around these bounding boxes (Figure 1E). We adopted the two-stage approach due to its flexibility in accommodating new experimental setups and the relatively small size of marmosets in each camera view. To train these deep neural networks, we labeled the Marmoset3K dataset (see STAR Methods) based on 1,527 images containing one marmoset and 1,646 images

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

CellPress
OPEN ACCESS

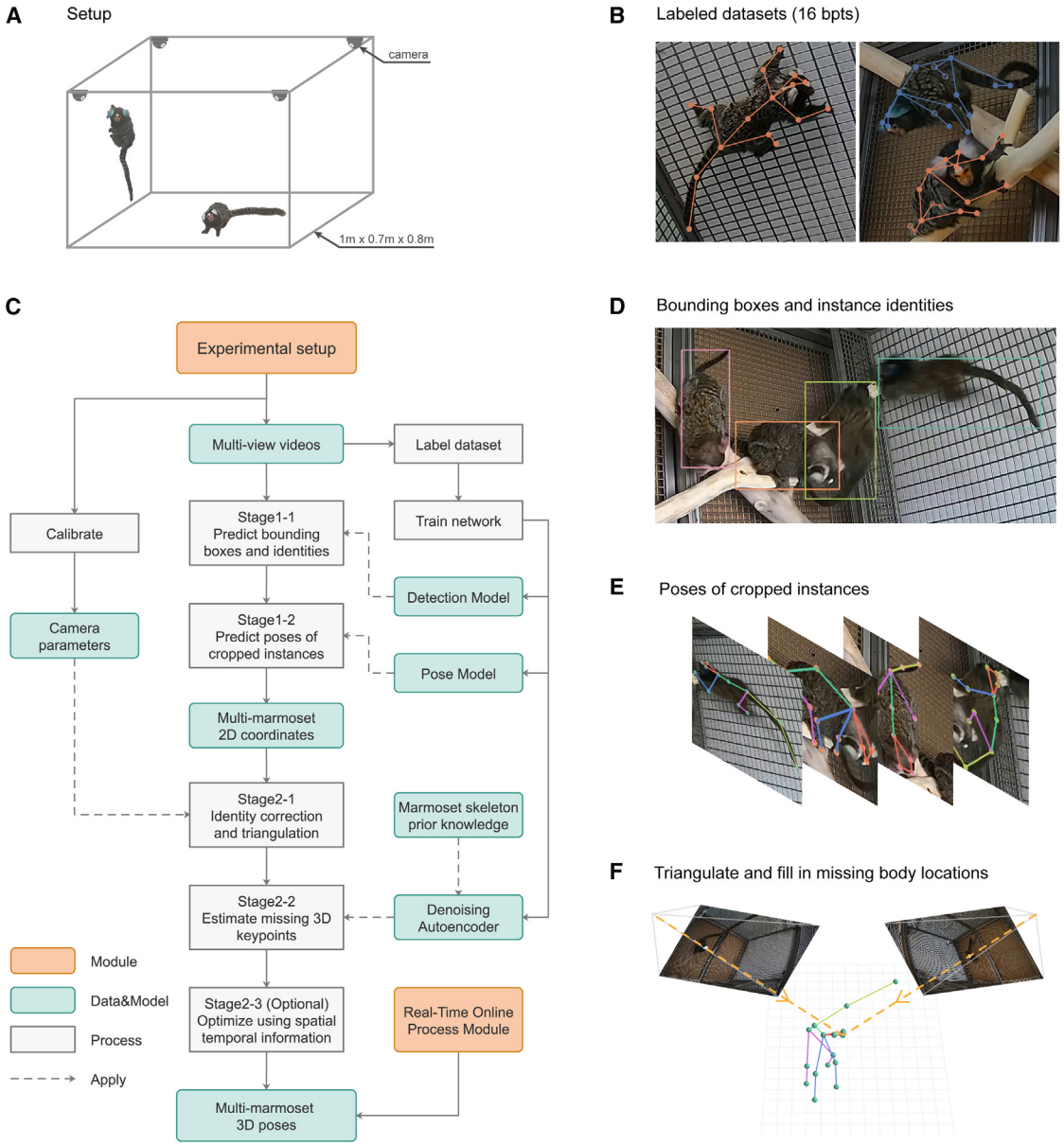


Figure 1. MarmoPose is a complete system for multi-marmoset 3D pose tracking

(A) Experimental setup of MarmoPose. Four cameras are mounted on the upper corners of the marmoset home cage (1 × 0.7 × 0.8 m), and wooden perches and wire mesh platforms are placed in the cage as their daily environment. Marmosets can freely move without any interference.

(B) Example (cropped) images with manual annotations for the Marmoset3K dataset. Each dot represents one of the 16 body locations (transparent dots with white outlines denote invisible body locations from this camera view), and the lines indicate the marmoset skeleton.

(C) Diagram of the whole workflow of MarmoPose.

(D) Example (cropped) image with predicted bounding boxes, where the instance identity of the marmoset is denoted by the color of the bounding box.

(E) Example (cropped) images with predicted body locations, where invisible body locations are omitted.

(F) Illustration of 3D triangulation with 2D predictions.

containing two marmosets across different camera views. Each image was annotated with the 16 body locations and identity information for each instance. In the second stage, instances with the same identity from all camera views were integrated and triangulated into 3D poses using camera parameters (Figure 1F), which were calibrated once the cameras were fixed (see STAR Methods). Subsequently, we employed a DAE[38] integrated with prior knowledge of a marmoset skeleton model to reconstruct invisible body locations (see Figure 3C). Finally, an optional step could be performed to refine the 3D poses further using an iterative optimization method (see STAR Methods). As a result, MarmoPose can accurately

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

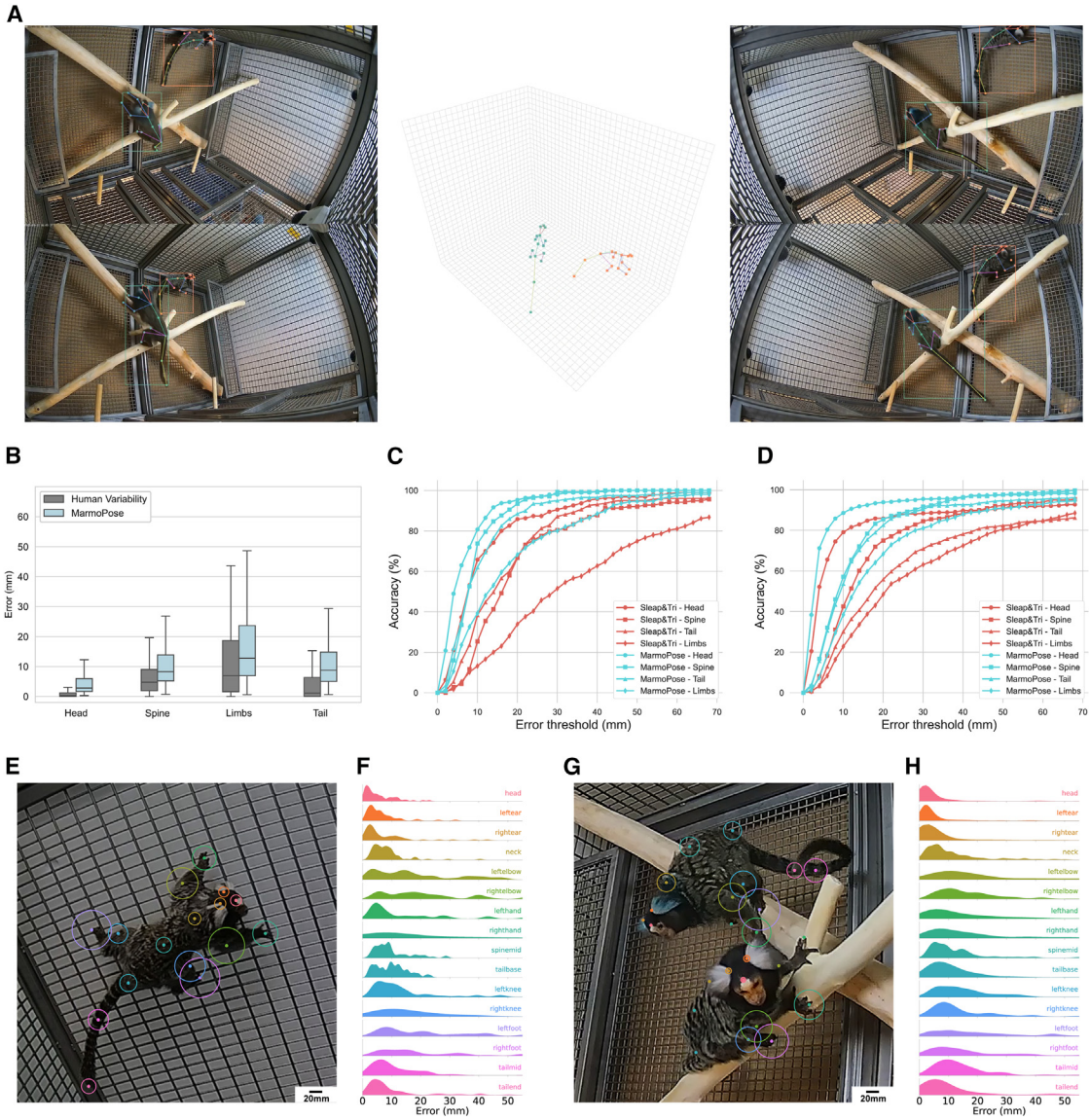


Figure 2. MarmoPose offers greater accuracy and robustness in tracking 3D poses of multiple marmosets compared to previous methods

(A) An example demonstrating one frame of the visualized videos generated by MarmoPose. Four corners show images from 4 camera views with predicted 2D body parts and identities, while the central part shows the reconstructed 3D poses.

(B) Boxplots of 3D Euclidean errors for MarmoPose and human variability evaluated on the Marmoset3D test dataset. 16 body locations are grouped into 4 categories based on their anatomical locations (head: head and left/right ear; spine: neck and spinemid; limbs: left/right elbow/hand/knee/foot; and tail: tailbase, tailmid, and tailend). Human variability indicates the error between the hand-labeled ground truths on the same data from two different people. (n = 258 instances). In the boxplots, the horizontal line within each box represents the median, the box spans the interquartile range (IQR), and whiskers extend to the farthest data points within 1.5×IQR.

(C) 3D reconstruction accuracy as a function of error threshold evaluated on the test data (n = 71 instances) for a single marmoset. Body locations are broken down to the same categories as in (B).

(D) 3D reconstruction accuracy as a function of error threshold evaluated on the test data (n = 152 instances) for paired marmosets.

(E–H) 3D reconstruction errors for a single marmoset and paired marmosets. Each dot represents one of the body locations, with surrounding circles denoting the 75th percentile of the 3D Euclidean errors, and histograms correspond to the full error distribution evaluated on the test data (n = 84 instances for single marmoset dataset; n = 146 instances for paired marmosets dataset).

estimate 3D poses of the marmosets with both visible and invisible body locations. To help visualize the estimated 3D poses, MarmoPose also generates videos combining images from camera views with predicted body locations and identity information (Figure 2A). In addition, MarmoPose provides an online process module to enable real-time experimental control based on the 3D poses of marmosets. This module will be discussed in detail below.

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

MarmoPose offers greater accuracy and robustness in tracking 3D poses of multiple marmosets compared to previous methods

Figure 2A shows a visualized frame generated by MarmoPose. Original images from the four cameras are annotated with the predicted identities, bounding boxes, and 2D poses of two marmosets in the cage. The central plot shows the reconstructed 3D poses of the two marmosets, reflecting the marmosets' 16 body positions and poses as they roam freely in the cage. This figure demonstrates that MarmoPose can track the 3D poses of multiple marmosets (see Videos S1 and S2).

In order to evaluate the accuracy of MarmoPose quantitatively, we constructed a dataset named Marmoset3D (see STAR Methods) by triangulating hand-labeled 2D coordinates from multiple camera views at the same time point. The Marmoset3D dataset contained 522 3D ground-truth instances with 8,352 body locations (16 body locations per each instance), consisting of 140 instances collected from 140 time points with single marmoset (from four marmosets) and 382 instances collected from 191 time points with paired marmosets (from three pairs of marmosets). The hand-labeled 2D coordinates were first annotated by one person and then proofread by a second person to ensure their accuracy. However, because marmosets are relatively small in each camera view and there were no obvious landmarks on their bodies for precise localization, there was inevitably variability in the 2D coordinates annotated by the two people. Theoretically, this human variability is the lower bound of the error of MarmoPose. We computed the human variability by measuring the error between hand-labeled ground truths on the same data from two different people. Figure 2B shows the 3D Euclidean errors of MarmoPose and human variability, in which 16 body locations are grouped into 4 categories based on their anatomical locations for clarity (head: head and left/right ear; spine: neck and spinemid; limbs: left/right elbow/hand/knee/foot; and tail: tailbase, tailmid, and tailend). MarmoPose can achieve comparable 3D errors to human variability: the median errors of human variability for these 4 groups are 0.29 (head), 4.77 (spine), 6.92 (limbs), and 1.12 (tail) mm, and the median errors of MarmoPose for these 4 groups are 2.82 (head), 8.25 (spine), 12.76 (limbs), and 8.75 (tail) mm.

Typically, multi-marmoset 3D pose tracking can be achieved by combining multi-animal 2D pose tracking systems, such as DeepLabCut or SLEAP, with triangulation. However, such an approach faces practical challenges encountered in real experiments, such as estimating occluded body locations, and lacks the necessary robustness and accuracy. MarmoPose is the first comprehensive system designed to track the 3D poses of multiple marmosets and optimized for practical use and higher accuracy. For comparison, we trained and selected the best model using SLEAP on the same dataset to estimate the 2D poses of multiple marmosets, followed by the same 3D reconstruction process to obtain 3D poses. Figures 2C and 2D illustrate the 3D reconstruction accuracy as a function of the error threshold for both MarmoPose and SLEAP combined with triangulation (abbreviated as Sleap&Tri) evaluated on the Marmoset3D test dataset of single marmoset and paired marmosets, respectively, where the accuracy is defined as the percentage of 3D body locations with 3D Euclidean errors below the error threshold, with 16 body locations grouped into four categories. As illustrated in Figures 2C and 2D, MarmoPose shows significantly higher accuracy than Sleap&Tri across all body locations. Quantitatively, the typical length of an adult marmoset body is about 20 cm (excluding the tail, which is also approximately 20 cm long). Using a threshold of 20 mm (about 10% of body length), MarmoPose achieves accuracies of 95% for head, 93% for spine, 68% for limbs, and 89% for tail on the single marmoset test data. In contrast, Sleap&Tri exhibits significantly lower accuracies, with 85% for head, 65% for spine, 35% for limbs, and 65% for tail (Figure 2C). Similarly, on the paired marmosets test data, MarmoPose achieves accuracies of 94% for head, 85% for spine, 68% for limbs, and 83% for tail, while Sleap&Tri exhibits significantly lower accuracies of 83% for head, 76% for spine, 48% for limbs, and 56% for tail (Figure 2D). The improved performance of MarmoPose is primarily attributed to two factors. First, we trained state-of-the-art neural networks for detection (RTMdet)[36] and pose estimation (RTMPose).[37] Due to their elaborate architectures, these networks have demonstrated higher accuracy and faster inference speed in multi-person pose estimation scenarios compared to the series of neural networks used in SLEAP. Second, we incorporated prior knowledge of the marmoset skeleton model to optimize the 3D poses, which includes estimating the missing data caused by occlusion and refining outliers for greater accuracy (see the following section and STAR Methods). These key advantages make MarmoPose more accurate and robust.

Figures 2E and 2G display the 3D reconstruction errors with two representative images. The 75th percentile of the 3D Euclidean errors for each body location is represented by a circle. The error distributions are shown in Figures 2F and 2H. Generally, body locations on the head have smaller errors due to the existence of a boundary between brown and white hairs, while body locations on the limbs have larger errors due to high degrees of freedom and frequent occlusion.

MarmoPose employs a collection of post-processing algorithms to further improve accuracy and robustness

Accurately reconstructing the 3D poses of multiple marmosets typically involves three common challenges: (1) marmoset misidentification, (2) missing values in the 3D poses due to occlusions, and (3) inaccuracy in some predicted body locations. MarmoPose mitigates each of these issues using a collection of specialized post-processing algorithms.

First, Figure 3A illustrates an example of marmoset misidentification and correction: in camera 2, the identities of the two marmosets are incorrectly recognized due to occlusion of their marked ears, but the identities are correctly identified in the remaining cameras (only camera 4 is shown here), and this kind of misidentification can be corrected through post-processing. To correct the identities, MarmoPose groups the detected bounding boxes across multiple videos using epipolar geometry constraints and then addresses potential misidentifications with low confidence scores by integrating the initially predicted IDs with the newly grouped IDs (see STAR Methods). This approach increases the identification accuracy from 95.1% to 96.5% for two marmosets and from 93.2% to 95.0% for four marmosets, as depicted in Figure 3B.

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

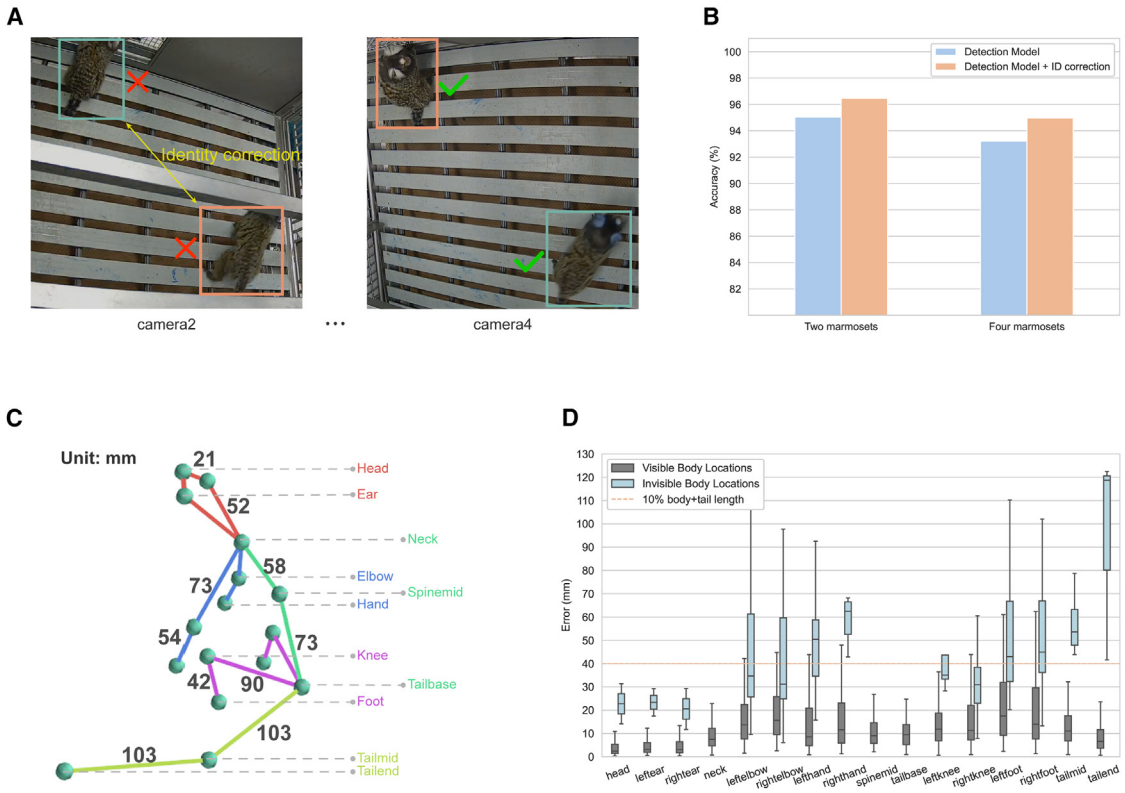


Figure 3. MarmoPose employs a collection of post-processing algorithms to further improve accuracy and robustness

(A) An example illustrating identity misidentification: in camera 2, the identities of the two marmosets are incorrectly recognized due to occlusion of their marked ears, but the identities are correctly identified in the remaining cameras (only camera 4 is shown here), and this kind of misidentification can be corrected through post-processing.

(B) Identification accuracy of the detection model with and without ID correction algorithm, evaluated on the paired marmosets dataset.

(C) Illustration of the marmoset skeleton model used as one of the terms in the loss function of DAE to guide the reconstruction of missing body locations. Numbers (mm) indicating the median length of the distances between two body locations measured on real marmosets, with different weights assigned based on their degrees of freedom.

(D) Boxplots of 3D Euclidean errors in test data for visible and invisible body locations. Note that neck, spinemid, and tailbase have no invisible body locations because they are used to normalize the poses. In the boxplots, the horizontal line within each box represents the median, the box spans the interquartile range (IQR), and whiskers extend to the farthest data points within 1.5×IQR.

Second, in the scenarios involving multiple marmosets freely moving in the home cage, some of their body locations will inevitably be self-occluded or occluded by other animals and objects like logs or shelves in the cage, leading to missing data in the reconstructed 3D poses. Inspired by a previous study on 2D human pose estimation,[39] we trained a DAE to reconstruct the missing 3D body locations, which receives the 3D coordinates of 16 body locations with missing data as input and outputs the estimated complete coordinates. We trained the model by randomly masking some body locations to simulate inputs with missing data based on the Marmoset3D dataset (see STAR Methods).

In order to guide the DAE to estimate missing coordinates better, we added an extra loss term to constrain the lengths between two body locations in the reconstructed 3D poses based on prior knowledge of marmosets. We measured the median length of distances between two body locations on three normal adult marmosets to form a skeleton model (Figure 3C). Incorporating this prior knowledge of marmosets into the model enhances its ability to capture the underlying structure of marmosets. As a result, the DAE can more accurately reconstruct the missing body locations, even with a limited amount of training data. Figure 3D shows the 3D error distributions for each body location when it is visible (computed by triangulation of multiple 2D predictions) or invisible (estimated by DAE) evaluated on the Marmoset3D test data. Note that neck, spindmid, and tailbase do not have corresponding invisible values because they are used to normalize the poses during the application of the DAE. It can be observed that for body locations on the head or spine, the errors of estimated coordinates are below 10% of the body and tail lengths (40 mm), whereas the errors of tail positions are the highest. This is reasonable because marmosets' tails are long and flexible, making their precise locations more difficult to predict. Although the missing data predicted by the DAE might not achieve the accuracy level of visible body locations, they may still offer a reasonable estimation of the missing body locations within a controlled experimental setting.

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

CellPress
OPEN ACCESS

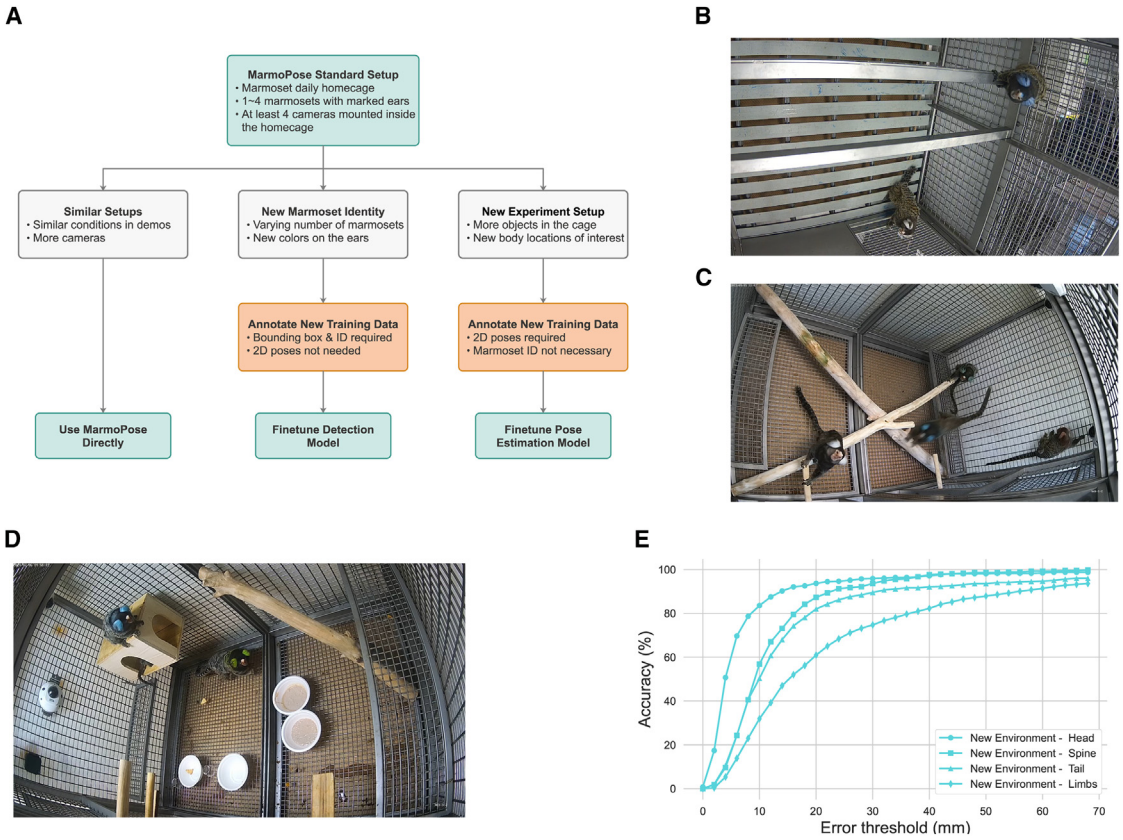


Figure 4. MarmoPose can be easily adapted to new experimental setups with minimal modifications

(A) Schematic illustrating the process of adapting MarmoPose to new environments. In similar setups, MarmoPose could be used directly. For new experiment setups, the detection model and the pose estimation model can be fine-tuned independently to adapt to new conditions with minimal additional training data.

(B) An example of similar setups. Two marmosets are in a home cage approximately half the size of the one used in the standard setup.

(C) An example of new marmoset identities: among the four marmosets, two of them (the green one and the red one) are not included in the training data. Therefore, the detection model needs to be fine-tuned to recognize them correctly.

(D) An example of new experimental setup: additional objects are placed in the home cage to reflect their daily environment. Therefore, the pose estimation model needs to be fine-tuned to further improve accuracy.

(E) The 3D reconstruction error of the fine-tuned model under a new environment.

The third issue involves inaccuracy in the 2D poses predicted by the pose estimation model, especially when multiple marmosets are close to each other. Simply adding more training data to improve the model's accuracy is not sufficient to effectively address this problem. Inspired by a previous study,[24] we incorporated an optimization step at the final stage that refines the 3D poses using spatial and temporal constraints (see STAR Methods). This approach reduces the influence of outliers, generates smoother coordinates, and improves the overall accuracy and plausibility of the final 3D poses.

MarmoPose can be easily adapted to new experimental setups with minimal modifications

In practical behavioral experiments with marmosets, researchers may want to place the animals into customized environments rather than just standard setups. This typically involves changes in two aspects: adding new marmosets with different identities or adding more objects to the home cage for experimental purposes. Unlike previous animal pose tracking methods that require a comparable amount of training data to retrain the entire model from scratch, the two-stage prediction design of MarmoPose allows for easy adaptation to such new environments with minimal additional training data for fine-tuning.

Figure 4A summarizes the steps required for researchers to employ MarmoPose in new experimental setups. If the new setup is similar to the standard configuration, then MarmoPose can be used directly without additional modifications. Figure 4B provides an example: two marmosets are in a home cage approximately half the size of the standard setup, with no new identities introduced. Video S3 demonstrates the seamless application of MarmoPose in this scenario. For more complex setups, additional fine-tuning data might be necessary. In the setup shown in Figure 4C, a family of four marmosets are in the home cage, with two of them (the green one and the red one) not included in the original training data; naturally, the detection model is unable to recognize their identities by default. In this case, only a small amount of training data needs to be annotated, requiring just the bounding boxes and identities (pose

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

annotations for each instance can be skipped) to fine-tune the detection model. In fact, we annotated only 100 images (which took about 1 h) to fine-tune the detection model, achieving comparable accuracy to the standard setup (see Video S4). In another scenario, shown in Figure 4D, a subset of marmosets from Figure 4C are placed in a more complex environment with additional objects. In this case, the marmosets might interact with the objects in ways that exceed the generalization capability of the pose estimation model, necessitating more training data to further improve the accuracy. Actually, we annotated 100 complete images to fine-tune the pose estimation model. As shown in Figure 4E, the average 3D reconstruction error in this new environment increased slightly by 0.8 mm compared to the original setup. This increase is acceptable given the higher complexity of the environment, and accuracy can be further improved with more training data for fine-tuning. Video S5 provides a demonstration of MarmoPose in this environment. These three examples highlight the flexibility and adaptability of MarmoPose to new experimental setups with minimal modifications. Fine-tuning the models typically requires only a few hundred new training data, which can be annotated in a few hours by researchers familiar with the annotation pipeline.

MarmoPose enables real-time experimental control based on events detected from 3D poses

Real-time experimental control triggered by events detected from 3D poses is highly desired by marmoset and other non-human primate researchers, as it enables a fully automated behavioral experiment pipeline. To meet this need, MarmoPose provides a ready-to-use online processing module that handles images from multiple real-time video streams and outputs 3D poses frame by frame with minimal latency. It supports user-customized event triggers based on the 3D poses and positions, allowing real-time modification of experimental stimuli.

To achieve this goal, we first adopted multi-processing and multi-threading to handle different tasks in parallel. As the workflow chart shows in Figure 5A, process-1 (prediction process) reads images from real-time video streams (cached by multiple threads) and performs 2D detection and 3D reconstruction. The images and pose data are sent to process-2 (display process) for visualization and process-3 (main process) for triggering customized events and performing experimental control. Executing different tasks in separate processes ensures that resource-intensive tasks like neural network inference and 3D image rendering do not clock each other, thereby minimizing overall system latency. Secondly, we utilized TensorRT to deploy the PyTorch model in half-precision, which significantly boosts the inference speed with minimal accuracy loss.

To evaluate the real-time performance of MarmoPose, we first benchmarked the combined inference speed of the detection model and pose estimation model. The benchmarks were performed across various numbers of marmosets in the videos and different inference batch sizes. As illustrated in Figure 5B, with a batch size of 4, the original PyTorch model achieved 68 fps for videos with 1 instance, 53 fps for 2 instances, and 39 fps for 4 instances. After the deployment via TensorRT, the inference speed was significantly increased to 112 fps for videos with 1 instance, 100 fps for 2 instances, and 82 fps for 4 instances. To process multiple frames in real time, we combined the frames from different cameras into a pseudo-batch for inference (see STAR Methods). Then, we evaluated the latency distribution of each part across different processes. Figure 5C shows the latency breakdown of the standard setup, which includes 4 cameras ($1,920\times1,080$ resolution) and 2 marmosets present in each video. In the predict process, the mean latency for predicting 2D poses from multiple cameras is 37 ms, with an additional 3 ms required for 3D pose reconstruction. In the display process, data generated by the predict process are received, and the 3D image is rendered, captured, and displayed on the screen for real-time monitoring. The rendering step takes approximately 23 ms, while the display step takes around 5 ms. Consequently, the overall computation latency for each process is under 40 ms, enabling MarmoPose to perform real-time experimental control based on events detected from the 3D poses. It is important to note that deploying the model in half-precision inevitably results in some loss of accuracy. We evaluated the 3D distance errors of both the original and the deployed model across 4 groups of body locations on the Marmoset3D test dataset. The average accuracy loss (less than 7%) is acceptable, given the significant speed improvement (Figure 5D).

We also evaluated the performance of the real-time processing module in a demonstration control experiment: playing different sounds based on the detected gaze direction from 3D poses. When the marmoset looks forward, white noise is played (indicated by a lighting icon), and when the marmoset looks left, music is played (indicated by a note icon). Since we reconstructed the 3D coordinates of the head, left ear, and right ear, we could easily compute the marmoset's gaze direction (see STAR Methods) based on real-time 3D poses and perform corresponding sound control. The poses of the marmoset at different time points are shown in Figure 5E. The marmoset initially looks forward during the first 40 ms and then turns its head to left, causing the sound stimuli to switch from white noise to music accordingly. With an average latency of less than 40 ms, MarmoPose can reliably track such quick movements in real time and trigger events rapidly. This demonstration highlights the powerful real-time processing ability of MarmoPose. In practical applications, researchers can set different trigger events based on the 3D poses and control various types of stimuli, such as sound, image, and more.

MarmoPose enables quantitative analysis of behaviors for multiple marmosets

A significant aspect of neuroscience research involves quantifying animal's natural behaviors.[22,28,30,40] With the accurate 3D poses reconstructed by MarmoPose, researchers can study marmoset behavioral patterns and potential preferences across various experimental setups by extracting features from high-dimensional data using machine learning methods. Here, we demonstrate two typical applications: identifying marmoset behavioral patterns using unsupervised learning and quantifying specific behaviors, like gaze, between marmosets.

With the 3D coordinates of 16 body locations of multiple marmosets reconstructed by MarmoPose, additional features like movement velocity can be easily computed. Generally, more complex behavioral states can be represented by nonlinear

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

CellPress
OPEN ACCESS

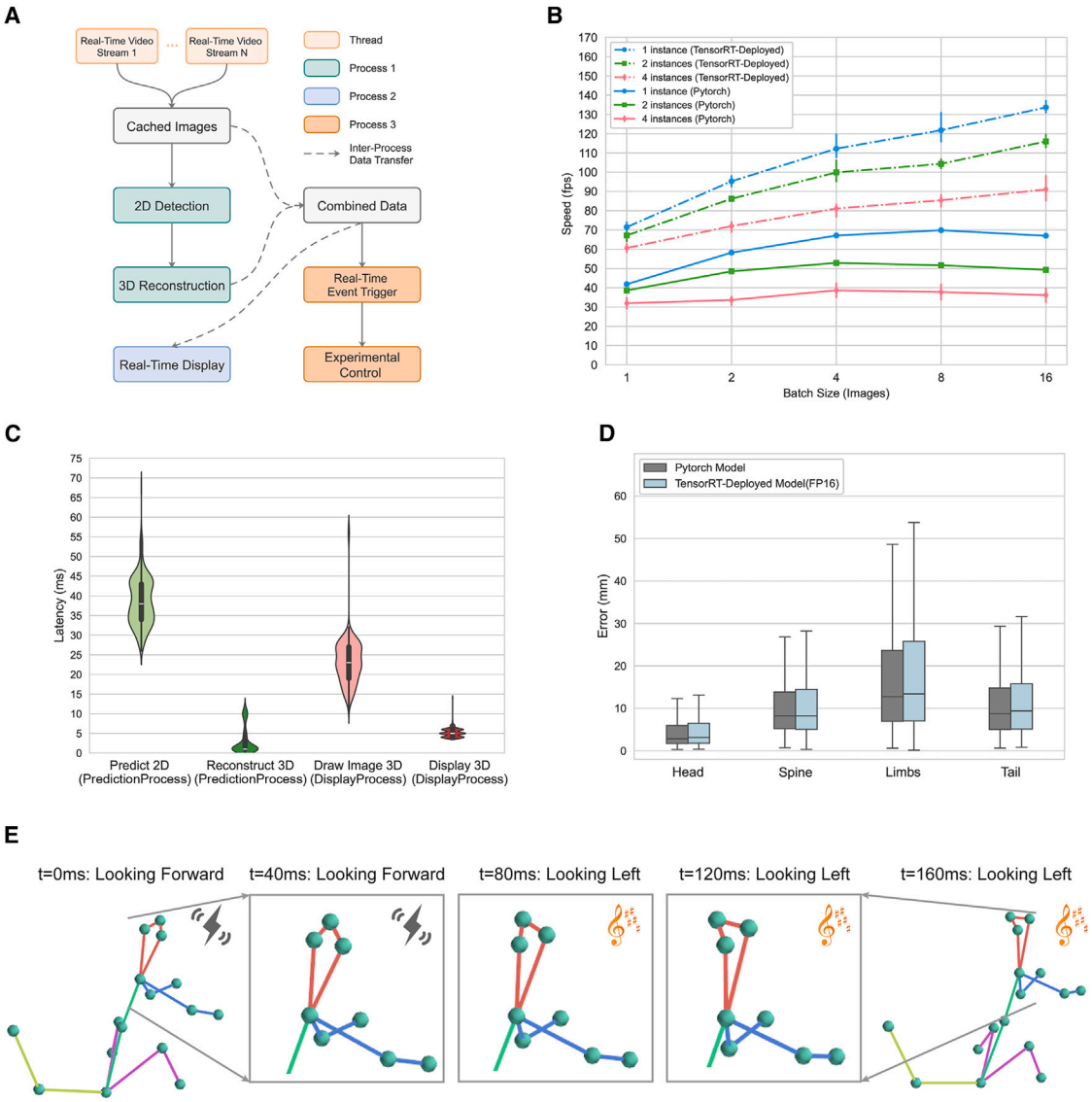


Figure 5. MarmoPose provides an online process module to enable real-time experimental control based on events detected from 3D poses

(A) Schematic of real-time experimental control. Process-1 (predict process) reads images from real-time video streams (cached by multiple threads) and perform 2D detection and 3D reconstruction. The images and pose data are sent to process-2 (display process) for visualization and process-3 (main process) for triggering customized events and performing experimental control.

(B) Inference fps variation across different number of instances in videos, evaluated over varying batch sizes, where each condition is evaluated on four different videos containing 1,000 frames. Solid lines represent the original PyTorch model, while dashed lines denote the model deployed using TensorRT in FP16 mode. Each point represents the mean fps, with error bars indicating 95% confidence interval.

(C) Latency distribution of each part across different processes, with two instances present in the video. In the violin plots, the central white line represents the median latency, the black box spans the interquartile range (IQR), and the upper and lower bounds of each violin represent the estimated density distribution based on kernel density estimation.

(D) Boxplots illustrating the 3D Euclidean errors of both the PyTorch model and the TensorRT-deployed model evaluated on the Marmoset3D test dataset. 16 body locations are grouped into 4 categories based on their anatomical locations. In the boxplots, the horizontal line within each box represents the median, the box spans the interquartile range (IQR), and whiskers extend to the farthest data points within 1.5×IQR.

(E) An example demonstrating real-time event triggering and experimental control: when the marmoset looks forward, white noise is played (indicated by a lighting icon); when the marmoset looks left, music is played (indicated by a note icon). In the example, the marmoset initially looks forward during the first 40 ms (1 frame) and then turns its head to the left, causing the sound stimuli to switch from while noise to music accordingly. The real-time control module enables MarmoPose to perform timely control with minimal latency based on accurately detected events.

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

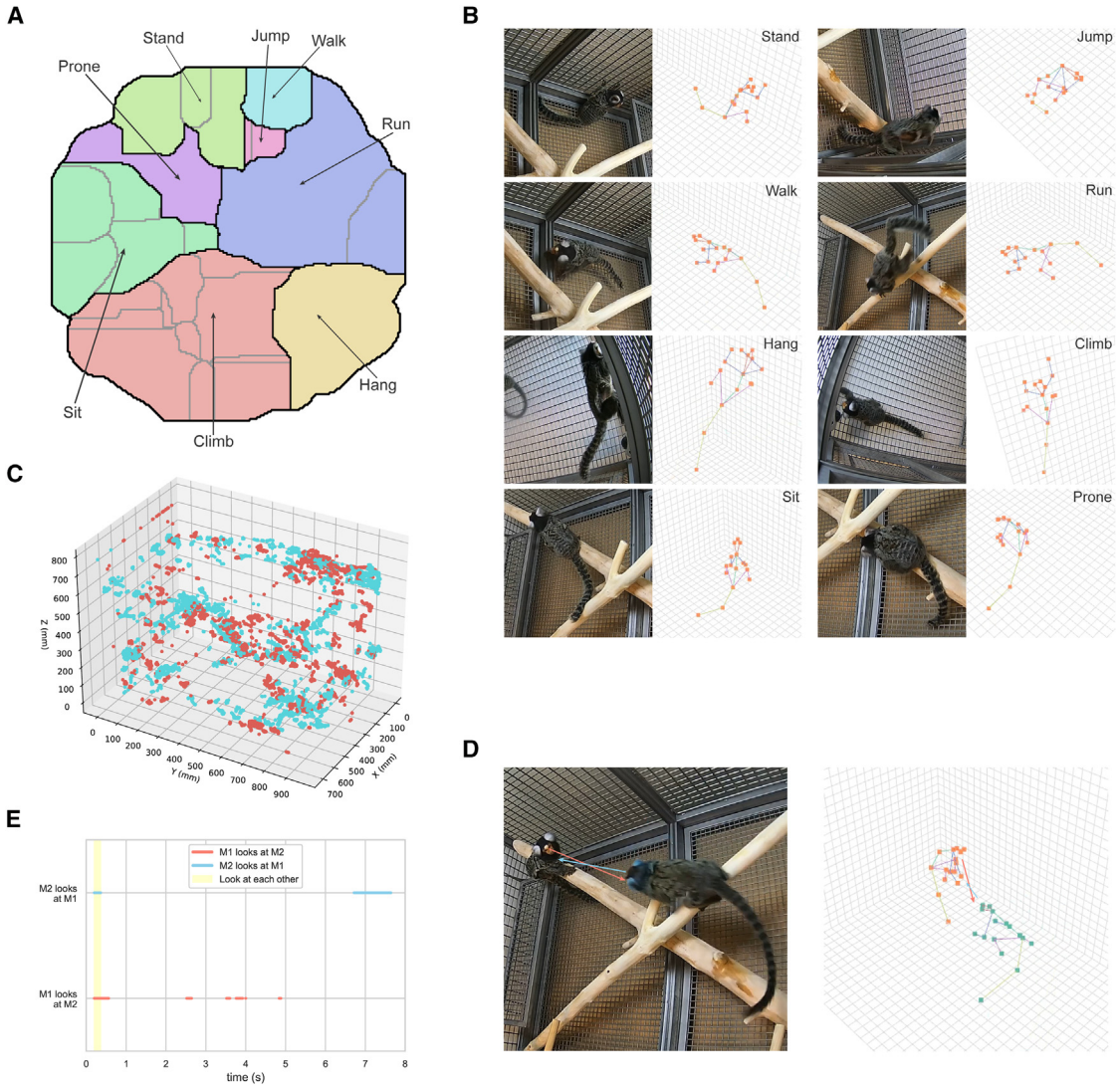


Figure 6. MarmoPose enables quantitative analysis of behaviors for multiple marmosets

(A) Posture clusters generated by applying watershed transform to the t-SNE density map. Each cluster represents a typical marmoset behavior.

(B) An illustration of 8 distinct poses. For each case: left, cropped raw image from one camera; right: reconstructed 3D pose.

(C) The spatial distribution of two freely roaming marmosets (red, marmoset 1; blue, marmoset 2) within a typical home cage over 1 h.

(D) An example of two marmosets looking at each other. Left: cropped raw image from one camera; right: reconstructed 3D poses. Red and blue arrows indicate the gaze direction of each marmoset.

(E) The distribution of mutual gaze between two marmosets over an 8-s time range. M1, marmoset 1; M2, marmoset 2. The red timeline indicates when M1 is looking at M2, and the blue timeline indicates when M2 is looking at M1. Yellow shading highlights the periods when both marmosets are looking at each other.

combinations of these basic features, which can be clustered and recognized through unsupervised learning. In Figure 6A, using the 3D poses of two marmosets freely roaming in a home cage within 1 h, we first reduced the high-dimensional features into 16 dimensions using principal-component analysis (PCA), and then we generated behavioral density maps and identified clusters by applying watershed transform over the density representation of the further reduced t-distributed stochastic neighbor embedding (t-SNE) space (see STAR Methods). Each cluster represents a typical behavior of the marmosets in daily life, including standing, jumping, walking, running, hanging, climbing, sitting, and proning (Figure 6B). This behavior map can further be utilized to characterize the differences in behavior between individual marmosets or for a single marmoset under varying conditions.

Position and head direction are the two most straightforward features we can extract from the original 3D poses. In Figure 6C, we display the positions of two marmosets freely roaming in the home cage, each dot representing a spatial location where the marmoset stays longer than 1 s. We observed that both marmosets prefer to stay on the wooden sticks and wire mesh platforms or climb and stay at the top corners of the home cage. Figure 6D provides an example of two marmosets looking at each other.

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

CellPress
OPEN ACCESS

MarmoPose can accurately detect this behavior by calculating the vector direction from the midpoint between the left and right ears to the head (see STAR Methods). Figure 6E shows the gaze interaction between two marmosets during an 8-s clip, where the red timeline indicates when marmoset 1 is looking at marmoset 2, and the blue timeline indicates when marmoset 2 is looking at marmoset 1. Yellow shading highlights the periods when both marmosets are looking at each other. In general, we found that even paired marmosets rarely look directly at each other: over a 1-h period, marmoset 1 looked at marmoset 2 only 4.1% of the time, and marmoset 2 looked at marmoset 1 only 7.9% of the time, with mutual gaze occurring only 0.3% of the time.

DISCUSSION

Here, we described MarmoPose, which enables real-time 3D pose tracking for multiple freely moving marmosets in a housing cage environment. The primary considerations in designing MarmoPose are to make it efficient, cost effective (using minimal hardware), real time, and easily deployable in typical housing cages. This system is specifically designed for marmosets in order to optimize its performance, but it could be adapted for other large animal species with specific modifications. It has several distinct features comparing to other published pose detection systems for freely moving animals. First, MarmoPose leverages prior knowledge of a marmoset skeleton model to estimate invisible body locations and refine 3D poses. Second, MarmoPose is a user-friendly system that can be deployed in a typical marmoset home cage environment and be easily adapted to new experimental setups with minimal modifications. Third, MarmoPose provides an online process module for users to perform real-time closed-loop experimental control based on the events detected from the 3D poses, which can be integrated with other experimental methodologies, such as stimulus playback and neural recording.

To the best of our knowledge, MarmoPose is the first system to enable practical real-time 3D pose tracking. Previous systems have either lacked the speed necessary for real-time 3D pose tracking or only supported 2D real-time pose tracking. Achieving 3D real-time pose tracking is inherently challenging because it requires synchronizing multiple video streams, managing the heavier computational burden from processing multiple frames simultaneously, and performing complex optimizations for data storage and real-time 3D visualization. To minimize latency, we implemented several key optimizations, including adopting multi-processing and multi-threading to handle different tasks in parallel and deploying the model using TensorRT for accelerated inference. These optimizations ensure that MarmoPose can deliver efficient and reliable performance in real-time applications. While MarmoPose represents a significant advancement in 3D real-time pose tracking, it does have certain limitations. For instance, deploying the models in half-precision results in a slight loss of accuracy. Additionally, some optimization steps are omitted to achieve faster inference speed. However, the accuracy required for real-time experimental control is typically lower than that needed for precise offline analysis, making these trade-offs acceptable in most scenarios.

In scenarios involving multiple marmosets, particularly when they are in close proximity, accurately distinguishing individual body locations is a challenging task. This is a common problem in the field of multi-object pose tracking that has not been addressed very well. Since each marmoset occupies only a small part of the whole image and their rapid movements often lead to motion blur, it is challenging even for human annotators to accurately locate specific body locations. MarmoPose has employed a list of techniques to address this issue, including a DAE integrated with the marmoset skeleton to estimate missing data and optimization of the 3D poses with spatial and temporal constraints. Additionally, a practical solution is to keep a balance between specific experimental demands and system accuracy. In many experimental situations, the focus might be only on a subset of body locations. Therefore, we can choose to ignore the misidentified data and concentrate on the relevant body locations. Furthermore, in social scenarios, the overall interaction of the marmosets might be more significant than the precise poses of each body location. Therefore, it might be more effective to treat marmosets that are close to each other as a single entity without attempting to distinguish every body part clearly.

In the current setup, four cameras are fixed on the upper corners inside the home cage to maximize coverage and minimize maintenance and potential damage by the animals. This is a simple and effective arrangement, which can be easily deployed in other housing cages without extra modifications, retaining minimal computational costs at the same time. While adding more cameras to cover more blind spots is a straightforward solution to reduce invisible body locations and improve system accuracy, it increases computing loads and slows down the processing speed. Practically, the number of cameras used should be determined by the specific experimental demands. The current arrangement with four cameras is sufficient for a range of research purposes. For researchers aiming for higher precision, more cameras can be added to cover more blind spots, which is also supported by MarmoPose. However, it is important to consider that increasing the number of cameras would lead to longer processing time and output latency.

The ability to automatically track the poses of freely moving marmosets in 3D space in real time using MarmoPose could significantly improve studies of natural behaviors in this field. Traditionally, studies on vocal communications[4,41] and visual directional preference [9–11] have relied on manually processed audio and video recordings. MarmoPose could be used to boost these studies by providing more accurate and comprehensive behavioral quantifications. Given their social nature and ease of handling, marmosets are ideal for comparative behavioral experiments with humans. For example, MarmoPose could be used to describe and compare the behavioral evolution and natural preferences between marmosets and humans.[1,17,42] Moreover, due to their relatively high reproductive cycles compared to other primates, marmosets are well suited for transgenic modifications[14,43] and disease modeling.[44–47] Integrating MarmoPose with other methodologies, such as neural recording and optogenetics, offers possibilities to explore neural mechanisms underlying natural behaviors in both normal and genetically modified marmosets.

Limitations of the study

In scenarios involving multiple marmosets, the ears of the marmosets need to be marked by the same color schemes as used in our standard setup for MarmoPose. Alternatively, researchers can fine-tune the detection model to recognize new colors. We adopted this simple yet effective approach because relying solely on spatial and temporal information for consistent individual identification is impractical during prolonged recording without an absolute identifier, like a particular color. Besides, when the marmosets are in close proximity, MarmoPose may fail to accurately detect individual body locations due to indistinguishable boundaries caused by similar hair colors. Although we have employed a series of optimizations algorithms to mitigate this issue, it may still occasionally occur.

When using the online processing module of MarmoPose for real-time experimental control, the system will record raw videos, rendered video of 3D poses, and predicted 2D and 3D pose data. However, due to fluctuations in the real-time video stream and system processing speed, some frames may be skipped to ensure the latest frames are processed in a timely manner, resulting in missing data at certain time points. Additionally, as discussed above, the half-precision TensorRT-deployed model exhibits an average accuracy loss of less than 7%. Therefore, it is recommended to use the online processing module to perform real-time experimental control and rely on the offline processing pipeline for detailed behavioral analyses.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Xiaoqin Wang (xiaoqin.wang@jhu.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The Marmoset3K dataset is publicly available at https://doi.org/10.5281/zenodo.14672425. Other data reported in this study can be shared by the lead contact upon request.
- Code for MarmoPose has been released at https://github.com/Leoswordy/MarmoPose. All the code can also be accessed at https://doi.org/10.5281/zenodo.14672988.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This research was supported by the Tsinghua University Non-Human Primate Research Center (THU-NPRC). We thank Dr. Li Luo and the veterinary and husbandry staff at the THU-NPRC for help with this research. We thank Chenggang Chen and Bing Yuan for their assistance in deploying this system in new setups at Johns Hopkins University and Tsinghua University.

AUTHOR CONTRIBUTIONS

C.C. and X.W. designed the study. C.C. developed the system. Z.H. helped develop the setup for video recording. C.C., R.Z., G.H., H.W., and L.T. labeled the dataset for model training and testing. C.C. and X.W. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Datasets
  - Training 2D detection model and pose estimation model
  - SLEAP models training for comparison
  - Identity correction
  - Camera calibration and coordinate system alignment
  - Triangulation
  - Marmoset skeleton model
  - Pose normalization
  - Denoising autoencoder
  - Real-time control module
  - Model deployment
  - Behavioral mapping and clustering
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2025.100986.

Received: September 16, 2024
Revised: November 28, 2024
Accepted: January 27, 2025
Published: February 17, 2025

REFERENCES

1. Prins, N.W., Pohlmeyer, E.A., Debnath, S., Mylavarapu, R., Geng, S., Sanchez, J.C., Rothen, D., and Prasad, A. (2017). Common marmoset (Callithrix jacchus) as a primate model for behavioral neuroscience studies. J. Neurosci. Methods *284*, 35–46. https://doi.org/10.1016/j.jneumeth.2017.04.004.
2. Schiel, N., and Souto, A. (2017). The common marmoset: an overview of its natural history, ecology and behavior. Dev. Neurobiol. *77*, 244–262.
3. Miller, C.T., and Wang, X. (2006). Sensory-motor interactions modulate a primate vocal behavior: antiphonal calling in common marmosets. J. Comp. Physiol. *192*, 27–38.
4. Miller, C.T., Beck, K., Meade, B., and Wang, X. (2009). Antiphonal call timing in marmosets is behaviorally significant: interactive playback experiments. J. Comp. Physiol. *195*, 783–789.
5. Miller, C.T., Mandel, K., and Wang, X. (2010). The communicative content of the common marmoset phee call during antiphonal calling. Am. J. Primatol. *72*, 974–980.
6. Agamaite, J.A., Chang, C.-J., Osmanski, M.S., and Wang, X. (2015). A quantitative acoustic analysis of the vocal repertoire of the common marmoset (Callithrix jacchus). J. Acoust. Soc. Am. *138*, 2906–2928.
7. Gultekin, Y.B., Hildebrand, D.G.C., Hammerschmidt, K., and Hage, S.R. (2021). High plasticity in marmoset monkey vocal development from infancy to adulthood. Sci. Adv. *7*, eabf2938.
8. Osmanski, M.S., and Wang, X. (2023). Perceptual specializations for processing species-specific vocalizations in the common marmoset (*Callithrix jacchus*). Proc. Natl. Acad. Sci. USA *120*, e2221756120. https://doi.org/10.1073/pnas.2221756120.

9. de Boer, R.A., Overduin-de Vries, A.M., Louwerse, A.L., and Sterck, E.H.M. (2013). The behavioral context of visual displays in common marmosets (Callithrix jacchus). Am. J. Primatol. *75*, 1084–1095.
10. Mitchell, J.F., Reynolds, J.H., and Miller, C.T. (2014). Active vision in marmosets: a model system for visual neuroscience. J. Neurosci. *34*, 1183–1194.
11. Mitchell, J.F., and Leopold, D.A. (2015). The marmoset monkey as a model for visual neuroscience. Neurosci. Res. *93*, 20–46.
12. Han, H.-J., Powers, S.J., and Gabrielson, K.L. (2022). The Common Marmoset—Biomedical Research Animal Model Applications and Common Spontaneous Diseases. Toxicol. Pathol. *50*, 628–637.
13. Perez-Cruz, C., and de Dios Rodriguez-Callejas, J. (2023). The common marmoset as a model of neurodegeneration. Trends Neurosci. *46*, 394–409.
14. Bert, A., Abbott, D.H., Nakamura, K., and Fuchs, E. (2012). The marmoset monkey: a multi-purpose preclinical and translational model of human biology and disease. Drug Discov. Today *17*, 1160–1165.
15. Okano, H., Hikishima, K., Iriki, A., and Sasaki, E. (2012). The common marmoset as a novel animal model system for biomedical and neuroscience research applications. Semin. Fetal Neonatal Med. *17*, 336–340. https://doi.org/10.1016/j.siny.2012.07.002.
16. Okano, H., and Mitra, P. (2015). Brain-mapping projects using the common marmoset. Neurosci. Res. *93*, 3–7. https://doi.org/10.1016/j.neures.2014.08.014.
17. Miller, C.T., Freiwald, W.A., Leopold, D.A., Mitchell, J.F., Silva, A.C., and Wang, X. (2016). Marmosets: a neuroscientific model of human social behavior. Neuron *90*, 219–233.
18. Burkart, J., and Heschl, A. (2006). Geometrical gaze following in common marmosets (Callithrix jacchus). J. Comp. Psychol. *120*, 120–130.
19. Spadacenta, S., Dicke, P.W., and Thier, P. (2019). Reflexive gaze following in common marmoset monkeys. Sci. Rep. *9*, 15292.
20. Pandey, S., Simhadri, S., and Zhou, Y. (2020). Rapid head movements in common marmoset monkeys. iScience *23*, 100837.
21. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., and Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. *21*, 1281–1289. https://doi.org/10.1038/s41593-018-0209-y.
22. Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M.M., Di Santo, V., Soberanes, D., Feng, G., et al. (2022). Multi-animal pose estimation, identification and tracking with DeepLabCut. Nat. Methods *19*, 496–504. https://doi.org/10.1038/s41592-022-01443-0.
23. Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., and Mathis, M.W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nat. Protoc. *14*, 2152–2176. https://doi.org/10.1038/s41596-019-0176-0.
24. Karashchuk, P., Rupp, K.L., Dickinson, E.S., Walling-Bell, S., Sanders, E., Azim, E., Brunton, B.W., and Tuthill, J.C. (2021). Anipose: A toolkit for robust markerless 3D pose estimation. Cell Rep. *36*, 109730. https://doi.org/10.1016/j.celrep.2021.109730.
25. Kane, G.A., Lopes, G., Saunders, J.L., Mathis, A., and Mathis, M.W. (2020). Real-time, low-latency closed-loop feedback using markerless posture tracking. eLife *9*, e61909. https://doi.org/10.7554/eLife.61909.
26. Pereira, T.D., Aldarondo, D.E., Willmore, L., Kislin, M., Wang, S.S.-H., Murthy, M., and Shaevitz, J.W. (2019). Fast animal pose estimation using deep neural networks. Nat. Methods *16*, 117–125.
27. Pereira, T.D., Tabris, N., Matsliah, A., Turner, D.M., Li, J., Ravindranath, S., Papadoyannis, E.S., Normand, E., Deutsch, D.S., Wang, Z.Y., et al. (2022). SLEAP: A deep learning system for multi-animal pose tracking. Nat. Methods *19*, 486–495.
28. Dunn, T.W., Marshall, J.D., Severson, K.S., Aldarondo, D.E., Hildebrand, D.G.C., Chettih, S.N., Wang, W.L., Gellis, A.J., Carlson, D.E., Aronov, D., et al. (2021). Geometric deep learning enables 3D kinematic profiling across species and environments. Nat. Methods *18*, 564–573. https://doi.org/10.1038/s41592-021-01106-6.
29. An, L., Ren, J., Yu, T., Hai, T., Jia, Y., and Liu, Y. (2023). Three-dimensional surface motion capture of multiple freely moving pigs using MAMMAL. Nat. Commun. *14*, 7727.
30. Bala, P.C., Eisenreich, B.R., Yoo, S.B.M., Hayden, B.Y., Park, H.S., and Zimmermann, J. (2020). Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio. Nat. Commun. *11*, 4560. https://doi.org/10.1038/s41467-020-18441-5.
31. Günel, S., Rhodin, H., Morales, D., Campagnolo, J., Ramdya, P., and Fua, P. (2019). DeepFly3D, a deep learning-based approach for 3D limb and appendage tracking in tethered, adult Drosophila. eLife *8*, e48571.
32. Zimmermann, C., Schneider, A., Alyahyay, M., Brox, T., and Diester, I. (2020). FreiPose: a deep learning framework for precise animal motion capture in 3D spaces. Preprint at bioRxiv. https://doi.org/10.1101/2020.02.27.967620.
33. Sun, G., Lyu, C., Cai, R., Yu, C., Sun, H., Schriver, K.E., Gao, L., and Li, X. (2021). DeepBhvTracking: a novel behavior tracking method for laboratory animals based on deep learning. Front. Behav. Neurosci. *15*, 750894.
34. Yabumoto, T., Yoshida, F., Miyauchi, H., Baba, K., Tsuda, H., Ikenaka, K., Hayakawa, H., Koyabu, N., Hamanaka, H., Papa, S.M., et al. (2019). MarmoDetector: A novel 3D automated system for the quantitative assessment of marmoset behavior. J. Neurosci. Methods *322*, 23–33.
35. Yurimoto, T., Kumita, W., Sato, K., Kikuchi, R., Oka, G., Shibuki, Y., Hashimoto, R., Kamioka, M., Hayasegawa, Y., Yamazaki, E., et al. (2024). Development of a 3D tracking system for multiple marmosets under free-moving conditions. Commun. Biol. *7*, 216.
36. Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., and Chen, K. (2022). RTMDet: An empirical study of designing real-time object detectors. Preprint at arXiv. https://doi.org/10.48550/arXiv.2212.07784.
37. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., and Chen, K. (2023). RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.07399.
38. Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103.
39. Carissimi, N., Rota, P., Beyan, C., and Murino, V. (2018). Filling the gaps: Predicting missing joints of human poses using denoising autoencoders. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 364–379.
40. Wei, P., Han, Y., Chen, K., Wang, Y., Liu, W., Wang, Z., Wang, X., Han, C., Liao, J., and Huang, K. (2023). Social Behavior Atlas: A few-shot learning framework for multi-animal 3D social pose estimation, identification, and behavior embedding. Preprint at Research Square. https://doi.org/10.21203/rs.3.rs-3020951/v1.
41. Eliades, S.J., and Miller, C.T. (2017). Marmoset vocal communication: behavior and neurobiology. Dev. Neurobiol. *77*, 286–299.
42. Pereira, T.D., Shaevitz, J.W., and Murthy, M. (2020). Quantifying behavior to understand the brain. Nat. Neurosci. *23*, 1537–1549.
43. Sasaki, E., Suemizu, H., Shimada, A., Hanazawa, K., Oiwa, R., Kamioka, M., Tomioka, I., Sotomaru, Y., Hirakawa, R., Eto, T., et al. (2009). Generation of transgenic non-human primates with germline transmission. Nature *459*, 523–527.
44. Kobayashi, Y., Okada, Y., Itakura, G., Iwai, H., Nishimura, S., Yasuda, A., Nori, S., Hikishima, K., Konomi, T., Fujiyoshi, K., et al. (2012). Pre-evaluated safe human iPSC-derived neural stem cells promote functional recovery after spinal cord injury in common marmoset without tumorigenicity. PLoS One *7*, e52787.
45. Carrion, R., Jr., and Patterson, J.L. (2012). An animal model that reflects human disease: the common marmoset (Callithrix jacchus). Curr. Opin. Virol. *2*, 357–362.

46. Anwar Jagessar, S., Fagrouch, Z., Heijmans, N., Bauer, J., Laman, J.D., Oh, L., Migone, T., Verschoor, E.J., and 't Hart, B.A. (2013). The different clinical effects of anti-BLyS, anti-APRIL and anti-CD20 antibodies point at a critical pathogenic role of γ-herpesvirus infected B cells in the marmoset EAE model. J. Neuroimmune Pharmacol. *8*, 727–738.

47. Smith, D., Trennery, P., Farningham, D., and Klapwijk, J. (2001). The selection of marmoset monkeys (Callithrix jacchus) in pharmaceutical toxicology. Lab. Anim. *35*, 117–130.

48. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C.L. (2014). Microsoft Coco: Common Objects in Context (Springer), pp. 740–755.

49. Zhang, Z. (2000). A flexible new technique for camera calibration. IEEE T Pattern Anal. *22*, 1330–1334. https://doi.org/10.1109/34.888718.

50. Sheshadri, S., Dann, B., Hueser, T., and Scherberger, H. (2020). 3D reconstruction toolbox for behavior tracked with multiple cameras. J. Open Source Softw. *5*, 1849.

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Marmoset3K	This study	https://doi.org/10.5281/zenodo.14672425.
Experimental models: Organisms/strains		
Common marmoset	Tsinghua University	N/A
Software and algorithms		
MarmoPose	This study	https://github.com/Leoswordy/MarmoPose; https://doi.org/10.5281/zenodo.14672988
Python 3.8	Open source	N/A
PyTorch 2.1	Open source	https://github.com/pytorch/pytorch
Other		
Recording cameras	Hikvision	N/A

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

All marmosets used in this study were healthy adults with normal body size and weight, ranging in age from 2 to 9 years. Specifically, 10 marmosets (6 males and 4 females) were used to create the training dataset. 6 marmosets (3 males and 3 females) were used to create the finetuning dataset across three experimental scenarios; 3 marmosets (1 male and 2 females) were used to build the marmoset skeleton model; 1 marmoset (male) was used to test the real-time experimental control module. All experimental procedures were approved by the Science and Technology Ethics Committee at Tsinghua University. The study complied with institutional guidelines for the care and use of animals in research. Food and water were freely available to encourage natural behaviors.

METHOD DETAILS

Datasets

To develop and evaluate MarmoPose, a set of videos of both single and multiple marmosets were collected. All videos were recorded using four synchronized cameras (HIKVISION DS- 2CD252CZY-ZFR) mounted at the upper corners of a typical marmoset home cage (1 m × 0.7 m x 0.8m). The cameras captured videos at a resolution of 1920 × 1080 × 3 pixels with 25 fps. Each home cage was equipped with daily objects, including wooden perches, wire mesh platforms, food boxes and small sticks to encourage naturalistic behaviors in the marmosets.

Ground truth annotations were initially labeled using tools provided by SLEAP, including the 2D coordinates of 16 body locations for each marmoset along with their corresponding identities. These annotations were subsequently converted into the COCO[48] format using a custom Python script, enabling the training of modes incorporated within OpenMMLab.

Marmoset3K

For the subset of Marmoset3K containing single marmoset, two human annotators labeled 1527 frames (1527 instances) from 4 different camera views across 16 videos capturing a single marmoset freely moving in the home cage. Four adult marmosets without additional modifications were used in this subset. For each marmoset instance, 16 body locations were labeled (head, leftear, rightear, neck, spinemid, leftelbow, lefthand, rightelbow, righthand, leftknee, leftfoot, rightknee, rightfoot, tailbase, tailmid, tailend) carefully. If a body part is occluded from a camera view, it is labeled as invisible and will not be included in the training and testing process.

For the subset of Marmoset3K containing paired marmosets, three human annotators labeled 1646 frames (3292 instances) from 4 different camera views across 12 videos capturing a pair of marmosets freely moving in the home cage. Six (three pairs) adult marmosets were used in this subset. Since the consistence of marmosets' identities across different camera views is essential for accurate 3D triangulation, we dyed the ears of one marmoset in each pair with harmless blue color to significantly distinguish them. For each pair of marmosets, the one dyed blue was annotated ID '1' and the other one with normal white ears was annotated ID '2', and each instance was annotated with the same 16 body locations.

In total, the Marmoset3K dataset contains 3173 labeled images comprising 4819 instances.

Marmoset3D

Ground truth of 3D poses are required in training denoising autoencoder and evaluating the accuracy of MarmoPose in 3D space, which were obtained by triangulating precisely hand-labeled 2D coordinates from multiple camera views at the same timepoint. To establish the Marmoset3D dataset, three human annotators labeled 522 3D ground truth instances, consisting of 140 instances

triangulating from 560 images containing single marmoset and 382 instances triangulating from 191 images containing paired marmosets. Two annotators labeled the images first and the third annotator proofread the labels to ensure the accuracy.

Finetuning dataset

As illustrated in Figure 4, deploying MarmoPose in a new environment with different setups requires a small amount of dataset for finetuning. For the scenario involving a family of four marmosets (Figure 4C), 100 frames (400 instances) from 4 different cameras views were annotated, and the ears of marmosets were dyed in blue, red, green with one left uncolored for identification. For the scenario involving more complex setup (Figure 4D), 100 frames (200 instances) from 4 different cameras views were annotated.

Training 2D detection model and pose estimation model

MarmoPose employes a two-stage design for 2D pose tracking, utilizing a detection model adopted from RTMdet[36] and a pose estimation model adopted from RTMPose.[37] The detection model predicts both the bounding box and the identity of each instance, while the pose estimation model predicts 16 body locations for each cropped instance. This two-stage approach minimizes the additional data required for finetuning when deploying MarmoPose in a new environment.

Both models were trained on the Marmoset3K dataset, with a division of 80% (2624 images, 3978 instances) for training and 20% (549 images, 841 instances) for testing. The detection model was trained for 300 epochs with a batch size of 16, and the training procedure took 6.5 h. The original images (1920 × 1080 × 3) were downscaled to 640 × 640 × 3 as input for the detection model. The pose estimation model was trained for 400 epochs with a batch size of 8, and the training procedure took 7.5 h. For each instance, a 512 × 512 × 3 cropped image centered on the bounding box was used as input for the pose estimation model. Training for both networks was conducted on a single NVIDIA RTX 4090 GPU using PyTorch 2.1.2.

SLEAP models training for comparison

We trained top-down models using SLEAP v1.3.0 on the same Marmoset3K dataset and selected the best configuration to estimate 2D poses of multiple marmosets. The resulting 2D poses were then processed through the same 3D reconstruction pipeline used in MarmoPose to obtain 3D poses. For the centroid model, the parameters were configured as *Max Stride = 32*, *filters rate = 1.5* and *sigma = 2.5*. For the centered-instance model with identity, the parameters were configured as *Max Stride = 64*, *filters rate = 1.5* and *sigma = 2.5*. Other parameters were set to default value.

Identity correction

To address potential errors in predicted animal identities caused by occlusion or invisibility in certain camera views, an identity correction post-processing step is applied. Each detected instance with a predicted identity in a camera view is treated as a graph node. The average peripolar distance of visible body locations across different views are computed and used as the edge weight between nodes. Then the graph is clustered into N_t groups, where N_t is the number of animals in the setup. The clustering is performed by minimizing the group cost, ensuring that instances belonging to the same animal are grouped together. Finally, the identity of each instance is updated based on the group identity, which is determined by the majority of the predicted identities within each group.

Camera calibration and coordinate system alignment

To estimate the intrinsic and extrinsic parameters of multiple cameras for triangulation, we adopted camera calibration methods similar to those provided by Anipose.[24,49] A standard checkerboard (11 × 9 squares with 45 mm square size) was placed in the home cage and rotated at various angles to collection calibration videos from all camera views.

In order to align the reconstructed coordinate system with real-world spatial positions and dimensions, we developed a custom labeling tool for users to define a new coordinate system. This tool allows users define a new coordinate system by marking three specific points in at least two camera views: the original point, a point on the x axis, and a point on the y axis. Then these points are triangulated into 3D coordinates, and the cameras' extrinsic parameters are updated accordingly to fit the newly defined coordinate system.

Triangulation

Random sample consensus (RANSAC) triangulation is employed to obtain accurate 3D coordinates by minimizing the impact of outliers in 2D predictions. Specifically, multiple possible combinations of predicted body locations from different camera views are triangulated using linear least-squares,[24,50] then the combination with the smallest reprojection error is selected as the final result.

Marmoset skeleton model

The marmoset skeleton model serves as a crucial prior knowledge for optimizing the 3D poses. As shown in Figure 3C, the skeleton model defines the reference distances (mm) between each pair of joints. These reference values were derived by manually measuring the distances between joints in three normal adult marmosets and averaging the results. These reference distances are divided into two types of constraints: strong constraints, applied to rigid body location such as the head-ear pair, where the distance remains nearly constant; and week constraints, applied to flexible body locations such as the knee-foot pair and the tail, where the distance

Please cite this article in press as: Cheng et al., A real-time, multi-subject three-dimensional pose tracking system for the behavioral analysis of non-human primates, Cell Reports Methods (2025), https://doi.org/10.1016/j.crmeth.2025.100986

varies due to soft tissue and rotation. These constraints are incorporated into several stages of the processing pipeline, including the 3D coordinates filtering, training of the denoising autoencoder and final skeleton optimization.

Pose normalization

To effectively train the denoising autoencoder (DAE) model and perform subsequent analysis, it is essential to normalize the 3D poses into a egocentric coordinate system, as in real-world spatial space, similar poses might have entirely different coordinates depending on their absolute positions and orientations. Given the original pose data P with a shape of $(16,3)$, normalization is performed by translating, rotating, and scaling P into an egocentric coordinate system. This process consists of the following steps.

First, set the middle of the spine ('spinemid') as the origin, translating all coordinates accordingly:

$$\tilde{P}_i = P_i - P_{spinemid}$$

Second, align the pose within a defined coordinate system. The upper end of spine ('neck') is aligned to the x axis, and the lower end of spine ('tailbase') lies on the x-z plane. Using these points, construct the rotation matrix R and rotate the translated coordinates:

$$P'_i = R\tilde{P}_i$$

Third, divide each coordinate by the distance L between 'spinemid' and 'neck':

$$\widehat{P}_i = \frac{P'_i}{L}$$

The resulting normalized pose $\widehat{P}$ serves as both the input and output for the DAE, ensuring consistency and eliminating the effects of translation and rotation during training and analysis.

Denoising autoencoder

Inspired by a pose estimation work on human,[39] we adopted denoising autoencoder to fill in the missing data caused by occlusion, which receives incomplete 3D poses as input and output the predicted complete 3D poses. Specifically, each pose is represented by a matrix $M_{n\times 3}$, where n is the number of body locations (n = 16 by default), and the elements are the 3D coordinated of each body location in the real spatial space. We first normalized them into a egocentric coordinated system, and then flatten the matrix into a vector $V_{1\times 3n}$ as the input of the DAE. With the Marmoset3D dataset, we generated incomplete 3D poses by randomly masking i ($1 \leq i \leq 4$) body locations to simulate the occlusion in real scenarios. These masked 3D poses and corresponding complete 3D poses are used as input and output of the DAE respectively during training.

In this study, the encoder of DAE is composed of 2 fully connected hidden layers, with 256 and 128 hidden units respectively. Symmetrically, the decoder also consists of 2 fully connected hidden layers with 128 and 256 hidden units. We used 32 latent dimensions to represent the input. The loss function consists of two parts: MSE calculating the difference between ground truth and reconstructed poses; and joint loss constraining the length between some body locations. Joint loss incorporates prior knowledge of the marmoset skeleton model into the model for guiding the reconstruction of missing body locations, making the DAE work better with limited training data.

Real-time control module

The real-time control module in MarmoPose consists of three processes: (1) Prediction process. It reads the latest images from multiple live video streams cached by separate threads, then perform 2D detection and 3D triangulation and transmit the results to the main process. (2) Display process. It reads the 2D images and 3D poses passed by the prediction process, then draw 3D poses combined with 2D predictions for display. (3) Main process. It reads poses and images from the prediction process and display the real-time results, and also provides an interface for users to perform customized event detection and corresponding experimental control. An interface, receiving the latest 2D and 3D poses of marmosets, is provided in the real-time module, allowing users to perform customized event detection and corresponding experimental control. In the scenario of head orientation detection, we first got the head orientation of the marmoset by computing the vector from the middle of 'left ear' and 'right ear' to 'head' in the real space, then we defined the head orientation as 'left' if the angle between the head orientation vector and the normal vector of left side of the cage (i.e., x-z plane in the default coordinate system) is less than 45°, and 'forward' if the angle is between 45 and 135°, and 'right' if the angle is larger than 135°.

Cameras produced live video streams at 25fps with 1920 × 1080×3 frame size, and frames are read by Real-Time Stream Protocol (RTSP). Latencies were evaluated on a computer with Intel Core i7- 13700K CPU, NVIDIA GeForce GTX 4090 GPU and 64GB Ram.

Model deployment

The detection and pose estimation models were deployed with TensorRT using tools provided by MMDeploy to further enhance inference speed for the purpose of real-time feedback. To support inputs of varying batch sizes, dynamic input shapes were configured during deployment. For the detection model, input shapes ranged from (1, 3, 640, 640) to (16, 3, 640, 640). For the detection model, input shapes ranged from (1, 3, 512, 512) to (64, 3, 512, 512) for the pose estimation model. In addition, FP16 precision was enabled to ensure real-time processing capabilities.

Behavioral mapping and clustering

To create the behavioral map, a 60-min video clip containing 180000 poses of two marmosets was analyzed. For each 3D pose, 103-dimensional features were selected, including non-locomotor movement (3x16 = 48 dimensions), derived by subtracting the 'spinemid' coordinates from each of the 3D pose and rotating the entire pose to ensure its head faces a specific direction, the locomotion velocity (3x16 = 48 dimensions), the direction from neck to head (3 dimensions), the direction from midpoint of left and right ears to head (3 dimensions), and the height of the 'spinemid' body location (1 dimension). Then principal component analysis (PCA) was performed to reduce the features to 16 dimensions which account for 97.1% of the data variance. Subsequently, t-Distributed Stochastic Neighbor Embedding (t-SNE) was employed to generate the 2-dimensional embedding of all samples. In the t-SNE space, closer point representing more similar samples after linear and non-linear transformations of the original features. We then estimate the density map and perform watershed transformation for unsupervised clustering, with the resulting clusters shown as blocks with gray boundaries. Each cluster now represents a type of representative behavior. We examine the 3D poses corresponding to the density peak of each cluster, and manually merge similar behaviors to group them into 8 typical postures of the marmoset, including standing, jumping, walking, running, hanging, climbing, sitting and proning.

QUANTIFICATION AND STATISTICAL ANALYSIS

Experiments were randomized, and no data were excluded from the analysis. Sample sizes (n) for each experiment are specified in the figure legends. Values are reported as mean ± SD unless otherwise stated. Plot and analyses were performed using Python 3.8.