A 3D whole-face movement analysis system to uncover underlying physiology in mice

Kyle Daruwalla[1,2], Irene Nozal Martin[1,2], Andrew Frankel[1], Diana Naglič[1], Zainab Ahmad[1], and Xun Helen Hou[1,*]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[2]Equal contribution

[*]Correspondence: hou@cshl.edu

Abstract

Synchronous movements of the entire face, from chewing to grimacing, offer significant insights into internal physiological processes. Mice, with discernible facial responses and evolutionarily conserved mammalian facial movement control circuits, provide an ideal model to unravel the link between facial movement and internal physiological states in mammals. However, existing frameworks lack the spatial or temporal resolution to track motion of the entire mouse face, due to its small and conical form factor. We introduce Cheese3D, a computer vision system that first captures high-speed 3D motion of the entire mouse face (including ears, eyes, whisker pad, jaw, while covering both sides of the face) using a calibrated six-camera array. The interpretable framework extracts dynamics of anatomically-meaningful 3D facial features in absolute world units at sub-millimeter precision. The precise face-wide motion data generated by Cheese3D provides clear physiological insights, as shown by proof-of-principle experiments predicting time under anesthetic from subtle facial patterns, and inferring tooth and muscle anatomy from fast chewing motions. Cheese3D can serve as a discovery tool that renders facial movements highly interpretable as a readout of otherwise hidden internal states.

1 Introduction

Facial expressions and movements, from grimacing to chewing, are a powerful reflection of our internal states in health and disease [1, 2]. Studying how coordinated movement of individual facial regions gives rise to multi-functional whole-face movements can therefore provide unique insights into internal physiological processes [3]. Work to-date suggests we can infer pain, distress, and sensory input based on subtle facial movement patterns in humans as well as rodents [4–10]. Being able to precisely and sensitively track facial dynamics has the potential to expand our understanding of how animals experience and respond to various interventions.

Mice share evolutionarily conserved facial movement control circuits with other mammals, including humans. Facial muscles controlling eyes, ears, whiskers, nose, and mouth receive direct commands from a motor control network in the brainstem, bypassing the spinal cord, and thus, are positioned relatively close to processing centers in the brain [11–13]. This shared circuit architecture makes laboratory mice ideally suited to serve as a model system for studying the link between facial movements and internal brain and body states. A sensitive method to characterize mouse facial motion will allow us to reveal how internal physiological, cognitive, and emotional states drive overt dynamics of the face. Although recent advances in computer vision have fueled state-of-the-art methods for human facial movement tracking [14, 15], similar approaches to characterizing face-wide movements in mice encounter unique technical challenges. Mouse faces are an order of magnitude smaller than human faces, and the conical shape of their head makes it difficult to capture face-wide movement using a single camera (**Figure 1a**). Existing methods rely on zooming into motion of a single facial region (e.g. whiskers, tongue) or a subset of facial regions on one side of the face [3, 16, 17]. Alternative methods have discarded temporal dynamics by focusing on still images of the face [6, 18]. Recent 3D methods hold promise to capture movements of the whole animal [19–21], but the approach has not been evaluated at the resolution required to examine the face of mouse.

2 Results

2.1 Cheese3D captures robust 3D whole-face movement in mice

The Cheese3D pipeline captures and analyzes synchronous movement of the entire mouse face at 100 Hz temporal resolution. The three pairs of high-speed video cameras (six total) are positioned compactly to capture the frontal view (Top Center and Bottom Center cameras), the profile view (Left and Right cameras), and an elevated half-profile view (Top Left and Top Right cameras) (**Figure 1b**). To acquire high-resolution facial video while maintaining comfort with the aim of obtaining more natural behavior, mice are acclimated to sitting in a tunnel with the head secured using a lightweight headpost custom-designed to allow unobstructed viewing of all facial areas (**Figure 1b** *inset*). Individual views from the six-camera array are temporally synchronized, and spatial alignment between views is captured through ChArUco calibration [22]. We identified a set of 27 facial keypoints that covers all facial areas on C57BL6/J mouse (**Figures 1c**–**1e**, **Supplementary Video 1**). Each keypoint is in sharp focus and visible by at least two cameras (see **Supplementary Table 1**), and reproducibly labeled by different researchers following written guidelines. The calibrated hardware setup and labeling protocol enables us to adapt existing markerless pose estimation techniques, such as Anipose [20] and DeepLabCut

[21], to create a unified 3D view of the whole mouse face at the spatial and temporal resolutions necessary to study facial movements.

As facial movement is inherently constrained in 3D space, existing 2D methods for mouse facial analysis either require a single camera view, limiting the type of movement studied to those that can be captured in a single plane, or relies on principal component analysis or hidden Markov models to integrate keypoints across multiple uncalibrated views, hindering direct interpretation. Moving from 2D to 3D calibrated face space is critical to enable interpretable feature selection that is physically grounded and verifiable in world units: we selected a set of 17 3D geometric features—distances, angles, areas, and volumes in 3D space—constructed from shapes defined by facial keypoints (**Figure 1f**). Furthermore, the features are localized to facial regions based upon known muscular anatomy and descriptors of rodent facial movements [5, 23].

We evaluated the accuracy of Cheese3D by comparing its resulting 3D geometric features with those measured statically using a 3D scanner (resolution: $50\,\mu$m) for the same mouse (**Figures 1g–1i**, **Supplementary Figure 1**, **Supplementary Video 2**). (Mean $\pm$ RMSE. Eye height: 2.62 ± 0.52 mm; Eye width: 3.71 ± 0.63 mm; Ear height: 12.45 ± 1.13 mm; Ear width: 6.54 ± 0.43 mm; Ear angle: $161.45 \pm 4.86°$; Eye area: 8.14 ± 2.27 mm^2; Ear area: 71.18 ± 7.39 mm^2; Nose bulge: 7.74 ± 4.75 mm^3; Whisker pad bulge: 43.87 ± 13.57 mm^3; $n = 7$ mice) (**Figure 1i**).

To validate the utility and necessity of having all six cameras, we omitted different pairs of cameras and measured changes in accuracy in corresponding facial regions (**Supplementary Figure 2**). Omitting frontal cameras resulted in skewed measurements of midline facial features (e.g. whisker pad bulge), whereas omitting elevated half-profile cameras resulted in errors of the most lateral features (e.g. ear). The six-camera array is also essential as it builds in redundancy which ensures measurements are still possible even when part of the face is obstructed in some views, as is often the case when the mouse paws (e.g. during grooming) or experimental apparatuses (e.g. to deliver food, drugs, or olfactory stimuli) come into close proximity to the face. Collectively, the synchronized and calibrated array of six cameras, combined with geometric facial features in 3D, reduces the tradeoff between compromising spatial versus temporal resolution in characterizing rodent facial movement.

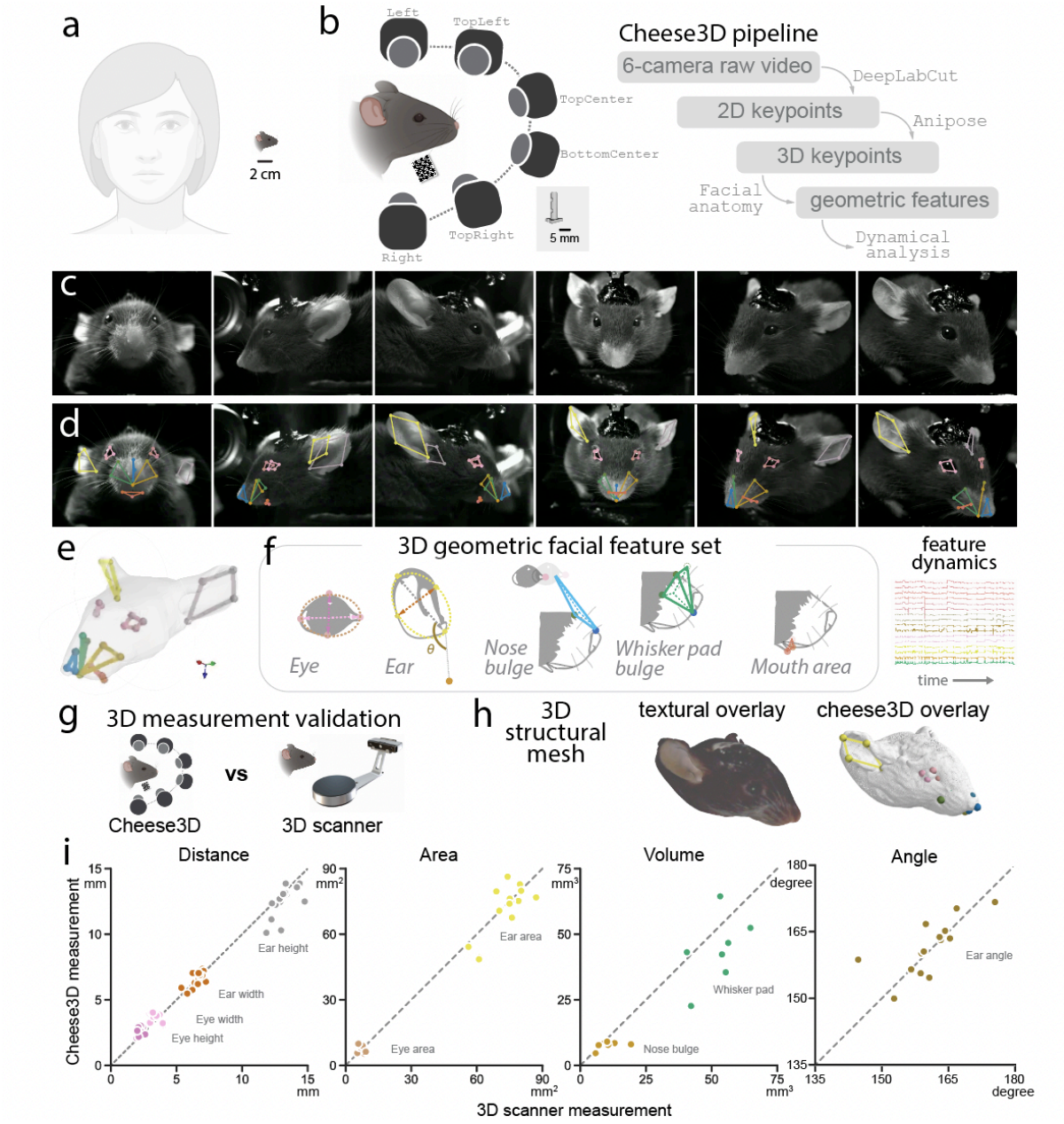


Figure 1: Framework and validation of capturing face-wide movement as 3D geometric features in mice

(a) The form factor of mouse face poses technical challenges to track mouse face-wide movement, compared to existing technology tailored for the human face.

(b) Schematic of the hardware and software framework. *Left*: The six-camera facial movement capture setup. The ChArUco board shown below the mouse is required for camera calibration. *Inset*: a head-post designed to image mouse face without occluding any facial features. *Right*: the analysis pipeline which inputs six-camera raw video and outputs the dynamics of a geometric facial feature set.

(c) Example synchronized frames from the six-camera setup.

(d) 3D facial keypoints visualized as projections onto the frames shown in **(c)**.

(e) 3D facial keypoints overlaid onto a 3D template mouse face from [24] used purely as a visual aid.

(f) Output of Cheese3D. *Left*: Illustrations of the set of anatomically-based facial features including 3D distances, areas, volumes, and angles across facial regions (see text and methods for details). *Right*, example time series of the 3D feature set.

(g) Experimental design to validate Cheese3D facial feature measurement (anesthetized) compared to 3D scanner.

Figure 1: (continued)

(h) Example mouse face 3D mesh obtained via 3D scanner *Left*: with texture overlay, showing fur and color details. *Right*: the same mesh overlaid with 3D keypoints obtained from Cheese3D to compare the two.

(i) Comparison of Cheese3D facial feature measurement with 3D scanner data, grouped by distances, areas, volumes, and angle, from left to right. Measurements for lateralized facial features (eyes, ears) contain both left and right sides and thus have twice the amount of data points compared to midline features (nose bulge, whisker pad bulge). Mouth area is the only feature excluded from the comparison since it cannot be reliably measured on the 3D scanner due to the orientation of the mouse face relative to the projector.

2.2 Measuring subtle movements across facial regions associated with anesthesia

As a proof-of-principle test that Cheese3D is able to capture subtle and rapid facial movements with physiological significance, we designed experiments to monitor mice emerging from ketamine-induced anesthesia. Small localized facial movements, including whisker deflections, both appear during anesthesia as well as signal early stages of recovery [25]. This poses unique challenges to sensitively track subtle movements while covering the entire face, compared to other overt body movements such as locomotion and reaching, where limbs and appendages undergo large translations and rotations relative to their size.

To measure the sensitivity of Cheese3D to detect and measure small, localized facial movements, we sought to explicitly quantify keypoint jitter in our setup in a control experiment using motionless periods. Keypoint jitter is a known issue whereby local fluctuations in keypoint tracking are unrelated to genuine movement [26]. This can happen due to image noise, inadequate lighting, low contrast/texture, label noise in the training data, or keypoint-specific uncertainty in the model. Across the facial keypoints selected, human labelers utilize not only texture, but also color and shape to determine the location of keypoints. Convolutional neural networks often focus on texture to solve object recognition tasks [27], thus it should be expected that certain keypoints which rely primarily on texture for their location are learned more confidently than others. Keypoint jitters are often mitigated using low-pass filters, but this can attenuate dynamics and reduce temporal resolution of detection [26]. A critical benefit of 3D multi-view calibration compared to single 2D uncalibrated views is that view redundancy reduces the amplitude of keypoint jitter, allowing us to detect more subtle fast movements. Studying motionless periods, we detected jitters of 3D keypoints without any filtering (**Figures 2a**, **2b**. Mean $\pm$ std, grouped by facial regions. Ear (left): 0.24 ± 0.10 mm/sec; Ear (right): 0.22 ± 0.11 mm/sec; Eye (left): 0.11 ± 0.06 mm/sec; Eye (right): 0.08 ± 0.04 mm/sec; Nose: 0.09 ± 0.04 mm/sec; Whisker pad: 0.17 ± 0.09 mm/sec; Mouth: 0.17 ± 0.06 mm/sec; $n = 5$ mice), and measured the reduction in jitter between 2D keypoints and 3D keypoints projected onto 2D views (**Supplementary Figure 3**). We further examined the effect of keypoint jitter on geometric features, which informs the mouse-specific threshold between keypoint tracking noise and bona fide movements that Cheese3D can detect (**Figure 2c**, **Supplementary Figure 4**. 99.9th percentile of jitter noise: Ear angle (left): $5.01 \pm 2.39\,°$/sec; Ear angle (right): $4.50 \pm 2.22\,°$/sec; Eye area (left): 1.63 ± 0.86 mm^2/sec; Eye area (right): 1.24 ± 0.66 mm^2/sec; Mouth area: 0.88 ± 0.41 mm^2/sec; Whisker pad bulge: 7.71 ± 2.33 mm^3/sec; Nose bulge: 2.18 ± 0.83 mm^3/sec; $n = 5$ mice).

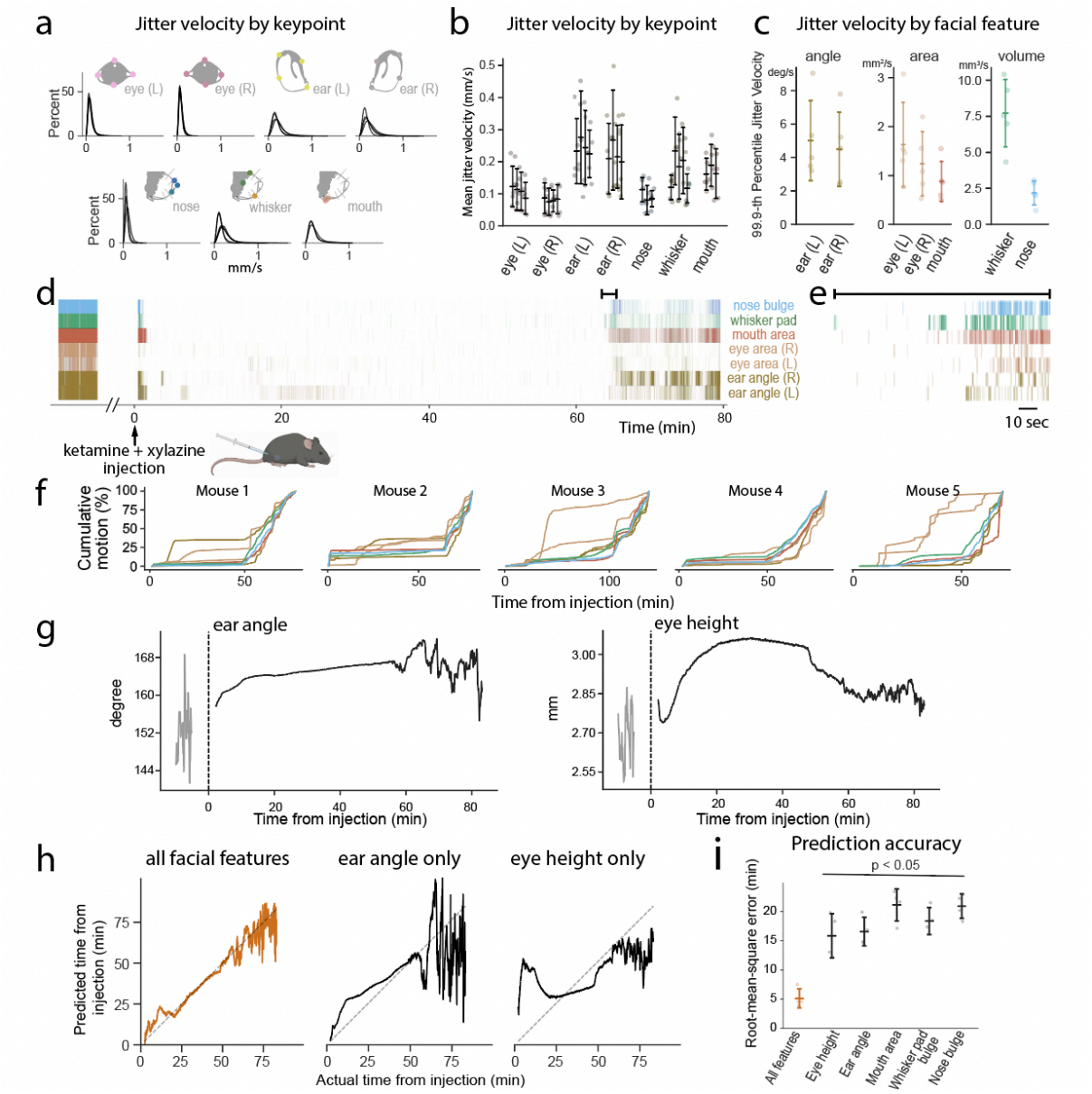


Figure 2: Measuring subtle facial movements fast and slow during anesthesia

(a) Distribution of keypoint-specific jitter (frame-to-frame velocity) during a motionless period for an example mouse. Each subpanel indicates a different facial region, and each curve indicates a different keypoint.

(b) Summary of keypoint-specific jitter across mice, where each column group indicates a facial region and each column indicates a keypoint.

(c) Summary of facial feature-specific jitter across mice.

(d) Example movement raster plot across facial regions, where each tick corresponds to velocity above the 99.9-th percentile jitter threshold as shown in **(c)**.

(e) Zoom in of movement raster plot from **(d)** to show the early moments of movement recovery following anesthesia.

(f) Cumulative movement as measured by cumulative percentage of motion raster during anesthesia across mice, where each subpanel indicates a mouse and each curve indicates a facial feature.

(g) The moving average (over a 10-sec window) of ear angle and eye height (shown for only the left side) prior to (grey) and after (black) anesthesia injection (dotted line) in an example mouse.

(h) Output from a quadratic model (fit across mice) predicting time since injection using the current facial feature value (e.g. eye height) and value at time of injection as input. Models are trained on unfiltered traces of the facial feature. The dotted identity line indicates an optimal prediction.

(i) Root-mean-square error for predicting time since injection where each column indicates a model trained on either all facial features (orange) or a particular facial feature (black). Paired two-sided t-test with Bonferroni correction (vs. all-features value, orange), $n = 4$ mice.

We next applied the thresholds defined by the jitter analysis to facial movements recorded by Cheese3D during anesthesia onset and offset, a common procedure that is associated with subtle facial movements [25]. We selected one representative geometric feature per each of the seven facial regions. Motions within different facial regions are visualized in the movement raster plot, in which each vertical line represents displacement above jitter threshold in the corresponding video frame. A detailed examination of facial movement velocity during induction, maintenance, and recovery of anesthesia reveals temporal patterns across different facial regions (**Figures 2d**–**2f**). These data demonstrate that Cheese3D can be used to detect small movements associated during anesthesia and recovery in mice.

2.3 Uncovering underlying physiology from external facial movements

We examined if Cheese3D can be used to study facial movements associated with physiological processes that are not otherwise externally visible, in addition to detecting movements that are fast and subtle. Visualizing the anesthesia data over the entire period (1 h to 2 h) revealed gradual changes in ears and eyes that are stereotyped across mice (**Figure 2g**, **Supplementary Video 3**). This suggests that certain facial features can be used as a "stopwatch" to track time since anesthesia induction. To test this hypothesis, we fit a single model across all mice to predict the time elapsed since anesthesia induction (intraperitoneal injection of ketamine and xylazine) using only the initial and current value of unfiltered facial features (**Figures 2g**–**2i**). Using features across facial regions yielded the most accurate results (RMSE: 5.12 ± 1.63 min, $n = 4$ mice) compared to single-feature models (**Figure 2i**. Single-feature model RMSE. Eye height: 15.83 ± 3.78 min; Ear angle: 16.56 ± 2.44 min; Mouth area: 21.13 ± 2.76 min; Whisker pad bulge: 18.37 ± 2.31 min; Nose bulge: 20.93 ± 2.10 min; $n = 4$ mice). We further assessed models where a single feature was omitted systematically and found that they did not significantly impact the accuracy (**Supplementary Figure 5**. Omit-one-feature model RMSE. Eye height: 5.70 ± 1.98 min; Ear angle: 6.24 ± 2.20 min; Mouth area: 5.30 ± 1.77 min; Whisker pad bulge: 5.37 ± 1.68 min; Nose bulge: 5.16 ± 1.73 min; $n = 4$ mice). In short, combining motions from different facial regions provides a useful visual indicator to track time elapsed in anesthesia.

We tested Cheese3D with facial movements that are vigorous in amplitude: chewing in rodents is difficult to characterize externally as teeth, along with food that has entered the mouth, cannot be seen. However, being able to track and measure chewing is essential to studies of nutrient absorption and efficient digestion [28]. Existing techniques to characterize chewing rely on invasive methods such as electromyography, to infer what is happening inside the mouth [29]. We hypothesized that Cheese3D would enable more direct assessment of chewing dynamics from careful examination of external facial movements during food consumption. Using the same Cheese3D multi-camera array and 17 geometric facial feature identification system with no modifications, we recorded mice as they ate crunchy food (3 mm diameter precision pellets), and visualized 3D trajectory of mouth keypoints (upper lip corners and lower lip, forming a triangle in 3D space; **Figures 3a**, **3b**). Plotting the area of this triangle (i.e., mouth opening) over time revealed two distinct modes of eating, with either elevated or reduced lower signal envelope, corresponding respectively to a food pellet obstructing the mouth opening or the mouth shut (**Figure 3c**). The transition between the two modes is abrupt and reliably identifiable across all mice (5.20 ± 1.71 sec; ranging from 2.77 sec to 8.12 sec; $n = 7$ mice. **Figure 3d**, **Supplementary Video 4**). The clear separation is also evident in movements within the facial

area close to the back of the mouth (**Figures 3e**, **3f**). This finding is consistent with the unique tooth anatomy of rodents, in which a distinct gap, termed diastema, separates the incisors (for ingestion) from the molars (for mastication), as labeled in microCT images (**Figure 3g**). Whole-face movement analysis also revealed temporally correlated eye protrusion with chewing during mastication but not ingestion or other spontaneous facial movement for every mouse examined (peak cross-correlation: 38.05 ± 19.45 for mastication; 2.80 ± 2.81 for ingestion; $n = 7$ mice, **Figures 3h–3k**, **Supplementary Video 5**). This could potentially be attributed to the anatomy of the rodent muscles of mastication since they wrap around the base of the eye socket [23]. The phenomenon has been frequently observed and named “eye boggling” in the pet rodent community, but to our knowledge has not been quantified in the scientific literature. Our data indicate that Cheese3D detects facial movements during rodent food consumption consistent with known characteristics of food placement, tooth anatomy, and muscle engagement.

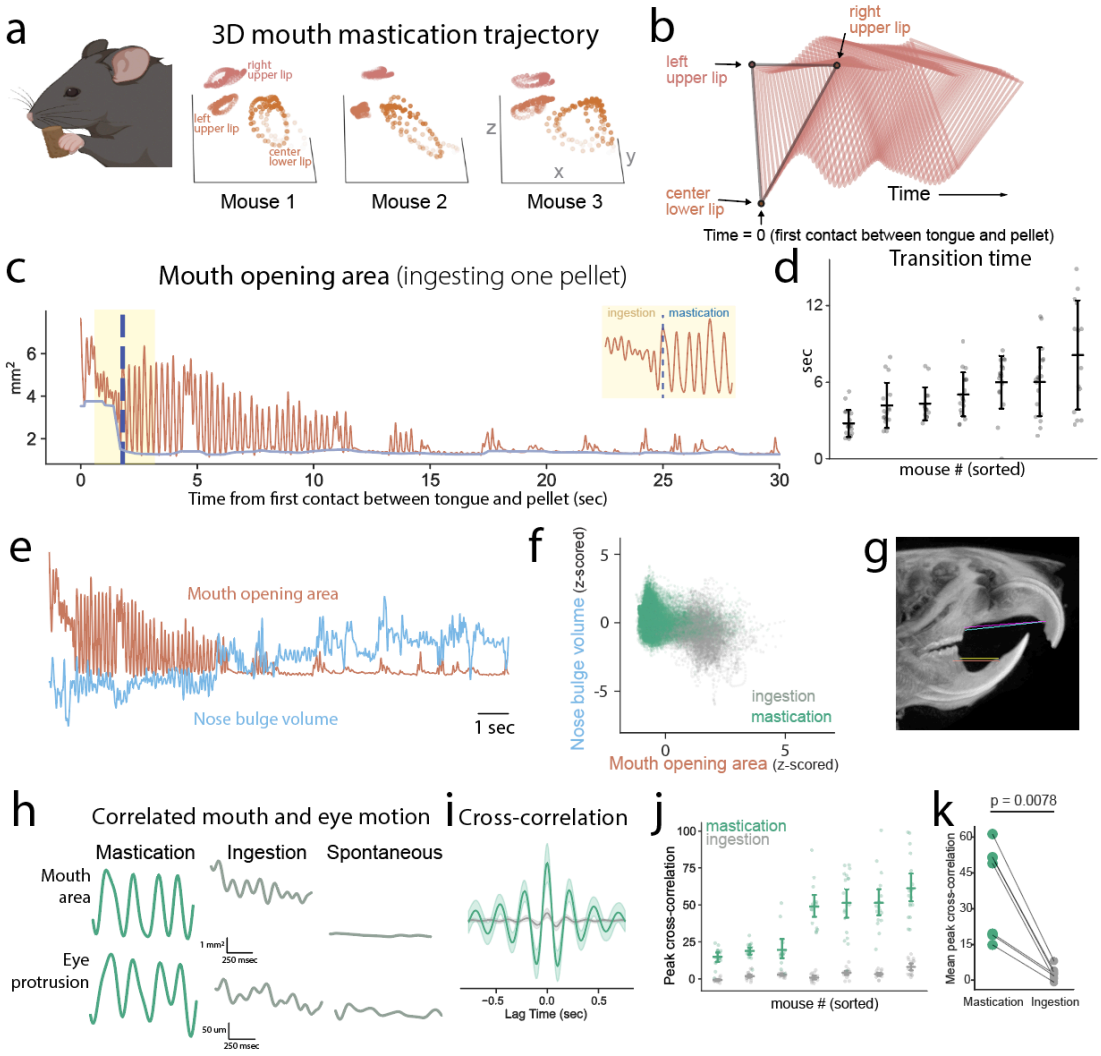


Figure 3: Chewing kinematics in mouth and surrounding facial areas

(a) 3D trajectory of three mouth keypoints for 1 sec of chewing motion for three example mice.

(b) Time evolution of the mouth opening triangle formed by the three mouth keypoints in **(a)** for 0.5 sec from the moment the food pellet comes into contact with the tongue for an example mouse.

(c) Area of the mouth opening triangle over time during consumption of one pellet for an example mouse. Blue dashed vertical line indicates putative transition time from ingestion (incisor chewing) to mastication (molar chewing), with a zoomed in view around the transition time (yellow shaded area) shown in *inset*.

(d) Summary of ingestion to mastication transition times. Each column is one mouse; each dot is one food pellet.

(e) Visualizing mouth opening area concurrent with nose bulge volume (z-scored per feature) while an example mouse consumes a single pellet (same pellet as shown in **(c)**).

(f) Same data as in **(e)** where each dot represents a time point, colored based on before the transition time (gray, putative ingestion phase) or after (green, putative mastication phase).

(g) MicroCT image of the mouse with diastema, the gap between incisors (for ingestion) and molars (for mastication), labeled in color lines.

(h) Example time segments of mouth area with eye protrusion, during putative mastication (green), ingestion (gray), and during spontaneous movement outside of chewing (gray).

(i) Cross-correlation between mouth area and eye protrusion for one example pellet for one mouse for putative mastication (green) and ingestion (gray) phases.

(j) Summary of peak cross-correlation (computed as shown in **(i)**) across pellets, where each column is one mouse.

(k) Summary of mean peak cross-correlation (computed as shown in **(h)**) across pellets, where each point is one mouse. One-sided Wilcoxon matched-pairs test (mastication mean value $>$ ingestion mean value), $n = 7$ mice.

3 Discussion

The goal of Cheese3D is to provide an interpretable framework for using mouse face-wide movement to discover underlying physiological functions across a wide range of applications. Recognizing the unique and underexplored potential to use whole-face dynamics as a noninvasive readout of moment-to-moment changes of body and brain states in mice, we crafted Cheese3D as a specialized high-resolution tool to study mouse facial movements, compared to and built upon emerging animal behavioral tracking methods aimed to generalize across body parts and species [19–21, 30–35]. Moreover, in contrast to existing methods that focus on static facial images, motion of a subset of facial features, or aggregates of orofacial behavior optimized to predict cortical neural activities [3, 6], Cheese3D is specifically designed to capture and represent whole-face movement while maintaining spatial and physical interpretability. Recording the motion of both individual facial regions and their spatial and temporal relationship to the whole face could be meaningful, since the building blocks of facial movement, i.e., compartments of facial musculature and the brainstem nuclei that directly control them, are highly topographically arranged [11, 12]. The multi-camera array setup facilitates reliable, markerless identification of facial keypoints in 3D space, counteracting occlusion and distortion found in single-camera setups, and reduces keypoint jitter compared to 2D methods. These precise spatial locations relative to other facial regions are preserved in the 3D geometric features.

The proof-of-principle work demonstrates the utility of Cheese3D in detecting and characterizing both subtle movements (anesthesia) as well as significant and temporally variable movements (food ingestion and mastication). Our analysis revealed informative synchronous facial movement patterns that could be used to infer unseen (internal anatomy and physiological functions) from seen (external synchronized facial motion). Although not in the scope of the current work, the framework described in detail here can be adapted for different strains of mice, in freely moving setup, as well as for tracking development. We anticipate the method will enable important discoveries across fields in biology and medicine by allowing for noninvasive readout of moment-to-moment changes in body states in mice. The potential applications of high-resolution, whole-face kinematics data made possible by Cheese3D are vast and are likely to inspire a new era of quantitative studies linking facial movements to changes in internal states brought on by disease, drug exposure, neural processes, or other physiological functions we would otherwise have limited access to based on external observations.

References

[1] Charles Darwin. *The expression of the emotions in man and animals / with photographic and other illustrations*. eng. OCLC: 20345883. London: John Murray, 1873.

[2] Stephen V. Shepherd, Marco Lanzilotto, and Asif A. Ghazanfar. "Facial Muscle Coordination in Monkeys during Rhythmic Facial Expressions and Ingestive Movements". en. In: *The Journal of Neuroscience* 32.18 (May 2012), pp. 6105–6116. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.6136-11.2012.

[3] Atika Syeda et al. "Facemap: a framework for modeling neural activity based on orofacial tracking". en. In: *Nature Neuroscience* 27.1 (Jan. 2024), pp. 187–195. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-023-01490-6.

[4] Chaona Chen et al. "Distinct facial expressions represent pain and pleasure across cultures". en. In: *Proceedings of the National Academy of Sciences* 115.43 (Oct. 2018). Publisher: National Academy of Sciences Section: PNAS Plus, E10013–E10021. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1807862115.

[5] Dale J. Langford et al. "Coding of facial expressions of pain in the laboratory mouse". en. In: *Nature Methods* 7.6 (June 2010), pp. 447–449. ISSN: 1548-7105. DOI: 10.1038/nmeth.1455.

[6] Nate Dolensek et al. "Facial expressions of emotion states and their neuronal correlates in mice". In: *Science* 368.6486 (Apr. 2020). Publisher: American Association for the Advancement of Science, pp. 89–94. DOI: 10.1126/science.aaz9468.

[7] Kameron K. Clayton et al. "Sound elicits stereotyped facial movements that provide a sensitive index of hearing abilities in mice". English. In: *Current Biology* 34.8 (Apr. 2024). Publisher: Elsevier, 1605–1620.e5. ISSN: 0960-9822. DOI: 10.1016/j.cub.2024.02.057.

[8] P. Ekman. "Facial expressions of emotion: an old controversy and new findings". eng. In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 335.1273 (Jan. 1992), pp. 63–69. ISSN: 0962-8436. DOI: 10.1098/rstb.1992.0008.

[9] Erwin B. Defensor et al. "Facial expressions of mice in aggressive and fearful contexts". en. In: *Physiology & Behavior*. International Behavioral Neuroscience Society (IBNS) 107.5 (Dec. 2012), pp. 680–685. ISSN: 0031-9384. DOI: 10.1016/j.physbeh.2012.03.024.

[10] Kathryn Finlayson et al. "Facial Indicators of Positive Emotions in Rats". en. In: *PLOS ONE* 11.11 (Nov. 2016), e0166446. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0166446.

[11] K W Ashwell. "The adult mouse facial nerve nucleus: morphology and musculotopic organization." In: *Journal of Anatomy* 135.Pt 3 (Oct. 1982), pp. 531–538. ISSN: 0021-8782.

[12] Jun Takatoh et al. "Constructing an adult orofacial premotor atlas in Allen mouse CCF". In: *eLife* 10 (Apr. 2021). Ed. by Alexander Theodore Chesler, Catherine Dulac, and Alexander Theodore Chesler. Publisher: eLife Sciences Publications, Ltd, e67291. ISSN: 2050-084X. DOI: 10.7554/eLife.67291.

[13] David Kleinfeld et al. "Low- and high-level coordination of orofacial motor actions". In: *Current Opinion in Neurobiology* 83 (Dec. 2023), p. 102784. ISSN: 0959-4388. DOI: 10.1016/j.conb.2023.102784.

[14] Tianye Li et al. "Learning a model of facial shape and expression from 4D scans". In: *ACM Transactions on Graphics* 36.6 (Nov. 2017), 194:1–194:17. ISSN: 0730-0301. DOI: 10.1145/3130800.3130813.

[15] Wojciech Zając et al. *NeRFlame: FLAME-based conditioning of NeRF for 3D face rendering*. arXiv:2303.06226 [cs]. Nov. 2023. DOI: 10.48550/arXiv.2303.06226.

[16] Tejapratap Bollu et al. "Cortex-dependent corrections as the tongue reaches for and misses targets". en. In: *Nature* 594.7861 (June 2021). Number: 7861 Publisher: Nature Publishing Group, pp. 82–87. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03561-9.

[17] Jun Takatoh et al. "The whisking oscillator circuit". en. In: *Nature* 609.7927 (Sept. 2022). Number: 7927 Publisher: Nature Publishing Group, pp. 560–568. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05144-8.

[18] Eric S. McCoy et al. *Development and validation of PainFace, a software platform that simplifies and standardizes mouse grimace analyses*. en. Pages: 2022.08.12.503790 Section: New Results. Aug. 2022. DOI: 10.1101/2022.08.12.503790.

[19] Timothy W. Dunn et al. "Geometric deep learning enables 3D kinematic profiling across species and environments". en. In: *Nature Methods* 18.5 (May 2021). Publisher: Nature Publishing Group, pp. 564–573. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01106-6.

[20] Pierre Karashchuk et al. "Anipose: A toolkit for robust markerless 3D pose estimation". In: *Cell Reports* 36.13 (Sept. 2021), p. 109730. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2021.109730.

[21] Tanmay Nath et al. "Using DeepLabCut for 3D markerless pose estimation across species and behaviors". en. In: *Nature Protocols* 14.7 (July 2019). Publisher: Nature Publishing Group, pp. 2152–2176. ISSN: 1750-2799. DOI: 10.1038/s41596-019-0176-0.

[22] *OpenCV: Calibration with ArUco and ChArUco*.

[23] Hester Baverstock, Nathan S. Jeffery, and Samuel N. Cobb. "The morphology of the mouse masticatory musculature". en. In: *Journal of Anatomy* 223.1 (2013), pp. 46–60. ISSN: 1469-7580. DOI: 10.1111/joa.12059.

[24] Luis A. Bolaños et al. "A three-dimensional virtual mouse generates synthetic training data for behavioral analysis". en. In: *Nature Methods* 18.4 (Apr. 2021). Publisher: Nature Publishing Group, pp. 378–381. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01103-9.

[25] Asako Noguchi et al. "Whisker electromyograms signify awake and anesthetized states in mice". eng. In: *Neuroscience Research* 148 (Nov. 2019), pp. 61–65. ISSN: 1872-8111. DOI: 10.1016/j.neures.2018.12.002.

[26] Caleb Weinreb et al. *Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics*. en. Pages: 2023.03.16.532307 Section: New Results. Dec. 2023. DOI: 10.1101/2023.03.16.532307.

[27] Robert Geirhos et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". en. In: *International Conference on Learning Representations*. 2019.

[28] Myra F. Laird et al. "Introduction: food processing and nutritional assimilation in animals". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 378.1891 (Oct. 2023). Publisher: Royal Society, p. 20220559. DOI: 10.1098/rstb.2022.0559.

[29] Yoshio Nakamura and Nobuo Katakura. "Generation of masticatory rhythm in the brainstem". In: *Neuroscience Research* 23.1 (Aug. 1995), pp. 1–19. ISSN: 0168-0102. DOI: 10.1016/0168-0102(95)90003-9.

[30] Alexander Mathis et al. "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning". en. In: *Nature Neuroscience* 21.9 (Sept. 2018). Publisher: Nature Publishing Group, pp. 1281–1289. ISSN: 1546-1726. DOI: 10.1038/s41593-018-0209-y.

[31] Anqi Wu et al. “Deep Graph Pose: a semi-supervised deep graphical model for improved animal pose tracking”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 6040–6052.

[32] Jacob M Graving et al. “DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning”. In: *eLife* 8 (Oct. 2019). Ed. by Ian T Baldwin et al. Publisher: eLife Sciences Publications, Ltd, e47994. ISSN: 2050-084X. DOI: 10.7554/eLife.47994.

[33] Jens F. Tillmann et al. “A-SOiD, an active-learning platform for expert-guided, data-efficient discovery of behavior”. en. In: *Nature Methods* 21.4 (Apr. 2024). Publisher: Nature Publishing Group, pp. 703–711. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02200-1.

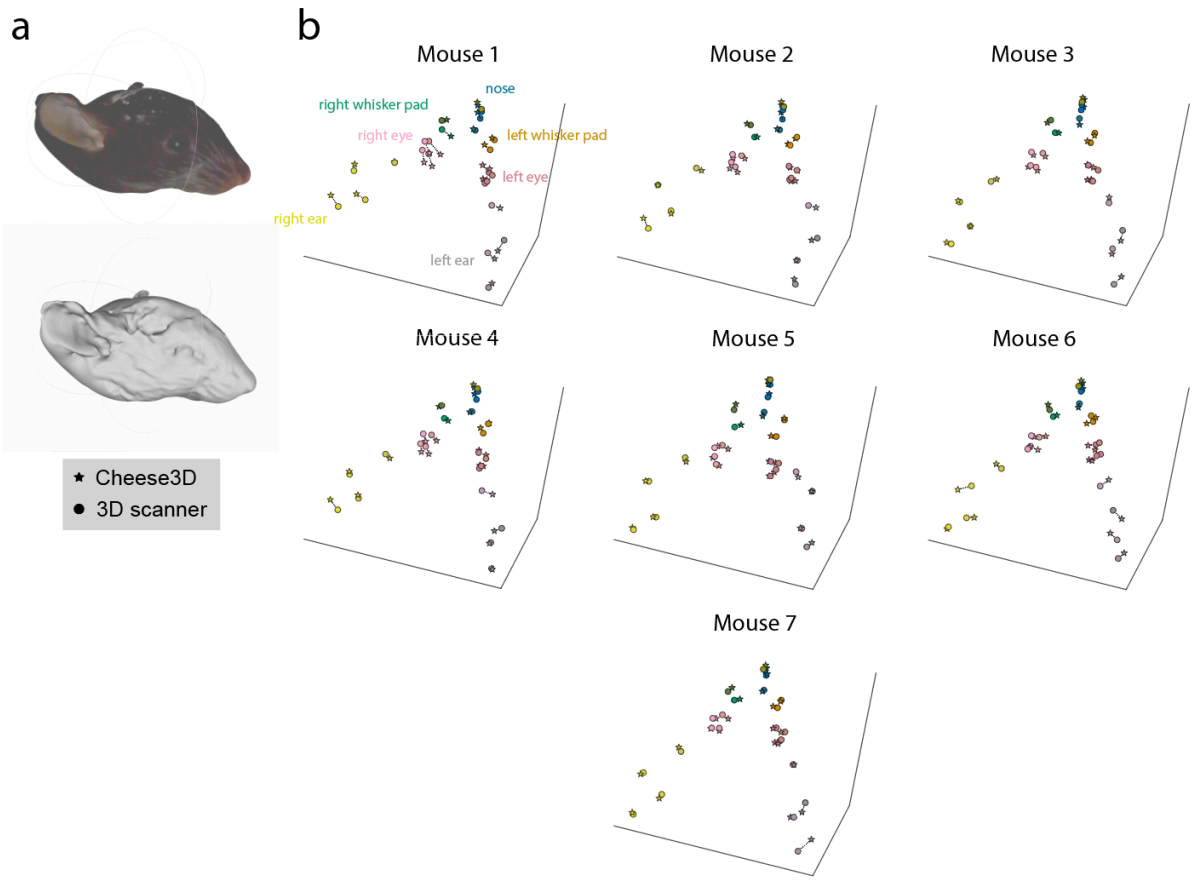
[34] Talmo D. Pereira et al. “SLEAP: A deep learning system for multi-animal pose tracking”. en. In: *Nature Methods* 19.4 (Apr. 2022). Publisher: Nature Publishing Group, pp. 486–495. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01426-1.

[35] Dan Biderman et al. *Lightning Pose: improved animal pose estimation via semi-supervised learning, Bayesian ensembling, and cloud-native open-source tools*. en. Apr. 2024. DOI: 10.1101/2023.04.28.538703.

[36] Bridget A Matikainen-Ankney et al. “An open-source device for measuring food intake and operant behavior in rodent home-cages”. In: *eLife* 10 (Mar. 2021). Ed. by Denise Cai, Kate M Wassum, and Daniel Aharoni. Publisher: eLife Sciences Publications, Ltd, e66173. ISSN: 2050-084X. DOI: 10.7554/eLife.66173.

Supplementary Information

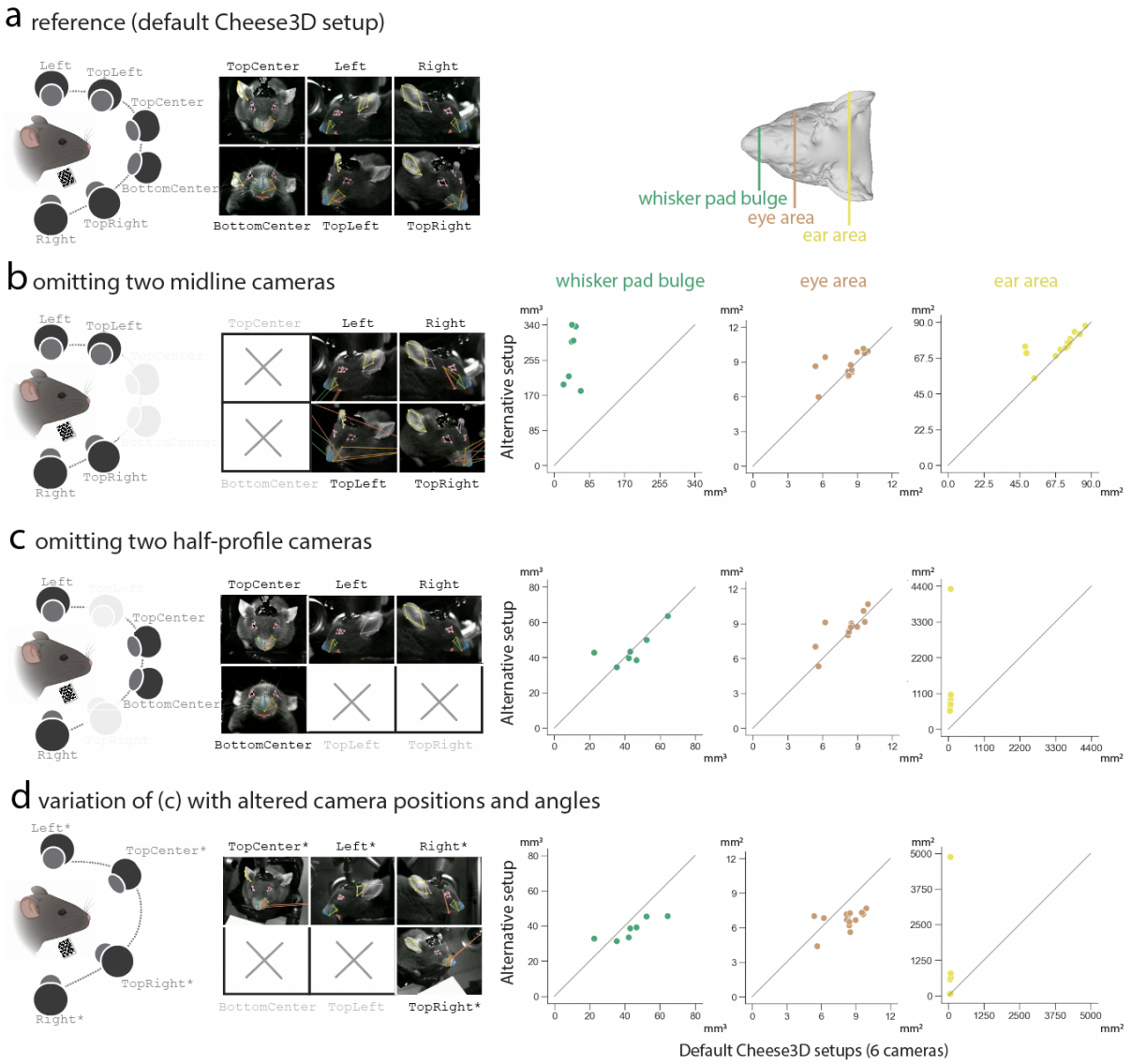
Supplemental Figures



Supplementary Figure 1: 3D facial keypoint comparison between Cheese3D and 3D scanner

(a) Example textured (above) and untextured (below) mesh of mouse face obtained in 3D scanner used for validating Cheese3D keypoint placement.

(b) 3D scatter plot showing corresponding 3D scanner (solid circle) and Cheese3D (star) triangulated keypoints for all mice shown in **Figure 1i**.



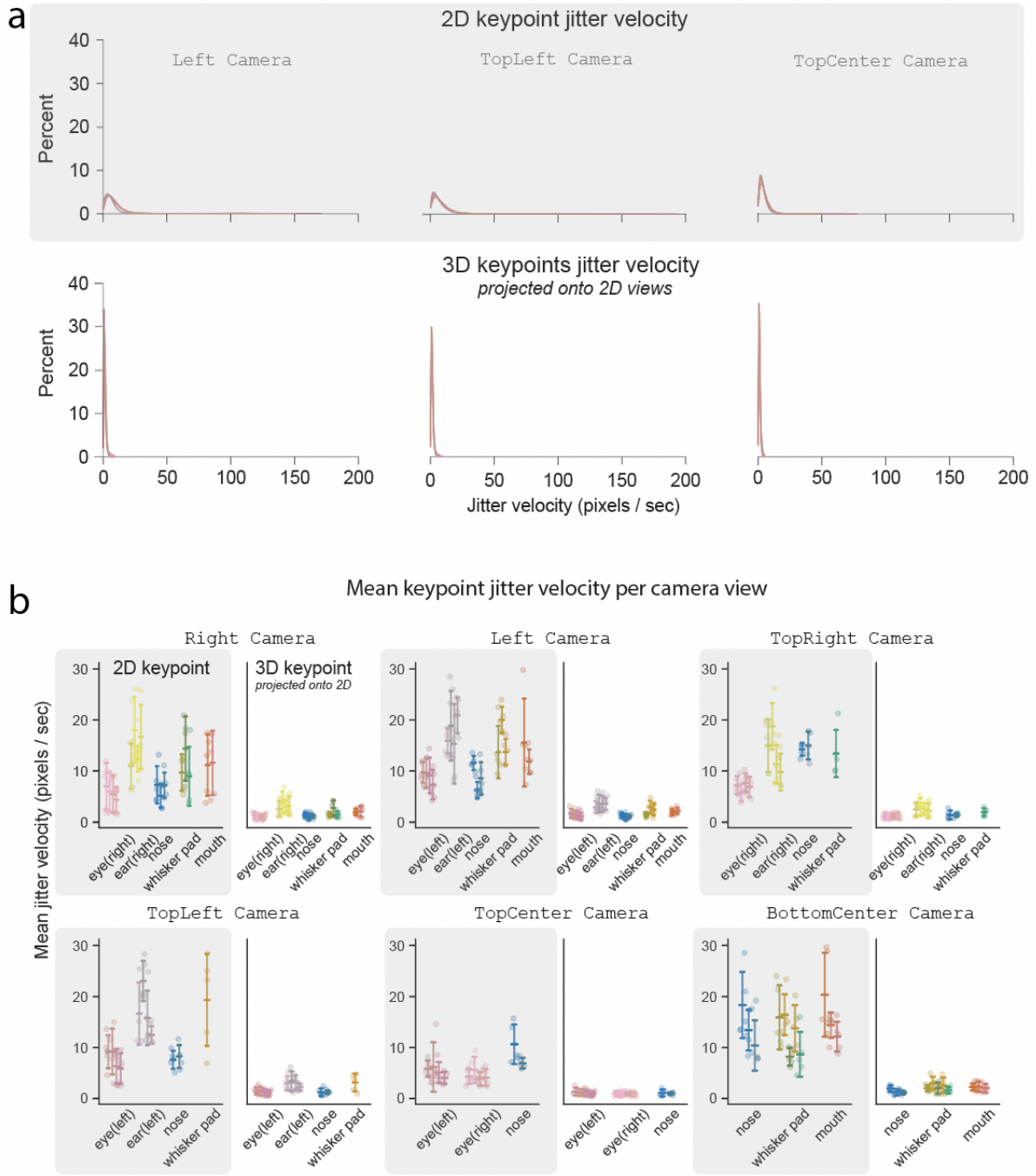
Supplementary Figure 2: Utility and necessity of six cameras in capturing mouse face

(a) Default setup for Cheese3D with a calibrated array of six (three pairs of) cameras: hardware schematic (*Left*) and 3D facial keypoints from the Cheese3D model projected onto example video frames (*Right*), same as **Figure 1**.

(b) Omitting a pair of midline cameras results in distorted keypoint inference and measurements of midline structures (Whisker pad bulge RMSE: $231.92\,\mu m^3$) but intact lateralized structures (Eye area RMSE: $1.27\,\mu m^2$; Ear area RMSE: $9.22\,\mu m^2$) compared to the default Cheese3D setup (**Figure 1i**).

(c) Omitting a pair of half-profile cameras results in altered keypoint inference and measurements of the most lateralized structures (Ear area RMSE: $1349.22\,\mu m^2$) but intact midline structures (Whisker pad bulge RMSE: $8.34\,\mu m^3$) and eye area (RMSE: $0.95\,\mu m^2$) compared to the default Cheese3D setup.

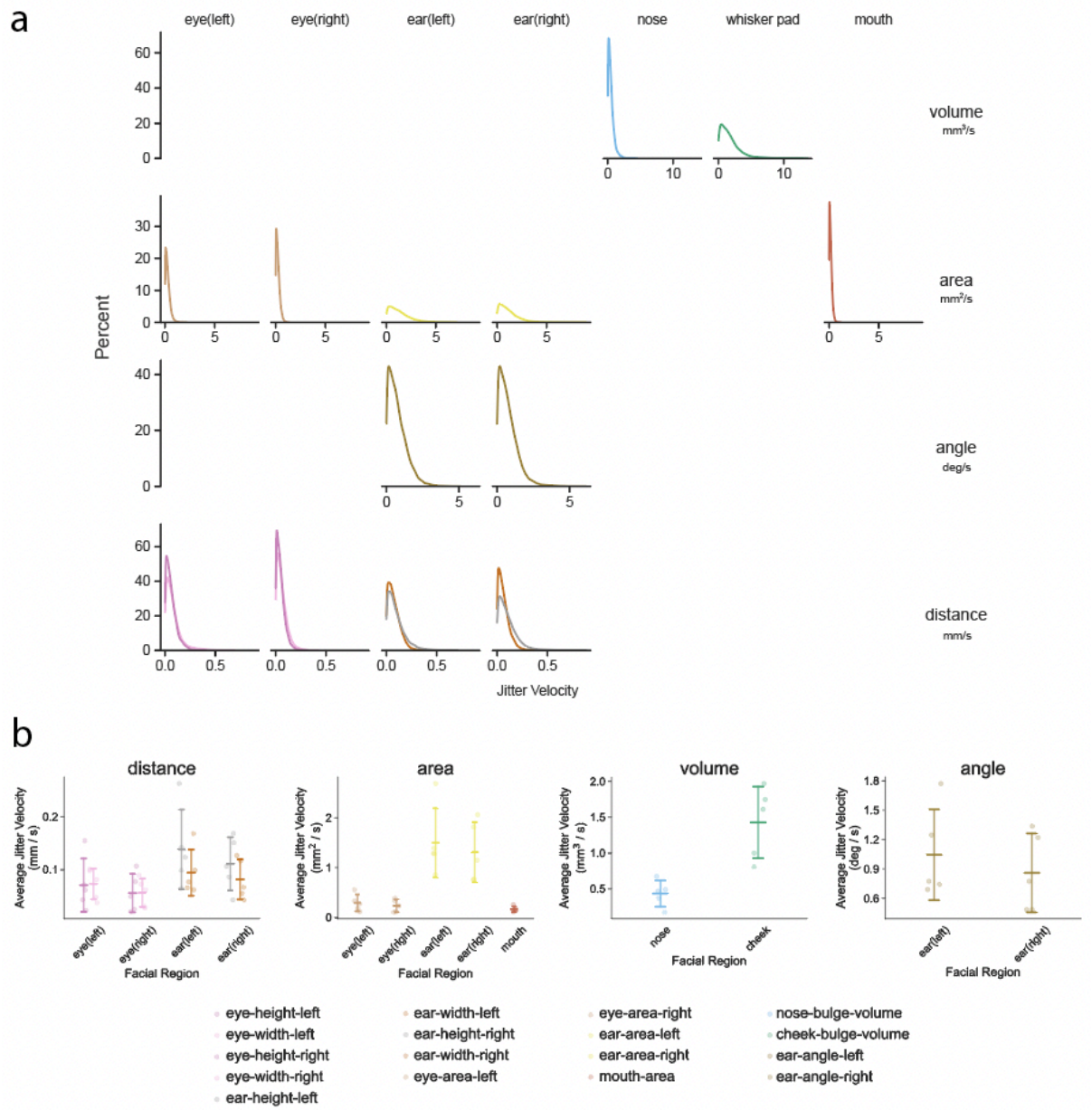
(d) A variation of setup of **(c)** with altered camera positions and angles does not rescue keypoint inference and measurements of the most lateralized structures (Ear area RMSE: $1311.45\,\mu m^2$) but has similarly intact midline structures (Whisker pad bulge RMSE: $9.89\,\mu m^3$) and eye area (RMSE: $1.90\,\mu m^2$).



Supplementary Figure 3: Keypoint jitter comparison between 2D and 3D

(a) Distribution of keypoint-specific jitter (frame-to-frame velocity) during a motionless period for an example mouse (shown for four keypoints from the left eye region). Each column indicates a different camera view. The top row (shaded) indicates 2D keypoints (prior to triangulation) and the bottom row indicates 3D keypoints (after triangulation) reprojected onto the 2D camera view planes.

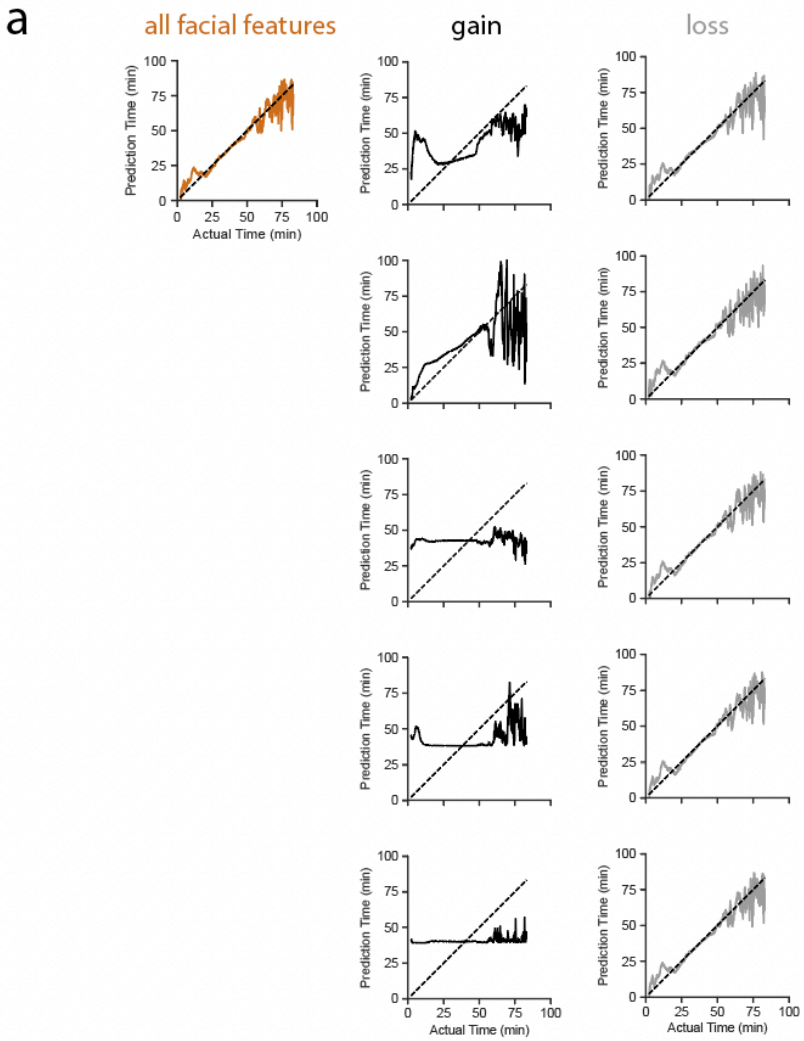
(b) Summary of keypoint-specific jitter where each pair of panels represents a different camera view, and each pair shows the jitter pre- (shaded) and post-triangulation (Mean jitter velocity, pre-triangulation. Ear (left): 17.37 ± 5.86 px/sec, Ear (right): 14.12 ± 5.26 px/sec, Eye (left): 7.29 ± 3.39 px/sec, Eye (right): 5.67 ± 2.67 px/sec, Nose: 10.05 ± 4.68 px/sec, Whisker pad: 13.55 ± 5.86 px/sec, Mouth: 13.86 ± 6.12 px/sec; Mean jitter velocity, post-triangulation. Ear (left): 3.16 ± 1.37 px/sec, Ear (right): 2.85 ± 1.50 px/sec, Eye (left): 1.29 ± 0.67 px/sec, Eye (right): 1.02 ± 0.50 px/sec, Nose: 1.23 ± 0.55 px/sec, Whisker pad: 2.21 ± 1.15 px/sec, Mouth: 2.10 ± 0.74 px/sec; all $p < 0.0001$, one-sided Wilcoxon matched-pairs test (pre-triangulation $>$ post-triangulation)).



Supplementary Figure 4: Tracking jitter by 3D facial feature

(a) Distribution of facial feature-specific jitter (frame-to-frame velocity) relative to the mean jitter during a motionless period for an example mouse. Each column indicates a different facial region, and each row indicates a different type of measurement.

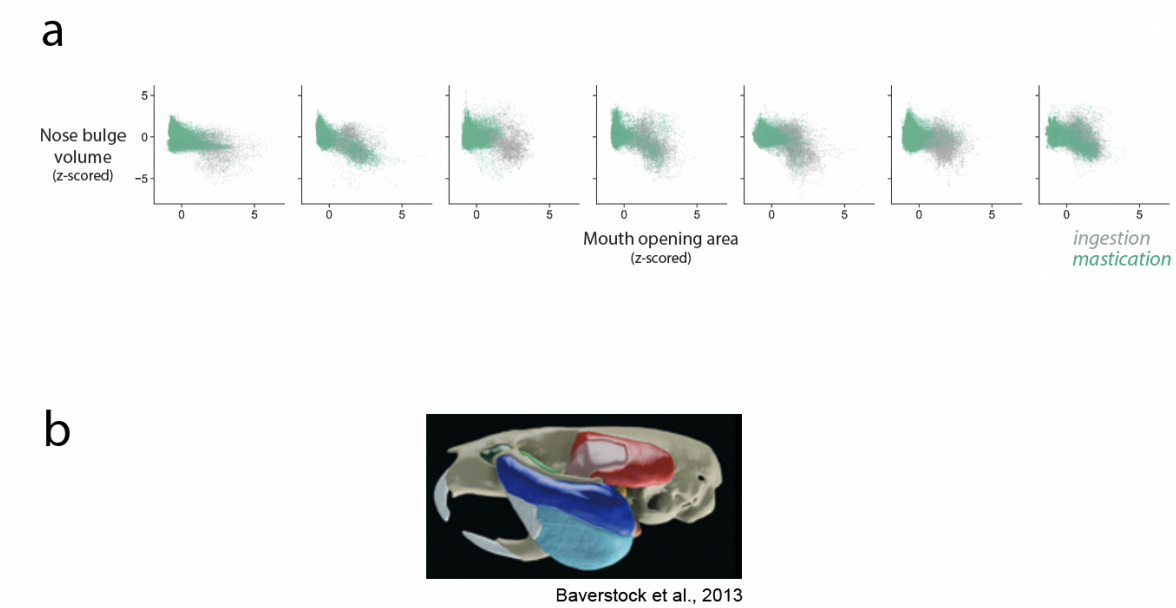
(b) Summary of facial feature-specific jitter where each panel represents a different type of measurement.



Supplementary Figure 5: Prediction accuracy of anesthesia time using different facial features

(a) We consider one feature from each facial region—eye height, ear angle, mouth area, whisker pad bulge volume, and nose bulge volume. The "whole face" feature set includes features from all regions. The "gain" condition indicates a model trained on only the specified region, and the "loss" condition indicates a model trained on all regions except the specified region.

(b) Summary of RMSE for different models where each column indicates a different feature set as described in **(a)**. For all gain condition $p < 0.05$, paired two-sided t-test with Bonferroni correction (vs. all-features value, orange), $n = 4$ mice. For all loss condition, $p > 0.1$, paired two-sided t-test with Bonferroni correction (vs. all features value, orange), $n = 4$ mice.

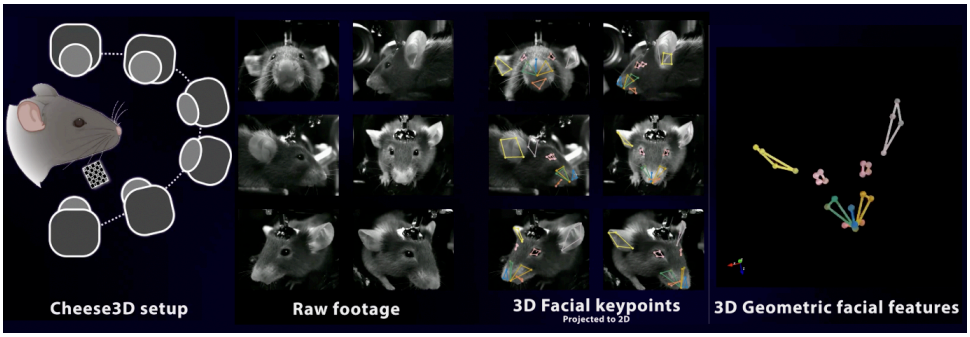


Supplementary Figure 6: Visualizing putative division between ingestion and mastication across mice

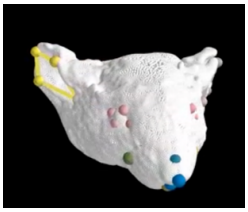
(a) Mouth opening area and nose bulge volume (z-scored) scatter plot across mice, sorted by ascending putative transition time between ingestion (gray) and mastication (green) (see **Figure 3d**). The example mouse shown in **Figure 3f** is second from the right.

(b) Segmentation of muscles of mastication shown wrapping around the eye socket [23].

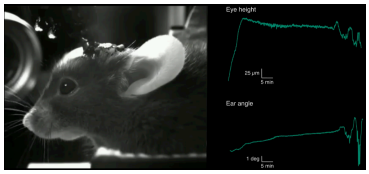
Supplemental Videos



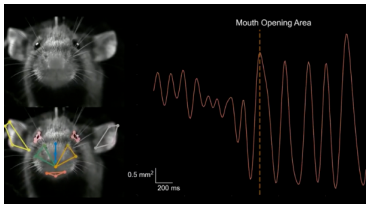
Supplementary Video 1: Cheese3D tracks whole-face movement in mouse



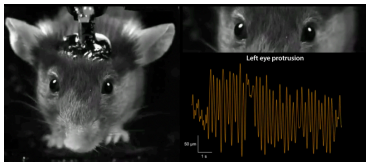
Supplementary Video 2: Facial keypoints from Cheese3D model overlaid on mesh from 3D scanner



Supplementary Video 3: Example of gradual change in eye height and ear angle during anesthesia (sped up $500\times$)



Supplementary Video 4: Example transition from ingestion (using incisors) to mastication (using molars) as capitulated in mouth opening area in 3D (slowed down $4\times$)



Supplementary Video 5: Example eye protrusion during chewing (slowed down $2\times$)

Supplemental Tables

Facial keypoint	Left	Right	Top Left	Top Right	Top Center	Bottom Center
nose(bottom)	Y	Y				Y
nose(tip)	Y	Y	Y	Y	Y	Y
nose(top)	Y	Y	Y	Y	Y	Y
pad(top)(left)	Y		Y			Y
pad(side)(left)	Y					Y
pad(top)(right)	Y			Y		Y
pad(side)(right)	Y					Y
pad(center)	Y	Y				Y
lowerlip	Y	Y				Y
upperlip(left)	Y	Y				Y
upperlip(right)	Y	Y				Y
eye(front)(left)	Y		Y		Y	
eye(top)(left)	Y		Y		Y	
eye(back)(left)	Y		Y		Y	
eye(bottom)(left)	Y		Y		Y	
eye(front)(right)		Y		Y	Y	
eye(top)(right)		Y		Y	Y	
eye(back)(right)		Y		Y	Y	
eye(bottom)(right)		Y		Y	Y	
ear(base)(left)	Y		Y			
ear(top)(left)	Y		Y			
ear(tip)(left)	Y		Y			
ear(bottom)(left)	Y		Y			
ear(base)(right)		Y		Y		
ear(top)(right)		Y		Y		
ear(tip)(right)		Y		Y		
ear(bottom)(right)		Y		Y		

Supplementary Table 1: Keypoints labeled per camera view

4 Methods

4.1 Mouse

All experiments were performed in compliance with protocols approved by the Institutional Animal Care and Use Committee at Cold Spring Harbor Laboratory (protocol number 22-6). Both female and male C57BL/6Jax mice 2-8 months of age were used for experiments. Unless stated otherwise, animals were housed in an inverse light:dark cycle with constant temperature (68 °F to 72 °F) and humidity (54-59%), and had ad libitum access to water and food.

4.2 Video capture, synchronization, and 3D calibration system

Six high-speed monochrome cameras (FLIR CM3-U3-13Y3M-CS 1/2" Chameleon®3) were used to record the video data at 100 fps. Based on their location relative to the face, the cameras are labeled LEFT (L), RIGHT (R), TOP LEFT (TL), TOP RIGHT (TR), TOP CENTER (TC) and BOTTOM CENTER (BC) (see **Figure 1b**). The camera location and orientation is selected such that each facial keypoint is in the focused view of at least 2 cameras (see **Supplementary Table 1** for details). The lateral cameras (L, R, TL, TR) are equipped with an 8 mm EFL, *f*/1.4 lens (MVL8M23, Thorlabs) and the center cameras (TC and BC) with a 12 mm EFL, *f*/1.4 lens (MVL12M23, Thorlabs). Lenses are connected to the body of the camera through a C-to-CS-mount (03-618 Edmund Optics) and 3D-printed 1.1 mm (L, R, TL, TR cameras) or brass 2 mm spacer rings (TC, BC cameras) (03-633, Edmund Optics) for fine focal adjustment. The face is illuminated using two infrared lamps (CMVision IR30 WideAngle) with a piece of Kimwipe (Kimtech Science) covering the LED surface acting as a light diffuser to minimize glare.

Cameras were synchronized using Bonsai (v2.8.1) and an Arduino Mega 2560 REV3, which sends a start signal to Bonsai through the serial port. Upon receiving the trigger signal, Bonsai begins recording frames from all cameras as well as associated metadata for each frame. To verify that the camera frames are synchronized, a miniature infrared LED (SML-S13RTT86, Mouser Electronics) is positioned to appear in the field of view of all cameras. As a synchronization signal, the LED is on for 10 ms every 10 sec, and verified post hoc in video analysis.

We calibrate camera views using a manufactured calibration board with a standard ChArUco template imprinted on its surface. A vectorized template for the ChArUco board was created using `https://github.com/dogod621/OpenCVMarkerPrinter`. The template used is for a 7×7 ChArUco board (4.5 mm marker length, 6 mm square side length, ArUco dictionary `DICT_4x4_50`). Prior to recording any experimental data, an experimenter held and rotated the ChArUco board in the focused view of all cameras for at least one minute. This calibration video acquisition step is repeated upon completion of the experiment. These calibration videos were used in Anipose to calibrate the pipeline for triangulation.

4.3 Headpost design and surgery

The custom-designed stainless steel headpost for head-fixation consists of a 6 mm $\times$ 4 mm $\times$ 1 mm rectangular base and a small 10 mm $\times$ 3 mm post that fits into the headpost holder. A groove was added on each lateral end of the base design to facilitate metabond adhesion during implant surgery. The headpost has a conical notch etched on the side to secure in

the headpost holder with a screw fastener. The headpost holder is angled at $27.9°$ following observation of the natural head angle of mouse eating to maximize comfort.

To implant the headpost, 2-month-old mice were anesthetized with isoflurane (SomnoFlo, Kent Scientific; 3–5% induction, 1–2% maintenance). Once anesthetic depth was achieved, mice were placed onto a stereotaxic apparatus where body temperature was maintained using a heating pad. After flattening the skull using skull landmarks, the base of the headpost is positioned above the medial-lateral midline, and immediately anterior to lambda, and secured using adhesive cement (Metabond, C&B). Following surgery, animals were administered buprenorphine ($0.1\,\mathrm{mg\,kg^{-1}}$) and allowed to recover on a heating pad before returning to their home cages, where the mice continue to recover for one week before being acclimated to sitting in a tunnel and head-fixation for one to two weeks.

4.4 Neural network keypoint detections and validations

We utilize video data from across all mice and experimental conditions (feeding experiments, awake recordings from the anesthesia experiment, and recordings from the structure experiment) to train a single DeepLabCut (DLC) model to track 2D keypoints. A total of 491 frames are selected using the K-means clustering algorithm for frame extraction provided by DLC, as well as selected manually (136 frames are manually taken from the feeding experiment). Random uniform sampling is used to separate 20% of the frames for testing, while the remaining 80% are used to train the model. Following the standard guidelines provided by DLC, we select the built-in ResNet-50 model architecture and image augmentation pipeline for our training procedure. The model is trained for $1\,030\,000$ iterations using a learning rate schedule of 0.005 for $10\,000$ iterations, 0.02 for $420\,000$ iterations, 0.002 for $300\,000$ iterations, and 0.001 for $300\,000$ iterations. After training, the train set error was 2.16 px and the test set error was 4.6 px.

Back-to-back 3D scanner and Cheese3D recordings in anesthetized mice were used to measure the spatial accuracy and resolution of keypoint detection (see **Figure 1**, Supplementary Fig. 1). Each mouse underwent intraperitoneal injection of Ketamine ($100\,\mathrm{mg\,kg^{-1}}$) and Xylazine ($10\,\mathrm{mg\,kg^{-1}}$) cocktail to induce anesthesia, scanned first on the 3D scanner (Einscan-SP, SHINING 3D) and then immediately on the Cheese3D setup. To test the robustness of Cheese3D in detecting 3D keypoints, an alternative set-up was constructed using only four cameras with altered positions and angles (**Supplementary Figure 2d**). The four cameras were equipped with an 8 mm EFL, *f*/1.4 lens (MVL8M23, Thorlabs), a C-to-CS-mount adaptor (03-618 Edmund Optics), and 3D-printed 1.1 mm spacer ring.

4.5 Triangulation and 3D tracking optimization

We use the trained DLC model to track keypoints in videos for each camera view separately per experiment. No post-processing is applied to the tracked keypoints. Anipose is used to triangulate 2D keypoints from multiple cameras into a single 3D keypoint per frame. Next, Anipose optimized the 3D keypoint tracking for the full recording by reprojecting the 3D keypoints to 2D in each camera view and minimizing the mean squared error of the reprojected points. Concurrently, the frame to frame velocity of the 3D keypoints is minimized to prevent spurious tracking errors. No post-processing or filtering is applied to the optimized 3D keypoints. To evaluate the performance of the tracking pipeline, we overlaid the optimized 3D keypoints re-

projected onto each camera view, and an experimenter curated the accuracy and precision of the tracking results.

4.6 Anatomical-based interpretable feature selection

Features are selected and calculated in five tiers with increasing spatial dimension. First, 3D facial keypoints (see **Figure 1**, **Supplementary Figure 1**) are selected based on the following criteria: 1) can be unambiguously and correctly pinpointed by at least three experimenters independently; 2) (for the purpose of 3D calibration) in focused view by at least two cameras; 3) reflect natural facial features and anatomy. Second, Euclidean distances between 3D keypoints within a localized facial region (e.g. the left eye) are calculated; Third, areas are calculated for the sets of keypoints that form a closed polygon; these include the eye, ear, and mouth areas. Fourth, the angle between the ear and snout is calculated as a measure for how forward-orienting the ears are with respect to the whole face. Fifth, the volumes of the nose bulge and whisker pad bulge are calculated to reflect anatomically relevant volumes [5].

The area of the eye and ear groups are calculated based on a flattened 2D ellipse. Each group consists of four points defining the major and minor axis endpoints of the ellipse. Since all four points are not necessarily coplanar, we assume that the ellipse can be bent along the minor axis. To compute the area of this bent ellipse, we begin by defining the major axis (using the front and back of the eye or the base and tip of the ear). Next we compute the midpoint of the major axis and calculate the Euclidean distance from this midpoint to each of the remaining two minor axis endpoints. The sum of these two distances defines the length of the minor axis after a potential bend has been flattened. Using the major and minor axis lengths, we compute the final ellipse area as the standard area of a 2D ellipse in Euclidean space. The area of the mouth can be computed as the standard area of a triangle in Euclidean space. The right and left upper lip points and one central lower lip point form the vertices of the triangle. The volume of the nose bulge is calculated for an irregular tetrahedron defined by the nose top, left and right pad top, and the midpoint between the front of the eyes. We use the standard volume for an irregular tetrahedron in Euclidean space. The volume of the whisker pad bulge is calculated for an irregular pyramid defined by the nose bottom, left and right pad top, and left and right pad side points. We compute the convex hull defined by these points, then calculate the volume of the hull by dividing the hull into smaller tetrahedrons. The specific choice of tetrahedrons used is determined by the SciPy library.

4.7 Analysis of kinematics during anesthesia

For the anesthesia experiments (see **Figure 2**), awake spontaneous movements were recorded in Cheese3D for 5 min, followed by intraperitoneal injection of Ketamine ($100\,\mathrm{mg\,kg^{-1}}$) and Xylazine ($10\,\mathrm{mg\,kg^{-1}}$) cocktail to induce anesthesia, before returning to Cheese3D to record facial movement during and recovery from anesthesia. Temperature was maintained on a heating pad, and the exact time of injection was recorded.

Prior to analyzing the kinematics during the anesthesia experiment, we quantified the tracking jitter of 3D keypoints and facial features using a five-minute video segment where the experimenter identified no rapid movement (referred to as the 'motionless period'). Next, we calculated the magnitude of the frame to frame velocity of each keypoint during the selected

periods which we refer to as the jitter velocity of a keypoint. We use frame to frame velocity as our metric for jitter so that we focus on short time scale noise in the tracking instead of slow moving trends in the tracking that may occur over minutes or hours. To visualize the distribution of keypoint jitter velocity in **Figure 2a**, we compute a Gaussian kernel density estimate (KDE) using the `histplot` function in the Seaborn plotting library (v0.13.2). The bin size is set to $0.05\,\mathrm{mm/sec}$, and the KDE bandwidth is set using the `scotts_factor` function in the SciPy library (v1.10.1). We summarize the distribution of jitter velocity during the motionless period by computing the average velocity over the entire period per mouse in **Figure 2b**.

To assess how the jitter velocity of keypoints affects our anatomical features, we computed the absolute frame to frame velocity of each feature during the selected periods which we refer to as the jitter velocity of an anatomical feature. We selected the 99.9th-percentile of the anatomical jitter velocity distribution per mouse as our motion threshold. Any movement with a frame to frame velocity below this threshold will be considered noise. The motion threshold across mice is summarized in **Figure 2c**.

To measure the wakefulness of each mouse during anesthesia, we compute the magnitude of the frame to frame velocity of each anatomical feature over the entire recording. We labeled each time point as movement if the frame to frame velocity crosses the previously computed motion threshold, while time points where the velocity is below the threshold is labeled as no movement. **Figure 2d** shows an example raster plot of time points labeled as movement for one mouse. In **Figure 2e**, the cumulative motion during anesthesia was calculated by counting the number of time points labeled as movement from the start of anesthesia until a given time point (normalized by the total number of time points labeled as movement for the entire period post-injection).

We analyzed slow drift of the anatomical features during anesthesia using a moving average of each feature during the entire recording period. The moving average is computed using a $10\,\mathrm{sec}$ wide sliding window average. **Figure 2f** shows exemplar filtered features for one mouse over the entire recording period. We visualized the filtered features across all mice and selected one feature per facial region—ear angle, eye height, mouth area, whisker pad bulge, and nose bulge. We trained a model across mice to predict time since injection using a subset of the selected unfiltered features during anesthesia. Our model's input consists of quadratic terms of the feature at the current time point and initial time point (quadratic terms computed using Scikit Learn's (v1.4.2) `PolynomialFeatures` class) as well as a constant bias. We performed a linear regression from our quadratic input terms to the current time since injection using the `LinearRegression` class from Scikit Learn (v1.4.2). A separate model is trained for features from all facial regions, a single facial region at a time, and all but one facial region at a time. We assessed the performance of each model by predicting the time since injection for each mouse individually. A moving average filtered (using the same filter as **Figure 2f**) prediction for a single mouse and exemplar feature sets is shown in **Figure 2g**. We compute the root mean squared error of each model's prediction per mouse in **Figure 2h**.

4.8 Analysis of chewing kinematics

FED3 [36] was used to dispense chocolate-flavored $20\,\mathrm{mg}$ pellets (Dustless Precision Pellets, F05301, Bio-Serv) on demand during the feeding experiment (see **Figure 3**). A funnel and tubings are placed underneath the FED3 spout to collect the dispensed pellet and deposit it

on a translucent plastic spoon (Measuring Scoop S378, Parkell). The spoon was attached to a servo motor connected to a 3D printed linear actuator to bring the pellet to the mouth, and then retracted to await the next pellet. Animals in the feeding experiments were gently food-restricted and acclimated for two days to eating from the spoon while head-fixed, to facilitate food consumption during the experiment. Each mouse was recorded eating 10 to 13 pellets in one session, and allocated 30 sec per pellet. Dropped pellets were excluded from subsequent analysis.

We distinguished the ingestion and mastication phases of chewing based on the shape of the lower envelope of the mouth area during the consumption of each pellet per mouse. An example lower envelope is shown in **Figure 3c**. To compute the envelope, we invert the mouth area by negating it, then identifying the peaks of the negated signal using the `find_peaks` function in SciPy (v1.10.1) with a window of 200 m sec. The lower envelope is defined by linearly interpolating the calculated peaks, then median filtering the interpolated curve with a window of 1.49 sec. We defined the transition time from ingestion to mastication as when the lower envelope drops sharply as shown in **Figure 3c**. To quantify the time when the envelope drops, we computed the cumulative area under the envelope during the consumption of each pellet. The cumulative area quickly increases during ingestion, then sharply transitions to a slower increase during mastication. The "knee" in the cumulative area under the envelope was used to quantitatively define the transition time. We used the Kneedle algorithm (with the sensitivity parameter set to 1) to identify the knee point (transition time) for each pellet per mouse shown in **Figure 3d**. The Python kneed (v0.8.5) library was used as our Kneedle implementation.

In **Figures 3e**, **3f**, we compared the mouth area and nose bulge during the consumption of pellets by z-scoring each anatomical feature separately per pellet per mouse. For **Figure 3f**, we plot the normalized mouth area and nose bulge against each other for an example mouse where each point constitutes a single frame. We color each point based on whether it occurs before or after the transition time for the corresponding pellet.

We defined the eye protrusion in **Figures 3h–3k** as the Z coordinate of the left eye back keypoint (we observed similar behavior for the right eye back). To quantify the degree of co-ordination between the mouth area and eye protrusion, we z-scored each feature per pellet per mouse. Next, we computed the cross-correlation between the normalized features per pellet per mouse separately for the ingestion and mastication phases. **Figure 3i** shows the mean cross-correlation taken across pellets for a single mouse. We identified the peak cross-correlation by selecting the time point with the largest absolute cross-correlation per pellet per mouse as shown in **Figures 3j**, **3k**.

4.9 Data and Code Availability

Data presented in this paper and code to reproduce the reported results are available from the corresponding author upon request. Following acceptance of the manuscript they will be archived in a permanent public repository.

4.10 Acknowledgements

We would like to thank Y. Zhang for double-blinded 3D scan annotation, R. Eifert for manufacturing ChArUco board and headpost, X. Zhang for assembling FED3, S. Rahman for double-

blinded measurements, J. Merrill and S. Lyons of Animal Imaging Core for microCT imaging. L. Corcos for video editing, CSHL facilities for constructing the behavioral experiment room. In addition, we thank M. Bagnall, B. Cowley, D. Klindt, B. Sabatini, and S. Shea for comments on the manuscript. This work was funded by a Brain and Behavior Research Foundation NARSAD grant to X.H.H., a CSHL NeuroAI Scholarship funded by the Schmidt Foundation to K.D., and a Fulbright Scholarship to I.N.M.

4.11 Competing Interests

The authors declare no competing interests.