Pupil-DLC: an open-source deep learning pipeline for scalable, markerless tracking of pupil dynamics across conscious and unconscious states

Parsa Seyfourian[1, 2 *], Lydia C. Marks [2], Leslie D. Claar [2], Yasmeen Nahas [3], Miles Keating [4], Christof Koch [2, 5], Irene Rembado [2 *]

[1] University of British Columbia, Vancouver, Canada; [2] Brain and Consciousness, Allen Institute, Seattle, US; [3] University of Chicago, Chicago, US; [4] University of Tübingen, Germany; [5] Tiny Blue Dot Foundation, Santa Monica, 90069 California

* Correspondence: parsa.sey@yahoo.com, irene.rembado@alleninstitute.org

Abstract

Pupil diameter provides a powerful, non-invasive biomarker of brain state, correlating with arousal, attention, cognitive processing, and level of consciousness. Despite its widespread use, pupillometry remains limited by software tools that lack scalability, robustness, and flexibility across experimental conditions. Here we introduce *Pupil-DLC*, an open-source, DeepLabCut-based pipeline for scalable, markerless tracking of pupil dynamics. Pupil-DLC is trained on 20,550 manually annotated frames selected from over 130 pupil videos of head-fixed and behaving mice spanning wakefulness to diverse drug-induced states of consciousness, including psychedelics, and anesthesia. The pipeline robustly captures state-dependent pupil dynamics, achieving high agreement with human-annotated ground truth data and outperforming an established automated method in both accuracy and detection reliability, while maintaining computational efficiency. Pupil-DLC incorporates a dual-model framework comprising a General Model (GM) for high-throughput analysis of infra-red recorded mouse pupils and an Individual Model (IM) tailored for session-specific optimization, as well as interpretable confidence metrics that enable principled trade-offs between accuracy and data retention. Notably, the model generalizes without retraining to human infrared pupil recordings and, with retraining, extends to human videos acquired under diverse daylight and camera conditions, enabling cross-species pupillometry. Pupil-DLC provides a flexible, reproducible platform for quantifying pupil-linked brain state dynamics across experimental paradigms and species.

Introduction

Changes in pupil diameter, when corrected for luminance and depth of focus, provide a robust index of internal brain state across species[1,2]. Pupil fluctuations track arousal, attention, emotional processing, cognitive load, and autonomic physiology, reflecting strong coupling between central and peripheral

nervous system dynamics[3–9]. During monotonous tasks, such as simulated driving, pupil diameter shows large fluctuations that accompany drowsiness and performance lapses, whereas emotionally salient states including fear, stress, or anxiety elicit pronounced pupillary responses[8,10,11]. Even in the absence of overt behavioral or emotional changes, pupil dynamics track substates of wakefulness[7,12] and capture moment-to-moment variations in cognitive effort, decision timing, and perceptual selection[13], and can reveal which of multiple competing sensory streams is being attended[14–16]. Clinically, pupillometry is increasingly used as a non-invasive biomarker of impairment or loss of consciousness, enabling objective neurological monitoring and prognostic assessment[17–19]. Together, these findings establish pupil dynamics as a powerful, non-invasive biomarker of neural state.

Despite the potential role of pupil dynamics as a biomarker for different experimental and clinical conditions, this measurement is often limited by proprietary or rigid software pipelines that lack scalability, reproducibility, and flexibility across experimental paradigms. Many existing approaches are sensitive to lighting variability, require manual correction, or fail to integrate with modern deep learning frameworks[20]. To address these limitations, we present a customizable, open-source pupil-tracking pipeline built on DeepLabCut[21] and trained on 20,550 manually annotated frames collected from over 130 videos of head-fixed and behaving mice spanning diverse drug-induced states of consciousness, including psychedelics and anesthesia. Our method enables high-resolution, markerless pupil tracking with post-processing tools that improve accuracy and reduce noise, and generalizes robustly to videos from other species, including humans. By combining this advanced computational method with the sensitivity of pupillometry to brain state, this platform provides a flexible foundation for studying arousal, cognition, conscious and unconscious states, and neural dynamics across species.

Methods

Experimental procedures followed closely to those described previously in Claar, Rembado, et al., 2023 and Russo et al., 2025[22,23]. The methods are summarized below, and additional details are provided for novel procedures.

Mice

C57BL/6J wild-type (n=65) were purchased from The Jackson Laboratory (JAX stock #000664) on postnatal day 28-35. Mice were housed in the Allen Institute animal facility and used in accordance with protocols approved by the Allen Institute's Institutional Animal Care and Use Committee (IACUC) under protocol 25105, 2212 and 2003. All mice were housed in a shared facility with room temperatures between 20 and 22°C and humidity between 30 and 70%. Mice were maintained on a reverse 12-h light cycle, and were single-housed following surgery, with *ad libitum* access to food and water.

Headframe Implantation Surgery

Prior to undergoing experimental procedures, all mice were implanted with a titanium headframe to facilitate head fixation *in vivo* [24]. Surgery took place when mice were 4-11 weeks old. Pre-operative injections of dexamethasone (3-4 mg/kg, IM) and ceftriaxone (100-125 mg/kg, SC) were administered one to three hours prior to surgery. Mice were deeply anesthetized with isoflurane (5%) in an induction chamber and then placed on a stereotaxic rig (Model# 1900, KOPF; Tujunga, CA) and maintained at a surgical level of anesthesia using isoflurane (1.5-2.5%) via nose cone. Breathing was monitored and body temperature was maintained at 37.5 °C with a heating pad under the animal (TC-1000 temperature controller, CWE, Inc). Ocular lubricant (Systane, Alcon Inc., Geneva, Switzerland) was applied to the eyes

during anesthesia to maintain hydration. During surgery, skin was removed to expose the skull, and the skull was leveled with reference to bregma. Using white C&B Metabond (Parkell, Inc., Edgewood, New York), a custom titanium headframe was secured to the skull, and the rest of the exposed skull was covered. After surgery, the mouse was given an injection of lactated Ringer's solution (up to 1mL, SC) and was placed on a heating pad to recover. All animals received analgesics and antibiotics for two days post-surgery.

Habituation

Five days after headframe implantation surgery, all mice began a 3-week training schedule. Researchers first habituated mice to handling for 2 days. On subsequent days, mice were placed on a running wheel where the headframe was fixed in place with two sets of screws[24]. During training, mice were kept head-fixed on the running wheel for an increasing amount of time, starting with 5-minute sessions and ending with 90-minute sessions. At the end of the second week of training, the mice scheduled to receive psilocybin on the experimental day began receiving 2 intraperitoneal (IP) injections of saline during each training session (method described below). All training and experimental sessions were performed during the dark cycle.

Experimental Sessions

After mice completed 3 weeks of habituation to the experimental setup, they underwent one recording a day for three consecutive days. Half of the animals received isoflurane on the first day and psilocybin on the second day, and the other half experienced the reverse order. The urethane recording always took place on the third day as it was considered a terminal procedure under our IACUC protocol. During each recording session, mice were head-fixed on a wheel, and data was collected for 45-120 minutes. On the isoflurane and psilocybin recording days, there was a period of 30-60 minutes of baseline activity recorded during the awake state before administering the drug, followed by 30-60 minutes after drug administration. On the urethane day, there was no awake state baseline period pre-sedation given that the urethane was administered intravenously via a catheter inserted in its tail vein before setting the animal on the rig (see below for details). On the experimental day the angular position of the running wheel was acquired by a dedicated computer with a National Instruments card acquiring digital inputs at 100 kHz, which was considered the master clock. A 32-bit digital "barcode" was sent with an Arduino Uno (DEV-11021, SparkFun Electronics, Niwot, Colorado) every 30 seconds to synchronize all devices[25].

Psilocybin and Ketanserin Administration

To administer psilocybin, a protocol was developed to give an intraperitoneal (IP) injection while the animal was head-fixed on the wheel. The researcher would scruff the loose skin over the mouse's back and immobilize the two hind legs to prevent the mouse from moving. While the animal was restrained, the researcher administered the injection into the peritoneal cavity on the lower right quadrant of the mouse's abdomen. During psilocybin recordings, mice received one injection of sterile saline (1 mg/kg, IP) followed 10-12 minutes later by an injection of psilocybin (1mg/kg, IP; Usona Institute). During psilocybin+ketanserin recordings, the saline injection was substituted with a ketanserin (1 mg/kg, IP; S006, Sigma-Aldrich) injection.

Anesthesia Scale

To ensure a reliable and reproducible state of consciousness during both isoflurane and urethane states, a behavioral scale was developed based on Devor and Zalkind, 2001 to evaluate depth of anesthesia[26]. The scale quantifies the mouse's posture and response to noxious stimuli on a scale of 0-4, where a

higher number corresponded to a minimal or no response, generally observed during deep levels of anesthesia. The four tests included an assessment of muscle tone and voluntary movement (posture), withdrawal reflexes of tail and foot when a firm pinch was applied (300-500 grams of force) through a rodent pincher analgesia meter (IITC Life Science), and response to the presence of a noxious alcohol swab held near the nose. The sum of the scores from each individual test ranged from 0-16 and represented 4 states on the continuum of consciousness. On the assumption that a loss of responsiveness corresponds to a loss of consciousness, a summed score < 6 corresponded to a "conscious" state, while a score of 6-10 represented a "mildly anesthetized" state. To be considered unconscious, a mouse needed a score ≥ 11, which corresponds to either a "completely unconscious" state (11-13) or a "deep anesthesia" state (14-16).

Isoflurane Administration

To induce isoflurane anesthesia while head-fixed on the wheel, the mouse received inhalant isoflurane delivered through a small tube placed in front of the mouse's nose. After 2-5 minutes at 5% isoflurane, the flow was reduced to 1-2% for the remainder of the recording. 10 minutes after the start of isoflurane, the mouse was given a score based on the anesthesia scale. If the score was lower than 11, the percentage of isoflurane was increased, and the test was administered again 10 minutes later until the mouse scored 11 or higher on the scale. Throughout the recording, the precise concentration of isoflurane being administered was recorded. During anesthesia, a heating pad was placed under the animal and set at 37.5 °C to maintain thermoregulation. At the end of the recording, the isoflurane was switched off, and the mouse recovered before being placed back in its home cage.

Urethane Administration

Prior to intravenous (IV) injection of urethane, the mouse was warmed under an infrared heat lamp for 4 minutes to increase body temperature and promote vasodilation. The animal was then anesthetized using isoflurane inhalant (3-5%) and placed on a gel heating pad to maintain thermoregulation. The tail was dipped in warm water to increase vasodilation, and an alcohol swab was administered to sterilize the tail and increase visibility of the lateral tail vein. A beveled 31G needle attached to infusion catheter tubing (SAI, Infusion Technologies) was held parallel to the mouse's tail and inserted into the lateral vein. Correct placement of the needle was confirmed by the presence of blood in the infusion tube and smooth injection of fluid without any resistance on the plunger of the syringe. Once the needle was in the vein, *Vetbond Tissue Adhesive* (Patterson Veterinary, Loveland, CO) was applied to the exposed area of the needle to hold it in place. Isoflurane was discontinued, and the mouse was moved to a restrainer where the urethane (1.5-1.8 mg/kg) was administered via the tail vein catheter at a rate of 0.02mL/min. The depth of anesthesia was tested 60-90 minutes after the initial urethane injection and if the animal scored 10 or below, supplemental doses (20% of initial dose) were administered (up to 60% of initial dose) until the animal achieved a score of 11 or higher. When the IV injection was unsuccessful (10 out of 32 animals), urethane (1.5-1.8 mg/kg) was administered in two IP injections, each containing half of the full dose, 15 minutes apart. During anesthesia, a heating pad was placed under the animal and set at 37.5 °C to maintain thermoregulation.

Pupil Video Recordings

Videos of the right eye were acquired with an infrared camera (Mako G-032B PoE 552; Proximity Series 130mm/0.73x; MAX: 000F315C1496) at a sampling rate of 30Hz with a resolution of 640x480 pixels. The camera was pointing to an infrared dichroic mirror positioned at around 9 cm from the camera,

reflecting the right eye of the animal positioned 3 cm away from the mirror (Fig. 1A). The pupil radius ranged from a minimum of 10 pixels (radius) to a maximum of 125 pixels, with 77 pixels corresponding to 1 mm for our specific setup. A black curtain was lowered over the front of the rig, placing the mouse in complete darkness.

Pupil Video Processing

We employed *DeepLabCut* (DLC, https://www.mackenziemathislab.org/deeplabcut), a non-commercial marker-less pose estimation toolbox based on deep convolutional neural networks[21], to track pupil dynamics of mouse eye videos. DLC leverages transfer learning on deep neural networks (ResnNet-50 backbone) to track user-defined keypoints with near human accuracy. To train the network, each frame was annotated with 12 equidistant points around the pupil's contour, arranged in a clockwise orientation, and 4 additional reference points along the horizontal and vertical axes of the eye identifying the medial and lateral commissures of each frame. These annotations served as training data for two model types: a General Model (**GM**) and an Individual-specific Model (**IM**).

DLC General Model Training

The GM of pupil-DLC was trained on a large, diverse dataset comprising 20,550 manually annotated frames selected from over 130 videos of different head-fixed mice undergoing distinct conscious and unconscious states, including: urethane anesthesia, isoflurane anesthesia, awake, psilocybin, and psilocybin with ketanserin pretreatment. This training dataset included both active (e.g., running) and quiescent (e.g., resting) conscious states, thereby equipping the model with the ability to generalize across different eye sizes and shapes. GM training was performed using DLC's ResNet-50 architecture for 1,000,000 iterations. The GM allows one to analyze new pupil videos without requiring per-session retraining, offering fast, high-throughput and reproducible pupil tracking across experimental paradigms.

DLC Individual Model Training

For experiments requiring animal or session-specific precision, a tailored IM was used. For each new video, 150 frames were randomly sampled throughout the whole recording and manually annotated using the same 12-point pupil scheme described above. A separate ResNet-50 model was then fine-tuned for 500,000 iterations. This approach ensures high accuracy even in videos with atypical lighting, eye positioning, or artifact presence. The resulting model was then used to analyze the full video.

Post-training and Ellipse Fitting

Once the full video was processed through the IM or GM, DeepLabCut produced, for each frame, the predicted pupil boundary keypoints (x_i, y_i) together with a confidence value p_i. Only keypoints with confidence $p_i \geq 0.9$ were retained for geometric estimation. For each frame, ellipse fitting proceeded only when at least six high confidence pupil boundary points were available. Frames that did not meet these criteria were excluded from ellipse fitting and all ellipse derived outputs for those frames were recorded as Not a Number (NaN).

The pupil contour in each frame was approximated by fitting a single ellipse to the retained boundary points using a least squares estimator. An ellipse in the image plane is fully defined by five geometric parameters: the center coordinates (x_c, y_c) , the semi major and semi minor radii (a, b) and the orientation angle θ. Because five independent parameters must be estimated, a minimum of five non

collinear points is required to uniquely determine an ellipse. In practice, using six or more points yields a reliable approach that improves robustness to noise and local boundary deviations.

Ellipse fitting was performed using the *EllipseModel* implementation in scikit image[27], which estimates ellipse parameters by minimizing the algebraic distance between the observed boundary points and the implicit ellipse model. Given a set of $n \geq 5$ observed points (x_i, y_i), the ellipse is represented as a conic section

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

subject to the constraint that the conic corresponds to an ellipse. The model parameters are estimated using a direct least squares solution, after which the conic representation is converted into the geometric ellipse parameters (x_c, y_c, a, b, θ).

From the fitted ellipse, pupil size was quantified using the ellipse radii. In all primary analyses we reported pupil size as the pupil radius defined as the largest of the two semi axis lengths:

$$r = \max(a, b)$$

which provides a robust estimate of pupil size, insensitive to small deviations from circularity. Pupil diameter can be approximated as $2r$ when required, and pupil area can be computed as $A = \pi 2r$.

For training the models and fitting the ellipse a Windows 11 Pro desktop computer with the following specifics was used: AMD Ryzen Threadripper 3960X 24-Core CPU @ 3.80 GHz; RAM: 128GB; Graphics card: 2x NVIDIA GeForce RTX 3090 Ti.

Accuracy Assessment and EyeLoop

To assess the accuracy of the outputs from either GM or IM, we compared the extracted pupil diameters with ground-truth (GT) manually measured pupil diameters obtained using imageJ[28]. Notably, due to the manual nature of ground-truth assessment, these measurements are inherently subject to variability in the order of a few pixels. Moreover, we compared the outputs of our pipeline with the output of EyeLoop[29], an open-source, Python-based, real-time eye-tracking toolkit designed for high-speed pupil detection using a vectorized contour-detection algorithm, operating entirely on consumer-grade CPUs. Its modular architecture allows integration of custom importers and extractors, supports multiple species (including rodents), and is optimized for closed- and open-loop experimental paradigms. In our study, EyeLoop served as a benchmark to demonstrate comparison to our pupil-DLC pipeline outputs in terms of accuracy and processing speed.

Statistical differences between estimates were assessed using a Student's *t*-test when assumptions of normality and homogeneity of variance were satisfied, as verified by Shapiro–Wilk and Levene's tests, respectively. When the normality assumption was not met, differences were evaluated using the Mann–Whitney U test.

Results

Pupil-DLC robustly captures pupil dynamics across conscious and unconscious states

We trained Pupil-DLC on a large and diverse library of pupil videos from head-fixed mice (Figure 1A) spanning multiple drug-induced states of consciousness. Across conditions, pupil size (reported as the largest radius of the estimated ellipses, see Methods) exhibited clear and reproducible state-dependent differences. During awake and psilocybin conditions, pupils were larger on average and showed pronounced temporal fluctuations throughout the recordings, with a more dilated pupil during running compared to resting conditions (Figure 1C). In contrast, urethane and isoflurane anesthesia were associated with smaller pupil diameters and markedly reduced variability (Figure 1C). During urethane sedation, pupil size alternated between two distinct substates: a more aroused state characterized by pupil dilation and a less aroused state marked by sustained constriction (Figure 1C, D), consistent with previous reports[30]. These state-dependent differences were evident at the level of individual animals (Figure 1C) and remained robust when pooled across sessions and subjects (Figure 1D).

Probability density distributions of pupil diameter further highlighted these effects, with conscious states exhibiting broader distributions shifted toward larger diameters, whereas unconscious states showed narrower distributions concentrated around smaller pupil sizes (Figure 1D). These distributions were consistent across mice, indicating that the observed effects were not driven by individual variability. Together, these results demonstrate that the Pupil-DLC pipeline reliably captures physiologically meaningful pupil dynamics across distinct levels of consciousness and underscores the breadth of the dataset used to train GM. In total, the dataset comprised 95 recordings during awake condition pre drug administration, 35 recordings collected during isoflurane anesthesia, 32 during urethane sedation, and 35 following psilocybin administration (with or without ketanserin pre-treatment) (Figure 1E).

Pupil-DLC achieves high accuracy relative to human ground truth and existing methods

To assess measurement accuracy, pupil size estimates obtained with the Pupil-DLC GM and IM were compared against human-annotated ground truth and an established automated method, EyeLoop[29]. Across extended recordings, both Pupil-DLC models closely tracked GT measurements over time, faithfully capturing rapid pupil dynamics and remaining aligned across behavioral transitions. When comparing the raw pupil traces detected by the different methods, EyeLoop produced noisier traces with systematic deviations from GT (Figure 2A). Quantitative analyses confirmed these observations: median pupil size estimates from both Pupil-DLC models were substantially closer to GT values than those obtained with EyeLoop, with error distributions showing minimal deviation from GT for Pupil-DLC and markedly and significantly larger discrepancies for EyeLoop (mean±STD difference in pixels to GT for GM: 1.28±1.09; IM: 1.22±1.04; EL: 23.81±16.85; Figure 2B). All the methods detected the pupil in nearly all frames, resulting in negligible frame loss (Figure 2C).

Despite its superior accuracy, IM incurred a ~100-fold increase in computational cost. In contrast, GM preserved computational efficiency comparable to EyeLoop while maintaining high accuracy (mean±STD

minutes for GM: 25.57±9.14; IM: 423.84±33.53; EL: 31.00±11.50; Figure 2D). Together, these results demonstrate that Pupil-DLC enables accurate, robust, and computationally efficient pupil tracking over long-duration recordings.

Pupil-DLC confidence metrics reliably capture measurement quality and accuracy

We next examined how the internal confidence estimates provided by the Pupil-DLC pipeline relate to measurement quality and accuracy. Across all frames, both GM and IM assigned consistently high confidence values, indicating reliable pupil detection (mean±STD confidence value for GM: 0.97±0.035; IM: 0.98±0.033; Figure 3A).

Model confidence values closely matched human-based assessments of frame quality. Frames labeled as high quality by human observers were associated with higher confidence scores, whereas lower-quality frames showed systematically reduced confidence, validating the confidence metric as a meaningful proxy for visual measurement reliability (Figure 3B).

Systematic variation of ellipse-fitting confidence thresholds revealed a clear trade-off between frame retention and measurement accuracy. By adjusting restraining parameters as defined by the minimum number of points required for ellipse fitting and the confidence threshold assigned to each point, we evaluated performance in terms of frame loss and deviation from ground truth. Increasing confidence thresholds and the minimum number of points reduced the number of retained frames (i.e. increased frame loss) but resulted in smaller average pixel deviation from ground truth, whereas more permissive thresholds increased frame retention at the expense of greater measurement error (Figure 3C). Together, these results demonstrate that Pupil-DLC enables flexible tuning of analysis parameters to balance data completeness and accuracy according to experimental needs.

Pupil-DLC generalizes to human pupil videos across recording conditions

Finally, we evaluated whether Pupil-DLC pipeline generalizes beyond its training domain by applying it to human pupil recordings acquired under infrared illumination and publicly available[31] (Figure 4A). The dataset was recorded using a state-of-the-art dark-pupil, head-mounted eye tracker equipped with a *PointGrey Chameleon3* USB3.0 camera, modified to capture high-speed video at 95 frames per second. Recordings were collected across a variety of everyday indoor and outdoor environments, encompassing diverse illumination conditions (natural and artificial lighting) and a wide range of camera–eye angles. Some participants wore contact lenses or glasses, introducing realistic challenges such as reflections and partial occlusions, thereby increasing both the diversity and ecological validity of the dataset. We randomly selected ten human pupil recordings of various quality from this dataset to test Pupil-DLC performance. When video quality was high and reflections and occlusions were minimal, GM reliably detected and tracked the pupil in these human videos despite substantial variability in eye morphology and recording conditions, producing stable ellipse fits across frames (Figure 4A). Under these good conditions, both GM and IM yielded pupil size estimates that closely matched ground truth values (mean±STD pupil size value in pixels for GM: 57.55 ± 11.15; IM: 57.50 ± 11.09; GT: 59.18 ± 11.64; Figure

4A). In contrast, for very low-quality videos at the limit of human visibility, GM failed to detect the pupil entirely (Figure 4B). However, IM still provided estimates closely aligned with ground truth, with a mean difference of only 3 pixels, well within the variability of the ground truth assessment (mean±STD pupil size value in pixels for IM: 24.92 ± 3.08; GT: 27.20 ± 2.90; Figure 4B). To further test the pipeline's generalizability, we evaluated pupil tracking in two markedly different human videos recorded under natural light conditions using common smartphones (Figure 4C). The videos differed in duration (10 vs. 30 seconds) and sampling rate (60 vs. 30 frames per second). As expected, GM failed to track the pupil under these conditions, whereas IM closely matched ground-truth measurements (Figure 4C). We next asked whether retraining GM using a small subset of frames (40 frames) from each video could recover performance. Notably, the retrained model (GM-RT) not only outperformed the original GM but also achieved accuracy comparable to the IM (mean ± SD pupil size difference from ground truth, in pixels: GM, 11.32 ± 7.39; IM, 0.89 ± 0.79; GM-RT, 1.51 ± 1.70; Figure 4C).

Together, these results demonstrate that GM trained exclusively on mouse data generalizes to previously unseen recordings from different species without additional retraining when video quality closely matches the original training conditions. Under lower-quality or heterogeneous recording conditions, IM provides a reliable alternative, and retraining GM with a small, video-specific dataset substantially extends its tracking fidelity. Collectively, these findings support the broad applicability of Pupil-DLC for cross-species pupillometry across diverse recording conditions.

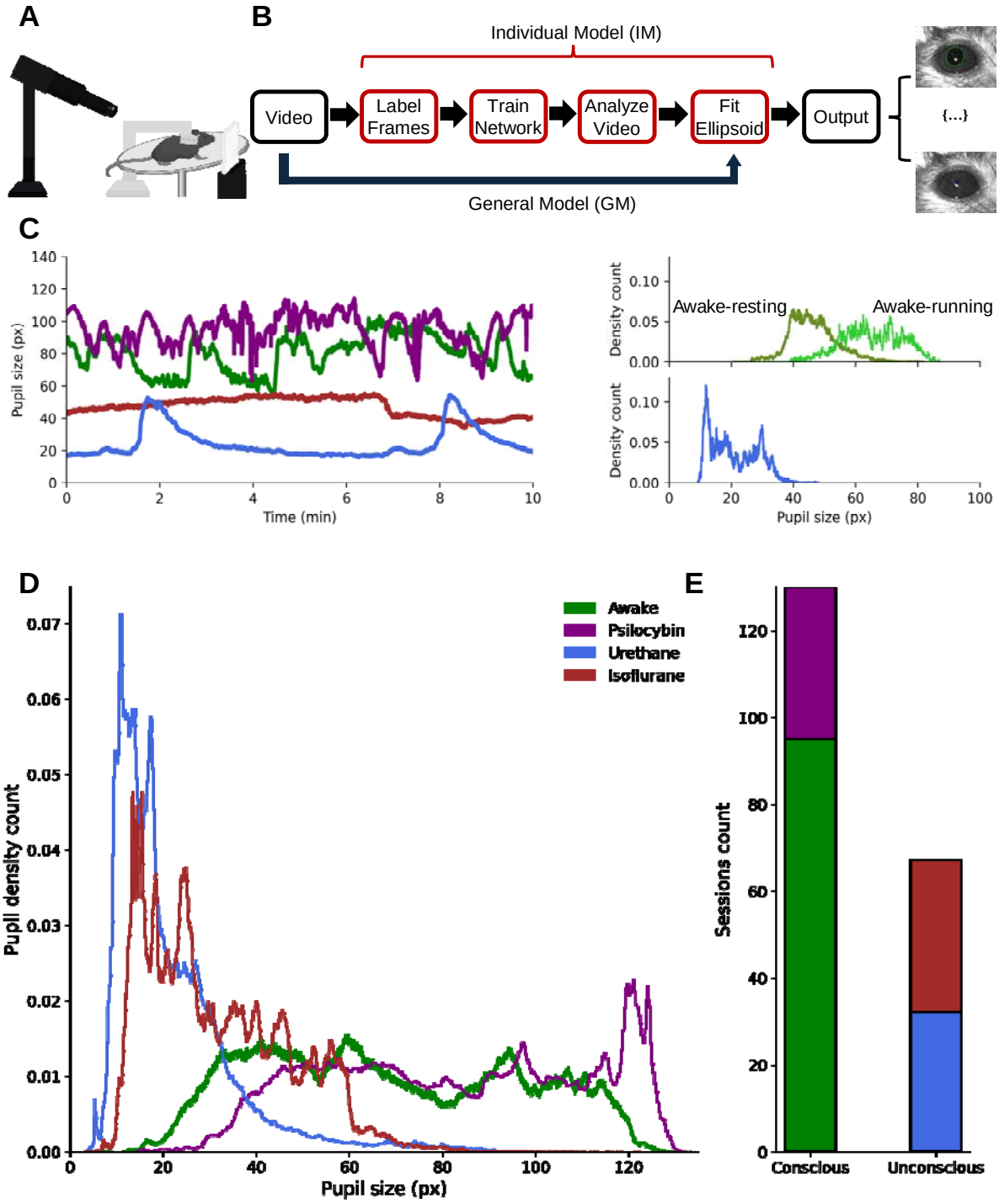

Figure 1. Pupillometry measurements in head-fixed mice undergoing different conscious and unconscious states. A) Experimental setup schematic of the head-fixed mouse on a rotating wheel (to monitor locomotion). The infrared camera points to a dichroic mirror reflecting the right eye of the animal utilized to track pupillometry. B) Pupil-DLC pipeline overview. Two main paths are provided by the pipeline: a general model (GM), trained on 20,550 manually annotated frames from over 130 videos of 65 mice, and an individual model (IM) trained on 150 frames of any one session. C) Left: Time courses of pupil size (in pixels, px) over a 10-minute recording session under four conditions: awake (green), after administration of psilocybin (purple), under urethane anesthesia (blue), and under isoflurane anesthesia (brown). Right: Probability density distributions of pupillometry across awake-resting, awake-running, and urethane states from a representative mouse. D) Probability density

distributions of pupillometry across conscious and unconscious states. E) Summary bar-plot showing the total number of sessions in each state used to train the pupil-DLC general model pooled over 130 videos from 65 mice. Note that for our mouse setup, the minimal pupil diameter of 10 pixels corresponds to 1/8 of a mm.

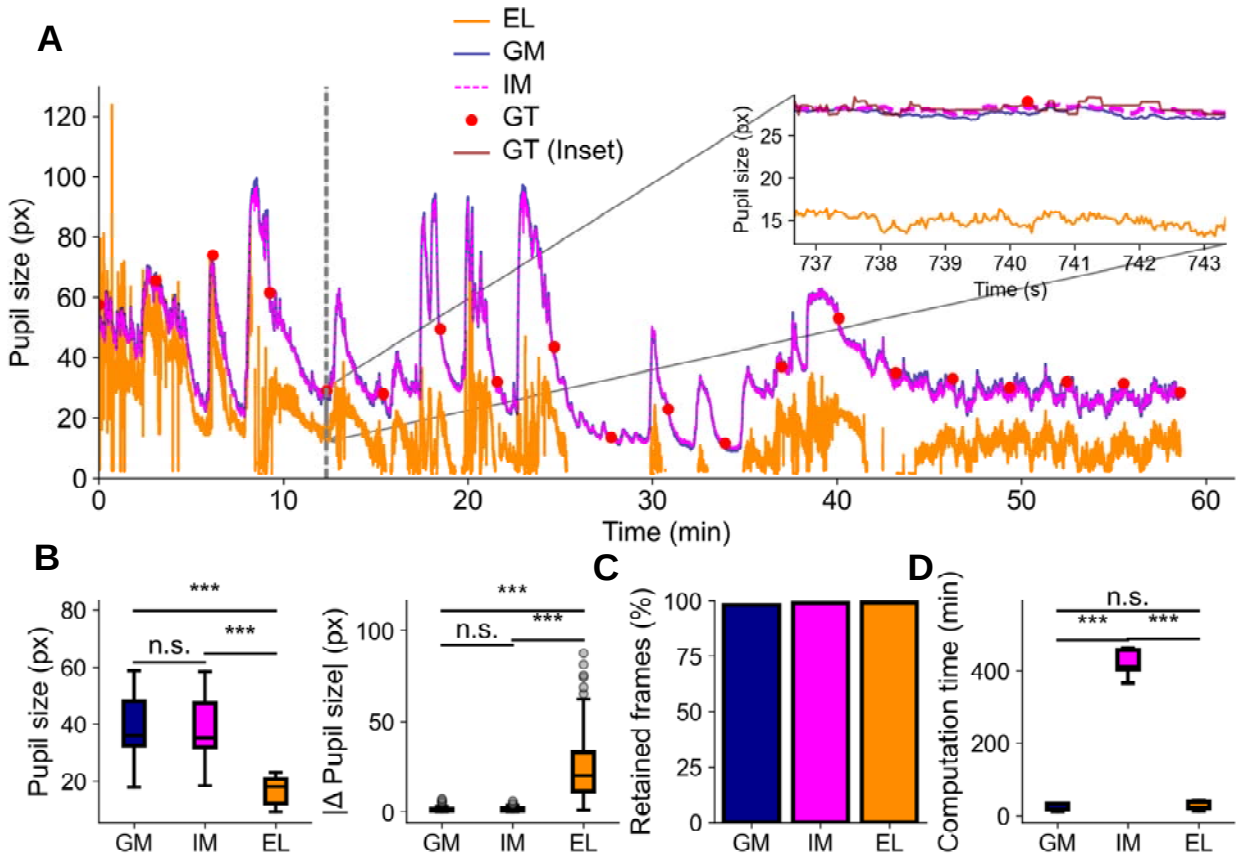

Figure 2. Performance evaluation of Pupil-DLC pipeline in measuring pupillometry across states compared to other approaches. A) Representative 60-minutes long pupillometry trace obtained from EyeLoop (EL, orange), General Model (GM, blue), and Individual Model (IM, magenta), compared to human-measured ground truth (GT, red). The inset shows a zoomed-in segment (4 seconds corresponding to 120 frames) to illustrate the close agreement between GM and IM with GT. B) Left: Absolute pupil size values (pixels) measured by GM, IM, and EL. Statistical test: Student t-test. Right: Difference between GT measurements of frames compared to other methods. Statistical test: Mann-Whitney U test. C) Percentage of retained frames (i.e., frames in which the algorithm measured the pupil) for each method. D) Computation time per video for each method. Statistical test: Mann-Whitney U test. Statistical significance: n.s $p>0.05$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

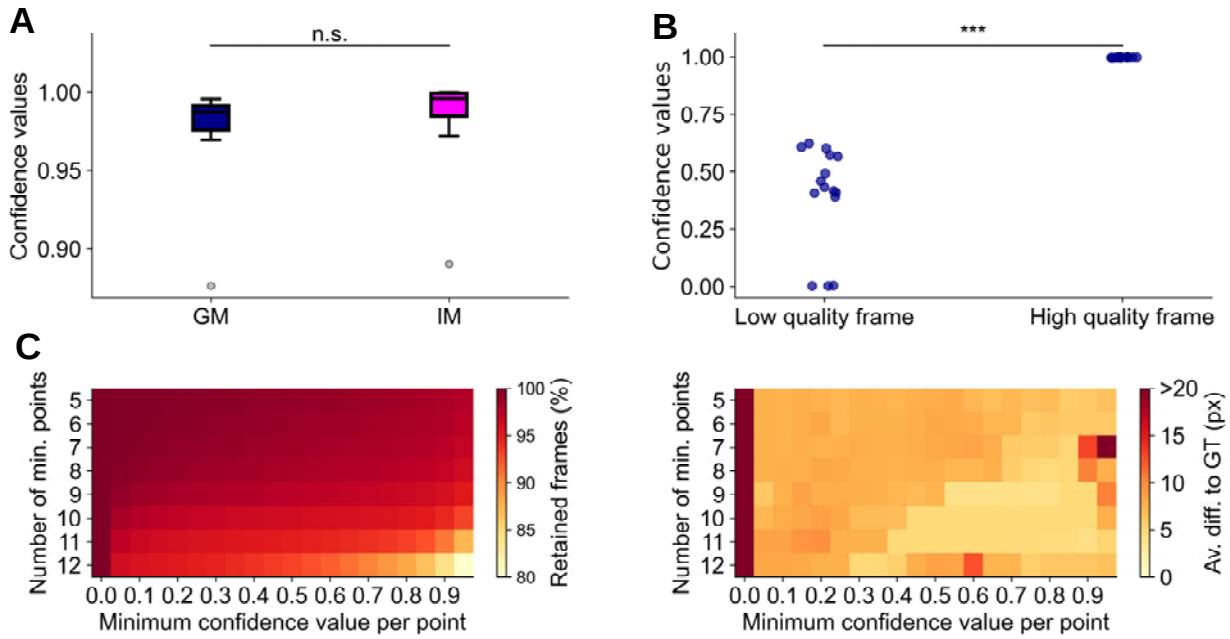

Figure 3. Performance evaluation as function of Pupil-DLC pipeline quality parameters. A) Confidence values assigned by the General Model (GM, blue) and the Individual Model (IM, magenta) across all frames averaged per video (n = 9 videos). B) GM confidence values as function of binary human-based frame quality assessment (n= 15 frames per each quality category). C) Left: Percentage of frames retained based on the minimum number of annotated points as a function of confidence value per point. Right: Pixel difference to GT as a function of the ellipse fitting parameters (number of minimum points and minimum confidence value per point). Statistical test: Mann-Whitney U test. Statistical significance: n.s p>0.05; * p<0.05; ** p<0.01; *** p<0.001.

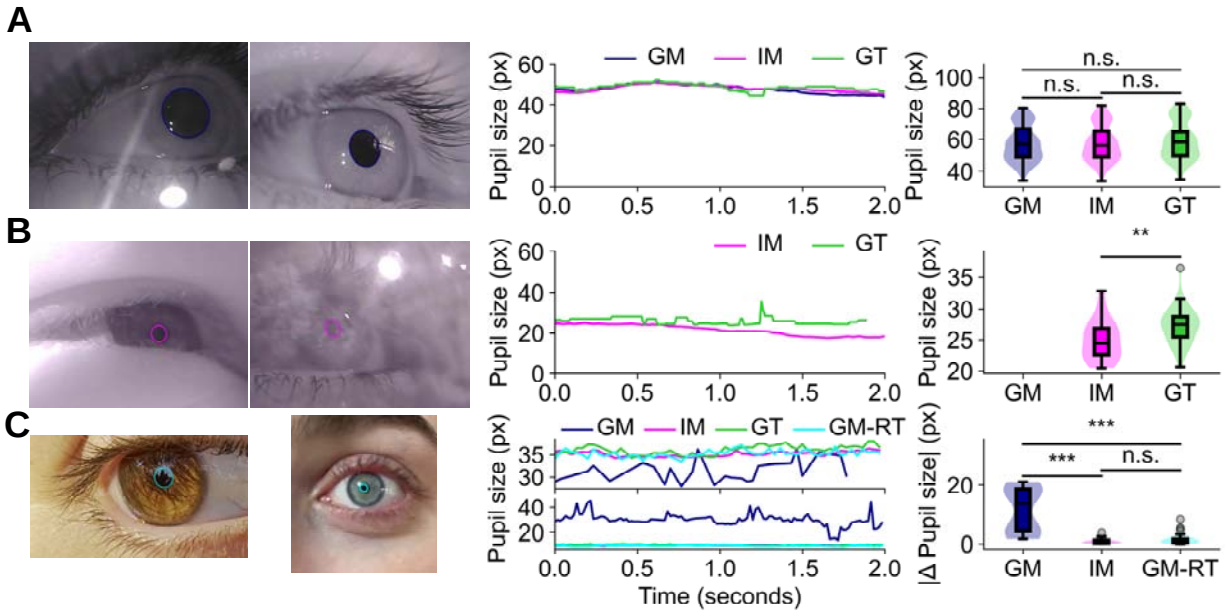

Figure 4. Pupil-DLC generalizes to unseen human pupil videos. A) Left: Representative frames from two good quality human pupil video recordings[31], showing the pupil tracked by Pupil-DLC (GM) (dark blue ellipse overlayed). Middle: Short time-series (2 seconds, 180 frames) of estimated pupil size (in pixels) by Pupil-DLC (GM, dark blue; IM, magenta), compared to ground truth (GT, green). Right: Pupil size estimation by GM, IM and GT across eight good quality human pupil videos (n=20 frames per

video). B) Left: Representative frames from two bad quality human pupil video recordings[31], showing the pupil tracked by Pupil-DLC (IM) (magenta ellipse overlayed). Middle: Short time-series (2 seconds, 180 frames) of estimated pupil size (in pixels) by Pupil-DLC (IM, magenta), compared to ground truth (GT, green). Right: Pupil size estimation by IM and GT across 2 bad quality human pupil videos (n=20 frames per video). Note GM failed to track the pupil. C) Left: Representative frames from two self-recorded human pupil videos, showing the pupil tracked by Pupil-DLC re-trained GM (GM-RT) (cyan ellipse overlayed). Middle: Short time-series (2 seconds, 180 frames) of estimated pupil size (in pixels) by Pupil-DLC (GM, dark blue; IM, magenta; GM-RT, cyan; GT, green). Right: Pupil size estimation difference from GT by GM, IM, and GM-RT across 2 human pupil videos (n=20 frames per video). Statistical test: Mann-Whitney U test. Statistical significance: n.s. $p>0.05$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

Discussion

In this study, we introduce *Pupil-DLC*, an open-source, DeepLabCut-based pipeline designed to overcome key limitations of existing pupillometry software and enable accurate extraction of pupil dynamics from video recordings. Pupil-DLC is trained on a large dataset of head-fixed mouse pupil videos spanning multiple drug-induced states of consciousness, with the explicit goal of providing scalable, markerless tracking that generalizes across experimental conditions and video quality.

A central finding is that Pupil-DLC robustly captures state-dependent pupil dynamics across awake, psychedelic, and anesthetized conditions, consistent with the established view that pupil diameter reflects brain state, autonomic regulation, and state of consciousness[3–7,9]. In our dataset, pupil size distributions differed systematically between conscious and unconscious states, where more aroused states such as during running show a dilated pupil compared to resting states, and urethane sedation exhibited alternating substates of dilation and constriction (Figure 1), in line with prior reports[12,30].

Beyond capturing biologically meaningful structure, Pupil-DLC achieved high agreement with human-annotated ground truth and outperformed a standard method, *EyeLoop*, in accuracy and precision (Figure 2). Across the heterogeneous conditions in our dataset, EyeLoop produced noisier estimates, and systematic deviations from ground truth, as opposed to the accuracy provided by both the Pupil-DLC General and Individual models, referred to as GM and IM respectively (Figure 2B). These performance differences likely reflect the advantages of a learned keypoint-based representation of pupil geometry, which remains robust to changes in pupil shape, partial eyelid occlusions, cornea reflections and imaging artifacts, conditions that commonly challenge classical contour-based methods.

A key practical contribution of this framework is the explicit separation between a GM, optimized for high-throughput analysis, and an IM, designed for session-specific optimization. Note that IM can be sufficiently trained on only 150 frames, in videos that can last over an hour or more, depending on the quality variability of the recording. GM is trained to analyze new videos without per-session retraining, enabling rapid, scalable, and reproducible processing across experiments (Figure 1B). However, in cases where recording conditions or video quality deviate substantially from the training data, performance of the GM may degrade. To address this, Pupil-DLC provides an IM that can be fine-tuned to new recordings, achieving high accuracy at the cost of increased computational time (Figure 2D). In practice, this design supports a tiered analysis strategy in which GM is used for large-scale screening and routine processing, while IM is reserved for challenging datasets or analyses requiring maximal precision.

An additional strength of Pupil-DLC is the availability of interpretable confidence metrics that closely align with human judgments of frame quality and can be leveraged to tune analysis stringency (Figure

3). Confidence values corresponded well with binary human-based assessments of frame quality, and systematic variation of ellipse-fitting thresholds revealed a clear trade-off between measurement accuracy and frame retention (Figure 3). This provides a principled framework for parameter selection depending on whether the priority is maximizing data yield or maximizing agreement with ground truth. Such flexibility is particularly valuable for long-duration recordings, where missing data can bias brain-state estimates, and for studies requiring stringent filtering to resolve fine-grained pupil dynamics.

An unexpected and notable result was the cross-species generalization observed when applying the Pupil-DLC pipeline, trained exclusively on mouse data, to human pupil recordings acquired under infrared illumination and publicly available[31]. Despite substantial differences in eye morphology and recording conditions, GM produced stable ellipse fits and pupil size estimates that closely matched human-annotated ground truth across multiple high-quality human videos (Figure 4A). In contrast, for poor-quality videos in which pupil detection was challenged by reflections and occlusions, GM failed to detect the pupil. However, IM successfully tracked the pupil even under these extreme conditions, deviating from manually estimated ground truth (itself subject to variability) by an average of only 3 pixels (Figure 4B). Furthermore, we tested Pupil-DLC on two diverse and self-recorded human videos under natural light. While the GM failed, IM closely matched ground truth, and retraining GM on a small subset of frames (GM-RT) restored performance to levels comparable to IM (Figure 4). These results highlight the flexibility of Pupil-DLC in bridging controlled laboratory recordings and real-world human data. The ability of a model trained exclusively on mouse pupil videos to generalize to human recordings underscores shared visual and geometric features across species when imaging conditions are comparable. Importantly, the complementary roles of GM and IM, together with rapid retraining using minimal data, provide a practical strategy for maintaining high tracking fidelity under heterogeneous recording conditions. This adaptability positions Pupil-DLC as a scalable tool for extending pupillometry beyond tightly controlled settings and into diverse experimental and translational contexts.

A few limitations should be acknowledged. First, the present study focuses on head-fixed recordings acquired under specific infrared imaging geometry. Performance may differ in freely moving conditions, at extreme gaze angles, or under variable illumination. However, the availability of the IM provides a flexible mechanism to adapt the pipeline to new recording conditions through targeted retraining. Second, pupil size outputs are reported in pixel units. While videos from any setup can be directly analyzed, conversion to physical units and comparisons across rigs require calibration of camera geometry and appropriate scaling.

In summary, Pupil-DLC is an open-source, scalable pupillometry pipeline that combines high accuracy, minimal frame loss, tunable quality control, and cross-species generalization. By lowering technical barriers to robust pupil tracking, this framework enables reproducible measurement of pupil-linked brain state dynamics across diverse experimental contexts, spanning basic neuroscience in mice and translational applications in humans.

Code and Data Availability

Pupil-DLC is freely available online: https://github.com/Valyriverse/Pupil-DLC.

The manually annotated frames used to train the General Model of *Pupil-DLC* are available in the Figshare repository: Manually annotated frames: Seyfourian P., et al. (2026), Pupil-DLC: an open-source deep learning pipeline for scalable, markerless tracking of pupil dynamics across conscious and unconscious states (https://doi.org/10.6084/m9.figshare.31282714).

Acknowledgments

We sincerely thank Peter Ledochowitsch for generous feedback and extended discussions. We thank the Allen Institute Animal Care, the Neurosurgery and Behavior, and the Lab Animal Services teams for mouse husbandry, care, and habituation; the Allen Institute Manufacturing and Process Engineering team for experimental hardware and software support. We wish to thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement, and support.

We gratefully acknowledge funding from the Tiny Blue Dot Foundation (Santa Monica, California), the Templeton World Charity Foundation (TWCF-2022-30262 (I.R.), and the Allen Institute.

Competing Interests Statement

CK holds an executive position, and has a financial interest, in *Intrinsic Powers, Inc.,* a company whose purpose is to develop a device that can be used in the clinic to assess the presence of consciousness in patients. The remaining authors declare no competing interests.

References

1. Laeng, B., Sirois, S. & Gredebäck, G. Pupillometry: A Window to the Preconscious? *Perspect. Psychol. Sci.* **7**, 18–27 (2012).
2. Mathôt, S. Pupillometry: Psychology, Physiology, and Function. *J. Cogn.* **1**, 16 (2018).
3. Aston-Jones, G. & Cohen, J. D. AN INTEGRATIVE THEORY OF LOCUS COERULEUS-NOREPINEPHRINE FUNCTION: Adaptive Gain and Optimal Performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
4. Joshi, S., Li, Y., Kalwani, R. M. & Gold, J. I. Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron* **89**, 221–234 (2016).
5. McGinley, M. J., David, S. V. & McCormick, D. A. Cortical Membrane Potential Signature of Optimal States for Sensory Signal Detection. *Neuron* **87**, 179–192 (2015).
6. Reimer, J. *et al.* Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nat. Commun.* **7**, 13289 (2016).

470 7. Reimer, J. *et al.* Pupil Fluctuations Track Fast Switching of Cortical States during Quiet Wakefulness.
471 *Neuron* **84**, 355–362 (2014).
472 8. Bradley, M. M., Miccoli, L., Escrig, M. A. & Lang, P. J. The pupil as a measure of emotional arousal
473 and autonomic activation. *Psychophysiology* **45**, 602–607 (2008).
474 9. Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M. & Cohen, J. D. Pupil diameter tracks changes in control
475 state predicted by the adaptive gain theory of locus coeruleus function. *Cogn. Affect. Behav.*
476 *Neurosci*. **10**, 252–269 (2010).
477 10. Beatty, J. Task-evoked pupillary responses, processing load, and the structure of processing
478 resources. *Psychol. Bull.* **91**, 276–292 (1982).
479 11. Unsworth, N. & Robison, M. K. Pupillary correlates of lapses of sustained attention. *Cogn. Affect.*
480 *Behav. Neurosci.* **16**, 601–615 (2016).
481 12. McGinley, M. J. *et al.* Waking State: Rapid Variations Modulate Neural and Behavioral Responses.
482 *Neuron* **87**, 1143–1161 (2015).
483 13. Einhäuser, W., Stout, J., Koch, C. & Carter, O. Pupil dilation reflects perceptual selection and predicts
484 subsequent stability in perceptual rivalry. *Proc. Natl. Acad. Sci.* **105**, 1704–1709 (2008).
485 14. Einhäuser. Pupil dilation betrays the timing of decisions. *Front. Hum. Neurosci.*
486 https://doi.org/10.3389/fnhum.2010.00018 (2010) doi:10.3389/fnhum.2010.00018.
487 15. Naber, M., Alvarez, G. A. & Nakayama, K. Tracking the allocation of attention using human pupillary
488 oscillations. *Front. Psychol.* **4**, (2013).
489 16. Murphy, P. R., O'Connell, R. G., O'Sullivan, M., Robertson, I. H. & Balsters, J. H. Pupil diameter
490 covaries with BOLD activity in human locus coeruleus. *Hum. Brain Mapp.* **35**, 4140–4154 (2014).
491 17. Oddo, M. *et al.* Quantitative versus standard pupillary light reflex for early prognostication in
492 comatose cardiac arrest patients: an international prospective multicenter double-blinded study.
493 *Intensive Care Med.* **44**, 2102–2111 (2018).

494 18. Traylor, J. I. *et al.* Quantitative pupillometry in patients with traumatic brain injury and loss of
495 consciousness: A prospective pilot study. *J. Clin. Neurosci.* **91**, 88–92 (2021).
496 19. Figueroa, S. A., Olson, D. M., Kamal, A. & Aiyagari, V. Quantitative Pupillometry: Clinical Applications
497 for the Internist. *Am. J. Med.* **137**, 825–831 (2024).
498 20. Kret, M. E. & Sjak-Shie, E. E. Preprocessing pupil size data: Guidelines and code. *Behav. Res.*
499 *Methods* **51**, 1336–1342 (2019).
500 21. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep
501 learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
502 22. Claar, L. D. *et al.* Cortico-thalamo-cortical interactions modulate electrically evoked EEG responses
503 in mice. *eLife* **12**, RP84630 (2023).
504 23. Russo, S. *et al.* Thalamic feedback shapes brain responses evoked by cortical stimulation in mice and
505 humans. *Nat. Commun.* **16**, 3627 (2025).
506 24. Groblewski, P. A. *et al.* A standardized head-fixation system for performing large-scale, in vivo
507 physiological recordings in mice. *J. Neurosci. Methods* **346**, 108922 (2020).
508 25. Siegle, J. H. *et al.* Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*
509 **592**, 86–92 (2021).
510 26. Devor, M. & Zalkind, V. Reversible analgesia, atonia, and loss of consciousness on bilateral
511 intracerebral microinjection of pentobarbital. *Pain* **94**, 101–112 (2001).
512 27. Van Der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
513 28. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis.
514 *Nat. Methods* **9**, 671–675 (2012).
515 29. Arvin, S., Rasmussen, R. N. & Yonehara, K. EyeLoop: An Open-Source System for High-Speed, Closed-
516 Loop Eye-Tracking. *Front. Cell. Neurosci.* **15**, 779628 (2021).

30. Blasiak, T., Zawadzki, A. & Lewandowski, M. H. Infra-Slow Oscillation (ISO) of the Pupil Size of Urethane-Anaesthetised Rats. *PLoS ONE* **8**, e62430 (2013).

31. Tonsen, M., Zhang, X., Sugano, Y. & Bulling, A. Labelled pupils in the wild: a dataset for studying pupil detection in unconstrained environments. in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* 139–142 (ACM, Charleston South Carolina, 2016). doi:10.1145/2857491.2857520.