Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

PREPRINT

Validated inference of respiration from videography in mice

VrForaging Platform, Bruno Cruz, Laura Driscoll, Suyee Lee, Tiffany Ona, Trevor Ortega, Cindy Poo, Sharmishtaa Seshamani, Kyle Severson, Scott Sterrett, Olivia Oritse-SeunDede Unokesan, and Sonya Vasquez

Allen Institute for Neural Dynamics, Seattle, WA, USA [authors]
Authors after the VrForaging Platform are listed in alphabetical order.

Rodents rely on olfaction to survive. Odors guide them to find food, avoid predators, and recognize mates and kin. As mice actively sample odors via sniffing, the timing of each inhalation shapes odor perception. Ground-truth methods for measuring respiration in mice typically rely on sensors implanted in the nasal cavity, which require invasive surgery, and often fail to yield a reliable signal. Here, we developed a ground-truth-validated deep-learning method and show that the respiration waveform and the timing of each inhalation can be recovered directly from videography. We also provide a dataset of 560 curated sessions from 50 head-fixed mice performing an olfactory foraging task, in which behavioral videos are synchronized with a nasal epithelial temperature sensor that provided ground-truth measurement of respiration. Our method - Zephyr - accurately extracted the timing of inhalation events and recovered the respiration waveform with high fidelity in animals held out from training. This performance was robust across different animals, recording conditions, and camera angles. Trained on as few as 4 mice, Zephyr already generalized to new animals, and fine-tuning it on video of 2 mice adapted it to a new task and new camera views, making it feasible for adoption by standard behavioral and systems neuroscience experiments and labs. We propose that Zephyr, along with our curated dataset and quality-control tools, is a viable alternative to existing methods for measuring respiration.

Introduction

Sniffing is a fundamental form of active sensing. Each inhalation draws odors into the nose, so the timing of sniffing sets when olfactory information is sampled (Kepecs et al. 2006; Verhagen et al. 2007; Wachowiak 2011). Respiration in mice is approximately 2–4 Hz at rest and up to 12–15 Hz during active sniffing (Wesson et al. 2008). Olfactory bulb neurons respond at precise phases of this cycle (Shusterman et al. 2011). Respiration also reflects an animal's internal state and changes with arousal and emotion (Del Negro et al. 2018; Homma & Masaoka 2008). Neurons of the breathing rhythm generator themselves promote arousal (Yackle et al. 2017), and fear-conditioned mice breathe at a steady 4 Hz while they freeze (Bagur et al. 2021). The time of inhalation is therefore a natural reference event for olfactory coding, and respiration would provide valuable state information in many experiments on awake mice beyond studies of olfaction.

Measuring respiration with this precision in a behaving mouse usually requires an implanted sensor. The standard methods place either a temperature-sensitive probe or a pressure cannula in the nasal epithelium or cavity (Grimaud & Murthy 2018; McAfee et al. 2016; Reisert et al. 2014; Wesson et al. 2008). These methods, along with plethysmography, piezoelectric sensors, and electromyography,

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

each involve trade-offs in timing precision, freedom of movement, or invasiveness (Grimaud & Murthy 2018). In addition, signals from chronically implanted sensors degrade over time, and are often discarded after inspection (Rafilson et al. 2025; Weinreb et al. 2024). A measure of respiration from video would spare animals this surgery, a refinement in the sense of the 3Rs (Tannenbaum & Bennett 2015). It would also allow extraction of respiration signals from head-fixed experiments that already have readily available videography of the face.

The face carries signatures of respiration. Respiration and orofacial movements are coordinated (Kleinfeld et al. 2014; Moore et al. 2013), and machine vision now allows quantitative measurements of behaviour (Mathis et al. 2018; Pereira et al. 2022; Weinreb et al. 2024). Motion from face video has also been leveraged to explain neural activity (Musall et al. 2019; Stringer et al. 2019; Syeda et al. 2024). Yet not all facial motion is respiration. The nose, whisker pad, tongue, and jaw are also involved in various orofacial movements, such as turning, whisking, licking, chewing, and grooming, and the nose can make rhythmic movements independent from inhalation (Kurnikova et al. 2017). Current tools that label a "sniffing" component of facial motion do so without validation against ground-truth sensor data (Hakim et al. 2025; Syeda et al. 2024). Thermal imaging detects temperature changes associated with inhalations in head-restrained mice but requires specialized thermal cameras (Esquivelzeta Rabell et al. 2017; Mutlu et al. 2018), and conventional cameras have yielded only respiration rate in anaesthetized or sleeping animals (Breuer et al. 2023; Geuther et al. 2022; Kunczik et al. 2019). Networks that recover pulse and respiration from human face video (Chen & McDuff 2018; Liu et al. 2020; Liu et al. 2023; Yu et al. 2019; Yu et al. 2022; Zou et al. 2025) were built for slow, steady signals, whereas a mouse changes its respiration rate several-fold within a second. To our knowledge, no validated method has been shown to reconstruct the respiration waveform and timing of each inhalation from conventional videography in rodents, and no public dataset exists to develop and test such a method.

To fill this gap, we assembled a dataset that pairs multi-view videos with synchronized nasal epithelial temperature sensor (thermistor) recordings from 560 sessions in 50 head-fixed mice on 15 behavioral rigs. We tested computer-vision and deep learning methods for recovering respiration from the video. Our network, which we call Zephyr, not only recovered the timing of each inhalation in held-out animals but was also able to faithfully reconstruct the thermistor waveform. Zephyr performed well when trained across several camera views, and, after fine-tuning on a small amount of new data, generalized to new experimental conditions. Finally, we show that curated quality-control scores computed from Zephyr's output alone can be used to evaluate each prediction, generating confidence metrics of respiration signals extracted from videography in place of surgically implanted sensors. We release the dataset, the code, as well as trained models, for others to use and fine-tune on their own data.

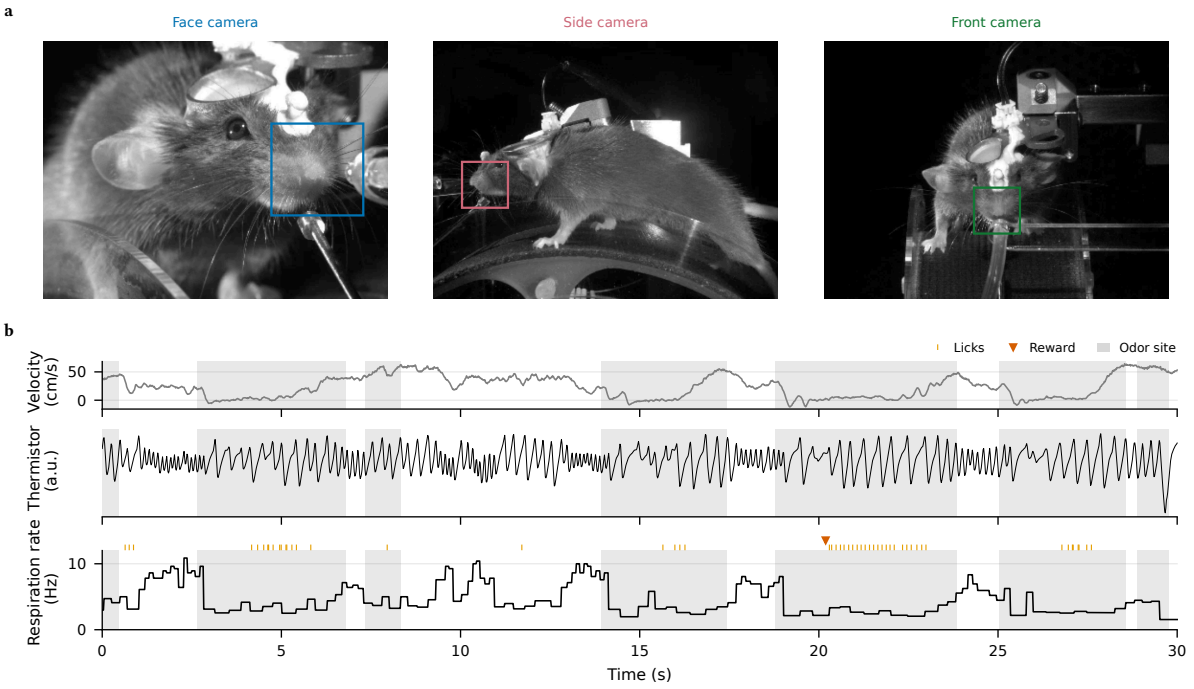
Results

A dataset of orofacial video paired with nasal epithelial temperature sensor recordings

We trained head-fixed mice to perform an olfactory virtual-reality foraging task on a circular treadmill: thirsty mice traversed a virtual corridor with odor-cued sites and could stop at a site to collect a probabilistic reward (Methods). A site was rewarded if the mouse stopped and received water, unrewarded if it stopped and received none, and skipped if it did not stop. Three cameras filmed the animal, the face and front cameras through a hot mirror in front of the screen (Figure 1a; Figure S1), and a thermistor implanted over the nasal epithelium recorded respiration (Methods). The task drew a wide range of respiration within each session, from breathing at rest to fast sniffing (Figure 1b).

After quality control of the thermistor (Methods, Figure S2), the dataset holds 560 sessions from 50 mice on 15 rigs, about 93 h of video per camera (Figure 1c). Respiration spanned the range of the mouse: 26% of cycles were at 2–4 Hz, 54% at 4–8 Hz and 20% at 8–15 Hz (Figure 1d).

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9



c

Dataset	
Animals	50 head-fixed mice
Sessions	560, two 300-s clips each
Video	≈93 h per camera
Rigs, span	15 rigs over 13 months
Cameras	face, side and front; 240 or 360 Hz, read at 60 Hz
Thermistor	nasal epithelium, 200–250 Hz
Respiration	26% at 2–4 Hz, 54% at 4–8 Hz, 20% at 8–15 Hz
Each fold	
Train	31–33 animals
Validation	9 animals
Held out	9–10 new animals; new sessions of 26–30 training animals

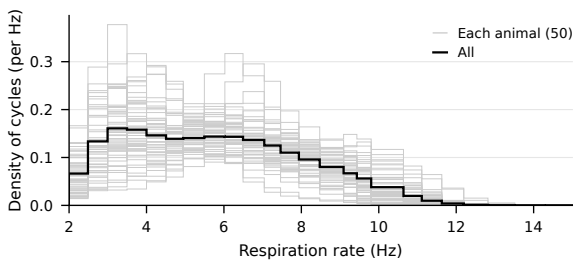


Figure 1: Behavioral video paired with measurements of respiration from the temperature of the nasal cavity. a, One session seen by the face, side and front cameras. Boxes, the crop each camera's network reads. **b,** Thirty seconds of one session of a held-out animal: running velocity, the thermistor, and the respiration rate counted from its inhalations, with licks (ticks), reward (triangle) and odor sites (shading). **c,** The dataset and the sets of each cross-validation fold (Methods Table 3). **d,** Respiration rate over the thermistor's cycles at 2–15 Hz, for the whole dataset (black; n = 1,614,211 cycles) and each animal (grey; n = 50).

Neural networks recover respiration from video

Our goal was to infer respiration from conventional videography alone, so that respiration can be measured in animals without an implanted sensor. A useful method must recover both the respiration waveform and the timing of each inhalation in animals it was never trained on. First, we asked whether the signals that current tools extract from face video can track respiration. We scored inhalation timing with the F1 score, which balances the fraction of detected inhalations that were real (precision) against the fraction of real inhalations that were detected (recall); a detection counted as correct if it fell within 25 ms of a thermistor inhalation, as defined by a peak detector on the thermistor signal (Methods). The most common summary of facial movement is motion energy, the mean absolute difference between frames (Musall et al. 2019; Stringer et al. 2019). Even when we fit its sign and time lag to the thermistor on the training clips, motion energy failed to predict respiration in held-out mice (r = 0.03, inhalation F1 0.32 Figure 3a,b; Table 1). A linear readout of the leading facial motion principal components, as in Facemap (Syeda et al. 2024), did better but still recovered only part of the signal (r = 0.48, inhalation F1 0.51).

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Deep-learning methods have been shown to recover periodic physiological signals from human video. We thus wondered if this class of methods would perform better on held-out mice and sessions. We started from the state-of-the-art in camera-based physiological measurement methods, four networks developed to recover pulse and respiration from human face video: TS-CAN (Liu et al. 2020), PhysNet (Yu et al. 2019), PhysFormer (Yu et al. 2022) and RhythmFormer (Zou et al. 2025). We trained each on the same clips under the same protocol, reading in a 96 × 96 grayscale crop over the snout from the face camera view, and predicting the thermistor waveform (Methods). All of them performed significantly above FaceMap and managed to recover the respiration waveform to a high degree of accuracy. Neverthless, these networks are designed for long, slowly varying signals. The two transformers are, in addition, large and slow to run.

We therefore propose a smaller and more efficient architecture, Zephyr (Figure 2a; Table 4) we reason would be better suited for our goals of recovering inhalation times. Zephyr takes a 96 × 96 grayscale crop over the snout, and its first convolutional layer spans three consecutive frames, so that the encoder sees short-range motion and not only appearance. A small convolutional neural network (CNN) embeds each crop, and the embeddings are interpolated by their timestamps onto a uniform 60 Hz grid. A non-causal temporal convolutional network (TCN) with dilated residual blocks then models respiration over time. Its receptive field of 4.2 s lets each prediction draw on the breaths before and after it. The TCN feeds two heads, one that outputs the respiration waveform and one that predicts inhalation onsets directly, trained against Gaussians centred on the thermistor inhalations. The networks we compare against predict only the waveform, and they need a separate digital signal processing step to extract onsets (Methods).

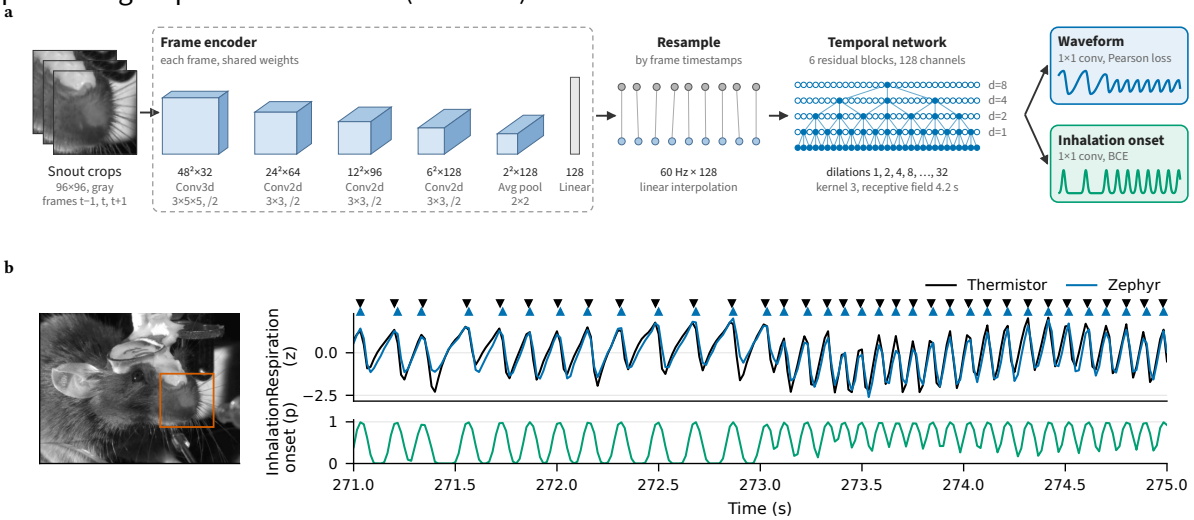


Figure 2: Zephyr. a, The network (Methods). A frame encoder and a temporal network feed two heads: the waveform head predicts the respiration waveform, and the inhalation-onset head the probability that an inhalation starts at each time step. **b,** Four seconds of a clip of a held-out animal. The face frame beside it shows the crop the network sees (orange). The predicted respiration (blue) is drawn over the thermistor (black), above the inhalation-onset probability (green). Ticks mark the thermistor's and Zephyr's inhalations.

To measure how well methods generalize, for all analyses, we held out whole animals, and blocks of sessions of animals from the training set (Methods, **Data splits**).

We scored each method on held-out animals, never seen in training, in a five-fold cross-validation in which each of 48 animals is held out once (Methods, Table 3). Because generalization depends on both the animal and the recording setup (Figure S3), we treat each animal × rig combination as one unit of comparison: scores were averaged within each one of these units before comparison (107 units; Methods, **Statistics**). Every network recovered respiration on held-out animals (Figure 3b; Table 1; an example in Figure 2b), so respiration can indeed be learned from face video in mice. Zephyr and PhysNet showed similar inhalation detection performance. Their F1 scores within 25 ms at 60 Hz were 0.96 (Zephyr; 95% interval 0.95–0.97) and 0.94 (PhysNet; paired difference 0.014, 0.002–0.028). Their

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

waveform correlations with the thermistor were also similar: 0.81 for Zephyr (95% interval 0.79–0.83) and 0.79 for PhysNet (difference 0.02, −0.01 to 0.05); TS-CAN, RhythmFormer, and PhysFormer reached 0.65–0.78. Importantly, while PhysNet and Zephyr nearly matched each other in accuracy, Zephyr processed 66% more frames per second than the full 3D CNN PhysNet on a consumer GPU, and 509–544% more than the two transformers (Figure 3b; Table S2). Its inhalations were both complete (0.96) and precise (0.96). Thus, neural networks can accurately and efficiently recover respiration waveforms from face video.

Mouse respiration spans a wide range of rates, so we asked whether accuracy held across that range (Figure 3c). Zephyr found almost every inhalation between 6 and 12 Hz (F1 0.99) and slightly fewer above 12 Hz (0.95). It was least accurate in slow breathing, at 2–3 Hz (0.89). Splitting F1 into its two parts shows different errors at the two ends: at 2–3 Hz Zephyr both added inhalations (precision 0.88) and missed them (recall 0.91), while above 12 Hz its inhalations were nearly all real (precision 0.99) but some were missed (recall 0.91 Figure 3d). This is not for lack of slow cycles to learn from: as a share of the cycle, Zephyr times slow inhalations more precisely than fast ones, and the fixed 25-ms tolerance is what is strictest at slow rates (see below and Figure S7). The other networks lost more accuracy at both ends of the range.

The ability to recover inhalation events with precise timing is crucial for behavior and physiology studies. Over the held-out animals of all folds, Zephyr's inhalations lagged the thermistor's by 0.2 ms on average (95% interval over folds −1.6 to 1.7 ms), well within one cycle at the fastest sniffing rates (Figure 3e). The lag hardly changed with respiration rate, from +1.5 ms at 2–3 Hz to −1.1 ms at 11–12 Hz. We also compared respiration rate, counted in 10-s windows, with the thermistor's, a measurement commonly reported for camera-based respiration in humans (Liu et al. 2023; Villarroel et al. 2019). Zephyr's rate was off by 0.07 Hz (4.2 cycles per minute) on average, or 1.6% of the thermistor's rate, with virtually no overall bias (+0.007 Hz; Figure 3f).

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

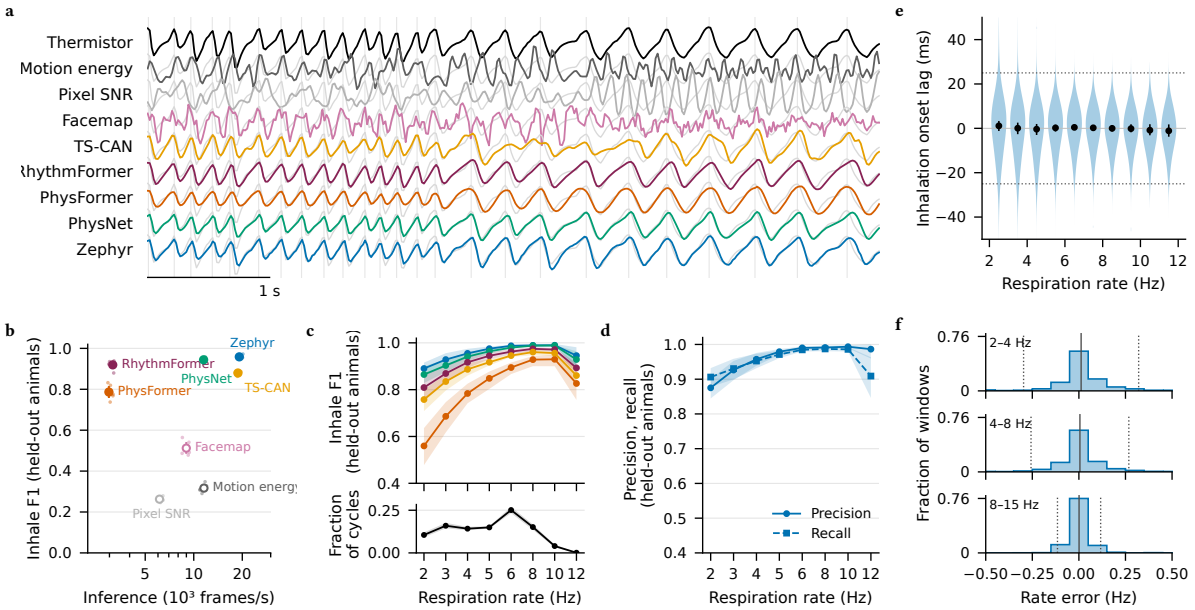


Figure 3: Neural network methods faithfully recover respiration dynamics. a, Six seconds of one clip of a held-out animal (face camera), predicted by each method. The thermistor is at the top and each method's waveform is drawn over it in grey. Vertical lines mark the thermistor's inhalations. Panels **b–f** show the results from held-out animals across folds, 1,038 clips from 519 sessions of 48 animals. **b,** Inhalation F1 on new animals against inference speed. Filled symbols are networks on a consumer GPU and open symbols are methods on a CPU. **c,** Inhalation F1 on held-out animals by respiration rate for each network, mean over folds with its 95% interval. Colors are as in **b**. The bottom panel shows the distribution of respiration rates in the held-out data. **d,** Zephyr's inhalation precision and recall by respiration rate, mean over folds with its 95% interval, on the same axis as **c**. **e,** Timing of Zephyr's inhalations relative to the thermistor's (predicted − thermistor) by respiration rate. Violins show every thermistor inhalation that had a prediction within 50 ms, pooled over the held-out animals of all folds (n = 1,505,851). Dots show the mean over folds with its 95% interval. Dotted lines mark the ±25-ms matching tolerance. **f,** Error of Zephyr's respiration rate (Zephyr − thermistor) in 10-s windows, pooled over the held-out animals of all folds (n = 31,136 windows), split by the thermistor's rate: 2–4 Hz (top), 4–8 Hz and 8–15 Hz (bottom). Because rates are counted in 10-s windows, errors come in steps of 0.1 Hz. Solid lines mark the mean error (bias) and dotted lines the 95% limits of agreement.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Table 1: Method comparison on held-out animals from the face camera view. Each score was averaged within each animal × rig unit, then over the held-out-animal units of a split, then over splits. Values are the mean with its 95% interval over the 5 cross-validation folds (107 units of 48 animals, each animal new in one fold). Respiration rate is counted in 10-s windows. MAE and RMSE are in cycles per minute, MAPE is relative to the thermistor's rate, and r is the Pearson correlation of the windowed rates. Waveform r is the correlation of the predicted respiration waveform with the thermistor. F1 counts events within 25 ms. Inference is the time in seconds for a 5-min clip from the preprocessed crops on a consumer GPU, or on a CPU where marked (*).

			Respiration rate (10-s windows)				Waveform and events		
Method	**Splits**	**Inference (s)**	**MAE ↓**	**RMSE ↓**	**MAPE (%) ↓**	**r ↑**	**Waveform r ↑**	**Inhale F1 ↑**	**Exhale F1 ↑**
Zephyr (ours)	5	**0.94**	**4.4** 3.8–5.0	**7.3** 6.4–8.2	**1.7** 1.4–1.9	**0.99** 0.99–0.99	**0.81** 0.79–0.83	**0.96** 0.95–0.97	**0.76** 0.68–0.82
PhysNet (Yu et al. 2019)	5	1.57	4.6 3.8–5.8	7.7 6.6–9.2	**1.7** 1.4–2.2	**0.99** 0.98–0.99	0.79 0.76–0.82	0.94 0.93–0.95	0.74 0.66–0.80
PhysFormer (Yu et al. 2022)	5	6.07	10.6 9.9–11.2	14.9 14.0–15.7	3.8 3.6–4.0	0.96 0.96–0.97	0.71 0.70–0.72	0.79 0.76–0.82	0.62 0.59–0.65
RhythmFormer (Zou et al. 2025)	5	5.74	5.3 4.7–5.9	8.4 7.6–9.3	1.9 1.7–2.2	0.98 0.98–0.99	0.78 0.75–0.81	0.92 0.90–0.93	0.71 0.66–0.76
TS-CAN (Liu et al. 2020)	5	0.96	11.9 10.1–13.7	16.2 14.0–18.6	4.6 3.8–5.4	0.95 0.92–0.97	0.65 0.62–0.68	0.88 0.87–0.88	0.68 0.64–0.74
Facemap (Stringer et al. 2019; Syeda et al. 2024)	5	2.01*	148.7 139.3–160.5	164.0 154.8–175.0	59.2 54.4–64.9	0.31 0.27–0.36	0.48 0.44–0.51	0.51 0.48–0.54	0.47 0.45–0.49
Pixel SNR	5	2.94*	98.0 92.9–102.9	112.4 106.9–117.8	38.3 34.9–41.6	0.04 −0.02–0.09	0.00 −0.01–0.00	0.26 0.25–0.27	0.26 0.25–0.27
Motion energy (Musall et al. 2019; Stringer et al. 2019)	5	1.56*	157.8 152.3–163.0	170.1 164.9–175.0	62.4 59.1–65.7	0.19 0.13–0.26	0.03 0.02–0.03	0.32 0.30–0.34	0.32 0.31–0.33

Zephyr generalizes to new data

Because our goal is to measure respiration without implanted sensors, a useful videography-based method for inferring respiration must generalize to non-implanted animals absent from its training dataset. We first considered a typical scenario, in which a laboratory has recorded animals in a similar setup and wishes to apply a trained network to new data acquired under similar, experimentally controlled conditions. Besides whole animals, each cross-validation fold held out blocks of consecutive sessions of its training animals (Methods, Table 3). All networks recovered respiration from these held-out sessions with very high waveform correlation and inhalation F1, and, although relatively lower, their scores on held-out animals remained high (Figure 4a; Table 1, Table S3). We emphasize that this generalization depends on both the animal and the recording setup (Figure S3). Zephyr scored the highest among the networks on both new sessions (inhalation F1 0.98, waveform correlation 0.90) and held-out animals (0.96 and 0.81). Inhalation F1 generalized more robustly than waveform correlation: from new sessions to new animals, Zephyr's F1 fell only from 0.98 to 0.96, against 0.90 to 0.81 for its waveform correlation, and the vast majority of held-out samples scored above 0.9 (Figure 4b).

Since not all experiments record the same camera view, we next asked whether respiration can be extracted from the other recorded angles. Each session of our dataset was filmed by three cameras, allowing us to compare the performance of Zephyr across viewpoints. Besides the face camera, a side and a front camera film the whole animal from farther away and at different angles (Figure S6). Trained with the same protocol, Zephyr recovered respiration from both views on new animals, though less accurately than from the face camera (Figure 4d,e; Figure S5, Table S4).

Networks trained on large and diverse data often generalize better than specialized ones, as foundation models for vision (Oquab et al. 2024) and for animal pose estimation pooled across many laboratories and setups (Ye et al. 2024) have shown. We therefore asked whether one network could learn all three views. Performance of a single network trained on the pooled dataset of all three

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

cameras was not different from the network trained on each camera alone, on any of the three cameras (Figure 4f) [fold cv_0 only so far; add the numbers once all folds are scored].

We next experimented with how much data Zephyr required to generalize effectively to new animals. Keeping the held-out animals of one fold fixed, we varied the number of training animals from 2 to 32, each with all their sessions (Figure 4c). Despite a clear trend of improvement with more animals, 2 animals already gave an inhalation F1 of 0.88 against 0.97 with all 32. As the number of animals decreased to 2, average performance did not drop dramatically but became more variable across the 5 draws (F1 0.91, 0.94, 0.90, 0.89 and 0.75), suggesting that the space of inputs was not adequately represented in such a small training set. We also note that the duration of video mattered less than the number of unique animals. All 32 animals with a quarter of their sessions matched the full training set (F1 0.96). Together, these results show that Zephyr can be trained from different views and generalizes effectively to unseen data from similar experimental conditions without requiring extensive training data.

However, whether the network generalizes to data from different experimental setups remained an open question. We used an independent dataset (Su et al. 2026), in which mice performed a dynamic foraging task where the faces of head-fixed mice were filmed from the side and from below (Figure 4g), to evaluate how well Zephyr pre-trained on our dataset performed on these sessions. On a held-out animal, it was not able to directly track respiration (inhalation F1 0.10 from the side and 0.06 from below). Given our previous success on generalizing to new animals with little data, we next asked whether a small amount of additional training videos from this new setup could improve performance. We fine-tuned the network: training continued from its weights on the original training set together with two animals recorded in the dynamic foraging task, and we predicted on the third, held-out animal (Methods). Both measures improved markedly from both camera angles, to an inhalation F1 of 0.80 and 0.76 and a waveform correlation of 0.82 and 0.80 (Figure 4h,i), while the fine-tuned network did not lose performance on the initial dataset (Figure S9). This result supports the idea that while Zephyr may not generalize to unseen setups out of the box, it can be fine-tuned with little data to adapt to new experimental conditions.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

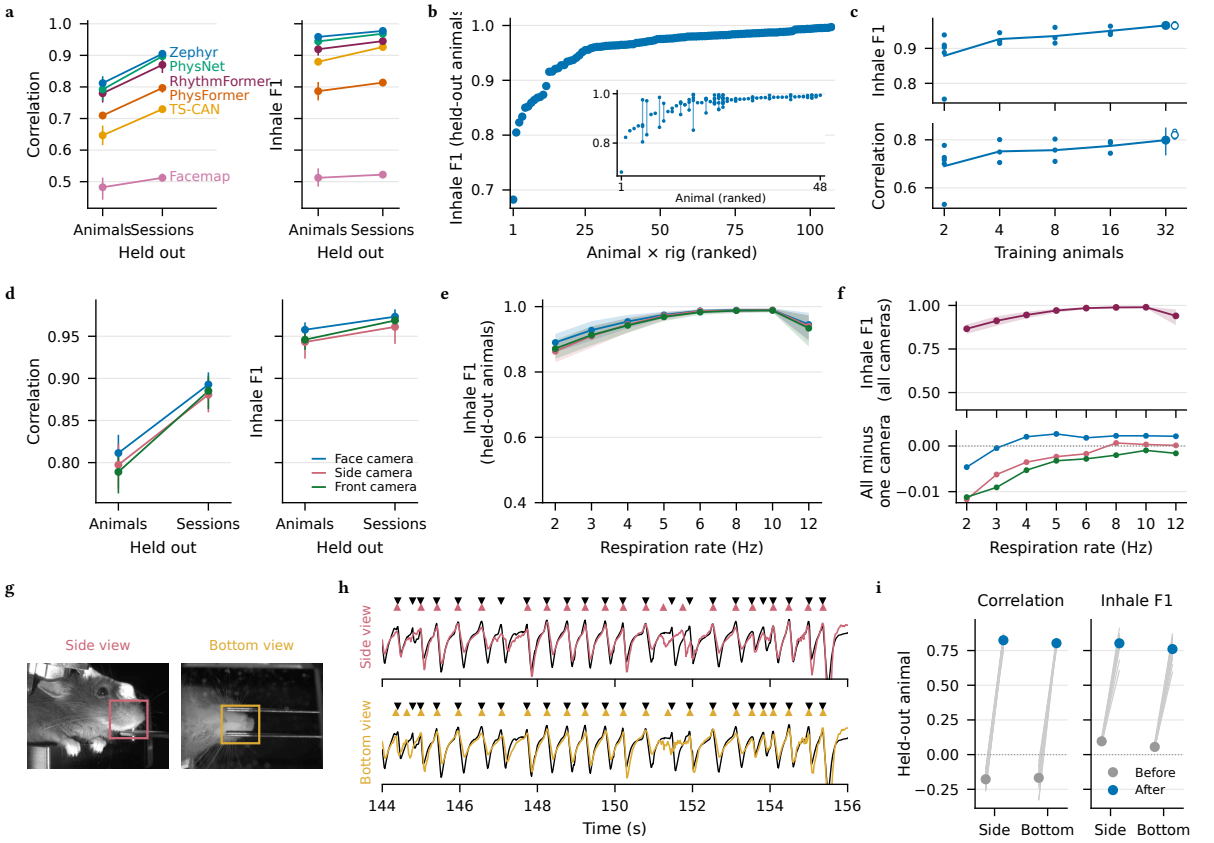


Figure 4: Generalization to new animals, camera views, and experiments. a, Waveform correlation and inhalation F1 of each method on held-out animals and held-out sessions (107 and 86 animal × rig units). Colors as in Figure 3. **b,** Inhalation F1 of held-out videos, ranked; inset ranks animals, and lines join one animal's rigs. **c,** F1 scores on the same held-out data for networks trained on increasingly more animals of the same fold. The open circles indicate networks trained on all animals with a quarter of their sessions. **d,** Zephyr trained on the face, side or front camera (Figure S5). **e,** Inhalation F1 of each camera's network by respiration rate (colors as in **d**). **f,** One network trained on all three camera views. Inhalation F1 by respiration rate (top) and its difference from each single-camera network (bottom). **g–i,** The network of **f** fine-tuned on a dynamic-foraging experiment. **g** F1 scores of a held-out animal (3 animals in total with "leave-one-out") on its side and bottom views with their crops. **h** 12 s of one clip with the ground-truth thermistor (black), and the output of the fine-tuned network (colored as **g**). **i** correlation and inhalation F1 before and after fine-tuning, per held-out animal.

Quality-control on predictions without ground-truth sensor data

A method that replaces a sensor will be used where no sensor can check it, so errors must be predictable and calibrated. We therefore sought to understand the conditions under which Zephyr's predictions are reliable and how to quantify its confidence in the absence of ground-truth data.

By far the most common errors were relatively small timing mismatches between predicted and thermistor inhalations, rather than missed or hallucinated inhalations. A predicted inhalation counts as correct if it falls within a matching tolerance (by default 25 ms) of a thermistor inhalation. Relaxing this tolerance from 25 to 50 ms raised Zephyr's inhalation F1 from 0.96 to 0.99, and the remaining errors fell from 8.2% to 2.4% of cycles (Figure 5a,b).

Timing also explained most of the loss of accuracy in slow breathing (Figure 3c). At 2–3 Hz, Zephyr found 91% of the thermistor's inhalations within 25 ms but 98% within 50 ms. Its timing spread was wider in slow breathing in milliseconds (18 ms at 2–3 Hz against 13 ms at 6–8 Hz) but narrower as a share of the respiratory cycle (5% against 9%): with a tolerance of 10% of each cycle, Zephyr found 97% of inhalations at 2–3 Hz and 66% at 10–12 Hz, where 10% of the cycle is shorter than one video

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

frame. A fixed 25-ms tolerance is thus strictest for slow breathing, which Zephyr times more precisely than fast breathing relative to the cycle. Some of the remaining false inhalations in slow breathing may be shallow cycles that the thermistor detector does not label reliably. Relatedly, lowering the ground-truth detector's prominence threshold added 3.0% more inhalations at 2–3 Hz and 0.0% at 6–8 Hz (Figure S7). In other words, depending on the scientific question, temporal precision may be sacrificed for overall detection sensitivity, especially at lower breathing rates.

Even if the sources of most errors are known, we still need ways to assess the quality of the predictions. For instance, without having the thermistor raw data, one would like to be able to distinguish between the predictions provided by pre-trained and fine-tuned networks on the dynamic foraging experiments (add figure).

We calculated metrics (see methods) that we find reliably predict inference quality. Briefly, on each 30-s window of a prediction we computed four measures that need no ground truth: 1. how well the inhalation onset head's inhalations agree with the peaks of the network's own waveform (head vs waveform F1); 2. how often inhalations and exhalations fail to alternate; 3. how high the inhalation onset peaks are; and 4. how high the inhalation onset probability stays between them (Methods, Table 7). Using validation data, we calibrated these scores. Next, we generated surrogate "bad" predictions by systematically adding Gaussian pixel noise to the videos (standard deviation of 16, 32 and 64 gray levels), which progressively lowers the network's accuracy. Finally, for comparison we also add the scores of the dynamic foraging dataset. We found that the scores tracked the true accuracy. With noise of 16 gray levels, inhalation F1 barely changed (0.93 against 0.97 on validation animals), and the scores stayed close to their validation range. At 32 gray levels and beyond, F1 collapsed (0.22), and every score moved away from that range: head vs waveform F1, for instance, fell from 0.98 to 0.53 (Figure 5d,e). The dynamic-foraging predictions behaved the same way. Before fine-tuning, the network found almost none of that experiment's inhalations (true F1 0.07), its head vs waveform F1 fell to 0.67 and the inhalation onset probability between inhalations rose from 0.02 to 0.18. After fine-tuning (as above), the same network scored as well on these videos as it did on our dataset (Figure 5c,e). The proposed metrics can thus separate predictions that work from known failures without implanting a thermistor.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

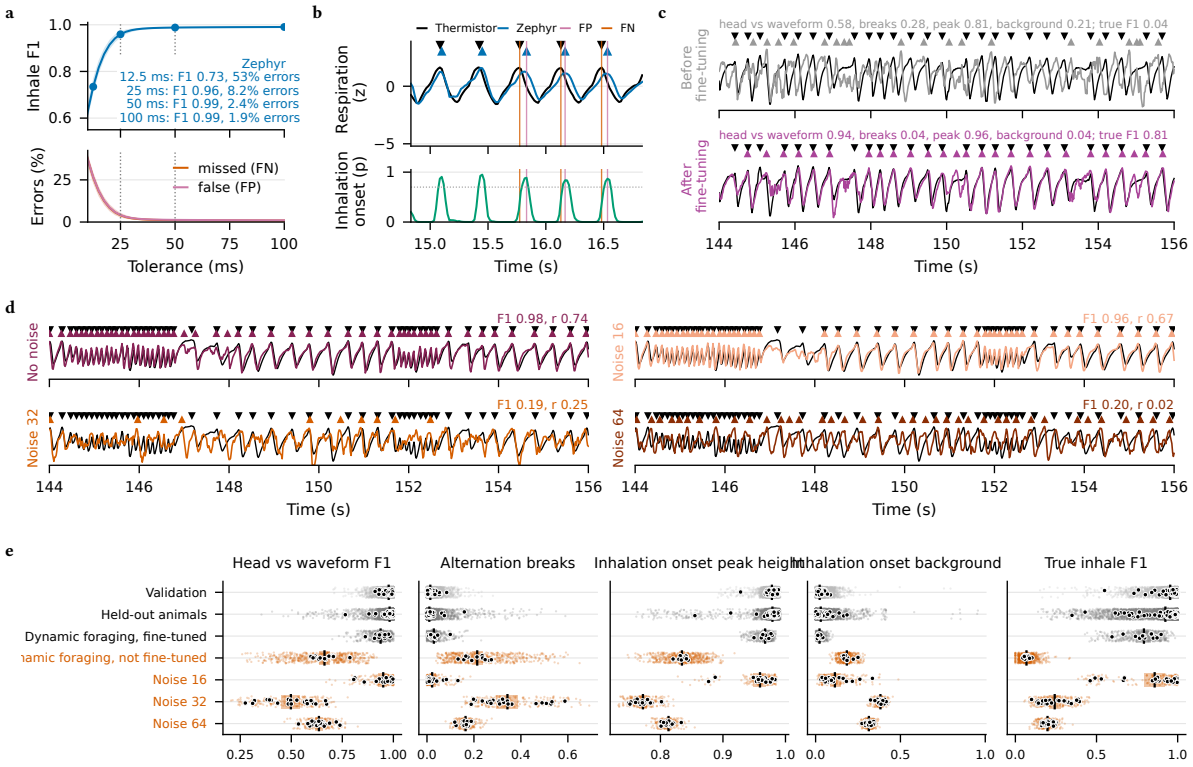


Figure 5: Zephyr predictions can be scored without ground truth. a, Zephyr's inhalation F1 on held-out animals as the matching tolerance is relaxed (top), and the misses (FN) and false inhalations (FP) left (bottom). **b,** A clip with many near misses, predictions 25–100 ms from a thermistor inhalation, which count as both a miss and a false inhalation (r = 0.89, F1 0.86). **c,** Twelve seconds of one dynamic-foraging clip predicted by the pooled network before and after fine-tuning, over the thermistor (black). **d,** Twelve seconds of one clip of a held-out animal with progressively more Gaussian noise added (std = 16, 32 and 64). **e,** The quality-control scores and the true inhalation F1 over 30-s windows of the pooled network's validation animals. Faint points are single windows, black dots each session's median across its windows, black ticks the mean of those medians, and shaded bars its 95% bootstrap interval over sessions. Bands in **a** are 95% intervals over the 5 folds (Methods).

Respiration in experiments without a thermistor

Finally, we used the network for its intended purpose - measuring respiration in experiments without a thermistor. We ran inference on all views from 19 animals never implanted with a thermistor, but otherwise running in the same experimental conditions. Over their 2,274 30-s windows, the quality-control scores lay largely within the range of good predictions and away from the known failures (Figure 6a). Nevertheless, outlier 30s windows were observed occasionally (Figure 6a). Anecdotally, these scores were related to individual windows and not entire 1-hour sessions. The example in Figure 6b shows how performance in an individual window drops during a grooming bout where the snout is not fully visible. This highlights how the metrics can not only be used to evaluate full sessions, but also reject individual windows with relatively lower confidence.

We then asked whether the predicted respiration carries the same behavioural structure as the thermistor's. In one session (60 min), we aligned the predicted inhalations to the task as we had the thermistor's (Figure 1b). The predicted waveform followed the transitions between running, stopping at odor sites and licking (Figure 6c), and the respiration rate around odor onset differed by the site's outcome as in the thermistor session, both in this session and on average over all 19 animals (Figure 6d,e). The predictions thus carry the relationship between respiration and behavior.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

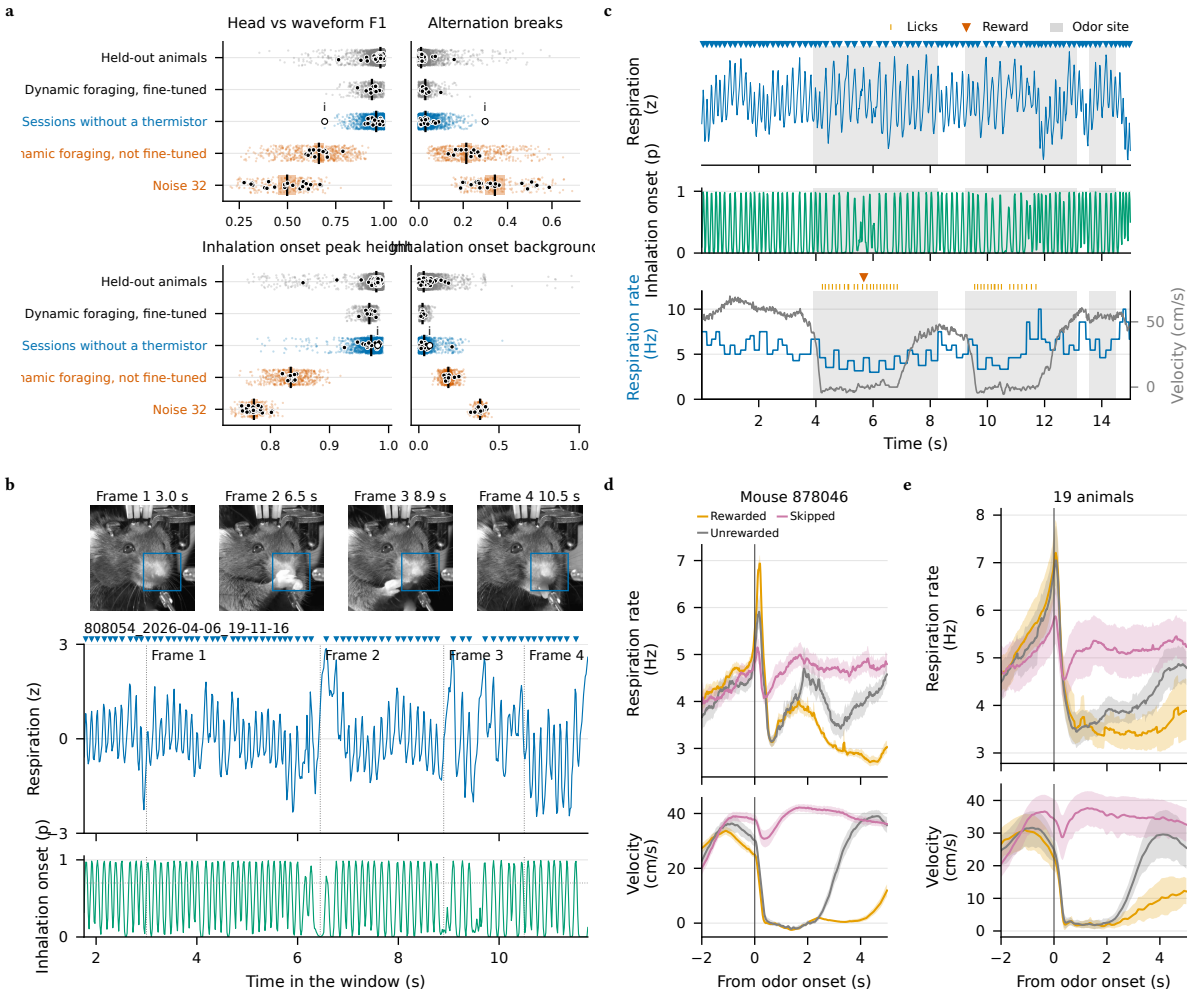


Figure 6: Inferring respiration without ground-truth data. a, Quality-control scores over the 2,274 30-s windows of the 19 sessions - from 19 animals, beside predictions as in Figure 5e. A white circle marks the window that scored the lowest on "head vs waveform" F1 metric (i). **b,** 10 s of the window marked in **a**, centred on four frames of its face video with the crop box (top; frame 1, the face at rest; frames 2 and 3, paw strokes over the snout; frame 4, a lick, its tongue out). Dotted lines mark the frames' times. **c,** Fifteen seconds of one animal's session showing (top) predicted waveform, (middle) predicted inhalation onset probability, and (bottom) the predicted respiration rate (blue) aligned to features of behavior: running velocity (grey), licks (yellow tick), reward (orange) and odor inhalation onset (gray shading). **d, e,** Respiration rate and running velocity around odor onset, broken down by site outcome. **d,** The session of **c** (n = 188, 90 and 150 sites, respectively; mean over sites with its 95% bootstrap interval). **e,** All sessions, averaged over sessions (n = 2,772, 1,973 and 3,171 sites; mean over sessions of each session's mean, with its 95% bootstrap interval over sessions).

Discussion

We provide a dataset that pairs orofacial video with a nasal epithelial temperature sensor in behaving mice, and a network Zephyr that successfully recovers inhalation in animals it has never seen. It recovered the thermistor waveform and the timing of each inhalation onset with high accuracy and temporal precision. We show that signals that current behavioural tools extract from facial motion did not faithfully recover respiration, so a "sniffing" component of facial motion should not be equated with respiration without a sensor to validate it.

Our method has implications past the immediate context of neuroscience experiments. Reading respiration from video also bears on animal welfare. A nasal implant requires invasive surgery on the

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

snout, can obstruct airflow and damage the nasal epithelium, and can fail before an experiment even begins (Grimaud & Murthy 2018; McAfee et al. 2016). In our own hands, several sessions had to be discarded once the thermistor's signal degraded and was no longer reliable. Once trained for a rig, the network generalizes to new animals, which then need no implant. This refines the procedure for those animals and, where implants would otherwise fail, reduces the number of animals used (Tannenbaum & Bennett 2015), and simplifies otherwise complex experimental procedures. Respiration can be inferred from face videos from prior experiments without recording any new animals, provided the experimental conditions match those of the training data.

The proposed network architecture Zephyr is a small network trained to predict respiration data from head-fixed rigs and is not intended to be a general model of facial motion. Video from one rig varies little: the camera, lens, lighting and head position stay the same from session to session, and a small network has the capacity for it. Zephyr matched larger networks from camera-based physiology measurement tools in accuracy and with improved inference speeds. The same architecture also performed well on input from the side-view and front-facing cameras, which see the snout from farther away and at other angles, and a single network trained on all three cameras matched each single-camera network on its own camera. Accuracy depended more on the number of training animals than on the amount of video. Nevertheless, we achieved relevant accuracy with as few as 2 or 4 animals. Whether one network can maintain this accuracy across rigs, lenses, strains and tasks, or each rig needs a network of its own, remains open. Our rigs share one design and place the cameras in similar positions, yet their lighting and the cameras' exact position, angle and lens still differ (Figure S6), and the network scored lower on some rigs and animals than on others (Figure S3). The quality-control measures score each window of a recording without a sensor and, calibrated against scores of predictions known to work and known to fail, allow the user to set their own thresholds.

Adopting Zephyr therefore requires little beyond the videography many head-fixed laboratories already collect. A laboratory first runs the pre-trained network on its own recordings and uses the quality-control measures to check whether the predictions are reliable for its setup. If they are not, the network can be fine-tuned with a small amount of ground-truth data from the laboratory's own rig: a few animals recorded with video and a nasal epithelial temperature sensor (2 to 4 animals sufficed in our experiments). Once fine-tuned, Zephyr runs efficiently on a consumer-grade GPU, and no sensor is needed in subsequent animals.

The dataset also limits what we can claim. All mice were head-fixed, so the camera saw the face from a fixed distance and angle, and the animal could not turn away, rear or groom freely. Whether respiration can be read from video in freely moving mice remains untested. Head-mounted cameras, which already film the eyes and face of freely moving mice during natural behaviour and electrophysiology (Meyer et al. 2018; Michaiel et al. 2020; Parker et al. 2022; Sattler & Wehr 2021), keep the face at a fixed distance and angle, and are a natural setting in which to test it. The thermistor measures the temperature of the airflow, not the airflow itself, so its own delay relative to nasal pressure bounds the timing error we can measure.

As opposed to pose-estimation methods, where a human can manually correct and validate labels, annotation of breathing from video is much less straightforward and, at least in our hands, impossible to perform reliably. This presents a challenge since predictions cannot be easily verified against ground truth in cases where the thermistor is not implanted. To address this, our approach relies on quality-control measures that score each window against a calibrated reference, ensuring that only reliable predictions are considered. These are not hard cutoffs but rather provide a continuous measure of confidence that can guide the user in selecting trustworthy segments of the video and should be considered against the scientific context of the experiment.

Future methods may need fewer labels and generalize better to new experimental rigs. Thermistor recordings are costly to collect, whereas video of the face is recorded in many experiments that measure no respiration. Self-supervised methods, which in humans recover pulse from unlabeled face video by exploiting the periodicity of the signal (Gideon & Stent 2021; Sun & Li 2022), could learn from such video and need a sensor only to check the result. In pose estimation, models trained on data pooled from many laboratories already track animals in new settings with little or no further

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

labelling (Ye et al. 2024), and general-purpose visual foundation models (Oquab et al. 2024) may provide features from which respiration can be read with fewer labels. Camera physiology in humans advanced on shared datasets and benchmarks recorded across groups (Liu et al. 2023), and networks that excel on one dataset often fail on another. We release the dataset in a format upon which other labs' recordings can be added, so that a benchmark spanning rigs, tasks, strains and freely moving animals can be built from recordings that pair a sensor with video of the face.

Data availability

We release the dataset (face, side and front video, frame timestamps and the thermistor, on one clock) with the splits we used and the trained weights.

Code availability

The code for Zephyr (training, inference, scoring and the quality-control measures) is at https://github.com/AllenNeuralDynamics/zephyr, with an application that runs the network on new recordings: it places a crop box on each recording, predicts respiration and reports the quality-control measures. The reimplementations of the comparison methods, the experiment configurations, an evaluation that reproduces the scores reported here, and the notebooks that generate every figure, table and reported value of this paper from the results are at https://github.com/AllenNeuralDynamics/zephyr-manuscript. Data were acquired with the VrForaging platform (https://github.com/AllenNeuralDynamics/Aind.Behavior.VrForaging), using Bonsai (Lopes et al. 2015) and Harp devices (https://harp-tech.org) (Carvalho et al. 2024).

AI declaration

The code for Zephyr, the reimplementations of the comparison methods, and the notebooks used to analyse the data and generate the figures (all written in Python) were developed iteratively with Claude Opus (Anthropic) acting as a coding assistant. The authors designed the network, the experiments, the comparisons and the analyses, and specified them in detail; Claude implemented them in code under the authors' direction. The authors reviewed and tested this code and verified that it performs the intended analyses and that the outputs accurately represent the underlying data. The scientific questions, the interpretation of the results and the conclusions are the authors' own. The authors take full responsibility for the content of this manuscript.

Acknowledgments

We thank Lukas Braun for discussions on the design and implementation of the experiments. We also thank Shalaika Akella for reviewing a draft of the manuscript.

Funding

This work was funded by the Allen Institute [grants, if any].

Competing interests

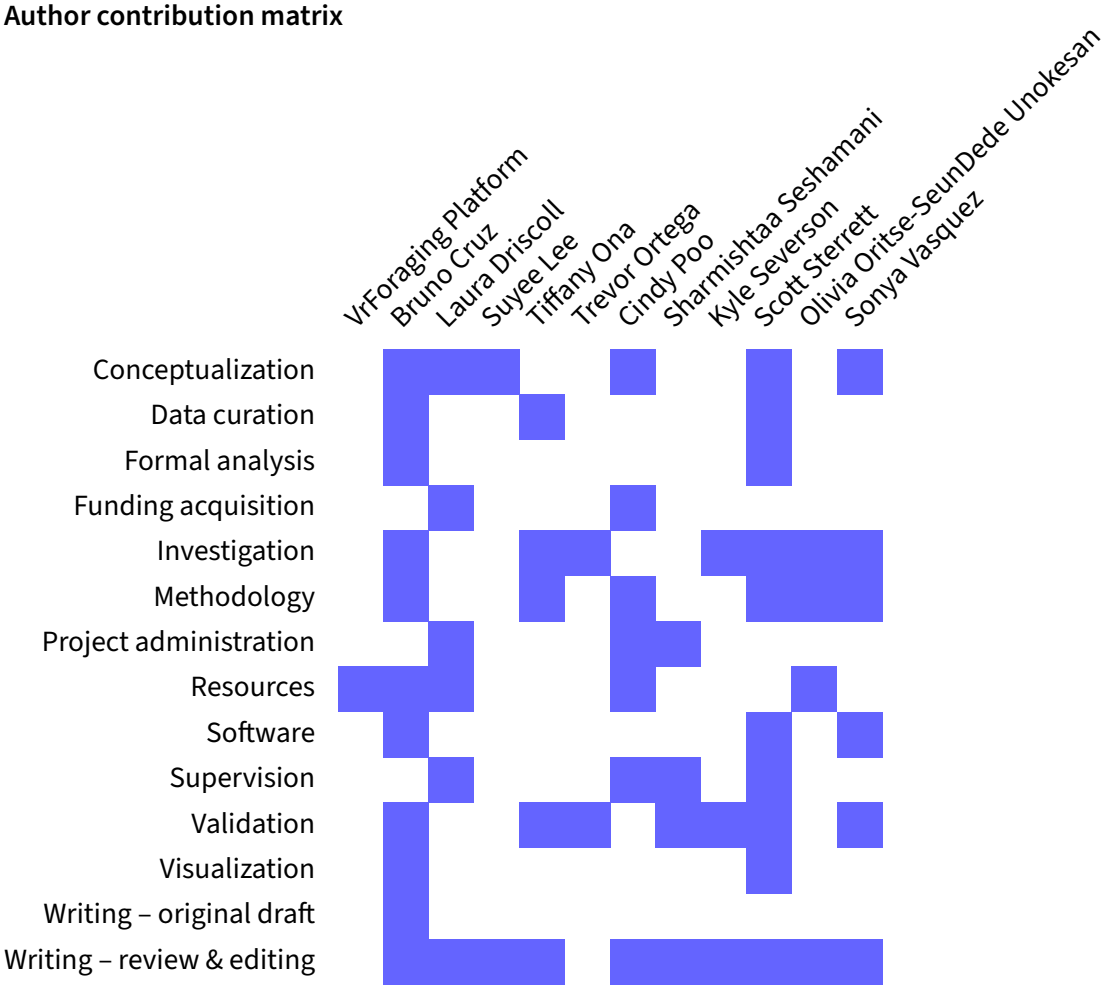
The authors declare no competing interests [confirm with every author].

Author contributions

Fill in the author contributions using the CRediT taxonomy.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Author contribution matrix



Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

References

Bagur, S., J. M. Lefort, M. M. Lacroix, G. de Lavilléon, C. Herry, M. Chouvaeff, C. Billand, H. Geoffroy, & K. Benchenane. 2021. "Breathing-driven prefrontal oscillations regulate maintenance of conditioned-fear evoked freezing independently of initiation." *Nature Communications* 12: 2605.

Breuer, L., L. Mösch, J. Kunczik, & others. 2023. "Camera-based respiration monitoring of unconstrained rodents." *Animals* 13: 1901.

Carvalho, F., A. Silva, B. Cruz, J. Frazão, & G. Lopes. 2024. *Harp: A standard for reactive and self-synchronising hardware for behavioural research.* Zenodo.

Chen, W., & D. McDuff. 2018. "DeepPhys: video-based physiological measurement using convolutional attention networks." In *European Conference on Computer Vision (ECCV)*.

Del Negro, C. A., G. D. Funk, & J. L. Feldman. 2018. "Breathing matters." *Nature Reviews Neuroscience* 19: 351–367.

Esquivelzeta Rabell, J., K. Mutlu, J. Noutel, & others. 2017. "Spontaneous rapid odor source localization behavior requires interhemispheric communication." *Current Biology* 27: 1542–1548.

Geuther, B., M. Chen, R. J. Galante, & others. 2022. "High-throughput visual assessment of sleep stages in mice using machine learning." *Sleep* 45: zsab260.

Gideon, J., & S. Stent. 2021. "The way to my heart is through contrastive learning: remote photoplethysmography from unlabelled video." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 3975–3984.

Grimaud, J., & V. N. Murthy. 2018. "How to monitor breathing in laboratory rodents: a review of the current methods." *Journal of Neurophysiology* 120: 624–632.

Hakim, R., A. Jaggi, G. Heo, & others. 2025. "Spectral envelopes of facial movements predict intention, cortical representations, and neural prosthetic control." *bioRxiv*.

Homma, I., & Y. Masaoka. 2008. "Breathing rhythms and emotions." *Experimental Physiology* 93: 1011–1021.

Kepecs, A., N. Uchida, & Z. F. Mainen. 2006. "The sniff as a unit of olfactory processing." *Chemical Senses* 31: 167–179.

Kleinfeld, D., M. Deschênes, F. Wang, & J. D. Moore. 2014. "More than a rhythm of life: breathing as a binder of orofacial sensation." *Nature Neuroscience* 17: 647–651.

Kunczik, J., C. Barbosa Pereira, L. Zieglowski, & others. 2019. "Remote vitals monitoring in rodents using video recordings." *Biomedical Optics Express* 10: 4422–4436.

Kurnikova, A., J. D. Moore, S.-M. Liao, M. Deschênes, & D. Kleinfeld. 2017. "Coordination of orofacial motor actions into exploratory behavior by rat." *Current Biology* 27: 688–696.

Liu, X., J. Fromm, S. Patel, & D. McDuff. 2020. "Multi-task temporal shift attention networks for on-device contactless vitals measurement." In *Advances in Neural Information Processing Systems*.

Liu, X., G. Narayanswamy, A. Paruchuri, & others. 2023. "rPPG-Toolbox: deep remote PPG toolbox." In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.

Lopes, G. et al. 2015. "Bonsai: an event-based framework for processing and controlling data streams." *Frontiers in Neuroinformatics* 9: 7.

Mathis, A., P. Mamidanna, K. M. Cury, & others. 2018. "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning." *Nature Neuroscience* 21: 1281–1289.

McAfee, S. S., M. C. Ogg, J. M. Ross, & others. 2016. "Minimally invasive highly precise monitoring of respiratory rhythm in the mouse using an epithelial temperature probe." *Journal of Neuroscience Methods* 263: 89–94.

Meyer, A. F., J. Poort, J. O'Keefe, M. Sahani, & J. F. Linden. 2018. "A head-mounted camera system integrates detailed behavioral monitoring with multichannel electrophysiology in freely moving mice." *Neuron* 100: 46–60.

Michaiel, A. M., E. T. T. Abe, & C. M. Niell. 2020. "Dynamics of gaze control during prey capture in freely moving mice." *eLife* 9: e57458.

Moore, J. D., M. Deschênes, T. Furuta, & others. 2013. "Hierarchy of orofacial rhythms revealed through whisking and breathing." *Nature* 497: 205–210.

Musall, S., M. T. Kaufman, A. L. Juavinett, S. Gluf, & A. K. Churchland. 2019. "Single-trial neural dynamics are dominated by richly varied movements." *Nature Neuroscience* 22: 1677–1686.

Mutlu, K., J. Esquivelzeta Rabell, P. Martin del Olmo, & S. Haesler. 2018. "IR thermography-based monitoring of respiration phase without image segmentation." *Journal of Neuroscience Methods* 301: 1–8.

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

Parker, P. R. L., E. T. T. Abe, E. S. P. Leonard, D. M. Martins, & C. M. Niell. 2022. "Joint coding of visual input and eye/head position in V1 of freely moving mice." *Neuron* 110: 3897–3906.

Pereira, T. D., N. Tabris, A. Matsliah, & others. 2022. "SLEAP: a deep learning system for multi-animal pose tracking." *Nature Methods* 19: 486–495.

Rafilson, S., N. Gonzales Hess, T. M. Findley, & M. C. Smear. 2025. "Challenges in inferring breathing rhythms from olfactory bulb local field potentials." *Chemical Senses* 50.

Reisert, J., G. J. Golden, K. Matsumura, M. Smear, D. Rinberg, & A. Gelperin. 2014. "Comparing thoracic and intra-nasal pressure transients to monitor active odor sampling during odor-guided decision making in the mouse." *Journal of Neuroscience Methods* 221: 8–14.

Sattler, N. J., & M. Wehr. 2021. "A head-mounted multi-camera system for electrophysiology and behavior in freely-moving mice." *Frontiers in Neuroscience* 14: 592417.

Shusterman, R., M. C. Smear, A. A. Koulakov, & D. Rinberg. 2011. "Precise olfactory responses tile the sniff cycle." *Nature Neuroscience* 14: 1039–1044.

Stringer, C., M. Pachitariu, N. Steinmetz, & others. 2019. "Spontaneous behaviors drive multidimensional, brainwide activity." *Science* 364: eaav7893.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Su, Z. et al. 2026. “Topographic structure and function of locus coeruleus noradrenaline neurons.” *Nature*.

Sun, Z., & X. Li. 2022. “Contrast-Phys: unsupervised video-based remote physiological measurement via spatiotemporal contrast.” In *European Conference on Computer Vision (ECCV)* 492–510.

Syeda, A., L. Zhong, R. Tung, & others. 2024. “Facemap: a framework for modeling neural activity based on orofacial tracking.” *Nature Neuroscience* 27: 187–195.

Tannenbaum, J., & B. T. Bennett. 2015. “Russell and Burch's 3Rs then and now: the need for clarity in definition and purpose.” *Journal of the American Association for Laboratory Animal Science* 54: 120–132.

Verhagen, J. V., D. W. Wesson, T. I. Netoff, J. A. White, & M. Wachowiak. 2007. “Sniffing controls an adaptive filter of sensory input to the olfactory bulb.” *Nature Neuroscience* 10: 631–639.

Villarroel, M., S. Chaichulee, J. Jorge, & others. 2019. “Non-contact physiological monitoring of preterm infants in the Neonatal Intensive Care Unit.” *npj Digital Medicine* 2: 128.

Wachowiak, M. 2011. “All in a sniff: olfaction as a model for active sensing.” *Neuron* 71: 962–973.

Weinreb, C., J. E. Pearl, S. Lin, & others. 2024. “Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics.” *Nature Methods* 21: 1329–1339.

Wesson, D. W., T. N. Donahou, M. O. Johnson, & M. Wachowiak. 2008. “Sniffing behavior of mice during performance in odor-guided tasks.” *Chemical Senses* 33: 581–596.

Yackle, K., L. A. Schwarz, K. Kam, J. M. Sorokin, J. R. Huguenard, J. L. Feldman, L. Luo, & M. A. Krasnow. 2017. “Breathing control center neurons that promote arousal in mice.” *Science* 355: 1411–1415.

Ye, S., A. Filippova, J. Lauer, S. Schneider, M. Vidal, T. Qiu, A. Mathis, & M. W. Mathis. 2024. “SuperAnimal pretrained pose estimation models for behavioral analysis.” *Nature Communications* 15: 5165.

Yu, Z., X. Li, & G. Zhao. 2019. “Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks.” In *British Machine Vision Conference (BMVC)*.

Yu, Z., Y. Shen, J. Shi, H. Zhao, P. Torr, & G. Zhao. 2022. “PhysFormer: facial video-based physiological measurement with temporal difference transformer.” In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zou, B., Z. Guo, J. Chen, & others. 2025. “RhythmFormer: extracting patterned rPPG signals based on periodic sparse attention.” *Pattern Recognition*.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Methods

Data collection

Surgery. All procedures were approved by the Allen Institute Institutional Animal Care and Use Committee [protocol number; mice: strain, sex, age at surgery, housing and light cycle; analgesia and post-operative care]. Thermistors were implanted over the nasal epithelium as described (McAfee et al. 2016). Briefly, under isoflurane anaesthesia (1–5%), a 0.5-mm burr hole was thinned through the nasal bone 3.1 mm anterior to the frontonasal suture and 0.5 mm right of the midline, and the thermistor tip was lowered onto the epithelium without pressing on it. The hole was sealed with silicone (Kwik-Cast), and the probe and a head bar were fixed with dental cement (Metabond).

Data acquisition. Data were acquired on the VrForaging platform (https://github.com/AllenNeuralDynamics/Aind.Behavior.VrForaging), whose devices are synchronized by the Harp protocol (https://harp-tech.org) (Carvalho et al. 2024). For this dataset, a Behavior board (https://github.com/harp-tech/device.behavior) triggered the 240- or 360-Hz FLIR cameras, a custom-built olfactometer (https://github.com/harp-tech/device.olfactometer) delivered odors and a sniff-detector board (described below) recorded the thermistor. Three cameras filmed the animal: the side camera directly, and the face and front cameras through a hot mirror placed between the animal and the visual-stimulus screen, about 14 cm in front of the animal. Video was encoded in real time with H.264 on the GPU (NVENC, constant quality 18) and re-encoded offline on the CPU (libx264, CRF 18). Clips were cut from these files by stream copy.

For the “dynamic foraging” experiment, a similar setup to the one described previously (Su et al. 2026) was used. Animals rested in a metal tube and were asked to lick left or right to collect probabilistic rewards. Cameras were triggered at 504 Hz by a similar Behavior board and respiration data collected with an identical sniff-detector.

Sniff-detector board

TODO SONYA

Dataset construction

Sessions were selected with the rules of Table 2, in order. Thermistor quality asserted in 30 second windows. 5min clips were made from stretches where all windows cleared the quality criteria.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Table 2: Dataset selection. Rules are listed in the order they were applied, with the number of sessions each removed. Of 4,120 recorded sessions, 1,591 passed the thermistor rules and 1,506 the clip and camera rules; capping each animal's sessions on each rig, so that no animal dominates a set, removed most of the rest. A repeated event is an inhalation followed by another inhalation, or an exhalation by another exhalation.

Stage	Rule	Threshold	Removed
Window (30 s)	Signal amplitude (5th–95th percentile range)	≥ 100 raw units	2,038
	Clipped samples	0% at ≤ 20 or ≥ 4,090	—
	Cycles outside 2–15 Hz	≤ 10%	—
	Repeated events (I→I or E→E)	≤ 5%	—
Session	Dead sensor (low SNR)	Heads ≥ 3 dB apart	—
	Repeated events over the session	≤ 0.3%	257
	Consistent events (rate, alternation, amplitude)	Robust $z \leq 3$	1
	Longest run of good windows	≥ 300 s	229
	Trace rejected on manual		4
Clips: camera	Rule 10, validation 3, held-out animals 5; held-out sessions up to 5, at most a third of		910

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Thermistor events. Inhalation and exhalation onsets were detected on the thermistor at its native rate (200-250 Hz) with an adaptive peak detector. The signal was filtered (zero-phase second-order Butterworth, high-pass 1 Hz, low-pass 40 Hz) and z-scored over the recording. Maxima at least 50 ms apart were candidate inhalation onsets, and minima candidate exhalation onsets. A candidate was kept when its prominence was at least 0.2 of the local range of the signal (5th to 95th percentile over a centred 1.5-s window). These events were used for training, quality control and scoring.

Data splits

To test generalization to held-out data, for each fold, we split the data into training, validation and test. Test data was further split into two categories: held-out sessions (sessions from animals seen during training but on different days) and held-out animals (animals never seen during training). We drew five splits that remained constant throughtout the manuscript for all methods/networks. The cross-validation folds are described in Supplementary Note 1.

Table 3: Split sizes. Values are the minimum and maximum over the 5 cross-validation folds. Each of the 48 animals is a held-out animal in one fold.

Set	Held out as	Animals	Sessions
Train	—	31–33	213–268
Validation	whole animals	9	87–144
Held-out animals	whole animals	9–10	75–128
Held-out sessions	sessions of training animals	26–30	70–119
Total		50	560

Preprocessing

For the face camera, frames were downsampled to 360 × 270 pixels, cropped to a 96 × 96-pixel box over the snout (one per session), converted to gray and temporally resampled to 60 Hz by taking every Nth frame (4 or 6). We note that the network also performs well at 30 Hz (Figure S4) but with slightly reduced accuracy. For the side and front cameras, which film the animal from farther away, frames were kept at the native 720 × 540 resolution instead, with the same 96 × 96-pixel box over the snout.

Network architecture

A convolutional encoder, whose first layer spans three frames, embeds each crop (Table 4); embeddings are interpolated by their timestamps onto a uniform 60-Hz grid, and a non-causal temporal convolutional network (receptive field 4.2 s) outputs the respiration waveform and an inhalation onset logit at each step.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Table 4: Network architecture. BN, batch normalization.

Stage	Layer	Output
Input	gray crop, three consecutive frames	3 × 96 × 96
Encoder	Conv3D 3×5×5, stride (1, 2, 2), BN, GELU	32 × 48 × 48
	Conv 3×3, stride 2, BN, GELU	64 × 24 × 24
	Conv 3×3, stride 2, BN, GELU	96 × 12 × 12
	Conv 3×3, stride 2, BN, GELU	128 × 6 × 6
	average pool to 2×2, flatten, linear, GELU	128 per frame
Resampling	linear interpolation by frame timestamps	128 at 60 Hz
Temporal	Conv 1×1	128
	6 residual blocks of 2 × (Conv 3, dilation d, BN, GELU) and dropout; d = 1, 2, ..., 32	128
Heads	Conv 1×1 (waveform), Conv 1×1 (inhalation onset logit)	1 + 1

Training

The loss was 1 − r between the predicted and target waveforms, averaged over three time scales (the waveform, and copies average-pooled by 4 and 16 samples), plus 0.5 times a binary cross-entropy on inhalation onsets against Gaussians (s.d. 20 ms) centred on the thermistor inhalations. Other settings are in Table 5; the checkpoint with the highest mean correlation on the validation clips was kept. Each network was trained once per fold on a NVIDIA Tesla T4. The comparison networks were trained the same way, each with the loss of its original paper (Supplementary Note 2).

Table 5: Training settings of Zephyr.

Setting	Value
Input	gray, 96 × 96 crop, 60-Hz selection grid
Window, batch	512 samples (8.5 s), 16
Optimizer	AdamW, learning rate 6×10^{-4}, weight decay 10^{-4}
Schedule	200-step linear warm-up, cosine decay to 1%
Regularization	dropout 0.1 (temporal network), gradient clip 1, weight EMA 0.999
Loss	multi-scale 1 − r (scales 1, 4, 16) + 0.5 × inhalation onset BCE
Inhalation onset target	Gaussians, s.d. 20 ms, per-batch positive weight
Augmentation	frame jitter ±0.25 frame, shift ±4 px, zoom ±10%, brightness ±15%, contrast ±20%, noise 2 grey levels
Selection	mean correlation on validation animals, every 4 epochs
Stopping	patience 12 scorings, 40–200 epochs of 100 steps
Precision	fp16 mixed precision (bf16 is not native on the T4)

Inference and event detection

Clips were predicted in overlapping windows of 1,024 samples, discarding 126 samples at each edge, so that the stitched waveform equals a single pass over the clip. Inhalation onsets were detected as the peaks of the inhalation onset probability above a threshold, at least 50 ms apart. The threshold was chosen for each trained network on its own validation animals by sweeping the threshold from 0.20 to 0.85 in steps of 0.05, and scoring by the mean inhalation F1 over the validation clips (**Scoring**). Exhalations, and the inhalations of methods without an inhalation onset output (“DSP”), were found on the predicted waveform by the detector used on the thermistor, with the same settings (the 40-Hz low-pass does not apply at 60 Hz).

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Benchmarked methods

Every method read the same 96 × 96 gray crops at 60 Hz. The two pixel methods compute a waveform from each clip's pixels with no For each clip we set the sign and lag (up to 0.5s) that maximized the mean correlation with the thermistor over the training clips. We kept the most correlated sign and average lag and applied them to the test clips.

The methods, their re-implementation and what we changed are in Table 6. The four networks were trained in our training loop on the losses of their own papers (Supplementary Note 2); none has an inhalation onset output, so their inhalations were found on the predicted waveform.

Table 6: Benchmarked methods. For methods that are not networks, the details give their settings. For the networks, they give only what we changed; otherwise the networks follow their papers. All networks read one gray channel instead of color, and a respiration band replaces the heart-rate band.

Method	Source	Details
Motion energy	(Musall et al. 2019; Stringer et al. 2019)	Motion energy: mean absolute difference between consecutive frames over the crop
Pixel SNR	this work	Band-passed pixels weighted by their share of variance in the respiration band, signed to agree, averaged
Facemap	(Stringer et al. 2019; Syeda et al. 2024)	Top 100 principal components of each of the crop's motion energy and raw frames (binned by 2), at 11 lags (±250 ms); ridge readout, strength chosen leaving one training recording out
TS-CAN	(Liu et al. 2020)	One gray channel; output band-passed to 1–15 Hz
PhysNet	(Yu et al. 2019)	One gray channel; 96 × 96 crop
PhysFormer	(Yu et al. 2022), authors' code	One gray channel; crop upsampled to 128 pixels; rate spectrum over 1–15 Hz
RhythmFormer	(Zou et al. 2025), authors' code	One gray channel; crop upsampled to 128 pixels; rate spectrum over 1–15 Hz

Timing. Inference time was the wall-clock time from preprocessed crops in memory to the 60-Hz waveform, median over 8 5-min clips after a warm-up, at batch size 1 in bf16 on an NVIDIA RTX 4060 Ti (non-network methods on the CPU, an Intel Core i9-13900K with 24 cores and 64 GB of RAM).

Scoring

We scored predictions against ground truth thermistor signals over the span that both covered. Waveforms were scored by taking the Pearson correlation between the two signals. Events were scored by matching predicted and thermistor events within a tolerance window and computing precision, recall, and F1 score. An event was considered a match if it fell within 1.5 frames of the prediction (25 ms at 60 Hz). We paired each event one to one by matching with an optimal assignment on their time differences using the Hungarian algorithm. Matched pairs were true positives (TP), unmatched predictions false positives (FP; false inhalations) and unmatched thermistor events false negatives (FN; missed inhalations; an example of both in Figure 5b). Precision, TP /(TP + FP), is the share of predicted inhalations that are real; recall, TP /(TP + FN), is the share of the thermistor's inhalations found; and their harmonic mean, F_1 = 2TP /(2TP + FP + FN), was computed per clip. Finally, respiration rate was simply the number of inhalations counted in 10-s windows.

Spatial and temporal receptive fields

For occlusion, a 12 × 12-pixel square of the 96 × 96-pixel crop, moved in 6-pixel steps, was set to the training mean values in every frame and recorded the drop in performance. For gradients, we took the absolute gradient of the inhalation onset logit at each thermistor inhalation with respect to every input pixel, averaged over frames for the spatial map and summed over pixels for the temporal profile. The same network with its weights re-initialised served as a control. The temporal profile was calculated as the share of an inhalation's gradient carried by each frame, in 1/60-s bins (Figure S8).

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Generalization experiments

Training-set size experiments. We trained Zephyr from scratch on parts of the training set of fold cv_4 (Figure 4c): nested subsets of 2, 4, 8 and 16 of its 32 training animals with all their sessions (15, 33, 65 and 130 sessions), and 65 sessions drawn from all 32 animals. Each experiment was trained 3 times on a different draw of subset of trials. For the 2 animals condition, 5 networks were trained instead to account for the larger variability.

Fine-tuning on a different task. The pooled network of the first fold was fine-tuned on video of "dynamic-foraging" experiments. Training started from the pooled network's checkpoint, with half of each batch from the fold's training clips of all three cameras of the dataset, and half from the 8 sessions of 2 animals of that experiment, at a learning rate of 1.5×10^{-4}, with the other settings of Table 5. Early stopping and the inhalation onset threshold used the fold's validation animals. Both the fine-tuned network and the untouched pooled network were scored on the 52 clips (13 sessions) of a single third animal of that experiment.

Quality control

Each prediction was cut into 30-s windows, and each window was scored with the measures of Table 7. Each network was calibrated against validation animals by computing the distribution of the metrics against the F1 scores of the validation dataset. A report gives each window's measures and their percentiles among the reference's windows.

Table 7: Quality-control measures. *Shown in Figure 5e and Figure 6a every measure is in Figure S10.

Measure	What it measures in a window	Better
Inhalations	Number of inhalations from the inhalation onset output	—
Respiration rate	Median rate between consecutive inhalations (Hz)	—
Breaths too fast	Share of breaths faster than a mouse breathes (15 Hz); slower rates are not judged, as they change with running, odors and task	Lower
Alternation breaks*	Share of events (inhalations from the inhalation onset output, exhalations from the waveform, in time order) that repeat the kind before them; breathing alternates	Lower
Flat breaths	Share of breaths (inhalation to inhalation) whose waveform swings less than a quarter of the clip's median breath: stretches where the waveform sees nothing	Lower
Head vs waveform F1*	Agreement between the inhalation onset output's inhalations and the peaks of the network's own waveform (thermistor's detector, 1.5-frame tolerance): the two outputs should agree	Higher
Inhalation onset peak height*	Median inhalation onset probability at the inhalations: how sure the inhalation onset output is	Higher
Ambiguous inhalation onset peaks	Share of inhalation onset-probability peaks above 0.2 that stay below the inhalation onset threshold	Lower
Inhalation onset background*	Mean inhalation onset probability further than 100 ms from any inhalation: how quiet it is between breaths	Lower
Expected precision	Mean probability that an inhalation is real given its peak height, from a calibration fitted on the reference's validation clips	Higher

Sessions without a thermistor

The face videos of 19 task sessions of 19 mice without a thermistor were predicted with the network trained on all three cameras of the first fold, at its inhalation onset threshold. Other than the thermistor signal, all data acquisition progressed identically to the original dataset. The side and front videos of the same sessions (Figure S11) were predicted the same way.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Statistics

We choose to use the animal × rig unit as the primary unit of analysis since we found significant variability between rigs of the same animal.(Figure S3) Scores were averaged over a unit's clips, then over a fold's units, and reported as the mean over the 5 folds with a 95% bootstrap interval over folds; differences between methods are paired by fold. Analyses of a single fold use a percentile bootstrap over units (2,000 draws).

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Supplementary information

Supplementary Note 1: Cross-validation folds

Table S1 gives the number of animals and sessions in each set of the five cross-validation folds (Methods, **Data splits**).

Table S1: Cross-validation folds. Each cell gives the number of animals / sessions in one set of one fold. Animals with held-out sessions are also training animals, so animals do not add up across sets.

Set	cv_0	cv_1	cv_2	cv_3	cv_4
Training	31 / 240	32 / 265	32 / 213	33 / 268	32 / 260
Validation	9 / 88	9 / 104	9 / 144	9 / 88	9 / 87
Held-out animals	10 / 124	10 / 75	10 / 128	9 / 103	9 / 89
Held-out sessions	28 / 103	27 / 111	26 / 70	27 / 96	30 / 119
Not scored (359 Hz)	2 / 5	2 / 5	2 / 5	2 / 5	2 / 5

Supplementary Note 2: Comparison networks

TS-CAN (Liu et al. 2020), PhysNet (Yu et al. 2019), PhysFormer (Yu et al. 2022) and RhythmFormer (Zou et al. 2025) were reimplemented from their papers and the authors' code, with their published layers, and trained through the same training loop as Zephyr: the same gray 96 × 96 crops on the 60-Hz grid, optimizer, warm-up and cosine schedule, weight averaging, augmentation (without the frame jitter of Zephyr) and early stopping on the validation animals. Each was trained on the loss of its own paper. TS-CAN, whose input is resized to 36 × 36 pixels, is trained with the mean squared error against the standardized first difference of the target, and PhysNet with the negative Pearson correlation. PhysFormer and RhythmFormer read the crops upsampled to the 128 pixels they were built for and are trained with a negative Pearson term plus a cross-entropy over each window's spectrum (one bin per cycle per minute over 1–15 Hz) against the thermistor's rate in that window: $0.1r + b(\mathrm{CE} + \mathrm{KL})$, with b rising from 1 to 5 over training, for PhysFormer, and $0.2r + \mathrm{CE}$ for RhythmFormer. Their attention spans a whole window, so they were also run in windows of their training length. Window, batch, steps per epoch and learning rate are in Table S2 [confirm how the learning rates were chosen: on the validation animals, or from the papers (PhysFormer 10^{-4}, RhythmFormer 9×10^{-3})].

Table S2: Settings of the neural networks. Training and inference windows are in samples at 60 Hz, and the batch is the number of training windows per optimizer step. Scales are the time scales of the correlation loss, used by Zephyr and PhysNet only. The last two columns give the number of parameters and the inference time for a 5-min clip on a consumer GPU (NVIDIA RTX 4060 Ti; median of 8 clips; Methods).

Network	Window	Batch	Steps	LR	Inference window	Scales	Params (M)	GPU (s)
Zephyr	512	16	100	0.0006	1024	1/4/16	0.86	0.94
TS-CAN	512	4	400	0.003	1024	—	0.53	0.96
PhysNet	128	4	400	0.0003	1024	1	0.77	1.57
PhysFormer	320	4	400	0.0001	320	—	7.38	6.07
RhythmFormer	320	2	800	0.009	320	—	3.24	5.74

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Table S3: Method comparison on held-out sessions of training animals. As in Table 1, but on held-out sessions (86 animal × rig units of 41 animals). Values are the mean over the 5 cross-validation folds with its 95% interval.

			Respiration rate (10-s windows)				Waveform and events		
Method	**Splits**	**Inference (s)**	**MAE ↓**	**RMSE ↓**	**MAPE (%) ↓**	**r ↑**	**Waveform r ↑**	**Inhale F1 ↑**	**Exhale F1 ↑**
Zephyr (ours)	5	**0.94**	3.6 3.1–4.6	6.3 5.5–7.4	1.4 1.2–1.8	**0.99** 0.99–0.99	**0.90** 0.90–0.91	**0.98** 0.97–0.98	**0.94** 0.92–0.95
PhysNet (Yu et al. 2019)	5	1.57	**3.5** 3.3–3.9	**6.2** 5.7–6.7	**1.3** 1.2–1.4	**0.99** 0.99–0.99	**0.90** 0.89–0.90	0.97 0.96–0.98	**0.94** 0.91–0.96
PhysFormer (Yu et al. 2022)	5	6.07	10.7 9.2–12.4	14.6 12.7–16.8	3.8 3.2–4.5	0.97 0.96–0.97	0.80 0.78–0.81	0.81 0.80–0.83	0.79 0.76–0.82
RhythmFormer (Zou et al. 2025)	5	5.74	5.1 4.5–5.7	8.0 7.2–8.8	1.8 1.6–2.1	**0.99** 0.98–0.99	0.87 0.84–0.89	0.95 0.93–0.96	0.91 0.87–0.94
TS-CAN (Liu et al. 2020)	5	0.96	11.3 10.8–11.6	15.2 14.8–15.7	4.3 4.1–4.5	0.97 0.97–0.97	0.73 0.72–0.74	0.93 0.91–0.94	0.86 0.83–0.89
Facemap (Stringer et al. 2019; Syeda et al. 2024)	5	2.01*	138.0 131.9–144.0	152.6 146.5–158.8	54.1 51.6–56.6	0.31 0.26–0.35	0.51 0.50–0.53	0.52 0.51–0.53	0.48 0.46–0.50
Pixel SNR	5	2.94*	93.6 90.6–96.0	107.6 105.1–109.5	36.1 35.1–37.2	0.05 −0.01–0.11	0.00 −0.00–0.01	0.27 0.26–0.29	0.27 0.26–0.28
Motion energy (Musall et al. 2019; Stringer et al. 2019)	5	1.56*	154.8 151.3–158.5	166.6 163.2–170.2	60.4 59.0–62.0	0.21 0.14–0.27	0.03 0.02–0.04	0.32 0.31–0.34	0.32 0.31–0.32

Table S4: Zephyr on each camera. Zephyr trained on the face, side or front camera, and scored on the recording spans all three cameras share, against the same thermistor events. Scores are on held-out animals (1,018 clips) and held-out sessions (986 clips) of 5 of the cross-validation folds, as the mean over animal × rig units with its 95% bootstrap interval. Rate MAE is the error in respiration rate over 10-s windows, in cycles per minute.

Set	**Camera**	**Units**	**Waveform r**	**Inhale F1**	**Exhale F1**	**Rate MAE**
Held-out animals	Face	106	0.81 (0.79–0.83)	0.96 (0.95–0.97)	0.77 (0.72–0.81)	4.5 (4.0–5.2)
Held-out animals	Side	106	0.80 (0.77–0.82)	0.94 (0.92–0.96)	0.75 (0.70–0.79)	6.6 (4.6–10.1)
Held-out animals	Front	106	0.79 (0.76–0.81)	0.95 (0.93–0.96)	0.72 (0.68–0.77)	5.2 (4.3–6.3)
Held-out sessions	Face	86	0.89 (0.88–0.91)	0.97 (0.96–0.98)	0.91 (0.88–0.94)	4.0 (3.2–5.1)
Held-out sessions	Side	86	0.88 (0.86–0.90)	0.96 (0.94–0.98)	0.89 (0.86–0.92)	5.7 (3.7–9.3)
Held-out sessions	Front	86	0.89 (0.86–0.90)	0.97 (0.96–0.98)	0.90 (0.86–0.93)	4.1 (3.4–4.9)

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

Supplementary figures

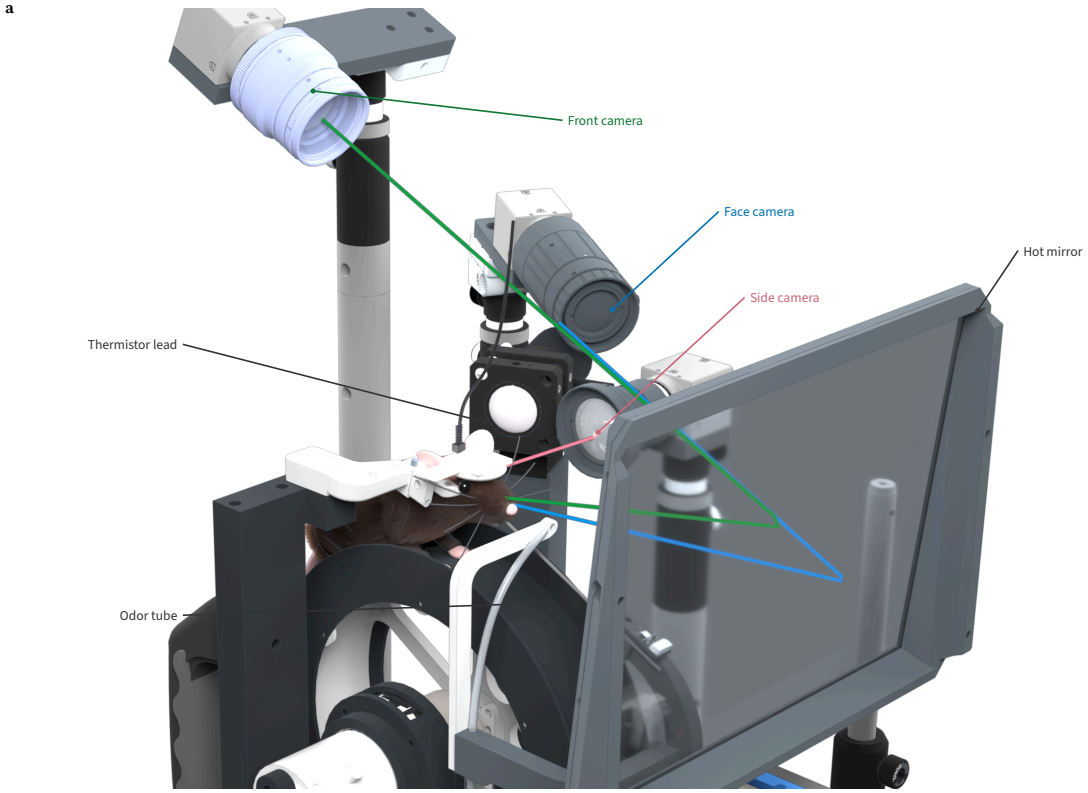


Figure S1: The behavioral rig. A rendering of the rig from the front left, the monitors hidden. The side camera films the animal directly; the face and front cameras film it through the hot mirror in front of the screen, which reflects near-infrared light and transmits the visible light of the screen (Methods). Lines trace each camera's view of the animal. Odors reach the nose through the tube held by the olfactory-line holder, and the thermistor's lead leaves the head-plate cement.

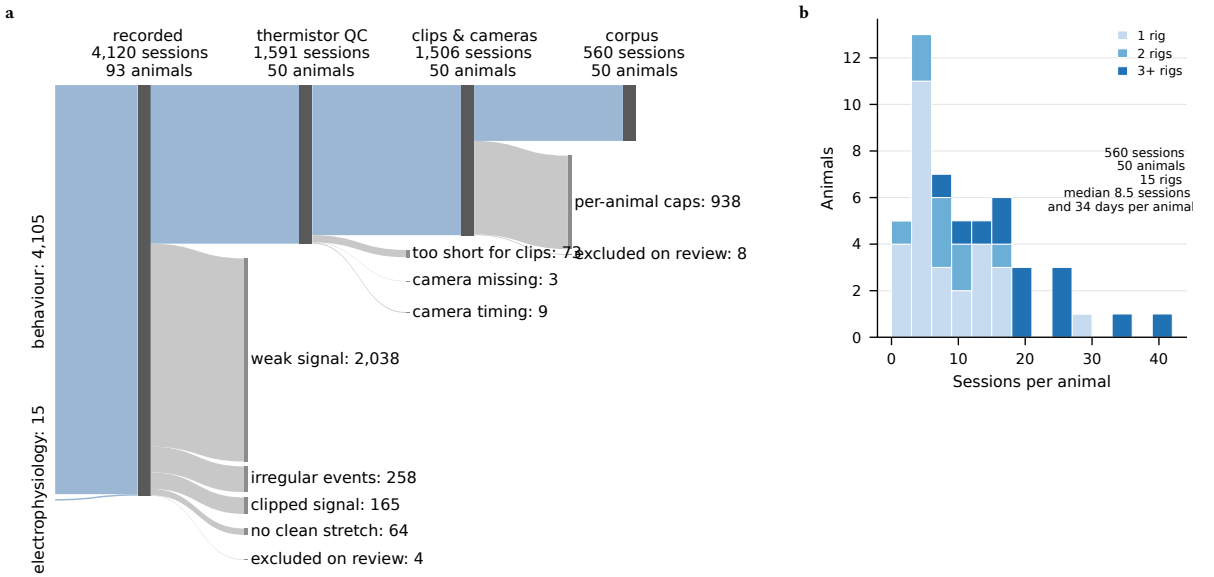


Figure S2: Selection of the dataset. a, Selection of the recorded sessions. Each step removes the sessions that failed a rule, labelled by the first rule they failed (rules and thresholds in Table 2). **b,** Number of sessions per animal, colored by the number of rigs the animal was recorded on (n = 50 animals, 560 sessions).

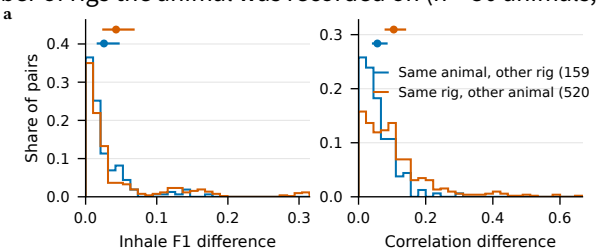


Figure S3: Scores differ less between rigs than between animals. Zephyr on the held-out animals of the five folds. A unit is an animal on one rig, with its score averaged over its clips. Histograms show the absolute difference in inhalation F1 (left) and waveform correlation (right) between pairs of units that share the animal but not the rig (159 pairs of 22 animals), and between pairs that share the rig but not the animal (520 pairs of 48 animals). Points above give each kind's mean with its 95% bootstrap interval over animals.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

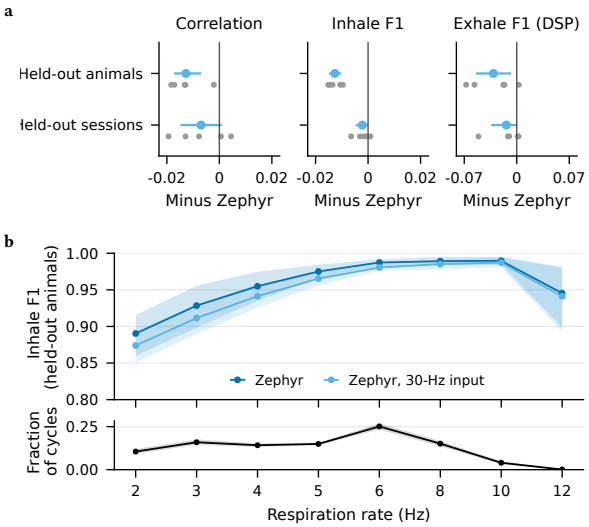


Figure S4: Zephyr works on 30-Hz video, with slightly lower accuracy. Zephyr trained on every other frame (30 Hz) of each of the five folds, still predicting on the 60-Hz grid. **a,** The 30-Hz network minus the 60-Hz network on each test set, in waveform correlation, inhalation F1 and exhalation F1. Grey points are single folds; the mean is shown with its 95% bootstrap interval over folds. **b,** Inhalation F1 on held-out animals by respiration rate, mean over folds with its 95% interval. The bottom panel shows the share of cycles in each rate bin, mean over animal × rig units with its 95% bootstrap interval.

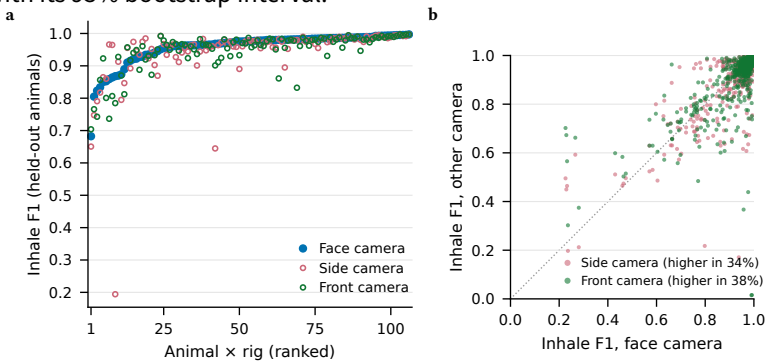


Figure S5: Zephyr recovers respiration from the side and front cameras. Zephyr trained on each camera's video in 5 of the cross-validation folds, and scored on the recording spans all three cameras share. The held-out animals have 1,018 clips from 106 units of 48 animals, and the held-out sessions 986 clips from 86 units. Every score is in Table S4. **a,** Inhalation F1 of each held-out-animal unit, ranked by the face camera, as in Figure 4b. **b,** Inhalation F1 of each held-out-animal clip from the side and front cameras, plotted against the face camera. The dotted line marks equality, and the legend gives the share of clips that scored higher than with the face camera.

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9



Figure S6: Representative frames from the dataset. First frames of the face, side and front cameras in 16 representative sessions, each labelled by its session name, with the network's crop box outlined (Methods).

Built from: 33ccd35bf3b5f3196f09cd24b7aabc6be21ad3c9

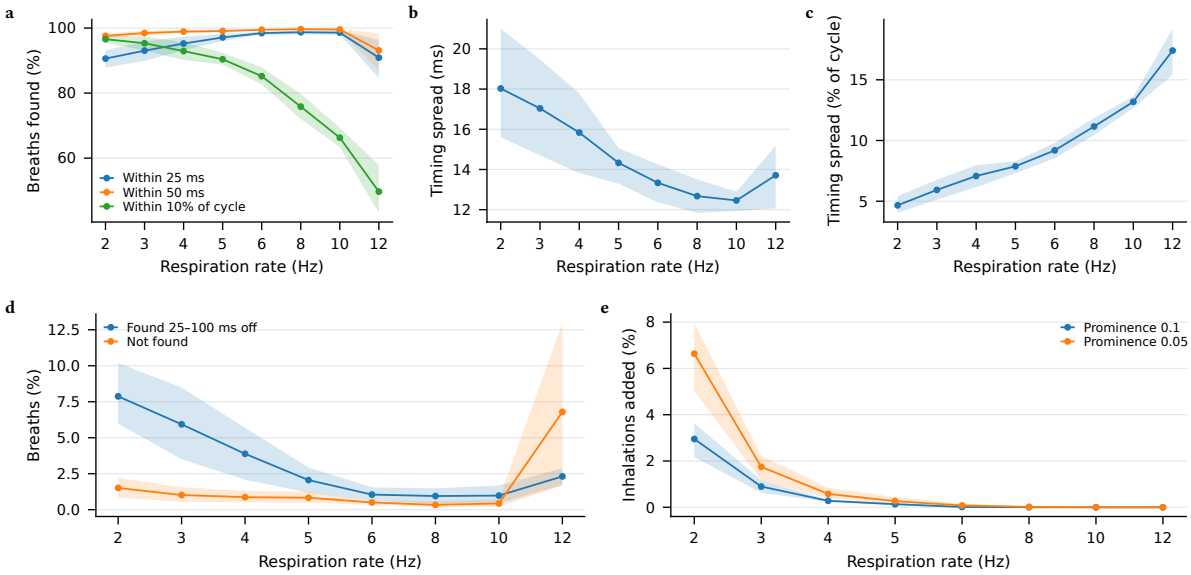


Figure S7: A fixed tolerance is strictest for slow breathing. Zephyr on the held-out animals of the 5 cross-validation folds, by the thermistor's respiration rate (bins labelled by their lower edge; 12 is 12–15 Hz). Lines show the mean over folds and bands its 95% interval. **a,** Share of thermistor inhalations found within 25 ms (the tolerance used throughout), within 50 ms, and within 10% of the cycle. **b,** Spread of Zephyr's inhalation times around the thermistor's (interquartile range of the lag). **c,** The same spread as a percentage of the cycle. **d,** Share of thermistor inhalations found 25–100 ms off, which count as both a miss and a false inhalation, and of those with no inhalation within 100 ms. **e,** Inhalations added by the thermistor's detector, as a percentage of the labelled cycles in each bin, when its prominence fraction is lowered from 0.2 to 0.1 or 0.05.

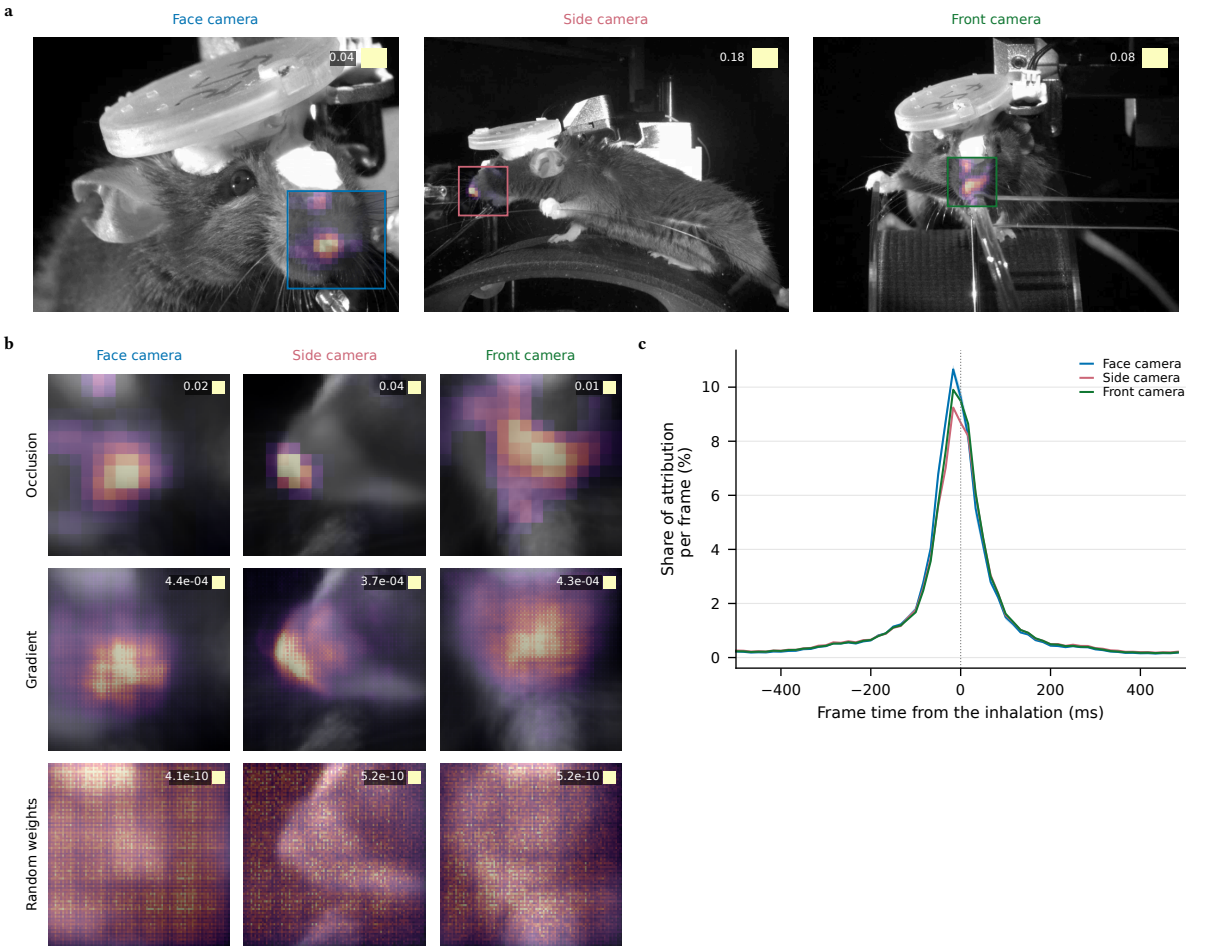


Figure S8: Spatial and temporal receptive fields. The face network of each cross-validation fold on its held-out animals (107 animal × rig units), and the side and front networks of the first fold on the same recording spans (28 units; Methods). **a,** One held-out animal seen by the three cameras, with the occlusion map inside each crop box: how much hiding each part of the crop lowered the inhalation onset probability at the thermistor's inhalations. **b,** Maps averaged over each camera's units: occlusion, the gradient of the inhalation onset output with respect to the input pixels, and the same gradient for the network with random weights. Each map is scaled to its own range. **c,** Share of an inhalation's gradient carried by each frame, plotted against the frame's time from the inhalation. Lines show the mean over folds; for the face network, the band shows the 95% interval over the 5 folds.

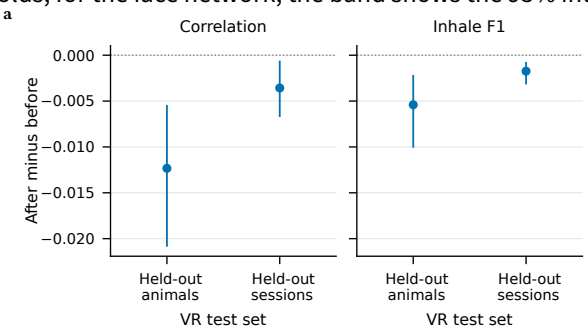


Figure S9: Fine-tuning on a different task costs little on the original dataset. The network of Figure 4f after fine-tuning on the dynamic-foraging experiment (Figure 4g–i), minus the same network before fine-tuning, on the held-out animals and held-out sessions of the first fold, over all three cameras. Points show the mean over animal × rig units with its 95% paired bootstrap interval.

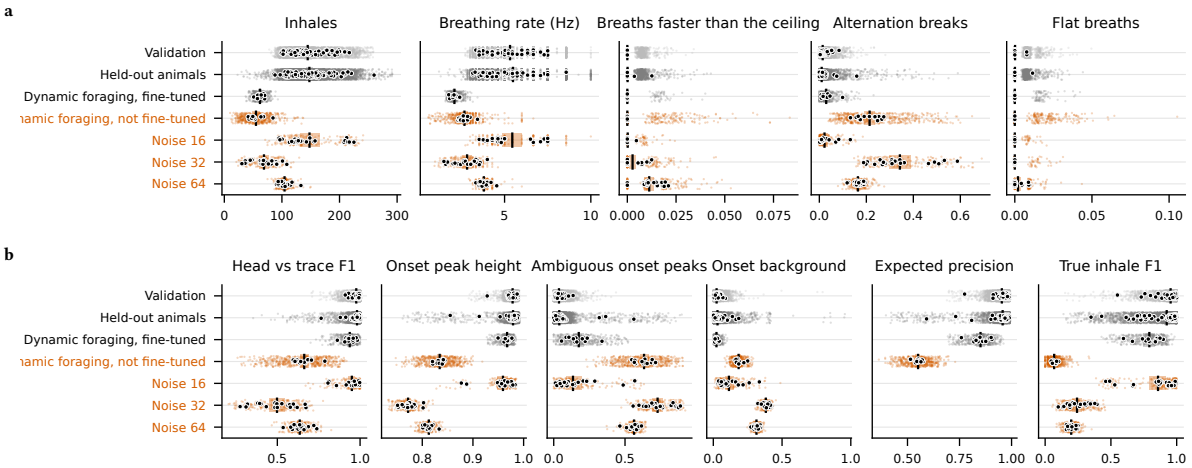


Figure S10: Every quality-control measure. Each measure of Table 7 over the 30-s windows of the sets of Figure 5e, drawn as there. **a,** Measures of the respiratory rhythm. **b,** Measures of the inhalation onset output, beside the true inhalation F1. Expected precision is not computed for the noise sets.

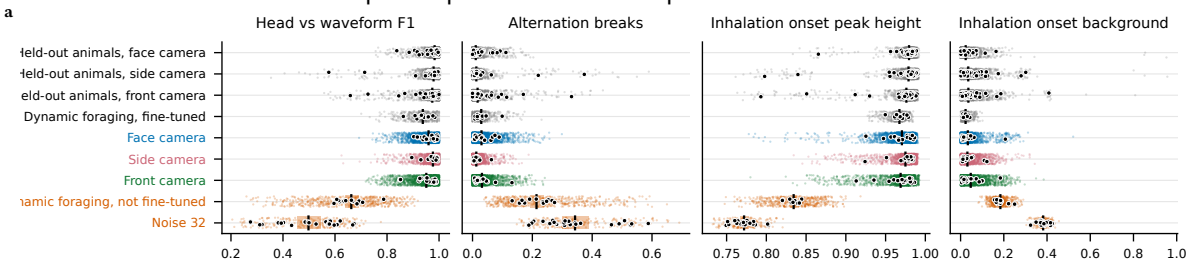


Figure S11: Quality control of the sessions without a thermistor on each camera. The 19 sessions of Figure 6, predicted from the face, side and front videos by the network trained on all three cameras. Each camera's quality-control scores over the 30-s windows of every session are shown beside predictions that work and known failures, drawn as in Figure 5e. The held-out animals' windows are split by the camera of their clips.