Title

Not all cross-modal responses are explained by face movements

Author names and affiliations

Timothy Olsen[1], Andrea Hasenstaub[1]

[1]Coleman Memorial Laboratory, Department of Otolaryngology–Head and Neck Surgery, University of California, San Francisco, 94143

Abstract

Recent work has claimed that most apparently cross-modal responses in sensory cortex are instead caused by the face movements evoked by stimuli of the non-dominant modality. We show that visual stimuli rarely trigger face movements in awake mice; when they occur, such movements do not explain visual responses in auditory cortex; and in simultaneous recordings, face movements drove artifactual cross-modal responses in visual but not auditory cortex. Thus face movements do not broadly explain cross-modal activity across all stimulus modalities.

Main

Recent work has claimed that in sensory cortex, most apparently cross-modal responses are in fact artifacts of movements, mostly of the face and whiskers, evoked by stimuli of the non-dominant modality. Consistent with this, considerable evidence has been presented that in awake mice, the intermittent sound stimuli used to probe neural responses do typically evoke face movements, that responses to these movements can be misinterpreted as being driven by the sound stimuli themselves, and that in the visual cortex, many apparently sound-evoked responses are of this nature (Clayton et al., 2024; Bimbard et al., 2023; Williams et al., 2023). We have previously reported visual responses in the auditory cortex (AC) (e.g. Morrill and Hasenstaub, 2018). Here, we tested whether these cross-modal responses in the AC are also artifacts of movement.

We first tested whether similarly infrequent visual stimuli also tend to evoke face movements. Awake mice (Fig. 1a) were, on randomly interleaved trials, presented with contrast-modulated noise movies (Fig. 1b, red, 100 trials) or repeating noise burst stimuli (Fig. 1b, blue, 30 trials), with each trial separated by 4-12 seconds. Face movements were quantified using Facemap software (Fig. 1c), which applies singular value decomposition to capture face movements over time using multiple principal components (PCs). We initially focused on PC1, with further analysis of PCs 2-3 provided in the supplementary material to confirm our findings.

Face movements were much less likely to occur during visual stimulus trials than during sound stimulus trials (Fig. 1d-f). To quantify this, we calculated the percentage of auditory and visual trials with stimulus-induced face movements. A trial was considered to have

stimulus-induced face movement if the mean face movement within 500 ms of stimulus onset exceeded 2 standard deviations from baseline (threshold shown by the green dashed line in figure 1d; see supplementary figure 1 for threshold rationale), where baseline is defined as the 500 ms prior to stimulus onset (gray shading in figure 1d). The percentage of trials with face movement was substantially higher for auditory stimuli than for visual stimuli (Fig. 1g, visual: 16.5 ± 9.6%, auditory: 58.0 ± 13.9%, two-sided Wilcoxon signed-rank test: $*P = 2.9 \times 10^{-4}$). We also examined whether the face movements elicited by visual and auditory stimuli were of similar magnitudes, and found that they were (Fig. 1h, visual: 31.0 ± 14.4, auditory: 32.4 ± 14.8, two-sided Wilcoxon signed-rank test: $P = 0.94$).

To ensure that our results were not dependent on our specific experiment and analysis choices, we ran further analyses using additional face movement principal components (supplementary figure 2), and a different auditory/visual stimulus paradigm (supplementary figure 3). In these complementary analyses, visual stimuli remained much less likely to evoke face movements than sound stimuli.

Although visually evoked face movements were rare, they were not nonexistent, leading us to question how much these movements might explain visually-evoked activity in the auditory cortex. While recording face movements from head-fixed mice in response to visual stimuli (Fig. 1a), we also recorded electrophysiological data from the primary auditory cortex (A1) (Fig. 2a-c). Consistent with our previous studies (Morrill & Hasenstaub, 2018; Bigelow et al, 2022), about 8% of recorded cells in the auditory cortex responded with significantly elevated firing in response to visual stimuli (Fig. 2d, red; n=39 of 499 cells, $P<0.01$, FDR corrected, two-sided Wilcoxon signed rank tests), most of which were located in the deep layers. When comparing visually evoked firing and face movements on the same trials (Fig. 2e), in most AC cells, face movements appeared unrelated to visual responses (i). Yet for a smaller subset of cells, visually evoked firing did appear to accompany face movements (ii). To quantify the prevalence of such face motion artifacts and the effect of face movement on visually evoked firing, we compared the change in firing rate following the visual stimulus between trials with and without face movement, for all visually responsive cells (Fig. 2f). As in Figure 1, we used a 2x SD threshold to detect trials with face movement. In only 20% of cells were visual responses significantly different between trials with and without face movement (Fig. 2f, left, orange, 20%, $P<0.01$, FDR corrected, two-sided Mann-Whitney U tests). In the vast majority (80%) of cells, the presence or absence of face movement made no difference to the visual responses (Fig. 2f, left, gray). We further split these data by recording and mouse to determine if the effect of face movements on visually evoked firing depended on these factors (Fig. 2f, right), but we found no such dependency; the visually responsive cells affected by face movements were randomly distributed across recordings ($P = 0.21$, Kruskal-Wallis test) and mice ($P = 0.16$, Kruskal-Wallis test).

We confirmed that these results remained consistent when incorporating additional face movement principal components (Supplementary Fig. 4), and they did not differ when only including recording sites with onset latencies consistent with primary auditory fields (Supplementary Fig. 5). We also performed an additional analysis using a generalized linear model to compare the effects of visual stimuli and face movements on visual responses in the auditory cortex (Supplementary Fig. 6). This analysis showed that visual responses in the auditory cortex were more strongly related to visual stimuli than to face movements, further supporting our findings in figure 2f.

In summary, most visual responses in the auditory cortex are not explained by face movements.

Our findings indicate that stimulus-driven face movements do not explain cross-modal activity in the auditory cortex, while previous research (e.g. Bimbard et al., 2023) indicated that face movements are major drivers of apparently cross-modal activity in the visual cortex. To verify that this difference is not due to differences between labs in experimental design or stimuli, we recorded simultaneously from the primary auditory and visual cortices and compared the effect of face movement on visually-evoked activity in the auditory cortex (Fig. 3a) with that on sound-evoked activity in the visual cortex (Fig. 3b), in interleaved trials during the same experiment. Plotting the difference in firing rate between trials with and without face movement revealed that visually-evoked firing in auditory cortex mostly remained stable across trials with and without face movements (Fig. 3c, left), whereas sound-evoked firing in visual cortex consistently increased on face movement trials (Fig. 3c, right).

To quantify this difference, we plotted the stimulus-evoked firing rate change for trials with (y-axis) and without (x-axis) face movement (Fig. 3d, left), for auditory cortex cells with significant visually evoked firing (red) and visual cortex cells with significant sound-evoked firing (blue). This showed that while visually evoked firing in the auditory cortex typically remained stable across trials with and without face movements (red), sound-evoked firing in the visual cortex almost always increased during face movement trials. When comparing the difference in firing rate change between trials with and without face movement ('with' minus 'without'), this difference was near zero for visually responsive cells in the auditory cortex (Fig. 3d, right, red) but nearly always above zero for sound-responsive cells in the visual cortex (Fig. 3d, right, blue), and this difference was statistically significant (AC: 0.4 ± 0.7, VC: 1.2 ± 0.9, two-sided Mann-Whitney *U* test: $^*P = 1.6 \times 10^{-4}$).

Taken together, these results show that while many apparently cross-modal responses in the visual system are explained by face movements, visually driven responses in the auditory system are generally not related to face movements.

While sensory cortices are predominantly driven by stimuli in their preferred modality, their firing can in some cases be driven by stimuli of other modalities: visual cortical firing may be initiated by sound (Ibrahim et al., 2016; Lurilli et al., 2012), auditory cortical activity may be driven by vision (Morrill & Hasenstaub, 2018), or somatosensation (Lohse et al, 2021), and so on. Recent work has demonstrated that in visual cortex, these apparently cross-modal responses are often caused by stimulus-triggered movements, and proposed more generally that apparently cross-modal activity is an artifact of changes in movement or internal state rather than a genuine response to the stimulus itself (Bimbard et al., 2023). The present results show that this is not always the case: rather, the relationship between the neural firing apparently driven by cross-modal stimulation and movement differs significantly by brain region and modality. In visual cortex, cross-modal responses are strongly influenced by face movements, while cross-modal responses in the auditory cortex are not. Mechanistically, these differences may arise from variation in the effects of movement and arousal across primary sensory cortices (Shimaoka et al., 2018), which may be associated with variations in the projection patterns and post-synaptic effects of noradrenaline, acetylcholine, and other relevant neuromodulators (Kim et al., 2016). Teleologically, these

differences could reflect the sensory ecology of mice; given their limited reliance on visual cues (Huberman & Neill, 2011), conventional visual stimuli may be less salient and thus less likely to trigger changes in arousal or body positioning than conventional sound stimuli. More salient visual stimuli, such as looming stimuli (Yilmaz & Meister, 2013), or quieter sounds that rarely evoke face movements (Bimbard et al., 2023, could yield different outcomes.

Nevertheless, these findings demonstrate that cross-modal responses can be genuine rather than artifactual, and that motor-arousal-sensory interactions in one sensory modality do not generalize across all sensory systems.

Acknowledgements

We would like to thank Professor Michael Stryker and Professor Christoph Schreiner for helpful comments on this manuscript. This work was supported by NIH grant R01NS116598.

Methods

Subjects. All procedures were approved by the Institutional Animal Care and Use Committee at the University of California, San Francisco. All mice were housed in cages of 2-5 under a 12 hr light/dark cycle, where food and water were provided *ad libitum*.

Experiments in all figures apart from supplementary figure 3 were conducted on 6 mice (3 males, 3 females; mean age: 88 days, age range: 83-96 days). Of these, 4 were PV-Cre:Ai32 mice, which we obtained by crossing female PV-Cre mice (Jax strain: 012358, C57 background) with male Ai32 mice (Jax strain: 012569), and 2 were wild-type C57 mice (Jax strain: 000664; all on C57 background). For supplementary figure 3, an additional set of 6 mice were used to validate our behavioral results with a different stimulus paradigm (3 males, 3 females; mean age: 98 days, age range: 74-106 days). This set included 4 PV-Cre:Ai32 mice and 2 SST-Cre mice (Jax strain: 013044, both on C57 background). Mice were optogenetically stimulated on randomly interleaved trials during these experiments, but the optogenetic stimulus trials are not analyzed here.

Surgical procedures. For all mice, an initial surgery was performed under isoflurane anesthesia (4% for induction, 1-2% for maintenance) to fix a custom metal headbar to the skull with dental cement. Following this surgery, mice were allowed to recover for at least 2 days. Mice were then habituated to the recording setup for 20-90 mins at a time for 2-5 days. Mice then underwent a second surgery under isoflurane anesthesia where two ~1-2 mm craniotomies were made. One was made to expose the primary auditory cortex (A1), and was made under the squamosal ridge at a location ~2.2-3.2 mm posterior to bregma. The other was made to expose the primary visual cortex (V1), and was centered at 3.0-3.5 mm posterior and 2.3-2.5 mm lateral to bregma (localization of the craniotomy above primary visual cortex was later verified stereotactically and based on responses to drifting gratings). Following this surgery, mice were allowed to recover overnight before electrophysiological recordings began. When not performing recordings, the craniotomies were covered with

agarose (2%) and subsequently sealed with silicone elastomer (Kwik-cast, World Precision Instruments). Recordings were performed daily for up to 3 days following the craniotomies.

Recording setup. All recordings were performed inside a sound attenuation chamber (Industrial acoustics), from awake mice placed on a spherical treadmill (Phillips & Hasenstaub 2017). Visual stimuli were presented to the mice using a 19 inch monitor (ASUS VW199T-P or Dell P2016t, 60 frames per second, 1440 by 900 pixels) placed ~25 cm directly in front of them, while auditory stimuli were delivered from a free-field electrostatic speaker (ES1, TDT) at 192 kHz. The speaker was placed immediately to the left of the monitor (relative to the mice), also ~25 cm from the mice. To record face movements, we used a Mako U-130B camera (Allied Vision) with a 13-130 mm varifocal lens (B&H Photo Video), infrared light for illumination (Mightex SLS-0208-A), and custom software written in Matlab. The camera and infrared light were placed to the immediate right of the monitor (relative to the mice), and face videos were recorded at 30 frames per second.

Visual stimuli. Visual stimuli consisted of contrast modulated noise (CMN) movies, which are broadband stimuli designed to maximally activate the visual cortex (Neill & Stryker 2008). The CMN was generated at 60x60 pixels and then interpolated to 900x900 pixels, covering an area of approximately 54 x 54° in the field of view. CMN stimuli were ramped from zero contrast (gray screen) to full contrast over 250 ms. The monitor during the inter-trial interval was black, thus giving an abrupt stimulus onset (black to gray). The monitor brightness was calibrated using a SpectraScan PR-670 device to 31 - 44 cd/m^2 at gray screen. For experiments in all figures apart from supplementary figure 3, visual stimuli were presented 100 times; for supplementary figure 3, visual stimuli were presented 50 times.

Sound stimuli. For most experiments, sound stimuli consisted of a repeated noise burst stimulus, which consisted of four 150 ms noise bursts, each separated by 150 ms of silence. The noise bursts were white noise band-pass filtered between 4 and 64 kHz, and were ramped on and off using 5 ms linear ramps. The same repeating noise burst stimulus was used for all trials, and was presented 30 times. In some experiments (supplementary figure 3), we used a random double sweep stimulus, which is a continuous, spectrally sparse sound that effectively drives neurons with diverse tuning properties in the primary auditory cortex (Gourévitch et al., 2015). For these experiments, 50 different random double sweep stimuli were presented, with one for each trial. Both sound stimuli were presented at ~65 dB, calibrated using a Brüel & Kjær 2209 and model 4939 microphone.

Electrophysiology. We recorded neural activity from the brain of awake mice using Neuropixels 1.0 probes. Probes were implanted into the primary auditory (A1) and visual (V1) cortices. The probes were continuously lowered into the brain at a speed of ~1 µm/sec until silent channels were observed both above (pia) and below (white matter) activity in the cortex – typically to a depth of ~1400 µm. Once this depth was reached, the recording sites were covered with agarose (2%) and the probes were left to settle for ≥30 minutes before recordings started. Data was recorded at a sampling rate of 30 kHz using a Neuropixels PXIe acquisition module (PXIE_1000), National instruments chassis (NI PXIe-1071), and SpikeGLX software (v.20230411). Single unit activity was obtained using Kilosort 2.0 (Pachitariu et al., 2016) and results were curated using Phy (https://github.com/cortex-lab/phy). Splitting single units into broad and narrow spiking types, and determining their cortical depths, was performed using methods identical to those in our previous publication

(Olsen & Hasenstaub 2022). All spiking post-stimulus time histograms (PSTHs) were generated using 50 ms bins.

Identifying recording sites with firing properties consistent with primary auditory fields. We only analyzed recordings showing clear signs of frequency tuning. Multi-unit (MU) firing was measured in response to 100 ms tones, using methods from Olsen & Hasenstaub (2022), across 14 logarithmically spaced frequencies (4–64 kHz) and 7 attenuation levels (35–65 dB in 5 dB steps). For each recording channel, we calculated the frequency tuning curve (FTC), collapsed across attenuation, and measured its variance. This variance was compared to a null distribution generated by shuffling frequency labels (1000 iterations) to determine if it exceeded chance. A recording was classified as frequency-tuned if at least one-third of tone-responsive channels showed significant FTC variance ($P < 0.01$). Tone-responsive channels were identified by significantly elevated tone-evoked firing (collapsed across frequency and attenuation) compared to baseline, using a two-sided Wilcoxon signed-rank test ($P < 0.01$) on single-trial spike counts 100 ms before and during the tones.

We also assessed whether recordings had onset latencies consistent with primary auditory fields, which includes our target area (A1). These latencies are typically shorter than those in secondary auditory fields (5-18 ms vs. 8-32 ms, Joachimsthaler et al., 2014). For each channel, we measured the onset latency of MU responses to tones (collapsed across frequency and attenuation) by identifying when the response first exceeded 2.5 SDs of baseline firing (baseline: 100 ms prior to tone onset). As in our previous reports (Olsen & Hasenstaub 2022; Bigelow et al., 2022), a recording was considered to be within a primary auditory field if the median onset latency across channels with significantly elevated tone-evoked firing was ≤14 ms. Since 82% of our recordings met this criterion, our results predominantly reflect firing from units in primary auditory fields. However, we also verified that our results remained unchanged when restricting analyses to these recordings (Supplementary figure 5).

Face movement analysis. We used Facemap software to process face videos (Syeda et al., 2024), which applies singular value decomposition to quantify facial movement over time using multiple principal components (PCs). We used PC1 for most of our analyses, and then confirmed our results using additional principal components (PCs 2 & 3, supplementary figures 2 and 4). For all analyses and plots, face movements were first aligned to visual and/or auditory stimuli for each trial. We then separately z-scored each trial by subtracting the mean baseline movement and dividing by the baseline standard deviation, using the baseline periods shown in each figure by gray shading. Trials with face movements exceeding 2 SDs in response to the stimuli were classified as face movement trials (see supplementary figure 1 for our rationale for choosing 2x SD). All face movement plots have a time resolution of 33.3 ms, determined by the video frame capture rate of 30 frames per second.

Generalized linear model. To confirm our main results, we used a generalized linear model (GLM) to assess how face movements contribute to visually evoked firing in the auditory cortex. The GLM attempted to predict spike counts within 500 ms windows, either during visual stimuli or during the inter-trial / baseline periods, using two predictor variables: a binary variable (“Vstim”) indicating whether spike counts came from visual (1) or baseline

(0) windows, and the mean face movement (PC1) during each window. The model equation is defined as follows, where 'win' represents a 500 ms window:

$$Num.spikes_{win} = \beta_0 + \beta_1 Vstim_{win} + \beta_2 FM_{win} + \beta_3 (Vstim_{win} \cdot FM_{win}) + \epsilon_{win}$$

We ran a GLM for each recorded cell using Matlab's (v.2024a) function fitglm, specifying 'Poisson' for the 'distribution' argument and 'Log' for the 'Link' argument.

Statistics. All statistical analyses were performed in Matlab (v.2024a). We used a two-sided Wilcoxon signed-rank test for comparing two groups of paired data, and a two-sided Mann-Whitney *U* test for comparing two groups of unpaired data. For analyses involving multiple comparisons using these tests, such as across multiple recordings (e.g., Figure 1g) or multiple cells (e.g., Figure 2d), we adjusted p-values using the Benjamini-Hochberg procedure to control for the false discovery rate (FDR). When comparing across more than two groups of unpaired data, we used a Kruskal-Wallis test.

References

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

Clayton, K.K., Stecyk, K.S., Guo, A.A., Chambers, A.R., Chen. K., Hancock, K.E. & Polley, D.B. Sound elicits stereotyped facial movements that provide a sensitive index of hearing abilities in mice. *Curr. Biol*. **34**, 1605-1620 (2024).

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C.B., Carandini, M. & Harris, K.D. Spontaneous behaviors drive multidimensional, brain-wide activity. *Science*. **364** (2019).

Bimbard, C., Sit, T.P.H., Lebedeva, A., Reddy, C.B., Harris, K.D. & Carandini, M. Behavioral origin of sound-evoked activity in mouse visual cortex. *Nat. Neuroscience*. **26**, 251-258 (2023).

Bizley, J.K., Nodal, F.R., Bajo, V,M., Nelken, I. & King, A.J. Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cereb. Cortex*. **17**, 2172-2189 (2007).

Cappe, C. & Barone, P. Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *Eur. J. Neuroscience*. **22**, 2886-2902 (2005).

Zhao, M., Ren, M., Jiang, T., Jia, X., Wang, X., Li, A., Li, X., Luo, Q. & Gong, H. Whole-brain direct inputs to and axonal projections from excitatory and inhibitory neurons in the mouse primary auditory area. *Neurosci. Bull*. **38**, 576-590 (2022).

Banks, M.I., Uhlrich, D.J., Smith, P.H., Krause, B.M. & Manning, K.A. Descending projections from extrastriate visual cortex modulate responses of cells in primary auditory cortex. *Cereb. Cortex*. **21**, 2620-2638 (2011).

Morrill, R.J. & Hasenstaub, A.R. Visual information present in infragranular layers of mouse auditory cortex. *J. Neurosci*. **38**, 2854-2862 (2018).

King, A.J. & Walker, K.M. Integrating information from different senses in the auditory cortex. *Biol. Cybernet*. **106**, 617-625 (2012).

Bigelow, J., Morrill, R.J., Olsen, T. & Hasenstaub, A.R. Visual modulation of firing and spectrotemporal receptive fields in mouse auditory cortex. *Current Research in Neurobiology*. **3**, 100040 (2022).

Huberman, A.D. & Niell, C.M. What can mice tell us about how vision works? *Trends. Neurosci*. **34**, 464-473 (2011).

Williams, A.M., Angeloni, C.F. & Geffen, M.N. Sound improves neuronal encoding of visual stimuli in mouse primary visual cortex. *J. Neurosci*. **43**, 2885-2906 (2023).

Lohuis, M.N.O., Marchesi, P., Olcese, U. & Pennartz, C.M.A. Triple dissociations of visual, auditory and motor processing in mouse primary visual cortex. *Nat. Neuroscience*. **27**, 758-771 (2024).

Yilmaz, M. & Meister, M. Rapid innate defensive responses of mice to looming stimuli. *Curr. Biol*. **23**, 2011-2015 (2013).

Phillips, E.A.K., Schreiner, C.E. & Hasenstaub, A.R. Cortical interneurons differentially regulate the effects of acoustic context. *Cell Rep*. **20**, 771-778 (2017).

Pachitariu, M., Steinmetz., Kadir, S., Carandini, M. & Harris, K, D. Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. Preprint at *bioRxiv*. https://doi.org/10.1101/061481 (2016).

Olsen, T. & Hasenstaub, A.R. Offset responses in the auditory cortex show unique history dependence. *J. Neurosci*. **42**, 7370-7385 (2022).

Joachimsthaler, B., Uhlmann, M., Miller, F., Ehret, G. & Kurt, S. Quantitative analysis of neuronal response properties in primary and higher-order auditory cortical fields of awake house mice (*Mus musculus*). *Eur. J. Neuroscience*. **39**, 904-918 (2014).

Neill, C.M. & Stryker, M.P. Highly selective receptive fields in mouse visual cortex. *J. Neurosci*. **28**, 7520-7536 (2008).

Gourevitch, B., Occelli, F., Gaucher, Q., Aushana, Y. & Edeline, J-M. A new and fast characterization of multiple encoding properties of auditory neurons. *Brain Topography*. **28**, 379-400 (2015).

Syeda, A., Zhong, L., Tung, R., Long, W., Pachitariu, M. & Stringer, C. Facemap: a framework for modeling neural activity based on orofacial tracking. *Nat. Neuroscience.* **27**, 187-195 (2024).

Shimaoka, D., Harris, K.D. & Carandini, M. Effects of arousal on mouse sensory cortex depend on modality. *Cell Reports*. 22, 3160-3167 (2018).

Kim, J-H., Jung A-H., Jeong, D., Choi, I., Kim, K., Shin, K., Kim, A.J. & Lee. S-H. Selectivity on neuromodulatory projections from the Basal Forebrain and Locus Ceruleus to primary sensory cortices. *J. Neurosci*. **36**, 5314-5327 (2016).

Ibrahim, L.A., Mesik, L., Ji, X-Y., Fang, Q., Li, H-F., Li, Y-T., Zingg, B., Zhang, L.I. & Tao, H.W. Cross-modality sharpening of visual cortical processing through layer-1-mediated inhibition and disinhibition. *Neuron*. **89**, 1031-1045 (2016).

Lurilli, G., Ghezzi, D., Olcese, U., Lassi, G., Nazzaro, C., Tonini, R., Tucci, V., Benfenati, F. & Medini, P. Sound-drive synaptic inhibition in primary visual cortex. *Neuron*. **73**, 814-828 (2012).

Lohse, M., Dahmen, J.C., Bajo, V.M. & King, A. Subcortical circuits mediate communication between primary sensory cortical areas in mice. *Nat Commun*. **12**, 3916 (2021).

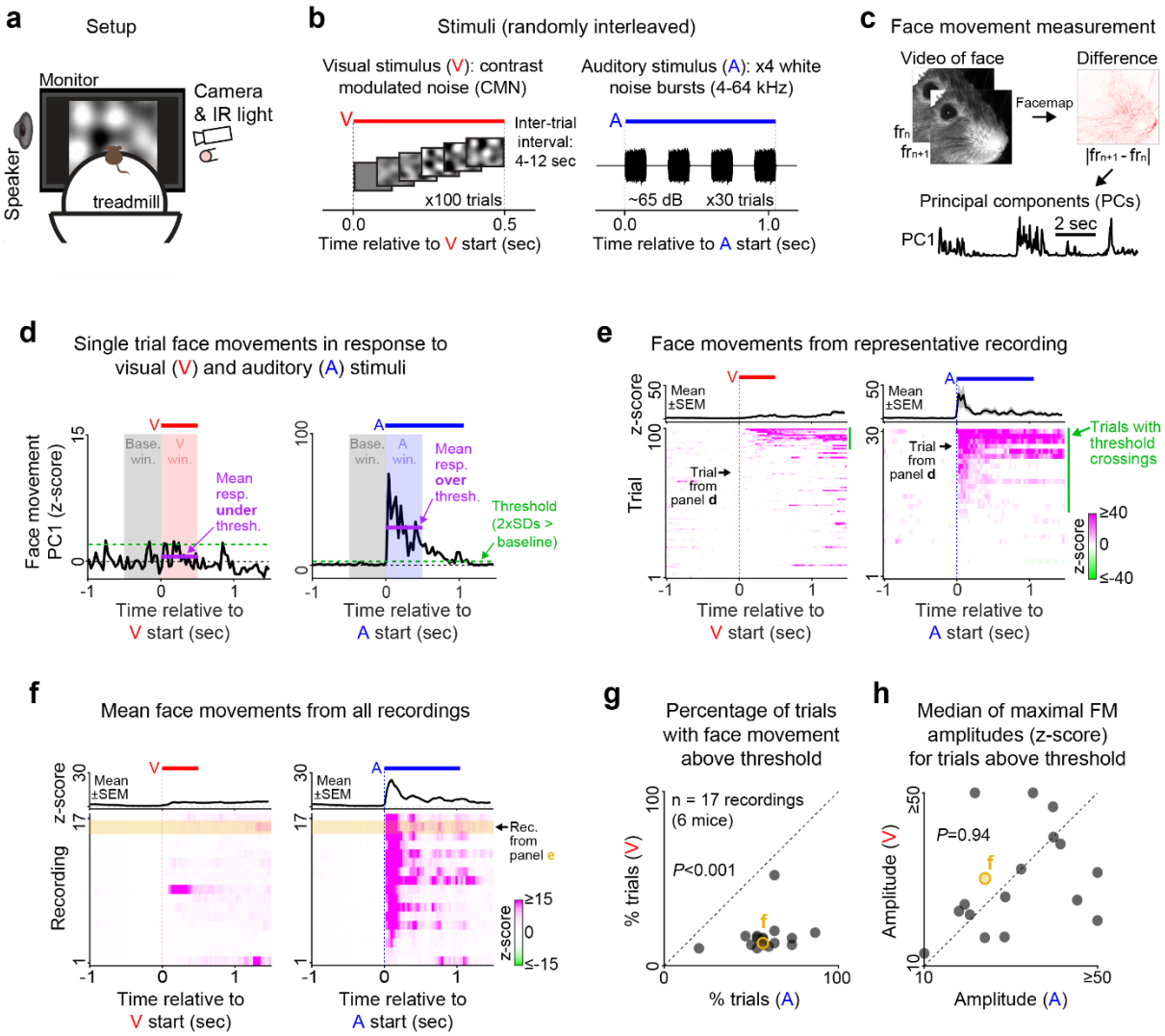


Fig. 1: Visual stimuli are much less likely to evoke face movements than auditory stimuli. **a**, Experimental setup. Mice were head-fixed and placed on a treadmill. Visual and auditory stimuli were presented while their face movements were recorded using an infrared camera and light. **b**, Visual stimuli (V) were 0.5-second contrast modulated noise movies (left) and auditory stimuli (A) were 1.05-second repeating noise burst stimuli. **c**, Face movements were measured using Facemap software, which calculated the absolute difference in pixel intensity between consecutive frames and extracted the first 500 principal components over time from this difference movie. **d**, Face movements (PC1) in response to both stimulus types from representative trials. The red (left, V. win) and blue (right, A. win) shading shows the visual and auditory response windows, where visually- and sound-evoked face movements were measured. **e**, Trial-aligned face movements (PC1) from representative recordings (bottom) and the mean of these face movements across trials (top). **f**, Mean face movements for each recording (bottom), and the grand means across recordings (top). **g**, Percentage of auditory versus visual trials with face movements (PC1) above threshold (n = 17 recordings, 6 mice). Auditory stimuli were significantly more likely than visual stimuli to evoke above-threshold face movements (*P = 2.9 x10^{-4}, two-sided Wilcoxon signed-rank test). **h**, Median face movement amplitudes for above-threshold trials, compared between visual and auditory stimuli. There was no significant difference between stimulus types (P = 0.94, two-sided Wilcoxon signed-rank test).

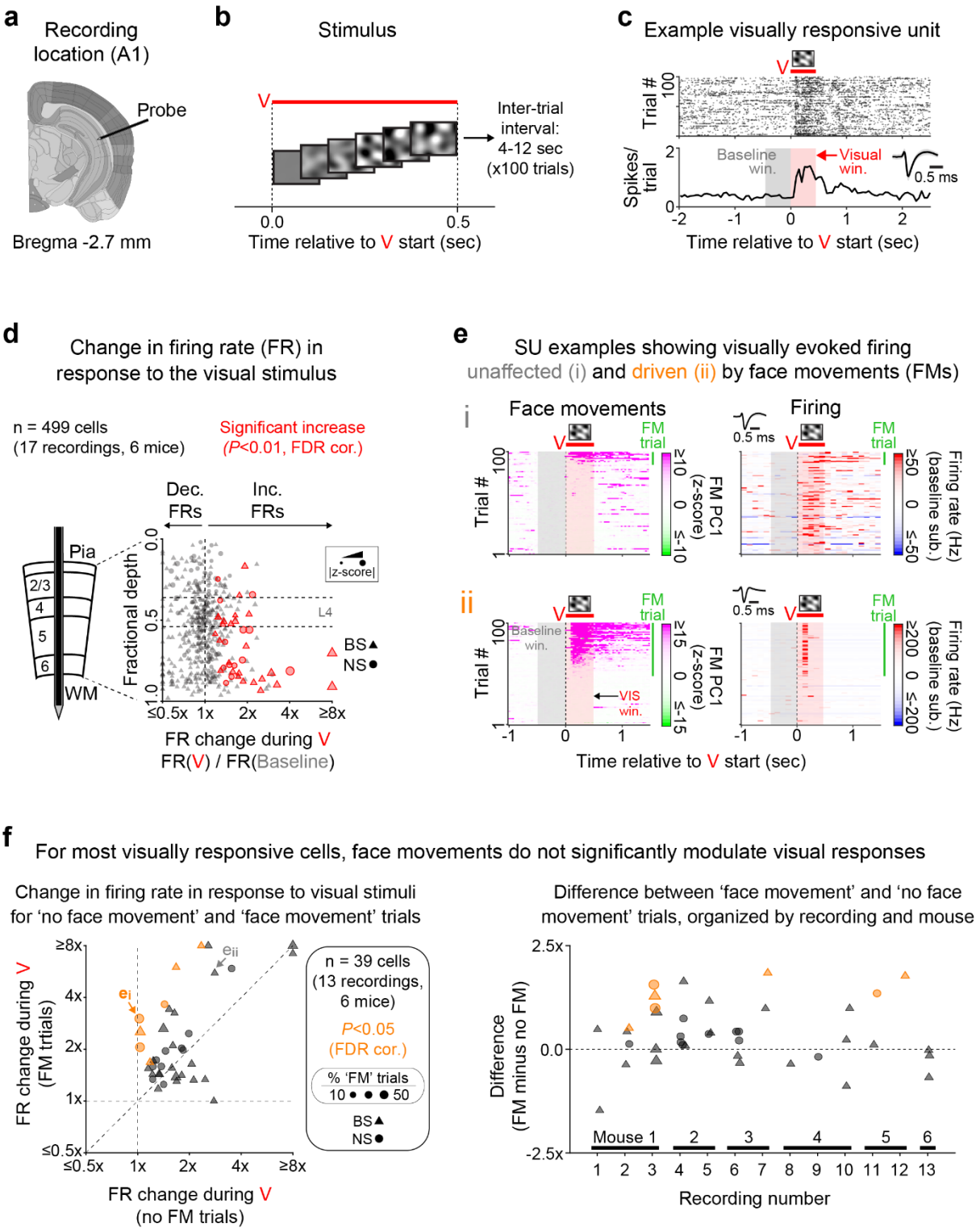


Fig. 2: Most visual responses in auditory cortex are not explained by face movements. **a**, Recording location. SU firing was recorded from all layers of the primary auditory cortex. **b**, For these experiments, we used visual stimuli only. These consisted of 100 contrast-modulated noise movies, each lasting 500 ms and separated by 4-12 seconds. **c**, Raster plot (top) and PSTH (bottom) showing significantly elevated firing in response to the visual stimulus for an example cell. Significance was determined using a two-sided Wilcoxon

signed-rank test comparing single-trial spike counts in the baseline (gray) and visual (red) windows ($P<0.01$, FDR corrected). **d**, Change in firing rate in response to the visual stimulus plotted against cell depth for all recorded cells (n = 499 cells, 17 recordings, 6 mice). Cells with significant increases in firing are shown in red ($P<0.01$, FDR corrected). **e**, Face movements (left) and SU spike counts (right) for two example recordings/cells (i and ii). The top example (i) shows a cell with firing largely unrelated to face movements. The bottom example (ii) shows a cell with firing strongly related to face movements. The green bars to the right of each plot indicate trials where face movements (PC1) in the visual window (red shading) exceed the 2x SD threshold. Trials (rows) are ordered by maximal face movement amplitudes in the visual window. **f**, (left) Change in firing rate for trials with (y-axis) and without (x-axis) face movement, for all cells with significant increases in firing in response to the visual stimulus (red cells in panel d). Cells showing a significant difference between the two trial types are marked in orange ($P<0.05$, FDR corrected, two-sided Mann-Whitney U tests). Most cells did not show a significant difference between trials with and without face movements (gray data points, 32 out of 39: 82%). (right) Effects of face movements split by recording and mouse. The plot shows the difference in firing rate change for face movement minus no face movement trials (y- minus x-axis values in figure 1f, left). We found no significant effect of recording (P = 0.21, Kruskal-Wallis tests) or mouse (P = 0.16, Kruskal-Wallis tests).

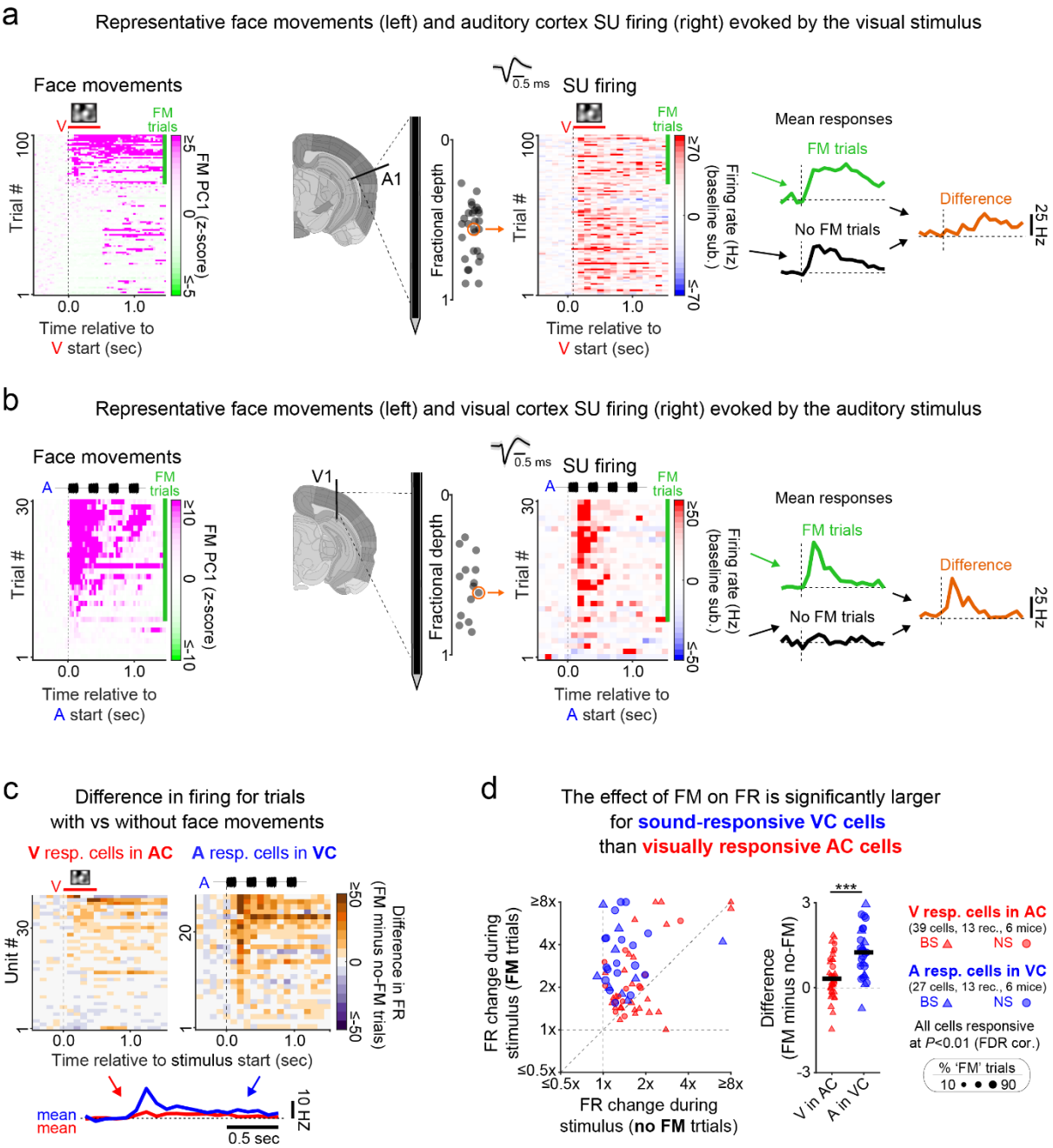
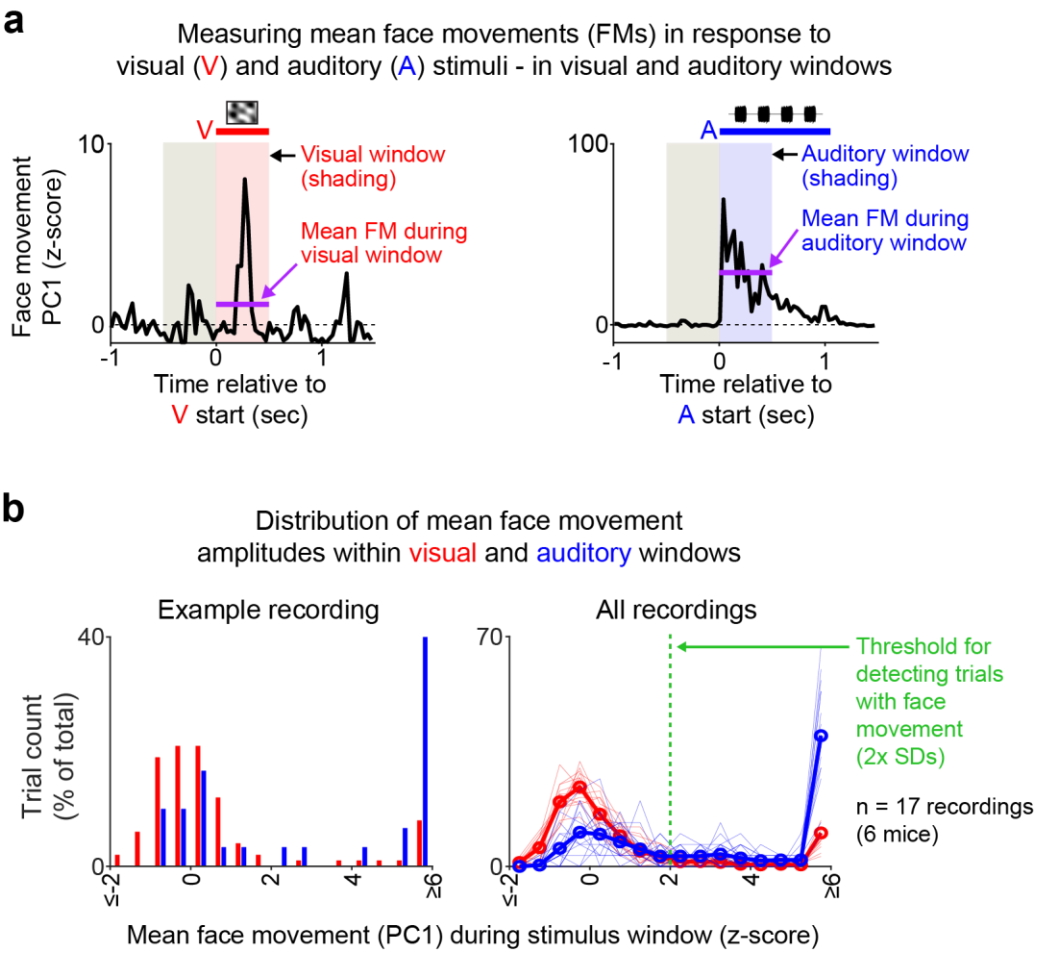
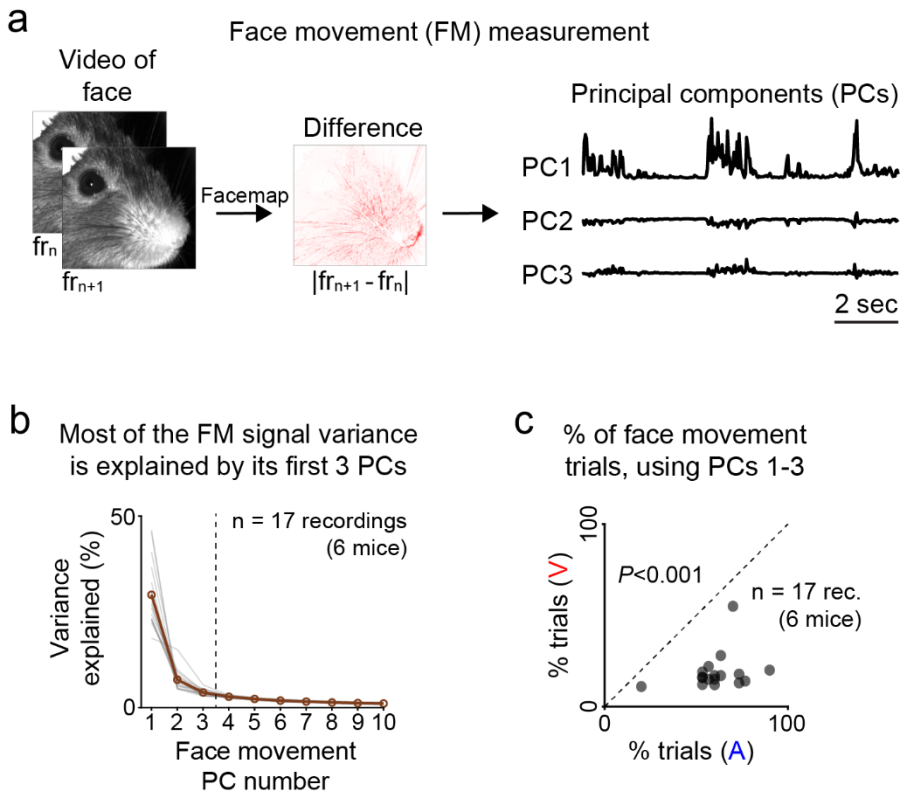


Fig. 3: The effect of face movements on cross-modal activity differs between auditory and visual cortices. **a**, Face movements (left) and SU firing (right) for a visually responsive unit in the auditory cortex. Mean firing differences between trials with and without face movement (far right, orange line) were minimal for this unit. **b**, Same as **a**, but for a sound-responsive unit in the visual cortex, where face movement had a substantial effect on stimulus-evoked firing. **c**, Firing rate differences between trials with and without face movement. Most visually responsive cells in the auditory cortex showed minimal differences (left), while sound-responsive cells in the visual cortex typically fired more during trials with face movement (right). **d**, Firing rate changes in response to the stimuli, for visually responsive cells in the auditory cortex (red) and sound-responsive cells in the visual cortex (blue), plotted for trials with (y-axis) and without (x-axis) face movement (left). For most

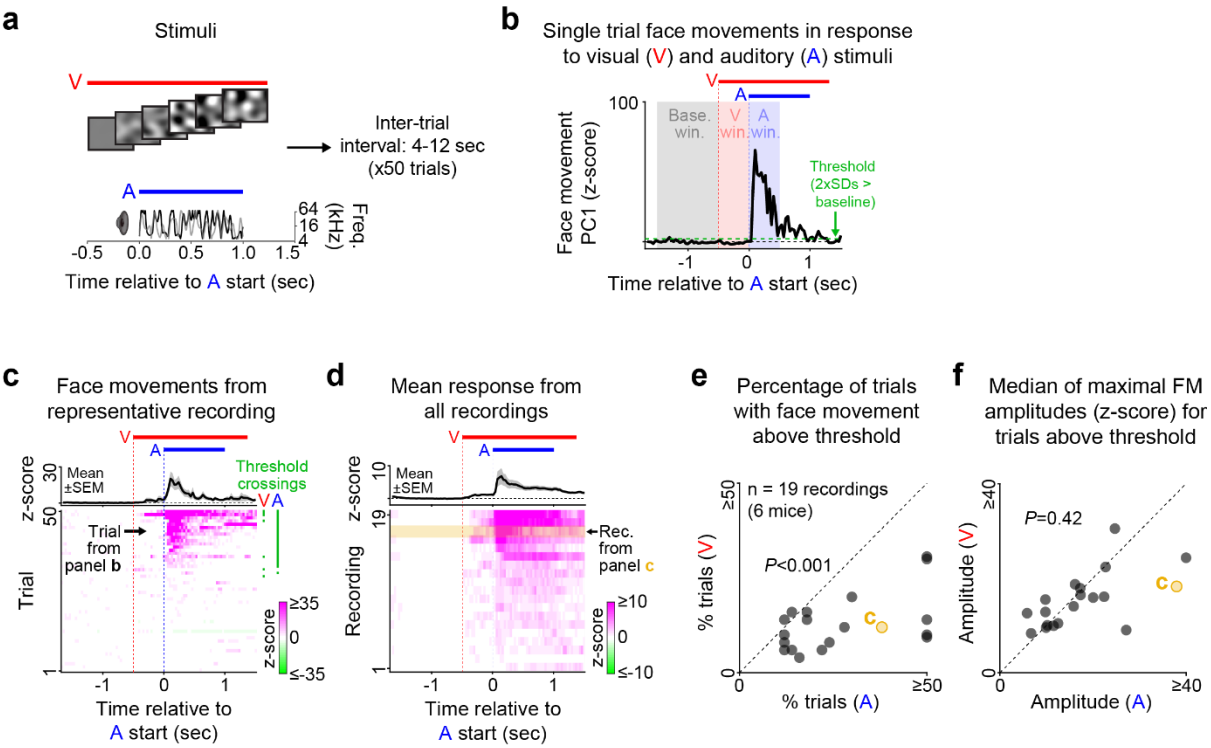
visually responsive cells in the auditory cortex, firing rate changes clustered near the diagonal, indicating a minimal effect of face movement. In contrast, firing rate changes landed substantially above the diagonal for nearly all sound-responsive cells in the visual cortex, indicating a positive, modulatory effect of face movements. Differences in firing rate changes (right) between face movement and no face movement trials were near zero for visually responsive cells in the auditory cortex, but significantly greater and predominantly positive for sound-responsive cells in the visual cortex (AC: 0.38 ± 0.74, VC: 1.2 ± 0.9, $^{*}P = 1.6 \times 10^{-4}$, two-sided Mann-Whitney *U* test).



Supplementary Fig. 1: Threshold selection for detecting trials with face movement. **a**, Single-trial face movements (PC1) from an example behavioral session showing mean stimulus-evoked responses (purple lines) during visual (red shading) and auditory (blue shading) trials. **b**, Histograms showing the percentage of trials with various mean response amplitudes, for an example behavioral session (left) and for all sessions (right; thin lines, histograms for individual sessions; thick line, grand mean across sessions). Nearly all sessions showed a bimodal distribution of face movement magnitudes, with one peak near zero (presumably corresponding to noise) and another peak corresponding to stimulus-evoked face movements. Based on these, we chose a threshold of +2 SD for separating trials with and without stimulus-evoked face movements (green dashed line).

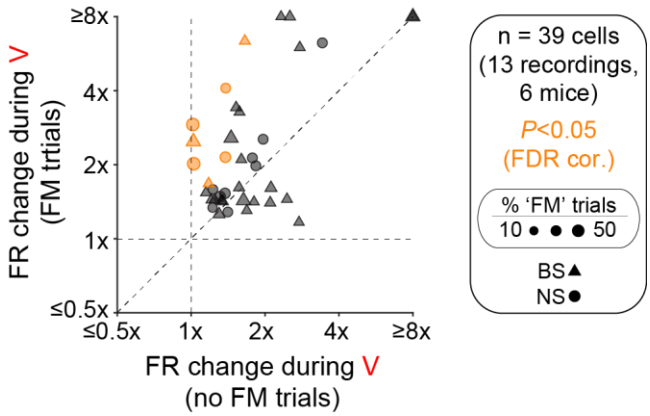


Supplementary Fig. 2: Even with more face movement PCs analyzed, visual stimuli remain less likely to evoke face movements than auditory stimuli. **a**, Multiple face movement principal components (PCs) were obtained from face videos using Facemap software. **b**, The percentage of variance explained by each face movement PC (up to 10), for individual behavioral sessions (thin gray lines) and their grand mean (thicker brown line with scatter points). PC1 consistently explained most of the variance, followed by far smaller contributions from PCs 2 and 3, and negligible contributions for PCs 4 and above. Following this observation, we tested if our results changed using PCs 1-3 rather than PC1 only (Fig. 1). **c**, Percentage of trials with above-threshold face movement (for at least one of PCs 1-3) for visual versus auditory stimuli. Even with additional PCs analyzed, the percentage of trials with above-threshold face movements remained significantly higher for auditory than for visual stimuli (V: 18.8 ± 10.2%, A: 61.0 ± 14.7%, two-sided Wilcoxon signed-rank test: $^*P = 2.9 \times 10^{-4}$).

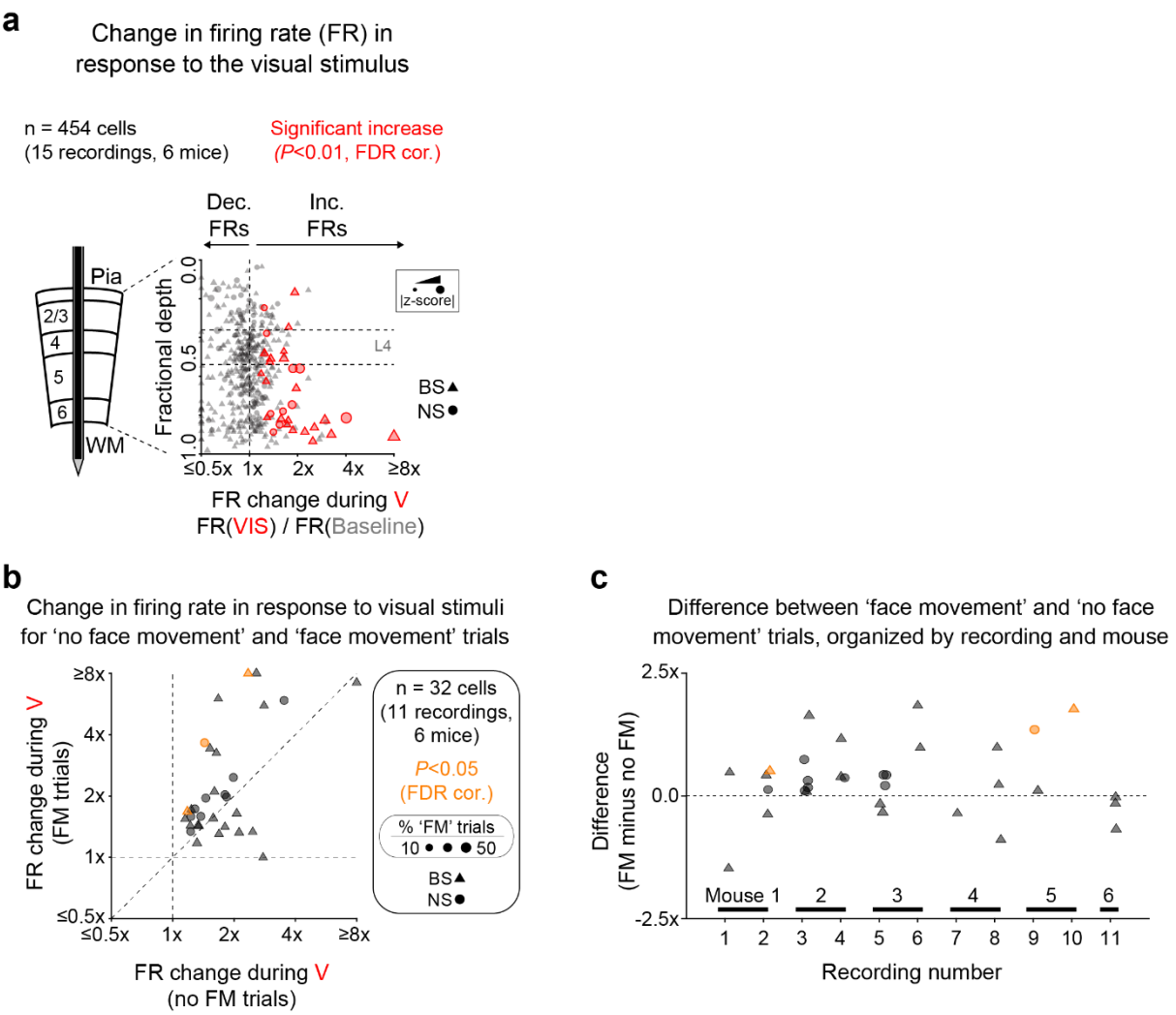


Supplementary Fig. 3: Visual stimuli remain less likely than auditory stimuli to evoke face movements in a different stimulus paradigm. We tested whether our main results from figure 1 held true when using a different stimulus paradigm. **a**, Stimulus trials included both visual (V) and auditory (A) stimuli. The visual stimuli were 50 different contrast-modulated noise movies, each lasting 1.75 seconds. The auditory stimuli were 50 different random double sweep sounds, each starting 0.5 seconds after the visual stimulus onset and lasting 1.0 second. Trials were separated by intervals of 4–12 seconds. **b**, Face movement (PC1) in response to the visual + auditory stimulus from a representative trial. Shaded areas indicate the baseline window used to z-score the face movement traces (gray; separate z-scoring for each trial), and the visual (red) and auditory (blue) windows in which visual- and sound-evoked face movements were measured. **c**, Trial-aligned face movements from a representative behavioral session, ordered by the maximal face movement. Green bars indicate trials on which movements exceeded the 2x SD threshold during the visual (V) or auditory (A) windows. The plot at the top shows the mean response across trials. **d**, Mean face movements for all behavioral sessions (bottom), and their grand mean (top). Rows are ordered by maximal face movement. **e**, Percentage of trials with face movements in the visual versus auditory windows. Similar to our previous stimulus paradigm (figure 1g), this percentage was significantly higher for auditory than for visual stimuli (V: 12.9 ± 7.3%, A: 30.8 ± 21.6%, $^*P = 2.8 \times 10^{-4}$, two-sided Wilcoxon signed-rank test). **f**, Median face movement amplitudes for above-threshold trials, compared between visual and auditory windows. Results are similar to our previous stimulus paradigm (figure 1h): although visual and auditory stimuli differ in their propensity to evoke face movements, there is no significant difference between the magnitude of face movements evoked during the auditory and visual response windows (V: 15.4 ± 5.8, A: 18.4 ± 11.5, $P = 0.42$, two-sided Wilcoxon signed-rank test).

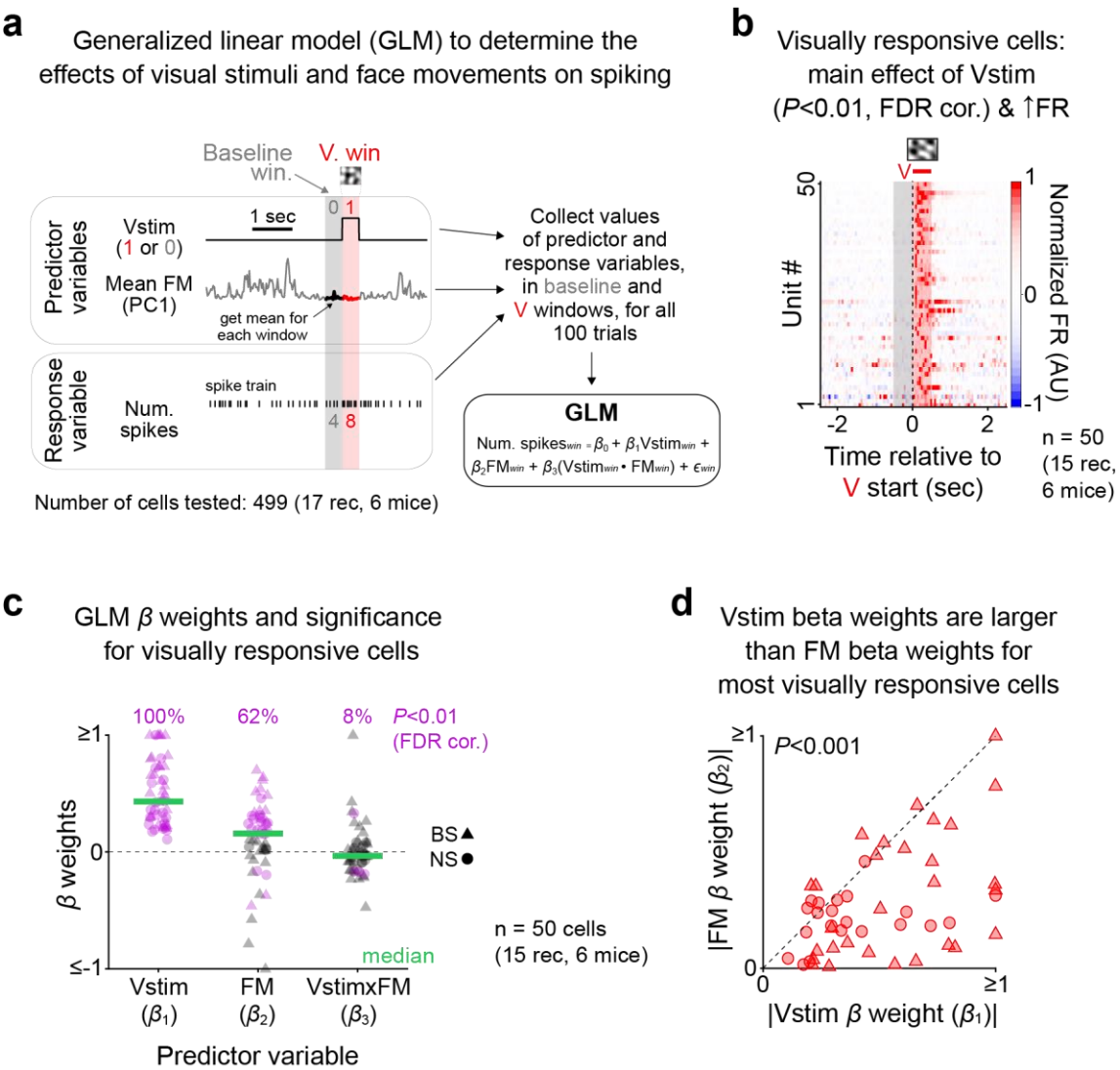
a Change in firing rate in response to the visual stimulus for 'no face movement' and 'face movement' trials, using the first **3** PCs to detect face movement



Supplementary Fig. 4: Even when face movements are analyzed using more principal components, they still do not account for most visual responses. **a**, As in figure 2f (left), but now face movement trials are defined based on movement in **any** of PCs 1-3. Consistent with our main results, most visually responsive cells did not show a difference in visually evoked firing between trials with and without face movement (gray data points, 31 out of 39, 79%).



Supplementary Fig. 5: Visual responses in primary auditory fields are largely independent of face movements. To assess whether our results changed when only including units from recordings with firing properties consistent with primary auditory fields (see Materials and Methods for criteria), we reanalyzed our main findings using data from these recording sites only. In total, 15 out of 17 recordings (from 6 mice) had firing properties consistent with primary auditory fields. **a**, From a total of 454 cells from these recordings (down from 499), 32 of these cells showed significantly elevated firing in response to the visual stimulus (down from 39 in figure 2d), and these were still mostly found in the deep layers. **b**, Out of these 32 visually responsive cells, 3 showed visually evoked firing that differed significantly between trials with and without face movements (orange data points; 9%, compared to 18% in figure 2f, left). **c**, In addition, similar to figure 2f (right), the effects of face movement on visually evoked firing were not dependent on recording (P=0.22, Kruskal-Wallis test) or mouse (P=0.20, Kruskal-Wallis test). These findings show that face movements still had little effect on visually evoked firing when only considering units from primary auditory fields.



Supplementary Fig. 6: A generalized linear model shows that visual responses are more related to the visual stimulus than to face movements. To validate our main findings in figure 2f using a complementary analysis we used a generalized linear model to analyze our data. **a**, Generalized linear model design for each unit. The model attempted to predict the number of spikes during 500 ms windows (win), either during the visual stimulus (red shading) or during the preceding baseline periods (gray shading), based on two predictor variables: a binary variable (Vstim) indicating the window type, with 0 for baseline windows and 1 for visual windows, and the mean face movement (FM) in each window (PC1). The GLM equation is shown in the bottom right. We tested for main effects of Vstim and FM, and for an interaction between these two variables. **b**, PSTHs showing firing in response to the visual stimulus for all visually responsive units, as determined by a significant main effect of Vstim (*P*<0.01, FDR corrected), and an increase in firing relative to baseline (*n* = 50 cells, 15 recordings, 6 mice). Firing rates are normalized by their absolute maximum in the visual response window (red shading). **c,** Model beta weights for the two predictor variables (Vstim and FM), and their interaction variable (Vstim x FM), are shown for all visually responsive cells (cells in panel **b**). The percentage of cells with significant beta weights (purple data points; *P*<0.01, FDR corrected) was highest for Vstim (100%), lower for FM (62%), and minimal for Vstim x FM (8%). **d,** Beta weights for the predictor variables

Vstim and FM (absolute values) were compared on a cell-by-cell basis for all visually responsive cells. Vstim beta weights were significantly larger than FM beta weights (*P = 4.7×10^{-7}, two-sided Wilcoxon signed-rank test). These results, along with those in panel **c**, further indicate that face movements do not account for most visually evoked firing.