What you hear is what you see: Sounds alter the contents of visual perception

Jamal R. Williams[1*], Yuri A. Markov[2,3], Natalia A. Tiurina[2,3], and Viola S. Störmer[1,4]

[1] Department of Psychology, University of California, San Diego
[2] HSE University, Russia
[3] Laboratory of Psychophysics, Brain Mind Institute, Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
[4] Department of Brain and Psychological Sciences, Dartmouth College
*Correspondence: jrwilliams@ucsd.edu

Visual object recognition in the real world is not performed in isolation, but is instead dependent on contextual information such as the visual scene an object is found in. And our perceptual experience is not just visual: objects generate specific and unique sounds which can readily predict which objects are outside of our field of view. Here, we test whether and how naturalistic sounds influence visual object processing and demonstrate that auditory information both accelerates visual information processing and modulates the perceptual representation of visual objects. Specifically, using a visual discrimination task and a novel set of ambiguous object stimuli, we find that naturalistic sounds shift visual representations towards the object features that match the sound (Exp. 1a-1b). In a series of control experiments, we replicate the original effect and show that these effects are not driven by decision- or response biases (Exp. 2a-2b) and are not due to the high-level semantic content of sounds generating explicit expectations (Exp.3). Instead, these sound-induced effects on visual perception appear to be driven by the continuous integration of multisensory inputs during perception itself. Together, our results demonstrate that visual processing is shaped by auditory context which provides independent supplemental information about the entities we encounter in the world.

When we look around the world, pertinent visual information is often ambiguous or indeterminate. To overcome this problem and to form meaningful representations, the visual system relies not only on the visual features of an object itself but also incorporates prior knowledge and concurrently available contextual information (1–7). This integration of available information is not exclusive to unimodal sources as available information from every sensory system is evaluated, weighed, and integrated to form a complete perceptual experience (8–16). However, the underlying mechanisms of how other sensory information affects visual processing—especially for complex and naturalistic stimuli—are not well understood. Here, we investigate whether and how naturalistic sounds alter our perceptual experience of visual objects. Specifically, we examine two questions: whether sounds that are related to visual objects speed the perceptual processing of them, and most importantly, whether these sounds alter visual

representations—shifting them away from the veridical visual features and towards those that are shared with the object that would generate that sound.

Previous work has shown that sounds enhance visual processing (17–23), and alter perception of simple visual stimuli (24, 25) for example, when two sounds produce the perception of two visual flashes, despite the presence of only a single visual stimulus (26). And, while the bulk of this work has used fairly simple stimuli, such as noise bursts and light flashes, naturalistic sounds have also been found to affect visual processing of objects, often reflected in faster response times or higher accuracy in object recognition tasks when sight and sound are congruent relative to maximally incongruent (12, 27). However, in this work, it is unclear whether sounds simply speed perceptual processing—leading to a more rapidly achieved but identical perception across conditions—or whether hearing particular sounds leads to changes in the visual representations themselves. For example, imagine you catch a glimpse of something rapidly flying by a window. It could be any number of things—and since you only saw it for a split second—auditory information could be incredibly useful at resolving this uncertainty: A constant buzzing would suggest it was likely a drone, whereas a stout caw would have you believe it was a crow. Does hearing the sound influence what you perceive? Does the sound of a drone make this dubious object appear more drone-like, whereas a caw makes the same object appear more crow-like? Or does a related sound simply accelerate object recognition without altering visual perception itself?

We addressed these questions by investigating how naturalistic sounds modulate the visual processing of ambiguous, real-world objects. We used a visual discrimination task with a perceptual locus (27, 28), and designed a novel set of object stimuli that were paired at random with related or unrelated sounds. Since the influence of sound on vision seems particularly effective when visual information is noisy or dubious—where sounds provide independent and unequivocal clues about the visual environment (8, 25, 29, 30)—we used ambiguous visual stimuli paired with clear and distinct sounds. Specifically, we created a set of ambiguous visual stimuli by morphing together the features of two visual objects (Objects A and B, e.g., a hammer and a seal, Fig 1a), and presented these stimuli with naturalistic sounds that were congruent with one of these progenitor objects. Visual objects and sounds were presented simultaneously, and participants looked for a target object in visual noise, after which they precisely reported that object using continuous report. We examined whether these sounds influenced how quickly participants recognized objects and, most importantly, whether the sounds—while not predictive of the target object in the visual discrimination phase—would alter the visuo-perceptual representation of a target object.

Results

Experiment 1: Incidental real-world sounds were paired with visual objects from a novel stimulus set that we created by morphing together the features of two distinct anchor objects (Object A and B, e.g., a hammer and a seal). By modulating the proportion of each anchor present in each object, we were able to create a series of novel, stepwise morphs that retained features from the original anchor objects (Fig 1A). The perceived

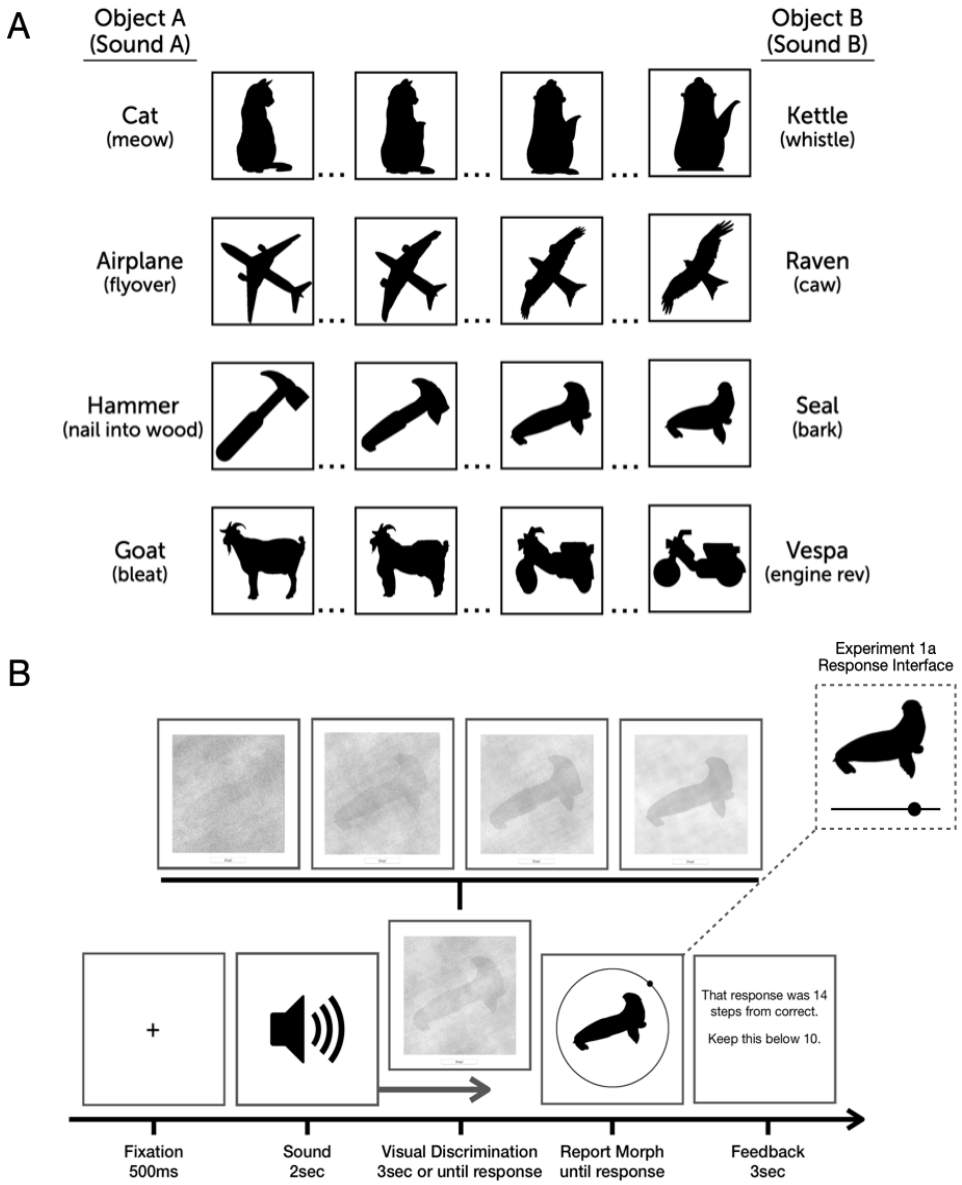

Figure 1. *Stimuli and Task. (A) The four object-pairs used in the experiments. The leftmost column shows anchor-objects A while the rightmost column shows anchor-objects B (with anchor-object sounds in parentheses). Between each anchor object were 98 unique morphs that maintained features of both anchor objects. (B) General task design. Sounds played while a noisy object slowly faded into view (example of denoising process above visual discrimination task panel). Experiment 1a used a linear response slider while Experiment 1b used a circular response wheel.*

ambiguity across the different morph levels was validated in an independent study in which we measured psychophysical functions for each stimulus pair.

On each trial, one of these morphs slowly faded into view from visual noise, while the sound of a real-world object played. Initially, the target morph was completely obscured by visual noise and as the trial progressed, this noise was continuously reduced such that participants were able to gather increasing amounts of visual information about

the target object. Participants were instructed to click with the mouse as soon as they were confident that they could accurately recreate the target morph using a continuous response interface that immediately followed the mouse click. Critically, the sounds could be either related or unrelated to the target morph: unrelated sounds were highly dissimilar from the target morph (e.g., a whistling train for the Hammer-Seal morphs) while related sounds matched the identity of one of the target morph's anchor objects[1]. After the mouse click, participants used a continuous response interface to precisely recreate the target morph from the visual discrimination phase and received feedback on their error.

Sounds influenced report error and response time (RT) on the visual discrimination and the continuous report phases, respectively. Specifically, we found that continuous report responses were influenced by which sound was played (sound A, B, or an unrelated sound; $F(2,36) = 10.05$, $p < 0.001$, $\eta^2 = 0.36$). Of main interest was whether the related sounds A and B differentially affected the same visual stimulus; thus, we next compared the mean error for each related sound to the error on unrelated sound trials – which matched the complexity and naturalistic properties of the related sounds, thus effectively serving as a neutral condition. These subsequent pairwise comparisons revealed that the sounds corresponding to anchor Object A shifted responses towards that side of the object-morph continuum and away from neutral ($t(18) = -2.16$, $p = 0.044$; *Cohen's* $d_z = 0.50$), while sounds corresponding to Object B pulled responses in the opposite direction ($t(18) = 2.57$, $p = 0.019$; $d_z = 0.59$; see Fig 2b).

We next focused on RT during the visual discrimination task, which reflects the rate by which visual information is meaningfully integrated into a complete object. Participants were faster, on average, when they heard a related sound (1638ms) compared to an unrelated sound (1682ms; $t(18) = 2.47$, $p = 0.023$, $d_z = 0.39$). This difference suggests that, on unrelated trials, participants required roughly 10% more visual evidence than on related trials to perform the task with roughly equal levels of accuracy (mean absolute error, 6.00 vs 6.07; $t(18) = 0.39$, $p = 0.7$, $d_z = 0.09$, $BF_{01} = 3.94$). Thus, auditory information accelerated visual feature extraction from the noisy images and possibly increased observers' confidence in their visual judgments as well (27).

Experiment 1a used a linear response interface where the leftmost edge corresponded to anchor object A (morph step 1) and the rightmost edge corresponded to anchor object B (morph step 100). It is therefore possible that participants used these reliable positions along the response slider as a cue when responding—instead of focusing on the visual features of the response morph itself. To mitigate these concerns, and to replicate Experiment 1a, in Experiment 1b, we implemented a response wheel that was rotated randomly on every trial so that across trials there was no correspondence between positions on the wheel and the visual response morph (Fig 1b). Results from Experiment 1b replicated Experiment 1a: we found that sounds had a reliable effect on report error ($F(2,78) = 11.23$, $p < 0.001$, $\eta^2 = 0.22$) and that related sounds pulled responses away from the average error on unrelated trials and towards the visual features of anchor Object A ($t(39) = -2.58$, $p = 0.013$, $d_z = 0.41$) and Object B ($t(29) = 2.77$, $p = 0.011$, $d_z = 0.43$; Fig 2). RT in the visual discrimination task was again faster when sounds

[1] Note that the anchor objects were never targets, and the visual and auditory stimuli were presented concurrently to capitalize on the tight temporal integration window during multisensory integration (34, 63).

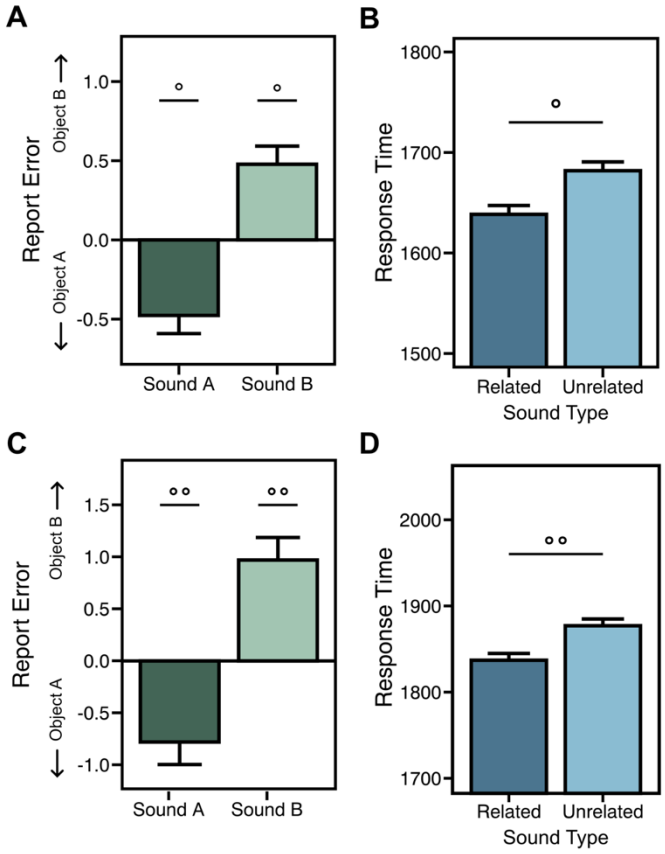

Figure 2; *Data from Experiments 1a (top row) and 1b (bottom row). Average report error (difference from neutral sounds) for Exp 1a (A) and Exp 1b (C) shows that related sounds influenced report error such that the response morph appeared more like the sound's anchor-object identity. Right column demonstrates that, for both Exp 1a (B) and Exp 1b (D), sounds influenced response time such that participants were faster when they heard a related sound compared to an unrelated one.*

were related to the target morph (1840ms) compared to unrelated sounds (1881ms; $t(39) = 2.55$, $p = 0.014$; $d_z = 0.40$) and, like before, this difference in RT did not result in a reliable difference in accuracy (6.83 vs 6.84; $t(39) = 0.03$, $p = 0.98$; $d_z = 0.004$ $BF_{01} = 5.86$). Taken together, the results from these experiments demonstrate that related auditory information speeds visual object processing, while also altering visual object representations by shifting them towards features that match the surrounding auditory context. This occurs even though the auditory information is non-predictive of the visual object.

Experiment 2: Based on the results from Experiment 1, we hypothesize that sounds influence concurrent visual processing by pulling ambiguous visual inputs towards visual features that are congruent with the auditory object; thus, exerting an effect on the

visual representations themselves. However, one alternative account is that the sounds influence later, non-perceptual processing stages, such as decisional and response processes. While such a post-perceptual account seems incompatible with faster RT for related sounds, we directly tested this alternative in Experiments 2a-b by presenting sounds where they should have the greatest impact over decisional processes: during the continuous report phase. Each trial began with the same visual discrimination phase, except with no sound and, after a button press, the visual input stopped, a real-world sound began to play, and the continuous report interface was presented (Fig 3). If the effect is largely driven by a decisional process (such as response bias or low-confidence responses), we would expect a similar, or perhaps even larger, effect of sound on visual perception relative to Experiments 1a-b. If, however, real-world sounds primarily affect perceptual, and not decisional processes, then this manipulation should eliminate or reduce the effect since perceptual processing is likely complete by the time participants begin reporting the target item.

In Experiment 2a, we found that sounds had little to no impact on report error *(F*(2,78) = 0.38, p = 0.69, η^2 = 0.009) and, as expected, RT on related (1911ms) and unrelated trials (1906ms) was not significantly different (t(39) = 0.29, p = 0.77, d_z = 0.04, BF_{01} = 5.63). A closer analysis of report error found no significant impact of sound: Error on unrelated trials was not significantly different than error on sound A trials (t(39) = 0.82, p = 0.42, d_z = 0.12) nor sound B trials (t(39) = 0.24, p = 0.81, d_z = 0.03) and we found compelling evidence to embrace these null findings (BF_{01} = 4.28 and 5.70, respectively).

In Experiment 2b, we combined manipulations from Experiment 1b and 2a in a within-subject design and varied whether sounds were played during the continuous report phase (like Experiment 2a) or instead, were played during the visual discrimination phase (like Experiments 1a-b). We found a main effect of sound (F(2,84) = 11.31, p < 0.001, η_p^2 = 0.12), no main effect of sound onset (during, or after, visual discrimination; F(2,84) = 0.16, p = 0.69, η_p^2 = 0.001) and a significant interaction *(F*(2,84) = 3.39, p = 0.035, η_p^2 = 0.04). To explore the interaction, we compared the effect of sound on report error, and found that sounds produced a significantly larger effect when they were played during the visual discrimination phase compared to when they were played during the continuous report phase (t(84) = 2.34, p = 0.021, d_z = 0.25; see Fig 3c). We next analyzed report error independently for each sound onset condition. When participants heard sounds during the visual discrimination task, we found that related sounds pulled responses away from neutral and towards anchor Object A (t(84) = 2.30, p = 0.024, d_z = 0.25) and Object B (t(84) = 2.96, p = 0.004, d_z = 0.32). However, and in contrast to these findings, when participants heard sounds during the continuous report task (Fig 3d), we

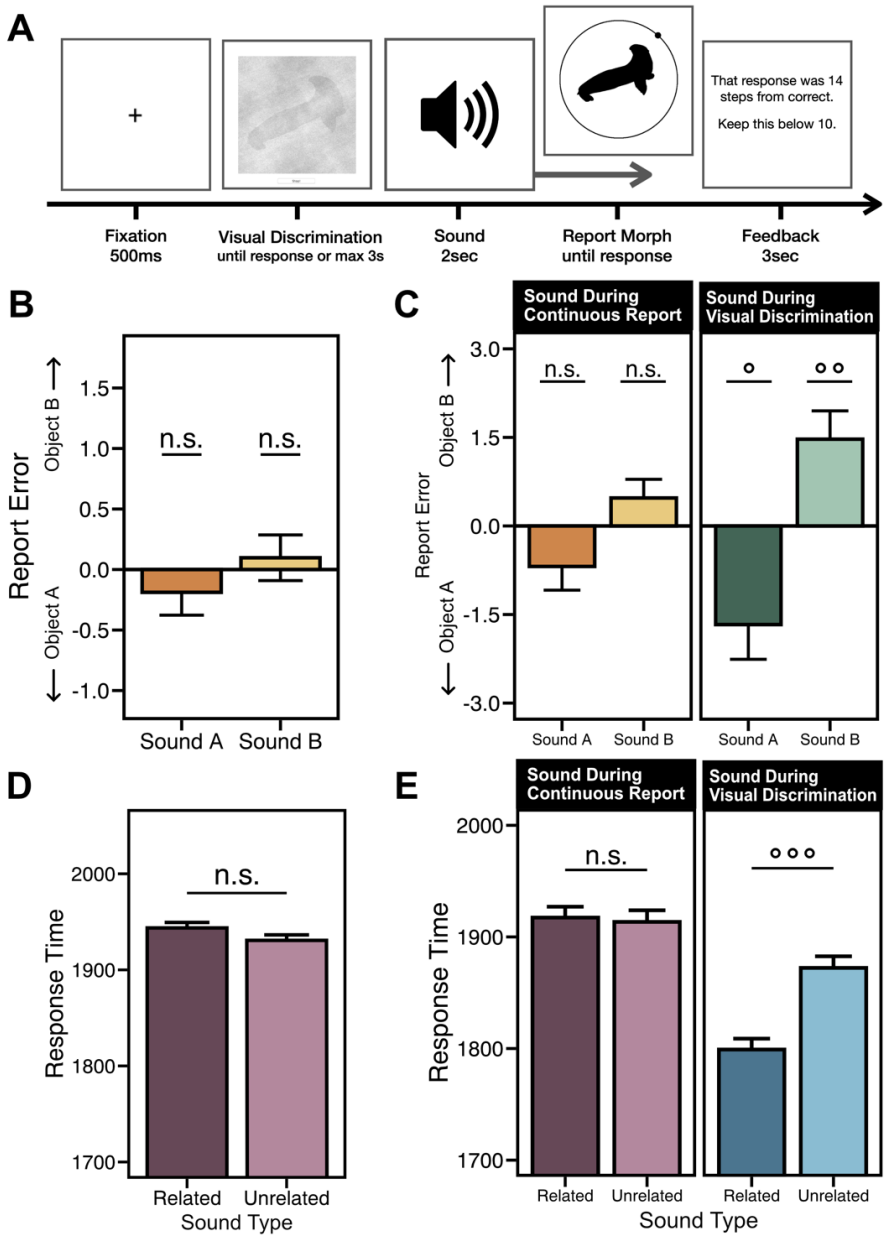

Figure 3 *Results and Task Design from Experiments 2a-b. (A) Task design: sounds were always played during the continuous report phase in Experiment 2a, and on half of all blocks in Experiment 2b. (B & C) Average report error (difference from neutral) for Experiment 2a and Experiment 2b (separated by when the sound onset began). These results show that related sounds influenced report error such that the response morph appeared more like the sound's anchor-object when the sound was played during the visual discrimination phase (C, green bars) and not when played during the continuous report phase (B-C, orange bars). (D) RT for related and unrelated trials in Experiment 2a. (E) RT for related and unrelated trials in Experiment 2b, separated by when the sound was played: during continuous report (purple bars) or visual discrimination phases (blue bars). Results show that RT was only reliably affected when sounds were heard during the visual discrimination phase.*

found that error on unrelated trials was not significantly different from error on sound A trials ($t(84) = 1.56$, $p = 0.12$, $d_z = 0.16$, $BF_{01} = 2.61$) and sound B trials ($t(84) = 1.42$, $p = 0.16$, $d_z = 0.15$, $BF_{01} = 3.18$).

Participants were significantly faster on related (1779ms) compared to unrelated trials (1852ms; $t(84) = 4.05$, $p < 0.001$, $d_z = 0.44$) when sounds played during the visual discrimination phase, and this difference in RT did not lead to differences in accuracy (7.76 vs 7.31; $t(84) = 1.21$, $p = 0.23$, $d_z = 0.13$, $BF_{01} = 4.13$; see Fig 3e). As expected, we observed no significant difference in RT between related (1899ms) and unrelated (1903ms) conditions when sounds were played during the continuous report phase ($t(84) = 0.25$, $p = 0.80$, $d_z = 0.03$, $BF_{01} = 7.99$). Overall, RT was on average slower when sounds were played during the continuous report task compared to the visual discrimination task but this difference in RT (i.e., having target images with lower levels of noise) did not lead to a significant difference in report error across sound onset conditions (7.38 vs 7.54, respectively; $t(84) = 0.62$, $p = 0.54$, $d_z = 0.07$, $BF_{01} = 6.94$; Fig 3). These results replicate the previous experiments and demonstrate that sounds have their greatest influence when they are presented concurrently with visual information and can thus be integrated directly with incoming visual information.

Experiment 3: Thus far, we have suggested that auditory and visual information are continuously integrated during sensory processing, and that biases at the decisional or response stages do not play a large role. However, another possibility is that the semantic content of these naturalistic sounds drives top-down influences on visual perception. Under this account, sounds may activate high-level semantic representations that subsequently influence sensory processing perhaps by biasing participants to voluntarily attend to or search specific parts of the visual object. To test for the contribution of such top-down effects, in Experiment 3, we presented the full length of a sound prior to the onset of the visual discrimination task (cf., 37, 38). Thus, this manipulation provides the same audio-semantic content as in previous experiments but should primarily induce pre-perceptual mechanisms that have been shown to require a longer delay between sound and target onset to exert effects on visual processing (31, 33–35).

We found that sounds did not have a significant impact over report error ($F(2,78) = 2.08$, $p = 0.13$, $\eta^2 = 0.05$) and we did not find a significant RT benefit for related sounds like we found in Experiments 1a-b and 2b ($t(39) = 1.73$, $p = 0.09$, $d_z = 0.27$, $BF_{01} = 1.50$). Preplanned t-tests of report error further demonstrated that error on unrelated trials was not significantly different than sound A ($t(39) = 1.19$, $p = 0.24$, $d_z = 0.18$, $BF_{01} = 3.04$) nor sound B trials ($t(39) = 0.64$, $p = 0.53$, $d_z = 0.10$, $BF_{01} = 4.84$). These results suggest that the multisensory effects we observed in Exp. 1a-b and Exp. 2b are driven by the continuous integration of auditory and visual information as it enters the senses, and that temporal overlap of the incoming information is critical, as predicted by multisensory integration accounts (34, 36–38).

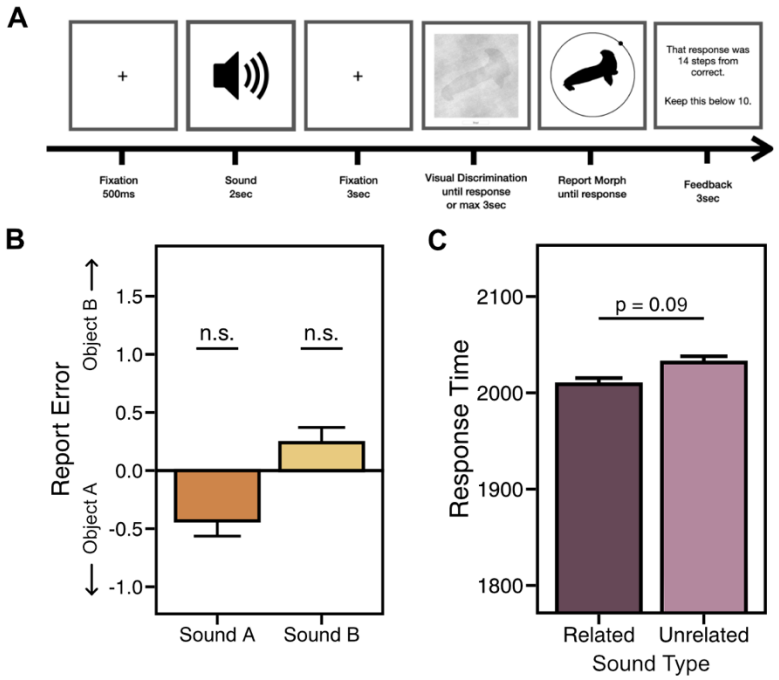

Figure 4 *Results from Experiment 3: Sounds were played prior to the onset of the visual discrimination phase. (B) Report Error: we found a substantially reduced, non-significant, effect (Cohen's d_z = 0.13) for sound A and sound B, suggesting that expectation and attention did not play a substantial role in the effects observed in Experiments 1a-b & 2b. (C) RT was also non-significant.*

Discussion

Our results show that naturalistic auditory information hastens the accumulation of related visual information and alters the perceptual representation such that it is shifted towards visual features that are congruent with the auditory stimulus: the same ambiguous object (e.g., a morph of 50% seal and 50% hammer) is perceived as more hammer-like when paired with a hammer sound, and more seal-like when paired with the sound of seal barking. In a series of control experiments, we demonstrate that these cross-modal effects are not due to biases at decision- nor response stages (Exp. 2a-b), nor are they driven by top-down effects (e.g., a volitional search for specific features; Exp. 3). Instead, sounds exert a reliable effect on visual perception only when both stimuli overlap temporally; consistent with previous work demonstrating that multisensory processing requires a tight temporal integration window (11, 35–37).

How might sounds exert influence over visual perception? In the natural world, sounds are causally predictive of the object that generated them—cats cannot bark, for example—and thus, sounds provide independent and informative cues about the visual world. This reliable and highly predictive relationship between audiovisual events can drive changes in early visual processing regions of the brain (18, 39) leading to selective processing of congruent compared to incongruent visual features. Previous work has shown a multitude of auditory effects on visual processing: that auditory information can rapidly affect the earliest stages of visual processing which results in downstream effects

on perception (18), that audition can dominate visual perception (8, 30, 40–43), that predictive relationships between stimuli lead to a selective reweighting of probabilistically relevant features (44–46), and that these effects are largely driven by prior knowledge (47, 48). For example, Kok and colleagues (2012) showed that when sensory information predicts an event, processing of probabilistically irrelevant features is suppressed relative to relevant features—those that are more likely to be observed—ultimately sharpening the processing of relevant sensory information. Taken together, these results lead us to hypothesize that the clear sounds presented in our study exerted a dominant influence over early visual processing which then led to a selective modulation of visual features that were inferred to come from the same generative object (i.e., ambiguous features are presumed dog-like when co-occurring with the sound of a barking dog). Additionally, within this framework, such sharpening of sensory processing can also lead to a facilitation of visual feature extraction for expected features, as evidenced by faster response times for related relative to unrelated sounds.

Another possible source of this effect may be that high-level semantic knowledge influences visual perception (12, 49). For example, presenting linguistic labels prior to a visual object has been shown to boost perceptual processing (31). However, while it is possible that perception is facilitated by high-level semantics and top down influences (but see 67), the present results are inconsistent with the hypothesis that activating semantic knowledge underlies the perceptual changes we observed here. Since the semantic content of real-world sounds alone did not reliably shift perceptual representations (Experiment 3) our results support a more implicit and low-level process of probabilistic inference (45, 47) where the purported effects of semantics and top-down goals on visual perception operate through separate mechanisms (32, 51, 52). Furthermore, finding that audiovisual events need to overlap temporally to exert an effect is also in line with the notion that the learned structure from the world—here, that sounds are exclusively produced by appropriate objects and that audiovisual events co-occur in time—influences how we perceive novel sensory information (49, 53–55).

What might be the functional role of sounds altering visual percepts? In real-world vision, where rapidly recognizing objects is imperative, where specific sounds are inextricably related to specific objects, and where sensory information is inherently noisy, it seems particularly beneficial to integrate cross-modal inputs with visual processing. Previous work has shown that prior knowledge and contextual information alter how we perceive novel sensory information and that this could be accomplished through probabilistic inference that is devoid of explicit top-down influences (1, 5, 47, 53, 56–58). If prior knowledge and context is used to interpret information as it enters the senses, as we propose here—and where the most likely cause of this sensory information is presumed as the source (59, 60)—it seems plausible that the resultant perceptual representation is shifted towards one that is congruent with these inferences (30, 42). Thus, the shifts in perceptual representations that we observed here represent inferences about impending sensory information: hearing a bark implies the presence of a dog and leads to perceptual inferences about certain dog-like visual features. And while an excessive shift of visual information might be maladaptive, our data demonstrate that these shifts are relatively small, at least when ample visual information is available, and that such small distortions likely represent an acceptable tradeoff between sensory processing demands and accuracy, especially considering that these inferences are

rarely wrong in the real world (61). Thus, in a world where sounds are causally predictive of visual objects and where preparing other sensory modalities for congruent experiences can lower the burden of sensory processing, integrating the auditory input with visual information might help to efficiently extract relevant features from noisy visual inputs, which ultimately aids object recognition.

Our results broadly relate to work that has shown influences of auditory context on visual-perceptual processing for real-world objects. For example, Chen & Spence (2011) showed that visual stimuli are processed more rapidly and more accurately when sounds (a dog barking) are directly congruent with the visual stimulus (line drawing of a dog) and presented in close temporal proximity. This is consistent with our finding that participants also responded faster during the visual discrimination phase when the sounds were related to the visual object[2]: participants appeared to accumulate visual information more rapidly or felt more confident about their visual representation when the sound matched one of the anchor objects. However, in previous work, these effects were often observed after explicit familiarization or training with the audiovisual stimuli, often through a task that required participants to report whether the sound and image were congruent, and typically involved rapid presentation of the visual stimulus—where some trials represent uncertain or low-confidence perception, possibly resulting in biases or specific response strategies (12, 13, 35). Here, we avoided these potential limitations and designed a novel task with a unique stimulus set that allowed us to measure naturally occurring cross-modal effects and directly assess the visual representations themselves. In particular: (1) Participants received no training and had no direct experience with the experimental stimuli prior to participating. (2) The task entailed accurately reporting the visual target, irrespective of the audiovisual relationship, thus avoiding any potential congruency biases. (3) Participants were in control over the amount of visual information they accumulated, thus allowing us to more confidently presume that participants had sufficient visual information to complete each trial accurately. Importantly, this demonstrates that this cross-modal effect is not limited to especially noisy perceptual representations, nor are they the product of uncertainty at response (especially since participants were encouraged on every trial to keep their error on the continuous report task as low as possible); suggesting that perceptual representations which participants feel confident in are nonetheless influenced by auditory context.

Overall, our findings demonstrate that the ongoing perceptual processing of novel, noisy, and ambiguous stimuli is altered by related auditory context, such that the ultimate perceptual representation is pulled towards sound-congruent features. Our results favor a multisensory, rather than a decisional or strategic account, in which visual and auditory information are continuously integrated such that inputs from one modality—in our case audition—trigger inferences about the world that the visual system uses to interpret concurrent ambiguous information. Most broadly, our study demonstrates the importance of investigating visual processing as an integrative rather than an isolated process (10), and that multisensory integration plays a critical role in forming visual object representations.

[2] Note that in the present study, sounds and target objects were never directly congruent as they were in previous work. Instead, the sound of an object would be paired with a visually ambiguous object that was a morph of related and unrelated visual features.

Materials and Methods

Participants

All participants (1) were undergraduates from UC San Diego, (2) participated in these online studies in exchange for course credit, (3) gave informed consent and were recruited in accordance with the procedures approved by the Institutional Review Board at UC San Diego, (4) were between 18 and 34 years old, and (5) reported having normal hearing and normal or corrected-to-normal vision. For Experiment 1a: 25 undergraduates participated (14 women; mean age 20.6 years), and six were removed for failing to meet predetermined inclusion criteria (see analysis below); Experiment 1b: n = 49 (35 women; mean age 20.52 years; nine removed); Experiment 2a: n = 50 (38 women, mean age 20.57 years; ten removed); Experiment 2b: n = 95 (75 women; mean age: 20.71 years; ten removed); Experiment 3: n = 47 (29 women; mean age 20.13 years; seven removed).

Stimuli

Eight real-world sounds were selected from online repositories and were edited to be two seconds in length and have equivalent amplitudes (when played at roughly 70dB SPL). For each sound, we collected or created a silhouette of a visual object that matched the object identity of the sound. Using each silhouette as the endpoints, we generated a set of 100 novel silhouettes by morphing between the two objects (Object A and Object B), using a morphing program to fuse objects together and create morph pairs (62).

Since the morphing process creates relatively arbitrary, psychologically non-uniform steps between 1 and 100, individual morph steps were rated in a separate online study to assess the psychometric functions for each of the morph pairs. From these data, we generated psychophysical curves and selected three morphs from each object-pair continuum that corresponded to the points where roughly 20, 50 and 80% of responses indicated the morph appeared more like Object A relative to Object B. Note that while we aimed to introduce variability in the stimulus set by selecting three different steps for each object-pair, we always chose morphs with some degree of ambiguity, and planned to collapse across these different morph levels for our main analysis to obtain enough trials in each condition of interest. In sum, the image set contained four unique object-pairs, each with three unique morphs (12 images total). For each object-pair, we collected and edited an additional unique sound that was unrelated to the object-pair and was selected to be as distinct as possible from the object-pair while related sounds were selected to closely match sounds made by either anchor-object A or B (see FIG 1A).

Visual discrimination task: First, to create the initial noise mask, we overlaid all 12 silhouette images, and completely randomized the phase of this averaged image. Second, we created a simple noise mask that consisted of random pixel flips and overlaid it above the initial phase-scrambled noise mask. Together, this resulted in a mask that effectively obscured the target morph silhouettes with both phased and random noise (see Fig 1B). Throughout each trial, as the mask slowly became more transparent, the randomized phase of the target image—which was initially randomized to 100%—slowly decreased until only 40% random phase remained.

General Procedure

Every trial began with a real-world sound that could be related to Object A, Object B, or completely unrelated to either object (33% of trials for each condition). 500ms after sound onset, the visual discrimination task (400x400 pixels) appeared centrally as the sound continued to play. The visual target object started completely obscured by visual noise which decreased by 1% every 5ms for 3sec or until the participant pressed a button on the mouse (indicating that they were confident enough to accurately perform the subsequent continuous report task). If participants did not press the button within 3s they received feedback encouraging them to accumulate visual information more quickly (these "wait" trials were removed from further analysis). The button press stopped the visual discrimination task immediately and began the continuous report task where one of the ambiguous morphs (300x300 pixels) was randomly selected as the starting point. This task was unspeeded and, once participants locked their response by clicking a mouse button, they received feedback on their error: number of steps from the correct answer for 3sec and were encouraged to respond accurately (less than 10 steps). Participants then pressed a button to initiate the next trial.

Procedure Experiment 1a

Participants completed 240 trials and the continuous report task was implemented with a linear response slider (400x10 pixels) which ranged from Object A (morph 1), leftmost point, to Object B (morph 100), rightmost point. This slider was positioned centrally and placed below the response morph (300x300 pixels) which changed continuously as the mouse moved along the slider (Fig 1a).

Procedure Experiment 1b

(120 trials) Exp. 1b was identical to Exp. 1a, except the continuous report task used a response wheel that was rotated randomly on every trial so that there was no correspondence between positions on the response wheel and the response morph, across trials (see Fig 1b). When the response screen appeared, a black ring (400x400 pixels) with a small position dot (50x50 pixels) appeared, surrounding the response morph (300x300 pixels).

Procedure Experiment 2a

(120 trials) The task was identical to Exp 1b except that sounds now started to play immediately following the visual discrimination phase and during continuous report; specifically, they started 500ms before the response interface appeared and continued to play as participants made their responses (up to 2s, or until a response was made).

Procedure Experiment 2b

(120 trials) In Exp. 2b, on half of all trials the sound started playing shortly before the visual discrimination task (as in Exp. 1a-b), and on the remaining half of the trials the sound was played after the visual discrimination task and during the continuous response

task (as in Exp. 2a). These sound-onset-conditions were blocked (30 trials per block), and the order of blocks was random and counterbalanced across participants.

Procedure Experiment 3

(120 trials) Each trial started with a real-world sound (2sec) and after a 3sec delay in which only the fixation point was shown to participants, the visual discrimination task began.

Analysis

For each sound condition we calculated median RT on the visual discrimination task and mean report error on the continuous report task. When comparing RT, we first checked to see whether RT differed between related conditions Sound A and Sound B and, across all experiments, we found no difference and thus collapsed RT estimates across Sound A and B. Report error, the number of morph steps between the correct response (target morph) and the provided response, could be negative, closer to morph 1 (Object A) than the correct response or positive, closer to morph 100 (Object B). We calculated a participant's mean error per sound condition (Sound A, B, and Unrelated) and submitted these data to an ANOVA. Report error in each figure is represented as the difference in average error between Related and Unrelated conditions.

Exclusion criteria were decided in advance based on pilot data. Data from participants were excluded if their average report error or average RT exceeded 3 standard deviations from the group mean. For each individual participant, all trials where report error or RT exceeded 4 standard deviations from their mean were excluded. Lastly, any trials where participants did not respond in the visual discrimination task—instead, opting to wait the entire duration of the trial—were excluded. Participants were excluded from further analysis if greater than 10% of trials were missing from their data set.

Acknowledgements

Thanks to Timothy Brady, Edward Vul, Maria Robinson, and Jonathan Keefe for their helpful conversations, comments, and suggestions.

Author Contributions

J.R.W, Y.M., N.T., and V.S. conceived of the research; J.R.W. and V.S. designed the experiments; J.R.W. performed the research and analyzed data; J.R.W, Y.M., N.T., and V.S. prepared the manuscript.

References

1. J. L. Davenport, M. C. Potter, Scene Consistency in Object and Background Perception. **15**, 559–564 (2004).
2. I. Biederman, R. J. Mezzanotte, J. C. Rabinowitz, Scene perception: Detecting and judging objects undergoing relational violations. *Cogn. Psychol.* **14**, 143–177 (1982).
3. I. Biederman, A. L. Glass, E. W. Stacy, Searching for objects in real-world scenes. *J. Exp. Psychol.* **97**, 22–27 (1973).
4. D. Draschkow, M. L. H. Võ, Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Sci. Rep.* **7**, 1–12 (2017).
5. M. Bar, Visual objects in context. *Nat. Rev. Neurosci.* **5**, 617–629 (2004).
6. T. Stein, D. Kaiser, M. V. Peelen, Interobject grouping facilitates visual awareness. *J. Vis.* **15**, 1–11 (2015).
7. S. Trapp, M. Bar, Prediction, context, and competition in visual recognition. *Ann. N. Y. Acad. Sci.* **1339**, 190–198 (2015).
8. D. Alais, D. Burr, The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Curr. Biol.* **14**, 257–262 (2004).
9. M. O. Ernst, M. S. Banks, Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
10. K. P. Körding, *et al.*, Causal inference in multisensory perception. *PLoS One* **2** (2007).
11. Y. C. Chen, C. Spence, When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition* **114**, 389–404 (2010).
12. Y. C. Chen, C. Spence, The Crossmodal Facilitation of Visual Object Representations By Sound: Evidence From the Backward Masking Paradigm. *J. Exp. Psychol. Hum. Percept. Perform.* **37**, 1784–1802 (2011).

13. T. R. Schneider, A. K. Engel, S. Debener, Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Exp. Psychol.* **55**, 121–132 (2008).

14. A. Thelen, D. Talsma, M. M. Murray, Single-trial multisensory memories affect later auditory and visual object discrimination. *Cognition* **138**, 148–160 (2015).

15. N. Van Atteveldt, M. M. Murray, G. Thut, C. E. Schroeder, Multisensory integration: Flexible use of general operations. *Neuron* **81**, 1240–1253 (2014).

16. D. R. Wozny, U. R. Beierholm, L. Shams, Human trimodal perception follows optimal statistical inference. *J. Vis.* **8**, 1–11 (2008).

17. F. Frassinetti, N. Bolognini, E. Làdavas, Enhancement of visual perception by crossmodal visuo-auditory interaction. *Exp. Brain Res.* **147**, 332–343 (2002).

18. M.-H. Giard, F. Peronnet, Aditory-Visual Integration during Multimodal object recognition in humans: A behavioral and electrophysiological study. *J. Cogn. Neurosci.* **11**, 473–490 (1999).

19. J. Vroomen, B. De Gelder, Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 1583–1590 (2000).

20. E. Van der Burg, C. N. L. Olivers, A. W. Bronkhorst, J. Theeuwes, Pip and Pop: Nonspatial Auditory Signals Improve Spatial Visual Search. *J. Exp. Psychol. Hum. Percept. Perform.* **34**, 1053–1065 (2008).

21. E. Van der Burg, D. Talsma, C. N. L. Olivers, C. Hickey, J. Theeuwes, Early multisensory interactions affect the competition among multiple visual objects. *Neuroimage* **55**, 1208–1218 (2011).

22. J. J. McDonald, W. A. Teder-Saälejärvi, S. A. Hillyard, Involuntary orienting to sound improves visual perception. *Nature* **407**, 906–908 (2000).

23. V. S. Störmer, J. J. McDonald, S. A. Hillyard, Cross-modal cueing of attention alters appearance and early cortical processing of visual stimuli. *Proc. Natl. Acad. Sci.* **106**,

22456–22461 (2009).

24. R. Sekuler, A. B. Sekular, R. Lau, Sound alters visual motor perception. *Nature* **385**, 308 (1997).

25. K. Watanabe, S. Shimojo, When Sound Affects Vision: Effects of Auditory Grouping on Visual Motion Perception. *Psychol. Sci.* **12**, 109–116 (2001).

26. S. Shams, L; Kamitani, Y.; Shimojo, What you see is what you hear. *Nature* **408**, 788 (2000).

27. J. R. Williams, V. S. Störmer, Auditory information facilitates sensory evidence accumulation during visual object recognition. *J. Vis.* **19**, 20c (2019).

28. J. Sadr, P. Sinha, Object recognition and Random Image Structure Evolution. *Cogn. Sci.* **28**, 259–287 (2004).

29. J. Heron, D. Whitaker, P. V. McGraw, Sensory uncertainty governs the extent of audio-visual interaction. *Vision Res.* **44**, 2875–2884 (2004).

30. T. Rohe, U. Noppeney, Sensory reliability shapes perceptual inference via two mechanisms. *J. Vis.* **15**, 1–16 (2015).

31. G. Lupyan, E. J. Ward, Language can boost otherwise unseen objects into visual awareness. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14196–14201 (2013).

32. D. Cox, S. W. Hong, Semantic-based crossmodal processing during visual suppression. *Front. Psychol.* **6**, 1–9 (2015).

33. B. Boutonnet, G. Lupyan, Words Jump-Start Vision: A Label Advantage in Object Recognition. *J. Neurosci.* **35**, 9329–9335 (2015).

34. Y. C. Chen, C. Spence, Audiovisual semantic interactions between linguistic and nonlinguistic stimuli: The time-courses and categorical specificity. *J. Exp. Psychol. Hum. Percept. Perform.* **44**, 1488–1507 (2018).

35. Y. C. Chen, C. Spence, Dissociating the time courses of the cross-modal semantic priming effects elicited by naturalistic sounds and spoken words. *Psychon. Bull. Rev.* **25**,

1138–1146 (2018).

36. Y. C. Chen, C. Spence, Crossmodal Semantic Priming by Naturalistic Sounds and Spoken Words Enhances Visual Sensitivity. *J. Exp. Psychol. Hum. Percept. Perform.* **37**, 1554–1568 (2011).

37. M. A. Meredith, J. W. Nemitz, B. E. Stein, Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *J. Neurosci.* **7**, 3215–3229 (1987).

38. H. Colonius, A. Diederich, Multisensory interaction in saccadic reaction time: A time-window-of- integration model. *J. Cogn. Neurosci.* **16**, 1000–1009 (2004).

39. N. M. van Atteveldt, B. S. Peterson, C. E. Schroeder, Contextual control of audiovisual integration in low-level sensory cortices. *Hum. Brain Mapp.* **35**, 2394–2411 (2014).

40. D. Burr, M. S. Banks, M. C. Morrone, Auditory dominance over vision in the perception of interval duration. *Exp. Brain Res.* **198**, 49–57 (2009).

41. S. Deneve, A. Pouget, Bayesian multisensory integration and cross-modal spatial links. *J. Physiol. Paris* **98**, 249–258 (2004).

42. C. R. Fetsch, A. Pouget, G. C. Deangelis, D. E. Angelaki, Neural correlates of reliability-based cue weighting during multisensory integration. *Nat. Neurosci.* **15**, 146–154 (2012).

43. M. Aller, U. Noppeney, *To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian causal inference* (2019).

44. P. Kok, J. F. M. Jehee, F. P. de Lange, Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron* **75**, 265–270 (2012).

45. F. P. de Lange, M. Heilbron, P. Kok, How Do Expectations Shape Perception? *Trends Cogn. Sci.* **22**, 764–779 (2018).

46. A. H. Bell, C. Summerfield, E. L. Morin, N. J. Malecek, L. G. Ungerleider, Encoding of Stimulus Probability in Macaque Inferior Temporal Cortex. *Curr. Biol.* **26**, 2280–2290 (2016).

47. P. Seriès, A. R. Seitz, Learning what to expect (in visual perception). *Front. Hum.*

Neurosci. **7**, 1–14 (2013).

48. A. A. Stocker, E. P. Simoncelli, Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* **9**, 578–585 (2006).

49. S. A. Los, E. Van der Burg, Sound speeds vision through preparation, not integration. *J. Exp. Psychol. Hum. Percept. Perform.* **39**, 1612–1624 (2013).

50. C. Firestone, B. J. Scholl, Cognition does not affect perception: Evaluating the evidence for top-down effects. *Behav. Brain Sci.* **39** (2015).

51. H. B. Helbig, M. O. Ernst, Visual-haptic cue weighting is independent of modality-specific attention. *J. Vis.* **8**, 1–16 (2008).

52. N. Gordon, N. Tsuchiya, R. Koenig-Robert, J. Hohwy, Expectation and attention increase the integration of top-down and bottom-up signals in perception through different pathways. *PLoS Biol.* **17**, 1–28 (2019).

53. C. Summerfield, T. Egner, Expectation (and attention) in visual cognition. *Trends Cogn. Sci.* **13**, 403–409 (2009).

54. C. Summerfield, F. P. De Lange, Expectation in perceptual decision making: Neural and computational mechanisms. *Nat. Rev. Neurosci.* **15**, 745–756 (2014).

55. A. Zuanazzi, U. Noppeney, Modality-specific and multisensory mechanisms of spatial attention and expectation. *J. Vis.* **20**, 1–16 (2020).

56. M. M. Chun, Y. Jiang, Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cogn. Psychol.* **36**, 28–71 (1998).

57. A. Oliva, A. Torralba, The role of context in object recognition. *Trends Cogn. Sci.* **11**, 520–527 (2007).

58. A. Yuille, D. Kersten, Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).

59. R. L. Gregory, Perceptions as hypotheses. *Philos. Trans. R. Soc. London. B, Biol. Sci.* **290**, 181–197 (1980).

60. D. C. Knill, A. Pouget, The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).

61. B. J. Fischer, J. L. Peña, Owl's behavior and neural representation predicted by Bayesian inference. *Nat. Neurosci.* **14**, 1061–1066 (2011).

62. J. Liao, *et al.*, Automating image morphing using structural similarity on a halfway domain. *ACM Trans. Graph.* **33** (2014).

63. P. Edmiston, G. Lupyan, What makes words special? Words as unmotivated cues. *Cognition* **143**, 93–100 (2015).