Psychonomic Bulletin & Review (2023) 30:2272–2281
https://doi.org/10.3758/s13423-023-02308-z

BRIEF REPORT

Audiovisual correspondence facilitates the visual search for biological motion

Li Shen[1,2,3] · Xiqian Lu[1,2,3] · Ying Wang[1,2,3] · Yi Jiang[1,2,3]

Accepted: 5 May 2023 / Published online: 25 May 2023

Abstract

Hearing synchronous sounds may facilitate the visual search for the concurrently changed visual targets. Evidence for this audiovisual attentional facilitation effect mainly comes from studies using artificial stimuli with relatively simple temporal dynamics, indicating a stimulus-driven mechanism whereby synchronous audiovisual cues create a salient object to capture attention. Here, we investigated the crossmodal attentional facilitation effect on biological motion (BM), a natural, biologically significant stimulus with complex and unique dynamic profiles. We found that listening to temporally congruent sounds, compared with incongruent sounds, enhanced the visual search for BM targets. More intriguingly, such a facilitation effect requires the presence of distinctive local motion cues (especially the accelerations in feet movement) independent of the global BM configuration, suggesting a crossmodal mechanism triggered by specific biological features to enhance the salience of BM signals. These findings provide novel insights into how audiovisual integration boosts attention to biologically relevant motion stimuli and extend the function of a proposed life detection system driven by local kinematics of BM to multisensory life motion perception.

Keywords Biological motion · Crossmodal · Audiovisual integration · Temporal correspondence · Visual search · Attention

Introduction

Vision and audition are two primary sensory modalities that work in concert to guide our attention in the dynamic world (Driver & Spence, 1998; Santangelo & Spence, 2007; ten Oever et al., 2016). For example, in a visual search task, a transient auditory cue synchronized with an abrupt change in a predefined visual target could facilitate the search performance, mostly reflected by faster reaction time (Chamberland et al., 2016; Gao et al., 2021; Van der Burg et al., 2008; Zou et al., 2012). Studies using a spatial cueing task further demonstrated that the tone-induced attentional facilitation in the visual search was independent of top-down task demands, pointing to a stimulus-driven mechanism whereby synchronous audiovisual cues enhance the salience of sensory events to capture attention (Matusz & Eimer, 2011; Turoman et al., 2021). Most of these findings are based on artificial, nonbiological stimuli and demonstrated a crossmodal effect elicited by synchronized but transient audiovisual events. However, relatively little is known about the audiovisual attentional facilitation for natural, biologically relevant stimuli, which usually have continuous and complex dynamic profiles and are likely to involve distinct mechanisms from those for nonbiological ones.

Biological motion (BM) is one of the naturally occurring dynamic stimuli of great evolutionary significance. It has a unique kinematic profile caused by muscle activity under the constraint of gravity (Vallortigara & Regolin, 2006; Wang et al., 2022). Empirical evidence has suggested that the visual processing of BM signals is modulated by the corresponding sounds based on spatial, temporal, or semantic congruency (Brooks et al., 2007; Mendonça et al., 2011; Meyer et al., 2013; Saygin et al., 2008; Schouten et al., 2011; Thomas & Shiffrar, 2010, 2013; van der Zwan et al., 2009; Wuerger et al., 2012). More importantly, the perception of

✉ Ying Wang
wangying@psych.ac.cn

[1] State Key Laboratory of Brain and Cognitive Science, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Psychology, Chinese Academy of Sciences, 16 Lincui Road, Chaoyang District, Beijing 100101, China

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China

[3] Chinese Institute for Brain Research, Beijing 102206, China

audiovisual temporal relations for BM differs from that for inverted BM that lacks the characteristic kinematic features but maintains low-level properties of normal BM (Saygin et al., 2008), and the temporal window of perceptual audiovisual synchrony is broader for BM than for artificial motion with constant speed (Arrighi et al., 2006). These findings raise the possibility that, different from the processing of non-BM stimuli, a specialized mechanism drives the processing of temporally congruent audiovisual BM cues and may yield a crossmodal attentional facilitation effect.

Notably, the specificity of BM perception arises mainly from the processing of two critical BM cues—namely, the global configuration cue representing the skeletal structure of static bodies and the local motion cue capturing the moving traces of critical joints (Hirai & Senju, 2020; Troje & Westhoff, 2006; Wang et al., 2010). Early studies suggest that global configuration cues are essential to BM perception, as observers can spontaneously recognize moving human figures from visual displays without local image motion (Beintema & Lappe, 2002; Bertenthal & Pinto, 1994). Nonetheless, recent findings highlight the pivotal role of local motion cues in visual BM perception. Above all, humans and some other species (e.g., chicks) exhibit an innate preference toward visual BM over non-BM signals, even when the BM stimuli had no identifiable global configuration but only local kinematic cues (Bardi et al., 2011; Simion et al., 2008; Vallortigara et al., 2005). Moreover, individual variations in the ability to discriminate locomotion direction from the local kinematic cues are genetically determined (Wang et al., 2018), and such abilities stem from one's sensitivity to the characteristic acceleration patterns carried by the feet motion (Chang & Troje, 2009; Troje & Westhoff, 2006). These findings hint at the existence of a neural mechanism selectively tuned to local BM information in the primitive brain system, which may serve as a 'life detector' for legged vertebrates (see reviews by Hirai & Senju, 2020; Lemaire & Vallortigara, 2022; Troje & Chang, 2023). While this 'life detector' responds to the dynamics of visual BM stimuli, whether it is sensitive to the temporal correspondence of auditory and visual BM signals and therefore causing a crossmodal attentional facilitation effect remains unexplored.

In the current study, we investigated whether and how concurrent auditory signals influence the selective attention of BM stimuli using an adapted visual search task. In Experiment 1, we examined whether listening to temporally congruent footstep sounds, as compared with listening to incongruent sounds, would promote the search for point-light walker (PLW) targets embedded in a crowd of PLW distractors. If audiovisual congruency could facilitate the attentional selection of BM signals, it might enhance search performance (e.g., result in shorter reaction time or higher search accuracy) in the audiovisual congruent condition than in the incongruent condition. Besides, by adopting a no-sound condition as the baseline, we examined whether the congruent sounds would confer benefits and the incongruent sounds would cause impairments to the search relative to the baseline. Results from Experiment 1 revealed a significant congruency effect mainly driven by the facilitation of congruent sounds.

To further examine the roles of the critical BM cues in the observed effect, we removed the local motion cue (Experiment 2) and the global configuration cue (Experiment 3) from the visual stimuli, respectively, while keeping the low-level temporal correspondence of the audiovisual signals unchanged. If the temporal congruence of sounds and local BM cues plays an essential role in the crossmodal effect, we should expect to observe this effect when the local BM cue is present (Experiment 3) but not when it is absent (Experiment 2). Otherwise, if the global BM configuration is crucial to the crossmodal effect, we might observe the effect only when the global configuration is intact (Experiment 2) but not when it is deprived (Experiment 3). Finally, we removed both global and local BM cues (Experiment 4) to assess whether crossmodal facilitation arose simply from low-level audiovisual correspondence that remained identical in all the experiments. If so, the crossmodal effect should occur in Experiment 4; otherwise, such an effect might disappear in Experiment 4.

Methods

Participants

A total of 80 participants (45 female, mean age$\pm SD$ = 21.6$\pm$2.2 years) took part in the study, 20 (11–12 females) in each Experiment. A two-tailed power analysis using G*Power (Faul et al., 2007) indicated that a sample size of at least 15 participants would afford 80% power to detect an audiovisual integration effect with a high effect size (Cohen's $d \geq 0.8$), based on the results of previous studies (Chamberland et al., 2016; Saygin et al., 2008). We have further increased the sample size to 20 per experiment to adequately detect the potential interactions across experiments. All participants reported normal hearing and normal or corrected-to-normal vision. They were naïve to the purpose of the study and gave informed consent according to procedures and protocols approved by the institutional review board of the Institute of Psychology, Chinese Academy of Sciences.

Stimuli

Visual Stimuli Figure 1b depicts the experimental design and the visual stimuli used in each experiment. In Experiment 1,

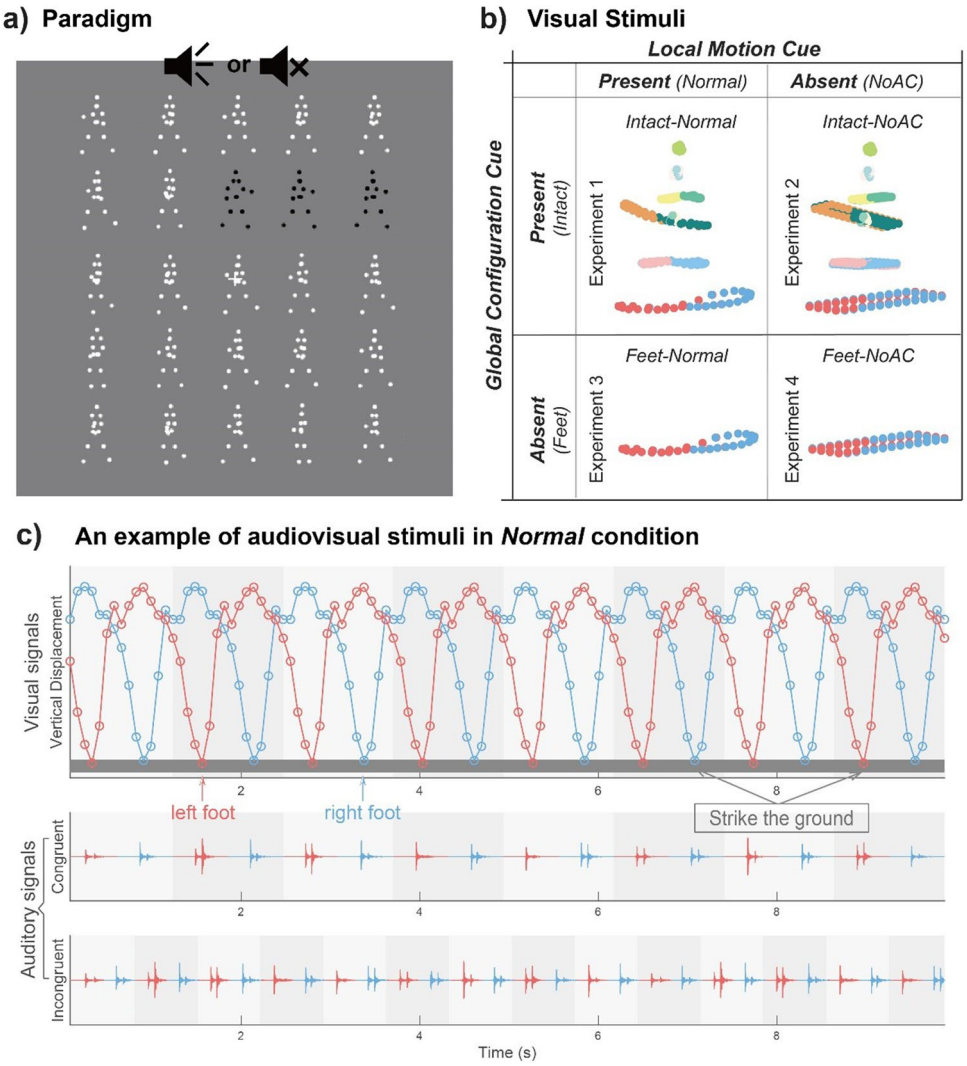

Fig. 1 Illustrations of the experimental paradigm and stimuli. **a)** An example of the search display used in Experiment 1. The targets consisted of three collinear PLWs (shown in black for illustration only) whose walking direction differs from that of the distractors. Observers were required to judge whether the visual target appeared in a row or a column while listening to footstep sounds or no sound. The frequency of the footstep sounds could be congruent or incongruent with that of the gait cycle of the target walkers in different trails. **b)** Visual point-light stimuli used in Experiments 1–4. Individual joints of the stimuli were rendered in different colors to depict the motion trajectories of the joints within one gait cycle. All dots were white in the formal experiment. Different rows show BM stimuli with/without intact global configuration cues (*Intact/Feet*); different columns show BM stimuli with/without normal local kinematic cues (*Normal/NoAC*). **c)** An example of the visual and auditory signals in the *Normal* acceleration condition. Upper panel: The vertical displacement of the left foot (red line) and the right foot (blue line) changes nonlinearly over time within each gait cycle (highlighted by grey background colors), resulting in an acceleration pattern that occurs periodically at a constant frequency. The grey bar at the bottom of the figure (or the minimum y value of the dots) represents the position of the ground. Lower panel: waveforms of the temporally congruent and incongruent footstep sounds displayed on the same timeline. (Color figure online)

the visual stimuli were normal PLWs (*Intact-Normal*) that consisted of 13 point-light dots attached to the head and the major joints of a human walker (Vanrie & Verfaillie, 2004). The temporal frequency of the walker is defined by the number of gait cycles (two steps) per second. In Experiment 2, we removed the vertical and horizontal acceleration cues from each limb-joint of the normal PLWs to disrupt the natural, gravity-compatible kinematic profile of BM (*Intact-NoAC*), generating point-light stimuli with intact global configuration but impaired local motion cues (Chang et al., 2018; Chang & Troje, 2009). In Experiment 3, we displayed feet-only sequences with normal acceleration profiles (*Feet-Normal*), given that such stimuli convey the most critical local motion information in BM while lacking intact global configuration (Bardi et al., 2014; Troje & Westhoff, 2006; Wang et al., 2014). In Experiment 4,

we further removed acceleration cues from the feet-only sequences (*Feet-NoAC*) to disrupt both the local and global BM cues, creating nonbiological motion stimuli that were matched with BM only in low-level properties.

Auditory stimuli Auditory stimuli were continuous footstep sounds (10 s) with a sampling rate of 44100 Hz. The sound sequences contain periodic impulses whose peak amplitudes occur around the points when the foot strikes the ground, with little variations in the waveforms of each step (Fig. 1c). We manipulated the duration interval between the time points when the two feet hit the ground to control the temporal frequency of the footstep sounds. Note that the auditory signals were always the same in all the experiments and were equally congruent with the visual stimuli regardless of the manipulation of accelerations, as we maintained the alignment between the sounds and the visual footstep events when removing the acceleration cues from the visual stimuli.

Stimuli presentation All stimuli were displayed using MATLAB (MathWorks, Inc) together with PsychToolbox extensions (Brainard, 1997; Pelli, 1997). Participants sat 60 cm from the screen in a dim room, with their heads stabilized on a chinrest. The visual stimuli were displayed on a 37.5 cm × 30 cm CRT monitor with a resolution of 1,280 × 1024 pixels. The refresh rate was 85 Hz. The sounds were presented binaurally through headphones.

Procedure and design

Experiment 1 Each trial began with a white fixation cross (0.6° × 0.6°) displayed at the center of the screen. Participants were required to maintain fixation on that cross throughout the trial. After 1,000~1,500 ms, a visual search display with 25 PLWs appeared (Fig. 1a). Participants were required to search for the target—three collinear PLWs with a predefined walking direction different from those of the distractors—among a 5 × 5 array and judge whether the target appeared in a row or a column by pressing different keys as quickly as possible while minimizing errors. Each PLW subtended approximately 0.84°×2.18°. The distance between the centers of every two PLWs was 2.71° in horizontal and 3.05° in vertical. The target-PLW walked at a higher or lower speed (i.e., at the temporal frequency of 1.42 Hz or 0.81 Hz) compared with the distractors. To ensure that any three aligned distractors would not walk at the same pace, we assigned two different frequencies to the distractors: 0.81 & 1.13 Hz when the target frequency was 1.42 Hz, and 1.42 & 1.13 Hz when the target was 0.81 Hz. The visually displayed PLWs were paired with footstep sounds, which had the congruent (1.42 Hz or 0.81 Hz) or incongruent (0.81 Hz or 1.42 Hz) frequency relative to the visual target. In addition to the audiovisually congruent and incongruent conditions, we adopted a no-sound condition as the baseline. There were 40 trials in each condition. We divided these trials into two baseline blocks and four sound blocks based on the frequency of the footstep sounds, with each baseline block present in between two sound blocks.

Experiments 2–4 Experiments 2–4 had the same procedures as Experiment 1, except for the visual stimuli (see the Visual Stimuli section for details), and we used a 4 × 4 stimulus array to control for task difficulty.

Data Analyses

For each participant, correct trials with reaction time (RT) within 3 standard deviations from the mean were included in further analysis. The percentages of trials excluded from the analyses were 0.77% for Experiment 1, 0.31% for Experiment 2, 0.83% for Experiment 3, and 0.78% for Experiment 4. For each experiment, a one-way repeated-measures analysis of variance (ANOVA) was used to test the influence of three audiovisual conditions (congruent, baseline, incongruent) on reaction time and accuracy. In the cross-experiment analysis, to deal with the weakness of adopting a between-subject design in such analyses, we calculated a normalized audiovisual congruency effect to correct for individual variances in search performance across experiments. Meanwhile, considering the significant sound effect on both reaction time and accuracy in Experiment 1 and any potential speed–accuracy trade-offs in other experiments, we calculated the normalized audiovisual congruency effect based on the inverse efficiency (IE) scores that take into account both reaction time and accuracy (Ngo & Spence, 2012; Spence et al., 2001), as follows $\left(\frac{IE_{Incongruent}-IE_{Congruent}}{(IE_{Congruent}+IE_{Incongruent}+IE_{Baseline})/3}\right)$. We also calculated the normalized benefit effect induced by congruent sounds $\left(\frac{IE_{Baseline}-IE_{Congruent}}{(IE_{Congruent}+IE_{Incongruent}+IE_{Baseline})/3}\right)$ and the cost effect induced by incongruent sounds $\left(\frac{IE_{Incongruent}-IE_{Baseline}}{(IE_{Congruent}+IE_{Incongruent}+IE_{Baseline})/3}\right)$ relative to the baseline in the similar way. Then, a two-way ANOVA, with global configuration (present vs. absent) and local motion (present vs. absent) cues as between-subjects factors, was conducted to compare each normalized sound effect (i.e., the congruency, benefit, or cost effect) across experiments.

Results

Experiment 1: Temporally congruent footstep sounds facilitate the visual search of BM signals

A repeated-measures ANOVA on RT revealed a significant main effect of sound conditions (congruent, incongruent, baseline; Fig. 2a), $F(2, 38) = 4.864$, $p = .013$, $\eta_p^2 = 0.204$.

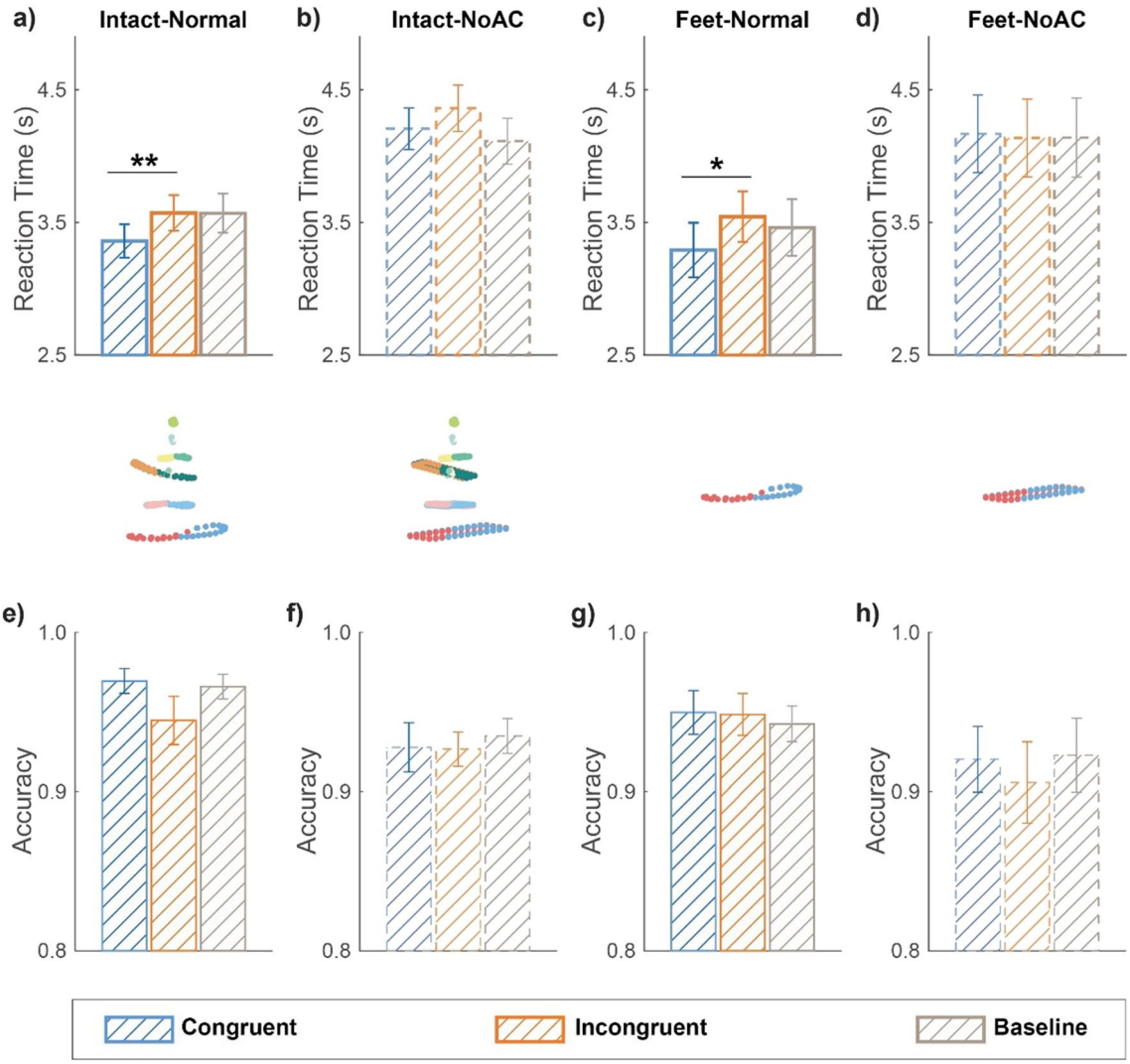

Fig. 2 Mean reaction times **(a–d)** and accuracies **(e–h)** for different sound conditions in Experiments 1–4, where the BM stimuli have intact or not intact global configuration cues (Intact vs. Feet) and normal or disrupted local kinematic cues (Normal vs. NoAC). $^{**}p < .01$, $^{*}p < .05$, after Bonferroni correction. Error bars represent ±1 standard error of means. (Color figure online)

A post hoc test showed that RT for incongruent trials was significantly slower than that for congruent trials, $t(19) = 3.621$, $p_{\text{uncorrected}} = .002$, $p_{\text{Bonferroni}} = .005$, indicating the presence of a crossmodal facilitation effect. Moreover, compared with the baseline condition, congruent audiovisual signals marginally facilitated the visual search, $t(19) = -2.358$, $p_{\text{uncorrected}} = .029$, $p_{\text{Bonferroni}} = .088$, while incongruent audiovisual signals did not impair the performance, $t(19) = 0.033$, $p_{\text{uncorrected}} = .974$, $p_{\text{Bonferroni}} = 1.000$.

The same analysis was performed on accuracy. There was a significant main effect of sound conditions (Fig. 2e), $F(2, 38) = 3.356$, $p = .045$, $\eta_p^2 = 0.150$, and no significant difference was observed among the three conditions, except that the accuracy in the congruent condition was higher than that in the incongruent condition before Bonferroni correction, congruent vs. incongruent, $t(19) = 2.266$, $p_{\text{uncorrected}} = .035$, $p_{\text{Bonferroni}} = .106$; baseline vs. congruent, $t(19) = -0.556$, $p_{\text{uncorrected}} = .585$, $p_{\text{Bonferroni}} = 1.000$; baseline vs. incongruent, $t(19) = 1.672$, $p_{\text{uncorrected}} = .111$, $p_{\text{Bonferroni}} = .333$.

Experiment 2: No crossmodal facilitation effect with disrupted local BM cues

To further determine the specific contributions of local and global BM cues to the observed audiovisual facilitation effect, we carried out Experiments 2–4. Research suggests that the specificity in visual BM perception is driven by gravity-compatible acceleration cues in the local BM signals (Chang & Troje, 2009; Troje & Chang, 2023; Troje &

Westhoff, 2006; Vallortigara & Regolin, 2006). If these local kinematic cues are critical to triggering the audiovisual integration of BM information, we should expect that removing such cues would eliminate the crossmodal facilitation effect.

Results from Experiment 2 supported this hypothesis. When acceleration cues were removed, despite the significant main effect of sound condition (Fig. 2b), $F(2, 38) = 3.763$, $p = .032$, $\eta_p^2 = 0.165$, the difference between the incongruent and congruent conditions was no longer evident, $t(19) = 1.656$, $p_{uncorrected} = .114$, $p_{Bonferroni} = .343$. The RTs in the incongruent and congruent conditions were both greater than that in the baseline, yet only the incongruent condition significantly slowed the search performance relative to the baseline, $t(19) = 3.085$, $p_{uncorrected} = .006$, $p_{Bonferroni} = .018$. In addition, no significant effect of auditory conditions was observed on accuracy (Fig. 2f), $F(2, 38) = 0.242$, $p = .786$, $\eta_p^2 = 0.013$. These results suggest that local BM information is critical for the crossmodal search facilitation effect observed in Experiment 1.

Experiment 3: Local motion alone could induce the audiovisual facilitation effect

In Experiment 3, we further explored whether local BM information alone was sufficient to explain the audiovisual facilitation effect by using the feet-only BM sequences, as the feet motions convey the most critical local BM cues but without intact global configuration (Chang & Troje, 2009; Troje & Westhoff, 2006). In line with the results from Experiment 1, the main effect of sound conditions was significant on RT (Fig. 2c), $F(2, 38) = 4.257$, $p = .021$, $\eta_p^2 = 0.183$. In particular, search performance was faster in the congruent condition relative to the incongruent condition, $t(19) = -3.365$, $p_{uncorrected} = .003$, $p_{Bonferroni} = .010$, suggesting that the crossmodal effect can occur even without intact global configuration information. Compared with the baseline condition, congruent sounds showed a tendency to speed up the search response, $t(19) = -1.939$, $p_{uncorrected} = .067$, $p_{Bonferroni} = .202$, and incongruent sounds did not yield an evident cost, $t(19) = 0.824$, $p_{uncorrected} = .420$, $p_{Bonferroni} = 1.000$. There was no significant main effect of sound conditions on accuracy (Fig. 2g), $F(2, 38) = 0.283$, $p = .755$, $\eta_p^2 = 0.015$.

Experiment 4: The absence of facilitation effect without characteristic BM cues

If the BM-specific local motion cues (the characteristic acceleration patterns in the feet movements) account for the effect observed in Experiment 3, removing such acceleration cues from the feet motion sequences should eliminate the crossmodal facilitation effect despite that such manipulation did not alter the low-level temporal correspondence in the audiovisual stimuli. To test this assumption, we conducted Experiment 4. As expected, a repeated-measures ANOVA showed no significant main effect of the sound conditions on RT (Fig. 2d), $F(2, 38) = 0.049$, $p = .953$, $\eta_p^2 = 0.003$. In addition, no significant sound effect was observed for accuracy (Fig. 2h), $F(2, 38) = 0.770$, $p = .470$, $\eta_p^2 = 0.039$. These results show that temporally congruent sounds could not facilitate the search of non-BM signals, suggesting that the facilitation effects observed in Experiments 1 and 3 were specific to BM processing.

Cross-experiment analysis: audiovisual facilitation hinges on local motion cues independent of global configuration

To further compare the crossmodal effects across Experiments 1-4 and examine whether there is an interaction between local and global BM cues, we calculated a normalized audiovisual congruency effect (see Methods) to control for the influence of individual variances in search performance across experiments and performed a two-way ANOVA based on this normalized effect, with global configuration (present vs. absent) and local motion (present vs. absent) as between-subject variables. The analysis revealed a significant main effect of local motion (Fig. 3a), $F(1, 76) = 5.223$, $p = .025$, $\eta_p^2 = 0.064$, but no significant main effect of global configuration cues, $F(1, 76) = 0.207$, $p = .651$, $\eta_p^2 = 0.003$, or interaction between local and global cues, $F(1, 76) = 0.112$, $p = .738$, $\eta_p^2 = 0.001$, suggesting that the audiovisual congruency effect mainly relies on the local motion cues independent of the global configuration.

We further performed ANOVA analyses on the normalized benefit effect of congruent sounds and the normalized cost effect of incongruent sounds relative to the baseline, in a way similar to that for the overall congruency effect (see Methods). The results for congruent sounds (Fig. 3b) were similar to that for the congruency effects, with a significant main effect of Local motion, $F(1, 76) = 12.149$, $p < .001$, $\eta_p^2 = 0.138$, but no significant main effect of Global or Local × Global interaction ($ps > .60$). For the effects of the incongruent sound, neither the main effects of local and global BM cues nor their interaction was significant (Fig. 3c, $ps > .46$). Altogether, these results suggest that local motion rather than global configuration cues are crucial to the crossmodal search facilitation induced by congruent sounds.

Discussion

Theories and knowledge about visual BM perception have developed for nearly half a century (Blake & Shiffrar, 2007; Hirai & Senju, 2020; Johansson, 1973). Whereas research on the audiovisual processing of BM information has emerged

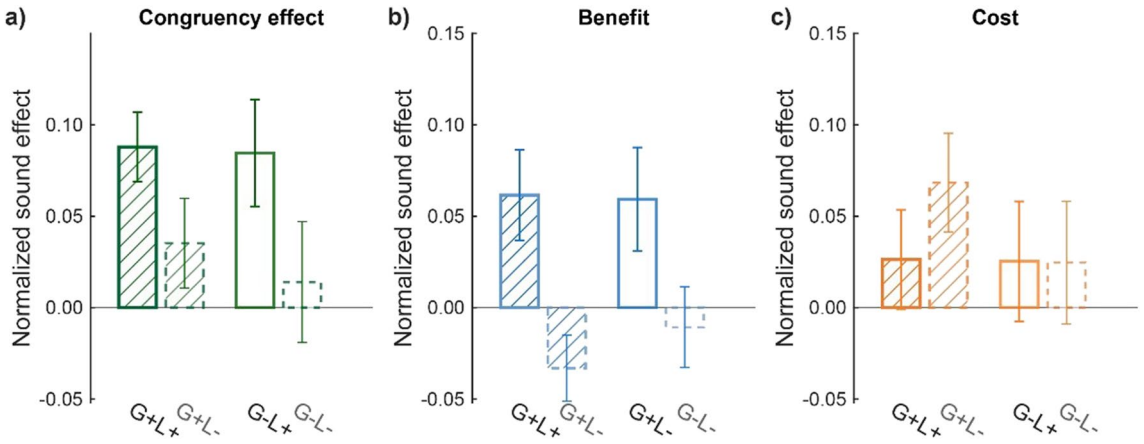

Fig. 3 The normalized sound effects across experiments. **a)** The overall audiovisual congruency effect. **b)** The benefit effect of congruent sounds relative to the baseline. **c)** The cost effect of incongruent sounds relative to the baseline. G = Global configuration; L = Local motion; + : present; − : absent. Error bars represent ±1 standard error of means. (Color figure online)

more recently, and the underlying mechanisms remain to be clarified (Arrighi et al., 2009; Brooks et al., 2007; Saygin et al., 2008; Schouten et al., 2011; Thomas & Shiffrar, 2010, 2013; van der Zwan et al., 2009). Here we found that the visual search for continuous dynamic BM signals in a crowded visual display was facilitated by temporally congruent yet spatially noninformative sounds.

Remarkably, these findings may reflect a domain-specific mechanism different from that underlying the crossmodal attentional facilitation of nonbiological artificial stimuli. The visual search facilitation driven by synchronized nonspatial auditory cues, known as the Pip and Pop effect, has been widely observed in studies using simple visual patterns (Chamberland et al., 2016; Gao et al., 2021; Staufenbiel et al., 2011; Van der Burg et al., 2008, 2010; Zou et al., 2012). Such crossmodal facilitation usually relies on the synchronization of auditory events with abrupt changes or onsets of the visual stimuli (e.g., Van der Burg et al., 2010). A plausible explanation is that integrating multisensory cues from meaningless information within a transient window may avoid spurious interactions of unrelated signals from a wider temporal range (Fujisaki et al., 2004; Van der Burg et al., 2008). These findings from nonbiological stimuli may partially account for the absence of crossmodal facilitation in the *NoAC* conditions, especially in Experiment 4, where the visual stimuli lacked biological features but still conveyed complex motion signals unfolding periodically without abrupt onsets or changes. However, critical biological features may modulate the time window of audiovisual integration and lead to crossmodal facilitation in normal BM conditions. Specifically, BM perception relies on the spatiotemporal summation and continuous tracking of rhythmic kinematic signals over time rather than at any single time point (Giese & Poggio, 2003; Neri et al., 1998; Shen et al., 2023). In addition, the multisensory processing of BM and non-BM has different properties in the temporal dimension (Arrighi et al., 2006; Saygin et al., 2008), and the audiovisual integration of BM depends not only upon timing but also on meaningful associations, like those between natural footstep sounds and motion (Arrighi et al., 2009; Thomas & Shiffrar, 2013). These findings, together with our observations, indicate that a mechanism based on both temporal correspondence and biological associations may underlie the audiovisual integration of BM information, thereby specifically facilitating the search for BM signals.

We further isolated the contributions of two fundamental BM cues (i.e., the global configuration and the local motion) to the crossmodal search facilitation effect in a series of experiments. We found that the facilitation effect was evident no matter whether the global configuration cue was present (Experiment 1) or absent (Experiment 3). By contrast, when we destroyed the local motion cue in limb movements, such effects disappeared regardless of whether the global configuration was intact (Experiment 2) or not (Experiment 4). These findings provide compelling evidence that local motion rather than global configuration is both necessary and sufficient for triggering the audiovisual binding of BM signals in a dynamic context and is responsible for the enhanced visual search performance. Moreover, they enrich our understanding of the dissociable roles of the global configuration and local motion cues in BM perception (Hirai & Senju, 2020). Previous research suggests that the dissociation may stem from the anatomically and functionally distinct neural responses to the form and motion cues in BM stimuli (Jastorff & Orban, 2009; Vaina et al., 2001; Vangeneugden et al., 2014), as well as the different genetic origins for local and global BM processing (Wang et al., 2018). In particular, researchers speculate

that a phylogenetically evolved neural mechanism tuned to the local BM cue may act as a life motion detector to help direct attention to these signals in humans and animals (Lemaire & Vallortigara, 2022; Troje & Chang, 2023). The current findings suggest that the function of this life-detection system may extend to the multisensory perception of BM information. More specifically, the local BM cue may interact with the temporally congruent auditory cues to enhance the salience of life motion signals and facilitate attentional selection in a crossmodal situation.

Functional neuroimaging studies have demonstrated that both auditory and visual BM stimuli could selectively activate the posterior superior temporal sulcus (pSTS) (Bidet-Caulet et al., 2005; Grossman & Blake, 2001). The pSTS also plays a crucial role in combing auditory and visual BM signals (Meyer et al., 2011) and is sensitive to audiovisual temporal correspondence (Noesselt et al., 2007). Given its significance to unimodal and multimodal BM processing, the pSTS is a possible neural substrate for the BM-specific audiovisual facilitation effect observed in the current study. Moreover, researchers propose that the processing of local BM cues involves subcortical regions, such as superior colliculus, pulvinar, and ventral lateral nucleus (Chang et al., 2018; Troje & Westhoff, 2006). Whether subcortical neural substrates mediate the predominant role of local BM cues in the audiovisual integration of BM signals is a question that warrants further research. Besides, BM stimuli convey rhythmic structures (i.e., gait cycles) akin to that embedded in continuous human speech. Recent electroencephalogram (EEG) and magnetoencephalography (MEG) studies have shown cortical tracking of rhythmic linguistic structures associated with language perception and comprehension (Ding et al., 2016, 2017; Keitel et al., 2018). Future studies could verify whether similar neural mechanisms underlie the visual, auditory, and multisensory processing of BM information, as well as the crossmodal attentional facilitation of BM signals.

Author contributions L.S. and Y.W. contributed to the design of the study. L.S. collected and analyzed the data under the supervision of Y.W. and Y.J. All authors contributed to the writing of the manuscript.

Funding This research was supported by grants from the Ministry of Science and Technology of China (STI2030-Major Project 2021ZD0203800), the National Natural Science Foundation of China (32171059, 31830037), the Strategic Priority Research Program (XDB32010300), the Interdisciplinary Innovation Team (JCTD-2021-06), the Youth Innovation Promotion Association of the Chinese Academy of Sciences, the Scientific Foundation of Institute of Psychology, Chinese Academy of Sciences (No. E2CX4325CX), and Fundamental Research Funds for the Central Universities.

Data availability All data generated during the current study are made available online (http://ir.psych.ac.cn/handle/311026/42953), and materials used during the study are available from the corresponding author upon request.

Code availability Not applicable.

Declarations

Competing interests The authors declared no competing interest relevant to the content of this article.

Ethics approval The protocols of the research were approved by the institutional review board of the Institute of Psychology, Chinese Academy of Sciences.

Consent to participate Participants provided written, informed consent before the experiments.

Consent for publication This manuscript has not been published and is not under review for publication elsewhere. All authors have approved the manuscript and this submission.

References

Arrighi, R., Alais, D., & Burr, D. (2006). Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. *Journal of Vision, 6*(3), 260–268.

Arrighi, R., Marini, F., & Burr, D. (2009). Meaningful auditory information enhances perception of visual biological motion. *Journal of Vision*, 9(4):25, 1–7. https://doi.org/10.1167/9.4.25

Bardi, L., Regolin, L., & Simion, F. (2011). Biological motion preference in humans at birth: Role of dynamic and configural properties. *Developmental Science, 14*(2), 353–359.

Bardi, L., Regolin, L., & Simion, F. (2014). The first time ever I saw your feet: Inversion effect in newborns' sensitivity to biological motion. *Developmental Psychology, 50*(4). https://doi.org/10.1037/a0034678

Beintema, J. A., & Lappe, M. (2002). Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences, 99*(8), 5661–5663.

Bertenthal, B. I., & Pinto, J. (1994). Global processing of biological motions. *Psychological Science, 5*(4), 221–225.

Bidet-Caulet, A., Voisin, J., Bertrand, O., & Fonlupt, P. (2005). Listening to a walking human activates the temporal biological motion area. *NeuroImage, 28*(1), 132–139.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology, 58*(1), 47–73.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436.

Brooks, A., van der Zwan, R., Billard, A., Petreska, B., Clarke, S., & Blanke, O. (2007). Auditory motion affects visual biological motion processing. *Neuropsychologia, 45*(3), 523–530.

Chamberland, C., Hodgetts, H. M., Vallières, B. R., Vachon, F., & Tremblay, S. (2016). Pip and pop: When auditory alarms facilitate visual change detection in dynamic settings. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*(1), 284–288.

Chang, D. H. F., Ban, H., Ikegaya, Y., Fujita, I., & Troje, N. F. (2018). Cortical and subcortical responses to biological motion. *NeuroImage, 174*, 87–96.

Chang, D. H. F., & Troje, N. F. (2009). Acceleration carries the local inversion effect in biological motion perception. *Journal of Vision, 9*(1), 19, 1–19, 17.

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in Human Neuroscience, 11*, 481.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19*(1), 158–164.

Driver, J., & Spence, C. (1998). Attention and the crossmodal construction of space. *Trends in Cognitive Sciences, 2*(7), 254–262.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7*(7), 773–778.

Gao, M., Chang, R., Wang, A., Zhang, M., Cheng, Z., Li, Q., & Tang, X. (2021). Which can explain the pip-and-pop effect during a visual search: Multisensory integration or the oddball effect? *Journal of Experimental Psychology: Human Perception and Performance, 47*(5), 689–703.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4*(3), 179–192.

Grossman, E. D., & Blake, R. (2001). Brain activity evoked by inverted and imagined biological motion. *Vision Research, 41*(10/11), 1475–1482.

Hirai, M., & Senju, A. (2020). The two-process theory of biological motion processing. *Neuroscience & Biobehavioral Reviews, 111*, 114–124.

Jastorff, J., & Orban, G. A. (2009). Human functional magnetic resonance imaging reveals separation and integration of shape and motion cues in biological motion processing. *Journal of Neuroscience, 29*(22), 7315–7329.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*(2), 201–211.

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features, *PLOS Biology, 16*(3), e2004473.

Lemaire, B. S., & Vallortigara, G. (2022). Life is in motion (through a chick's eye). *Animal Cognition, 26*(1), 129–140.

Matusz, P. J., & Eimer, M. (2011). Multisensory enhancement of attentional capture in visual search. *Psychonomic Bulletin & Review, 18*(5), 904.

Mendonça, C., Santos, J. A., & López-Moliner, J. (2011). The benefit of multisensory integration with biological motion signals. *Experimental Brain Research, 213*(2/3), 185–192.

Meyer, G. F., Greenlee, M., & Wuerger, S. (2011). Interactions between auditory and visual semantic stimulus classes: Evidence for common processing networks for speech and body actions. *Journal of Cognitive Neuroscience, 23*(9), 2291–2308.

Meyer, G. F., Harrison, N. R., & Wuerger, S. M. (2013). The time course of auditory–visual processing of speech and body actions: Evidence for the simultaneous activation of an extended neural network for semantic processing. *Neuropsychologia, 51*(9), 1716–1725.

Neri, P., Morrone, M., & Burr, D. (1998). Seeing biological motion. *Nature, 395*, 894–896.

Ngo, M. K., & Spence, C. (2012). Facilitating masked visual target identification with auditory oddball stimuli. *Experimental Brain Research, 221*(2), 129–136.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H.-J., & Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *Journal of Neuroscience, 27*(42), 11431–11441.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Santangelo, V., & Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1311–1321.

Saygin, A. P., Driver, J., & de Sa, V. R. (2008). In the footsteps of biological motion and multisensory perception: Judgments of audiovisual temporal relations are enhanced for upright walkers. *Psychological Science, 19*(5), 469–475.

Schouten, B., Troje, N. F., Vroomen, J., & Verfaillie, K. (2011). The effect of looming and receding sounds on the perceived in-depth orientation of depth-ambiguous biological motion figures. *PLOS ONE, 6*(2), e14725.

Shen, L., Lu, X., Yuan, X., Hu, R., Wang, Y., & Jiang, Y. (2023). Cortical encoding of rhythmic kinematic structures in biological motion. *NeuroImage, 268*, *119893*.

Simion, F., Regolin, L., & Bulf, H. (2008). A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences, 105*(2), 809–813.

Spence, C., Kingstone, A., Shore, D. I., & Gazzaniga, M. S. (2001). Representation of visuotactile space in the split brain. *Psychological Science, 12*(1), 90–93.

Staufenbiel, S. M., van der Lubbe, R. H. J., & Talsma, D. (2011). Spatially uninformative sounds increase sensitivity for visual motion change. *Experimental Brain Research, 213*(4), 457–464.

ten Oever, S., Romei, V., van Atteveldt, N., Soto-Faraco, S., Murray, M. M., & Matusz, P. J. (2016). The COGs (context, object, and goals) in multisensory processing. *Experimental Brain Research, 234*(5), 1307–1323.

Thomas, J. P., & Shiffrar, M. (2010). I can see you better if I can hear you coming: Action-consistent sounds facilitate the visual detection of human gait. *Journal of Vision, 10*(12), 14, 1–14, 11.

Thomas, J. P., & Shiffrar, M. (2013). Meaningful sounds enhance visual sensitivity to human gait regardless of synchrony. *Journal of Vision*, 14, 8, 1–*13*, 8, 13. https://doi.org/10.1167/13.14.8

Troje, N. F., & Chang, D. H. F. (2023). Life detection from biological motion. *Current Directions in Psychological Science, 32*(1), 26–32.

Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: Evidence for a "life detector"? *Current Biology, 16*(8), 821–824.

Turoman, N., Tivadar, R. I., Retsa, C., Maillard, A. M., Scerif, G., & Matusz, P. J. (2021). The development of attentional control mechanisms in multisensory environments. *Developmental Cognitive Neuroscience, 48*, *100930*.

Vaina, L. M., Solomon, J., Chowdhury, S., Sinha, P., & Belliveau, J. W. (2001). Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences, 98*(20), 11656–11661.

Vallortigara, G., & Regolin, L. (2006). Gravity bias in the interpretation of biological motion by inexperienced chicks. *Current Biology, 16*(8), R279–R280.

Vallortigara, G., Regolin, L., & Marconato, F. (2005). Visually inexperienced chicks exhibit spontaneous preference for biological motion patterns. *PLOS Biology, 3*(7), e208.

Van der Burg, E., Cass, J., Olivers, C. N. L., Theeuwes, J., & Alais, D. (2010). Efficient visual search from synchronized auditory signals requires transient audiovisual events, *PLOS ONE, 5*(5), e10664.

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1053–1065.

van der Zwan, R., MacHatch, C., Kozlowski, D., Troje, N. F., Blanke, O., & Anna, B. (2009). Gender bending: Auditory cues affect visual judgements of gender in biological motion displays. *Experimental Brain Research, 198*(2/3), 373–382.

Vangeneugden, J., Peelen, M. V., Tadin, D., & Battelli, L. (2014). Distinct neural mechanisms for body form and body motion discriminations. *Journal of Neuroscience, 34*(2), 574–585.

Vanrie, J., & Verfaillie, K. (2004). Perception of biological motion: A stimulus set of human point-light actions. *Behavior Research Methods, Instruments, & Computers, 36*(4), 625–629.

Wang, L., Yang, X., Shi, J., & Jiang, Y. (2014). The feet have it: Local biological motion cues trigger reflexive attentional orienting in the brain. *NeuroImage, 84*, 217–224.

Wang, L., Zhang, K., He, S., & Jiang, Y. (2010). Searching for life motion signals: Visual search asymmetry in local but not global biological-motion processing. *Psychological Science, 21*(8), 1083–1089.

Wang, Y., Wang, L., Xu, Q., Liu, D., Chen, L., Troje, N. F., He, S., & Jiang, Y. (2018). Heritable aspects of biological motion perception and its covariation with autistic traits. *Proceedings of the National Academy of Sciences, 115*(8), 1937–1942.

Wang, Y., Zhang, X., Wang, C., Huang, W., Xu, Q., Liu, D., Zhou, W., Chen, S., & Jiang, Y. (2022). Modulation of biological motion perception in humans by gravity. *Nature Communications, 13*(1), 2765.

Wuerger, S. M., Parkes, L., Lewis, P. A., Crocker-Buque, A., Rutschmann, R., & Meyer, G. F. (2012). Premotor cortex is sensitive to auditory–visual congruence for biological motion. *Journal of Cognitive Neuroscience, 24*(3), 575–587.

Zou, H., Muller, H. J., & Shi, Z. (2012). Non-spatial sounds regulate eye movements and enhance visual search. *Journal of Vision, 12*(5), 2.