Received: 15 November 2021 | Revised: 12 January 2022 | Accepted: 17 January 2022

DOI: 10.1002/hbm.25797

RESEARCH ARTICLE

WILEY

The interplay between audiovisual temporal synchrony and semantic congruency in the cross-modal boost of the visual target discrimination during the attentional blink

Song Zhao[1,2] | Chongzhi Wang[1] | Chengzhi Feng[1] | Yijun Wang[3] | Wenfeng Feng[1,4]

[1]Department of Psychology, School of Education, Soochow University, Suzhou, China

[2]Department of English, School of Foreign Languages, Soochow University, Suzhou, China

[3]Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China

[4]Research Center for Psychology and Behavioral Sciences, Soochow University, Suzhou, China

Correspondence
Wenfeng Feng, Department of Psychology, School of Education, Soochow University, Suzhou, Jiangsu 215123, China.
Email: fengwfly@gmail.com

Yijun Wang, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China.
Email: wangyj@semi.ac.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2021ZD0202600; National Natural Science Foundation of China, Grant/Award Numbers: 31771200, 32171048; Humanities and Social Science Project of Ministry of Education of China, Grant/Award Number: 17YJA880019; Strategic Priority Research Program of Chinese Academy of Science, Grant/Award Number: XDB32040200

Abstract

The visual attentional blink can be substantially reduced by delivering a task-irrelevant sound synchronously with the second visual target (T2), and this effect is further modulated by the semantic congruency between the sound and T2. However, whether the cross-modal benefit originates from audiovisual interactions or sound-induced alertness remains controversial, and whether the semantic congruency effect is contingent on audiovisual temporal synchrony needs further investigation. The current study investigated these questions by recording event-related potentials (ERPs) in a visual attentional blink task wherein a sound could either synchronize with T2, precede T2 by 200 ms, be delayed by 100 ms, or be absent, and could be either semantically congruent or incongruent with T2 when delivered. The behavioral data showed that both the cross-modal boost of T2 discrimination and the further semantic modulation were the largest when the sound synchronized with T2. In parallel, the ERP data yielded that both the early occipital cross-modal P195 component (192–228 ms after T2 onset) and late parietal cross-modal N440 component (424–448 ms) were prominent only when the sound synchronized with T2, with the former being elicited solely when the sound was further semantically congruent whereas the latter occurring only when that sound was incongruent. These findings demonstrate not only that the cross-modal boost of T2 discrimination during the attentional blink stems from early audiovisual interactions and the semantic congruency effect depends on audiovisual temporal synchrony, but also that the semantic modulation can unfold at the early stage of visual discrimination processing.

KEYWORDS
attentional blink, audiovisual, cross-modal interaction, ERPs, semantic congruency, temporal synchrony

1 | INTRODUCTION

Our ability to precisely extract important visual information from a rapidly changing environment is rather limited. One of the most striking examples illustrating this temporal limitation of attention is the attentional blink phenomenon (Raymond, Shapiro, & Arnell, 1992)—if two successive visual targets are embedded in a rapid serial visual presentation (RSVP) stream, observers often fail to

discriminate the second target (T2) when it appears 200–300 ms after the first target (T1). However, recent studies have consistently shown that a task-irrelevant, meaningless auditory stimulus delivered synchronously with T2 could substantially boost T2 discrimination during the attentional blink interval (Kranczioch & Thorne, 2013, 2015; Olivers & Van der Burg, 2008), indicating information from the auditory modality can help to overcome the temporal limitation of visual attention to some extent. Moreover, using event-related potential (ERP) recordings and line drawings of common objects (e.g., dogs) paired with semantically congruent or incongruent sounds (e.g., barks of dogs or beeps of cars), a more recent study explored the neural substrates of this cross-modal facilitation and the possible audiovisual semantic congruency effect on it (Zhao, Feng, Huang, Wang, & Feng, 2021). It was found that the semantically congruent sounds induced a larger T2 accuracy enhancement than the incongruent sounds, and the behavioral improvements induced by both the congruent and incongruent sounds were associated with a visual N1-like early ERP component (∼200 ms after T2 onset) over the occipital region, whereas the lower T2 accuracy for the incongruent sounds was correlated with an N400-like late ERP component (∼400 ms) over the parietal scalp (Zhao et al., 2021). These findings demonstrate that the cross-modal boost of T2 discrimination during the attentional blink has an early neural processing locus while the modulation of audiovisual semantic congruency occurs at a relatively late stage of processing.

Nevertheless, there is still an unresolved debate regarding whether the cross-modal boost of T2 discrimination during the attentional blink originates from genuine audiovisual cross-modal interactions or is merely a manifestation of sound-induced, modality-nonspecific alerting effect (Kranczioch & Thorne, 2013, 2015; Olivers & Van der Burg, 2008). The pioneering investigation showed that the cross-modal boost effect (i.e., the sound-induced T2 accuracy enhancement relative to the no-sound condition) occurred when the sound synchronized with T2 but not when delivered 250 ms before T2 onset (Olivers & Van der Burg, 2008). These findings were considered as in favor of the cross-modal interaction hypothesis, because the efficiency of cross-modal interaction generally decreases as the audiovisual temporal asynchrony increases (Meredith, Nemitz, & Stein, 1987; Spence, Shore, & Klein, 2001; Stone et al., 2001; van Wassenhove, Grant, & Poeppel, 2007; Zampini, Guest, Shore, & Spence, 2005), whereas the alerting effect is typically maximal when the alerting stimulus is presented 100–300 ms prior to a target stimulus (Bertelson, 1967; Los & Van den Heuvel, 2001; Niemi & Näätänen, 1981; Posner & Boies, 1971). In contrast, subsequent researchers found that although the cross-modal boost effect was not greater when the sound preceded T2 by 250 ms than when it synchronized with T2, the cross-modal boost effect in the preceding-sound condition was indeed significant (relative to the no-sound condition) rather than absent (Kranczioch & Thorne, 2013, 2015). Their results seem to suggest that the alerting hypothesis could still account for the cross-modal boost of T2 discrimination during the attentional blink to some degree.

The controversial findings regarding the psychological mechanisms of the cross-modal boost effect may be attributed to different experimental designs used in previous studies. First, since the probability for the synchronous-sound condition was four times higher than the preceding-sound condition in Olivers and Van der Burg's (2008) study (see their Exp. 4), it cannot rule out the possibility that low probability for the preceding-sound condition might have attenuated the potential sound-induced alerting effect. Second, given that the preceding-sound and synchronous-sound conditions were presented in *separate* blocks (sessions) in studies of Kranczioch and Thorne (2013, 2015), the sound would always precede T2 when delivered in a preceding-sound block. Previous studies have shown that the human brain can make rapid recalibration to repeatedly presented, temporally asynchronous audiovisual stimuli and increase the probability of audiovisual integration (Bhat, Miller, Pitt, & Shahin, 2015; Fujisaki, Shimojo, Kashino, & Nishida, 2004; Simon, Noel, & Wallace, 2017; Van der Burg, Alais, & Cass, 2013). Based on these prior findings, it is possible that audiovisual temporal recalibration may have occurred in response to T2 and the 100% preceding sound (when delivered) in the preceding-sound blocks of Kranczioch and Thorne (2013, 2015), leading to audiovisual integration. Thus, the origin of the cross-modal boost of T2 discrimination during the attentional blink is still equivocal and needs to be determined with improvements in the experimental paradigm.

Furthermore, it also deserves further investigation concerning whether the effect of higher-order audiovisual semantic congruency on the visual attentional blink depends on audiovisual temporal synchrony. Although a recent study has found that a semantically congruent sound led to higher T2 discrimination accuracy than a semantically incongruent sound even when these sounds preceded T2 by 210 ms (Adam & Noppeney, 2014), it should be noted again that similar to Kranczioch and Thorne's (2013, 2015) studies, the preceding-sound and synchronous-sound conditions were also presented in *separate* blocks in the study of Adam and Noppeney (2014). Meanwhile, there was no sound-absent condition in the study. Accordingly, in their preceding-sound blocks, although the sound could be either semantically congruent or incongruent with T2, it preceded T2 on *each* trial. As illustrated above, this kind of block-design may have triggered the audiovisual temporal recalibration in response to T2 and the 100% preceding sound (Bhat et al., 2015; Fujisaki et al., 2004; Simon et al., 2017; Van der Burg et al., 2013), thereby weakening the potential effect of audiovisual temporal synchrony. Hence, the existing evidence seems insufficient to answer whether the audiovisual semantic congruency effect on the visual attentional blink is genuinely independent of audiovisual temporal synchrony.

The current study investigated the questions mentioned above in an extended version of the visual attentional blink paradigm recently described by Zhao et al. (2021), under which a task-irrelevant but natural sound could synchronize with T2, precede T2 by 200 ms, be delayed relative to T2 by 100 ms or be absent, and could be either semantically congruent or incongruent with T2 when delivered (e.g., a bark of a dog with a drawing of a dog, or a beep of a car with a drawing of a dog). Notably, the temporal position of the sound was manipulated *within* each block and different sound temporal positions were kept *equally* probable in the present study. Accordingly, if the cross-

modal boost of T2 discrimination during the attentional blink originates from audiovisual cross-modal interactions and the audiovisual semantic congruency effect is contingent on audiovisual temporal synchrony, the present study should predict: (a) the cross-modal boost effect would be the largest when the sound synchronized with T2 and would be prominent but with a decreased magnitude when the sound was delayed by 100 ms, because a 100-ms audiovisual asynchrony is still within the temporal window of integration (Donohue, Roberts, Grent-'t-Jong, & Woldorff, 2011; Meredith et al., 1987; Spence et al., 2001; van Wassenhove et al., 2007; Zampini et al., 2005); (b) the audiovisual semantic congruency effect on the cross-modal boost would be the greatest when the congruent and incongruent sounds synchronized with T2. Importantly, in order to provide electrophysiological evidence for the present behavioral findings, high time-resolution ERP data were recorded concurrently with the behavioral task here, and ERP components that have been shown to underlie the cross-modal boost effect and the audiovisual semantic congruency effect (Zhao et al., 2021) were analyzed as functions of audiovisual temporal synchrony and semantic congruency.

2 | METHODS

2.1 | Participants

The pilot behavioral experiment ($n = 20$) showed that the 3 (sound temporal position: precede, synchronize, delay) × 2 (semantic congruency: congruent, incongruent) repeated-measures ANOVA on the cross-modal boost effect had a significant two-way interaction, with its effect size η^2_p being equal to 0.154. Thus, given an alpha level of .05, at least 28 subjects were required to achieve a power of 0.8 when focusing on the two-way interaction, which was computed using MorePower 6.0.4 (Campbell & Thompson, 2012). To obtain reliable results that were comparable to those reported in the study of Zhao et al. (2021) ($n = 34$) wherein the same stimuli and a similar experimental design were used, the current electroencephalogram (EEG) experiment recruited 42 healthy subjects (30 female and 12 males; age range of 18–28 years, mean age of 20.9 years; all right-handed). All subjects verbally reported normal or corrected-to-normal vision as well as normal audition, and could easily recognize the object category (dog, car, drum) of all visual and auditory stimuli used in the experiment, although more standardized tests for assessing visual and auditory functions should be adopted in future studies. They were naive as to the hypothesis of the experiment. In accordance with the Declaration of Helsinki, written informed consent as approved by the Human Research Protections Program of Soochow University was obtained from all subjects before their participation.

2.2 | Apparatus, stimuli and design

The experiment was performed in a dark and sound-attenuated room. Stimulus presentation was scripted using "Presentation" software (version 18.0, NeuroBehavioral Systems, Inc.). Visual stimuli were presented on a 27-in. LCD monitor (ASUS PG279Q, resolution 1,920 × 1,080, refresh rate 120 Hz) on which the background color was set to gray. Auditory stimuli were delivered by a pair of loudspeakers (HiVi X3) positioned on the left and right sides of the monitor symmetrically, so that a single sound presented by the two speakers simultaneously would be perceived as coming from the center of the monitor (Bertelson & Aschersleben, 1998). Subjects sat in front of the monitor with a viewing distance of approximately 80 cm, and were required to maintain their eyes fixated on a red cross (0.3° × 0.3°), which was displayed at the center of the screen throughout each RSVP stream. The visual stimuli consisted of 48 black-and-white line drawings (each 5.6° × 4.5°), including 30 unique drawings of houses used as distractors, nine unique drawings (three clothes, three cups, and three flowers) used as the first target (T1), and the remaining nine unique drawings (three dogs, three cars and three drums) used as the second target (T2). The line drawings for T1 and T2 were from two non-overlapping sets in order to avoid priming (Koelewijn, Van der Burg, Bronkhorst, & Theeuwes, 2008) or repetition blindness effects (Kanwisher, 1987). The auditory stimuli were comprised of nine unique natural sounds (three barks of dogs, three beeps of cars, and three beats of drums; all stereo) that were 200 ms in duration (with 20 ms rise and fall ramps) and approximately 75 dB in loudness at subjects' ears when delivered. These line drawings and natural sounds were all taken from the study of Zhao et al. (2021) wherein a similar basic experimental design was used.

The whole experiment consisted of 27 blocks of 60 trials each, resulting in a total of 1,620 trials, which were performed by each participant. Each trial began with the presentation of the red fixation for a fixed period of 1,000 ms, immediately followed by an RSVP stream presented at the center of the screen (Figure 1a). Each RSVP stream was comprised of 17 distinct line drawings, including T1, T2 and 15 distractors. The distractors for each trial were sampled randomly (without repetition) from the aforementioned 30 drawings of houses. Each drawing in the RSVP stream was presented immediately after the offset of the preceding drawing, and the duration of each drawing was 100 ms [i.e., the drawing-to-drawing stimulus onset asynchrony (SOA) was 100 ms]. T1 could be one of the nine drawings (three clothes, three cups and three flowers; Figure 1a, right) with equal probability, and was presented randomly from the third to the fifth position in the RSVP stream. T2 could be one of the remaining nine drawings (three dogs, three cars and three drums; Figure 1a, bottom right) equiprobably, with its presented position in the RSVP stream varying with different experimental conditions, listed below.

On 7/12 of all trials, T2 was presented three positions after T1 (i.e., at lag 3, T1-to-T2 SOA of 300 ms). Specifically, on 6/7 of these lag 3 trials (i.e., 6/12 of all trials), a task-irrelevant natural sound [i.e., one of the nine unique sounds (three barks of dogs, three beeps of cars, and three beats of drums) with equal probability] could either synchronize with T2 (labeled as *sync*, 2/12 of all trials), or precede T2 by 200 ms (labeled as *prec*, 2/12 of all trials), or be delayed relative to T2 by 100 ms (labeled as *delay*, 2/12 of all trials; see Figure 1a).

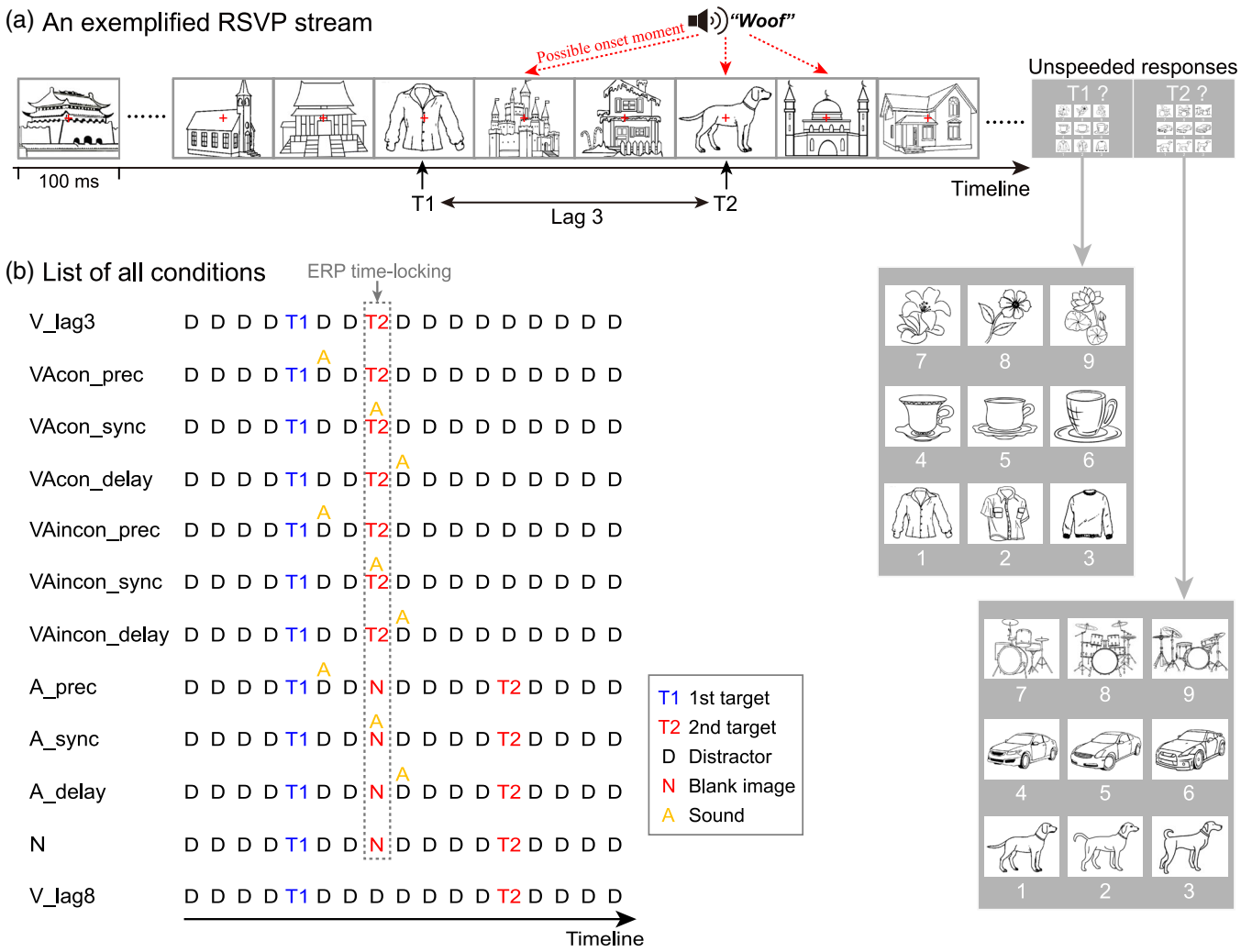

FIGURE 1 (a) Schematic illustration of the RSVP stream exemplified for a lag 3 trial on which T2 was presented 3 positions after T1, together with a semantically congruent but task-irrelevant sound that could synchronize with T2, precede T2 by 200 ms, or be delayed relative to T2 by 100 ms. The task for participants was to discriminate sequentially the exact identities of T1 and T2 without time limit after each RSVP stream, while ignoring all sounds. Note that the optional drawings for T1 and T2 and their corresponding button numbers were presented to the participants when they responded, as shown on the right side. (b) Detailed illustration of the RSVP sequences in all 12 experimental conditions (see section 2.2). The probability for each condition was 1/12, and ERP waveforms in all but V_lag8 condition were time-locked to the onset of visual stimulus at lag 3 position for further analysis

Meanwhile, the sound, regardless of its onset moment, could be either semantically congruent with T2 [labeled as *VAcon* (audiovisual congruent), 3/12 of all trials; e.g., a bark of a dog with a drawing of a dog] or semantically incongruent with T2 [labeled as *VAincon* (audiovisual incongruent), 3/12 of all trials; e.g., a beep of a car with a drawing of a dog]. On the remaining 1/7 of the lag 3 trials (i.e., 1/12 of all trials), T2 was presented without any nearing sound [labeled as *V* (visual-only)]. Accordingly, there were seven resulting experimental conditions for the lag 3 trials, namely, *V (V_lag3)*, *VAcon_prec*, *VAcon_sync*, *VAcon_delay*, *VAincon_prec*, *VAincon_sync* and *VAincon_delay* (see Figure 1b for full comprehension).

On another 4/12 of all trials, a white rectangle with the same size of line drawings (i.e., a blank drawing) was presented at lag 3 position. On 3/4 of these trials (i.e., 3/12 of all trials), a random one of the nine natural sounds could either synchronize with the blank drawing at lag 3 [labeled as *A_sync* (auditory-only_synchronize), 1/12 of all trials], or precede the blank drawing by 200 ms [labeled as *A_prec* (auditory-only_precede), 1/12 of all trials], or be delayed relative to the blank drawing by 100 ms [labeled as *A_delay* (auditory-only_delay), 1/12 of all trials; see Figure 1b]. On the remaining 1/4 of these trials (i.e., 1/12 of all trials), no sound was delivered near the blank drawing [labeled as *N* (no stimulus); see Figure 1b]. These four conditions (i.e., A_prec, A_sync, A_delay and *N*) were included in the experiment in order to isolate audiovisual cross-modal ERP components on the lag 3 trials when the sound preceded T2, synchronized with T2, and was delayed relative to T2, respectively (see section 2.4). Meanwhile, in order to generate a considerable number of lag 8 trials in terms of behavioral task, T2 was then presented at lag 8 position in these conditions. Importantly, since T2 appeared 500 ms after the blank drawing onset in these conditions, it could be ensured that subjects would engage in the task without response-related brain activity during the subsequent 500 ms interval after the blank drawing onset, which was the

epoch of interest in ERP analysis. In addition, similar to the aforementioned VA conditions, in A_prec, A_sync and A_delay conditions it was equally probable whether the sound nearing the blank drawing at lag 3 was semantically congruent or incongruent with the subsequent T2 at lag 8.

On the remaining 1/12 of all trials, T2 was presented at lag 8, a distractor instead of a blank drawing was presented at lag 3, and no sound was delivered. Thus, this type of trials was the standard visual-only lag 8 trials (labeled as *V_lag8*; see Figure 1b), which was included in order to check whether the basic visual attentional blink effect (i.e., much lower T2 discrimination accuracy in V_lag3 than V_lag8 condition) was successfully induced in the present study.

Accordingly, from the perspective of the behavioral task, there were seven types of lag 3 trials (V_lag3, VAcon_prec, VAcon_sync, VAcon_delay, VAincon_prec, VAincon_sync and VAincon_delay) and five types of lag 8 trials (A_prec, A_sync, A_delay, N and V_lag8). All these types of trials were presented in a pseudo-randomized order with equal probability [i.e., each 1/12 (135 trials)]. It is noteworthy that the experiment did not introduce the corresponding VA, A and N conditions for the standard V_lag8 condition because of the followings: (a) the focus of the present study was the effect of sound on T2 discrimination *during* the attentional blink (i.e., at lag 3) rather than *outside* the attentional blink (i.e., at lag 8); (b) the present design allowed the collection of as many lag 3 trials as possible without increasing the experiment duration, which could minimize the fatigue effect (cf., Maier & Rahman, 2018); (c) the study of Zhao et al. (2021) has shown that presenting a sound simultaneously with T2 at lag 8 had no effect on T2 discrimination.

The task for participants was to discriminate sequentially, after each RSVP stream, the exact identities of both T1 and T2 as accurately as possible in an unspeeded manner by pressing buttons on a keyboard's number pad (1–9 for T1, 1–9 for T2; see Figure 1a, right side) with the right hand, while ignoring all sounds if delivered. Note that the optional drawings for T1 and T2 (each 5.6° × 4.5°) and their corresponding button numbers were presented to the participants when they made their responses. Only if the exact identity of T1 or T2 was correctly discriminated would this discrimination be coded as a correct response. Importantly, in order to prevent subjects from responding based on sounds, subjects were informed that the identity of the sound was uninformative of the exact identity of T2, even in VAcon conditions (e.g., any single bark of dog was likely to be presented near any single drawing of dog). It was checked that the participants believed that the sounds provided no information on T2 identity (see Section B of Data S1, Supporting Information). The button-press for T2 then triggered the next trial. Participants were encouraged to have a rest between blocks in order to relieve fatigue.

2.3 | Electrophysiological recording and preprocessing

The electroencephalogram (EEG) was recorded continuously when subjects performed the behavioral task, using a SynAmps2 amplifier (NeuroScan, Inc.) and a custom-built 64-electrode elastic cap. The electrodes on the cap [FPz, FP1, FP2, AF3, AF4, Fz, F1, F2, F3, F4, F5, F6, F7, F8, FCz, FC1, FC2, FC3, FC4, FC5, FC6, Cz, C1, C2, C3, C4, C5, C6, T7, T8, CPz, CP1, CP2, CP3, CP4, CP5, CP6, TP7, TP8, Pz, P1, P2, P3, P4, P5, P6, P7, P8, POz, PO3, PO4, PO5, PO6, PO7, PO8, Oz, O1, O2, Iz, I3, I4, I5 (P9), I6 (P10) and M2 (right mastoid)] were positioned according to a modified 10–10 system montage (McDonald, Teder-Sälejärvi, Di Russo, & Hillyard, 2003). Two additional electrodes, AFz and M1 (left mastoid), served as the ground and reference electrodes during data acquisition, respectively. Horizontal eye movements were detected by a pair of bipolar electrodes positioned at the left and right outer canthi (horizontal electrooculogram, HEOG). Vertical eye movements and blinks were detected by another pair of bipolar electrodes placed above and below the left eye (vertical electrooculogram, VEOG). The impedances of all electrodes were kept below 5 kΩ. The online EEG and EOG signals were filtered by a band-pass filter of 0.05–100 Hz and digitized at a sampling rate of 1,000 Hz. The EEG recording was carried out using "Scan" software (version 4.5, NeuroScan, Inc.).

In offline preprocessing, the continuous EEG signals were firstly down-sampled to 500 Hz, and then low-pass filtered (half-amplitude cutoff = 33.75 Hz, transition band width = 7.5 Hz) using a zero-phase shifted (two-pass forward and reverse), Hamming-windowed sinc FIR filter to attenuate high-frequency noise triggered by muscle activities or external electrical sources. The filtered EEG data were re-referenced to the average of the left and right mastoid (M1 and M2) electrodes. The re-referenced EEG signals in all conditions except the V_lag8 condition were then segmented into 600-ms epochs time-locked to the lag 3 position (for V_lag3, VAcon and VAincon conditions, time-locked to T2 onset; for A and N conditions, time-locked to the blank drawing onset; see Figure 1b) with a 100-ms pre-lag3 baseline and were baseline-corrected. Automatic artifact rejection was performed based on a threshold of ±75 μV for both EEG and EOG electrodes, in order to discard epochs contaminated by eye movements, eye blinks or muscle activities. Based on previous EEG studies on the attentional blink (e.g., Haroush, Deouell, & Hochstein, 2011; Kranczioch, Debener, Maye, & Engel, 2007; Kranczioch & Thorne, 2015; Maier & Rahman, 2018; Vogel, Luck, & Shapiro, 1998; Zhao et al., 2021), only trials (epochs) on which T1 identity was correctly discriminated were further analyzed, hence leaving on average 106.7 (range 69–133) valid epochs per condition (see Table S1 for the average, minimal and maximal number of valid epochs in each condition). The remaining valid epochs in each condition were then averaged separately to obtain corresponding ERP waveforms. The EEG preprocessing and subsequent ERP analysis were performed using the EEGLAB toolbox (Delorme & Makeig, 2004) in combination with custom-built MATLAB scripts (The MathWorks, Inc.).

2.4 | Data analysis

To reveal neural activities underlying the cross-modal boost of T2 discrimination during the attentional blink and examine the roles of audiovisual temporal synchrony and semantic congruency, the present study isolated audiovisual cross-modal ERP components on the lag

3 trials when a semantically congruent or incongruent sound preceded T2, synchronized with T2, and was delayed relative to T2, respectively, by calculating cross-modal difference waveforms. The cross-modal difference waveforms were obtained by subtracting the summed ERPs elicited by the unimodal V and A stimuli (V + A) from ERPs elicited by the bimodal VA stimuli (VA), and statistically significant positive or negative waves (as compared with 0) in the difference waveforms have been considered as neural activities associated with audiovisual cross-modal interactions (Bonath et al., 2007; Gao et al., 2014; Giard & Peronnet, 1999; Kranczioch & Thorne, 2015; Mishra, Martínez, & Hillyard, 2008, 2010; Mishra, Martínez, Sejnowski, & Hillyard, 2007; Molholm et al., 2002; Molholm, Ritter, Javitt, & Foxe, 2004; Talsma, Doty, & Woldorff, 2007; Talsma & Woldorff, 2005; Teder-Sälejärvi, Di Russo, McDonald, & Hillyard, 2005; Teder-Sälejärvi, McDonald, Di Russo, & Hillyard, 2002; Van der Burg, Talsma, Olivers, Hickey, & Theeuwes, 2011; Yang et al., 2013; Zhao et al., 2021; Zhao, Wang, Feng, & Feng, 2020; Zhao, Wang, Xu, Feng, & Feng, 2018). In the present study, the cross-modal difference waveforms for different bimodal VA stimuli were calculated as follows:

1. When a semantically congruent sound preceded T2 by 200 ms: VAcon_prec − (V + A_prec).
2. When a semantically incongruent sound preceded T2 by 200 ms: VAincon_prec − (V + A_prec).
3. When a semantically congruent sound synchronized with T2: VAcon_sync − (V + A_sync).
4. When a semantically incongruent sound synchronized with T2: VAincon_sync − (V + A_sync).
5. When a semantically congruent sound was delayed relative to T2 by 100 ms: VAcon_delay − (V + A_delay).
6. When a semantically incongruent sound was delayed relative to T2 by 100 ms: VAincon_delay − (V + A_delay).

Prior to the aforementioned calculations, the time-locked ERPs in N condition (wherein a blank drawing was presented at lag 3 without any sound; see Figure 1b) were first subtracted from ERPs in each of the remaining ten conditions (i.e., V_lag3, VAcon_prec, VAcon_sync, VAcon_delay, VAincon_prec, VAincon_sync, VAincon_delay, A_prec, A_sync and A_delay), in order to cancel out not only any common pre-lag3 anticipatory activities that might extend into the post-lag3 period and lead to false discoveries of early cross-modal interactions (cf., Bonath et al., 2007; Mishra et al., 2007, 2008, 2010; Talsma & Woldorff, 2005; Van der Burg et al., 2011; Zhao et al., 2018, 2020, 2021), but also the systematic ERPs elicited by the pre- and post-lag3 distractors (cf., Kranczioch & Thorne, 2015; Luo et al., 2013; Luo, Feng, He, Wang, & Luo, 2010; Maier & Rahman, 2018; Sergent, Baillet, & Dehaene, 2005; Vogel et al., 1998; Zhao et al., 2021). In other words, the time-locked ERPs in N condition were used as a combined estimation of not only the distractor-elicited ERPs but also the prestimulus anticipatory ERPs.

In order to avoid the problem of multiple implicit comparisons that could inflate the Type I error rate (Luck & Gaspelin, 2017), the time windows and electrodes for measuring cross-modal ERP components associated with the cross-modal boost of T2 discrimination during the attentional blink were selected a priori in the current study based on the recent study conducted by Zhao et al. (2021) wherein the same stimuli and a similar experimental design were used. Specifically, the occipitally distributed **N195** component was measured as mean amplitude during the time window of 192–228 ms after T2 onset in the cross-modal difference waveform at the electrode O1, and the parietally distributed **N440** component was quantified as mean amplitude within the time interval of 424–448 ms in the cross-modal difference waveform at the electrode P1.

For statistical analysis, separate two-way repeated-measures ANOVAs with factors of sound temporal position (precede, synchronize, delay) and semantic congruency (congruent, incongruent) were first conducted on the mean amplitudes of each ERP component in the cross-modal difference waveforms, in order to examine the influences of audiovisual temporal synchrony and semantic congruency on the neural basis underlying the cross-modal boost of T2 discrimination during the attentional blink. When there was a main effect or interaction violating the sphericity assumption, the corresponding p value was corrected using the Greenhouse–Geisser method. Only when the two-way interaction was significant would it be further analyzed in the following two directions in order to better understand its pattern: (a) one-way repeated-measures ANOVAs with a factor of sound temporal position were conducted separately for semantically congruent and incongruent conditions, and pairwise comparisons by paired t-tests between ERP amplitude when sound synchronized versus that when sound preceded, and between ERP amplitude when sound synchronized versus that when sound was delayed, were performed only after finding a significant (i.e., $p < .05$) main effect of sound temporal position (cf., Fiebelkorn, Foxe, & Molholm, 2010; Fiebelkorn, Foxe, Schwartz, & Molholm, 2010); (b) separate paired t tests on ERP amplitudes between semantic congruent and incongruent conditions were conducted when sound preceded, synchronized and was delayed, respectively. To assess these follow-up results with caution, p values for all follow-up analyses in each direction above were further adjusted using the false discovery rate (FDR) correction (Benjamini & Hochberg, 1995), and the FDR-corrected p value was denoted as "p_{FDR}." For completeness, both the uncorrected and FDR-corrected p values were reported for each follow-up analysis. Moreover, one-sample t tests were conducted between 0 versus the mean amplitude of each ERP component in each of the six cross-modal difference waveforms, respectively, to examine the presence/absence of each cross-modal ERP in each VA condition. Again, both the uncorrected and FDR-corrected p values were reported for each one-sample t test.

Lastly, to further explore the relationship between the proposed psychophysiological processes indexed by the P195 and N440 components (for details, see section 4), the current study conducted two post hoc Pearson correlation analyses between the P195 and N440 amplitudes separately when a semantically incongruent sound synchronized with T2 and when a congruent sound synchronized with T2. The remaining four asynchronous-sound conditions were not considered because neither of the two ERP components was statistically

significant in any of the asynchronous-sound conditions (see section 3). Both the uncorrected and FDR-corrected p values were reported for each correlation coefficient.

3 | RESULTS

3.1 | Behavioral data

T2 discrimination accuracy (given correct T1 discrimination, the same below) was first compared between V_lag3 and V_lag8 conditions by a paired t test to check whether the basic visual attentional blink effect was successfully induced in the present study. Indeed, T2 accuracy was significantly lower at lag 3 than lag 8 [V_lag3: 43.0 ± 2.2% (M ± SE); V_lag8: 62.5 ± 2.8%; $t(41) = -13.90$, $p < .0001$, Cohen's $d = -2.14$], indicative of a robust attentional blink effect (Raymond et al., 1992).

Next, within lag 3 trials, T2 accuracy in V_lag3 condition was subtracted from T2 accuracy in each VA condition (i.e., VAcon_prec, VAcon_sync, VAcon_delay, VAincon_prec, VAincon_sync, VAincon_delay) to obtain the "cross-modal boost effect" (i.e., the beneficial effect of sound on T2 discrimination) in each of these conditions (Figure 2a). These difference measures were then subject to a 3 (sound temporal position: precede, synchronize, delay) × 2 (semantic congruency: congruent, incongruent) repeated-measures ANOVA in order to examine the influences of audiovisual temporal synchrony and semantic congruency on the cross-modal boost effect. The main effects of sound temporal position [$F(2,82) = 6.05$, $p = .004$, $\eta^2_p = 0.13$] and semantic congruency [$F(1,41) = 35.60$, $p < .0001$, $\eta^2_p = 0.47$] were both significant. Importantly, the two-way interaction was also significant [$F(2,82) = 3.24$, $p = .044$, $\eta^2_p = 0.07$]. Further analysis of the interaction showed that the main effect of sound temporal position was nonsignificant in semantically incongruent conditions [$F(2,82) = 0.26$, $p = .770$, $p_{FDR} = 0.770$, $\eta^2_p = 0.006$; precede: −0.7 ± 0.8%; synchronize: −0.1 ± 0.9%; delay: −0.2 ± 0.9%], but was significant in semantically congruent conditions [$F(2,82) = 8.84$, $p = .0003$, $p_{FDR} = 0.0006$, $\eta^2_p = 0.18$]. Pairwise comparisons revealed that the cross-modal boost effect was significantly larger when a congruent sound synchronized with T2 (4.7 ± 1.0%) than when that sound preceded T2 by 200 ms [1.2 ± 0.9%; $t(41) = 4.13$, $p = .0002$, $p_{FDR} = 0.0006$, $d = 0.64$] and when it was delayed by 100 ms [1.9 ± 0.7%; $t(41) = 3.16$, $p = .003$, $p_{FDR} = 0.004$, $d = 0.49$; Figure 2a].

Analyzing the two-way interaction above in another direction found that, although the larger cross-modal boost effect in semantically congruent than incongruent conditions occurred not only when sounds synchronized with T2s [$t(41) = 5.25$, $p < .0001$, $p_{FDR} < 0.0001$, $d = 0.81$] but also when sounds preceded [$t(41) = 2.04$, $p = .048$, $p_{FDR} = 0.048$, $d = 0.31$] and when sounds were delayed [$t(41) = 2.38$, $p = .022$, $p_{FDR} = 0.037$, $d = 0.37$; Figure 2a], the congruent-minus-incongruent magnitude difference in cross-modal boost effect (i.e., the semantic congruency effect; see Figure 2b) was significantly greater when sounds synchronized with T2s (4.8 ± 0.9%) than when sounds preceded [1.8 ± 0.9%; $t(41) = 2.47$, $p = .018$, $p_{FDR} = 0.037$, $d = 0.38$] and when sounds were delayed [2.1 ± 0.9%; $t(41) = 2.06$, $p = .046$, $p_{FDR} = 0.048$, $d = 0.32$]. In addition, one-sample t tests conducted between 0 versus the cross-modal boost effect in each VA condition yielded that the cross-modal boost effect was significant when a semantically congruent sound synchronized with T2 and when it was delayed relative to T2 by 100 ms [VAcon_sync: $t(41) = 4.50$, $p < .0001$, $p_{FDR} = 0.0003$, $d = 0.69$; VAcon_delay: $t(41) = 2.50$, $p = .017$, $p_{FDR} = 0.049$, $d = 0.39$; Figure 2a], but not in the remaining four VA conditions [VAcon_prec: $t(41) = 1.27$, $p = .213$, $p_{FDR} = 0.426$,

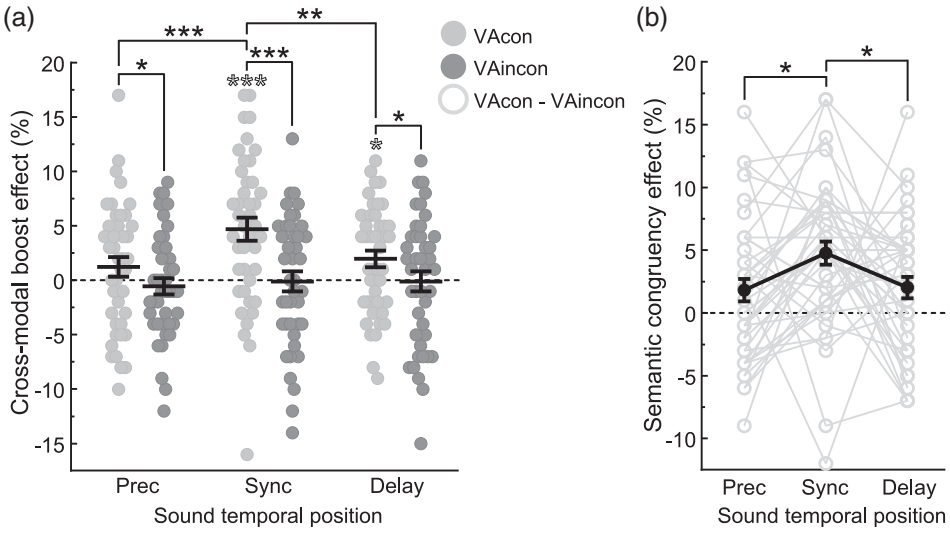

FIGURE 2 Behavioral results. (a) Cross-modal boost effect on lag 3 trials (quantified by the VA-minus-V difference in T2|T1 accuracy) as functions of sound temporal position (precede, synchronize, delay) and audiovisual semantic congruency (congruent, incongruent). The symbol "*" in white denotes a significant cross-modal boost effect against zero. (b) Audiovisual semantic congruency effect on lag 3 trials (quantified by the congruent-minus-incongruent magnitude difference in cross-modal boost effect) as a function of sound temporal position. In both graphs, the single-subject data are depicted by gray scatter dots, the group-averaged data are marked by black symbols, and error bars represent ±1 SE; *$p < .05$; **$p < .01$; ***$p < .001$

$d = 0.19$; VAincon_prec: $t(41) = -0.87$, $p = .389$, $p_{FDR} = 0.583$, $d = -0.13$; VAincon_sync: $t(41) = -0.07$, $p = .944$, $p_{FDR} = 0.944$, $d = -0.01$; VAincon_delay: $t(41) = -0.17$, $p = .868$, $p_{FDR} = 0.944$, $d = -0.03$]. These findings indicates that (a) the cross-modal boost of T2 discrimination during the attentional blink originates from audiovisual cross-modal interactions (which predicts the largest boost when a sound synchronizes with T2) rather than sound-induced alerting effect (which predicts the largest boost when a sound precedes T2 by 200 ms), which is in agreement with Olivers and Van der Burg (2008); (b) the semantic congruency effect on the cross-modal boost during the attentional blink is modulated by audiovisual temporal synchrony, which is inconsistent with findings reported by Adam and Noppeney (2014).

3.2 | ERP data

Figure 3 displays the extracted ERP waveforms (time-locked to lag 3 position) elicited by nondistractor stimuli in all but N condition [i.e., A_prec, A_sync, A_delay, V (V_lag3), VAcon_prec, VAcon_sync, VAcon_delay, VAincon_prec, VAincon_sync and VAincon_delay], which were obtained by subtracting the time-locked ERPs in N condition from ERPs in each of the remaining 10 conditions (see section 2.4). As shown, the auditory-only stimuli presented at lag 3 (A_sync) evoked classic auditory P1, N1 and P2 components with maxima over fronto-central electrodes (Figure 3a, electrode FCz), and the auditory-only stimuli presented 200 ms before and 100 ms after lag 3 position (A_prec and A_delay) elicited similar auditory ERPs but with a 200-ms precedence and a 100-ms delay in latency, respectively. The visual-only T2s presented at lag 3 (V) elicited characteristic visual P1 and N1 components over bilateral parieto-occipital electrodes (Figure 3a, electrode PO8). Among ERP waveforms elicited by the 6 types of bimodal VA stimuli, both the auditory and visual ERP components could be discerned, and the influence of audiovisual temporal asynchrony on the evolution of bimodal ERP waveforms could also be identified (Figure 3b). The patterns of these ERP waveforms confirm not only the precision of the manipulation of sound temporal position, but also the validity of introducing the N condition to cancel out distractor-evoked ERPs in RSVP streams (Luo et al., 2010, 2013; Maier & Rahman, 2018; Sergent et al., 2005; Zhao et al., 2021).

3.2.1 | Early cross-modal P195 component

To reveal neural activities underlying the cross-modal boost of T2 discrimination during the attentional blink and examine the roles of audiovisual temporal synchrony and semantic congruency, audiovisual cross-modal ERP components on the lag 3 trials when a semantically

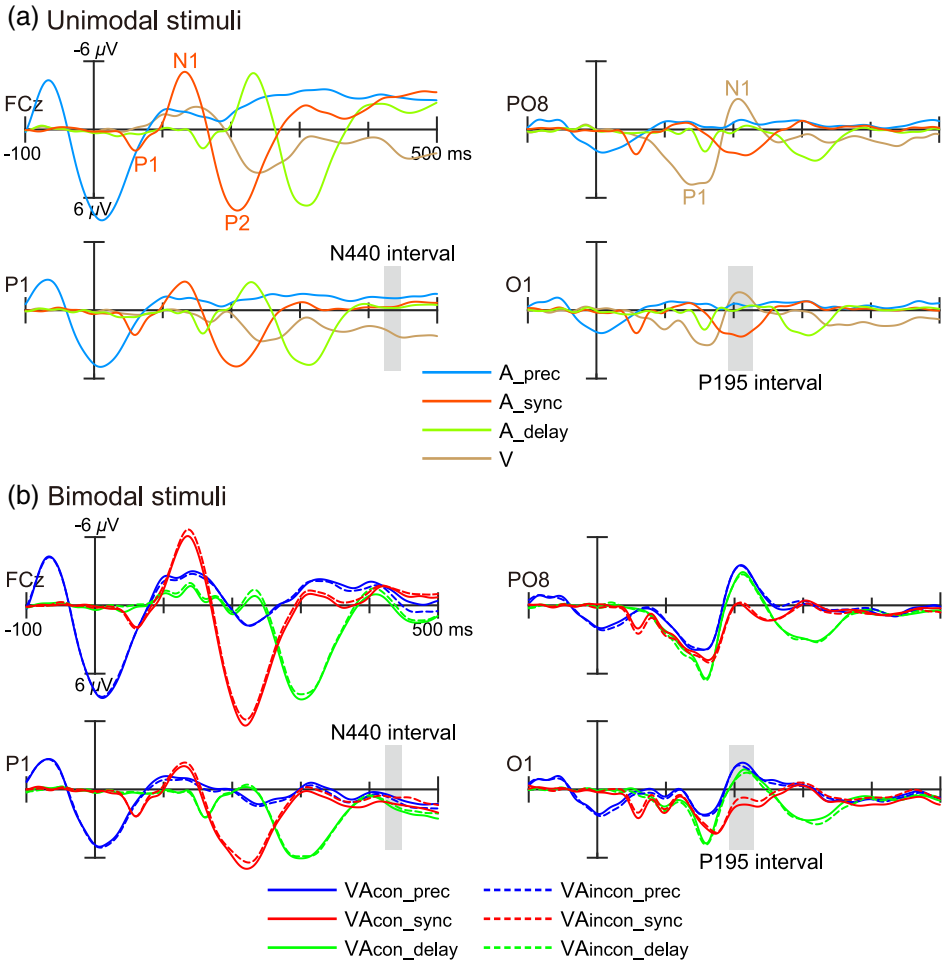

FIGURE 3 The grand-averaged, *extracted* ERP waveforms (time-locked to lag 3 position) elicited purely by non-distractor stimuli in all but N condition, which were obtained by subtracting ERPs in N condition from ERPs in each of the remaining 10 conditions (see Figure 1b for reference). ERP waveforms elicited by unimodal V and A stimuli (a) and bimodal VA stimuli (b) are shown from electrodes FCz, PO8, P1 and O1. The electrodes FCz and PO8 were where auditory-evoked ERP components (P1, N1 and P2) and visual-evoked ERP components (P1 and N1) were largest, respectively, whereas the electrodes P1 and O1 were where the crucial cross-modal N440 (424–448 ms) and P195 (192–228 ms) components were quantified, respectively

congruent or incongruent sound preceded T2, synchronized with T2, and was delayed relative to T2, were isolated respectively by calculating cross-modal difference waveforms (see section 2.4). The occipitally distributed early cross-modal negativity N195 (192–228 ms after T2 onset at the electrode O1), which has been shown to underlie the cross-modal boost during the attentional blink (Zhao et al., 2021), was manifested as a *positive* component in the present study (Figure 4b). Thus, this component was labeled as P195 in the present study for clarity. Then, a 3 (sound temporal position: precede, synchronize, delay) × 2 (semantic congruency: congruent, incongruent) repeated-measures ANOVA was conducted on the mean amplitudes of P195 component. The results showed a significant main effect of sound temporal position [$F(2,82) = 4.71$, $p = .012$, $\eta^2_p = 0.10$] and a nonsignificant main effect of semantic congruency [$F(1,41) = 0.38$, $p = .543$, $\eta^2_p = 0.01$]. Importantly, there was a significant two-way interaction [$F(2,82) = 8.38$, $p = .0005$, $\eta^2_p = 0.17$]. Further analysis of the interaction revealed that the main effect of sound temporal position was significant only in semantically congruent conditions [$F(2,82) = 9.32$, $p = .0002$, $p_{FDR} = 0.0004$, $\eta^2_p = 0.19$], with the P195 amplitude being significantly larger when a congruent sound synchronized with T2 [0.61 ± 0.20 μV (*M* ± *SE*); Figure 4b] than when the sound preceded T2 by 200 ms [−0.30 ± 0.21 μV; $t(41) = 4.50$, $p < .0001$, $p_{FDR} = 0.0002$, $d = 0.69$; Figure 4a] and when it was delayed by 100 ms [0.16 ± 0.23 μV; $t(41) = 2.15$, $p = .038$, $p_{FDR} = 0.050$, $d = 0.33$; Figures 4c and 6a]. In contrast, the main effect of sound temporal position was not significant in semantically incongruent conditions [$F(2,82) = 1.97$, $p = .146$, $p_{FDR} = 0.146$, $\eta^2_p = 0.05$; precede: −0.12 ± 0.21 μV; synchronize: 0.10 ± 0.22 μV; delay: 0.32 ± 0.21 μV].

Analyzing the two-way interaction above in another direction found that the P195 amplitude was significantly larger in the semantically congruent than incongruent condition only when sounds synchronized with T2s [$t(41) = 3.25$, $p = .002$, $p_{FDR} = 0.006$, $d = 0.50$; Figure 4b], but not when sounds preceded [$t(41) = -1.16$, $p = .254$, $p_{FDR} = 0.254$, $d = -0.18$; Figure 4a] or when sounds were delayed [$t(41) = -1.43$, $p = .161$, $p_{FDR} = 0.242$, $d = -0.22$; Figures 4c and 6a] relative to T2. Furthermore, one-sample *t* tests conducted between 0 versus the P195 amplitude in each of the six cross-modal difference waveforms yielded that the P195 component was elicited significantly only when a semantically congruent sound synchronized with T2 [VAcon_sync: $t(41) = 3.06$, $p = .004$, $p_{FDR} = 0.024$, $d = 0.47$; Figure 6a], but not in the remaining five difference waveforms [VAincon_sync: $t(41) = 0.43$, $p = .669$, $p_{FDR} = 0.669$, $d = 0.07$; VAcon_prec: $t(41) = -1.42$, $p = .163$, $p_{FDR} = 0.326$, $d = -0.22$; VAincon_prec: $t(41) = -0.56$, $p = .579$, $p_{FDR} = 0.669$, $d = -0.09$; VAcon_delay: $t(41) = 0.69$, $p = .494$, $p_{FDR} = 0.669$, $d = 0.11$; VAincon_delay: $t(41) = 1.55$, $p = .128$, $p_{FDR} = 0.326$, $d = 0.24$]. These findings demonstrate that the occurrence of early cross-modal P195 component is strongly contingent on the temporal synchrony between the sound and T2 and is further limited by audiovisual semantic incongruency, whose pattern is highly consistent with the pattern observed in the current behavioral results (see Figure 2a). In terms of scalp topography, although the significant P195 in the VAcon_sync condition seemed to have both occipital and central distributions (Figure 4b), the central one was also prominent in the two preceding-sound conditions (Figure 4a) wherein the occipital one was absent and no T2 accuracy improvement was observed. Therefore,

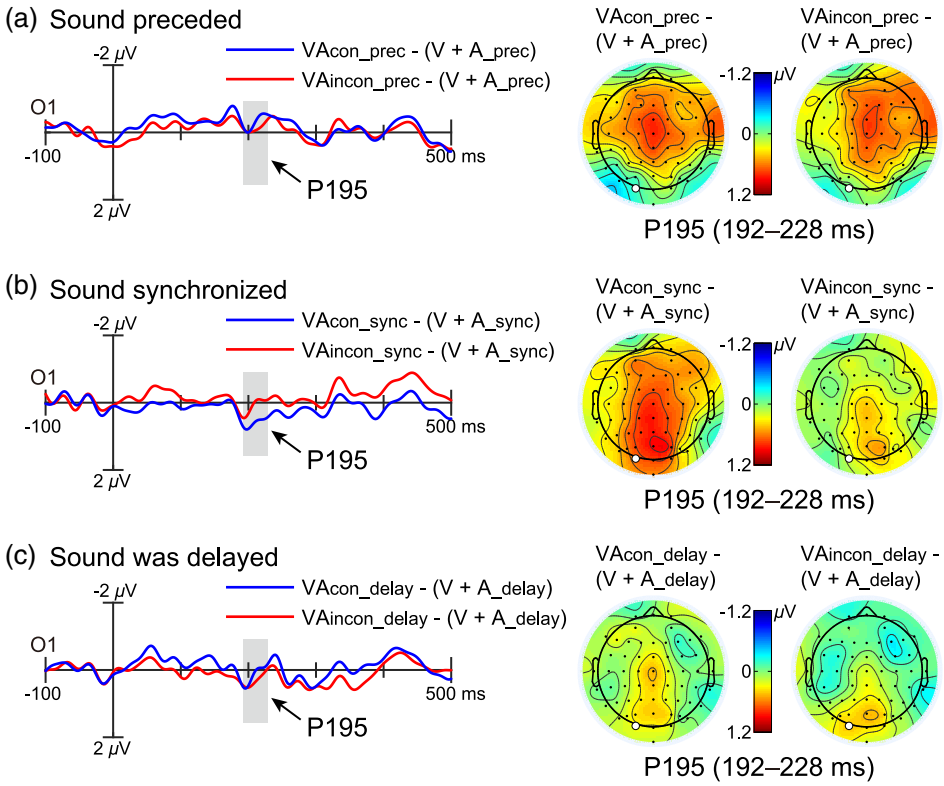

FIGURE 4 The grand-averaged cross-modal difference waveforms (left) and scalp topographies (right) for P195 component when semantically congruent and incongruent sounds preceded T2 by 200 ms (a), synchronized with T2 (b) and were delayed relative to T2 by 100 ms (c). The shaded areas on waveforms and the white dots on scalp topographies depict the time window (192–228 ms after T2 onset) and electrode (O1) where the mean amplitude of P195 component was quantified

the centrally distributed voltage was more likely to be unbalanced EEG noise or another cross-modal interaction unrelated to the sound-induced T2 accuracy improvement.

3.2.2 | Late cross-modal N440 component

The late cross-modal negativity N440 (424–448 ms after T2 onset at the electrode P1), which has been found to underlie the semantic congruency effect on the cross-modal boost during the attentional blink (Zhao et al., 2021), was also observed in the current study (Figure 5b). This difference component appeared to be maximal in the temporally-synchronous but semantically incongruent condition, wherein a parietal scalp distribution with a trend of left-hemispheric preponderance could be observed. The same 3 (sound temporal position: precede, synchronize, delay) × 2 (semantic congruency: congruent, incongruent) repeated-measures ANOVA conducted on the mean amplitudes of N440 component showed that neither the main effect of sound temporal position [$F(2,82) = 2.23$, $p = .115$, $\eta^2_p = 0.05$] nor the main effect of semantic congruency [$F(1,41) = 1.79$, $p = .188$, $\eta^2_p = 0.04$] reached significance, but the two-way interaction was significant [$F(2,82) = 4.75$, $p = .011$, $\eta^2_p = 0.10$]. Further analysis of the interaction revealed that the main effect of sound temporal position was not significant in semantically congruent conditions [$F(2,82) = 0.43$, $p = .655$, $p_{FDR} = 0.655$, $\eta^2_p = 0.01$; precede: −0.29 ± 0.26 μV; synchronize: −0.46 ± 0.27 μV; delay: −0.19 ± 0.29 μV], but was significant in semantically incongruent conditions [$F(2,82) = 4.34$, $p = .016$, $p_{FDR} = 0.032$, $\eta^2_p = 0.10$]. Pairwise comparisons found that the N440 amplitude was significantly greater when an incongruent sound synchronized with T2 (−1.04 ± 0.36 μV; Figure 5b) than when the sound preceded T2 by 200 ms [−0.03 ± 0.31 μV; $t(41) = -3.00$, $p = .005$, $p_{FDR} = 0.019$, $d = -0.46$; Figures 5a and 6b], whereas the N440 amplitude was not greater when the incongruent sound synchronized with T2 than when it was delayed by 100 ms [−0.45 ± 0.27 μV; $t(41) = -1.60$, $p = .117$, $p_{FDR} = 0.156$, $d = -0.25$; Figure 5c].

Analyzing the two-way interaction above in another direction yielded that the N440 amplitude was substantially larger in the semantically incongruent than congruent condition only when sounds synchronized with T2s [$t(41) = -2.87$, $p = .007$, $p_{FDR} = 0.020$, $d = -0.44$; Figure 5b], but not when sounds preceded [$t(41) = 1.24$, $p = .224$, $p_{FDR} = 0.265$, $d = 0.19$; Figure 5a] or when sounds were delayed [$t(41) = -1.13$, $p = .265$, $p_{FDR} = 0.265$, $d = 0.17$; Figures 5c and 6b] relative to T2. Lastly, one-sample t tests performed between 0 versus the N440 amplitude in each of the six cross-modal difference waveforms showed that the N440 component was elicited significantly only when a semantically incongruent sound synchronized with T2 [VAincon_sync: $t(41) = -2.86$, $p = .007$, $p_{FDR} = 0.040$, $d = -0.44$; Figure 6b], but not in the remaining five difference waveforms [VAcon_sync: $t(41) = -1.74$, $p = .089$, $p_{FDR} = 0.216$, $d = -0.27$; VAcon_prec: $t(41) = -1.11$, $p = .274$, $p_{FDR} = 0.411$, $d = -0.17$; VAincon_prec: $t(41) = -0.10$, $p = .923$, $p_{FDR} = 0.923$, $d = -0.01$; VAcon_delay: $t(41) = -0.67$, $p = .508$, $p_{FDR} = 0.610$, $d = -0.10$; VAincon_delay: $t(41) = -1.64$, $p = .108$, $p_{FDR} = 0.216$, $d = -0.25$]. These results demonstrate that the late occurring cross-modal N440 component is sensitive only to semantically incongruent audiovisual T2s and is further modulated by audiovisual temporal synchrony, whose pattern is in agreement with the current behavioral findings

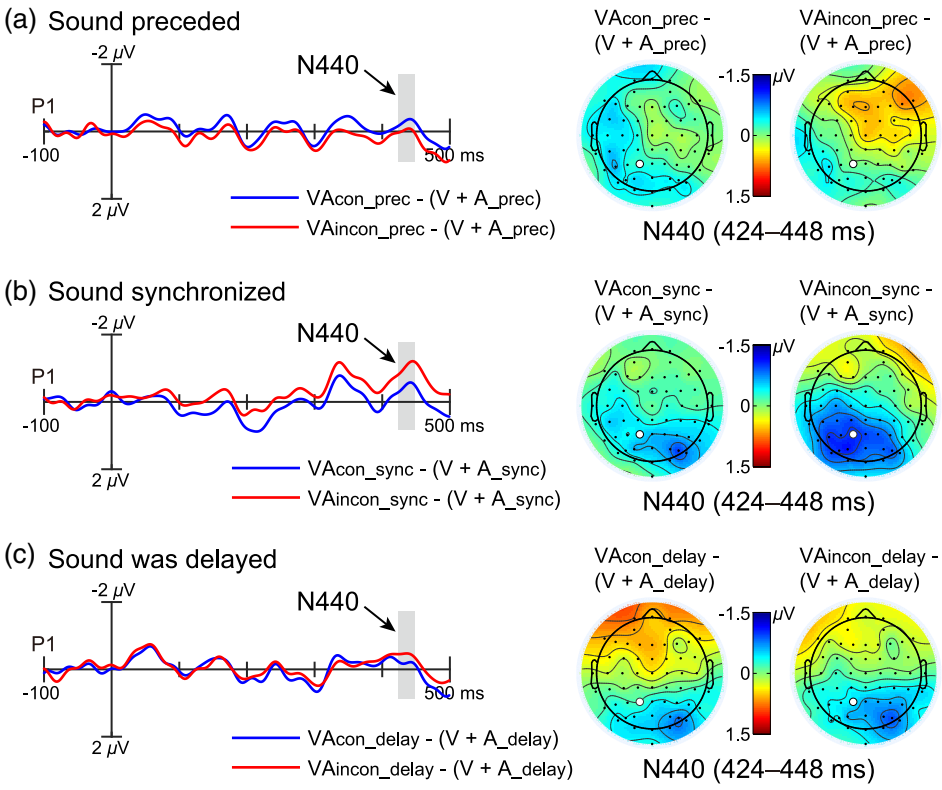

FIGURE 5 The grand-averaged cross-modal difference waveforms (left) and scalp topographies (right) for N440 component when semantically congruent and incongruent sounds preceded T2 by 200 ms (a), synchronized with T2 (b) and were delayed relative to T2 by 100 ms (c). The shaded areas on waveforms and the white dots on scalp topographies depict the time window (424–448 ms after T2 onset) and electrode (P1) where the mean amplitude of N440 component was quantified

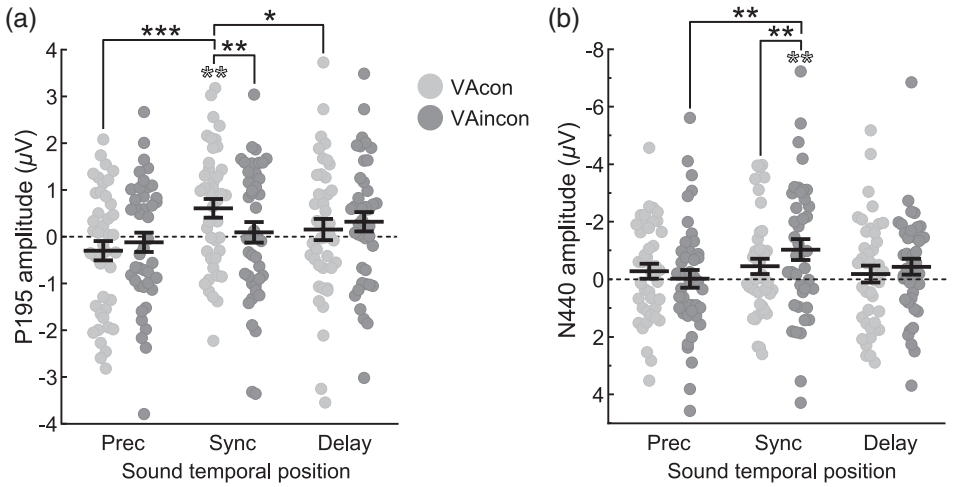

FIGURE 6 A summary for the mean amplitudes of P195 component (a) and N440 component (b) in cross-modal difference waveforms as functions of sound temporal position (precede, synchronize, delay) and audiovisual semantic congruency (congruent, incongruent), with the single-subject data being depicted by gray scatter dots and the group-averaged data being marked by black symbols. The symbol "*" in white denotes a significant cross-modal ERP amplitude against zero. Error bars correspond to ±1 *SE*; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

that the congruency effect was largest when the sound and T2 was synchronous (see Figure 2b), thereby providing clear electrophysiological evidence that the semantic congruency effect on the cross-modal boost of T2 discrimination during the attentional blink is contingent on audiovisual temporal synchrony.

3.2.3 | Post hoc correlation analyses between the P195 and N440 amplitudes

To further explore the relationship between the proposed psychophysiological processes indexed by the P195 and N440 components (for details, see section 4), the current study conducted the following two post hoc Pearson correlation analyses between the P195 and N440 amplitudes. To begin with, when a semantically *incongruent* sound synchronized with T2, the P195 amplitude was found to be correlated significantly with the N440 amplitude [$r(40) = 0.38$, $p = .012$, $p_{FDR} = 0.024$], indicating that subjects with smaller P195 positive amplitudes in this condition tended to have larger N440 negative amplitudes. In contrast, when a semantically *congruent* sound synchronized with T2, there was no reliable correlation between the statistically significant P195 amplitude and the nonsignificant N440 amplitude [$r(40) = 0.23$, $p = .142$, $p_{FDR} = 0.142$], confirming that the close linkage above is specific to the semantically incongruent audiovisual stimuli.

Incidentally, post hoc inspection of the cross-modal difference waveforms at electrodes O1 and P1 suggests that there might be some unexpected ERP effects earlier than 100 ms after T2 onset. However, further reasoning and statistical analysis indicate that these visually apparent effects were actually not reliable, which were detailed in Section C of Data S1.

4 | DISCUSSION

The current study aimed at revealing whether the cross-modal boost of T2 discrimination during the attentional blink originates from audiovisual cross-modal interactions or is merely a manifestation of sound-induced, modality-nonspecific alerting effect, and whether the audiovisual semantic congruency effect on the cross-modal boost during the attentional blink is contingent on audiovisual temporal synchrony. By ensuring that the temporal position of the sound was manipulated *within* each block and different sound temporal positions were *equally* probable, the behavioral data showed that when the task-irrelevant sound was semantically congruent with T2 presented at lag 3, the sound-induced enhancement of T2 discrimination accuracy relative to the no-sound condition was substantially larger when the sound synchronized with T2 than when it preceded T2 by 200 ms or when it was delayed by 100 ms, and the delayed sound also boosted T2 discrimination to a moderate degree whereas the preceding sound did not. This inverted U-shaped pattern of T2 accuracy improvement was highly consistent with the characteristic of audiovisual temporal binding window (Donohue et al., 2011; Meredith et al., 1987; Spence et al., 2001; Stone et al., 2001; van Wassenhove et al., 2007; Zampini et al., 2005), thus providing strong support for the proposal that the cross-modal boost of T2 discrimination during the attentional blink stems genuinely from audiovisual interactions rather than transient-induced alertness (Olivers & Van der Burg, 2008). Indeed, had the cross-modal boost effect resulted from the sound-induced alertness, T2 accuracy improvement should have been the greatest when the sound preceded T2 by 200 ms (wherein alerting effect is typically maximal: Bertelson, 1967; Posner & Boies, 1971; Niemi & Näätänen, 1981; Los & Van den Heuvel, 2001), and should have been absent when the sound was delayed by 100 ms (wherein alertness in advance is impossible; for similar logic, see Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008). In addition, a classic visual attentional blink study (Vogel et al., 1998) has shown that when a visual flash, but not a sound, was presented synchronously with T2 at lag 3 (wherein alertness is possible but audiovisual interaction is impossible), T2 discrimination was actually impaired instead of enhanced, which echoes the current findings and demonstrates that the modality-nonspecific alerting hypothesis cannot account for the cross-modal boost during the attentional blink.

On the contrary, when the task-irrelevant sound was semantically incongruent with T2 presented at lag 3, no T2 accuracy enhancement was found regardless of the temporal position of the sound. In fact, it was the different patterns between the semantically congruent and incongruent sounds in terms of the audiovisual temporal synchrony effect on T2 accuracy enhancement (see Figure 2a), that resulted in the audiovisual semantic congruency effect on T2 accuracy enhancement being modulated by audiovisual temporal synchrony (see Figure 2b). These observations are inconsistent with the finding reported by Adam and Noppeney (2014) that the semantic congruency effect was independent of sound temporal position. However, as noted in section 1, the preceding-sound and synchronous-sound conditions were presented using a block-design in the study of Adam and Noppeney (2014). Therefore, in their preceding-sound blocks, the semantically congruent or incongruent sound was delivered prior to T2 on *each* trial. It is possible that this kind of block-design may have triggered the audiovisual temporal recalibration in response to T2 and the 100% preceding sound (Bhat et al., 2015; Fujisaki et al., 2004; Simon et al., 2017; Van der Burg et al., 2013), thereby weakening the underlying modulation of sound temporal position. In contrast, the present study presented all conditions including the preceding-sound, synchronous-sound and delayed-sound conditions in a pseudo-randomized order *within* each block, hence the avoidance of this possibility. Accordingly, the current finding of the largest audiovisual semantic congruency effect on T2 accuracy enhancement when the sound synchronized with T2 clearly supports the proposal that the effect of higher-level audiovisual semantic congruency on the cross-modal boost during the attentional blink is nonetheless dependent on the low-level audiovisual temporal synchrony.

The aforementioned behavioral findings are also supported by the present ERP data. To begin with, although the early cross-modal difference component N195 (192–228 ms after T2 onset at the electrode O1), which has been shown to underlie the cross-modal boost of T2 discrimination during the attentional blink (Zhao et al., 2021), was manifested as a *positive* component in the present study (labeled as **P195**), the pattern of modulations of the P195 component (see Figure 6a) was in close accordance with that of the behavioral findings (see Figure 2a). Specifically, the cross-modal P195 component was elicited significantly only when a semantically congruent sound synchronized with T2, and the P195 amplitude was found to be substantially greater in the congruent-sound than incongruent-sound condition only when these sounds synchronized with T2. These results indicate that the temporal synchrony between the sound and T2 is a prerequisite for triggering early cross-modal interaction manifested by the P195 component, and that the semantic congruency between the sound and T2 would further limit the occurrence of this early cross-modal interaction. Importantly, the co-variation between ERP and behavioral data provides convergent evidence that the cross-modal boost of T2 discrimination during the attentional blink originates genuinely from audiovisual cross-modal interactions and the audiovisual semantic congruency effect is contingent on audiovisual temporal synchrony. However, it is noteworthy that although the semantically congruent sound that was delayed relative to T2 by 100 ms also boosted T2 discrimination but to a lesser extent, the P195 component in that condition did not reach significance. The reason for this discrepancy might be that the calculation of cross-modal difference waveforms, which was necessary to isolate the P195 component, decreased the signal-to-noise ratio of ERP waveforms and led to the P195 component being harder to reach significance than T2 accuracy improvement, especially when T2 accuracy improvement was merely 1.9% in that case (see Figure 2a). Additional research with more trials might be needed to achieve a higher signal-to-noise ratio when examining the significance of P195 component.

Interestingly, the current study showed that the occipitally distributed early cross-modal ERP component (i.e., P195) was positive in polarity, whereas this component was found to be negative (i.e., N195) in the recent study conducted by Zhao et al. (2021). Indeed, in the existing literature regarding multisensory integration, the polarity of the early cross-modal ERP component during 150–220 ms over the parieto-occipital region has been a controversial issue. Specifically, similar to the present P195 component, many previous studies found this component positive, manifested as a more positive amplitude in the bimodal ERP waveforms than the summed unimodal ERP waveforms (e.g., Gao et al., 2014; Giard & Peronnet, 1999; Molholm et al., 2002; Ren et al., 2018; Stekelenburg & Vroomen, 2005; Teder-Sälejärvi et al., 2002, 2005; Yang et al., 2013; Zhao et al., 2018, 2020). In contrast, some other prior studies found the component negative (e.g., Brandwein et al., 2011, 2013; Kaya & Kafaligonul, 2019; Molholm et al., 2004; Molholm, Murphy, Bates, Ridgway, & Foxe, 2020), consistent with the N195 component observed by Zhao et al. (2021). The factors affecting the polarity of this cross-modal ERP have not yet been examined so far, thus the reasons for its polarity reversal are currently unknown. As far as we can speculate, the polarity reversal here might be attributed to a subtle difference in experimental design between the current and Zhao et al.'s (2021) studies. Concretely, we noted that the crucial visual stimulus T2 could be *absent* (i.e., replaced by a blank drawing) in 40% of the RSVP streams in Zhao et al.'s (2021) study (see their fig. 1c). In contrast, since T2 would appear even when a blank drawing had been presented in the current study (see Figure 1b, A_prec, A_sync, A_delay and N conditions), it was actually presented in *every single* RSVP stream. Such difference could lead participants in the two experiments to utilize slightly different top-down strategies when encoding T2, which might result in the observed polarity reversal. Although further studies are definitely needed to test the speculation, this cross-modal ERP component, regardless of its polarity, has been generally considered as reflecting an influence of auditory signals on early visual discrimination processing, because its timing and scalp distribution were similar to visual N1 component elicited by unimodal visual stimuli (Giard & Peronnet, 1999; Kaya & Kafaligonul, 2019; Molholm et al., 2002, 2004; Stekelenburg & Vroomen, 2005; Teder-Sälejärvi et al., 2002, 2005), with their similarity in timing being substantiated also in the current study [see Figure 3a (visual N1) vs. Figure 4b (cross-modal P195)]. Moreover, the neural generators of this component have been localized consistently to the ventral occipito-temporal cortex regardless of its polarity

(Molholm et al., 2004; Teder-Sälejärvi et al., 2002, 2005). Therefore, despite the polarity reversal as compared with Zhao et al.'s (2021) study, we suggests that the specific psychophysiological meaning reflected by the present P195 component remains unaltered. That is, the cross-modal boost of T2 discrimination during the attentional blink stems from the synchronous sound cross-modally strengthening the early visual discrimination processing for T2, which might lead to the visual representation of T2 becoming more durable for the final conscious report (Zhao et al., 2021).

Another unexpected but important finding in the current study was that when the sound synchronized with T2, the cross-modal P195 component was elicited significantly only in the congruent-sound condition but was completely absent in the incongruent-sound condition, indicating the modulation of audiovisual semantic congruency on T2 discrimination has occurred at the early stage of visual discrimination processing. This finding is inconsistent with the pattern recently reported by Zhao et al. (2021) that the homologous N195 component was unaffected by audiovisual semantic incongruency. It is worth mentioning that in those prior studies wherein the occipitally distributed early cross-modal ERP was suppressed by audiovisual semantic incongruency (e.g., Molholm et al., 2004), participants had to attend voluntarily to *both* the visual and auditory modalities in order to ensure task performance. In contrast, in other previous studies wherein the occipital ERP around 200 ms was unaffected by audiovisual semantic incongruency (e.g., Sinke et al., 2014; Yuval-Greenberg & Deouell, 2007), participants were required to focus *only* on the visual modality while ignoring all sounds. Based on the two pieces of evidence, it seems that although our subjects were instructed to discriminate the visual T2 under the premise of ignoring all sounds, information from the auditory modality was still attended to a certain degree, which may have resulted in the audiovisual semantic congruency modulating T2 discrimination at the early stage of processing in the current study.

In terms of the reason for the inadequately ignored auditory information, the present study speculated that it might derive from the sound being delivered actually on 75% of the trials here (see Figure 1b for details) but only on 60% of the trials in the study of Zhao et al. (2021). Consequently, the more frequent presentation of sound in the current than Zhao et al.'s (2021) study could make it possible that the semantic content carried by the sound gained more attentional resources to some extent. This speculation could also explain why the current synchronously presented and semantically incongruent sound did not boost T2 discrimination (Figure 2a) but that in Zhao et al.'s (2021) study did (because the auditory content more difficult to be ignored here might have intensified the semantic conflict when it was semantically incongruent with T2). Meanwhile, it should be noted that being unable to entirely ignore the auditory modality does not mean that our participants simply guessed T2 identity based on the semantic content carried by the sound when T2 identity was subjectively blurry. Indeed, had such guesswork been engaged frequently in the current study, given that the identity of the sound was uninformative of the exact identity of T2 (see section 2.2): (a) T2 discrimination accuracy should have been worse when the semantically incongruent sound synchronized with T2 than when no sound was delivered, but was not (see Figure 2a); (b) the semantic conflict should have been weakened when the semantically incongruent sound synchronized with T2, resulting in the cross-modal N440 component sensitive to semantic incongruency being nonsignificant in this condition (but the N440 component was actually significant in this case; see below).

Following the occipitally distributed P195 component, the late cross-modal negativity **N440** with a parietal maximum (measured during 424–448 ms after T2 onset at the electrode P1), which has been shown to be responsible for the audiovisual semantic congruency effect on the cross-modal boost during the attentional blink (Zhao et al., 2021), was also prominent in the current study. Many previous ERP studies have also reported N400-like deflections similar to the current N440 component that were elicited by synchronously presented, semantically incongruent auditory and visual stimuli (Donohue, Todisco, & Woldorff, 2013; Kang, Chang, Wang, Wei, & Zhou, 2018; Molholm et al., 2004; Zimmer, Itthipanyanan, Grent't-Jong, & Woldorff, 2010). Furthermore, the neural generators of unisensory-evoked N400 effects have been localized mainly to the middle/superior temporal cortex (Khateb et al., 2007; Khateb, Pegna, Landis, Mouthon, & Annoni, 2010; Lau, Phillips, & Poeppel, 2008), which is also a well-documented brain region of multisensory integration (Beauchamp, 2005; Calvert, 2001). The statistical analysis yielded that the present cross-modal N440 component was elicited significantly only when a semantically incongruent sound synchronized with T2, in accordance with the recent finding that this component was sensitive only to audiovisual stimuli with semantic conflict (Zhao et al., 2021). Moreover, in parallel with the current behavioral findings (see Figure 2b), the magnitude of audiovisual semantic congruency effect (incongruent vs. congruent) on the N440 amplitude was found to be also dependent on the temporal co-occurrence of the sound and T2, thereby providing further electrophysiological evidence for the proposal that the influence of higher-level audiovisual semantic congruency on the cross-modal boost during the attentional blink is limited by the low-level audiovisual temporal synchrony.

With regard to the specific functional significance, if the present cross-modal N440 belongs under the N400 family of components, given that early classic investigations typically attribute the occurrence of the N400 effect to the *violation of preexisting semantic expectancy* established by a leading stimulus (Kutas & Federmeier, 2000; Kutas & Hillyard, 1980, 1984), and that the 200-ms preceding, semantically incongruent sound in the current study could establish such semantic expectancy that would be violated ultimately by the incoming T2, the N440 amplitude should have been larger in the incongruent- than congruent-sound condition even when these sounds preceded T2. However, the present study found that such audiovisual semantic congruency effect on the N440 amplitude was completely absent when sounds preceded but highly prominent when sounds synchronized with T2. These novel findings imply that the cross-modal N440 component observed in the current and recent studies (Zhao et al., 2021) may not reflect the violation of semantic expectancy but instead an *extra processing cost* when integrating the temporally synchronous but semantically conflicting auditory and visual inputs. Indeed, a series of recent investigations on the N400

component have demonstrated that within the N400 time range, there are very likely to be not only the classic subcomponent indicative of the violation of semantic expectancy but also another distinct subcomponent indexing the plausibility (or conflict) of semantic integration (Delong, Quante, & Kutas, 2014; Lau, Namyst, Fogel, & Delgado, 2016; Mantegna, Hintz, Ostarek, Alday, & Huettig, 2019; Nieuwland et al., 2020). Therefore, it is rather probable that the cross-modal N440 component here corresponds to the latter N400 subcomponent described above.

Since the appearances of both N440 and P195 components were contingent on the temporal co-occurrence of the sound and T2, and the N440 component occurred only when the sound was semantically incongruent with T2 whereas the P195 component occurred only when the sound was semantically congruent with T2, the current study proposed an updated model for the cross-modal boost of T2 discrimination during the attentional blink, which modified the hierarchical model recently described by Zhao et al. (2021). Specifically, when a semantically ***congruent*** sound synchronizes with T2, the sound-induced early cross-sensory interaction strengthens the visual discrimination processing for T2 at a relatively early stage (indexed by the presence of P195 that might originate from the ventral occipito-temporal cortex), whereby increasing the probability that T2 would ultimately escape the attentional blink. In contrast, when a semantically ***incongruent*** sound synchronizes with T2, given that the task-irrelevant auditory information was presumably harder to be ignored in the present than Zhao et al.'s (2021) study, the semantic conflict caused by the incongruent sound may be aggravated and in turn can interfere with the early facilitatory cross-modal interaction (indexed by the absence of P195). Subsequently, the audiovisual temporal synchrony enforces the semantically incongruent sound and T2 to be bound together for processing at the late semantic analysis stage (for similar binding mechanisms, see Fiebelkorn, Foxe, & Molholm, 2010; Zimmer et al., 2010), wherein the visual representation of T2 that has not obtained the early cross-modal processing gain is more susceptible to the semantically conflicting auditory representation, hence resulting in an extra processing cost (indexed by the presence of N440 that might stem from the middle/superior temporal cortex) that ultimately limits the cross-modal boost of T2 discrimination during the attentional blink. Notably, the main extension of the present model relative to the hierarchical model proposed by Zhao et al. (2021) is that when information in auditory modality receives more attentional resources, the effect of audiovisual semantic congruency would occur not only at the late semantic analysis stage but also at the early visual discrimination stage in advance. Future neuroimaging research with higher spatial resolution (e.g., fMRI) is required to further determine whether the brain regions suggested above are involved in the cross-modal boost of T2 discrimination during the attentional blink.

The updated model is also supported, at least in part, by the post hoc correlation analyses. To begin with, when a semantically *incongruent* sound synchronized with T2, the P195 amplitude was found to be correlated significantly with the N440 amplitude, indicating that subjects with smaller P195 positive amplitudes in this condition tended to have larger N440 negative amplitudes. This finding substantiates the close linkage between the inhibited early cross-modal interaction (nonsignificant P195) and the later cost of integrating semantic conflict (significant N440) when the sound was semantically incongruent with T2, hence suggesting that the ultimate absence of behavioral cross-modal boost of T2 discrimination during the attentional blink in this condition (Figure 2a) may well be a consequence of the interplay between early and late neural processes. In contrast, when a semantically *congruent* sound synchronized with T2, there was no reliable correlation between the statistically significant P195 amplitude and the nonsignificant N440 amplitude, which not only confirms that the close linkage above is specific to the semantically incongruent audiovisual stimuli, but also supports that the neural basis underlying the beneficial effect of the semantically congruent sound on the visual attentional blink lies mainly in the occipitally distributed early cross-modal interaction. Finally, the current findings might also suggest an interesting practical application: when you are scrolling your mouse wheel rapidly, trying to search for multiple key points of an article, for instance, playing a semantically relevant sound in time might help you decrease the likelihood of missing the second key point.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2021ZD0202600 to Wenfeng Feng), the National Natural Science Foundation of China (grant numbers 31771200 and 32171048 to Wenfeng Feng), the Ministry of Education of Humanities and Social Science Project (17YJA880019 to Chengzhi Feng), and the Strategic Priority Research Program of Chinese Academy of Science (XDB32040200 to Yijun Wang).

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

AUTHOR CONTRIBUTIONS

Song Zhao, Yijun Wang and Wenfeng Feng designed the research. Song Zhao and Chongzhi Wang performed the research. Song Zhao, Chongzhi Wang and Wenfeng Feng analyzed the data. Song Zhao, Chengzhi Feng, Yijun Wang and Wenfeng Feng wrote the paper.

DATA AVAILABILITY STATEMENT

The raw EEG data that were analyzed in this study are openly available in the OSF repository (https://osf.io/7dkuz/). The codes (MATLAB scripts) used to analyze these data are available from the corresponding authors by submitting a feasible research plan.

ORCID

Song Zhao https://orcid.org/0000-0001-6453-9214
Wenfeng Feng https://orcid.org/0000-0002-7664-5863

REFERENCES

Adam, R., & Noppeney, U. (2014). A phonologically congruent sound boosts a visual target into perceptual awareness. *Frontiers in Integrative Neuroscience*, *8*, 70. https://doi.org/10.3389/fnint.2014.00070

Beauchamp, M. S. (2005). See me, hear me, touch me: Multisensory integration in lateral occipital-temporal cortex. *Current Opinion in Neurobiology*, *15*(2), 145–153. https://doi.org/10.1016/j.conb.2005.03.011

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the*

Royal Statistical Society: Series B (Methodological), *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bertelson, P. (1967). The time course of preparation. *Quarterly Journal of Experimental Psychology*, *19*(3), 272–279. https://doi.org/10.1080/14640746708400102

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review*, *5*(3), 482–489. https://doi.org/10.3758/BF03208826

Bhat, J., Miller, L. M., Pitt, M. A., & Shahin, A. J. (2015). Putative mechanisms mediating tolerance for audiovisual stimulus onset asynchrony. *Journal of Neurophysiology*, *113*(5), 1437–1450. https://doi.org/10.1152/jn.00200.2014

Bonath, B., Noesselt, T., Martinez, A., Mishra, J., Schwiecker, K., Heinze, H. J., & Hillyard, S. A. (2007). Neural basis of the ventriloquist illusion. *Current Biology*, *17*(19), 1697–1703. https://doi.org/10.1016/j.cub.2007.08.050

Brandwein, A. B., Foxe, J. J., Butler, J. S., Russo, N. N., Altschuler, T. S., Gomes, H., & Molholm, S. (2013). The development of multisensory integration in high-functioning autism: High-density electrical mapping and psychophysical measures reveal impairments in the processing of audiovisual inputs. *Cerebral Cortex*, *23*(6), 1329–1341. https://doi.org/10.1093/cercor/bhs109

Brandwein, A. B., Foxe, J. J., Russo, N. N., Altschuler, T. S., Gomes, H., & Molholm, S. (2011). The development of audiovisual multisensory integration across childhood and early adolescence: A high-density electrical mapping study. *Cerebral Cortex*, *21*(5), 1042–1055. https://doi.org/10.1093/cercor/bhq170

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, *11*(12), 1110–1123. https://doi.org/10.1093/cercor/11.12.1110

Campbell, J. I., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods*, *44*(4), 1255–1265. https://doi.org/10.3758/s13428-012-0186-0

Delong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162. https://doi.org/10.1016/j.neuropsychologia.2014.06.016

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Donohue, S. E., Roberts, K. C., Grent-'t-Jong, T., & Woldorff, M. G. (2011). The cross-modal spread of attention reveals differential constraints for the temporal and spatial linking of visual and auditory stimulus events. *Journal of Neuroscience*, *31*(22), 7982–7990. https://doi.org/10.1523/JNEUROSCI.5298-10.2011

Donohue, S. E., Todisco, A. E., & Woldorff, M. G. (2013). The rapid distraction of attentional resources toward the source of incongruent stimulus input during multisensory conflict. *Journal of Cognitive Neuroscience*, *25*(4), 623–635. https://doi.org/10.1162/jocn_a_00336

Fiebelkorn, I. C., Foxe, J. J., & Molholm, S. (2010). Dual mechanisms for the cross-sensory spread of attention: How much do learned associations matter? *Cerebral Cortex*, *20*(1), 109–120. https://doi.org/10.1093/cercor/bhp083

Fiebelkorn, I. C., Foxe, J. J., Schwartz, T. H., & Molholm, S. (2010). Staying within the lines: The formation of visuospatial boundaries influences multisensory feature integration. *European Journal of Neuroscience*, *31*(10), 1737–1743. https://doi.org/10.1111/j.1460-9568.2010.07196.x

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience*, *7*(7), 773–778. https://doi.org/10.1038/nn1268

Gao, Y., Li, Q., Yang, W., Yang, J., Tang, X., & Wu, J. (2014). Effects of ipsilateral and bilateral auditory stimuli on audiovisual integration: A behavioral and event-related potential study. *Neuroreport*, *25*(9), 668–675. https://doi.org/10.1097/WNR.0000000000000155

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473–490. https://doi.org/10.1162/089892999563544

Haroush, K., Deouell, L. Y., & Hochstein, S. (2011). Hearing while blinking: Multisensory attentional blink revisited. *Journal of Neuroscience*, *31*(3), 922–927. https://doi.org/10.1523/JNEUROSCI.0420-10.2011

Kang, G., Chang, W., Wang, L., Wei, P., & Zhou, X. (2018). Reward enhances cross-modal conflict control in object categorization: Electrophysiological evidence. *Psychophysiology*, *55*(11), e13214. https://doi.org/10.1111/psyp.13214

Kanwisher, N. G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, *27*(2), 117–143. https://doi.org/10.1016/0010-0277(87)90016-3

Kaya, U., & Kafaligonul, H. (2019). Cortical processes underlying the effects of static sound timing on perceived visual speed. *NeuroImage*, *199*, 194–205. https://doi.org/10.1016/j.neuroimage.2019.05.062

Khateb, A., Pegna, A. J., Landis, T., Michel, C. M., Brunet, D., Seghier, M. L., & Annoni, J. M. (2007). Rhyme processing in the brain: An ERP mapping study. *International Journal of Psychophysiology*, *63*(3), 240–250. https://doi.org/10.1016/j.ijpsycho.2006.11.001

Khateb, A., Pegna, A. J., Landis, T., Mouthon, M. S., & Annoni, J. M. (2010). On the origin of the N400 effects: An ERP waveform and source localization analysis in three matching tasks. *Brain Topography*, *23*(3), 311–320. https://doi.org/10.1007/s10548-010-0149-7

Koelewijn, T., Van der Burg, E., Bronkhorst, A. W., & Theeuwes, J. (2008). Priming T2 in a visual and auditory attentional blink task. *Perception & Psychophysics*, *70*(4), 658–666. https://doi.org/10.3758/pp.70.4.658

Kranczioch, C., Debener, S., Maye, A., & Engel, A. K. (2007). Temporal dynamics of access to consciousness in the attentional blink. *NeuroImage*, *37*(3), 947–955. https://doi.org/10.1016/j.neuroimage.2007.05.044

Kranczioch, C., & Thorne, J. D. (2013). Simultaneous and preceding sounds enhance rapid visual targets: Evidence from the attentional blink. *Advances in Cognitive Psychology*, *9*(3), 130–142. https://doi.org/10.2478/v10053-008-0139-4

Kranczioch, C., & Thorne, J. D. (2015). The beneficial effects of sounds on attentional blink performance: An ERP study. *NeuroImage*, *117*, 429–438. https://doi.org/10.1016/j.neuroimage.2015.05.055

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Science*, *4*(12), 463–470. https://doi.org/10.1016/s1364-6613(00)01560-6

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/science.7350657

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

Lau, E. F., Namyst, A., Fogel, A., & Delgado, T. (2016). A direct comparison of N400 effects of predictability and incongruity in adjective-noun combination. *Collabra Psychology*, *2*(1), 13. https://doi.org/10.1525/collabra.40

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Los, S. A., & Van den Heuvel, C. E. (2001). Intentional and unintentional contributions to nonspecific preparation during reaction time foreperiods. *Journal of Experimental Psychology: Human Perception & Performance*, *27*(2), 370–386. https://doi.org/10.1037//0096-1523.27.2.370

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Luo, W., Feng, W., He, W., Wang, N., & Luo, Y. (2010). Three stages of facial expression processing: ERP study with rapid serial visual presentation. *NeuroImage*, *49*(2), 1857–1867. https://doi.org/10.1016/j.neuroimage.2009.09.018

Luo, W., He, W., Yang, S., Feng, W., Chen, T., Wang, L., … Luo, Y. (2013). Electrophysiological evidence of facial inversion with rapid serial visual presentation. *Biological Psychology*, *92*(2), 395–402. https://doi.org/10.1016/j.biopsycho.2012.11.019

Maier, M., & Rahman, R. A. (2018). Native language promotes access to visual consciousness. *Psychological Science*, *29*(11), 1757–1772. https://doi.org/10.1177/0956797618782181

Mantegna, F., Hintz, F., Ostarek, M., Alday, P. M., & Huettig, F. (2019). Distinguishing integration and prediction accounts of ERP N400 modulations in language processing through experimental design. *Neuropsychologia*, *134*, 107199. https://doi.org/10.1016/j.neuropsychologia.2019.107199

McDonald, J. J., Teder-Sälejärvi, W. A., Di Russo, F., & Hillyard, S. A. (2003). Neural substrates of perceptual enhancement by cross-modal spatial attention. *Journal of Cognitive Neuroscience*, *15*(1), 10–19. https://doi.org/10.1162/089892903321107783

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *Journal of Neuroscience*, *7*(10), 3215–3229. https://doi.org/10.1523/JNEUROSCI.07-10-03215.1987

Mishra, J., Martínez, A., & Hillyard, S. A. (2008). Cortical processes underlying sound-induced flash fusion. *Brain Research*, *1242*(4), 102–115. https://doi.org/10.1016/j.brainres.2008.05.023

Mishra, J., Martínez, A., & Hillyard, S. A. (2010). Effect of attention on early cortical processes associated with the sound-induced extra flash illusion. *Journal of Cognitive Neuroscience*, *22*(8), 1714–1729. https://doi.org/10.1162/jocn.2009.21295

Mishra, J., Martínez, A., Sejnowski, T., & Hillyard, S. A. (2007). Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. *Journal of Neuroscience*, *27*(15), 4120–4131. https://doi.org/10.1523/JNEUROSCI.4912-06.2007

Molholm, S., Murphy, J. W., Bates, J., Ridgway, E. M., & Foxe, J. J. (2020). Multisensory audiovisual processing in children with a sensory processing disorder (I): Behavioral and electrophysiological indices under speeded response conditions. *Frontiers in Integrative Neuroscience*, *14*, 4. https://doi.org/10.3389/fnint.2020.00004

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex*, *14*(4), 452–465. https://doi.org/10.1093/cercor/bhh007

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Cognitive Brain Research*, *14*(1), 115–128. https://doi.org/10.1016/s0926-6410(02)00066-6

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological Bulletin*, *89*(1), 133–162. https://doi.org/10.1037/0033-2909.89.1.133

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., … Wolfsthurn, S. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522

Olivers, C. N. L., & Van der Burg, E. (2008). Bleeping you out of the blink: Sound saves vision from oblivion. *Brain Research*, *1242*, 191–199. https://doi.org/10.1016/j.brainres.2008.01.070

Posner, M. I., & Boies, S. J. (1971). Components of attention. *Psychological Review*, *78*(5), 391–408. https://doi.org/10.1037/h0031333

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception & Performance*, *18*(3), 849–860. https://doi.org/10.1037//0096-1523.18.3.849

Ren, Y., Ren, Y., Yang, W., Tang, X., Wu, F., Wu, Q., … Wu, J. (2018). Comparison for younger and older adults: Stimulus temporal asynchrony modulates audiovisual integration. *International Journal of Psychophysiology*, *124*, 1–11. https://doi.org/10.1016/j.ijpsycho.2017.12.004

Sergent, C., Baillet, S., & Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, *8*(10), 1391–1400. https://doi.org/10.1038/nn1549

Simon, D. M., Noel, J. P., & Wallace, M. T. (2017). Event related potentials index rapid recalibration to audiovisual temporal asynchrony. *Frontiers in Integrative Neuroscience*, *11*, 8. https://doi.org/10.3389/fnint.2017.00008

Sinke, C., Neufeld, J., Wiswede, D., Emrich, H. M., Bleich, S., Münte, T. F., & Szycik, G. R. (2014). N1 enhancement in synesthesia during visual and audio-visual perception in semantic cross-modal conflict situations: An ERP study. *Frontiers in Human Neuroscience*, *8*, 21. https://doi.org/10.3389/fnhum.2014.00021

Spence, C., Shore, D. I., & Klein, R. M. (2001). Multisensory prior entry. *Journal of Experimental Psychology: General*, *130*(4), 799–832. https://doi.org/10.1037//0096-3445.130.4.799

Stekelenburg, J. J., & Vroomen, J. (2005). An event-related potential investigation of the time-course of temporal ventriloquism. *Neuroreport*, *16*(6), 641–644. https://doi.org/10.1097/00001756-200504250-00025

Stone, J. V., Hunkin, N. M., Porrill, J., Wood, R., Keeler, V., Beanland, M., … Porter, N. R. (2001). When is now? Perception of simultaneity. *Proceedings of the Royal Society B: Biological Sciences*, *268*(1462), 31–38. https://doi.org/10.1098/rspb.2000.1326

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: Is attending to both modalities a prerequisite for early integration? *Cerebral Cortex*, *17*(3), 679–690. https://doi.org/10.1093/cercor/bhk016

Talsma, D., & Woldorff, M. G. (2005). Selective attention and multisensory integration: Multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience*, *17*(7), 1098–1114. https://doi.org/10.1162/0898929054475172

Teder-Sälejärvi, W. A., Di Russo, F., McDonald, J. J., & Hillyard, S. A. (2005). Effects of spatial congruity on audio-visual multimodal integration. *Journal of Cognitive Neuroscience*, *17*(9), 1396–1409. https://doi.org/10.1162/0898929054985383

Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F., & Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research*, *14*(1), 106–114. https://doi.org/10.1016/s0926-6410(02)00065-4

Van der Burg, E., Alais, D., & Cass, J. (2013). Rapid recalibration to audiovisual asynchrony. *Journal of Neuroscience*, *33*(37), 14633–14637. https://doi.org/10.1523/JNEUROSCI.1182-13.2013

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Non-spatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception & Performance*, *34*(5), 1053–1065. https://doi.org/10.1037/0096-1523.34.5.1053

Van der Burg, E., Talsma, D., Olivers, C. N. L., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage*, *55*(3), 1208–1218. https://doi.org/10.1016/j.neuroimage.2010.12.068

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. https://doi.org/10.1016/j.neuropsychologia.2006.01.001

Vogel, E. K., Luck, S. J., & Shapiro, K. L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology: Human Perception & Performance*, *24*(6), 1656–1674. https://doi.org/10.1037//0096-1523.24.6.1656

Yang, W., Li, Q., Ochi, T., Yang, J., Gao, Y., Tang, X., … Wu, J. (2013). Effects of auditory stimuli in the horizontal plane on audiovisual integration: An event-related potential study. *PLoS One*, *8*(6), e66402. https://doi.org/10.1371/journal.pone.0066402

Yuval-Greenberg, S., & Deouell, L. Y. (2007). What you see is not (always) what you hear: Induced gamma band responses reflect cross-modal interactions in familiar object recognition. *Journal of Neuroscience*, *27*(5), 1090–1096. https://doi.org/10.1523/JNEUROSCI.4828-06.2007

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics*, *67*(3), 531–544. https://doi.org/10.3758/BF03193329

Zhao, S., Feng, C., Huang, X., Wang, Y., & Feng, W. (2021). Neural basis of semantically dependent and independent cross-modal boosts on the attentional blink. *Cerebral Cortex*, *31*(4), 2291–2304. https://doi.org/10.1093/cercor/bhaa362

Zhao, S., Wang, Y., Feng, C., & Feng, W. (2020). Multiple phases of cross-sensory interactions associated with the audiovisual bounce-inducing effect. *Biological Psychology*, *149*, 107805. https://doi.org/10.1016/j.biopsycho.2019.107805

Zhao, S., Wang, Y., Xu, H., Feng, C., & Feng, W. (2018). Early cross-modal interactions underlie the audiovisual bounce-inducing effect. *NeuroImage*, *174*, 208–218. https://doi.org/10.1016/j.neuroimage.2018.03.036

Zimmer, U., Itthipanyanan, S., Grent-'t-Jong, T., & Woldorff, M. G. (2010). The electrophysiological time course of the interaction of stimulus conflict and the multisensory spread of attention. *European Journal of Neuroscience*, *31*(10), 1744–1754. https://doi.org/10.1111/j.1460-9568.2010.07229.x

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Zhao, S., Wang, C., Feng, C., Wang, Y., & Feng, W. (2022). The interplay between audiovisual temporal synchrony and semantic congruency in the cross-modal boost of the visual target discrimination during the attentional blink. *Human Brain Mapping*, *43*(8), 2478–2494. https://doi.org/10.1002/hbm.25797