Attention, Perception, & Psychophysics (2022) 84:1625–1634
https://doi.org/10.3758/s13414-021-02437-4

# Semantically congruent audiovisual integration with modal-based attention accelerates auditory short-term memory retrieval

**Hongtao Yu[1] · Aijun Wang[2] · Ming Zhang[1,2] · JiaJia Yang[1] · Satoshi Takahashi[1] · Yoshimichi Ejima[1] · Jinglong Wu[1,3]**

Accepted: 28 December 2021 / Published online: 31 May 2022
© The Psychonomic Society, Inc. 2022

**Abstract**
Evidence has shown that multisensory integration benefits to unisensory perception performance are asymmetric and that auditory perception performance can receive more multisensory benefits, especially when the attention focus is directed toward a task-irrelevant visual stimulus. At present, whether the benefits of semantically (in)congruent multisensory integration with modal-based attention for subsequent unisensory short-term memory (STM) retrieval are also asymmetric remains unclear. Using a delayed matching-to-sample paradigm, the present study investigated this issue by manipulating the attention focus during multisensory memory encoding. The results revealed that both visual and auditory STM retrieval reaction times were faster under semantically congruent multisensory conditions than under unisensory memory encoding conditions. We suggest that coherent multisensory representation formation might be optimized by restricted multisensory encoding and can be rapidly triggered by subsequent unisensory memory retrieval demands. Crucially, auditory STM retrieval is exclusively accelerated by semantically congruent multisensory memory encoding, indicating that the less effective sensory modality of memory retrieval relies more on the coherent prior formation of a multisensory representation optimized by modal-based attention.

**Keywords** Audiovisual integration · Short-term memory · Modal-based attention · Semantic congruency

## Introduction

Combining inputs from individual sensory stimuli is essential for sufficiently perceiving the real-world environment. Multisensory integration describes the cognitive process in which signals derived from different sensory systems are integrated into a coherent percept, thereby leading to higher accuracy (Lehmann & Murray, 2005), faster reaction times (Talsma et al., 2007), or higher perception precision (Odegaard et al., 2016). Previous multisensory studies in animals indicate that integration efficiency is modulated by several constraints between different channels, such as low-level spatiotemporal congruency (Fort et al., 2002; Stein et al., 1994) and high-level semantic relationships (Doehrmann & Naumer, 2008). The facilitation effect of spatiotemporal congruence has been considered due to the increased neural firing rate of multisensory neurons in the superior colliculus. However, such a theoretical framework cannot account for the facilitated behavioral performance of multisensory inputs with congruent semantic contents.

Multisensory studies have shown that perceptual performance is enhanced or attenuated depending on whether visual- and auditory-channel shared semantic contents belong to the same object (Laurienti et al., 2004; Molholm et al., 2004; Suied et al., 2009). For instance, Laurienti et al. (2004) reported significantly faster visual discrimination when participants responded to congruent audiovisual stimuli (e.g., a blue circle with a sound "*blue*") and suggested that whether the human brain can bind individual visual and auditory signals to one perceptual unit depends on the congruent

✉ Aijun Wang
ajwang@suda.edu.cn

Hongtao Yu
przw2yui@s.okayama-u.ac.jp

Jinglong Wu
jl.wu@siat.ac.cn

[1] Cognitive Neuroscience Laboratory, Graduate School of Interdisciplinary Science and Engineering in Health Systems, Okayama University, Okayama, Japan

[2] Department of Psychology, Research Center for Psychology and Behavioral Sciences, Soochow University, Suzhou, People's Republic of China

[3] Research Center for Medical Artificial Intelligence, Shenzhen Institute of Advanced Technology, Chinese Academy of Science, Shenzhen, Guangdong, China

semantic relationship of the audiovisual pair. It is worth noting that semantically congruent audiovisual integration facilitates not only instant perception performance but also subsequent cognitive performance. Imagine that you must keep the phone number of a new friend in your mind. The memory-encoding process will be facilitated if this friend writes the number while repeating the number in the friend's own voice; alternatively, it will be suppressed if the friend writes the number while making an irrelevant joke.

Recently, using a delayed matching-to-sample paradigm (DMS), Xie et al. (2017) reported that visual working memory (WM) retrieval was accelerated by previous semantically congruent audiovisual encoding compared with the visual-only encoding condition. In particular, it must be noted that overall higher accuracy rates (i.e., 95%) were found under all encoding conditions, indicating that the DMS paradigm cannot sufficiently tax WM resources. The DMS paradigm might be an appropriate paradigm for evaluating short-term memory (STM) and has been widely investigated in recent STM studies (Almadori et al., 2021; Liu et al., 2021). In particular, in previous multisensory memory studies, participants were asked to divide their attention between visual and auditory stimuli during multisensory encoding (Xie et al., 2017; Xie et al., 2019; Xie et al., 2021). However, if the semantic information of visual and auditory stimuli is conflicting, divided attention toward two modalities (e.g., a cat picture with the sound of a dog) might increase susceptibility to a distractor (e.g., the sound of a dog) and lead to impaired encoding of the target modality (e.g., a cat picture) stimulus into memory (Craik et al., 1996), further impacting target modality memory retrieval. Importantly, such interference might be destructive for subsequent auditory memory retrieval according to previous studies reporting that auditory perceptual performance can be strongly affected by task-irrelevant visual stimuli, but not vice versa (visual dominance effect, e.g., Sinnett et al., 2007).

Additionally, previous studies showed that cross-modal semantic congruency could facilitate visual perception performance by reallocating attention resources to target stimuli (Mastroberardino et al., 2015), while attention can also directly modulate the integration efficiency of semantically congruent multisensory stimuli (Talsma et al., 2007; Mozolic et al., 2008). For example, Mastroberardino et al. (2015) reported that semantically congruent audiovisual pairs could positively facilitate subsequent visual Gabor discrimination only when the spatial location of Gabor was congruent with those of previous audiovisual pairs, indicating that cross-modal semantic congruence generates a processing bias associated with the location of congruent pictures by capturing visual attention. For the latter, previous multisensory studies reported that the integration efficiency was restricted when the attention focus was directed toward one modality (called "modal-based attention"; Mozolic et al., 2008) compared to the case of divided attention resources directed toward both modalities. Importantly, some previous studies have further indicated that unisensory behavioral performance differentially benefits from restricted multisensory integration (Mozolic et al., 2008; Thelen et al., 2015). Poorly perceptible unisensory signals, such as auditory signals, can gain more multisensory benefits from task-irrelevant visual signals, but not vice versa. For instance, one study reported that auditory object discrimination could benefit from previous semantically congruent audiovisual pairs with modal-based attention (Thelen et al., 2015). This evidence might indicate that semantically congruent multisensory integration with modal-based attention can also differentially modulate the subsequent unisensory STM performance.

The present study investigated the effect of semantically (in)congruent audiovisual integration on subsequent unisensory STM performance by manipulating the attention focus toward the visual or auditory modality. Participants were asked to selectively focus on one modality while ignoring another task-irrelevant stimulus during multisensory encoding. This method has been widely used in traditional multisensory integration (Yang et al., 2016; Yang et al., 2020) as well as multisensory recognition memory studies (Heikkila et al., 2015; Heikkilä et al., 2017). Considering that the available evidence suggests that perception and cognition processes share an overlapping resource pool, highly efficient perception processing (i.e., multisensory integration) may render more resources available for subsequent cognition performance (i.e., integrated perception-cognition theory; Schneider & Pichora-Fuller, 2000; Frtusova et al., 2013). We hypothesize that both unisensory visual and auditory STM retrieval can benefit from restricted multisensory encoding with semantically congruent relationships. In particular, previous multisensory studies reported that instant auditory discrimination was especially facilitated by the presentation of semantically congruent audiovisual pairs (Thelen et al., 2015). Therefore, similar to exclusively facilitated perceptual auditory discrimination performance, we hypothesized that auditory STM performance might also exclusively benefit from semantically congruent multisensory memory encoding.

## Methods

### Participants

A statistical power analysis in G*Power version 3.1.9.7 (Faul et al., 2007) was performed for sample size estimation. The projected partial $\eta^2$ was determined with reference to a similarly designed two factorial within-subject experiment, and the value was set as 0.1 (Zhang et al., 2021). The two-tailed alpha level was set to 0.05, the power value was set to

0.95, the number of groups was set to 1, and the number of measurements was set to 6. The calculations indicated that a sample size of 16 was required. In particular, to ensure that the example size was the same as that in a previous, very closely related multisensory memory study (Xie et al., 2017), we recruited 34 participants (14 women; age range = 21–34 years; mean age = 26.85 years, SD = 3.17) from campus to participate in this experiment. All participants had normal or corrected-to-normal vision and hearing, were right-handed, were reported to be not have mental illness, and had not participated in a similar experiment previously. Individuals were compensated $10 for their participation. After receiving a full explanation of the experiment and potential risks, all participants provided written informed consent, in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki), and the study protocol was approved by the Ethics Committee of Okayama University, Japan.

## Stimuli and apparatus

Visual stimuli were obtained from the standard set of outlined drawn pictures (Snodgrass & Vanderwart, 1980) with an 8° visual angle. The selected pictures contained an equivalent number of objects from six semantic categories (e.g., animals, tools, instruments, vehicles, dolls, and furniture) and were divided equally among each experimental condition. The auditory stimuli consisted of verbalizations that corresponded to the visual stimuli (the sound of a cat meowing was paired with the picture of a cat). All sound files were downloaded from a website (http://www.findsounds.com) and modified with audio-editing software (Adobe Audition version 5.0) according to the following parameters: 16 bit and 44,100 Hz digitization. Semantically related sounds were delivered binaurally at an intensity of 75 dB. A total of 48 line drawings (six semantic categories × eight stimuli) and 48 matching sounds were used in the task.

The visual stimuli were presented on a 24-in. VG 248 LCD computer monitor with a screen resolution of 1,920 × 1,080 and a refresh rate of 144 Hz (Taiwan, ASUS). The monitor was located 75 cm away from the subjects. Auditory stimuli were delivered binaurally at an intensity of 70 dB via headphones (Sony, MH-1000XM3).

## Experimental design and procedure

The present study evaluated the effects of semantically congruent (cAV) and incongruent (icAV) multisensory encoding on subsequent visual (V) and auditory (A) memory retrieval. The present experiment consisted of a 3 encoding pattern (unimodal, bimodal cAV, and bimodal icAV) × 2 unisensory retrieval modality (V and A) within-subject design. Participants performed a delay-matched task during the six experimental blocks. Half of the blocks evaluated unisensory visual STM retrieval performance under the unimodal encoding condition (V-TestV), bimodal semantically congruent encoding condition (cAV-TestV), and bimodal semantically incongruent encoding condition (icAV-TestV), and the other half of the blocks evaluated unisensory auditory STM retrieval performance under the unimodal encoding condition (A-TestA), bimodal semantically congruent encoding condition (cAV-TestA), and bimodal semantically incongruent encoding condition (icAV-TestA). The six conditions designed in the experiment are depicted in Fig. 1.

The study was conducted in a dimly lit, sound-attenuated, and electrically shielded laboratory room at Okayama University in Japan. In the experimental procedure, taking the cAV-TestV condition as an example, at the beginning of each trial, a white central fixation icon was presented on the screen for 500 ms, and then semantically congruent audiovisual stimuli were presented at the encoding stage for a duration of 600 ms, which was followed a 2,000-ms delay; then, a probe stimulus was presented for 600 ms, followed by a blank screen for 2,400 ms (i.e., within a 3,000-ms time window). During the memory-encoding stage, the participants were asked to selectively focus on the target modality and ignore another task-irrelevant modality stimulus according to different experimental introductions. During the memory retrieval stage, the participants were asked to determine whether the probe stimulus was the same as the target stimulus presented during the memory-encoding stage with a key response (for half of the participants, “yes” and “no” responses corresponded to the "1" and "3" number keys on the keypad, respectively, and for the other half of the participants, “yes” and “no” responses corresponded to the "3" and "1" number keys on the keypad, respectively), with presented and unpresented probe stimuli referenced equally. All visual and auditory stimuli were presented synchronously for 600 ms. The inter-trial interval (ITI) ranged from 1,500 to 3,000 ms. An experimental introduction was presented on the screen before each condition began. The stimulus delivery and behavioral response recordings were controlled using Presentation 0.71 software (Neurobehavioral Systems Inc., Albany, CA, USA). Each participant performed six blocks, and each block included 48 trials: 24 probe stimuli were presented, and 24 probe stimuli were unpresented. The order of the blocks was counterbalanced across the participants. After each block, the participants were asked to rest for 1 min. The completion time of the entire experiment was approximately 1 h.

Before the formal experiment, each participant was required to complete two practice experiments. For the two practice experiments, the stimulus duration time was the same as that in the formal experiment. In the first practice experiment, the participants were asked to fully familiarize themselves with the 48 audiovisual pairs used in the formal experiment. In the second practice experiment, the participants were asked to fully familiarize themselves with the six conditions. Each condition included four trials

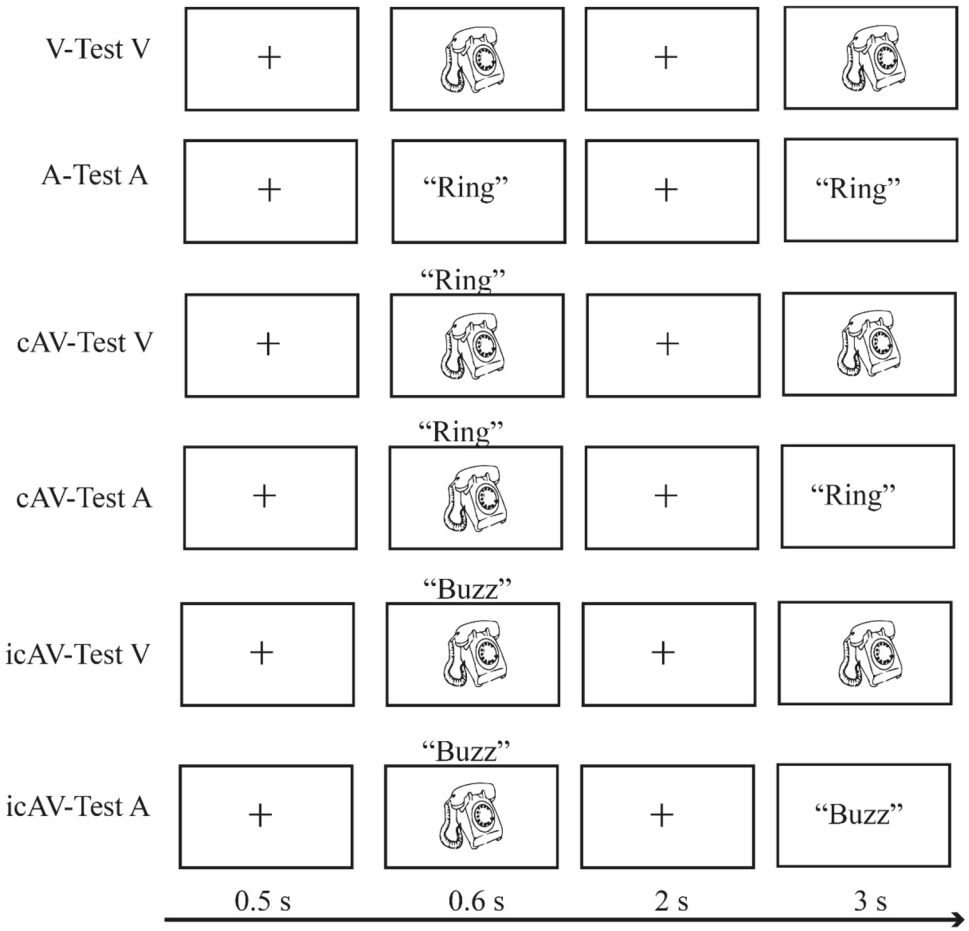


**Fig. 1** Six-block (condition) design of the experiment. In each trial of six blocks, a fixation cross was shown for 500 ms, and then a stimulus (a visual, auditory, or semantically congruent or incongruent audiovisual stimulus) with a duration of 600 ms was presented. A blank screen was shown after a 2,000-ms delay, and finally, a probe stimulus was presented for 600 ms, followed by a blank screen for 2,400 ms (i.e., within a 3,000-ms time window). V-TestV indicates that both the encoding and retrieval stimuli were visual modalities; cAV-TestV indicates that encoding semantically congruent audiovisual stimuli were used, and the retrieval probes were visual stimuli; icAV-TestV indicates that the encoding semantically incongruent audiovisual stimuli and retrieval probes were visual stimuli; A-TestA indicates that both the encoding and retrieval stimuli were auditory modalities; cAV-TestA indicates that encoding semantically congruent audiovisual stimuli were used, and the retrieval probes were auditory stimuli; and icAV-TestA indicates that encoding semantically incongruent audiovisual stimuli were used, and the retrieval probes were auditory stimuli

(i.e., two trials were the same as the previous multisensory presentations, and the other two trials were not the same as the previous multisensory presentations), and correct/error feedback followed each trial. The formal experiment did not begin until the participants understood and could accurately repeat the experimental requirements.

## Results

Accurate response rates (ACRs) and reaction times (RTs) were recorded for the six blocks. Trials with no responses or RTs ± 2 SDs (Ratcliff, 1993) beyond the mean RT were not included in the RT analysis. Additionally, trials with a failure to respond within the 3,000-ms time window were also considered incorrect and removed from further analysis. This resulted in the exclusion of 0.18% of trials for the V-Test V condition, 0.12% of trials for the A-Test A condition, 0.31% of trials for the cAV-Test A condition and 0.06% of trials for the icAV-Test V condition.

The ACRs for visual and auditory STM retrieval performance reached a ceiling in all encoding patterns (above 95%). A 3 encoding pattern (unimodal, bimodal cAV, and bimodal icAV) × 2 unisensory retrieval modality (V and A) repeated-measures analysis of variance (ANOVA) was conducted, and no significant main effect of the encoding pattern, with $F(1,33) = 0.78$, $p = 0.46$, and $\eta^2 = 0.02$, or unisensory retrieval modality, with $F(1,33) = 2.56$, $p = 0.12$,

**Table 1** Reaction time (RT) and accurate response rate (ACR) results for the six blocks of the experiment

| Block | Encoding | Test | RTs (mean ± SD ms) | ACRs (mean ± SD %) |
|---|---|---|---|---|
| 1 | V | V | 523 ± 76 | 96.5 ± 4.4 |
| 2 | cAV | V | 511 ± 67 | 96.6 ± 3.4 |
| 3 | icAV | V | 516 ± 68 | 97.6 ± 2.5 |
| 4 | A | A | 604 ± 103 | 97.1 ± 2.8 |
| 5 | cAV | A | 587 ± 98 | 96 ± 3.8 |
| 6 | icAV | A | 612 ± 105 | 95.8 ± 5.2 |

and $\eta^2 = 0.07$, was observed. Additionally, no significant interaction between the encoding pattern and unisensory retrieval modality was observed, with $F(1,33) = 2.64$, $p = 0.08$, and $\eta^2 = 0.07$. The details of the ACRs and RTs are shown in Table 1.

*SD* standard deviation, *V* visual, *A* auditory, *cAV* semantically congruent audiovisual, *icAV* semantically incongruent audiovisual

For the mean correct-response RT data, a 3 encoding pattern (unimodal, bimodal cAV, and bimodal icAV) × 2 unisensory retrieval modal (V and A) repeated-measures ANOVA was conducted, revealing a significant main effect of the encoding pattern, with $F(1,33) = 60.83$, $p < 0.001$, and $\eta^2 = 0.65$. The post hoc comparison results showed that unimodal encoding was faster than cAV encoding ($p < 0.001$) and icAV encoding ($p < 0.001$), and the RTs for cAV encoding stimuli were faster than those for icAV encoding stimuli ($p < 0.001$). The main effect of the unisensory retrieval modality was significant, with $F(1,33) = 43.73$, $p < 0.001$, and $\eta^2 = 0.57$, indicating that the STM retrieval speed was faster for the unisensory visual (542 ms) modality than for the unisensory auditory (576 ms) modality. Crucially, the interaction between the encoding pattern and unisensory retrieval modality was significant, with $F(1,33) = 37.42$, $p < 0.001$, and $\eta^2 = 0.53$. A subsequent paired t-test comparison with Bonferroni correction revealed that the unisensory visual STM retrieval RTs for bimodal cAV encoding were faster than those for unimodal encoding ($t = 2.0$, $p < 0.05$, $d = 0.17$) but not those for bimodal icAV ($t = -0.95$, $p = 0.35$, $d = 0.07$) encoding. Additionally, unisensory auditory STM retrieval RTs for the bimodal cAV were faster than those for the unimodal ($t = 2.12$, $p < 0.04$, $d = 0.17$) and bimodal icAV ($t = -2.59$, $p < 0.01$, $d = 0.25$) encoding conditions. Additionally, we compared the differences between unisensory visual and auditory STM retrieval under three different encoding patterns using a paired *t*-test, and the results revealed significant differences for the unimodal ($t = -7.64$, $p < 0.001$, and $d = 0.9$), cAV ($t = -7.6$, $p < 0.001$, and $d = 0.91$) and icAV ($t = -8.53$, $p < 0.001$, and $d = 1.1$) encoding conditions Fig. 2.

## Discussion

The present study aimed to investigate the impact of semantically multisensory integration with modal-based attention on subsequent unisensory STM retrieval performance. The RT results showed significantly faster visual STM retrievals than auditory STM retrievals under all encoding conditions, indicating that the visual modality played a dominant role in multisensory representation. Importantly, this study produced two novel findings. First, our results indicated that not only visual but also auditory STM retrieval was accelerated by semantically congruent multisensory STM encoding compared to unisensory STM encoding. More importantly, compared to visual STM retrieval, we found that only auditory STM retrieval performance exclusively benefited from semantically congruent rather than incongruent multisensory encoding.

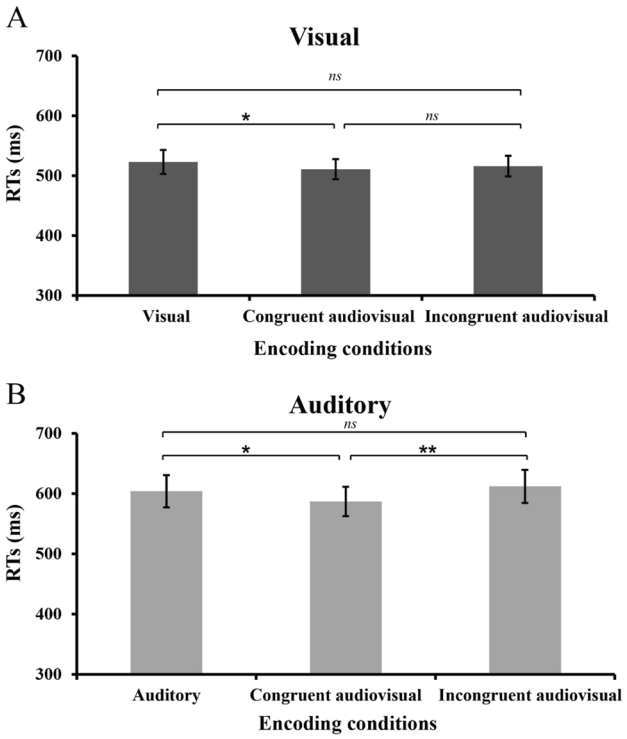


**Fig. 2** Mean reaction times (RTs) for unisensory visual (**a**) and auditory (**b**) short-term memory under the unimodal (visual or auditory), bimodal congruent audiovisual (cAV) and bimodal incongruent audiovisual memory-encoding conditions. The error bars represent 95% within-subject confidence intervals (Baguley, 2012). *$p$ <0.05. **$p$ <0.01

## General facilitation effect of restricted multisensory integration on subsequent unisensory STM retrieval performance

The facilitation effect of bimodal presentation (e.g., audiovisual pairs vs. visual-only) on subsequent visual STM recognition precision has been demonstrated in previous multisensory STM studies (Aizenman et al., 2018; Bigelow & Poremba, 2016). Experimental evidence suggested that visual recognition precision was improved by coherent multisensory representations constructed during semantically congruent multisensory memory encoding (Almadori et al., 2021). According to memory strengthening theory, ACRs are a useful index for evaluating the recognition content precision facilitated by previously constructed representations, while RTs are used to evaluate retrieval speeds; in other words, both ACRs and RTs are measures of the strength of information storage in memory (Kahana & Loftus, 1999). In particular, Kahana and Loftus (1999) suggested that researchers should consider RTs when ACRs reach the ceiling because higher ACRs cannot sufficiently account for memory representation strength. The present study failed to find an ACRs difference between semantically (in)congruent multisensory integration with modal-based attention during the encoding stage of STM; however, the results showed that both unisensory visual and auditory STM retrieval speeds were accelerated by restricted multisensory integration and suggested that both unisensory memory retrieval were generally facilitated by coherent multisensory representations, as long as the unisensory component belonged to the coherent multisensory representation. This explanation might also support the opinion that memory retrieval is closely associated with memory trace redintegration mechanisms, in which unisensory visual or auditory memory retrieval can reactivate prior whole multisensory memory traces (Moran et al., 2013).

Importantly, modality-based attention can ensure that task-relevant modality information is prioritized for multisensory memory encoding and that task-irrelevant modality distractors are filtered (Downing, 2000; Myers et al., 2017). Such selective multisensory memory encoding might facilitate coherent multisensory representation formation to some degree. It must be noted that some previous multisensory perception studies also indicated that coherent multisensory representation formation can be facilitated by modal-based attention (Xi et al., 2019). Evidence suggests that coherent multisensory representation formation is especially facilitated when modal-based attention is engaged in semantically congruent multisensory integration. Moreover, imaging has indicated that the anterior temporal lobe (ATL) might act as a central hub, linking the cortical networks that respond to top-down selective attention and semantically congruent multisensory integration (Kowialiewski et al., 2020; Lee et al., 2017; Ralph et al., 2017). In particular, a more recent multisensory STM study also indicated that the successful retrieval STM information is a function of attentional prioritization at the encoding stage, and coherent multisensory representation formation was facilitated by cross-modal semantic congruency with modal-based attention (Almadori et al., 2021).

Additionally, the results in this study were similar to closely related multisensory recognition memory studies, in which prior semantically congruent multisensory presentation improved subsequent unisensory recognition precision (Heikkila et al., 2015; Thelen et al., 2015). These studies support the conceptual short-term memory model provided by Potter (1976) and suggest that semantically congruent audiovisual stimuli can facilitate rapidly accessing the corresponding concept from the long-term memory network and activate higher-order multisensory memory networks, which can enhance subsequent unisensory recognition precision. The present study partially supports this opinion and suggests that unisensory probes can trigger constructed multisensory representations. In particular, it must be noted that selectively attending to one modality stimulus while ignoring the task-irrelevant modality stimulus during multisensory memory encoding might involve more complex cognitive processing rather than STM, such as WM. In recent multisensory WM studies, Xie et al. (2017, 2019) suggested that the central executive (CE) component of WM plays potential roles in not only allocating attention resources to task-relevant modality stimuli but also integrating semantically congruent information from different subordinate systems into a unified multisensory representation. Unlike the rapidly, unconsciously conceptual accesses in conceptual short-term memory (CSTM), standard WM tends to consciously, selectively allocate attention resources to encode information and influence later cognitive judgment (Potter, 2012). To some degree, this attention operation of memory encoding might explain why some studies suggested that the DMS paradigm was appropriate for investigating STM (Almadori et al., 2021; Bigelow & Poremba, 2016), while other studies suggested that the DMS paradigm was useful for investigating multisensory integration during the encoding stage of WM (Xie et al., 2017; Xie et al., 2021).

A further supplemental experiment was tentatively conducted to investigate the issue that whether semantically congruent audiovisual integration can also facilitate subsequent WM retrieval under three interference condition: non-interference (NI), distraction (DIS), and interruption (INT) (see Fig. S1 and Methods section of the Online Supplementary Material (OSM) as well as Aurtenetxe's study). Overall, the RT results for the INT condition showed a significant negative impact on unisensory WM retrieval compared with the DI and NI conditions (see OSM Fig. S2 and Table S1). In particular, for the INT condition, the RTs results revealed a

significant difference in visual WM retrieval between semantically congruent bimodal memory encoding and unimodal memory encoding. These results were consistent with our formal experiment described in the formal experiment, indicating that semantically congruent bimodal encoding provided an advantage for visual WM retrieval. Additionally, for the INT condition, the RTs revealed a significant difference in auditory WM retrieval between semantically congruent bimodal encoding and incongruent bimodal encoding. Partially consistent with our formal experiment, this result might indicate that a coherent multisensory representation was constructed during the encoding stage of WM, resisted external INT in the maintenance stage, and was then triggered by the less effective auditory probe. This hypothesis might partially support and extend Xie's opinion that CE can not only allocate limited attention resources to special modality stimuli but also integrate semantic congruent information from different channels (Xie et al., 2017), and even resist interference while maintaining a coherent multisensory representation during the maintenance stage. It must be noted that these results were only found under the INT condition, indicating that unisensory WM retrieval might not only depends on the optimal encoding pattern (e.g., bimodal cAV) but also requires adequate executive mechanisms to divide attention between remembered stimuli and interference. In comparing the DI and INT conditions, Hedden and Park (2001) suggested that handling DI during WM requires attentional and inhibitory control mechanisms that facilitate remembering the relevant information and voluntarily inhibiting irrelevant distractors (Hedden & Park, 2001). However, handling INT during WM requires attention-switching abilities that allow attention to be divided between the memory task and the secondary task (Aurtenetxe et al., 2016). Especially, considering the evidence has indicated that visual stimuli play a dominant role in object recognition because they provide more reliable object information (Molholm et al., 2004). We suspect that the auditory interference used in the maintenance stage in supplementary experiment might have been insufficient compared with the visual interference and thereby could not cause enough interference in multisensory representation. Therefore, auditory WM retrieval can also gain more benefits from the coherent multisensory representation. Future work is necessary to further investigate whether faster unisensory memory retrieval (especially concerning the auditory modality) demands the close interaction of multisensory integration in the encoding stage and attention allocation in the maintenance stage.

Overall, we suggest that unisensory STM retrieval performance benefits from the formation of a multisensory representation optimized by modal-based attention constructed during semantically congruent multisensory encoding. When a unisensory probe belongs to an element of multisensory representation, it can rapidly reactivate richer multisensory traces and enhance unisensory STM retrieval performance.

## Auditory STM retrieval exclusively benefited from restricted multisensory encoding

Crucially, the present study found that auditory STM retrieval was exclusively accelerated by a task-irrelevant, semantically congruent picture during memory encoding and impaired when the picture contained incongruent information. This facilitation of specifically auditory memory retrieval was partly consistent with several previous multisensory recognition memory findings. For example, Thelen et al. (2015) compared the effects of semantically congruent and incongruent multisensory presentations on later unisensory recognition and found that semantically congruent multisensory gains for auditory recognition precision were significantly higher (6.35% vs. -11.15%) than those for visual recognition precision (2.35% vs. -3.9%). In addition, Heikkilä et al. (2017) found that *d'* (discrimination ability between old/new objects) was significantly higher for auditory recognition with a picture/written word that carried object-related information than under other conditions. Moreover, Matusz et al. (2017) suggested that semantically congruent audiovisual pairings involving less effective inputs (e.g., auditory stimuli) trigger stronger multisensory processing during memory retrieval. Previous multisensory integration studies reported that less effective unimodal stimuli (i.e., auditory sensory stimuli) yielded larger-magnitude multisensory gains when accompanied by other high-stimulus intensity modal information (i.e., visual sensory information), which is called the "inverse effectiveness principle" (Stein et al., 1994). Typically, such inverse effectiveness principle-induced multisensory perceptual gains in both neuronal responses and behavior have been consistently found to depend on low-level perceptual saliency (Meredith & Stein, 1986; Stein & Meredith, 1993). However, in the present study, the possibility that auditory STM retrieval was improved by a salient visual stimulus cannot explain why auditory STM retrieval was not equally improved by a semantically incongruent visual stimulus. Thus, we tentatively suggested that semantic congruency was involved in visual-induced auditory inverse facilitation. This hypothesis was supported by a recent multisensory study suggesting that inverse effectiveness enhancement can be modulated by low-level stimulus association (e.g., spatial alignment and temporal synchrony) and high-level semantic congruency (van de Rijt et al., 2019). Thus, a less effective auditory stimulus might trigger a more multisensory process due to visual-induced auditory verse facilitation during memory retrieval.

Additionally, it must be noted that modal-based attention might play a positive role in coherent multisensory

representation formation. In the present study, under the cAV-TestA condition, participants were asked to pay attention to auditory stimuli while ignoring visual stimuli during multisensory memory encoding. However, visual sensory processing is more suitable for processing object-related information because pictures can provide richer, more reliable information than auditory sensory processing (Molholm et al., 2004; Molholm et al., 2007). Thus, the effect of task-irrelevant visual information on auditory memory encoding cannot be fully ignored. Schmid et al. (2011) explored the interaction mechanism between cross-modal competition and modal-based attention using fMRI measurements and found a significant visual dominance advantage only when attention was focused on the auditory modality. The authors suggested that cross-modal competition was modulated by modal-based attention and that poor auditory encoding could receive more redundant information compensation from a visual stimulus that was not the attention focus. This poor modality encoding compensation mechanism might reflect the flexible recognition necessary for the external environment. Thus, it is reasonable to assume that a coherent, robust multisensory representation was constructed during memory encoding because of task irrelevance, but semantically congruent visual stimuli provide more redundant information. Santangelo et al. (2015) suggested that memory representation formation could be modulated by low-level external (e.g., stimulus saliency) and high-level internal factors (e.g., conception and matching between complex scenes and objects). Importantly, context-incongruent visual information can capture attention resources, in turn increasing the probability of encoding this context-incongruent visual information into WM. Similarly, in the present study, a congruent, task-irrelevant visual stimulus also captured more attention resources for coherent multisensory representation formation. In contrast, when the task-irrelevant visual signal contained incongruent information, it also captured more attention, leading to strong semantic conflicts with auditory signals and failure to construct a coherent multisensory representation. This hypothesis might be partly supported by the predictive coding model (Friston, 2010), which suggests that stochastic models (i.e., representation) of the environment exist in the brain and can be continuously updated based on ongoing sensory information processing. In particular, semantically congruent multisensory stimuli can result in a stochastic model receiving consistent information and accelerate the information feedback for low-level areas. Stochastic internal models will be updated if top-down prediction conflicts with external incongruent semantic information, thereby leading to poor behavioral performance (Talsma, 2015).

In the present study, for the multisensory encoding stage, we suggested that although attention was selectively directed toward a less effective auditory modality, task-irrelevant but semantically congruent visual images produced a strongly cross-modal competition effect, which means that semantically congruent pictures that are not the attention focus can also provide more redundant information for auditory encoding and subsequently lead to a robust multisensory representation. When one less effective auditory probe was associated with previous robust multisensory representation, robust multisensory representation-related cortical networks could be rapidly triggered for the auditory STM retrieval process (i.e., even auditory WM retrieval, see OSM Fig. S2). However, for semantically incongruent multisensory encoding, coherent multisensory representation formation during the memory-encoding stage is strongly disturbed by a mismatching picture; thus, auditory STM retrieval cannot activate a coherent representation, leading to poor performance.

## Conclusion

In summary, we suggested that coherent multisensory representation formation might be optimized by semantically congruent multisensory integration with modal-based attention in memory encoding and can be rapidly triggered by subsequent unisensory memory retrieval demands. For exclusively accelerated auditory STM retrieval, we suggested that coherent multisensory representation formation is strengthened by a semantically congruent visual stimulus that is not the attention focus during the memory-encoding stage. During the memory retrieval stage, a less effective auditory stimulus can trigger optimized multisensory representation, thereby facilitating rapid memory retrieval processing.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13414-021-02437-4.

**Acknowledgments** This study was partially supported by the Japan Society for the Promotion of Science (JSPS) Kakenhi grant numbers 18K18835, 18H01411, 18K12149, 19KK0099, 20K04381 and 20K07722 and National Natural Science Foundation of China (31700939, 31871092). AW was also supported by the 14th five year plan of Jiangsu Province Education Science (B/2021/01/87), the Humanities and Social Sciences Research Project of Soochow University (22XM0017) and the Interdiscipline Research Team of Humanities and Social Sciences of Soochow University (2022). Additionally, the author gratefully acknowledges the financial support from the China Scholarship Council, No. 201708220080 and Shenzhen Overseas Innovation Team Project(KQTD20180413181834876).

## References

Aizenman, A. M., Gold, J. M., & Sekuler, R. (2018). Multisensory integration in short-term memory: Musicians do rock. *Neuroscience, 389*, 141-151.

Almadori, E., Mastroberardino, S., Botta, F., Brunetti, R., Lupianez, J., Spence, C., & Santangelo, V. (2021). Crossmodal Semantic Congruence Interacts with Object Contextual Consistency in Complex Visual Scenes to Enhance Short-Term Memory Performance. *Brain Sciences, 11*(9).

Aurtenetxe, S., Garcia-Pacios, J., Del Rio, D., Lopez, M. E., Pineda-Pardo, J. A., Marcos, A., Delgado Losada, M. L., Lopez-Frutos, J. M. and Maestu, F. (2016) 'Interference Impacts Working Memory in Mild Cognitive Impairment', *Frontiers in Neuroscience,* 10, pp. 443.

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. Behavior Research Methods, 44(1), 158-175.

Bigelow, J., & Poremba, A. (2016). Audiovisual integration facilitates monkeys' short-term memory. *Animal Cognition, 19*(4), 799-811.

Craik, F. I. M., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, 125(2), 159-180.

Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: how meaning modulates processes of audiovisual integration. *Brain Research, 1242*, 136-150.

Downing, P. E. (2000). Interactions between visual working memory and selective attention. *Psychological Science, 11*(6), 467-473.

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

Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002). Early auditory-visual interactions in human cortex during nonredundant target identification. *Brain Research. Cognitive Brain Research, 14*(1), 20-30.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127-138.

Frtusova, J. B., Winneke, A. H., & Phillips, N. A. (2013). ERP evidence that auditory-visual speech facilitates working memory in younger and older adults. *Psychology and Aging*, 28(2), 481-494.

Hedden, T. and Park, D. (2001) 'Aging and interference in verbal working memory', *Psychology and Aging,* 16(4), pp. 666-81.

Heikkila, J., Alho, K., Hyvonen, H., & Tiippana, K. (2015). Audiovisual semantic congruency during encoding enhances memory performance. *Experimental Psychology, 62*(2), 123-130.

Heikkilä, J., Alho, K., & Tiippana, K. (2017). Semantically Congruent Visual Stimuli Can Improve Auditory Memory. *Multisensory Research, 30*(7-8), 639-651.

Kahana, M., & Loftus, G. (1999). Response time versus accuracy in human memory. In R. J. Sternberg (Ed.), *The nature of cognition* (pp. 323-384). MIT Press.

Kowialiewski, B., Van Calster, L., Attout, L., Phillips, C., & Majerus, S. (2020). Neural Patterns in Linguistic Cortices Discriminate the Content of Verbal Working Memory. *Cerebral Cortex*, 30(5), 2997-3014.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research, 158*(4), 405-414.

Lee, H., Stirnberg, R., Stocker, T., & Axmacher, N. (2017). Audiovisual integration supports face-name associative memory formation. *Cognitive Neuroscience*, 8(4), 177-192.

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Brain Research. Cognitive Brain Research, 24*(2), 326-334.

Liu, J., Zhang, H., Yu, T., Ren, L., Ni, D., Yang, Q., . . . Xue, G. (2021). Transformative neural representations support long-term episodic memory. Science Advances, 7(41), eabg9715.

Mastroberardino, S., Santangelo, V., & Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Frontiers in Integrative Neuroscience, 9*, 45.

Matusz, P. J., Wallace, M. T., & Murray, M. M. (2017). A multisensory perspective on object memory. *Neuropsychologia, 105*, 243-252.

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology, 56*(3), 640-662.

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.

Molholm, S., Martinez, A., Shpaner, M., & Foxe, J. J. (2007). Object-based attention is multisensory: co-activation of an object's representations in ignored sensory modalities. *European Journal of Neuroscience, 26*(2), 499-509.

Moran, Z. D., Bachman, P., Pham, P., Cho, S. H., Cannon, T. D., & Shams, L. (2013). Multisensory encoding improves auditory recognition. *Multisensory Research, 26*(6), 581-592.

Mozolic, J. L., Hugenschmidt, C. E., Peiffer, A. M., & Laurienti, P. J. (2008). Modality-specific selective attention attenuates multisensory integration. *Experimental Brain Research, 184*(1), 39-52.

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. *Trends in Cognitive Sciences, 21*(6), 449-461.

Odegaard, B., Wozny, D. R., & Shams, L. (2016). The effects of selective and divided attention on sensory precision and integration. *Neuroscience Letters, 614*, 24-28.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*(5), 509-522.

Potter, M. C. (2012). Conceptual short term memory in perception and thought. *Frontiers in Psychology, 3*, 113.

Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews. Neuroscience*, 18(1), 42-55.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. Psychological Bulletin, 114(3), 510-532.

Santangelo, V., Di Francesco, S. A., Mastroberardino, S., & Macaluso, E. (2015). Parietal cortex integrates contextual and saliency signals during the encoding of natural scenes in working memory. *Human Brain Mapping, 36*(12), 5003-5017.

Schmid, C., Buchel, C., & Rose, M. (2011). The neural basis of visual dominance in the context of audio-visual object processing. *Neuroimage, 55*(1), 304-311.

Schneider, B. A., & Pichora-Fuller, M. K. (2000). Implications of perceptual deterioration for cognitive aging research. In: Handbook of Aging and Cognition, 2nd edn (eds) F. A. M. Craik and T. A. Salthouse. Mahwah, NJ: Lawrence Erlbaum, 155–219.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: the Colavita effect revisited. *Perception & Psychophysics*, 69(5), 673-686.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 6(2), 174-215.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. The MIT Press.

Stein, B. E., Meredith, M. A., & Wallace, M. T. (1994). Development and neural basis of multisensory integration. *The development of intersensory perception: Comparative perspectives*, 81-105.

Suied, C., Bonneel, N., & Viaud-Delmon, I. (2009). Integration of auditory and visual information in the recognition of realistic objects. *Experimental Brain Research, 194*(1), 91-102.

Talsma, D. (2015). Predictive coding and multisensory integration: an attentional account of the multisensory mind. *Frontiers in Integrative Neuroscience, 9*, 19.

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.

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

van de Rijt, L. P. H., Roye, A., Mylanus, E. A. M., van Opstal, A. J., & van Wanrooij, M. M. (2019). The Principle of Inverse Effectiveness in Audiovisual Speech Perception. *Frontiers in Human Neuroscience, 13*, 335.

Xi, Y., Li, Q., Gao, N., He, S., & Tang, X. (2019). Cortical network underlying audiovisual semantic integration and modulation of attention: An fMRI and graph-based study. *PLoS One, 14*(8), e0221185.

Xie, Y., Xu, Y., Bian, C., & Li, M. (2017). Semantic congruent audiovisual integration during the encoding stage of working memory: an ERP and sLORETA study. *Scientific Reports, 7*(1), 5112.

Xie, Y. J., Li, Y. Y., Xie, B., Xu, Y. Y., & Peng, L. (2019). The neural basis of complex audiovisual objects maintenances in working memory. *Neuropsychologia, 133*, 107189.

Xie, Y., Li, Y., Duan, H., Xu, X., Zhang, W., & Fang, P. (2021). Theta Oscillations and Source Connectivity During Complex Audiovisual Object Encoding in Working Memory. *Frontiers in Human Neuroscience, 15*, 614950. https://doi.org/10.3389/fnhum.2021.614950

Yang, W., Ren, Y., Yang, D. O., Yuan, X., & Wu, J. (2016). The Influence of Selective and Divided Attention on Audiovisual Integration in Children. *Perception*, 45(5), 515-526.

Yang, W., Li, S., Xu, J., Li, Z., Yang, X., & Ren, Y. (2020). Selective and divided attention modulates audiovisual integration in adolescents. *Cognitive Development*, 55, 100922.

Zhang, D., Yu, W., Mo, L., Bi, R., & Lei, Z. (2021). The brain mechanism of explicit and implicit processing of emotional prosodies: An fNIRS study. *Acta Psychologica Sinica, 53*(1), 15.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.