Title of Thesis

Study on Benefits of Semantically Audiovisual Interaction on Unisensory Working Memory

March 2022

Hongtao Yu

Graduate School of Interdisciplinary Science and Engineering

in Health Systems

(Doctor’s Course)

OKAYAMA UNIVERSITY

Abstract

In daily life, humans are often surrounded by stimuli from different sensory modalities (e.g., auditory, visual, tactile). To better master external circumstances, the human brain must integrate different sensory signals to sufficiently perceive the external environment. In particular, approximately 80% of information is derived from auditory and visual channels. The ability to integrate visual and auditory signals into complete and coherent cognition is highly dependent on audiovisual integration (AVI), which not only facilitates instant perception performance, but also enables 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 you write the number down while your friend repeats it aloud, or suppressed if you write it down while your friend makes an irrelevant joke. Although evidence has shown that visual working memory (WM) retrieval can be accelerated by previous semantically congruent AVI, it remains controversial whether faster memory retrieval is contributed to by a coherent multisensory representation or a modality-specific unisensory representation. The multisensory evidence indicates that the formation of a coherent multisensory representation is contributed to by semantically congruent AVI. Early multisensory studies reported that the benefits of coherent multisensory representation for unisensory perception performance were asymmetric. Compared with visual perception performance, less effective auditory perception performance can lead to more multisensory benefits. Some studies have also found that asymmetric perception performance can modulate subsequent cognitive processing. A possible research plan

for exploring this unresolved question is investigating unisensory (i.e., visual and auditory) WM performance in different multisensory memory encoding environments. The main aim of this present thesis was to explore how semantic audiovisual interactions differentially modulate subsequent unisensory WM performance.

For Part 1, we examine whether semantically congruent AVI during the encoding stage of short-term memory (STM) can differentially modulate subsequent unisensory visual and auditory STM performance by applying a delayed matching-to-sample paradigm (DMS). The reaction time (RT) results revealed significantly faster unisensory short-term retrieval performance under the semantically congruent audiovisual encoding condition. The findings of the present study suggest that the formation of coherent multisensory representation 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 short-term retrieval, we assert that the formation of a coherent multisensory representation is strengthened by a semantically congruent visual stimulus that is not the attentional focus during the memory encoding stage. Importantly, during the memory retrieval stage, a less effective auditory stimulus can trigger optimized multisensory representation, thereby facilitating rapid memory retrieval processing. Notably, DMS has been widely used in previous STM and WM studies. To further evaluate the possibility that unisensory memory retrieval was also involved in WM but not limited to STM, we evaluated the reliability of the results under three interference conditions: distractor, interruption, and no interference. For the interruption condition, the RT outcomes showed a significant difference in visual

WM retrieval between semantically congruent bimodal memory encoding and unimodal memory encoding. These findings indicate that semantically congruent bimodal encoding accelerates unisensory STM and WM retrieval.

For Part 2, based on Part 1, we further examine whether the interaction benefits between semantically congruent AVI and top-down attention can further modulate the subsequent unisensory visual and auditory WM retrieval performance. The results reconcile and extend previous multisensory WM studies by demonstrating that a semantically congruent, bimodal presentation with divided-modality attention can accelerate subsequent unisensory WM retrieval, especially less effective auditory WM retrieval. This outcome signals that a sufficient semantically congruent bimodal presentation (e.g., divided-modality attention) not only facilitates immediate behavioral perceptual performance, but can also strongly impact subsequent unisensory WM performance. Moreover, compared with insufficient multisensory integration (e.g., modality-specific selective attention), sufficient multisensory integration (e.g., divided-modality attention) requires more resources for an individual to fully encode and integrate visual and auditory information and maintain a robust multisensory representation, leading to fewer available resources for subsequent unisensory WM retrieval. In particular, we conducted a control experiment to evaluate whether participants remembered the visual or auditory stimulus by using the verbal naming method. In line with our previous experimental outcomes, the results of the control experiment also demonstrated faster auditory memory retrieval under semantically congruent AVI with divided-modality attention, indicating that the verbal naming effect was not an important factor for faster auditory memory retrieval. One possibility we

cautiously suggest is that dividing attentional resources into two modalities might lead to sufficient multisensory integration and then the formation of a robust multisensory representation.

For Part 3, based on Part 2, we further examine whether interaction between semantically congruent AVI and top-down attention can differentially modulate unisensory visual and auditory WM performance by affecting the memory encoding or retrieval stage. The first experiment evaluated whether unisensory WM retrieval benefited from semantically congruent AVI. The findings only point to a weak significant difference for auditory WM retrieval under the semantically congruent and incongruent multisensory retrieval conditions. Then, the second experiment assessed whether unisensory WM retrieval not only benefited from multisensory retrieval benefits, but also from multisensory encoding benefits. For visual WM retrieval, a significantly faster RT was noted when semantically congruent audiovisual pairs were presented during the memory encoding and retrieval stages of WM, indicating that the formation of a coherent multisensory representation was facilitated by semantically congruent audiovisual encoding, and that the visual probe triggered the multisensory representation even under the task-irrelevant, auditory stimulus interference condition. For auditory WM retrieval, faster memory retrieval was only observed in semantically congruent audiovisual encoding conditions. It is reasonable to assume that a coherent, robust multisensory representation was constructed during semantically congruent multisensory memory encoding because of task irrelevance, but semantically congruent visual stimuli provide more redundant information. Then, during the WM retrieval stage, a less effective auditory stimulus can trigger an optimized multisensory

representation and achieve rapid memory retrieval processing.

In sum, first, we found that unisensory WM retrieval (i.e., especially auditory modality) can be accelerated by previous semantically congruent AVI, signaling the possibility that the formation of a coherent multisensory representation was contributed to by semantically congruent AVI. Furthermore, we found that a semantically congruent bimodal presentation with divided-modality attention can accelerate subsequent unisensory WM retrieval, especially less effective auditory WM retrieval, highlighting the possibility that the formation of multisensory representation strongly depends on adequate attentional resources. Finally, we observed that auditory memory retrieval can gain more multisensory benefits from the memory encoding stage but not the retrieval stage, suggesting that the memory retrieval stage may depend more on the extent to which the probe information overlaps with the previously encoded information. In particular, a less effective auditory probe can trigger a coherent multisensory representation and then achieve rapid memory retrieval processing, regardless of whether the semantic information provided by a task-irrelevant visual stimulus is congruent or incongruent.

Key words: Audiovisual integration, Semantic congruency, Unisensory working memory, Top-down attention, Encoding, Retrieval

Table of Contents

Chapter 1 Introduction

Summary

First, this chapter introduces the concept of audiovisual integration (AVI) and its constraints, including low-level spatiotemporal congruency and high-level semantic congruency, as well as its interaction with top-down attention. Second, this chapter examines semantically congruent audiovisual benefits for subsequent memory performance. Third, this chapter delves into the unresolved question between multisensory integration and working memory. Finally, this chapter provides a research framework for exploring the unresolved question: Are auditory and visual information stored in modality-specific, unisensory storage (e.g., separate visual and auditory representation) or central storage (i.e., multisensory representation) in WM?

1.1 Audiovisual Integration

To sufficiently understand external circumstances, the human brain must integrate information from different channels to construct a unified perception. In daily life, multisensory experiences can provide useful information for perceiving the environment. For example, when a person must walk across the street, he or she must notice a car coming and hear the sound of its horn in traffic; such a multisensory experience could facilitate the individual's motor actions to move out of the way of the oncoming vehicle. The ability to integrate visual and auditory stimuli to identify a relevant and salient stimulus is highly dependent on mechanisms of AVI.

AVI involves the cognitive process in which signals derived from visual and auditory sensory systems are integrated into a coherent percept and then lead to higher accuracy [1], faster reaction times (RTs) [2], or greater perception precision [3]. In the last two decades, numerous studies have reported that integration efficiency is modulated by several constraints between different channels such as low-level spatiotemporal congruency [4, 5], high-level semantic relationships, [6] and top-down attention [7]. The facilitation effect of spatiotemporal congruence has been attributed 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 content or top-down attention.

1.1.1 Spatio-temporal congruency

Early multisensory studies report that the near-simultaneous presentation of visual and auditory stimuli in the same location results in multisensory integration of audiovisual stimuli

and then facilitates a behavioral response (e.g., the spatiotemporal rule). The rule of spatiotemporal congruency contains two important factors: spatial congruency and temporal synchrony. For spatial congruency, a famous example is the Ventriloquist effect (VE), which implies visual information influencing the perception of an auditory stimulus based on spatial congruency. During the VE experiment, sound is perceived as originating from a different source due to the perception of a visual stimulus, despite a spatial discrepancy in the sound source and visual stimulus [8]. For temporal synchrony, multisensory studies showed significantly faster RTs for audiovisual stimuli with congruent temporal relationships [9, 10]. A famous example is the Pip and Pop phenomenon whereby a visual object pops out from a complex environment by providing a sound with a synchronous temporal relationship [11]. One possible explanation is that the temporal information of the auditory signal is integrated with the visual signal, generating a relatively salient emergent feature that automatically draws attention. For the neural substrates of spatiotemporal congruency, electrophysiological studies suggest that deep layers of the superior colliculus cortex contain multisensory neurons that multiply their firing rate when two stimuli of different modalities are presented in close spatial and temporal proximity. If two stimuli are temporally asynchronous or spatially incongruent, the firing rates of multisensory neurons are greatly reduced or even inhibited [12, 13].

1.1.2 Semantic congruency

Semantic congruency means that the semantic content of visual and auditory stimuli belongs to one object (e.g., a picture of a cat along with the sound “meow”). Similarly, semantic incongruency indicates that the semantic content belongs to different objects (e.g., a

picture of a cat along with a barking sound). Numerous multisensory studies have reported that semantically congruent bimodal stimuli may produce better behavioral performance (e.g., a faster RT) than unimodal stimuli, and no enhanced effect has been found for semantically incongruent audiovisual stimuli [14, 15]. One possible explanation is the crossmodal semantic congruency involved in memory representation matching, which implies that semantically congruent audiovisual stimuli will gain faster feedback, while semantically conflicting audiovisual stimuli will update the representation and lead to weak behavioral performance (i.e., predictive coding theory, [16, 17]).

Some neuroimaging evidence also provides strong evidence by showing that the cortical network of AVI is closely associated with semantic processing-related cortex networks [18-20]. For example, Zheng et al. (2017) found that both superior temporal regions and the medial prefrontal cortex are involved in the integration of speech and lip movements. In particular, significant activations in the right middle and superior temporal gyri were found when the localization of sound sources was semantically congruent with visual stimuli [21]. Additionally, Beauchamp et al. discovered that the posterior superior temporal sulcus responded more strongly to audiovisual stimuli with congruent semantic relationships than to either auditory or visual stimuli [22].

Additionally, the evidence suggests that crossmodal semantic relevance involves higher-level cognitive processing and has deep interrelationships with attention [23, 24]. For example, some studies have shown that the temporoparietal junction (TPJ) [25] and the anterior temporal lobe (ATL) [18] have been widely considered to play a key role in crossmodal formation of semantic representations. Semantic information interacts within multiple cortical regions, some of which may act as hubs comprising a cortical network related to the stage of semantic integration. Thus, it is necessary to explore how semantic

congruency interacts with attention allocation mechanisms to influence crossmodal integration processing.

1.1.3 The effect of top-down attention on semantically congruent audiovisual integration

Top-down attention refers to the voluntary allocation of attention to special modality stimuli, features, or locations [26-28]. Imagine that you must voluntarily allocate limited attentional resources to finish your test paper within 10 minutes. Attention is not only voluntarily directed, but can also be attracted by a salient, task-irrelevant modality stimulus. For example, even though you must devote limited attentional resources to your test paper, your attention will be drawn by an unexpected alarm. In this thesis, we only focus on top-down, voluntary attention.

Previous multisensory studies have reported that behavioral facilitation effects for semantically associated stimulus components might also be modulated by top-down attention. Different attentional focuses have a differential modulatory effect on semantically congruent or incongruent multisensory perceptions [29, 30]. Evidence indicates that the multisensory benefit is weaker when attention is voluntarily directed toward one modality (i.e., modality-specific selective attention) compared with two modalities (i.e., divided-modality attention). For example, Mozolic et al. (2008) found that perceptual performance regarding semantically congruent multisensory stimuli was enhanced by divided-modality attention compared with modality-specific selective attention. In contrast, for semantically incongruent multisensory stimuli, behavioral decrements were greater for the divided-modality attention condition than for the modality-specific selective attention condition [31].

Additionally, some multisensory studies imply that the formation of a robust multisensory

memory representation also depends on sufficient multisensory integration with divided-modality attention [32, 33]. Research shows that simultaneously attending to two modalities guarantees sufficient resources for multisensory integration. However, attending to one modality can cause a reduced amount of information available in the task-irrelevant modality and can lead to insufficient multisensory integration [34]. Thus, stronger multisensory facilitation under divided-modality attention conditions can also positively influence the formation of an even more robust multisensory representation. Hence, sufficient multisensory integration can contribute to the formation of a robust multisensory representation during semantically congruent multisensory integration with divided-modality attention conditions. Neuroimaging studies also support this view and suggest that the formation of a coherent multisensory representation is especially facilitated when top-down attention is engaged in semantically congruent multisensory integration. For example, the ATL might act as a central hub, linking the cortical networks that respond to top-down attention and semantically congruent multisensory integration [18].

1.2 Working memory

WM is a capacity-limited system that can temporarily store and manipulate information within a short period [35-37]. The concept of WM was first coined by Baddeley and Hitch (1974), who proposed a multiple-component model where information is stored in two domain-specific subsystems (e.g., the phonological loop and the visuospatial sketchpad) that are directed by a supervisory attention system (e.g., the central executive) [38]. The phonological loop is responsible for the short-term maintenance of speech-based and acoustic

items. The visuospatial sketchpad maintains visually and/or spatially encoded items. Some recent studies have found that unisensory information comes from the phonological loop and that the visuospatial sketchpad can be integrated into a single unified representation [39].

1.2.1 The multisensory benefits for working memory

Imagine that you must keep the phone number of a new friend in your mind. The memory encoding process will be facilitated if you write the number down while your friend repeats it aloud, or suppressed if you write it down while your friend makes an irrelevant joke. This scenario indicates the possibility that the memory encoding process can be driven by multisensory integration.

Early research on multisensory WM highlighted a bimodal recall advantage, suggesting that a multisensory representation is more robust and easier to recall [40, 41]. For example, Thompson and Paivio (1994) showed that the recall accuracy of the bimodal presentation condition (i.e., when pictures and sounds are presented together) was significantly higher than that of the unimodal conditions in the context of incidental learning instructions [42]. This result supports the hypothesis that auditory and visual components of audiovisual objects are functionally independent in memory; the work of Thompson and Paivio was also one of the first studies to imply a cognitive advantage of the bimodal format of presentation with respect to the unimodal format. Further, multisensory integration is necessary to form a multisensory memory representation [43]. However, the issue of whether bimodal presentation with (in)congruent semantic relationships can further modulate subsequent WM performance remains poorly understood.

Recently, Xie et al. (2017) reported faster visual WM retrieval in a semantically congruent audiovisual WM encoding condition compared to a unisensory visual-only or auditory-only

WM encoding condition. Further standardized low-resolution brain electromagnetic tomography (sLORETA) outcomes revealed that the posterior parietal cortex (PPC) could play a central executive (CE) role that can integrate the initially processed sensory information from the visual-spatial sketchpad and phonological loop into a unified multisensory representation, leading to faster visual WM retrieval [44]. A more recent functional magnetic resonance imaging (fMRI) study by Xie et al. (2019) found separate brain networks for maintaining semantically congruent and incongruent audiovisual encoding information; for example, the left parietal cortex (e.g., left angular gyrus, supramarginal gyrus, and precuneus) responded exclusively to maintain semantically congruent audiovisual encoding information, while the bilateral angular, left superior parietal lobule, and middle temporal gyrus were exclusively activated while preserving semantically incongruent audiovisual encoding information [45].

In particular, according to the integrated perception-cognition theory developed by Schneider et al. (2000), highly efficient perception processing could leave more resources for subsequent higher-order cognitive function processes. In contrast, devoting too many processing resources to perception may result in insufficient resources being available for subsequent higher-order processing, such as WM [46]. This theory is appropriate for explaining the delayed benefits of bimodal presentation for subsequent memory performance. Frtusova et al. (2013 and 2016) suggested that improved WM performance is related to the degree of audiovisual speech integration [47, 48]. The author supports and extends the integrated perception-cognition theory, suggesting that audiovisual speech integration can efficiently facilitate perceptual processing, thus leaving more resources available for WM processing.

1.2.2 The open question in multisensory working memory studies

Although a bimodal audiovisual encoding advantage has been widely reported in previous multisensory WM studies, it remains an open question: Are auditory and visual information stored in modality-specific, unisensory storage (e.g., separate visual and auditory representation) or central storage (i.e., multisensory representation) in WM? See **Fig. 1**

Some studies support the notion of distinct storage systems for sensory information from separate modalities [42]. In terms of the dual code theory developed by Thompson and Paivio (1994), the bimodal advantage exists because bimodal stimuli are represented by two codes, while unimodal stimuli are represented by a single code [42]. Consequently, the representation of bimodal stimuli would appear to be more robust and also easier to recall than unimodal stimuli. Memory traces reflect all components of past experiences and, in particular, their sensory properties, actions performed on objects in the environment, and people's emotional states. Memory traces are therefore distributed across multiple neuronal systems that code the multiple components of experiences (the Act-In model, Versace, 2014) [49]. Additionally, according to the view of memory trace redintegration, memory retrieval is closely associated with the encoded operation. This means that the unisensory probe can activate the unisensory memory trace, but can also activate other channel memory traces [50].

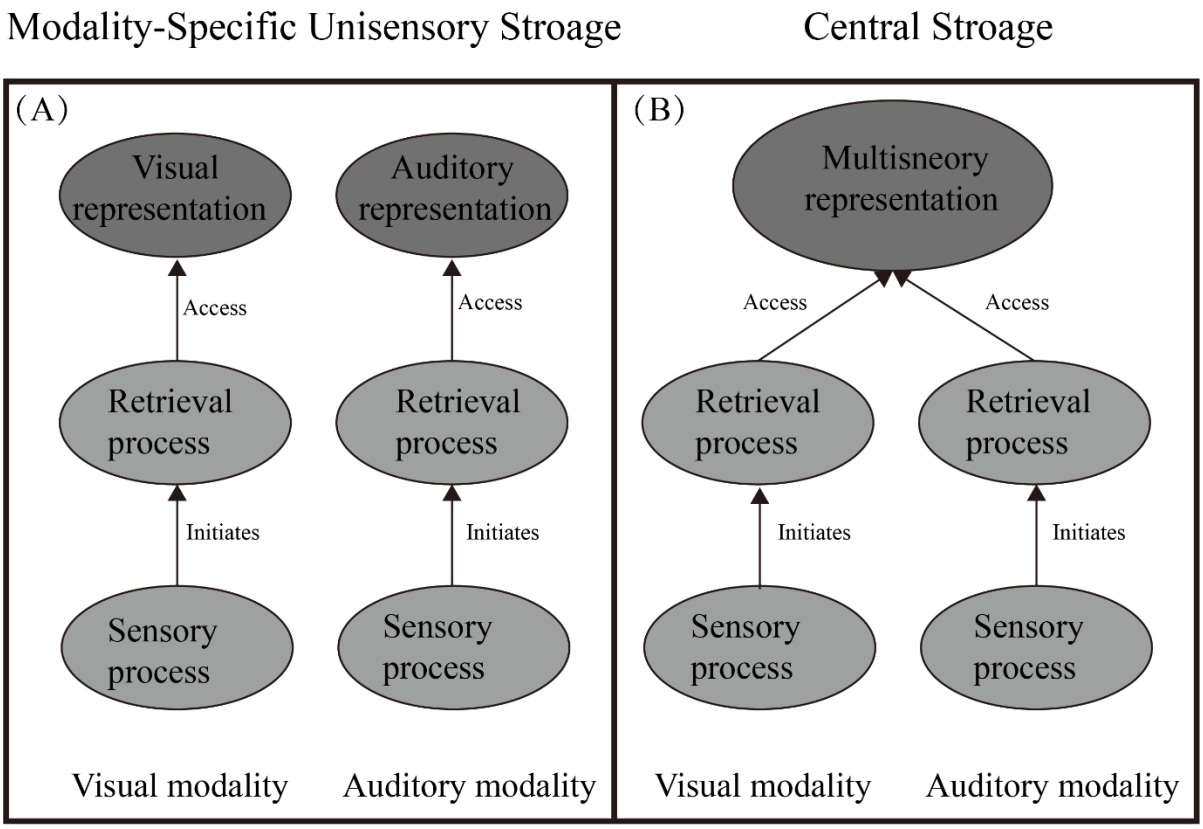

Fig.1 *Two classical opinions for memory representation storage. (A) Modality-specific unisensory storage opinion. Visual and auditory signals are processed at the sensory level during the retrieval stage. Then, visual and auditory probes of memory retrieval can separately trigger their own representation. (B) Central storage opinion. Visual and auditory signals are processed at the sensory level during the retrieval stage. Then, visual and auditory probes of memory retrieval can trigger a coherent multisensory representation.*

Some other studies support the notion of central storage; for example, Xie et al. (2017) investigated the neural substrates of semantically congruent audiovisual WM encoding by using event-related potential (ERP) methods [44]. For the behavioral results, visual WM retrieval speed was accelerated by semantically congruent audiovisual encoding compared with the visual-only encoding condition. The ERP evidence for simultaneous audiovisual

stimuli differed from the ERP for the sum of unisensory constituents during the encoding stage and occurred within a 236–530 ms timeframe over the frontal and parietal-occipital electrodes. The author suggested that the PPC might play a CE role as it not only allocated limited attentional resources for the two modalities, but also integrated the different pieces of sensory information into a single unified multisensory representation. In a recent study, by using fMRI measurements, Xie et al. (2019) obtained results that further support the central storage theory by highlighting a separate cortex network for storing a coherent multisensory representation (e.g., left angular gyrus, supramarginal gyrus, and precuneus), as well as a modality-specific unisensory representation (e.g., bilateral angular, left superior parietal lobule, and left middle temporal gyrus) [45].

1.3 A research framework for investigating the open question

Although previous studies provide possible evidence, central storage theory offers an explanation to account for semantically congruent AVI during the encoding stage of WM. However, faster visual WM retrieval contributed to by semantically congruent multisensory encoding cannot exclude the possibility that a visual probe could trigger a visual and auditory memory representation twice (i.e., modality-specific separate storage) but not a multisensory representation (i.e., central storage).

Perception and cognition processes may share an overlapping resource pool, and highly efficient perception processing (i.e., multisensory integration) may render more resources for subsequent cognitive performance (i.e., integrated perception-cognition theory, [46]). Some multisensory evidence has been reported whereby visual WM retrieval can be accelerated by multisensory integration with congruent semantic relationships. Based on this theory, it is

reasonable to assume that unisensory visual and auditory WM retrieval could also benefit from previous semantically congruent multisensory memory encoding. In particular, prior multisensory studies indicate that instant auditory discrimination is facilitated especially by semantically congruent audiovisual pairs [51]. Thus, similar to exclusively facilitated perceptual auditory discrimination performance, auditory WM performance may also exclusively benefit from a coherent multisensory representation compared to visual WM performance.

Additionally, multisensory evidence has shown that the multisensory benefit is weaker when attention is directed toward one modality compared with two modalities [31]. Furthermore, different attentional focuses have a differential modulatory effect on semantically congruent or incongruent multisensory integration [30]. For example, Mozolic et al. (2008) found that perceptual performance regarding semantically congruent multisensory stimuli was enhanced by divided-modality attention compared with modality-specific selective attention [31]. In contrast, for semantically incongruent multisensory stimuli, behavioral decrements were greater for the divided-modality attention condition than for the modality-specific selective attention condition. Therefore, it is reasonable to assume that attention modulates memory encoding by influencing representation formation, and both unisensory visual and auditory WM retrieval can differentially benefit from previous semantically multisensory integration with top-down selective and divided attention.

Finally, using the N-back paradigm, one study investigated the effect of audiovisual verbal integration benefits on encoding as well as the retrieval stage of WM [52]. The results revealed faster visual and auditory WM retrieval when semantically congruent audiovisual pairs were only presented during the retrieval stage, but not during the encoding stage. The author indicated that although audiovisual semantic congruency facilitates the formation of

multisensory representations, however, unisensory WM retrieval is exclusively benefited from the audiovisual semantic congruency of memory retrieval stage. Similar to the work of Brunetti, it is reasonable to assume that unisensory visual and auditory WM performance also differentially benefit from non-verbal AVI (i.e., pictures with corresponding sounds) during the encoding or retrieval stages of WM.

Thus, according to the content discussed above, we designed a research framework for exploring whether a semantically congruent audiovisual presentation can lead to central or modality-specific unisensory storage. We focused on unisensory visual and auditory WM retrieval performance under three conditions (experiments): (1) audiovisual semantic congruency, (2) the interaction of audiovisual semantic congruency and top-down attention, and (3) the interaction of audiovisual semantic congruency and top-down attention during the encoding or retrieval stages of WM. For Chapter 2, we investigate the benefits of semantically congruent AVI on unisensory WM retrieval, which can provide evidence that unisensory WM retrieval is asymmetric. This study is similar to research on perception evidence, which has reported that perceptual auditory discrimination exclusively benefits from multisensory presentation. If auditory WM retrieval exclusively benefits from AVI, we can tentatively assume that such asymmetric facilitation is caused by a coherent multisensory representation (i.e., central storage theory). Based on Chapter 2, Chapter 3 further focuses on the benefits of semantically congruent AVI with top-down attention on unisensory WM retrieval. The aim of Chapter 3 was to explore the possibility that a coherent multisensory representation can be modulated by attentional focus, and the strength of a multisensory representation can become more deeply associated with subsequent unisensory WM retrieval. Similarly, based on Chapter 3, Chapter 4 investigates whether unisensory WM retrieval is differentially modulated by semantically congruent AVI during the encoding or retrieval

stages of WM. We can further explore whether unisensory WM retrieval gains more benefits from a coherent multisensory representation during the encoding or retrieval stages of WM. For details, see the research framework in **Fig. 2**.

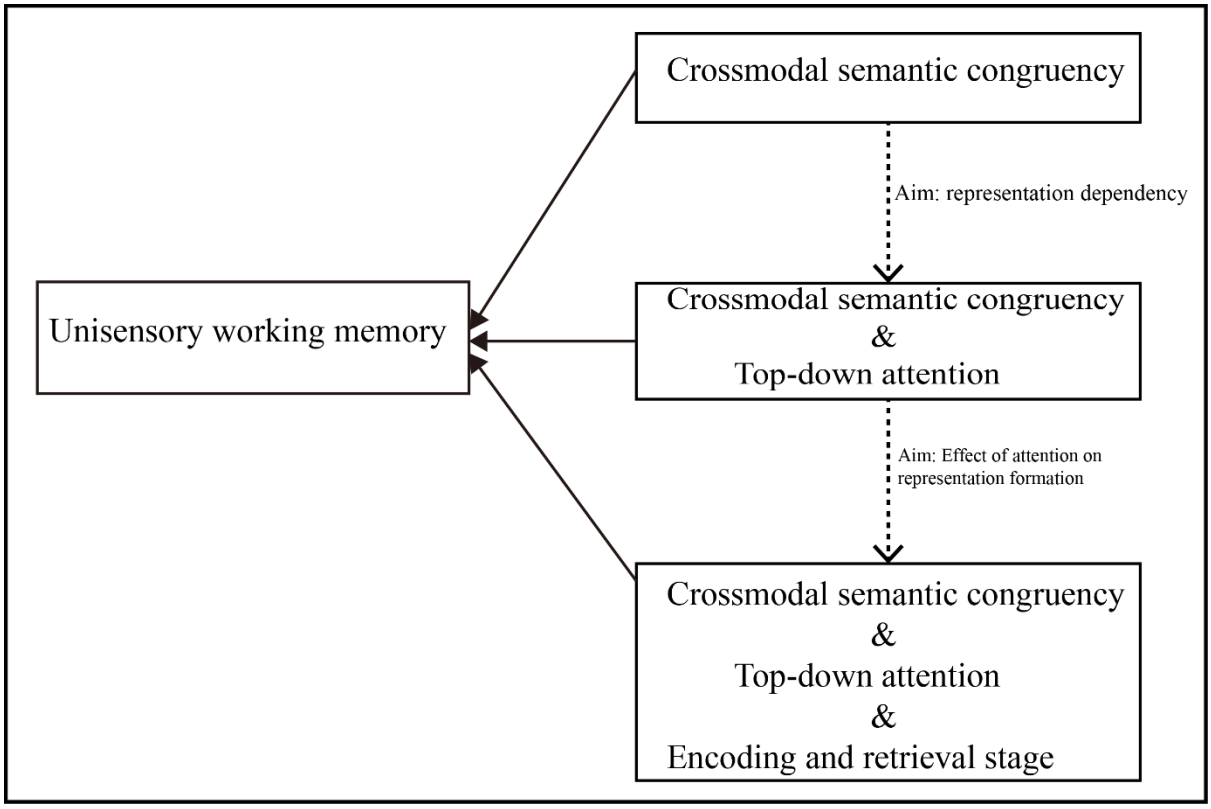

Fig.2 *The research framework of this thesis. First, the present study focuses on the effect of crossmodal audiovisual semantic congruency on unisensory WM retrieval. The aim of this step was to determine whether unisensory memory retrieval is associated with a coherent multisensory representation. Second, the thesis centers on the beneficial effects of attentionally mediated multisensory integration for subsequent unisensory WM retrieval. This aim of this step was to examine the possibility that coherent multisensory formation could be modulated by top-down attention. Finally, this thesis scrutinizes the effect of attentionally mediated multisensory integration during the encoding and retrieval stages of WM. This aim of this step was to establish whether unisensory WM retrieval receives more benefits from an*

attention-optimized multisensory representation during the encoding or retrieval stages.

1.4 The purpose of the present thesis

The main aim of this thesis was to investigate the benefits of semantic audiovisual interactions for subsequent unisensory WM.

Chapter 1 introduces the concept of AVI, WM, and the benefits of AVI for WM. In particular, an unresolved question between AVI and WM is mentioned. A research framework was designed to explore the unresolved question. Finally, the purpose and content of the thesis are briefly described.

Chapter 2 describes how semantically congruent AVI during the encoding stage of short-term memory (STM) can differentially modulate subsequent unisensory visual and auditory WM performance by applying the DMS paradigm, which has been widely used in previous studies on STM and WM. Additionally, we conducted a control experiment to evaluate the possibility that unisensory memory retrieval may be involved in WM, but not limited to STM.

Chapter 3 describes whether the interaction of semantically congruent AVI and top-down attention can further modulate subsequent unisensory visual and auditory WM performance. In particular, we conducted a control experiment to determine whether participants' memory could be affected by the visual or auditory stimulus by using a verbal naming method.

Chapter 4 describes whether the interaction of semantically congruent AVI and top-down attention can differentially modulate unisensory visual and auditory WM performance by affecting the encoding or retrieval stages. The first experiment assessed

whether unisensory WM retrieval benefits from multisensory retrieval but not multisensory encoding. Then, the second experiment evaluated whether unisensory WM retrieval not only benefits from multisensory retrieval, but also from multisensory encoding.

Chapter 5 presents general conclusions based on the findings of these experiments. Future challenges are also described.

Chapter 2 Benefits of Semantically Congruent Audiovisual Integration on the Encoding Stage of Unisensory Working Memory

Summary

Evidence has shown that the multisensory integration benefits of unisensory perception performance are asymmetric, and that auditory perception performance can receive more multisensory benefits, especially when attention is directed toward a task-irrelevant visual stimulus. At present, it remains unclear whether the benefits of semantically (in)congruent multisensory integration with modal-based attention for subsequent unisensory short-term memory (STM) retrieval are also asymmetric. Using a delayed matching-to-sample paradigm, we investigated this issue by manipulating the attentional focus during multisensory memory encoding. The results revealed that both visual and auditory STM retrieval reaction times (RTs) were faster under semantically congruent multisensory conditions than under unisensory memory encoding conditions. We suggest that the formation of a coherent multisensory representation 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. Additionally, a following control experiment indicate the delayed multisensory benefits also facilitate WM retrieval.

2.1 Background

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 [1], faster reaction times [2] or higher perception precision [3]. Previous multisensory studies in animals indicate that integration efficiency is modulated by several constraints between different channels, such as low-level spatiotemporal congruency [4] and high-level semantic relationships [6]. 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 [15]. 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 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 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 working memory resources. The DMS paradigm might be an appropriate paradigm for evaluating short-term memory (STM) and has been widely investigated in recent STM studies [53, 54]. In particular, in previous multisensory memory studies, participants were asked to divide their attention between visual and auditory stimuli during multisensory encoding [44]. 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 [55], 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, [56]).

Additionally, previous studies showed that crossmodal semantic congruency could facilitate visual perception performance by reallocating attention resources to target stimuli [57], while attention can also directly modulate the integration efficiency of semantically congruent multisensory stimuli [2, 31]. 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 crossmodal 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", [31]) 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 [31, 51]. 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 [51]. 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 [32, 33] as well as multisensory recognition memory studies [58, 59]. 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, [46]). 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 [51]. 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.

2.2 Methods

2.2.1 Participants

A statistical power analysis in G*Power version 3.1.9.7 [60] was performed for sample size estimation. The projected partial η^2 was referred to similarly designed two factorial within-subject experiment and then set the value as 0.1[61], 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 would be required. Especially, to ensure the example size was same to a previous, very closely related multisensory memory study [44], we finally 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 the participants had normal or corrected to normal vision and hearing, and were right -handed, were reported being without mental illness, and had not participated in a similar experiment before. 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.

2.2.2 Apparatus and materials

Before the formal experiment, we conducted a pre-experiment to select high-familiarity outline drawings as well as their matching sounds. Ten more subjects participated in the pre-experiment (3 women; mean age = 26.5 years, SD = 1.72). A total of 96 pictures contained an equivalent number of objects from six semantic categories (i.e., animals, tools, instruments, vehicles, dolls, and furniture; see the standard picture set of Snodgrass and Vanderwart, 1980). Similarly, a total of 96 matching sounds were downloaded from a website (http://www.findsounds.com). According to the picture judgment standard provided by Snodgrass and Vanderwart [63], familiarity was defined as the degree to which the object is usual or unusual in your common experience. A 5-point rating scale was adopted in which 1 indicated *very unfamiliar* (or mismatching) and 5 denoted *very familiar* (or well-matched). If the participants did not know what the object was, 1 point was assigned. If they understood the object very well, 5 points were assigned. There was a neutral point, 3, which signaled that the concept of familiarity was located between *familiar* and *unfamiliar*. Only outline drawings with high familiarity scores and a high audio-visual matching degree were used in the following experiments. See **Fig. 3**.

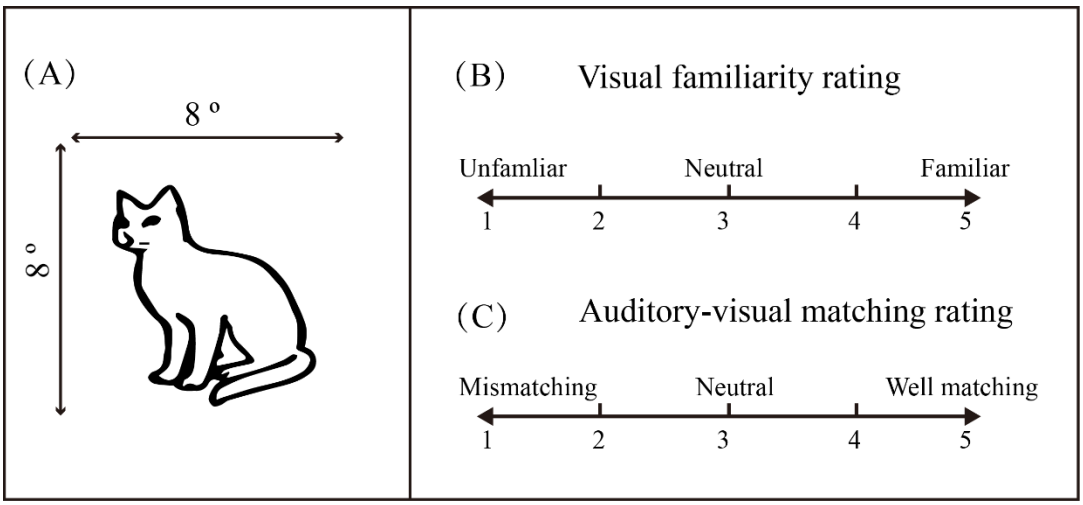

Fig. 3 *(A) Outline drawing of a standard picture set (Snodgrass and Vanderwart, 1980). (B) A 5-point rating test for selecting outline drawings with high visual familiarity. (C) A 5-point rating test for selecting the appropriate audio-visual pairs with a high degree of matching.*

Thus, in the formal experiment, 48 visual stimuli were obtained from the standard set of outlined drawn pictures (i.e., 6 semantic categories × 8 stimulus) with an 8° visual angle. The 48 matching 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 (6 semantic categories×8 stimuli) and 48 matching sounds were used in the task.

The visual stimuli were presented on a 24-inch VG 248 LCD computer monitor with a screen resolution of 1920×1080 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).

2.2.3 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. 4**.

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 2000 ms delay; then, a probe stimulus was presented for 600 ms, followed by a blank screen for 2400 ms (i.e., within a 3000 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 intertrial interval (ITI) ranged from 1500 to 3000 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, California, 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 (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.

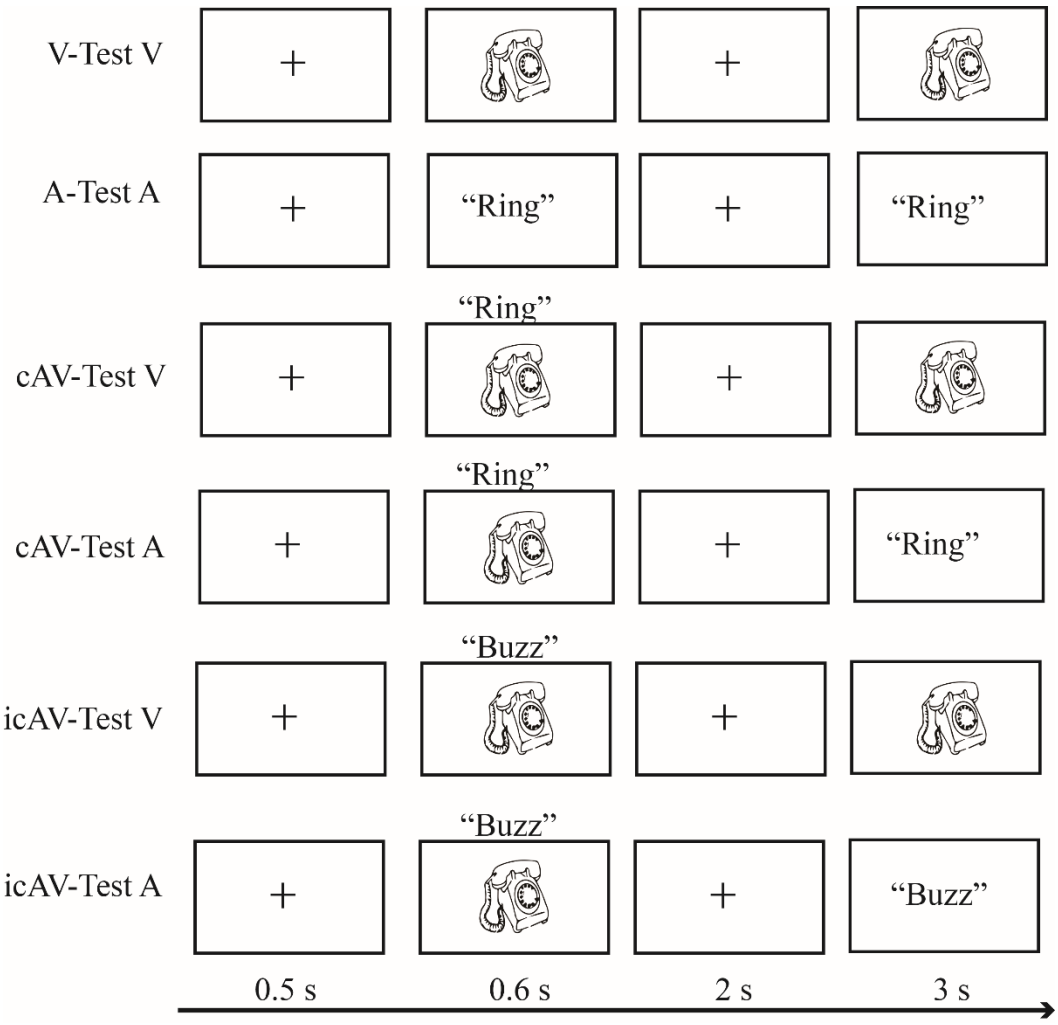

Fig. 4. *Six-block (condition) design of the experiment. In each trial of six blocks, a fixation cross was shown for 500 ms, and then a stimulus (a visual, auditory or semantically congruent or incongruent audiovisual stimulus) with a duration of 600 ms was presented. A blank screen was shown after a 2000 ms delay, and finally, a probe stimulus was presented for 600 ms, followed by a blank screen for 2400 ms (i.e., within a 3000 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.

2.3 Results

Accurate response rates (ACRs) and reaction times (RTs) were recorded for the six blocks. Trials with no responses or RTs ± 2 SDs [63] beyond the mean RT were not included in the RT analysis. Additionally, trials with a failure to respond within the 3000 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 $F(1,33) = 2.56$, $p = 0.12$, 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**.

Table 1. *RT and ACR results for the six blocks of the experiment. Notes: RTs, reaction times; ACRs, accuracy rates; SD, standard deviation; V, visual; A, auditory; cAV, semantically congruent audiovisual; and icAV, semantically incongruent audiovisual.*

Block	**Encoding**	**Test**	**RTs (M ± SD ms)**	**ACRs （M ± 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

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. See the **Fig. 5**.

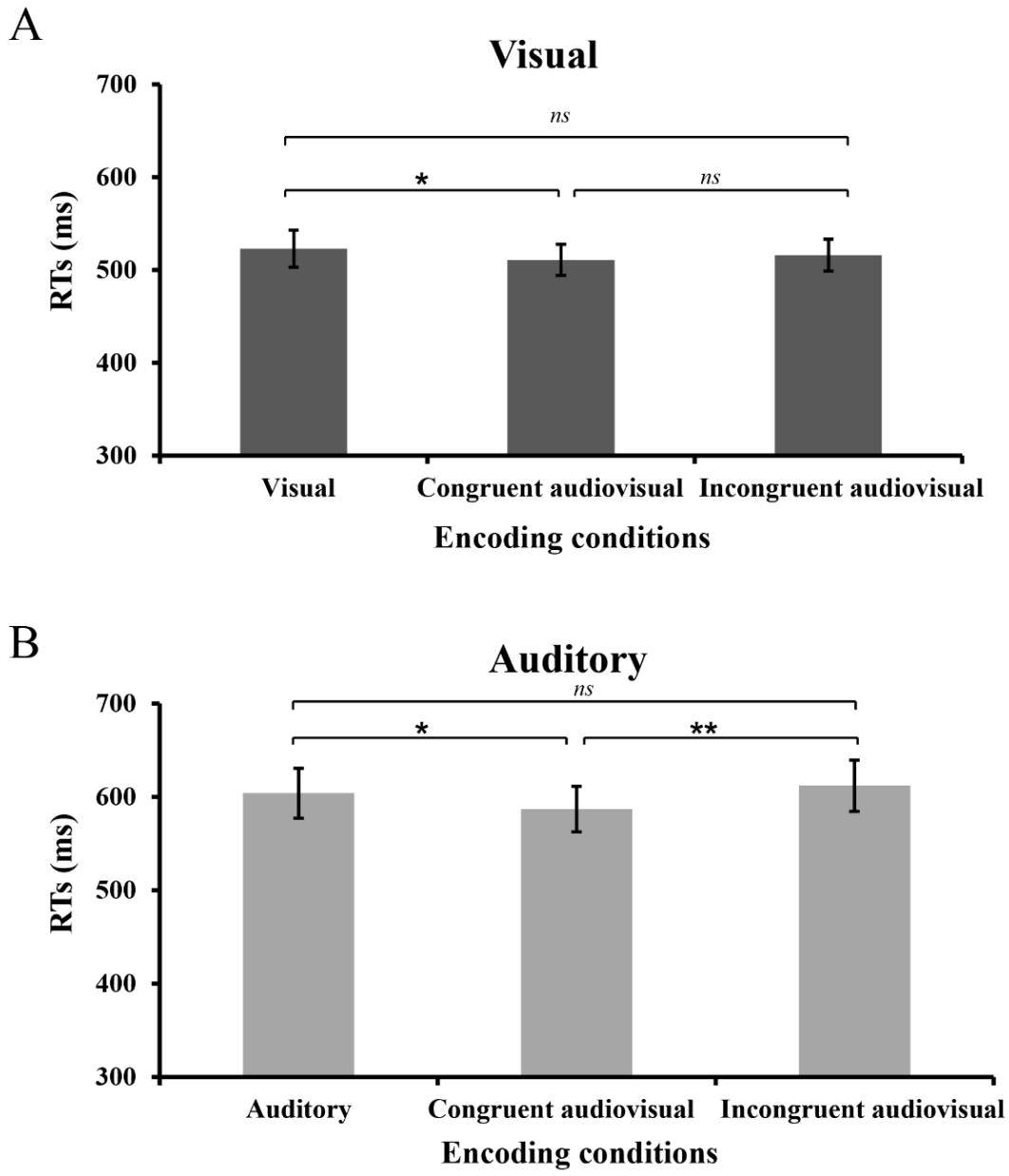

Fig. 5 *Mean RTs for unisensory visual (A) and auditory (B) short-term memory under the unimodal (visual or auditory), bimodal congruent audiovisual (cAV) and bimodal incongruent*

*audiovisual memory encoding conditions. The error bars represent 95% within-subject confidence intervals (Baguley, 2012, [64]). *p <0.05. **p <0.01.*

2.4 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.

2.4.1 General crossmodal semantic congruency benefits for unisensory memory 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 [53, 65]. Experimental evidence suggested that visual recognition precision was improved by coherent multisensory representations constructed during semantically congruent multisensory memory encoding [53]. 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 [66]. In particular, Kahana et al. (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 [50].

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 [67, 68]. 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 [18]. 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 [69-71]. 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 crossmodal semantic congruency with modal-based attention [53].

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 [58, 59]. These studies support the conceptual short-term memory model provided by Potter et al. (1976) [73, 73] 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 working memory. In recent multisensory working memory studies, Xie et al. (2017 & 2019) suggested that the central executive (CE) component of working memory 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 working memory tends to consciously, selectively allocate attention resources to encode information and influence later cognitive judgment [70]. To some degree, this attention operation of memory encoding might explain why some studies suggested that the DMS paradigm was appropriate for investigating STM [53, 65], while other studies suggested that the DMS paradigm was useful for investigating

multisensory integration during the encoding stage of working memory [44, 45]. Future work is necessary to investigate whether faster unisensory memory retrieval can be facilitated by multisensory working memory encoding.

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.

2.4.2 Auditory memory retrieval exclusively benefits from crossmodal semantic congruency

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%) [51]. 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 [59]. 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

[74]. 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" [75]. 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 [76, 77]. 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 [78]. 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 [79, 80]. 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 crossmodal competition and modal-based attention using fMRI measurements and found a significant visual dominance

advantage only when attention was focused on the auditory modality [81]. The authors suggested that crossmodal 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) [82]. Importantly, context-incongruent visual information can capture attention resources, in turn increasing the probability of encoding this context-incongruent visual information into working memory. 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 [16] , 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 [83].

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 crossmodal 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. 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.

2.5 Conclusions

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.

2.6 Control experiment 1: interference effect and working memory

Based on the limitation of Experiment 1, we referenced a similar WM study that investigated the interference effect on subsequent face recognition using the DMS paradigm [84] and designed a supplemental experiment to investigate whether unisensory WM memory retrieval can also benefit from semantically congruent WM encoding under different interference conditions. If the results indicated that unisensory WM retrieval (e.g., especially auditory modality) also benefited from the interference condition, we tentatively hypothesized that a robust, coherent multisensory representation might be constructed during the encoding stage of WM, resist interference in the maintenance stage, and then lead to faster memory retrieval. For the details, please see the following text:

2.7 Methods

2.7.1 Participants

Another 10 students (3 women; age range = 23-28 years; mean age = 25.4 years, SD = 1.78; all right-handed) with normal or corrected-to-normal vision and hearing and no history of mental illness who had not previously participated in our experiment were recruited randomly from campus. 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.

2.7.2 Apparatus and materials

Half of the 48 visual stimuli were the same as those used in the experiment described in the manuscript. The other half of the 48 new visual stimuli (e.g., interference) were also taken from the standard set of outlined drawn figures [4]. These interference-related visual stimuli consisted of 24 non-living and 24 living stimuli. Similarly, another 48 new interference-related auditory stimuli were also obtained from the internet (https://elements.envato.com/). Thus, a total of 96 line drawings and 96 matching sounds were used in the experiment. The stimulus parameter was the same as that in Aurtenetxe's study [84].

2.7.3 Experimental design and procedure

The supplemental experiment followed 3 encoding pattern (unimodal, semantically congruent bimodal and semantically incongruent bimodal) × 3 interference condition (no interference, distractor and interruption) × 2 unisensory retrieval modality (visual and auditory) within-subject design.

The control experiment consisted of three main stages: encoding, maintenance, and recognition. In the encoding phase, a unimodal stimulus (e.g., visual-only or auditory-only) or bimodal stimulus (e.g., audiovisual pair with/without a semantically congruent relationship) was displayed for a 1000 ms period. For the bimodal stimulus, according to the experimental instructions, the participants were asked to focus on one modality stimulus while ignoring another task-irrelevant modality stimulus. In the maintenance period, the participants were instructed to remember the encoded stimulus for a 4000 ms delay period. In the recognition phase, a unimodal stimulus was displayed for 1000 ms. The participants were asked to

determine whether the probe stimulus was the same as the target stimulus presented during the WM encoding stage with a key response (e.g., for "Yes", press the number key 1; for "No", press the number key 3). Please see **Fig. 6**.

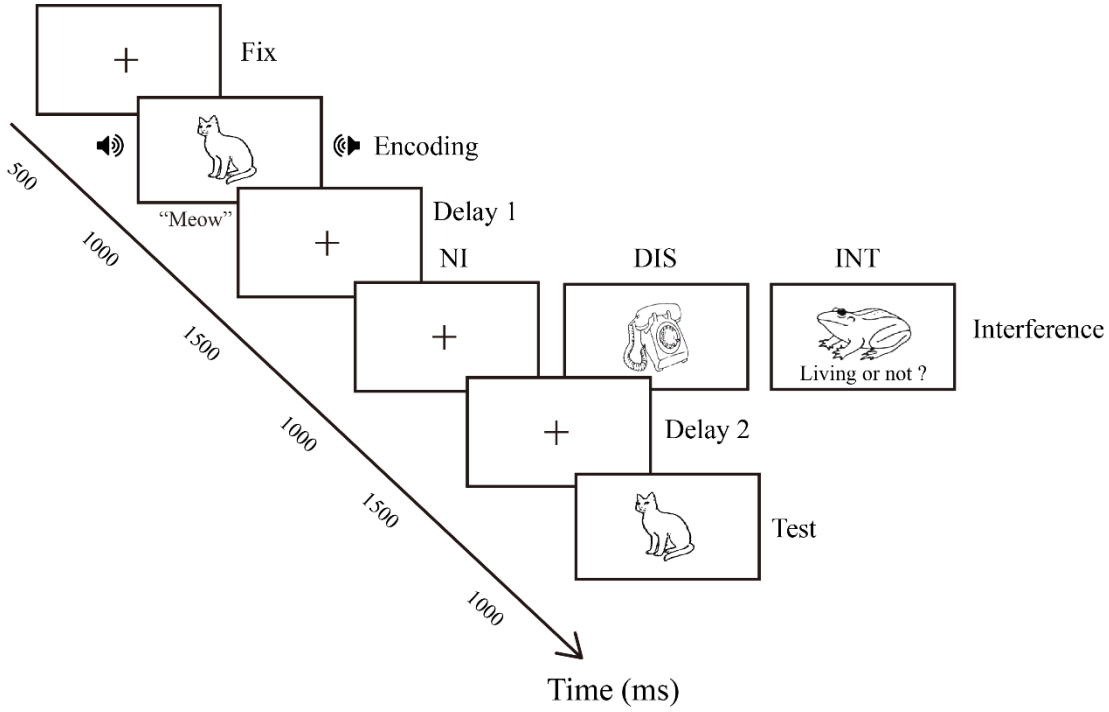

Fig. 6. *Example cAV-Test V under three different interference conditions. The working memory task consisted of three interference conditions: noninterference (NI), distraction (DIS), and interruption (INT). During the encoding stage, the participants were asked to focus on one modality stimulus while ignoring another task-irrelevant modality stimulus. In particular, during the delay stage, the participants were asked to ignore the distractor (DI) or judge whether the drawing belonged to the living category (INT) and to remember the encoded modality stimulus. In the recognition stage, the participants were asked to judge whether the probe was the same as the target stimulus.*

Importantly, the DMS task included three different interference conditions during the maintenance stage: no interference (NI), distraction (DI) and interruption (INT). Under the NI

condition, the participants were instructed to remember the encoded unimodal stimulus during this period, and no interference stimulus was presented during the delay stage. Under the DI condition, a unimodal interference stimulus was presented as a distractor for 1000 ms after the first 1500 ms of the maintenance period. The participants were instructed to ignore the distractor while continuing to remember the encoded unimodal stimulus. Under the INT condition, a unimodal stimulus was presented as an interruption after the encoding phase and was displayed for 1000 ms after the first 1500 ms of the maintenance period. The participants were instructed to press a key (i.e., numpad key "5") if the interference stimulus belonged to the living category. If the interference stimulus belonged to the nonliving category, they did not press any key.

Each participant was separately tested under the six encoding-recognition conditions (e.g., V-Test V, A-Test A, cAV-Test V, cAV-Test A, icAV-Test V and icAV-TestA) with three different interference conditions. Thus, the supplemental experiment consisted of 18 blocks (i.e., 3 encoding patterns × 3 interference conditions × 2 retrieval modalities). Each condition was presented in a block. Each block consisted of 48 randomly presented trials. The block presentation order was counterbalanced across subjects. After each block, the participants were asked to rest for 1 min. After each interference condition, the participants were asked to rest for 10 min. The completion time of the entire experiment was approximately 3 h and 30 min.

2.8 Results

Two paired t-tests were used to separately compare the visual or auditory WM performance under the six encoding-recognition conditions with different interference

conditions.For the accuracy rates (ACRs), the results showed no significant differences between the V-Test V and cAV-Test V (NI, $p = 0.19$; DI, $p = 0.87$; INT, $p = 0.23$), V-Test V and icAV-Test V (NI, $p = 0.17$; DI, $p = 0.33$; INT, $p = 0.50$), and cAV-Test V and icAV-Test V (NI, $p = 0.81$; DI, $p = 0.17$; INT, $p = 0.66$) conditions. Additionally, no significant differences were found between the A-Test A and cAV-Test A (NI, $p = 0.63$; DI, $p = 0.91$; INT, $p = 0.88$), A-Test A and icAV-Test A (NI, $p = 0.14$; DI, $p = 0.07$; INT, $p = 0.59$), and cAV-Test A and icAV-Test A (NI, $p = 0.75$; DI, $p = 0.18$; INT, $p = 0.56$) conditions.

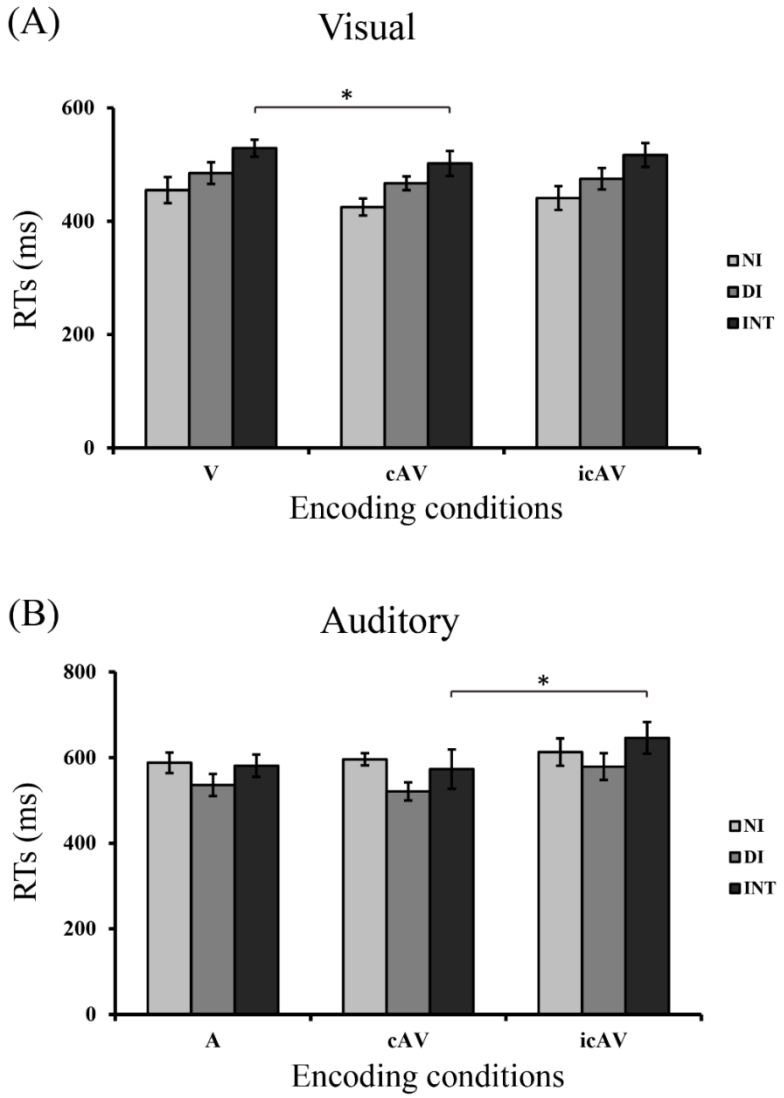

***Fig.** 7 Mean RTs for unisensory visual (A) and auditory (B) memory during the unimodal, bimodal cAV*

*and bimodal icAV encoding conditions with three interference conditions. The error bars represent the SE. *p <0.05.*

For the mean correct reaction times (RTs), only under the INT interference condition did the results reveal two significant differences, which were between the cAV-Test V and V-Test V conditions (p = 0.04) and between the cAV-Test A and icAV-Test A conditions (p = 0.02). Additionally, under the NI interference condition, two weakly significant differences between the cAV-Test V and V-Test V (p = 0.08) and the cAV-Test A and icAV-Test A (p = 0.09) conditions were found. No significant differences were found among any other conditions. For details, see **Fig.7** and **Table 2**.

Table 2. *RT and ACR results under six encoding conditions with different interference conditions*

Conditions	**No interruption (M±SD)**	**Distractor (M±SD)**	**Interference (M±SD)**
V-Test V			
RTs (ms)	455 ± 63	485 ± 49	528 ± 68
ACRs (%)	97.7 ± 1.8	96.3 ± 3.1	96.7 ± 4.1
cAV-Test V			
RTs (ms)	425 ± 48	467 ± 39	502 ± 59
ACRs (%)	95.4 ± 5.1	96.4 ± 2.8	94.8 ± 4.6
icAV-Test V			
RTs (ms)	441 ± 54	475 ± 69	517 ± 67

ACRs (%)	95.8 ± 4.1	96.4 ± 6.4	95.6 ± 3.6
A-Test A			
RTs (ms)	576 ± 63	536 ± 46	581 ± 101
ACRs (%)	97.3 ± 1.8	94.2 ± 4.2	94.8 ± 4.3
cAV-Test A			
RTs (ms)	596 ± 82	527 ± 66	582 ± 101
ACRs (%)	96.7 ± 4.1	94.0±3.8	95.0 ± 3.0
icAV-Test A			
RTs (ms)	613 ± 83	577 ± 146	646 ± 116
ACRs (%)	96.3 ± 2.5	91.3 ± 4.6	95.8 ± 4.4

2.9 Discussion

The RT results for the INT condition showed a significant negative impact on unisensory WM retrieval compared with the DI and NI conditions. In particular, for the INT condition, the RT results revealed a significant difference in visual WM retrieval between semantically congruent bimodal memory encoding and unimodal memory encoding. These results were partially consistent with our formal experiment described in the manuscript, indicating that semantically congruent bimodal encoding provided an advantage for visual memory retrieval. However, the failure of the results to reveal a significant difference in auditory WM retrieval between congruent bimodal encoding and unimodal encoding might indicate that the study

used an insufficient sample size or task. Importantly, for the INT condition, the RTs also revealed a significant difference in auditory WM retrieval between semantically congruent bimodal encoding and incongruent bimodal encoding. Consistent with our previous experiment, this result indicated 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. In particular, it must be noted that these results were only found under the INT condition, indicating that unisensory WM retrieval might depend on not only the encoding pattern but also the attention resources in the maintenance stage. In comparing the DI and INT conditions, Hedden et al. (2001) suggested that handling DI during WM requires attentional and inhibitory control mechanisms that facilitate remembering the relevant information and voluntarily inhibiting irrelevant distractors [85]. However, handling INT during WM requires attention-switching abilities that allow attention to be divided between the memory task and the secondary task [84]. Evidence has indicated that visual stimuli play a dominant role in object recognition because they provide more reliable object information [79]. We suspect that the auditory interference used in the maintenance stage in this 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 provide more multisensory integration benefits. 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, our results might partly support and extend Aurtenetxe's opinion that both visual and auditory WM performance can be affected by interference and that reaction time performance 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. This hypothesis might partially support Xie's opinion that CE may be necessary for semantically congruent multisensory memory encoding [44]. CE can not only allocate limited attention resources to special modality stimuli but also integrate information from different sensory stimuli and even resist interference while maintaining a coherent multisensory representation during the maintenance stage.

Additionally, Kahana et al. (1999) discussed the relationship between accuracy and RT in human memory in detail and suggested that both were useful measures for evaluating multisensory representation in human memory [66]. In particular, RTs provide a useful index for evaluating memory retrieval speed when accuracy reaches the ceiling, as clarified in the following text from Kahana et al. (1999):

"*This is one version of a strength theory of memory—accuracy and IRTs are just two measures of the strength of information stored in memory*" and "*Superficially, it appears that our review of theory and data concerning accuracy and RT in human memory supports the view that these two measures may reflect a single underlying dimension of information.*"

"*In these tasks, people rarely make errors, yet speed may be of the essence. Therefore, to study tasks that are performed essentially without errors, we must consider RTs. It is probably fair to say that almost all RT research is concerned with tasks where error rates are negligible.*"

Additionally, in the experiment, the participants were asked to selectively focus on one modality stimulus while ignoring another task-irrelevant modality stimulus during multisensory memory encoding. This operation was also used in previous, related multisensory recognition memory studies [58, 59].

Chapter 3 Benefits of Semantically Congruent Audiovisual Integration with Top-down Attention on the Encoding Stage of Unisensory Working Memory

Summary

Although previous studies have shown that semantic multisensory integration can be differentially modulated by attention focus, it remains unclear whether attentionally mediated multisensory perceptual facilitation could impact further cognitive performance. Using a delayed matching-to-sample paradigm, the present study investigated the effect of semantically congruent bimodal presentation on subsequent unisensory working memory (WM) performance by manipulating attention focus. The results showed that unisensory WM retrieval was faster in the semantically congruent condition than in the incongruent multisensory encoding condition. However, such a result was only found in the divided-modality attention condition. This result indicates that a robust multisensory representation was constructed during semantically congruent multisensory encoding with divided-modality attention; this representation then accelerated unisensory WM performance, especially auditory WM retrieval. Additionally, overall faster unisensory WM retrieval was observed under the modality-specific selective attention condition compared with the divided-modality condition, indicating that the division of attention to address two modalities demanded more central executive resources to encode and integrate crossmodal information and to maintain a constructed multisensory representation, leaving few resources for WM retrieval. Additionally, the present finding may support the central storage view that WM has

an amodal central storage component that is used to maintain modal-based attention-optimized multisensory representations. Additionally, a following control experiment evaluated the effect of verbal naming effect for result reliability.

3.1 Background

Working memory (WM) is typically considered a capacity-limited system that can temporally store and manipulate information in a short period [86]. WM involves the temporal maintenance of an active representation of external perception information so that it is available for subsequent retrieval processing [87, 88]. Previous WM evidence has demonstrated a bimodal recall advantage and has suggested that multisensory representation is more robust and easier to recall [89]. It must be noted that some evidence has revealed that multisensory integration is necessary to form a multisensory memory representation [43]. However, the issue of whether semantically congruent bimodal presentation can further modulate subsequent WM performance remains poorly understood.

Previous multisensory studies reported enhanced perceptual behavioral performance when visual and auditory stimuli shared common rather than conflicting semantic information [79, 90]. Furthermore, evidence has shown that such semantically congruent bimodal presentation can not only facilitate immediate behavioral perceptual performance but can also accelerate unisensory WM retrieval [44, 45]. For example, Xie et al. (2017) reported faster visual WM retrieval in a semantically congruent audiovisual WM encoding condition compared with a unisensory visual-only or auditory-only WM encoding condition. Further standardized low-resolution brain electromagnetic tomography (sLORETA) results revealed that the posterior parietal cortex (PPC) could play a central executive role that can integrate the

initially processed sensory information from the visual-spatial sketchpad and phonological loop into a unified multisensory representation and then lead to faster visual WM retrieval.

Despite robust evidence showing that the benefits of semantically congruent multisensory WM encoding contribute to faster WM retrieval, it remains unclear whether semantically congruent bimodal presentation with different attention focuses can differentially modulate subsequent unisensory WM retrieval. Previous multisensory evidence has shown that the multisensory benefit is weaker when the attention focus is directed to one modality compared with two modalities [2, 32-33]. Furthermore, a few studies have also found that different attention focuses have a differential modulatory effect on semantically congruent or incongruent multisensory perception [31, 91]. For example, Mozolic et al. (2008) found that perceptual performance regarding semantically congruent multisensory stimuli was enhanced by divided-modality attention compared with modality-specific selective attention. In contrast, for semantically incongruent multisensory stimuli, behavioral decrements were greater for the divided-modality attention condition than for the modality-specific selective attention condition.

Some WM studies have suggested that the formation of a unified multisensory representation is controlled in real time by the central executive, which can selectively allocate limited attention sources to the task-relevant target process while suppressing interference from task-irrelevant distractors or dividedly allocate attention sources to achieve dual-task processing [92-94]. Additionally, previous WM studies widely reported impaired task performance when the central executive must divide attention sources into secondary tasks compared with single-task performance [95, 96]. Such results may indicate that attention source allocation can differentially modulate subsequent WM performance. Additionally, some memory evidence also indicates that attention modulates memory encoding by influencing

representation formation [87, 97, 98]. Considering that multisensory representation formation can be differentially modulated by attention [3, 18], bimodal presentation with different attention modalities during WM encoding may also differentially modulate multisensory representation formation, consequently affecting subsequent unisensory WM performance.

The present study aimed to investigate the issue by adopting a delayed matching-to-sample (DMS) paradigm similar to that of Xie (2017). We manipulated the attention focus (e.g., divided-modality attention and modality-specific selective attention) during the semantically (in)congruent multisensory WM encoding stage and compared the subsequent unisensory visual and auditory WM retrieval reaction times (RTs) and accurate response rates (ACRs). Considering that multisensory enhancement is stronger under divided-modality attention conditions [32, 33], we hypothesized that unisensory WM retrieval may specifically benefit from a robust multisensory representation constructed in semantically congruent multisensory encoding with divided-modality attention.

3.2 Methods

3.2.1 Participants

The group of participants comprised 34 students (13 women; age range = 21–31 years; mean age = 25.5 years, SD = 2.94, all right-handed) randomly recruited from campus, with normal or corrected-to-normal vision and hearing and no history of mental illness. 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.

3.2.2 Apparatus and materials

Visual stimuli were obtained from a standard set of outlined drawing pictures [62] 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 experimental conditions. The auditory stimuli consisted of verbalizations that corresponded to the visual stimuli (e.g., the sound of a cat meowing was paired with a picture of a cat). All of the 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 44100 Hz digitization. Semantically related sounds were delivered binaurally at an intensity level of 75 dB. A total of 48 line drawings (6 semantic categories × 8 stimuli) and 48 matching sounds were used in the experiment. The visual stimuli were presented on a 24-inch VG 248 LCD computer monitor with a screen resolution of 1920 × 1080 and a refresh rate of 144 Hz on a black background (Taiwan, ASUS); the monitor was located 70 cm away from the subjects. Auditory stimuli were delivered binaurally at an intensity level of 75 dB via headphones (Sony, MH-1000XM3).

3.2.3 Experimental design and procedure

The experiment consisted of a 2 attention focus (modality-specific selective attention and divided-modality attention) × 2 semantic congruency (cAV and icAV) × 2 unisensory retrieval modality (V and A) within-subject design. Participants performed a delayed match-to-sample WM task during the four experimental blocks. The first block evaluated the

effect of a semantically congruent bimodal presentation with modality-specific selective attention on subsequent unisensory visual WM (cAV_s-TestV) and auditory WM (cAV_s-TestA) retrieval performance. The second block evaluated the effect of a semantically incongruent bimodal presentation with modality-specific selective attention on subsequent unimodal visual WM ($icAV_s$-TestV) and auditory WM ($icAV_s$-TestA) retrieval performance. The third block evaluated the effect of a semantically congruent bimodal presentation with divided-modality attention on subsequent unimodal visual WM (cAV_d-TestV) and auditory WM (cAV_d-TestA) retrieval performance. The fourth block evaluated the effect of semantically incongruent bimodal presentation with divided-modality attention on subsequent unimodal visual WM ($icAV_d$-TestV) and auditory WM ($icAV_d$-TestA) retrieval performance. The trials of the cAV_s-TestV condition were only presented in the first half of block 1, and the trials of the cAV_s-TestA condition were only presented in the latter half of block 1. Similarly, the stimuli of the $icAV_s$-TestV condition were only presented in the first half of block 2, and the stimuli of the $icAV_s$-TestA condition were only presented in the latter half of block 2. Contrary to the fixed-condition presentation of block 1 and block 2, the trials of the cAV_d-TestV and cAV_d-TestA conditions were intermixed randomly in block 3, and $icAV_d$-TestV and $icAV_d$-TestA trials were intermixed randomly in block 4. Each condition contained 48 trials, 24 in which probe stimuli were presented, and 24 in which probe stimuli were not presented. The order of blocks and the conditions in each block were counterbalanced across participants. The four conditions designed in the experiment are depicted in **Fig. 8 (A)**.

The study was conducted in a dimly lit, sound-attenuated, and electrically shielded laboratory room at Okayama University in Japan. In the modality-specific selective attention modulated multisensory WM encoding blocks (block 1 and block 2), taking the cAV_s-TestV condition as an example, at the beginning of each trial, a white central fixation icon was

presented on the screen for 500 ms. Then, semantically congruent audiovisual stimuli were presented at the encoding stage for a duration of 600 ms, which was followed by a 2000-ms delay and the presentation of a probe stimulus for a duration of 600 ms, with a 3000-ms response limit. During the WM encoding stage, the participants were asked to selectively attend to the target modality and ignore another task-irrelevant modality stimulus according to the particular experimental

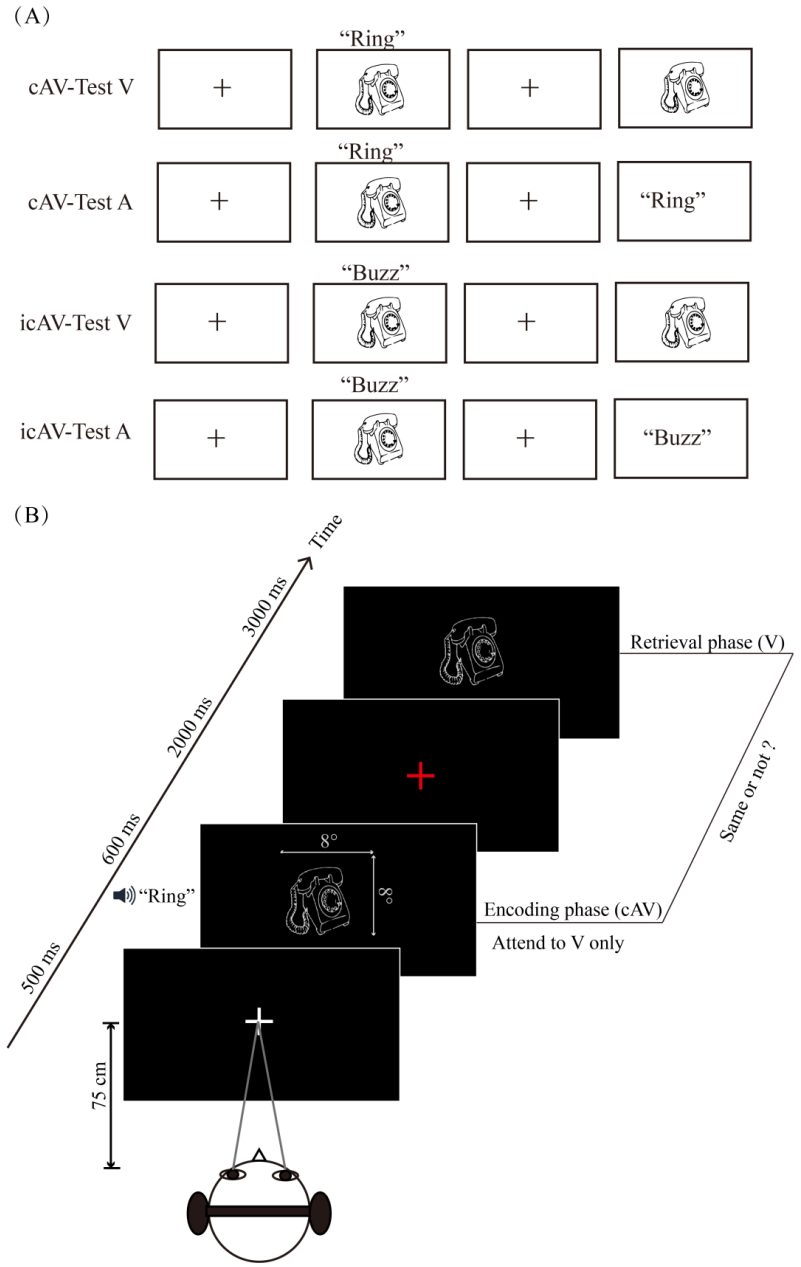

Fig. 8*. Four conditions (A) and an experimental procedure example of a cAV-TestV trial under the modality-specific selective attention condition (B). (A) Four conditions separately evaluating the bimodal presentation (i.e., cAV and icAV) for subsequent unimodal retrieval (i.e., V and A) under modality-specific selective attention and divide-modality attention conditions. Four conditions tested under modality-specific selective attention and divide-modality attention conditions. cAV-TestV means that semantically congruent audiovisual encoding stimuli and retrieval probes were visual stimuli, icAV-TestV means that semantically incongruent audiovisual encoding stimuli and retrieval probes were visual stimuli, cAV-TestA means that semantically congruent audiovisual encoding stimuli and retrieval probes were auditory stimuli, and icAV-TestA means that semantically incongruent audiovisual encoding stimuli and retrieval probes were auditory stimuli. (B) Experimental procedure example of a cAV-TestV trial under modality-specific selective attention conditions. A fixation cross was shown for 500 ms, and a stimulus (semantically congruent or incongruent audiovisual stimulus) with a duration of 600 ms was then presented. A blank screen was shown after a 2000-ms delay, and finally, the probe visual or auditory stimulus was presented for 600 ms within a 3000-ms time limit. The participants were asked to determine whether the probe stimulus was the same as the stimulus presented during the encoding stage.*

instructions (**Fig. 8 (B)**). During the WM retrieval stage, the participants were asked to determine whether the probe stimulus was the same as the target stimulus presented during the WM 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; for the other half of the participants, “yes” and “no” responses corresponded to the "3" and "1" number keys on the keypad, respectively); presented and unpresented probe stimuli were referenced equally. However, in the multisensory WM blocks with divided-modality attention (blocks 3 and 4), participants were asked to attend to both visual and auditory stimuli during the multisensory WM encoding stage because they could not predict whether the following retrieval probe

would be a visual modality or an auditory modality. All visual and auditory stimuli were presented synchronously for 600 ms, followed by a randomized intertrial interval (ITI) ranging from 1500 to 3000 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, California, USA). After each block, participants were asked to rest for 1 min. The completion time for the entire experiment was approximately 1 h and 20 min.

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 that would be used in the formal experiment. In the second practice experiment, the participants were asked to fully familiarize themselves with the four practice conditions under two different attention modalities. Each practice condition consisted of four trials (i.e., the probe stimuli were the same as those in the previous bimodal presentations in two trials and were not the same as those in the previous bimodal presentations in the other two trials), and correct/error feedback followed each trial. The formal experiment did not begin until the participants understood and could accurately repeat the experimental requirements.

3.3 Results

ACRs and RTs were recorded for four blocks. Accuracy rates were calculated as the percentage of correct responses (correct hits and correct rejections). Only RT values associated

with correct responses and within ± 2 SDs were considered for further analysis.

Regarding the ACRs, those for visual and auditory WM retrieval performance reached a ceiling in all encoding patterns (above 94%). A 2 attention focus (modality-specific selective attention and divided-modality attention) ×2 semantic congruency (cAV and icAV) × 2 unisensory retrieval modality (V and A) repeated-measures analysis of variance (ANOVA) was conducted, revealing a significant main effect for attention focus ($F(1, 33) = 14.25$, $p < 0.001$, $\eta^2 = 0.3$) and for unisensory retrieval modality ($F(1, 33) = 14.21$, $p < 0.001$, $\eta^2 = 0.3$). Additionally, only a significant two-way interaction between semantic congruency and unisensory retrieval modality was found ($F(1, 33) = 4.89$, $p = 0.034$, $\eta^2 = 0.13$). For semantic congruency, a post hoc analysis with Bonferroni correction only revealed a significant difference from the auditory modality under the cAV and icAV conditions ($p = 0.049$). This result indicates that semantically congruent multisensory encoding can facilitate subsequent unisensory modality retrieval compared with semantically incongruent multisensory encoding. For the unisensory retrieval modality, a significant difference was only found between the visual and auditory retrieval modalities under the icAV condition ($p < 0.001$). Such a result indicates that object recognition accuracy was visually dominated considering that visual pictures can provide more reliable information for object encoding (Molholm, Ritter, Javitt, & Foxe, 2004; Schmid, Büchel, & Rose, 2010). It must be noted that such a visual encoding advantage was more significant when the semantic contents of the two modalities conflicted, as visual object representations are more robust against the presence of object representations from the auditory domain and vice versa (Schmid, Büchel, & Rose, 2010). Additionally, no significant three-way interaction was found ($p = 0.44$).

Regarding the mean correct RTs, a 2 attention focus (modality-specific selective attention and divided-modality attention) ×2 semantic congruency (cAV and icAV) × 2 unisensory

retrieval modality (V and A) repeated-measures ANOVA was conducted and showed a significant main effect of attention focus ($F(1, 33) = 49.88$, $p < 0.001$, $\eta^2 = 0.6$), demonstrating a faster retrieval response under the modality-specific selective attention condition (500 ms) than under the divided-modality attention condition (561 ms). The results also showed a main effect of semantic congruency ($F(1, 33) = 22.59$, $p < 0.001$, $\eta^2 = 0.41$), with a faster response to cAV stimuli (521 ms) than to icAV stimuli (540 ms). In addition, a significant main effect of the unisensory retrieval modality was also found ($F(1, 33) = 98.32$, $p < 0.001$, $\eta^2 = 0.75$), showing a faster response to the visual retrieval modality (491 ms) than to the auditory retrieval modality (570 ms). A significant two-way interaction between semantic congruency and unisensory retrieval modalities was found ($F(1, 33) = 8.43$, $p = 0.007$, $\eta^2 = 0.2$). Another two-way interaction between attention focus and semantic congruency was found ($F(1, 33) = 9.12$, $p = 0.004$, $\eta^2 = 0.22$). Importantly, the interaction between the three factors was significant ($F(1, 33) = 4.49$, $p = 0.042$, $\eta^2 = 0.12$). Three-way interaction results revealed significant differences in visual or auditory modalities between cAV_s and cAV_d ($p < 0.001$) conditions as well as between cAV_s and $icAV_d$ ($p < 0.001$) conditions.

To evaluate the effect of bimodal presentation with different attention focuses on subsequent unisensory WM retrieval, two separate 2 semantic congruency (cAV and icAV) × 2 unisensory retrieval modality (V and A) repeated-measures ANOVAs were conducted. For the modality-specific selective attention condition, only a significant main effect of retrieval was found ($F(1, 33) = 83.99$, $p < 0.001$, $\eta^2 = 0.72$), indicating significantly faster WM retrieval for the visual modality (461 ms) than for the auditory modality (539 ms). There was no significant interaction between semantic congruency and the retrieval modality ($p = 0.48$). For the divided-modality attention condition, significant main effects of semantic congruency ($F(1, 33) = 20.54$, $p < 0.001$, $\eta^2 = 0.38$) and retrieval modality ($F(1, 33) = 66.95$, $p < 0.001$, $\eta^2 = 0.67$)

were found, indicating significantly faster WM.

Table 3. *RT and ACR results for visual and auditory WM retrieval under different attention-mediated multisensory WM encoding conditions. Notes: RTs, reaction times; ACRs, accuracy rates; SD, standard deviation; V, visual; A, auditory; cAV_s, semantically congruent audiovisual condition with modality-specific selective attention; $icAV_s$, semantically incongruent audiovisual condition with modality-specific selective attention; cAV_d, semantically congruent audiovisual condition with divided-modality attention; and $icAV_d$, semantically incongruent audiovisual condition with divided-modality attention.*

Encoding	Test	RTs (M ± SD ms)	ACRs (M ± SD %)
cAV_s	V	459 ± 86	98.0 ± 2.2
$icAV_s$	V	463 ± 87	98.2 ± 1.8
cAV_d	V	512 ± 113	97.1 ± 2.8
$icAV_d$	V	530 ± 110	97.5 ± 2.9
cAV_s	A	535 ± 121	97.7 ± 2.7
$icAV_s$	A	544 ± 119	97.1 ± 2.3
cAV_d	A	580 ± 155	95.8 ± 4.5
$icAV_d$	A	623 ± 144	94.4 ± 5.2

retrieval under the semantic congruency condition (546 ms) than under the incongruency condition (576 ms) and significantly faster WM retrieval in the visual modality (521 ms) than in the auditory modality (602 ms). Additionally, the interaction between semantic congruency and retrieval modalities was significant ($F(1, 33) = 13.04$, $p < 0.001$, $\eta^2 = 0.28$). For semantic congruency, significant differences between visual and auditory modalities were found under the cAV condition ($p < 0.001$) as well as under the icAV ($p < 0.001$) condition. For the retrieval

modality, significant differences in the visual modality between the cAV and icAV conditions ($p = 0.008$) and in the auditory modality between the cAV and icAV conditions were found ($p < 0.001$). The details of the ACRs and RTs are shown in **Table 3 and Fig. 9 (A).**

To further explore the effect size of attentionally mediated multisensory benefits (e.g., the RTs of the cAV condition minus the RTs of the icAV condition under two different attention conditions) for the unisensory retrieval modality, a 2 attentionally mediated multisensory benefit (modality-specific selective attention and divided-modality attention) × 2 unisensory retrieval modality (V and A) repeated-measures ANOVA was conducted. The results revealed a significant main effect of the attentionally mediated multisensory benefits ($F(1, 33) = 9.12$, $p = 0.005$, $\eta^2 = 0.22$) and of the unisensory retrieval modality ($F(1, 33) = 8.43$, $p = 0.007$, $\eta^2 = 0.2$), with a significant effect size of multisensory benefits under the divided modality

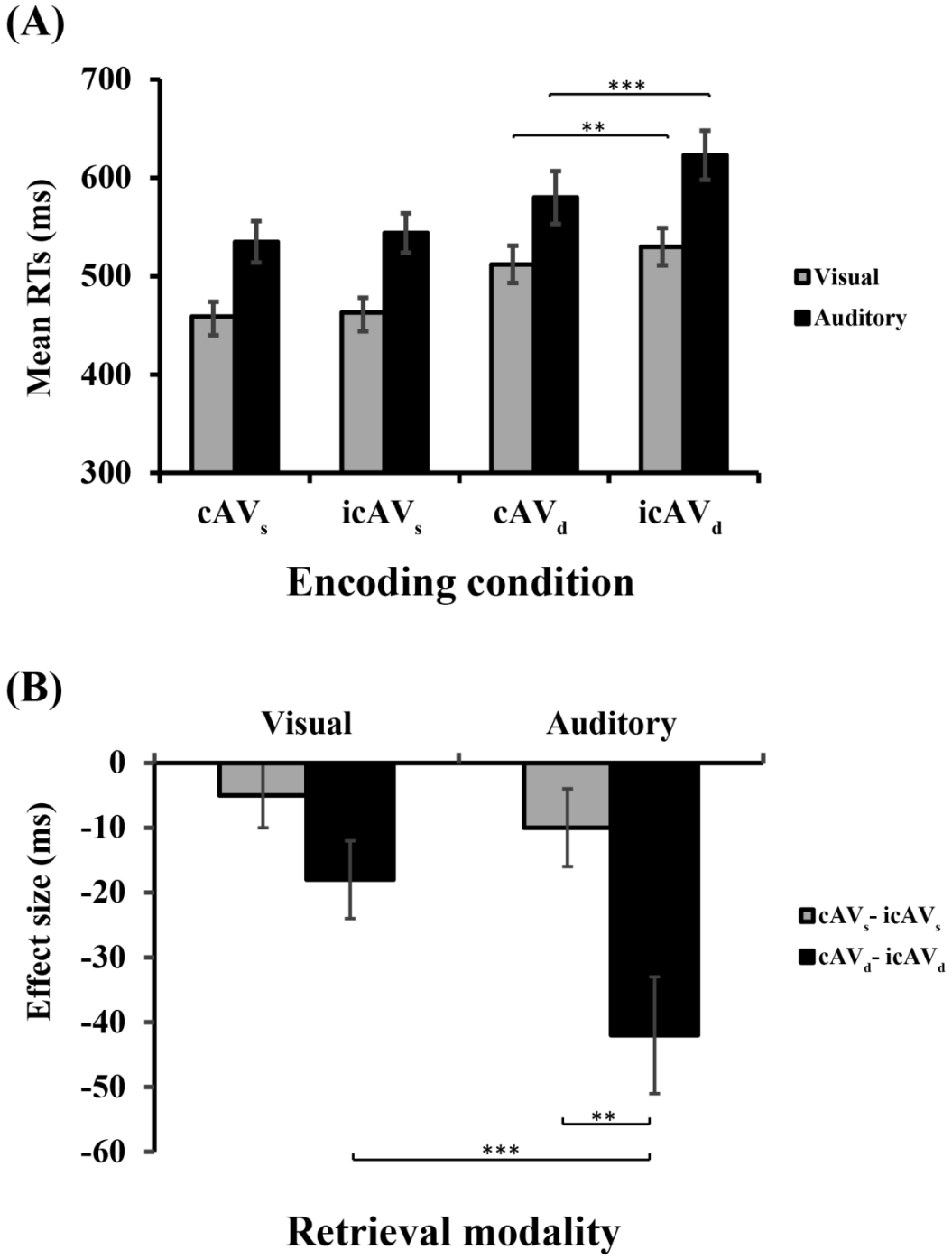

Fig. 9 *Mean RTs (A) and attentionally modulated semantically congruent multisensory benefits of unisensory WM retrieval (B). cAV_s, semantically congruent audiovisual condition with modality-specific selective attention; $icAV_s$, semantically incongruent audiovisual condition with modality-specific selective attention; cAV_d, semantically congruent audiovisual*

*condition with divided-modality attention; and $icAV_d$, semantically incongruent audiovisual condition with divided-modality attention. Error bars denote the SE. *p < 0.05, **p < 0.01, ***p < 0.001.*

condition (30 ms) compared with that under the modality-specific selective attention condition (7 ms) and a larger effect size in the auditory modality (26 ms) than in the visual modality (11 ms). Crucially, the interaction between attentionally mediated multisensory benefits and retrieval modality was significant ($F(1, 33) = 4.89$, $p = 0.042$, $\eta^2 = 0.12$). For the unisensory retrieval modality, a post hoc analysis with Bonferroni correction revealed a significant difference between modality-specific selective attention and divided-modality attention-mediated multisensory benefits in the auditory modality ($p = 0.003$). No significant difference in the visual modality under two different attention-mediated multisensory conditions was observed ($p = 0.086$). Regarding attentionally mediated multisensory benefits, a similar analysis revealed a significant difference between visual and auditory modalities due to the semantically congruent bimodal presentation with divided-modality attention ($p < 0.001$; **Fig. 9 (B))**. No significant difference was found between visual and auditory modalities under the modality-specific selective attention condition ($p = 0.48$). Additionally, two paired t-tests were used to evaluate the unisensory memory retrieval speed of six semantic categories under modality-specific selective attention or divided-modality attention conditions and failed to reveal any significant results among six semantic categories under all the conditions, regardless of whether the attention focus was modality-specific selective attention or bimodal-divided attention.

3.4. Discussion

The present study aimed to evaluate the effects of semantically (in)congruent bimodal presentation with different attention focuses on subsequent unisensory WM retrieval. There were two important results. First, for the bimodal presentation with divided-modality attention, unisensory WM retrieval was faster under semantically congruent than under semantically incongruent conditions. Moreover, auditory memory retrieval exclusively benefitted from semantically congruent multisensory encoding compared with incongruent multisensory encoding. Importantly, such a result was only found in the divided-modality attention condition. Second, overall faster unisensory WM retrieval was found in the modality-specific selective attention condition than in the divided-modality attention condition.

Additionally, the present study revealed overall faster visual WM retrieval than auditory WM retrieval in all conditions, and this result is in accordance with previous multisensory memory studies that suggested a dominant role of the visual modality in multisensory representations [81]. Alternatively, slower auditory memory retrieval may be caused by slower perceptual processing. A previous study reported auditory system decoding acoustic events that unfold over time, and a sequence of objects forms a unified auditory representation [99, 100]. Thus, transforming an auditory stimulus into a perceptual representation may be more costly with respect to time than transforming a visual stimulus. Such a slower auditory perception process may also affect subsequent auditory memory retrieval.

3.4.1 Semantically congruent bimodal presentation with divided-modality attention accelerates unisensory memory retrieval

A multisensory behavioral study reported that perceptual performance for semantically congruent multisensory stimuli was facilitated by divided-modality attention compared with the performance observed in a modality-specific selective attention condition. However, for incongruent stimuli, divided-modality attention can cause greater interference and then degrade performance [31]. This differentially modulated effect of divided-modality attention on semantically congruent bimodal presentation is similar to the present findings that faster (or slower) unisensory WM retrieval was found in semantically congruent (or incongruent) multisensory encoding with divided-modality attention. We tentatively suggest that semantic multisensory facilitation with divided-modality attention could differentially modulate subsequent unisensory WM retrieval by influencing multisensory representation formation during WM encoding. Previous multisensory WM studies indicated that the integration of initially processed visual and auditory information from different slave systems was controlled in real time by the central executive, which can allocate limited attentional resources to the formation of multisensory representations [44, 45, 101]. Crucially, some multisensory studies may indicate that robust multisensory representation formation depends on sufficient multisensory integration with divided-modality attention [7, 44]. The evidence showed that simultaneously attending to two modalities guarantees sufficient resources for multisensory integration. However, attending to one modality can cause a reduced amount of information available in the task-irrelevant modality and can lead to insufficient multisensory integration [32]. Thus, stronger multisensory facilitation under divided-modality attention conditions can also positively influence further robust multisensory representation formation. Thus, in the

present study, it is reasonable to assume that stronger multisensory facilitation contributed to robust multisensory representation formation during semantically congruent bimodal presentation with divided-modality attention conditions and then enhanced subsequent unisensory WM retrieval. However, under the divided-modality attention condition, semantically incongruent stimuli can cause greater semantic conflict and cannot contribute to robust multisensory representation formation. This explanation also partially supports the extended memory redintegration hypothesis that unisensory memory retrieval can reactivate previously constructed multisensory representations [50].

Alternatively, accelerated unisensory WM retrieval can also be interpreted as a faster trigger result of a consistent internal model constructed in the divided-modality attention-mediated semantically congruent multisensory WM encoding condition. According to the theoretical framework of the predictive coding model provided by Friston (2010) [16], stochastic models of the environment exist in the brain and can be continuously updated by constant sensory information. Stochastic internal models will be updated if a top-down prediction conflicts with the external sensory input. Talsma et al. (2015) suggested that semantically congruent bimodal presentations can result in high-order brain areas receiving consistent information; then, these brain areas produce a consistent internal model [83]. In contrast, incongruent multisensory stimuli may update an internal model update, which would in turn produce a weak internal model. Crucially, Talsma et al. (2015) also suggested that attention can boost the precision of predicted sensory input to determine whether the current internal model needs to be updated. Considering that acquiring information in a multisensory representation is an active process, a stimulus of one modality can activate itself and other modalities of information. Thus, in the present study, accelerated unisensory WM retrieval can also be interpreted as a faster trigger result of a robust consistent internal model constructed

during semantically congruent multisensory encoding with divided-modality attention.

Of note, the present study revealed that the semantically congruent multisensory benefits were significantly larger for auditory WM retrieval under the divided-modality attention condition than in other conditions ($p < 0.003$). Although previous studies have suggested that auditory memory performance is inferior to visual memory [51, 102], it must be noted that auditory memory retrieval can accrue more multisensory encoding benefits than visual memory, which has been indicated in some multisensory memory studies [50, 59, 103]. For example, Thelen et al. (2015) reported that semantically congruent and incongruent multisensory gains for auditory recognition memory performance were significantly higher (6.35% vs. -11.15%) than those for visual recognition memory performance (2.35% vs. -3.9%). For such special auditory memory facilitation, some multisensory memory studies have suggested that less effective auditory stimuli can trigger more multisensory benefits [74, 104]. Previous multisensory studies reported an inverse effectiveness relationship between visual and auditory signals in which poorly perceptible unisensory signals demonstrated strong multisensory enhancement if presented with another unisensory signal [12, 13]. A recent study revealed that inverse effectiveness can also play a role at the word level, in which ambiguous words accompanied by matching spoken sound produce greater multisensory integration [78]. This result indicated that the multisensory enhancement of inverse effectiveness was modulated not only by stimulus saliency but also by crossmodal semantic congruency. In the current study, exclusively facilitated auditory WM retrieval may indicate that less effective unisensory stimuli can trigger greater multisensory benefits from a robust multisensory representation constructed under the semantically congruent bimodal presentation with the divided-modality attention condition.

Sufficient multisensory integration may be an important factor in faster unisensory

memory retrieval. However, whether faster unisensory memory retrieval benefited from early perception facilitation or late semantic integration remains unclear. Electrophysiological evidence indicates that faster memory retrieval benefits from later semantic integration [44]. Previous studies have indicated that both early perception facilitation and late semantic integration are two parallel integration patterns in multisensory processing [18]. In the present study, we cannot distinguish whether faster unisensory memory retrieval is facilitated by early perception facilitation, later semantic integration or both. To investigate this unresolved question in future work, a possible method is to evaluate the multisensory encoding of meaningless audiovisual or meaningful audiovisual pairs by using high-temporal-resolution electrophysiology measures.

Additionally, it must be noted that participants could remember the visual or auditory stimulus by using verbal labels during multisensory encoding. For example, participants may be remembering a verbal label ("cat") for the stimuli instead of or in addition to the actual visual and auditory representations. Thus, faster auditory WM retrieval was contributed by the multisensory representation, or the verbal label effect was ambiguous. A possible method to weaken the verbal label effect was increasing the recognition difficulty [105]. For the present study, taking the cAV-Test V condition as an example, during the recognition stage, divided the visual probe into two types: there was a 50% possibility of the probe being the same as the previously presented cat drawing (original type) and a 50% possibility of it being similar to but different from the original cat (novel type). Participants may depend more on the actual visual representation considering the fact that both probe types have the same concept (i.e., cat, see Figure. S1 of Supplementary Material). A supplemental experiment was conducted to evaluate unisensory WM retrieval under weak verbal label conditions (see Supplementary Material for further details). The supplemental experimental results also revealed faster auditory WM

retrieval during semantically congruent multisensory encoding with divided-modality attention (see Figure. S2 and Table. S1 of Supplementary Material). The supplemental results might indicate that verbal labels might not be a critical factor for faster auditory WM retrieval. Importantly, dividing attention resources into semantically congruent visual and auditory modalities might be a critical factor for robust multisensory representation formation considering the fact that dividing limited resources into two modalities can lead to sufficient multisensory integration [7, 45]. Such an opinion might be partly supported by some multisensory memory studies that suggested that recognition memory was contributed by semantically congruent bimodal presentation but not the verbal label effect [58]. For example, Heikkilä et al. (2015) reported that significantly facilitated visual recognition memory performance benefited from semantically congruent bimodal presentation (i.e., pictures with natural sounds) but not unimodal presentation (i.e., pictures with written words). However, participants could easily remember the pictures by using verbal labels when the pictures were paired with written words. Heikkilä et al. (2015) suggested that bimodal presentation of congruent information during encoding contributes to multisensory representation formation.

3.4.2 Faster unisensory memory retrieval was found in the modality-specific selective attention condition but not in the divided-modality attention

Although numerous studies have suggested that multisensory integration is stronger when attentional resources are divided to address stimuli in both modalities compared with the integration of a single specific modality stimulus, sufficient integration also requires more resources. In the present study, overall faster unisensory WM retrieval was found in the

modality-specific selective attention condition than in the divided-modality attention condition. Such results may indicate that multisensory encoding with divided-modality attention, integrating crossmodal information and maintaining multisensory representations have a higher resource cost, leaving fewer resources for subsequent unisensory WM retrieval processing.

Previous studies have proposed that a high WM load interferes with executive control, reducing the capability of the brain to maintain the priorities of stimuli processing demands. Thus, task-irrelevant low-priority distractors would interfere more with the processing of task-relevant stimuli [106, 107]. It must be noted that such interference effects may be more serious under divided attention conditions considering that WM and divided attention share overlapping neural substrates [106, 108, 109]. Santangelo et al. (2013) investigated the neural substrates of WM loads and divided attention using functional magnetic resonance imaging (fMRI) measurements and found increased activity in the intraparietal sulcus (IPS) with higher WM loads, especially when subjects had to divide their attention to monitor multiple objects. The author suggested that WM and divided attention shared a common limited-capacity resource pool and that the IPS may play a modulation role that can not only divide attention to monitor multiple objects but also maintain these objects in WM. Additionally, Xie et al. (2017) suggested that the IPS may play a central executive role in that it can allocate attention to visual and auditory modalities but can also integrate these signals into a unified multisensory representation during the WM encoding stage. Thus, it seems that the IPS may play a central executive role in that it can allocate attentional resources, integrate crossmodal information and maintain multisensory representation. In the present study, slow unisensory WM retrieval under divided-modality attention conditions may reflect higher central executive demands (e.g., a higher WM load) to encode and integrate the crossmodal information from stimuli dividing

one's attention and maintain the multisensory representation, leaving few resources for unisensory WM retrieval.

Additionally, the result cannot exclude the possibility that divided attention to two modalities may weaken multisensory representation formation. Previous studies have indicated that the central executive must monitor the information coming from visual and auditory slave systems. In particular, Santangelo et al. (2013) revealed that working memory and divided attention utilize a common, limited-capacity pool of processing resources in the overlapping brain region (i.e., the left intraparietal sulcus). If too many cognitive resources are devoted to attention control operations, few resources may be available for multisensory representation formation. Such an alternative explanation is in accordance with Mastroberardino's opinion that the central executive is limited by increasing the difficulty of a concurrent task, especially in the central executive, which must allocate more cognitive resources to perform the concurrent task and thus cannot combine information from different sources [43].

Meanwhile, this explanation partially supports the integrated perception-cognition theory suggested by Schneider (2000), which suggested that highly efficient perception processing could leave more resources for subsequent high-order cognitive function processes. In contrast, devoting too many processing resources to perception may result in insufficient resource availability for subsequent higher-order processing, such as WM. Frtusova et al. (2013, 2016) suggested that improved WM performance is related to the degree of audiovisual speech integration. The author supports and further extends the theory, suggesting that audiovisual speech integration can efficiently facilitate perceptual processing, thus leaving more available resources for WM processing. In the present study, sufficient multisensory integration demanded divided, equal attentional resources to process two modalities, whereas insufficient multisensory integration indicated that fewer attentional resources were available to attend to

one specific modality. Although divided-modality attention contributed to sufficient multisensory integration, this attention modality also costs more resources than modality-specific selective attention and leaves fewer available resources for subsequent WM retrieval processing.

3.5. Conclusions

This study investigated the effect of semantic bimodal presentation with different attention focuses during the WM encoding stage on subsequent unisensory WM retrieval. The results reconcile and extend previous multisensory WM studies by demonstrating that semantically congruent bimodal presentation with divided-modality attention can accelerate subsequent unisensory WM retrieval, especially less effective auditory WM retrieval. This result indicated that sufficient semantically congruent bimodal presentation (e.g., divided-modality attention) not only facilitates immediate behavioral perceptual performance but can also strongly impact subsequent unisensory WM performance. Moreover, compared with insufficient multisensory integration (e.g., modality-specific selective attention), sufficient multisensory integration (e.g., divided-modality attention) requires more resources for an individual to fully encode and integrate visual and auditory information and maintain a robust multisensory representation, leading to fewer available resources for subsequent WM retrieval.

3.6. Control experiment 2: verbal naming effect

The results of experiment 2 indicated the multisensory representation formation could be facilitated by the divided-modality attention. However, there is an unsolved question: does the faster unisensory WM performance was contributed by the coherent multisensory representation or just verbal naming effect? For example, participants may be remembering a verbal label ("cat") for the stimuli instead of or in addition to the actual visual and auditory representations. Therefore, faster unisensory working memory might be caused by verbal naming but not the actual representation.

A possible method to weaken the verbal naming effect was increasing the recognition difficulty. Taken cAV-Test V condition for an example, during the recognition stage, the visual probe has two possible types: 50% possibility were same to previous presented cat drawing, called "original type" and 50% possibility were similar but different from original type, called "novel type". Participants were asked to judge whether the probe stimulus was same to previous present stimulus. Participants might be depending more on the actual visual representation considering the fact that both two probe types have the same concept (i.e., cat).

3.7 Methods

3.7.1 Participants

A total of 11 students (3 women; age range = 20-25 years; mean age = 22.1 years, SD = 1.58, all right-handed) recruited randomly from campus, with normal or corrected-to-normal vision and hearing and no history of mental illness. 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.

3.7.2 Apparatus and materials

Half visual stimuli were the same as those used in Experiment in the manuscript. Another half visual stimuli (e.g., line drawing) were taken from the internet (i.e., https://www.google.co.jp/). Each line drawing of six semantic category has a similar picture. For example, the out-lined drawing of cat contains two types: original type and novel type. Please see the **Fig. 10**. Similarly, another half auditory stimuli were also taken from the internet (http://www.findsounds.com) and have two types. Thus, a total of 96 line drawings (6 semantic categories × 8 stimuli× 2 types) and 96 matching sounds were used in the experiment. The stimulus parameter was the same as those used in Experiment in the manuscript.

3.7.3 Experimental design and procedure

All experimental designs and procedures were the same as those in the experiment of manuscript except the probe stimulus. During the retrieval stage, there are two probe types: original and novel. Original means the probe stimulus was same to previous present stimulus. In contrast, novel means the probe stimulus was similar but different from previous presented stimulus. Please see the red frame of **Fig.10**. Participants were asked to determine whether the probe stimulus was the same as the target stimulus presented during the WM encoding stage

with a key response (e.g., “Yes”, press number key 1;” No”, press number key 3).

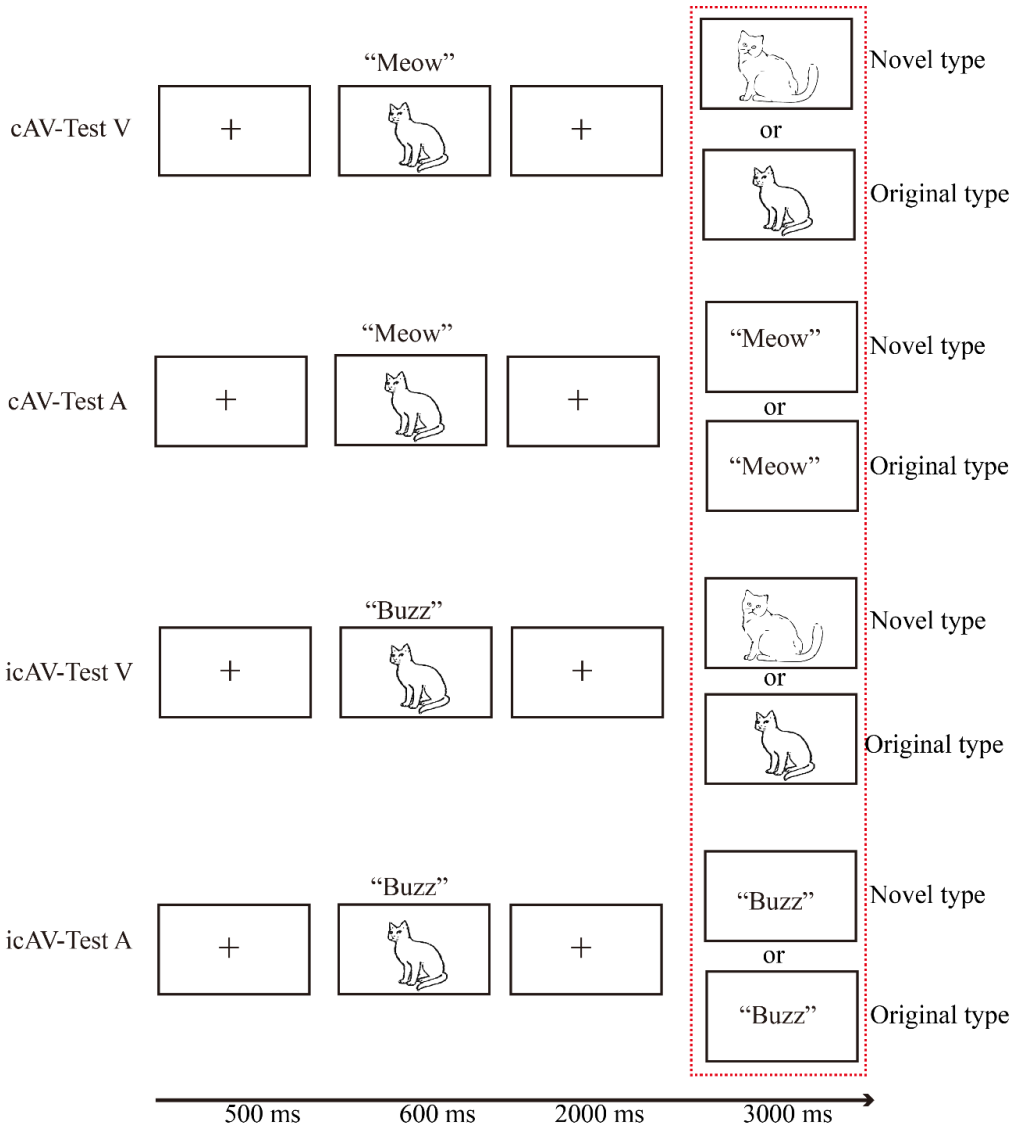

Fig. 10. *Experimental condition of control experiment. A fixation cross was shown for 500 ms, and then a stimulus (semantically congruent or incongruent audiovisual stimulus) with a duration of 600 ms was presented. A blank screen was shown after a 2000 ms delay, and finally, the probe visual or auditory stimulus (e.g., original type, 50% possibility; novel type, 50% possibility) was presented for 600 ms within a 3000 ms time limit. The participants were asked to determine whether the probe stimulus was the same as the stimuli presented during the encoding stage.*

3.8 Results

Accuracy rates were calculated as the percentage of correct responses (correct hits and correct rejections). Only RT values associated with correct responses and within ± 2 SDs were considered for further analysis.

With regards the accuracy rates (ACRs), the results showed no significant difference between cAV_s-Test V and $icAV_s$-Test V ($t = 0.47$, $p = 0.65$, $d = 0.09$) as well as cAV_s-Test V and icAVs-Test V ($t = -0.26$, $p = 0.8$, $d = 0.07$). Also, no significant difference between cAV_d-Test V and $icAV_d$-Test V ($t = -0.14$, $p = 0.89$, $d = 0.07$) as well as cAV_d-Test V and $icAV_d$-Test V ($t = -1.47$, $p = 0.17$, $d = 0.07$).

For mean correct reaction times (RTs), the results showed a significant difference between cAV_d-Test A and $icAV_d$-Test A condition ($t = -2.51$, $p < 0.03$, $d = -0.61$) as well as a weak significant difference between cAV_d-Test V and $icAV_d$-Test V condition ($t = 2.12$, $p = 0.06$, $d = -0.62$). No significant differences were found between cAV_s-Test V and $icAV_s$-Test V ($t = 0.11$, $p = 0.91$, $d = 0.03$) as well as cAV_s-Test A and $icAV_s$-Test A condition ($t = 0.98$, $p = 0.35$, $d = 0.07$). Details see the **Table 4** and **Fig.11**.

Table 4. *RTs and ACRs results during the control experiment*

Encoding	Test	RTs (M ± SD ms)	ACRs (M ± SD %)
cAV_s	V	537 ± 72	92.0 ± 5.0
$icAV_s$	V	535 ± 62	90.0 ± 6.5
cAV_d	V	554 ± 91	92.1 ± 4.6
$icAV_d$	V	636 ± 164	93.4 ± 4.6
cAV_s	A	656 ± 131	94.2 ± 4.9

$icAV_s$	A	646 ± 149	92.5 ± 5.3
cAV_d	A	628 ± 129	92.0 ± 6.0
$icAV_d$	A	724 ± 186	92.7 ± 6.2

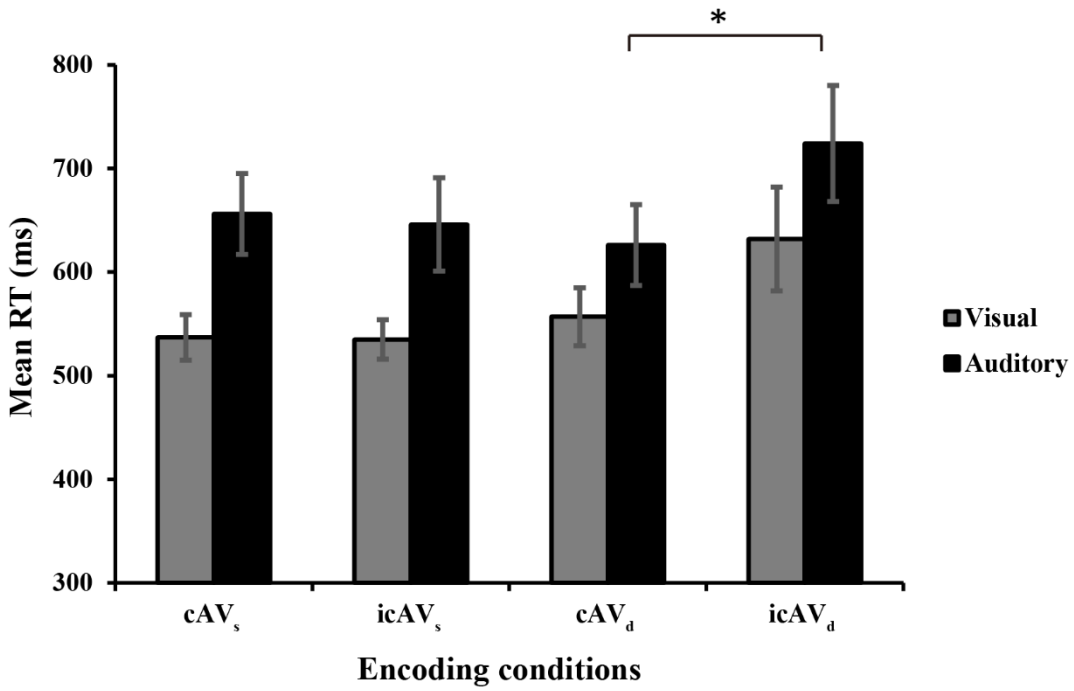

Fig. 11. *Mean RTs of visual and auditory memory retrieval under different encoding conditions. cAV_s, semantically congruent audiovisual condition with modality-specific selective attention; $icAV_s$, semantically incongruent audiovisual condition with modality-specific selective attention; cAV_d, semantically congruent audiovisual condition with divided-modality attention; and $icAV_d$, semantically incongruent audiovisual condition with divided-modality attention. Error bars denote the SE. *p < 0.05.*

3.9 Discussion

For the participants, verbal naming was a conventional method for remembering a picture by using words. However, the present results do not support the verbal naming hypothesis since

they showed that significantly faster auditory working memory (WM) retrieval indicates the possibility that matching a novel sound with the original sound can be facilitated by the semantically congruent visual stimulus. This outcome implies that a participant might depend more on the actual representation but not on verbal naming. Such findings also denote the possibility that dividing attentional resources into two modalities might lead to sufficient multisensory integration and then to the formation of a robust multisensory representation. These results also support the opinion that visual sensory processing is more suitable for processing object-related information because pictures can provide richer, more reliable information than auditory sensory processing [51].

Additionally, a weak significant difference between cAV_d-Test V and $icAV_d$-Test V (p=0.06) might reflect an insufficient sample size. Similar to faster auditory WM retrieval, both visual and auditory memory retrieval showed an accelerated tendency during the divided-modality attention mediated, semantically congruent multisensory encoding condition. This result also indicates that auditory WM retrieval could receive more semantically congruent multisensory benefits compared with visual WM retrieval.

Future work should deeply investigate the verbal naming effect for multisensory WM by using complex matching mechanisms such as word (encoding)–picture (retrieval) or picture (encoding)–word (retrieval). For example, Heikkilä et al. (2015) reported that significantly facilitated recognition memory performance benefited the semantically congruent bimodal presentation (i.e., pictures with natural sounds) but not the unimodal presentation (i.e., pictures with written words). However, the participant could easily remember the pictures by using verbal labels when the pictures were paired with written words. Heikkilä et al. (2015) suggested that the bimodal presentation of congruent information during encoding contributes to the formation of a multisensory representation.

Chapter 4 Benefits of Semantically Congruent Audiovisual Integration with Top-Down Attention on the Encoding and Retrieval Stages of Unisensory Working Memory

Summary

Although previous multisensory working memory (WM) studies have reported that semantically congruent multisensory memory encoding benefits subsequent unisensory WM retrieval, it remains unclear whether memory retrieval can also benefit from semantically congruent multisensory integration. Further, we examined whether unisensory WM retrieval concurrently benefits from semantically congruent multisensory benefits during the encoding or retrieval stages. For this chapter, we investigated the two issues by conducting two experiments. The results of the first experiment only revealed a weakly significant difference for auditory WM retrieval under the semantically congruent and incongruent conditions, indicating that less effective auditory memory retrieval was accelerated by congruent semantic information conveyed by a task-irrelevant visual stimulus. The outcomes of the second experiment showed that for visual WM retrieval, a significantly faster reaction time (RT) was found when congruent audiovisual pairs were presented during the memory encoding and retrieval stages of WM, indicating that the formation of a coherent multisensory representation was facilitated by semantically congruent audiovisual encoding, and that the visual probe triggered the multisensory representation even under the task-irrelevant, auditory stimulus interference condition. For auditory WM retrieval, it is reasonable to assume that a coherent, robust multisensory representation is constructed during semantically congruent

multisensory memory encoding because of task irrelevance, but semantically congruent visual stimuli provide more redundant information. Then, during the memory retrieval stage, a less effective auditory stimulus can trigger optimized multisensory representation and achieve rapid memory retrieval processing.

4.1 Background

Multisensory evidence has reported that unisensory visual or auditory perceptual performance, such as accuracy and response time, is facilitated by a concurrent, task-irrelevant sound or visual stimulus [1-3]. In particular, some studies suggest that such crossmodal facilitation is restricted by several rules, such as low-level temporal synchrony [4] and spatial colocation [5], as well as high-level semantic association [6] and top-down attention [7].

Crossmodal semantic congruency means that the semantic content of visual and auditory aspects belongs to the same object (i.e., a picture of a cat with the sound “meow”) or a different object (i.e., a picture of a cat with a barking sound). Studies have shown significantly faster visual response times to semantically congruent audiovisual pairs than to incongruent pairs [15]. Additionally, semantically congruent audiovisual stimuli not only facilitate visual perception performance, but can also accelerate WM retrieval [44, 45]. For instance, Xie et al. (2017) found that visual memory retrieval was accelerated by previous semantically congruent audiovisual memory encoding, indicating that a coherent multisensory representation was constructed during the encoding stage of WM and then triggered by a visual probe. These studies only focused on modality effects (i.e., bimodal vs. unimodal) or semantic congruency (i.e., congruent vs. incongruent) during the encoding stage

of WM and ignored the possibility that the memory retrieval process might also be affected by multisensory presentation.

Memory retrieval is the interaction process of an external perception signal and internal memory traces [111]. Neuroimaging studies on the relationship between encoding and retrieval [112, 113] indicate that similar cortical circuits are activated during encoding and retrieval. This indicates that improved WM performance might be caused by bimodal encoding or retrieval since previous memory studies cannot discriminate whether improved memory was caused by better encoding or retrieval operations. Prior WM studies have reported that the visual memory retrieval process can be impaired by another task: irrelevant visual information. In particular, Wais et al. (2011) investigated the impact of task-irrelevant auditory information (e.g., three conditions: white noise, ambient sound, and silence control) on visual memory retrieval and found that the presence of auditory distractions diminished the objective recollection of goal-relevant details relative to the silence and white noise conditions [113]. These findings suggest that the disruption of recollection by external stimuli is a domain-general phenomenon produced by interference between resource-limited, top-down mechanisms that guide the selection of mnemonic details and control processes that mediate our interactions with external distractors. Another study, in a PhD dissertation by Philippi, investigated crossmodal visual-haptic memory retrieval (e.g., cV-rV, cT-rT, cV-rVT, and cT-rVT) and found significant haptic memory retrieval benefits when haptic memory retrieval accompanies a visual stimulus [114]. The authors provide two possibilities for multisensory retrieval. First, the unisensory components of a multisensory retrieval cue could each initiate an attempt to retrieve information independently, followed by probability summation. Second, because of the multisensory integration of redundant information, the

unisensory components of a multisensory retrieval cue could interact and improve the signal-to-noise ratio of each retrieval attempt.

At present, it remains unclear whether crossmodal audiovisual multisensory retrieval can also lead to unisensory visual or auditory WM retrieval. We investigated the issue by using a delayed matching-to-sample paradigm. We hypothesized that unisensory WM retrieval could benefit from a semantically congruent multisensory presentation since the evidence showed that the memory retrieval process can be affected by the perception process.

4.2 Methods

4.2.1 Participants

A total of 34 paid participants (15 women; age range = 21-26 years; mean age = 24.5 years, SD = 1.46) were recruited randomly from campus to participate in experiment 1. All the participants had normal or corrected to normal vision and hearing, and both were right-handed, were without mental illness, and had not participated in a similar experiment before. 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.

4.2.2 Apparatus and materials

Visual stimuli were obtained from the standard set of outlined drawing 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, doll 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 of the 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; 44,100 Hz digitization. Semantically related sounds were delivered binaurally at an intensity level of 75 dB. A total of 48 line drawings (6 semantic categories×8 stimuli) and 48 matching sounds were used in the task.

The visual stimuli were presented on a 24-inch VG 248 LCD computer monitor with a screen resolution of 1920×1080 and a refresh rate of 144 Hz (Taiwan, ASUS); the monitor was located 75 cm away from subjects. Auditory stimuli were delivered binaurally at an intensity level of 70 dB via headphones (Sony, MH-1000XM3).

4.2.3 Experimental design and procedure

Experiment 1 consisted of a 2 memory retrieval pattern (bimodal cAV and bimodal icAV) × 2 unisensory encoding modal (V and A) within-subject design. The participants performed a delayed matching WM task during the two experimental blocks. Each block consisted of two conditions. For Block 1, the first condition evaluated the effect of task-irrelevant, semantically

congruent auditory information on unimodal visual WM retrieval performance (V-Test cAV). The second condition assessed the effect of task-irrelevant, semantically incongruent auditory information on unimodal visual WM retrieval performance (V-Test icAV). For Block 2, the first condition determined the effect of task-irrelevant, semantically congruent visual information on unimodal auditory WM retrieval performance (A-Test cAV). The second condition examined the effect of task-irrelevant, semantically incongruent visual information on unimodal auditory WM retrieval performance (A-Test icAV). Each condition contained 48 trials: 24 probe stimuli were presented, and 24 probe stimuli were not presented. The order for blocks and conditions in each block was counterbalanced across the participants. The four conditions designed for the experiment are depicted in **Fig. 12**.

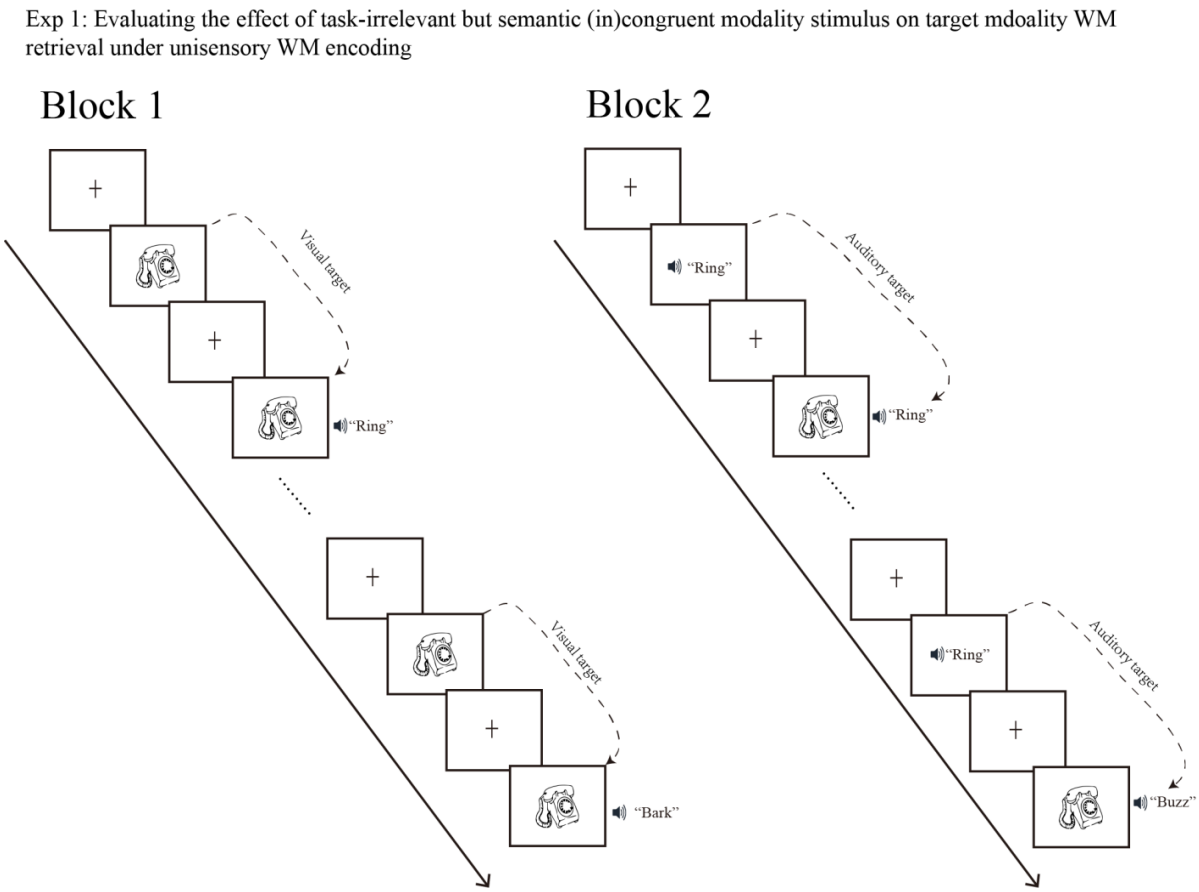

Fig.12 *Two blocks of multisensory working memory retrieval. (A) Evaluation of the effect of a task-irrelevant sound on visual memory matching. The participants were asked to judge whether the visual picture was the same as the previously presented one and to ignore the task-irrelevant, semantically congruent or incongruent sound. (B) Evaluation of the effect of task-irrelevant visual pictures on visual memory matching. The participants were asked to judge whether the sound was the same as the previously presented sound and to ignore the task-irrelevant, semantically congruent or incongruent picture.*

The study was conducted in a dimly lit, sound-attenuated, and electrically shielded laboratory room at Okayama University in Japan. For the first 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, followed by 2000 ms of delay. Next, a probe stimulus was presented for 600 ms with a 3000 ms response limit. During the WM encoding stage, the participants were asked to attend to the unisensory target modality and to remember the unisensory target stimulus. During the WM retrieval stage, the participants were asked to selectively attend to the target stimulus and ignore the task-irrelevant modality stimulus, and then determine whether the probe stimulus was the same as the target stimulus presented during the WM encoding stage with a key response (for half of the participants, the yes and no responses corresponded to the "1" and "3" number keys on the keypad, respectively; for the other half of the participants, the yes and no responses corresponded to the "3" and "1" number keys on the keypad, respectively). The presented and unpresented probe stimuli were referenced equally. All visual and auditory stimuli were presented synchronously for 600 ms, followed by a randomized intertrial interval (ITI) ranging from 1500 to 3000 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, California, USA). After each block, the participants were asked to rest for 1 min. The completion time for the entire experiment was approximately 35 min.

Before the formal experiment, each participant was required to complete two practice experiments for which 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 two conditions of each block. Each condition had four trials (i.e., two trials were the same as previous multisensory presentations, while the other two trials were not the same as the prior 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.

4.3 Results

Accurate response rates (ARS) and RTs were recorded for two blocks. The accuracy rates were calculated as the percentage of correct responses (correct hits and correct rejections). Only RT values associated with correct responses and within the mean ±2 SDs were considered for further analysis.

Regarding the ARS, a 2 modality (V and A)×retrieval congruency (bimodal cAV and bimodal icAV) repeated-measures analysis of variance (ANOVA) was conducted. No significant main effect was found for the unisensory retrieval modality $F(1, 33)=0.24$, $p=0.63$, $\eta^2=0.007$ or for retrieval congruency $F(1, 33)=0.09$, $p=0.76$, $\eta^2=0.003$. There was no

significant interaction between the unisensory retrieval modality and retrieval congruency $F(1, 33)=1.64$, $p=0.21$, $\eta^2=0.05$. For details, see **Table 5**.

Table 5*. The RT and ARS results for visual and auditory WM retrieval under different multisensory retrieval conditions. Notes: RTs = reaction times; ARS = accuracy response rates; SD = standard deviation; V = visual; A = auditory; bimodal cAV = semantically congruent audiovisual; Bimodal icAV= semantically incongruent audiovisual.*

Modality	Retrieval congruency	RTs (Me ± SD ms)	ACRs （M ± SD %）
V	Bimodal cAV	455 ± 101	94.2 ± 6.0
V	Bimodal icAV	457 ± 97	95.2 ± 5.0
A	Bimodal cAV	524 ± 112	94.3 ± 4.4
A	Bimodal icAV	533 ± 112	94.8 ± 4.7

Regarding the mean correct response RT data, a 2 modality (V and A) × retrieval congruency (bimodal cAV and bimodal icAV) repeated-measures ANOVA was conducted. The results revealed a significant difference for modality $F(1, 33) = 49.67$, $p < 0.001$, $\eta^2 = 0.6$, revealing a faster retrieval response for visual modality (456 ms) rather than auditory modality (528 ms). No significant main effect was found for the unisensory retrieval modality $F(1, 33) = 2.04$, $p = 0.16$, $\eta^2 = 0.06$. There was no significant interaction between the unisensory retrieval modality and retrieval congruency $F(1, 33) = 1.78$, $p = 0.19$, $\eta^2 = 0.05$. There was a weakly significant difference between the A-Test cAV and icAV conditions ($p = 0.09$).

4.4 Discussion

We investigated the effects of task-irrelevant, semantically congruent or incongruent modality stimuli on target unisensory WM retrieval. The results showed no significant difference for accuracy and RTs. Before the formal experiment, we hypothesized that semantically congruent, task-irrelevant modality stimuli would facilitate target unisensory WM retrieval. As one previous study reported significant visual-haptic memory retrieval benefits for unisensory visual and haptic memory retrieval, we suspect that four points may have led to these outcomes: (1) the sample size, (2) the stimulus, (3) the memory task, and (4) the evaluation index (accuracy vs. RT).

For the sample size, 18 participants took part in Philippi's study [114], while 34 participants took part in our study. Eighteen participants might be insufficient in a 2-factor within-subject design. G*power provided an appropriate sample size of 24 to research the minimum power 80% (i.e., for effect size, α parameter was used as the default value). Thus, an insufficient sample size was an important factor for the difference in the results.

The visual stimulus in Philippi's study [114] was black cards with dots or dashes, while the haptic stimulus was simple tactile vibration. In the present study, we used an outline drawing with a semantically congruent or incongruent sound. Semantic information may have been an important factor for the difference in the outcomes.

For the task, in Philippi's study [114], the participants were instructed to remember the content and location of each card during the encoding stage. In addition, the participants had to verbally repeat the letter string 'a,b,c' during the encoding phase. During the retrieval stage, the participants were asked to give the location of the object with the most resemblance to the object they just observed. In the present study, the participants were asked to memorize the picture or sound during the encoding stage of WM. Then, when the participant judged whether

the picture (i.e., visual probe) was the same as the previous memorized picture, a task-irrelevant sound accompanied the picture. The task difference might also be a critical factor for the difference in results. In particular, multisensory evidence has reported that visual perception performance can gain few multisensory integration benefits compared with auditory perception performance. Welch and Warren (1980) suggested that visual stimuli are efficient and reliable when processing object-related information [115]. Thus, visual WM matching performance can gain few benefits from semantically congruent auditory stimuli. For auditory WM matching performance, some studies indicated that maintenance of the auditory representation is easy for visual representation since evidence has shown that external visual input needs to be transformed into a corresponding phonological code and not vice versa [116].

For the evaluating index, accuracy was used in Philippi's study, while RT was evaluated in the present study. Kahana et al. (1999) discussed the relationship of accuracy and RTs in human memory in detail and suggested that both are useful measures for evaluating multisensory representations in human memory [66]. In particular, RTs were a useful index for evaluating memory retrieval speed when accuracy reached the ceiling; according to Kahana et al. (1999): "This is one version of a strength theory of memory—accuracy and IRTs [item response theories] are just two measures of the strength of information stored in memory." Kahana et al. further stated: "Superficially, it appears that our review of theory and data concerning accuracy and RT in human memory supports the view that these two measures may reflect a single underlying dimension of information" and that in "these tasks, people rarely make errors, yet speed may be of the essence." Therefore, to study tasks that are performed essentially without errors, we must consider RTs. Additionally, an overall 94% accuracy was found in our study, indicating that the task was too easy to sufficiently explore

WM performance. Thus, evaluating the RT of memory retrieval operations was an appropriate choice.

Although the interaction effect cannot enable research of the significance, we conducted a control plan comparison and found a possible significant tendency of auditory WM matching performance. This result implies that auditory memory retrieval benefits from task-irrelevant, semantically congruent visual stimuli and not vice versa. We support Philippi's opinion that less effective auditory WM retrieval might receive more benefits from a semantically congruent visual picture to some degree.

4.5 Research limitation

The present study suggests that auditory WM retrieval could be positively affected by semantically congruent visual stimuli and not vice versa. However, can auditory WM retrieval not only be affected by multisensory retrieval, but also modulate multisensory encoding benefits? In the present study, the auditory probe only triggered the unisensory auditory memory trace, even if this retrieval process was positively modulated by task-irrelevant, semantically congruent visual information. Considering previous multisensory WM studies (including our first two experiments), unisensory WM retrieval can be accelerated by prior semantically congruent audiovisual encoding. Such faster memory retrieval benefited the central storage opinion of multisensory representations, which means that memory retrieval was associated with a previous encoding operation. Hence, a subsequent experiment was necessary to investigate whether unisensory WM could concurrently benefit from semantically congruent multisensory encoding and multisensory retrieval.

4.6 Background

WM is a cognitive function that can temporarily maintain and manipulate a limited amount of information over a short period of time [93]. Memory encoding is an important stage that can transfer external perceptual information for temporary maintenance by building transient representations [87]. Different modal information that was initially processed is integrated into a coherent multisensory representation during the encoding stage of WM, and then facilitates subsequent visual WM performance [44, 45].

Multisensory studies have shown that perceptual behavioral performance is enhanced or attenuated depending on whether visual and auditory stimuli sharing semantic content belong to the same object ("semantically congruent") or not ("semantically incongruent") [15]. Not only can semantically congruent multisensory integration improve perceptual behavioral performance; it can also accelerate subsequent WM performance [44, 45]. For example, Xie et al. (2017) found a faster RT for visual WM retrieval under the semantically congruent multisensory encoding condition compared with the visual-only encoding condition. Further standardized low-resolution brain electromagnetic tomography (sLORETA) results indicate that initially processed and semantically congruent sensory information from the visual-spatial sketchpad and phonological loop are integrated into a unified multisensory representation in the posterior parietal cortex (PPC) [44].

In particular, evidence has shown that faster memory performance not only benefits from multisensory encoding, but can also be affected by multisensory retrieval. For example, Brunetti et al. (2017) investigated the impact of crossmodal correspondence (CC) on unisensory visual and auditory WM performance using an N-back paradigm [52]. They found

significantly faster unisensory WM retrieval when numerosity congruent audiovisual pairs were concurrently presented in the encoding as well as retrieval stages of WM. In particular, they found that WM retrieval was sensitive to audiovisual numerosity congruency compared with encoding operations, regardless of whether the modality was visual or auditory. The authors suggested that multisensory retrieval could be interpreted as a redundancy effect, which means that the target offers a crossmodal reinforcement of the information. Although Brunetti et al. (2017) reported such differentially multisensory benefits for memory encoding and retrieval, it remains unclear whether semantically congruent multisensory presentation can also modulate non-verbal unisensory WM performance.

Additionally, previous studies have focused on the effect of semantically congruent multisensory integration on visual memory retrieval and have overlooked the strong dependence of multisensory integration on one's attentional resources. For example, in Xie et al.'s study, the participants were asked to divide their attention between visual and auditory stimuli during multisensory encoding. If the semantic information of visual and auditory stimuli were conflicting, divided attention toward two modalities (e.g., a picture of a cat with a barking sound) may increase susceptibility to a distractor (e.g., the sound of a dog barking) and lead to impairments of the target modality (e.g., a picture of a cat) stimulus encoded into memory [55], thus further impacting target modality memory retrieval. Importantly, such interference might be destructive for subsequent auditory memory retrieval considering previous studies reporting that auditory perceptual performance can be strongly affected by a task-irrelevant visual stimulus but not vice versa [56]. A useful method to weaken the task-irrelevant inference during multisensory encoding is to selectively attend to one modality while ignoring another task-irrelevant stimulus. This method has been widely used in prior

research on multisensory integration [18, 32, 33] as well as in studies on multisensory recognition memory [58, 59]. Multisensory integration is weaker when attention is directed toward one modality compared with when attentional resources are divided between two modalities [32, 33]. Since visual WM retrieval is strongly dependent on the facilitation of a perception of crossmodal semantic multisensory integration, it is reasonable to assume that weak semantically congruent multisensory integration during WM encoding might be able to influence subsequent unisensory WM retrieval performance.

The purpose of this study was to investigate the effect of semantically congruent multisensory integration (i.e., with modal attention) in the memory encoding and retrieval stages of WM by manipulating attention to focus on the visual or auditory modality. We examined this issue by employing the delayed matching-to-sample paradigm (DMS) that was used by Xie et al. (2017). By using the DMS paradigm, we were able to directly evaluate unisensory memory retrieval under semantically (in)congruent multisensory encoding conditions without any frequent processing of representation updates, such as the N-back task [117]. We posited that both unisensory visual and auditory WM retrieval would benefit from semantically congruent multisensory encoding modulated by modal attention. In particular, auditory WM retrieval might receive more multisensory benefits than visual WM retrieval.

4.7 Methods

4.7.1 Participants

Another 34 paid participants (13 women; age range = 22-29 years; mean age = 24.9 years, SD = 1.97) were recruited randomly from campus to participate in experiment 2. All the

participants had normal or corrected to normal vision and hearing, and both were right-handed, were without mental illness, and had not participated in a similar experiment before. 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.

4.7.2 Apparatus and materials

All experimental stimuli and apparatus were the same as those in previous multisensory retrieval experiment.

4.7.3 Experimental design and procedure

All stimulus parameters and procedures were the same as those in the multisensory retrieval experiment, except for the experimental conditions. There were eight conditions (blocks) in this experiment. Unisensory visual and auditory WM retrieval were separately evaluated under the following four conditions: (1) semantically congruent audiovisual pairs were concurrently presented during the encoding and retrieval stages of WM; (2) semantically incongruent audiovisual pairs were concurrently presented during the encoding and retrieval stages of WM; (3) semantically congruent audiovisual pairs were only presented during the encoding stage of WM; (4) semantically congruent audiovisual pairs were only presented during the retrieval stage of WM. See **Fig. 13**.

During the WM encoding stage, the participants were asked to selectively attend to the target modality and to ignore another task-irrelevant modality stimulus according to different

experimental introductions. During the WM retrieval stage, the participants were also asked to selectively attend to the target modality and to ignore another task-irrelevant modality stimulus, and then to determine whether the probe stimulus was the same as the target stimulus presented during the WM encoding stage with a key response (for half of the participants, the yes and no responses corresponded to the "1" and "3" number keys on the keypad, respectively; for the other half of the participants, the yes and no responses corresponded to the "3" and "1" number keys on the keypad, respectively). Presented and unpresented probe stimuli were referenced equally.

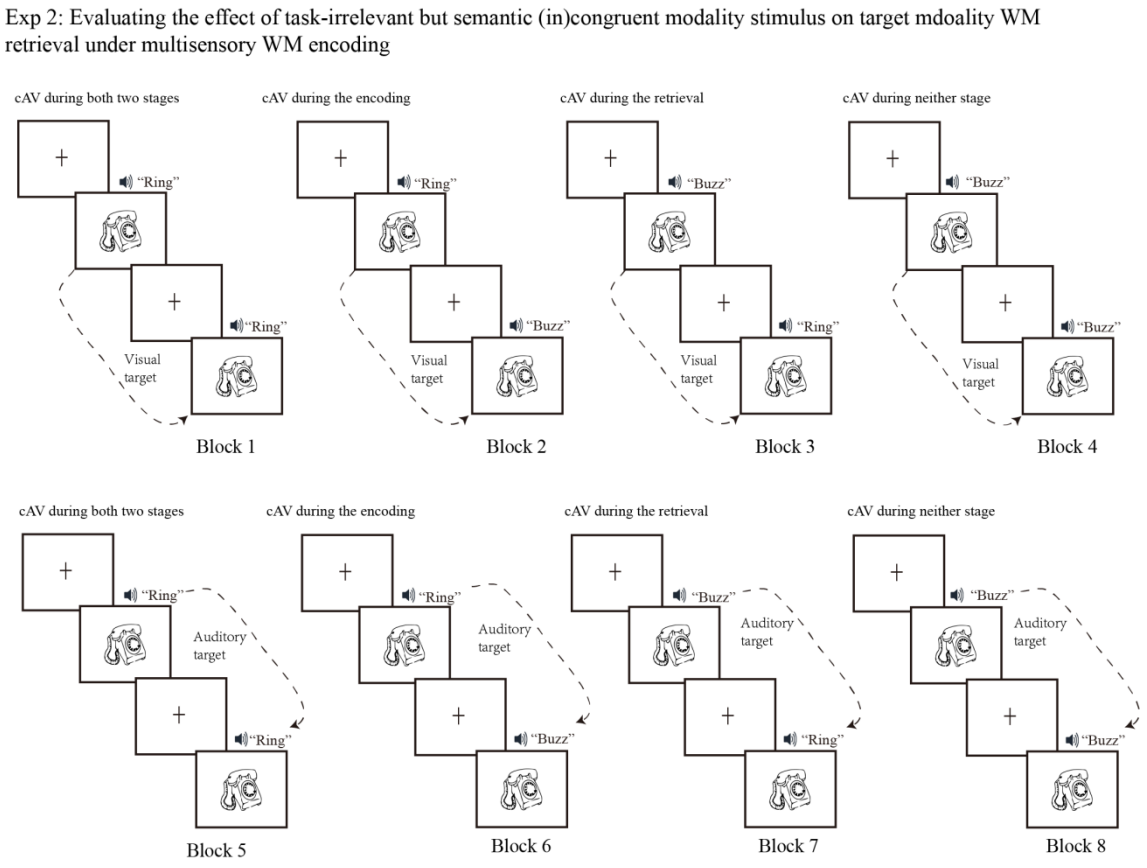

Fig.13. *Top panel: Examples of visual working memory retrieval. The participants were asked to judge whether the picture was the same as the previously presented one under four conditions: encoding and*

retrieval were semantically congruent (Block 1) or incongruent (Block 4), encoding was semantically congruent while retrieval was semantically incongruent (Block 2) and encoding was semantically incongruent while retrieval was semantically congruent (Block 3). In the bottom panel, the participants were asked to judge whether the sound was the same as the previously presented picture under four conditions: encoding and retrieval were semantically congruent (Block 5) or incongruent (Block 8), encoding was semantically congruent while retrieval was semantically incongruent (Block 6), and encoding was semantically incongruent while retrieval was semantically congruent (Block 7).

4.8 Results

Regarding the ARS, those for visual and auditory WM retrieval performance reached a ceiling in all encoding patterns (above 92%). A 2 encoding congruency (bimodal cAV and bimodal icAV) × 2 retrieval congruency (bimodal cAV and bimodal icAV) × 2 attended modality (V and A) repeated-measures ANOVA was conducted and revealed a significant main effect for attended modality $F(1, 33) = 4.73$, $p = 0.04$, $\eta^2 = 0.13$. There was no significant main effect for encoding congruency $F(1, 33) = 2.36$, $p = 0.13$, $\eta^2 = 0.07$ or retrieval congruency $F(1, 33) = 1.27$, $p = 0.27$, $\eta^2 = 0.04$. Additionally, there was no significant difference for the three-way interaction $F(1, 33) = 1.38$, $p = 0.25$, $\eta^2 = 0.04$. For details, see Table 6.

Table 6. *The RT and ARS results for visual and auditory WM retrieval under different multisensory encoding and retrieval conditions. Notes: RTs=reaction times; ARS=accuracy response rates; SD=standard deviation; V=visual; A=auditory; Bimodal cAV=semantically congruent audiovisual; Bimodal icAV=semantically incongruent audiovisual.*

Encoding	Retrieval	RTs (M ± SD ms)	ACRs (M ± SD%)
		Visual	
Bimodal cAV	Bimodal cAV	429 ± 71	95.4 ± 4.5
Bimodal cAV	Bimodal icAV	454 ± 85	95.1 ± 3.6
Bimodal icAV	Bimodal cAV	459 ± 84	95.1 ± 4.1
Bimodal icAV	Bimodal icAV	459 ±103	96.6 ± 2.8
		Auditory	
Bimodal cAV	Bimodal cAV	482 ± 89	96.7 ± 5.1
Bimodal cAV	Bimodal icAV	489 ± 97	92.7 ± 4.5
Bimodal icAV	Bimodal cAV	512 ± 120	95.2 ± 3.4
Bimodal icAV	Bimodal icAV	531 ± 128	94.8 ± 4.1

Regarding the mean correct response RT data, a 2 encoding congruency (bimodal cAV and bimodal icAV) × 2 retrieval congruency (bimodal cAV and bimodal icAV) × 2 attended modality (V and A) repeated-measures ANOVA was conducted and showed a significant main effect for encoding congruency $F(1, 33) = 22.28$, $p < 0.001$, $\eta^2 = 0.4$, demonstrating a faster response under the bimodal cAV condition (450 ms) than under the bimodal icAV condition (503 ms). The results also revealed a significant difference for retrieval congruency $F(1, 33) = 17.73$, $p < 0.001$, $\eta^2 = 0.35$, demonstrating a faster response under the bimodal cAV condition (463 ms) than under the bimodal icAV (490 ms). Additionally, the results indicated a significant main effect for attended modality $F(1, 33) = 7.3$, $p = 0.01$, $\eta^2 = 0.18$, showing a

faster response for the visual modality (470 ms) than for the auditory modality (484 ms). The interaction between these three factors was significant $F(1, 33) = 4.72$, $p = 0.04$, $\eta^2 = 0.13$. The details of the ARS and RTs are depicted in **Table 2**.

To evaluate the effect of crossmodal semantic congruency on subsequent unisensory visual and auditory WM retrieval, two separate 2 semantic congruency (bimodal cAV and bimodal icAV) × 2 retrieval congruency (bimodal cAV and icAV) repeated-measures ANOVAs were performed. For visual WM retrieval, the results only highlighted a significant main effect of encoding congruency (F (1, 33) = 5.34, p = 0.03, η^2 = 0.14), indicating an encoding advantage for the cAV condition (442 ms) over the bimodal icAV condition (459 ms). Additionally, there was a significant interaction between encoding congruency and retrieval congruency (F (1, 33) = 5.1, p = 0.03, η^2 = 0.13). A post hoc t-test showed that response time was significantly faster when encoding and retrieval were in the bimodal cAV rather than bimodal icAV conditions (t= -2.23, p = 0.03, d = 0.34). Additionally, response time was significantly faster when encoding and retrieval were in the bimodal cAV rather than the encoding bimodal cAV and retrieval icAV conditions (t = -3.26, p = 0.003, d = 0.32). Another significant difference was found when encoding and retrieval were in the bimodal cAV—compared with the encoding bimodal icAV and retrieval cAV—conditions (t = -3.77, $p <$ 0.001, d = 0.39). See **Fig. 14 (A)**.

For auditory WM retrieval, the results only revealed a significant main effect of encoding congruency (F (1, 33) = 15.93, $p < 0.001$, $\eta^2 = 0.33$), indicating an encoding advantage for the cAV condition (484 ms) over the bimodal icAV condition (521 ms). The interaction between encoding congruency and retrieval congruency was not significant (F (1, 33) = 0.42, p = 0.52, η^2 = 0.12). A plan comparison uncovered a significant difference between the cAV-Test cAV

condition and the icAV-Test cAV condition (p = 0.006), as well as the icAV-Test icAV condition ($p < 0.001$). Additionally, there was a significant difference between cAV-Test icAV and icAV-Test cAV (p = 0.03), as well as icAV-Test icAV ($p < 0.001$). See **Fig. 14 (B)**.

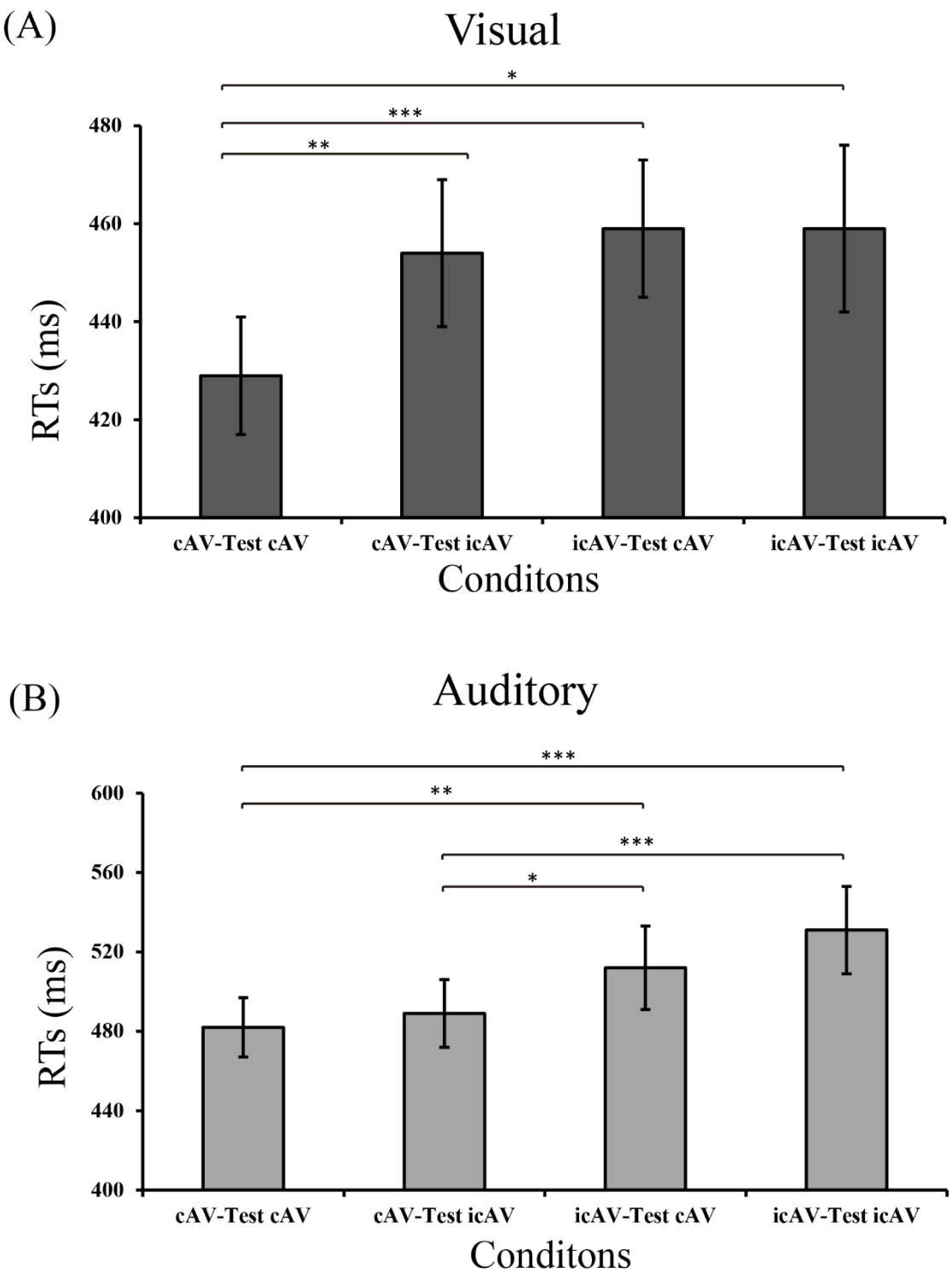

Fig. 14. *Mean RTs of visual and auditory working memory retrieval under different encoding and retrieval conditions. cAV=semantically congruent audiovisual condition; icAV=semantically incongruent audiovisual condition; cAV=semantically congruent audiovisual condition; and* $icAV_d$*=semantically*

*incongruent audiovisual condition. Top panel: Reaction time results of visual working memory retrieval under four conditions: encoding and retrieval were semantically congruent (cAV-Test cAV) or incongruent (icAV-Test icAV), encoding was semantically congruent while retrieval was semantically incongruent (cAV-Test icAV), and encoding was semantically incongruent while retrieval was semantically congruent (icAV-Test cAV). Bottom panel: Reaction time results of auditory working memory under four conditions: encoding and retrieval were semantically congruent (cAV-Test cAV) or incongruent (icAV-Test icAV), encoding was semantically congruent while retrieval was semantically incongruent (cAV-Test icAV), and encoding was semantically incongruent while retrieval was semantically congruent (icAV-Test cAV). Error bars denote the SE. *p<0.05, **p<0.005, ***p<0.001.*

4.9 Discussion

The purpose of this study was to investigate the beneficial effects of semantically congruent multisensory integration during the encoding and retrieval stages on unisensory visual and auditory WM performance. In agreement with Brunetti's study, our results showed significantly faster visual and auditory performance when semantically congruent multisensory presentation was presented in the memory encoding and retrieval stages compared with the semantically incongruent conditions. However, the results were different when semantically congruent audiovisual pairs were only presented during the encoding or retrieval stages of WM.

For unisensory visual WM retrieval, significantly faster memory retrieval speed was only found when semantically congruent audiovisual pairs were concurrently presented in the encoding *and* retrieval stages compared with only presented in the encoding (p=0.03) *or* retrieval stage (p=0.003), indicating that crossmodal semantic congruency strongly

contributed to multisensory facilitation for the memory encoding stage. One possible explanation is that the formation of a coherent multisensory representation was facilitated by semantically congruent AVI, and the visual probe then triggered the multisensory representation even though the task-irrelevant, auditory stimulus interfered with visual memory retrieval. Previous multisensory studies have reported that visual stimuli play a dominant role in object recognition because visual stimuli can provide more reliable information when processing objects (i.e., modality appropriateness theory [115]). In particular, the lack of a significant difference between the cAV-Test cAV and cAV-Test icAV conditions may signal that the incongruent auditory stimulus interferes with memory retrieval. This outcome supports the opinion of Wais that the disruption of recollection by external stimuli is a domain-general phenomenon produced by interference between resource-limited, top-down mechanisms that guide the selection of mnemonic details and control processes that mediate our interactions with external distractors [113]. Such an explanation might also support and extend Philippi's opinion that unisensory components of a multisensory retrieval cue could interact with a multisensory representation by improving the signal-to-noise ratio of each retrieval attempt. Such memory retrieval may have been strongly dependent on the semantic content of another task-irrelevant modality stimulus. Exclusively, this outcome may partially conflict with the findings of Brunetti, who showed strong multisensory benefits for memory retrieval but not memory encoding. This result is not surprising because Brunetti could not find a significant difference between multisensory encoding and multisensory retrieval when the stimulus type was a quantity, thereby underlining the strong dependency of the stimulus material.

For unisensory auditory WM retrieval, similar to visual WM, there was a significantly

faster RT for the cAV-Test cAV condition than for the icAV-Test icAV condition, implying that the behavioral benefits were concurrently contributed to by semantically congruent multisensory encoding and retrieval. In contrast to the visual modality, there were three novel outcomes. First, there was a significant difference between cAV-Test cAV and icAV-Test icAV, denoting that auditory memory retrieval was strongly dependent on the previous congruent audio-visual encoding. The specifically facilitated auditory WM retrieval performance is partially consistent with several findings on multisensory recognition memory. For example, Thelen et al. (2015) compared the effect of semantically congruent and incongruent multisensory presentations on later unisensory recognition and found that semantically congruent multisensory gains for auditory recognition performance were significantly higher than those for visual recognition memory [51]. In addition, Heikkilä et al. (2017) found that discrimination ability regarding old/new objects was significantly higher for auditory recognition with a picture/written word that can carry object-related information than for other conditions [59]. Second, significantly faster auditory memory retrieval under the cAV-Test icAV condition than under the icAV-Test cAV condition might indicate that the benefits of multisensory encoding are larger than those of multisensory retrieval. Auditory memory retrieval seems to be strongly associated with a previously constructed coherent multisensory representation, and even resistance to visual interference. Moreover, the lack of a significant difference between the icAV-Test cAV and icAV-Test icAV conditions might mean that the semantically congruent multisensory encoding benefits for auditory WM are larger, even though the semantically congruent visual stimulus can provide semantically congruent information for auditory WM retrieval. This outcome conflicts with the work of Brunetti, who showed that auditory WM retrieval is more sensitive to multisensory retrieval

but not multisensory encoding. This discrepancy might reflect the difference in stimulus type (e.g., a meaningless point and sound vs. a meaningful picture and matching sound) or paradigm (e.g., the N-back paradigm vs. the delayed matching-to-sample paradigm). Brunetti et al. (2017) could not report the difference between CC at sample stimulus (i.e., cAV-Test icAV) and CC at target stimulus (i.e., icAV-Test cAV) when the auditory stimulus type was a digit. Indeed, Brunetti also claimed that the nature of the numerosity CC is dependent on the type of information used (digits vs. quantities).

In particular, compared with semantically incongruent multisensory encoding, the results indicate that auditory WM retrieval can receive more semantically congruent multisensory encoding benefits (p=0.03) than visual WM retrieval (p=0.58). This hypothesis is supported by a recent multisensory study that suggests that inverse effectiveness enhancement can be modulated by low-level stimulus association (e.g., spatial alignment and temporal synchrony) and high-level semantic congruency [78]. Thus, a less effective auditory stimulus might trigger a more multisensory process due to visually induced auditory verse facilitation during memory retrieval. Another point to consider is modal-based attention; all the participants were asked to selectively attend to one modality and to ignore another task-irrelevant modality stimulus, regardless of multisensory encoding or multisensory retrieval. Prior multisensory evidence demonstrates that visual sensory processing is more suitable for processing object-related information because pictures can provide richer, more reliable details than auditory sensory processing [79, 80]. Hence, the effect of task-irrelevant visual information on auditory WM encoding should not be ignored. Schmid et al. (2011) explored the interaction mechanism between crossmodal competition and modal attention using fMRI measurements, and found a significant visual dominance advantage only when attention was

focused on the auditory modality [81]. The authors suggested that crossmodal competition was modulated by modal attention, and that poor auditory encoding can receive more redundant information compensation from an unattended visual stimulus. Attending to a poor modality encoding compensation mechanism might reflect the flexible recognition necessary for the external environment. Thus, for semantically congruent multisensory memory encoding, it is reasonable to assume that a coherent, robust multisensory representation would be constructed during WM encoding because of task irrelevance, but semantically congruent visual stimuli provide more redundant information. Then, during the WM retrieval stage, a less effective auditory stimulus can trigger an optimized multisensory representation and achieve rapid WM retrieval processing. For semantically incongruent multisensory encoding, the formation of a coherent multisensory representation during the WM encoding stage is strongly disturbed by a mismatching picture; hence, auditory WM retrieval cannot activate a coherent representation, even though the task-irrelevant visual stimulus can provide congruent semantic information for auditory WM retrieval.

4.10 Research limitation

Although we found that visual and auditory WM retrieval can receive more multisensory benefits from the memory encoding stage but not the retrieval stage, it remains unclear whether the multisensory benefit pattern could be changed under a higher memory load (e.g., the N-back paradigm). Importantly, the central executive might be involved in the formation of multisensory representations. For the N-back task, a higher memory load also means that the central executive must construct more multisensory representations during the semantically congruent encoding condition. Therefore, unisensory memory retrieval might be changed

because WM is a resource-limited system, and only a few multisensory representations are maintained for subsequent cognitive operations. As such, future work should investigate whether the advantage of semantically congruent multisensory encoding can be amplified or eliminated in light of a higher memory load.

4.11 Conclusions

We investigated whether unisensory WM retrieval could be concurrently modulated by semantically (in)congruent multisensory encoding and multisensory retrieval. For visual WM retrieval, we observed a significantly faster RT when congruent audiovisual pairs were presented during the memory encoding and retrieval stages of WM, indicating that the formation of a coherent multisensory representation was facilitated by semantically congruent audiovisual encoding, and that the visual probe triggered the multisensory representation even under the task-irrelevant, auditory stimulus interference condition. For auditory WM retrieval, it is reasonable to assume that a coherent, robust multisensory representation would be constructed during semantically congruent multisensory memory encoding because of task irrelevance, but semantically congruent visual stimuli provide more redundant information. Then, during the WM retrieval stage, a less effective auditory stimulus could trigger an optimized multisensory representation and achieve rapid WM retrieval processing.

Chapter 5 General Conclusion and Future Projections

Summary

Our main aim was to investigate the beneficial effect of semantic audiovisual interactions on subsequent unisensory WM retrieval. Our results support and extend the central storage opinion of memory representation by showing that unisensory WM retrieval (i.e., especially auditory) can be accelerated by crossmodal semantic congruency as well as top-down divided-modality attention. In addition, visual and auditory WM retrieval differentially benefited from multisensory encoding and retrieval. This chapter summarizes our findings. Furthermore, we propose some paths for future research.

5.1 General Conclusions

We investigated unisensory WM retrieval under three multisensory conditions: (1) semantically congruent audiovisual encoding benefits; (2) the interaction benefits of semantically congruent audiovisual integration (AVI) and top-down attention; and (3) the interaction benefits of semantically congruent AVI and top-down attention during the encoding and retrieval stages of WM.

Chapter 2 describes how semantically congruent AVI can differentially modulate subsequent unisensory visual and auditory short-term memory (STM) by applying the delayed matching-to-sample (DMS) paradigm. The results revealed significantly faster unisensory short-term retrieval performance under the semantically congruent audiovisual encoding condition. Our findings suggested that the formation of a coherent multisensory representation 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 short-term retrieval, we suggest that the formation of a coherent multisensory representation is strengthened by a semantically congruent visual stimulus that is not the focus during the memory encoding stage. During the memory retrieval stage, a less effective auditory stimulus can trigger an optimized multisensory representation, thereby facilitating rapid memory retrieval processing.

DMS task has been widely used in previous studies on STM and WM. To further evaluate the possibility that unisensory memory retrieval is involved in WM but not limited to STM, we conducted a control experiment. The RT outcomes for the control experiment showed a significantly negative impact on unisensory WM retrieval compared with the DI and

NI conditions. In particular, for the INT condition, the RT results indicated a significant difference in visual WM retrieval between semantically congruent bimodal memory encoding and unimodal memory encoding. These outcomes are partially consistent with our formal experiment described in the manuscript, signaling that semantically congruent bimodal encoding accelerates unisensory STM and WM.

Chapter 3 describes whether the interaction of semantically congruent AVI and top-down attention can further modulate subsequent unisensory visual and auditory WM performance. The findings reconcile and extend previous multisensory WM studies by demonstrating that the semantically congruent bimodal presentation with divided-modality attention can accelerate subsequent unisensory WM retrieval, especially less effective auditory WM retrieval. This outcome implies that a sufficient semantically congruent bimodal presentation (e.g., divided-modality attention) not only facilitates immediate behavioral perceptual performance, but can also strongly impact subsequent unisensory WM performance. Moreover, compared with insufficient multisensory integration (e.g., modality-specific selective attention), sufficient multisensory integration (e.g., divided-modality attention) requires more resources for the individual to fully encode and integrate visual and auditory information and maintain a robust multisensory representation, leading to fewer available resources for subsequent WM retrieval. In particular, we conducted a control experiment to assess whether participants' memory could be affected by the visual or auditory stimulus by using a verbal naming method. In line with our experimental results in the manuscript, the outcomes of the control experiment revealed a significant difference between the cAV_d-Test A and $icAV_d$-Test A conditions, suggesting that dividing attentional resources into two modalities might lead to sufficient multisensory integration and the formation of a robust multisensory representation.

Chapter 4 describes whether the interaction of semantically congruent AVI and top-down attention can differentially modulate unisensory visual and auditory WM performance by affecting the encoding or retrieval stages. The first experiment examined whether unisensory WM retrieval benefits from multisensory retrieval. We only found a weak significant difference for auditory WM retrieval under the semantically congruent and incongruent multisensory retrieval conditions. Then, the second experiment evaluated whether unisensory WM retrieval not only benefits from multisensory retrieval, but also from multisensory encoding. For visual WM retrieval, we noted a significantly faster RT when congruent audiovisual pairs were presented during the memory encoding and retrieval stages of WM, implying that the formation of a coherent multisensory representation was facilitated by semantically congruent audiovisual encoding, and that the visual probe triggered the multisensory representation, even under the task-irrelevant, auditory stimulus interference condition. For auditory WM retrieval, it is reasonable to assume that a coherent, robust multisensory representation would be constructed during semantically congruent multisensory memory encoding because of task irrelevance, but semantically congruent visual stimuli provide more redundant information. Then, during the WM retrieval stage, a less effective auditory stimulus could a trigger optimized multisensory representation and achieve rapid WM retrieval processing.

Overall, this thesis supports the view of the central storage of memory representation by showing that unisensory WM retrieval (e.g., especially auditory) can be accelerated by semantically congruent AVI. Such semantically congruent audiovisual encoding might lead to a coherent multisensory representation, which can be triggered by subsequent unisensory components. Importantly, this thesis further extends the central storage opinion by demonstrating that auditory memory retrieval can gain more semantically congruent

multisensory encoding benefits than visual memory retrieval. In particular, auditory memory retrieval can gain more benefits from the divided-modality attention-mediated multisensory encoding condition, highlighting a correlation between accelerated but less effective auditory memory retrieval and a divided-modality, attention-optimized multisensory representation. Additionally, auditory memory retrieval specifically benefited from semantically congruent audiovisual encoding compared with congruent audiovisual retrieval, thereby supporting the encoding-retrieval matching perspective of Byberg et al. (2000) that the memory retrieval stage depends more on the extent to which the probe information overlaps with previously encoded information [118]. Further, for the semantically congruent audiovisual encoding conditions, views on modality-specific unisensory storage might offer appropriate explanations considering the claim that semantic conflicting crossmodal signals interfere with the formation of coherent multisensory representations during the encoding stage of WM [45]. Future work should verify this hypothesis by exploring unisensory WM retrieval under different semantically incongruent multisensory encoding conditions.

5.2 Future Projections

We focused on the behavioral facilitation effect of audiovisual interactions on subsequent unisensory WM performance. However, several questions remain unresolved.

(1) *Stimulus material*. Since the visual stimuli are pictures, it is unclear whether the participants indeed engaged in multisensory integration to the extent or in the same way that they would if videos had been used instead. Rather, it seems like they may have engaged in a higher-level semantic-based association process for the semantically congruent AV stimuli.

Audiovisual video stimuli can minimize some stimulus-driven contributions (i.e., temporal synchrony) to multisensory integration and provide stronger semantic associations for possible semantic integration. For the present study, we used outline drawings and their matching sound, which have been widely used in related studies on multisensory memory. In the future, we would like to investigate whether faster unisensory retrieval speed is associated with early perception facilitation or later semantic integration by using dynamic audiovisual video stimuli.

(2) *A comparison of paradigms*. We used the DMS paradigm to evaluate unisensory WM performance in different multisensory encoding environments. We conducted two possible control experiments and found multisensory benefits for unisensory WM. However, this paradigm was too easy to use and could not sufficiently evaluate WM resources. In fact, the participants' performance was at a ceiling, meaning that accuracy (which should be the first measure to consider in a memory experiment) was unusable. Thus, future research should examine the audiovisual interaction benefits for unisensory WM performance under high load memory conditions, such as the N-back task.

(3) *The neural mechanism of multisensory encoding and multisensory retrieval*. Our study provides possible behavioral evidence for semantic multisensory encoding or multisensory retrieval benefits. However, the neural substrate of these experiments remains unclear. Future work should focus on the neural mechanism of multisensory encoding and multisensory retrieval. In particular, the neural mechanism difference of visual and auditory WM should be compared by using high-temporal resolution ERP measurements and high spatial-resolution fMRI measurements. The evidence has reported that WM performance was decreased in patients with mild cognition impairment compared to healthy adults [119]. We suspect that these patients also exhibited asymmetric memory retrieval. In the future, by

combining behavioral and neural evidence, we will explore the possibility of predicting different stages of Alzheimer's disease by evaluating auditory WM performance. This study may provide useful support for optimizing the present Alzheimer's Disease Assessment Scale–cognitive subscale (ADAS-cog, a neuropsychological assessment widely used to evaluate the severity of cognitive symptoms of dementia). Especially, considering the bimodal memory encoding advantage compared with unimodal encoding, this thesis tentatively suggested that bimodal audiovisual encoding training might be a possible method for improving the memory retrieval performance of older adults or mild cognitive impairment patients.

Appendix

実験参加報告書

(実験者(記入)→実験参加者(記入:実験日)→実験者(受取)→GL→高橋)

【実験者 記入欄】

実験タイトル：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

実験者：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

実施年月日：＿＿＿＿年＿＿＿＿月＿＿＿＿日

実験時間：＿＿＿＿：＿＿＿＿～＿＿＿＿：＿＿＿＿ (実験説明，休憩含む)

実験場所＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

実験項目(内容は簡単に，実験数に応じて記入)

実験１：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

実験２：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

実験３：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

実験４：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

実験５：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

【実験参加者 記入欄】(表面，裏面とも記入し，署名する)

○現在の状態について，次の７段階で最も当てはまるスケールの数字に○印をしてください．

↓項目	当てはまらない		少し当てはまる		かなりあてはまる		非常に当てはまる
まぶたが重いと感じる	1	2	3	4	5	6	7
眠い	1	2	3	4	5	6	7
緊張している	1	2	3	4	5	6	7
どきどきしている	1	2	3	4	5	6	7
くつろいだ気分だ	1	2	3	4	5	6	7
ゆったりした気分だ	1	2	3	4	5	6	7
思考が鈍っている	1	2	3	4	5	6	7
認知の集中ができにくい	1	2	3	4	5	6	7
活力がみなぎっている	1	2	3	4	5	6	7
積極的な気分だ	1	2	3	4	5	6	7

やる気が出ない	1	2	3	4	5	6	7
何かすることに気乗りがしない	1	2	3	4	5	6	7
昨日は十分に睡眠をとった	1	2	3	4	5	6	7
昨日今日は暴飲暴食，深酒はしていない	1	2	3	4	5	6	7
健康状態は良好だ	1	2	3	4	5	6	7
実験中に不安に感じることがあった	1	2	3	4	5	6	7
次回も実験に参加したい	1	2	3	4	5	6	7
実験１の方法は理解できた	1	2	3	4	5	6	7
実験２の方法は理解できた	1	2	3	4	5	6	7
実験３の方法は理解できた	1	2	3	4	5	6	7
実験４の方法は理解できた	1	2	3	4	5	6	7
実験５の方法は理解できた	1	2	3	4	5	6	7
(以下必要に応じて実験者が追加)							
	1	2	3	4	5	6	7
	1	2	3	4	5	6	7
	1	2	3	4	5	6	7
↑項目	当てはまらない		少し当てはまる		かなりあてはまる		非常に当てはまる

実験について，感想・コメントを１言(以上)書いてください.

【実験参加者 署名】

実験年月日：　　年　　月　　日

署名(自署)：

同意書への記入：　　有　　無

Publications

Journal Papers

[1] ***Yu, H***., Wang, A., Li, Q., Liu, Y., Yang, J., Takahashi, S., . . . Wu, J. (2021). Semantically Congruent Bimodal Presentation with Divided-Modality Attention Accelerates Unisensory Working Memory Retrieval. *Perception, 50*(11), 917-932. doi: 10.1177/03010066211052943.

[2] ***Hongtao Yu***, Aijun Wang*, Ming Zhang, Jiajia Yang, Satoshi Takahashi, Ejima Yoshimichi and Jinglong Wu*. Semantically congruent audiovisual integration with modal-based attention accelerates auditory short-term memory retrieval, *Attention, Perception, & Psychophysics,* (Accepted, 2021).

[3] ***Hongtao Yu***, Qiong Wu, Mengni Zhou, Qi Li, Jiajia Yang, Satoshi Takahashi, Ejima Yoshimichi and Jinglong Wu*. A Basic Psychophysics Study of Crossmodal Semantic Reliability for Optimizing the Multisensory Presence in Virtual Reality, *International Journal of Mechatronics and Automation (IJMA)* (Accepted, 2022).

Conference Papers

[1] ***Yu, H.***, Wu, Q., Zhou, M., Li, Q., Yang, J., Takahashi, S., ... & Wu, J. (2021, August). A Basic Psychophysics Study of Sound Reliability Effects on Audiovisual Integration for Developing New Virtual Reality Device. In *2021 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 1298-1303.

[2] Zhou Mengni, ***Yu Hongtao***, 上原翼，楊家家， 高橋智，江島義道，呉景龍. (2021). 視幼児顔の可愛さ知覚の空間周波数依存性に関する認知心理学的研究 (第18回日本ワーキングメモリ学会大会, 大会発表要旨).

Acknowledgements

First of all, I would like to express my sincerely gratitude to Prof. Jinglong Wu for the continuous support during my Ph.D studies. Prof. Jinglong Wu helped me in daily life, research design and writing of this thesis. I could not complete my study of doctor course and finish this thesis successfully without his enlightening instruction, impressive kindness and patience. Thanks to Prof. Jinglong Wu for giving me guidance during the very confused and helpless submission process, my paper can be accepted smoothly. Without Prof. Wu's encouragement and support, the long nights spent in the lab, and months of data collection, would not have been possible. Thank you for always having my back, and my best interests at heart. Your diligence gives me power not only during my present PhD scours, but also in my future research working.

Secondly, I also want to express my sincerely gratitude to Prof. Satoshi Takahashi. His rigorous scientific research attitude, serious and rigorous working style, as well as the concern and responsibility for the students' academic work will always be a role model for my future research work. I got a lot of comments for Prof. Satoshi Takahashi during the group meeting, including talking skills using native Japanese language, experiments design, prepare for conference papers, and prepare for graduation announcement and this thesis.

Thirdly, I also want to express my sincere gratitude to Prof. Ming Zhang, who

provided very valuable advice for revising my research plan, conducting experiments design, writing the dissertation, proposing kindly manuscript review and encouragement. As a freshman in cognitive psychology, I learn a lot from Prof. Ming Zhang. I could not complete my study of doctor course and finish this thesis successfully without his enlightening instruction, impressive kindness and patience. His diligence gives me power not only during my present PhD scours, but also in my future life.

Fourthly, I would like to express my sincerely my gratitude to Prof. Aijun Wang. Thank you for the experiment design and statistical analyses of psychology experiment. He has also been of great help in resolving technical and statistical conundrums, other than being somebody who always lends a helping hand. Especially, Prof. Wang read many drafts of my writing, and always being around to share ideas, frustrations, and good times.

Finally, I would also like to express my sincere thanks to Prof. Jiajia Yang and Prof. Yoshimichi Ejima, without your precious support, it would not be possible to conduct my research successfully. I also thank the students and staff in the Wu lab. Without their cooperation, I cannot imagine how I could have finished my experiments. I sincerely thank all those who contributed to my experiment, my paper, and dissertation. And I thank my fellow lab mates in for the stimulating discussions and for all the fun we have had together during these years. I would like to thank my parents and friends for supporting me spiritually throughout writing this thesis and my life in general.

References

[1] Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Brain Res Cogn Brain Res, 24*(2), 326-334.

[2] Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cereb Cortex*, 17(3), 679-690.

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

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

[5] Starke, J., Ball, F., Heinze, H. J., & Noesselt, T. (2020). The spatio-temporal profile of multisensory integration. *Eur J Neurosci*, 51(5), 1210-1223.

[6] Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain Res, 1242*, 136-150.

[7] Rohe, T., & Noppeney, U. (2018). Reliability-Weighted Integration of Audiovisual Signals Can Be Modulated by Top-down Attention. *eNeuro*, 5(1).

[8] Howard, I., & Templeton, W. (1966). Human spatial orientation. John Wiley & Sons

[9] Smith, E., Zhang, S., & Bennetto, L. (2017). Temporal synchrony and audiovisual integration of speech and object stimuli in autism. *Res Autism Spectr Disord*, 39, 11-19.

[10] Li, Q., Wu, Y., Yang, J., Wu, J., & Touge, T. (2015). The temporal reliability of sound modulates visual detection: an event-related potential study. *Neurosci Lett*, 584, 202-207.

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

[12] Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science*, 221 (4608), 389–391.

[13] Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. 1. Temporal factors.

Journal of Neuroscience, 7(10), 3215–3229

[14] Hein G, Doehrmann O, Müller NG, Kaiser J, Muckli L, Naumer MJ (2007) Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *J Neurosci* 27:7881–7887.

[15] 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.

[16] Friston, K. J. (2018). Does predictive coding have a future? *Nature Neuroscience*, 21(8), 1019–1021.

[17] Chan, J. S., Langer, A., & Kaiser, J. (2016). Temporal integration of multisensory stimuli in autism spectrum disorder: A predictive coding perspective. Journal of Neural Transmission, 123, 917–923.

[18] 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. doi: 10.1371/journal.pone.0221185

[19] Belardinelli MO, Sestieri C, Di Matteo R, Delogu F, Del Gratta C, Ferretti A, Caulo M, Tartaro A, Romani GL (2004): Audio-visual crossmodal interactions in environmental perception: An fMRI investigation. *Cogn Process* 5:167–174.

[20] Plank, T., Rosengarth, K., Song, W., Ellermeier, W., & Greenlee, M. W. (2012). Neural correlates of audio-visual object recognition: effects of implicit spatial congruency. *Hum Brain Mapp*, 33(4), 797-811.

[21] Ye, Z., Russeler, J., Gerth, I., & Munte, T. F. (2017). Audiovisual speech integration in the superior temporal region is dysfunctional in dyslexia. *Neuroscience*, 356, 1-10.

[22] Beauchamp MS (2005) See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr Opin Neurobiol* 15:145–153

[23] Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychol (Amst)*, 134(3), 372-384.

[24] Tang, X., Wu, J., & Shen, Y. (2016). The interactions of multisensory integration with endogenous and exogenous attention. *Neurosci Biobehav Rev*, 61, 208-224.

[25] Fujino, J., Tei, S., Itahashi, T., Aoki, Y.Y., Ohta, H., Izuno, T., Nakamura, H., Shimizu, M., Hashimoto, R.I., Takahashi, H., Kato, N., Nakamura, M., 2021. A single session of navigation-guided repetitive transcranial magnetic stimulation over the right anterior temporoparietal junction in autism spectrum disorder. *Brain Stimulation* 14 (3), 682–684.

[26] 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. *Eur J Neurosci*, 26(2), 499-509.

[27] Xia, J., Zhang, W., Jiang, Y., Li, Y., & Chen, Q. (2018). Neural practice effect during cross-modal selective attention: Supra-modal and modality-specific effects. *Cortex*, 106, 47-64.

[28] Spence, C., & Frings, C. (2020). Multisensory feature integration in (and out) of the focus of spatial attention. *Attention, Perception, & Psychophysics*, 82, 363–376

[29] Spilcke-Liss, J., Zhu, J., Gluth, S., Spezio, M., & Glascher, J. (2019). Semantic Incongruency Interferes With Endogenous Attention in Cross-Modal Integration of Semantically Congruent Objects. *Front Integr Neurosci*, 13, 53.

[30] Donohue, S. E., Todisco, A. E., and 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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Allen, R. J., Hitch, G. J., & Baddeley, A. D. (2009). Cross-modal binding and working memory. *Visual Cognition*, 17(1-2), 83-102.

[36] Baddeley, A. (1992). Working memory. *Science*, 255(5044), 556-559. doi: 10.1126/science.1736359

[37] Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology Section A*, 49(1), 5-28.

[38] Baddeley, A.D. and Hitch, G. (1974) Working Memory. In Psychology of Learning and Motivation 8 (Bower, G. H., ed), pp. 47–89, *Academic Press*

[39] Barutchu, A.; Sahu, A.; Humphreys, G.W.; Spence, C. Multisensory processing in event-based prospective memory. *Acta Psychol*. 2019, 192, 23–30.

[40] Goolkasian, P., & Foos, P. W. (2005). Bimodal format effects in working memory. American Journal of Psychology, 118(1), 61-77.

[41] Goolkasian, P. (2000). Pictures, words, and sounds: from which format are we best able to reason? *J Gen Psychol*, 127(4), 439-459.

[42] Thompson, V. A., & Paivio, A. (1994). Memory for pictures and sounds: Independence of auditory and visual codes. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 48(3), 380.

[43] Mastroberardino, S., Santangelo, V., Botta, F., Marucci, F. S., & Olivetti Belardinelli, M. (2008). How the bimodal format of presentation affects working memory: an overview. *Cogn Process*, 9(1), 69-76.

[44] 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. *Sci Rep*, 7(1), 5112.

[45] 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.

[46] Schneider, B. A., & Pichora-Fuller, M. K. (2000). Implications of perceptual deterioration for cognitive aging research.

[47] Frtusova, J. B., & Phillips, N. A. (2016). The Auditory-Visual Speech Benefit on Working Memory in Older Adults with Hearing Impairment. *Front Psychol*, 7, 490.

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

[49] Versace, R., Vallet, G. T., Riou, B., Lesourd, M., Labeye, É., & Brunel, L. (2014). Act-in: An integrated view of memory mechanisms. *Journal of Cognitive Psychology*, 26, 280–306.

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

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

[52] Brunetti, R., Indraccolo, A., Mastroberardino, S., Spence, C., & Santangelo, V. (2017). The impact of cross-modal correspondences on working memory performance. *Journal of Experimental Psychology: Human Perception and Performance*, 43(4), 819–831.

[53] 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 Sci*, 11(9).

[54] Liu, J., Zhang, H., Yu, T., Ren, L., Ni, D., Yang, Q., . . . Xue, G. (2021). Transformative neural representations support long-term episodic memory. *Sci Adv*, 7(41), eabg9715.

[55] 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.

[56] Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: the Colavita effect revisited. *Percept Psychophys*, 69(5), 673-686.

[57] Mastroberardino, S., Santangelo, V., & Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Front Integr Neurosci*, 9, 45.

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

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

[60] 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. *Behav Res Methods*, 39(2), 175-191.

[61] 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.

[62] 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.

[63] Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychol Bull*, 114(3), 510-532.

[64] Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behav Res Methods*, 44(1), 158-175.

[65] Bigelow, J., & Poremba, A. (2016). Audiovisual integration facilitates monkeys' short-term memory. *Anim Cogn*, 19(4), 799-811.

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

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

[68] Downing, P. E. (2000). Interactions between visual working memory and selective attention. *Psychol Sci*, 11(6), 467-473.

[69] Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nat Rev Neurosci*, 18(1), 42-55.

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

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

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

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

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

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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. *Hum Brain Mapp*, 36(12), 5003-5017.

[83] Talsma, D. (2015). Predictive coding and multisensory integration: An attentional account of themultisensory mind. *Frontiers in Integrative Neuroscience*, 9(MAR), 1–13.

[84] 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', *Front Neurosci*, 10, pp. 443.

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

[86] Baddeley, A. (1992). Working memory. *Science*, 255(5044), 556-559.

[87] Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends Cogn Sci*, 16(2), 129-135.

[88] Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1997). Transient and sustained activity in a distributed neural system for human working memory. *Nature*, 386(6625), 608-611.

[89] Delogu, F., Raffone, A., & Belardinelli, M. O. (2009). Semantic encoding in working memory: is there a (multi)modality effect? *Memory*, 17(6), 655-663.

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

[91] Mishra, J., & Gazzaley, A. (2012). Attention distributed across sensory modalities enhances perceptual performance. *J Neurosci*, 32(35), 12294-12302.

[92] Allen, R. J., Hitch, G. J., & Baddeley, A. D. (2009). Cross-modal binding and working memory. *Visual Cognition*, 17(1-2), 83-102.

[93] Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology Section A*, 49(1), 5-28.

[94] Loose, R., Kaufmann, C., Auer, D. P., & Lange, K. W. (2003). Human prefrontal and sensory cortical activity during divided attention tasks. *Hum Brain Mapp*, 18(4), 249-259.

[95] D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature*, 378(6554), 279-281.

[96] Della Sala, S., Baddeley, A., Papagno, C., & Spinnler, H. (1995). Dual-task paradigm: a means to examine the central executive. *Ann N Y Acad Sci*, 769, 161-171.

[97] Uncapher, M. R., Hutchinson, J. B., & Wagner, A. D. (2011). Dissociable effects of top-down and bottom-up attention during episodic encoding. *J Neurosci*, 31(35), 12613-12628.

[98] Uncapher, M. R., & Rugg, M. D. (2005). Effects of divided attention on fMRI correlates of memory encoding. *J Cogn Neurosci*, 17(12), 1923-1935.

[99] Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nat Rev Neurosci*, 14(10), 693-707.

[100] Cohen, M. A., Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). Auditory and visual memory in musicians and nonmusicians. *Psychon Bull Rev*, 18(3), 586-591.

[101] Gao, Z., Wu, F., Qiu, F., He, K., Yang, Y., & Shen, M. (2017). Bindings in working memory: The role of object-based attention. *Atten Percept Psychophys*, 79(2), 533-552.

[102] Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *Proc Natl Acad Sci U S A*, 106(14), 6008-6010.

[103] Heikkila, J., Fagerlund, P., & Tiippana, K. (2018). Semantically Congruent Visual Information Can Improve Auditory Recognition Memory in Older Adults. *Multisens Res*, 31(3-4), 213-225.

[104] Tatz, J. R., Undorf, M., & Peynircioglu, Z. F. (2020). Effect of impoverished information on multisensory integration in judgments of learning. *J Exp Psychol Learn Mem Cogn*.

[105] Tulving, E. (1981). Similarity relations in recognition. Journal of Verbal Learning and Verbal *Behavior*, 20, 479–496.

[106] Santangelo, V., & Macaluso, E. (2013). The contribution of working memory to divided attention. *Hum Brain Mapp*, 34(1), 158-175.

[107] Simon, S. S., Tusch, E. S., Holcomb, P. J., & Daffner, K. R. (2016). Increasing working memory load reduces processing of cross-modal task-irrelevant stimuli even after controlling for task difficulty and executive capacity. *Frontiers in human neuroscience*, 10, 380.

[108] Loose, R., Kaufmann, C., Auer, D. P., & Lange, K. W. (2003). Human prefrontal

and sensory cortical activity during divided attention tasks. *Hum Brain Mapp*, 18(4), 249-259.

[109] Nikolin, S., Lauf, S., Loo, C. K., & Martin, D. (2018). Effects of High-Definition Transcranial Direct Current Stimulation (HD-tDCS) of the Intraparietal Sulcus and Dorsolateral Prefrontal Cortex on Working Memory and Divided Attention. *Front Integr Neurosci*, 12, 64.

[110] Frankland, P.W., Josselyn, S.A., & Kohler, S. (2019). The neurobiological foundation of memory retrieval. *Nature Neuroscience*, 22, 1576–1585.

[111] Slotnick, S. D. (2004). Visual memory and visual perception recruit common neural substrates. *Behavioral and Cognitive Neuroscience Reviews*, 3, 207–221

[112] Kent, C., & Lamberts, K. (2008). The encoding-retrieval relationship: retrieval as mental simulation. *Trends in Cognitive Sciences*, 12, 92–98.

[113] Wais, P. E., & Gazzaley, A. (2011). The impact of auditory distraction on retrieval of visual memories. *Psychonomic Bulletin & Review*, 18(6), 1090–1097.

[114] Philippi, T. G. (2012). Benefits of multisensory presentation on perception, memory and navigation. Phd dissertation, Utrecht University.

[115] Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88, 638–667.

[116] Suchan, B., Linnewerth, B., Köster, O., Daum, I., & Schmid, G. (2006). Cross-modal processing in auditory and visual working memory. *Neuroimage*, 29(3), 853–858.

[117] Naghavi, H. R., & Nyberg, L. (2005). Common fronto-parietal activity in attention, memory, and consciousness: Shared demands on integration? *Consciousness and Cognition*, 14(2), 390–425.

[118] Nyberg, L., Habib, R., McIntosh, A. R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proc Natl Acad Sci U S A, 97*(20), 11120-11124.

[119] Saunders, N. L., & Summers, M. J. (2011). Longitudinal deficits to attention, executive, and working memory in subtypes of mild cognitive impairment. Neuropsychology, 25(2), 237-248. doi: 10.1037/a0021134