Proceedings of the 7th International Conference on Education Innovation and Philosophical Inquiries
DOI: 10.54254/2753-7048/2026.34105

# *The Visual-Auditory Dominance and Diverse Mechanisms in Multimodal Memory*

**Xingwan Bai**

*Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, China*
*t330016001@mail.bnbu.edu.cn*

***Abstract.*** As the two main perceptual modes in human cognition, vision and hearing have always been the core content of memory research. The development of experimental techniques and paradigms in cognitive science has led to a deeper understanding of memory in both visual and auditory channels in recent years. However, there is a lack of summary on the different conditions and interaction mechanisms of the dominant modes of memory in these two modalities in long-term memory and working memory. Based on recent studies and some classic literature, this paper uses a literature review to fill the research gap. The review found that visual and auditory working memory displayed a variable pattern where auditory perception predominates. That is, the dominant mode of perception changes according to the task requirements and the characteristics of the stimuli. The interaction mode of multimodal working memory was regulated by task type, attention allocation, and cognitive load. The interaction of different sensory channels in multimodal long-term memory was influenced by semantic consistency and precise identity matching.

***Keywords:*** Visual Dominance, Auditory Dominance, Working Memory, Long-Term Memory, Multimodal Interaction

## 1. Introduction

Vision and hearing, as the two most frequently utilized sensory and memory channels in human daily life, have received numerous academic attentions in the field of cognitive psychology. Research has shown that there are significant differences between the two memory channels of auditory and visual in terms of memory capacity and recognition accuracy rate. Among them, vision holds a dominant position [1, 2].

With the updates in experimental paradigms and theories in recent years, the studies of visual and auditory memory can now examine the differences and interactions between the two memory channels from a more detailed perspective. The review by Matusz et al. indicates that the interaction between visual and auditory cross-modal memory is related to semantics and cross-sensory asymmetry, and there may be multi-sensory integrated memory representations [3]. However, there is still a lack of a systematic discussion on the differences in the conditions under which the visual and auditory channels play a dominant role in memory, as well as the potential mechanisms of multimodal interaction. Furthermore, there remain scarce studies to systematically summarize the differentiation and integration patterns of multimodal channels in both working memory and long-

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2026.34105

term memory. This article intends to review some related literature to analyze and discuss the dominant role and internal mechanism of the visual and auditory channels in working memory and long-term memory, with the aim of inspiring future cognitive experiments and practical applications in related fields.

## 2. Auditory dominance and dynamic cross-modal mechanism in working memory

### 2.1. Auditory dominance

The sensory input mode of memory has long been widely explained by the visual-dominated theory, which states that humans have a tendency to preferentially use and allocate attention to the visual channel for information acquisition [4]. In the study published by Colavita, through an experimental paradigm involving multimodal presentation and responses, it was demonstrated that the visual pathway holds an absolute priority in the process of information processing [4]. When both visual and auditory stimuli were presented simultaneously, participants tended to react to the visual stimulus, and numerous portions of them were not even aware of the existence of the auditory stimulus altogether. This perceptual advantage of the visual channel is deemed to extend into working memory because attention supports both the processes of encoding and maintenance.

However, recent studies have challenged this traditional view by demonstrating that modality dominance varies in different task conditions and discovered that the dominant role of the visual and auditory senses in working memory is far more complicated than originally assumed. In certain circumstances, the temporal nature of sound can make it more dominant in working memory than vision.

Due to the high temporal nature of auditory perception, after eliminating individual differences based on the subjects' subjective perception, the experiment found that when visual and auditory stimuli were semantically incongruent, meaning that these two stimuli referred to different categories and auditory perception occurred before visual perception, the visual advantage difference reduced [5]. The visual advantage was primarily weakened by accelerating the retrieval of auditory memory, with the high temporal resolution and earlier onset of auditory stimuli further facilitating attention allocation. Therefore, visual superiority is not absolute in working memory; rather, it is a complex process influenced by multiple factors.

Uluç et al.'s research results demonstrate the potential conditions that visual superiority often encounters in laboratory perception experiments [6]. After eliminating the differences in the informational amount of visual and auditory stimuli, individual sensitivity, and task paradigm to control the differences in the extraction methods of the two modalities of memory, Uluç et al.'s research found that in cross-sensory competition of working memory, auditory is superior to visual, both during the input stage and the output stage [6]. This indicates that the visual superiority results discovered in previous perception experiments might be due to the mismatch between the auditory and visual stimuli and the difficulty of the tasks, rather than the inherent differences in auditory and visual memory.

Huang et al. only found the weakening effect of hearing, but in fact, vision still holds an absolute dominant position in working memory [5]. Nevertheless, Uluç et al. discovered after eliminating all mismatch and interference conditions that hearing is more advantageous in working memory [6]. The majority of the previous studies' differences stemmed from the fact that semantic cues could enhance visual encoding by facilitating conceptualization. In Uluç et al. 's experiment, the use of pure meaningless stimuli for hearing did not get affected by semantic conditions, while the stimulus presentation in Huang et al. 's experiment was more natural and daily-like [5, 6]. Visual perception is

DOI: 10.54254/2753-7048/2026.34105

more influenced by semantics. Thus, when it is impossible to simultaneously remove the semantics of both stimuli, it is still the visual perception that dominates working memory.

## 2.2. Dynamic cross-modal asymmetry and competitive interference

The cross-modal relationship between auditory and visual information in working memory is more complex than just dominance. The literature indicates that this relationship is highly influenced by the level of cognitive load, the asymmetry of attention resource allocation, and the modality of the task.

This cross-modal audio-visual influence mechanism is primarily related to the level of cognitive load. The dominance of the visual channel in working memory is not fixed. In the experiment, the intensity of cognitive load was adjusted by manipulating the type of stimuli, which was different from the classical experimental paradigm of simply increasing item number, which may mix cognitive load with capacity limits. The low-load condition involves alphanumeric stimuli, while the high-load condition features complex dashboard images that require heavier demands on visual-spatial analysis and pattern recognition [7]. This paradigm is more in line with the cognitive tasks encountered in real life. The result shows that different interference mechanisms will be adopted in the visual and auditory channels depending on the level of cognitive load. When the cognitive load is low, vision still dominates, and the interference effect of vision on hearing is much greater than that of hearing on vision, that is, a unidirectional asymmetric interference pattern. However, in the case of high cognitive load, shrunk executive resources may prevent either modality from maintaining a clear advantage. At this point, the dominant roles of the visual and auditory senses tend to be equal.

In another study, the attention allocation mechanism and the asymmetry of the interference effect in visual and auditory working memory were significantly emphasized [8]. Excluding the influence of perceptual expectations from sensory input, the experiment focused on the attentional resource allocation during the retention stage of working memory as the key mechanism. It turned out that, in terms of the reliance on attention resources, vision is much higher than hearing, while the auditory channel of working memory is more dependent on feature-based memory, which refers to relatively low-effort representations that can be maintained without continuous attention. As a result, when attention is impaired, the part most affected is visual working memory. This finding precisely explains why, in the study by He et al., the dominance of visual working memory decreased after the cognitive load increased [7]. As the cognitive load increases, the available attentional resources decrease. Therefore, the visual channel was more impaired while the auditory channel was less affected, thereby weakening the dominant position of vision in working memory.

Both studies have pointed to the conclusion that audio and visual information interfere with each other, even though the degree of interference is asymmetric [7, 8]. Nevertheless, the experiment conducted by Maezawa & Kawahara indicates that in terms of spatial information, the working memory of the visual and auditory channels shares a set of processing systems [9]. During the experiment, the direction cues of the space mission were switched between auditory and visual channels, but this did not result in the typical multimodal switching cost observed in task performance. When it comes to spatial tasks, the working memory channels for vision and hearing are continuous rather than mutually exclusive. This is contrary to the viewpoint presented in Li & Cowan, which explains the mechanism of the two sensory channels of working memory [8]. This might be because the perceptual patterns employed in space missions are not one-way, so the multimodal working memory invoked is not competitive but rather mutually supportive.

DOI: 10.54254/2753-7048/2026.34105

## 3. Long-term memory

### 3.1. Absolute visual dominance in long-term memory

In the field of long-term memory through visual and auditory channels, the prevailing view is still that the visual channel holds an absolute advantage. Most studies focus on the specific mechanisms of different long-term memory patterns of two channels.

One type of experiment on long-term memory involves comparing the memory of basic visual stimuli with auditory stimuli. A study employed a paradigm that combined images, sounds and semantics, aiming to compare the explanatory power of the fuzzy trace theory, the dual coding theory, the physical uniqueness hypothesis and the conceptual uniqueness hypothesis, and to explore the underlying mechanism for why visual memory is generally superior to auditory memory [10]. In the six experiments conducted, experiments 1A and 1B did not incorporate multimodal and semantic elements. Instead, they differentiate the influence of gist and verbatim traces through yes-no recognition and two-alternative forced-choice questions, and disassociate Gist Traces and Verbatim Traces, enabling them to be separately explored in the Exemplar Test and Novel Test paradigms. Experiments 2A and 2B incorporated visual semantic cues into the sounds to test the dual-coding theory, and finally supplemented with the subjective ratings of the participants' behaviors. The final results show that the visual long-term memory of pictures outperforms the auditory long-term memory in both the processing of concepts and the processing of perceptual details.

Another type of long-term memory that is frequently studied is scene memory. Scene memory, as a form of dynamic memory, naturally integrates audio and visual elements into a single entity, which triggers a more dynamic pattern of visual-auditory bimodal memory. In a study, 50 movie clips were used as the experimental materials [11]. With testing procedures held the same across different conditions, the study manipulated the duration and modality of stimuli during learning to isolate their respective effects on subsequent memory retrieval. The results show that even though the dual-modal long-term memory achieved by combining both visual and auditory information is the most effective, the single visual channel still outperforms the auditory channel.

In another study by the same first author, through modeling analysis and manipulation of the dual-modal and single-modal stimuli and extraction processes, a more comprehensive separation exploration of the audio-visual channels of long-term memory was conducted [12]. In the first three experiments, the visual channel performed significantly better in terms of memory. Furthermore, in Experiment 4, to eliminate the interfering variable of visual memory accuracy, the visual images were pixelated and degraded in quality. After such processing, the results still show that vision holds an absolute advantage in long-term memory, indicating that this is not due to the accuracy of the stimulus input through the visual channel. Besides, in the process of dual-modal extraction, the interference received by the visual and auditory systems is also asymmetric, which is similar to the results of the working memory research mentioned earlier [7, 8]. However, these studies on long-term memory have not further explored the underlying mechanisms of this asymmetric interference during scene memory retrieval—a gap that deserves future investigation, particularly regarding whether this asymmetry reflects differential prioritization of attentional allocation or distinct neural activity patterns for visual and auditory processing.

### 3.2. Multimodal long-term memory enhancement

In addition to focusing on visual advantages, just like in the case of working memory, researchers studying long-term memory also strive to explore the facilitating effect of multimodal channels on

DOI: 10.54254/2753-7048/2026.34105

long-term memory. Nevertheless, whether multimodal channels have a promoting effect on long-term memory remains controversial.

In the aforementioned study by Meyerhoff & Huff, various manipulations such as semantic mismatch, backward playback, asynchronous continuous playback, and random interleaved playback were employed in this study to manipulate the semantic consistency and temporal synchrony of the audio-visual stimuli [11]. The results show that the semantic consistency of visual and auditory cues has an enhancing effect on long-term memory, and this effect remains valid even under sequential time delays. This kind of semantic-consistent bimodal long-term memory enhancement is somewhat similar to the research results from Ahmad et al., but the two are completely different in terms of the types of stimuli and the natural degree of audio-visual combination [10]. This indicates that working memory capacity may serve as an essential mediator in the integration of bimodal stimuli during long-term memory encoding, as the blending of semantically congruent audio-visual information likely places demands on limited-capacity cognitive resources before solidification of memory can occur.

Besides, Duarte et al. used the accidental encoding paradigm, which did not inform the participants in advance that there would be a memory test, as well as tests for the recollection of familiarity and sound details, to investigate whether consistent sound in semantic enhance recognition via recollection or familiarity [13]. The results show that even if the task does not require remembering the sounds, sounds that are semantically consistent with the visual stimuli still significantly enhance the accuracy of recall and recognition. The author explained this result using the theory of predictive coding, stating that when the semantics are consistent, the prediction error is smaller. This saves the consumption of cognitive resources when feeling surprised, thus making it easier for visual and auditory information to be integrated and encoded together.

However, a major challenge regarding the idea that combining audio and visual elements can enhance long-term memory comes from the face-sound matching paradigm. Smith et al. employed three within-subject design conditions, integrating single-modal matching, full multimodal matching, and cross-modal matching within the facial-voice matching paradigm [14]. The final result shows that the long-term memory accuracy of dual-modal matching stimuli involving both sound and face does not improve the accuracy compared to that of single-modal matching. Nevertheless, the pairing of faces and voices cannot be compared with the pairing of object semantic consistency. For familiar objects, humans already associate their sounds with their visual representations; however, such cross-modal pairing is largely limited to the species level and cannot be proficiently refined to the individual-level identification, particularly for unfamiliar persons. This means that humans are still unable to accurately match the voices and facial images of unfamiliar specific individuals. This is the key reason why results from these two studies are so different [13, 14].

## 4. Discussion and suggestion

Research on working memory and long-term memory shows certain similarities and differences in terms of the dominance of the visual and auditory channels and the multimodal interaction. Therefore, the application of their research results involves both intersections and divergences.

Working memory is of great significance in the field of education, as it is directly related to the level of immediate acquisition of knowledge. In the process of language ability learning, the accuracy of auditory working memory is of crucial importance. It affects students' immediate response ability, which is different from the delayed response in visual learning. Huang et al. indicate that the sequence of visual and auditory presentation helps enhance the dominance of

DOI: 10.54254/2753-7048/2026.34105

auditory working memory [5]. This provides promising insights for listening instruction, namely that the teaching model of first obtaining auditory information and then presenting the corresponding text for listening training may more effectively open auditory working memory channels and improve listening comprehension abilities. The research findings of Uluç et al. prompt test developers to redesign the cognitive demands of listening tests appropriately in a standardized manner, particularly with respect to the concurrent processing burdens imposed by question-reading that may disturb auditory working memory encoding [6]. Furthermore, visual working memory is susceptible to the influence of attention allocation and cognitive load [7, 8]. Therefore, taking into account the students' fatigue levels, it is necessary to flexibly design teaching plans that are mainly visual or mainly auditory.

On the other hand, the implications of long-term memory for criminology are particularly noticeable. Whether it is static image memory or dynamic scene memory, the visual advantage is something that cannot be ignored. Whether it is static image memory or dynamic scene memory, under the multimodal condition, the advantage of visual memory over auditory memory is undeniable [11, 13, 14]. Therefore, in order to ensure the accuracy of the testimony, the court still needs to give more consideration to using eyewitnesses rather than hearsay evidence.

In addition, a series of studies on multimodal interference in long-term memory have shown that auditory information has a facilitating effect on visual long-term memory when it comes to the semantic consistency level [11, 13]. However, such a facilitating effect is insignificant in the task of face-sound matching [14]. The voiceprints required for identity recognition are often without semantics and are difficult to accurately match by strangers. Therefore, for witnesses, the investigation strategy might need to be re-focused on identifying the relevant items related to the crime or the visual details of specific scenes, rather than mainly relying on matching the voices and faces of the suspects – this approach fully utilizes the powerful ability of visual long-term memory, while avoiding the limitations of cross-modal individual-level recognition. Future research could fruitfully explore the performance of this object-centered recognition scheme in stressful environments or situations with time delays, as these are common features in real-world witnessing scenarios.

## 5. Conclusion

This paper focuses on analyzing the dominant role of the visual-auditory memory channel and the multimodal interaction from the perspectives of working memory and long-term memory. The results show that in working memory, a variable auditory dominance pattern is presented. In contrast, long-term memory exhibits an absolute visual dominance pattern. Moreover, the inhibitory and promoting integration methods of multimodal visual-auditory working memory also vary depending on task type, cognitive load, and attention allocation mechanism. The long-term memory of multimodal information is influenced by semantic consistency and identity recognition. This article synthesizes the research on working memory and long-term memory, revealing the dominance of the visual and auditory channels in both types of memory and the modality interactivity, thereby filling the gaps that existed in the existing literature when separately studying these two memory systems. Based on the analysis results of the literature, this paper finally proposes some possible future research directions and practical application suggestions in the fields of education and criminal investigation.

DOI: 10.54254/2753-7048/2026.34105

## References

[1] Standing, L. (1973). Learning 10, 000 pictures. The Quarterly Journal of Experimental Psychology, 25(2), 207–222.
[2] Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. Proceedings of the National Academy of Sciences of the United States of America, 106(14), 6008–6010.
[3] Matusz, P. J., Wallace, M. T., & Murray, M. M. (2017). A multisensory perspective on object memory. Neuropsychologia, 105, 243–252.
[4] Colavita, F. B. (1974). Human sensory dominance. Perception & Psychophysics, 16(2), 409–412.
[5] Huang, J., Wang, A., & Zhang, M. (2024). The audiovisual competition effect induced by temporal asynchronous encoding weakened the visual dominance in working memory retrieval. Memory, 32(8), 1069–1082.
[6] Uluç, I., Turpin, T., Kotlarz, P., Lankinen, K., Mamashli, F., & Ahveninen, J. (2024). Comparing auditory and visual aspects of multisensory working memory using bimodally matched feature patterns. Experimental Brain Research, 243(1), Article 38.
[7] He, Y., Yang, T., Zhang, Y., Sun, K., Guo, Q., Chen, Q., Wang, X., Xu, X., Wei, P., Wu, S., & Xu, T. (2025). Exploring the effects of audiovisual incongruence on working memory performance in the combined 2-back + Go/NoGo paradigm. Frontiers in Psychology, 16, 1578391.
[8] Li, Y., & Cowan, N. (2021). Attention effects in working memory that are asymmetric across sensory modalities. Memory & Cognition, 49(5), 1050–1065.
[9] Maezawa, T., & Kawahara, J. I. (2021). Commonalities of visual and auditory working memory in a spatial-updating task. Memory & Cognition, 49(6), 1172–1187.
[10] Ahmad, F. N., Tremblay, S., Karkuszewski, M. D., Alvi, M., & Hockley, W. E. (2024). A conceptual-perceptual distinctiveness processing account of the superior recognition memory of pictures over environmental sounds. Quarterly Journal of Experimental Psychology, 77(7), 1555–1580.
[11] Meyerhoff, H. S., & Huff, M. (2016). Semantic congruency but not temporal synchrony enhances long-term memory performance for audio-visual scenes. Memory & Cognition, 44(3), 390–402.
[12] Meyerhoff, H. S., Jaggy, O., Papenmeier, F., & Huff, M. (2023). Long-term memory representations for audio-visual scenes. Memory & Cognition, 51(2), 349–370.
[13] Duarte, S. E., Ghetti, S., & Geng, J. J. (2023). Object memory is multisensory: Task-irrelevant sounds improve recollection. Psychonomic Bulletin & Review, 30(2), 652–665.
[14] Smith, H. M. J., Ritchie, K. L., Baguley, T. S., & Lavan, N. (2025). Face and voice identity matching accuracy is not improved by multimodal identity information. British Journal of Psychology, 116(2), 367–385.